UNIVERSITY OF CALIFORNIA
SANTA CRUZ

**TRANSCRIPTIONAL SIGNATURES OF THE TUMOR AND THE TUMOR MICROENVIRONMENT PREDICT CANCER PATIENT OUTCOMES.**

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

**Verena Friedl**

September 2021

The Dissertation of Verena Friedl
is approved:

Professor Joshua M. Stuart, Chair

Professor Olena Morozova Vaske

Professor Christopher C. Benz

Peter Biehl
Vice Provost and Dean of Graduate Studies

# Table of Contents

# List of Figures

# List of Tables

## Abstract

Transcriptional Signatures of the Tumor and the Tumor Microenvironment Predict Cancer Patient Outcomes.

by

Verena Friedl

Predicting the most effective cancer therapy for patients is a challenging yet very important task. In my doctoral thesis, I describe new insights gained from using transcriptional signatures from gene expression data of tumors and the tumor microenvironment. Out of different multi-omics data types, gene expression is found to be the most useful in predicting cancer drug sensitivity in a data set of cancer cell lines. Gene expression data can also be used to predict the presence of cancer driver events, genetic abnormalities responsible for tumor growth and progression. I describe the detection of rare genomic driver events found by association with known driver events using transcriptional signatures.

Cancers are traditionally classified into types and subtypes by the organ- and cell-of-origin. However, more and more cancer subtypes are now being defined on a molecular basis using for example gene expression or mutation data. I perform a meta-analysis of molecular subtype classifiers for 26 different cancer cohorts that demonstrates which aspects of the input samples and input data are important to build an accurate molecular subtype classifier. In advanced prostate cancer, I use transcriptional signatures to reliably classify samples into subtypes. The gene expression data, in combination with histological review, is able to define the

most at-risk patients in this cohort. However, I show that the classification of cancers into distinct subtypes is not applicable to all samples in this cohort, because they exist on a continuous spectrum between the subtypes - a finding that I was able to recapitulate in a study of lung cancer samples. I define and describe this continuum between subtypes of advanced prostate cancer using gene expression data.

The tumor microenvironment, the normal cells that mix with cancer cells to form a tumor, plays an important role in cancer progression and treatment response. I built a landscape of about 10,000 cancer samples from immune cell infiltration estimated by deconvolution of gene expression profiles. However, immune cells are not the only cell types in the tumor microenvironment. I present a comprehensive deconvolution analysis of tumor patient samples using a cell type library defined from single-cell RNA sequencing data. These cell type estimates enable the detection of a pan-cancer high-risk sample group that is not detected by traditional gene expression analysis.

## Acknowledgments

I would like to thank the members of my thesis committee for their time and their leadership during my time at UCSC, especially my advisor Josh Stuart for his support and guidance. I would also like to thank every current and past member of the Stuart lab, it was a pleasure to work with every single one of you.

All chapters of this thesis are written about work I did in collaboration with other researchers. The project described in chapter 2 was led by Kiley Graim and is a partial and edited reprint of the publication in:

> Kiley Graim, Verena Friedl, Kathleen E. Houlahan, and Joshua M. Stuart. PLATYPUS: A Multiple-View Learning Predictive Framework for Cancer Drug Sensitivity Prediction. *Pacific Symposium on Biocomputing*, 24, 2019.

The project described in chapter 3 was led by David Haan and is a partial and edited reprint of the publication in:

> David Haan, Ruikang Tao, Verena Friedl, Ioannis N. Anastopoulos, Christopher K. Wong, Alana S. Weinstein, and Joshua M. Stuart. Using Transcriptional Signatures to Find Cancer Drivers with LURE. *Pacific Symposium on Biocomputing*, pages 343–354, 2020.

The work in chapter 4 is a contribution to the Tumor Tumor Molecular Pathology (TMP) Analysis Working Group (AWG) formed under the National Cancer Institutes Center for Cancer Genomics Genomic Data Analysis Network (GDAN), the research collaboration following the TCGA project. It will be published within the GDAN TMP AWG under the project lead of Andrew D. Cherniak and Peter W. Laird.

The projects described in chapters 5 and 6 were conducted as part of the research collaboration called the West Coast Dream Team (WCDT) from the Stand-Up-To-Cancer Prostate

Cancer Foundation under the leadership of Eric J. Small and Joshua M. Stuart. The work described in chapter 5 was led by Rahul Aggarwal and is a partial and edited reprint of the publication in:

> Rahul Aggarwal, Jiaoti Huang, Joshi J. Alumkal, Li Zhang, Felix Y. Feng, George V. Thomas, Alana S. Weinstein, Verena Friedl, Can Zhang, Owen N. Witte, Paul Lloyd, Martin Gleave, Christopher P. Evans, Jack Youngren, Tomasz M. Beer, Matthew Rettig, Christopher K. Wong, Lawrence True, Adam Foye, Denise Playdle, Charles J. Ryan, Primo Lara, Kim N. Chi, Vlado Uzunangelov, Artem Sokolov, Yulia Newton, Himisha Beltran, Francesca Demichelis, Mark A. Rubin, Joshua M. Stuart, and Eric J. Small. Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study. *Journal of Clinical Oncology*, 36(24):2492, 2018.

In chapter 5, I also describe parts of three additional WCDT studies and a TCGA (The Cancer Genome Atlas) study of lung adenocarcinomas to which I contributed:

1. Joshi J. Alumkal, Duanchen Sun, Eric Lu, Tomasz M. Beer, George V. Thomas, Emile Latour, Rahul Aggarwal, Jeremy Cetnar, Charles J. Ryan, Shaadi Tabatabaei, Shawna Bailey, Claire B. Turina, David A. Quigley, Xiangnan Guan, Adam Foye, Jack F. Youngren, Joshua Urrutia, Jiaoti Huang, Alana S. Weinstein, Verena Friedl, Matthew Rettig, Robert E. Reiter, Daniel E. Spratt, Martin Gleave, Christopher P. Evans, Joshua M. Stuart, Yiyi Chen, Felix Y. Feng, Eric J. Small, Owen N. Witte, and Zheng Xia. Transcriptional profiling identifies an androgen receptor activity-low, stemness program associated with enzalutamide resistance. *Proceedings of the National Academy of Sciences*, 117(22):12315–12323, 2020.

2. Rahul Aggarwal, Gustavo Rubio Romero, Verena Friedl, Alana S. Weinstein, Adam Foye, Jiaoti Huang, Felix Feng, Joshua M. Stuart, and Eric J. Small. Clinical and genomic characterization of Low PSA Secretors: a unique subset of metastatic castration resistant prostate cancer. *Prostate cancer and prostatic diseases*, pages 1–7, 2020.

3. Daniel H. Kwon, Li Zhang, David A. Quigley, Adam Foye, William S. Chen, Christopher K. Wong, Felix Y. Feng, Adina Bailey, Jiaoti Huang, Joshua M. Stuart, Verena

Friedl, Alana S. Weinstein, Tomasz M. Beer, Joshi J. Alumkal, Matthew Rettig, Martin Gleave, Primo N. Lara, George V. Thomas, Patricia Li, Austin Lui, Eric J. Small, and Rahul R. Aggarwal. Down-regulation of ADRB2 expression is associated with small cell neuroendocrine prostate cancer and adverse clinical outcomes in castration-resistant prostate cancer. *Urologic Oncology: Seminars and Original Investigations*, 38(12):931.e9–931.e16,2020.

4. Jian Carrot-Zhang, Xiaotong Yao, Siddhartha Devarakonda, Aditya Deshpande, Jeffrey S. Damrauer, Tiago Chedraoui Silva, Christopher K Wong, Hyo Young Choi, Ina Felau, Gordon A. Robertson, et al. Whole-genome characterization of lung adenocarcinomas lacking the RTK/RAS/RAF pathway. *Cell reports*, 34(5):108707, 2021.

The work described in chapter 6 will be published within the WCDT collaboration under the project lead of Eric J. Small and Joshua M. Stuart.

The work in chapter 7.1 is a contribution to the Tumor Deconvolution and Immunogenicity (TDI) Analysis Working Group (AWG) formed under the National Cancer Institutes Center for Cancer Genomics Genomic Data Analysis Network (GDAN), the research collaboration following the TCGA project. It will be published within the GDAN TDI AWG under the project lead of Bhavneet Bhinder, Paul Spellman, and Olivier Elemento.

The project described in chapter 7.2 is a collaboration with fellow graduate student Yuanqing Xue and other members of the Stuart lab and is currently in preparation for publication.

# Chapter 1

# Motivation and Introduction

The World Health Organization reports cancer as the second leading cause of death globally. This year, 2021, 1,898,160 new cases of cancer and 608,570 deaths from cancer are projected to occur in the United States alone [65]. Even though mortality rates for some cancers are improving, for others they remain unchanged.

Cancer therapy is often very aggressive and has many side effects, but does not always cure the disease or help patients. Therefore, it is an important challenge to predict the most effective therapy for each patient, weighing the benefits and risks. Traditionally, this is done by creating a standard of care treatment plan that, on average, promises the best outcome for certain cancer types and cancer subtypes. Cancer types and subtypes are usually classified by the organ and cell type the cancer originated from and how far the disease has already progressed. More recently, some cancer therapies were approved not for the treatment of a certain cancer type as defined by cell-of-origin, but for cancers with a certain molecular biomarker, irrespective of where in the body the cancer originated. Examples are Pembrolizumab approved in 2017 [44]

and Larotrectinib approved in 2018 [26].

Cancer is caused by changes in the genome and genetic cancer biomarkers describe specific genetic changes of the DNA that influence the overall behavior of the tumor cells. The altered genetic code of a cancer cell results in altered transcription of genes into messenger RNA (mRNA), called gene expression. The mRNAs are then translated into proteins and the altered protein levels ultimately determine cell behavior. RNA sequencing (RNA-Seq) is a laboratory technique that measures the expression level of genes by sequencing, mapping, and counting mRNA molecules in a biological sample. The gene expression can be used, with certain limitations, to estimate the phenotype of biological cells. The relatively easy measurement of mRNAs with high-throughput sequencing methods, and the biological implications of altered gene expression make RNA-Seq a widely used assay in cancer research.

In chapter 2, I present a project directly predicting cancer drug sensitivity from multi-omics data, like genetic alterations or gene expression, using the controlled environment of cancer cell lines. How successful drug sensitivity can be predicted, and which molecular data is useful in doing so, can present new insights on which molecular assays might be useful to consider when trying to choose the optimal therapy.

Genetic cancer biomarkers that are responsible for the cancer to start and continue growing are called cancer driver events. Finding cancer driver events can present an opportunity for therapy because it uncovers vulnerabilities to target. Genetic driver events, which can be e.g. mutations, copy number alterations, or gene fusions, often have a strong transcriptional signal measured in gene expression. I was involved in a project that uses transcriptional signatures of known driver events to find related but rare driver events for cancer samples that, until then, had

no or few driver events annotated (see chapter 3).

The recent advances in sequencing technology and molecular data generation led to an increased understanding of the molecular functioning of different tumors. A new way of grouping cancers on a molecular level has been found for many cancer types, and molecular subtypes that distinguish patients with different disease characteristics and survival outcomes have been defined. In chapter 4, I present a meta-analysis of molecular subtype classifiers for 26 different cancer types. I show that a balanced input sample set and subtypes that are well distinguishable in the data platform used for classification are the most important aspects for high model performance. A higher number of cancer subtypes presents a more complex prediction task and therefore reduces model performance, which, in our study, can not be improved by adding more samples to the training data.

Selective pressure by cancer treatment can influence the genetics and behavior of tumors. Together with clinical collaborators, I was involved in the description and analysis of a highly aggressive subtype of progressive metastatic castration-resistant prostate cancer (mCRPC) that predominantly emerges after treatment. This analysis is described in chapter 5.

In follow-up analyses, a third, intermediate subtype of mCRPC was discovered. I contributed to the description and analysis of these samples, resulting in the definition of a continuous spectrum between the subtypes of mCRPC (see chapter 6).

In my final chapter 7, I turn towards the tumor microenvironment: the normal cells that mix with cancer cells to form a tumor. The tumor microenvironment is often overlooked or even avoided in cancer sample analysis, for example by striving for a high tumor purity in biopsy samples [9]. However, not just the cancer cells are determining how tumors will develop,

progress, and respond to treatment, the cells in the tumor microenvironment play an important role too [79]. Certain cells in the tumor microenvironment promote tumorigenesis, whereas others were shown to keep cancer cells in check [34]. In section 7.1, I present an immune cell infiltration landscape for about 10,000 cancer tissue samples of 33 different cancer types. Section 7.2 describes a more comprehensive deconvolution of these cancer samples by using a large reference of cell types defined by single-cell RNA sequencing. Each cell type's influence on patient survival and disease progression is analyzed. A pan-cancer high-risk sample group emerges from this deconvolution approach, that is not detected by traditional gene expression analysis.

# Chapter 2

# Predicting Cancer Drug Sensitivity with a Semi-Supervised Learner

The project described in this chapter was conceptualized and led by Kiley Graim and was published in [30]. I was involved in the development and implementation of the iterative machine learning framework, especially the label-learning-validation and cross-validation processes and visualizations. I partially ran the experiments to optimize the method, measured and compared the performance of models, and investigated the most important features used by the models.

## 2.1 Introduction

Cancer treatment selection by screening compound libraries for a specific tumor sample is mostly prohibited by availability and cost. However, using genomic assays such as DNA and RNA sequencing in the clinic is on the rise. As the costs for these high-throughput assays

drop, applying genomic signatures from machine-learning trained on external data in place of the more expensive direct drug assay becomes an option. We present a multiple view learning (MVL) framework called PLATYPUS (**P**rogressive **LA**bel **T**raining b**Y** **P**redicting **U**nlabeled **S**amples), which learns such signatures that predict cancer drug sensitivity using multi-omics data.

PLATYPUS overcomes multiple obstacles that are often present in this learning task. 1) Clinical outcomes are expensive and time-consuming to collect and are therefore often missing. PLATYPUS uses an iterative semi-supervised learning framework that infers missing clinical outcomes and subsequently uses them to be able to train on more samples. 2) Methods that combine multiple data platforms often require all of those measurements to be present in a sample. But data sets are often incomplete and sample numbers considerably decrease when this requirement has to be met. PLATYPUS builds 'views' that are learned on a specific data type and are then combined in an ensemble while maximizing prediction agreement between multiple views. This way, PLATYPUS does not require all data types to be present for a sample, but it learns from all data types in cases where they are present. 3) Both points 1 and 2 help maximize sample usage for this low–sample, high–feature problem. This type of problem is typical in biological applications and can result in biologically irrelevant solutions [72]. Another way of improving on this issue is the incorporation of prior biological knowledge to guide feature selection for outcome prediction [38, 39, 62, 71]. PLATYPUS' views are able to directly incorporate biological priors, which reduces the number of features in a meaningful way and helps the interpretation of results.

## 2.2 Methods

### 2.2.1 Data

At the time of download the Cancer Cell Line Encyclopedia (CCLE) contained genomic, phenotype, clinical, and other annotation data for 1,037 cancer cell lines [12].

1. Drug sensitivity data were available for only 504 cell lines and 24 drugs. Drug response was converted to a binary label in order to transform the regression problem into a classification problem. For each compound, cell lines were divided into quartiles ranked by ActArea (the area over the dose-response curve); The bottom 25% were assigned to the 'non–sensitive' class and the top 25% to the 'sensitive' class. Cell lines lying in the middle were marked with 'intermediate' and considered unlabeled in this test.

2. CCLE includes mutation data in the form of Single Nucleotide Polymorphism (SNP) and insertion or deletion (Indel) events for 1,651 genes. SNPs and Indels were combined into a set of non-silent mutations that include all events changing the amino acid composition of the resulting protein, including Indels or missense SNPs in the coding region, splice site, and stop or start codon alterations.

3. CCLE gene expression data contains expression for about 18,900 genes.

4. CCLE Copy Number Variation (CNV) data covers 23,316 genes. Genes that have the same value for all cell lines lay on the same genome segment. These sets of genes were merged into one feature in order to reduce redundancy in the data, resulting in 20,247 features.

5. Clinical sample annotation data for the CCLE cell lines contain the biological sex of the cancer patient and information about the cancer origin, i.e. 24 different tissue types, 21 histology types, and 67 histology subtypes.

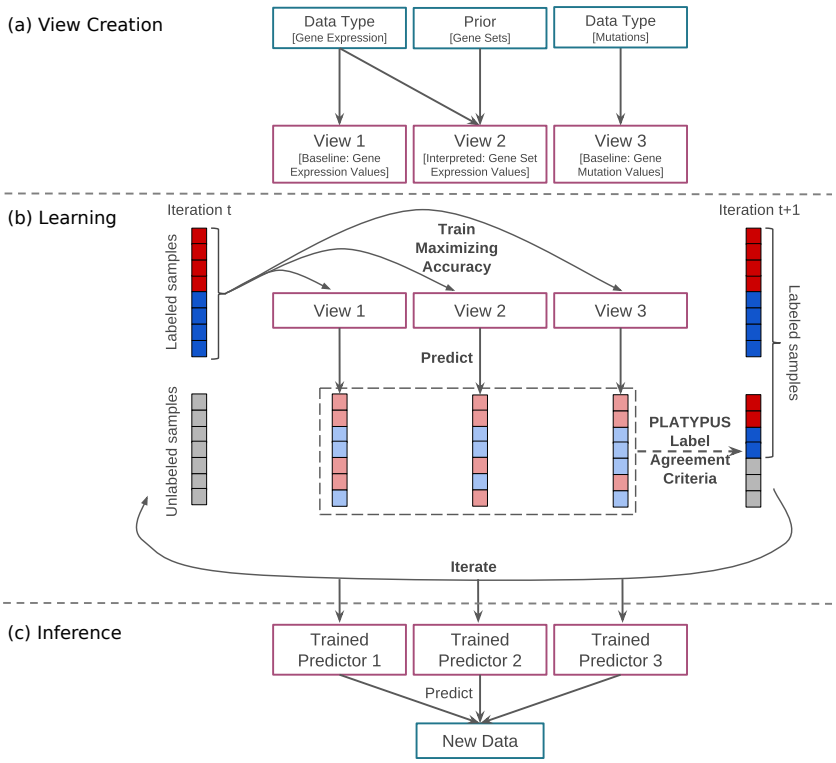


Figure 2.1: **PLATYPUS framework illustrated with three views. (a)** Creation of single views using sample data and optional prior knowledge. **(b)** Iterative Learning: Each view maximizes prediction accuracy on the labeled samples; unlabeled samples predicted with high confidence are added to the known sample set; repeat until no new samples are labeled. **(c)** Models from the final iteration of PLATYPUS training applied to new data.

### 2.2.2 Single views and co-training

PLATYPUS uses co-training (Fig. 2.1) between single views to learn labels for unlabeled samples. Single views are based on different feature sets. Genomic or clinical features can

be used directly (baseline views), or transformed using a biological prior (interpreted views). We built four baseline views from the CCLE data: expression, CNV, mutation, Sample- and Patient-Specific (SPS) information; and many interpreted views using e.g. biological gene sets like the Molecular Signatures Database (MSigDB) [66], drug target annotations, or inferred regulator activity from viper [8]. For each view, the machine learning algorithm and parameters can be optimized.

Co-training works by training a separate classifier using each view as a separate feature set to make independent predictions, then incorporating disagreement into the loss function. Each view trains on the labeled data then predicts labels for the common unlabeled set. High confidence labels are passed as truth in the next iteration. Co-training methods iterate until either convergence, some threshold (a minimal change in label definition on the unlabeled samples) is attained, or a maximum number of iterations is reached.

After co-training, each view can be used as a standalone classifier that incorporates learning from one or more data platforms without relying solely on that data platform. Since views are trained in conjunction, the trained models will incorporate the perspectives of all views. This also provides a measure of influence from all views when applying any of the classifiers to new data, without requiring data for those views when making predictions.

### 2.2.3 Maximizing agreement across views through label assignment

The key step in the PLATYPUS approach is the inference of outcome labels for a set of unlabeled data. Each training iteration seeks to improve the agreement of the assignments given to the unlabeled data across all views. Views are first created by applying machine learn-

ing methods using either the features directly, or from a transformation using a biological prior. Fig. 2.1 shows an overview of PLATYPUS using three views. Any number of views may be used - in this project, up to 10 views were used per experiment.

PLATYPUS searches iteratively for a label assignment that improves the agreement on unlabeled data (Fig. 2.1(b)). At each iteration $t$, the views are trained on labeled data and the labels for unlabeled samples are inferred. Because the set of labels can change across iterations, we denote the training data with sensitive labels as $T^+(t)$ and those with non-sensitive labels as $T^-(t)$ at iteration $t$. $T^+(0)$ and $T^-(0)$ are the given sets of sensitive and non-sensitive training samples before learning labels, respectively. The set of unlabeled samples is denoted $U(t)$, with all unlabeled samples before learning labels as $U(0)$.

$V$ is the set of views used in the PLATYPUS run. In iteration $t$, each view $v \in V$ is trained to maximize its prediction accuracy on the labeled samples $T^+(t)$ and $T^-(t)$. The accuracy of view $v$ at iteration $t$ is determined using cross-validation of the training samples and is written here as $a(v,t)$, where $a(v,0)$ is the single view accuracy before learning labels. A prediction is then made by the trained models for each unlabeled sample $s$. Let $l(v,s,t)$ be the prediction of sample $s$ by view $v$ in iteration $t$ where it is 1 if predicted sensitive and 0 otherwise. The single view votes are summarized to a sensitive ensemble vote $L^+(s,t)$ and non-sensitive ensemble vote $L^-(s,t)$ for each sample (Eq. 2.1 and 2.2).

$$L^+(s,t) = \sum_{v \in V^s} w(v,t)\, l(v,s,t) \tag{2.1}$$

$$L^-(s,t) = \sum_{v \in V^s} w(v,t)\, (1 - l(v,s,t)) \tag{2.2}$$

Only views with data to predict sample $s$ are taken into account: $V^s = \{v \in V :$

$v$ has data for $s\}$; and the different views are weighted by $w(v,t)$ (Eq. 2.3). View accuracies within $[0.5,1]$ are rescaled to $[0,1]$ and log-scaled. Views with an accuracy lower than 0.5 are given a weight of 0 since it indicates worse than random predictions.

$$w(v,t) = \begin{cases} -log(1-\frac{a(v,t)-0.5}{0.5}) & \text{if } a(v,t) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

To determine, which unlabeled samples are added to the training data for the next iteration, we define $L^{\max}(t)$, the strongest vote found between all samples in iteration $t$ (Eq. 2.4), and $\Psi(t)$, the set of samples reaching the strongest vote (Eq. 2.5).

$$L^{\max}(t) = \max_{s \in U(t)} \{\max\{L^+(s,t), L^-(s,t)\}\} \tag{2.4}$$

$$\Psi(t) = \{s \in U(t) : \max\{L^+(s,t), L^-(s,t)\} = L^{\max}(t)\} \tag{2.5}$$

In order to favor missing data for a sample over conflicting predictions, we define $L^{\min}(t)$ as $\min_{s \in \Psi(t)}\{\min\{L^+(s,t), L^-(s,t)\}\}$, the weakest contrary vote that is found between all samples in $\Psi(t)$.

All samples meeting both the strongest vote and the weakest contrary vote conditions (Label Agreement Criteria) build the set of new training samples $\mathcal{T}(t)$, which are added to $T^+(t)$ and $T^-(t)$ for the next iteration's training data:

$$\mathcal{T}(t) = \{s \in \Psi(t) : \min\{L^+(s,t), L^-(s,t)\} = L^{\min}(t)\} \tag{2.6}$$

$$T^+(t+1) = T^+(t) \cup \{s \in \mathcal{T}(t) : L^+(s,t) > L^-(s,t)\} \tag{2.7}$$

$$T^-(t+1) = T^-(t) \cup \{s \in \mathcal{T}(t) : L^+(s,t) < L^-(s,t)\} \tag{2.8}$$

To avoid adding predictions with low confidence, $L^{\max}(t)$ needs to stay above a certain value, otherwise no labels are added to the training data in iteration $t$. This can be adjusted by the learning threshold $\lambda$, which represents the fraction of the maximal reachable vote, i.e. when all views agree. By default $\lambda$ is 75%.

The training process continues until a convergence criterion is met: either all labels have been learned, no new labels have been learned in the last iteration, or a maximum number of iterations has been reached. After termination of the learning process, the trained single-view predictors can be used independently or as an ensemble via PLATYPUS (Fig. 2.1(c)).

## 2.3 Results

### 2.3.1 Preliminary experiments to optimize PLATYPUS performance

We ran 120 different PLATYPUS variants to predict drug sensitivity in the CCLE cell lines to identify the best way to combine the views for this application. As mentioned before, samples with intermediate levels of sensitivity for a particular drug were treated as unlabeled and used by the co-training to maximize agreement across views.

We first asked whether the interpretive views that use gene set information provide any benefit over using only the baseline views (Section 2.2.2). We then determined a weighting scheme for the ensemble to achieve better performance (Eq. 2.3). I ran PLATYPUS using the 4 baseline views and the 3, 5, 7, and 10 best-performing single views for each of the 24 CCLE drugs at a $\lambda = 75\%$ learning threshold, for a total of 120 different PLATYPUS variants (5 per drug). Fig. 2.2(a) shows the highest accuracy PLATYPUS models as well as each of

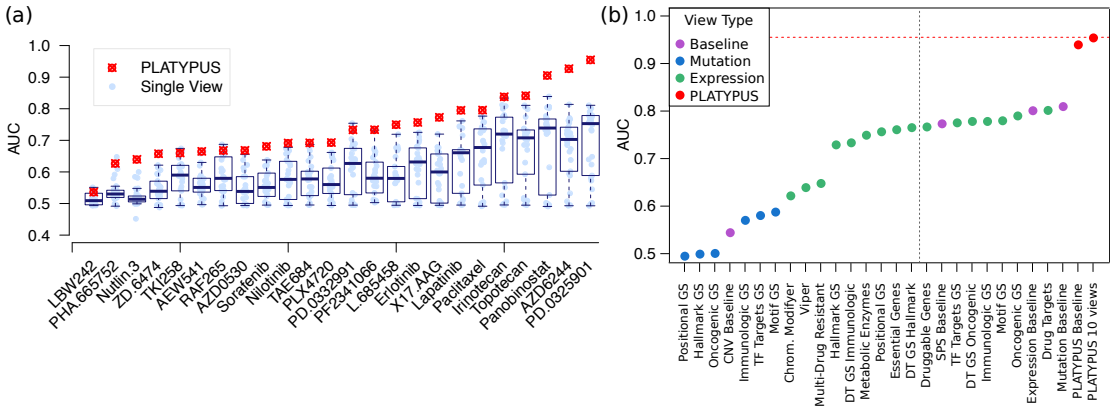


Figure 2.2: **PLATYPUS Performance. (a)** Boxplot showing performance (in AUC) sorted by PLATYPUS score, of all single views and the best PLATYPUS score. PLATYPUS score for each drug is the highest from the 3,5,7, and 10 view runs. **(b)** AUC for PD-0325901 sensitivity predictions for each single view, colored by view type. The 10 views to the right of the gray line are used in the PLATYPUS ensemble. See Fig. 2.5 for single view AUCs for all drugs. DT = Drug Target; GS = Gene Set.

the single view scores. In almost all cases PLATYPUS significantly outperforms single-view models, most notably for the MEK inhibitors AZD6244 and PD-0325901, and HDAC inhibitor Panobinostat. Adding interpreted views to PLATYPUS increased PD-0325901 AUC from 0.94 to 0.99 (Fig. 2.2(b)), motivating their continued inclusion in PLATYPUS models. Furthermore, within 10 iterations, most PLATYPUS runs added 90% or more of the unlabeled cell lines to the labeled set, effectively doubling the number of samples on which the models were trained. I look more closely at the results from the best overall performing PLATYPUS model, PD-0325901, as well as important features from each of its models, in Section 2.3.2.

Fig. 2.3 shows the process of label learning validation (LLV) for one example drug. Most of the drug models learn labels correctly. However, model AUC decreases once a model starts to learn labels incorrectly. Over many iterations, this can lead to a model where the majority of labels are learned incorrectly. We found that this risk can be minimized by setting a

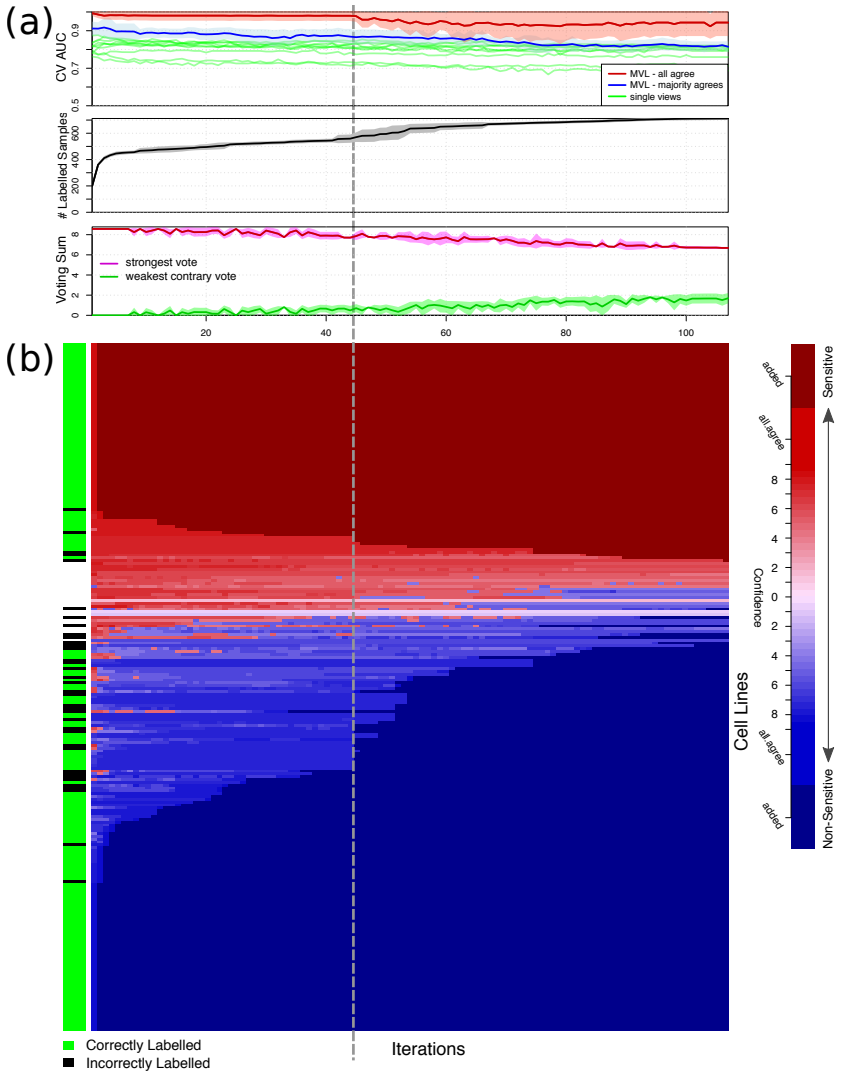


Figure 2.3: **PLATYPUS Validation.** Dashed vertical line shows user-defined stopping point for the method, where the overall disagreement in predictions between the views has started to increase and before the models area under the receiver operating characteristic curve (AUC) starts to significantly decrease. **(a)** Cross-validation accuracy mean (solid line) and standard deviation (colored area) plotted at each iteration (x-axis). Top plot: the prediction accuracy for single views (green lines), PLATYPUS ensembles with majority (75%) agreement (blue line), or all view votes agreeing (red line) on unlabeled samples. Middle plot: For each iteration, the number of samples for which labels have been learned (y-axis). Bottom plot: The votes summed up for the unlabeled sample with the highest vote (pink line) and smallest vote (green line) at each iteration across the cross-validation folds. **(b)** Label learning progress over successive iterations (x-axis) showing confidence of predictions for each unlabeled cell line (y-axis). Success or failure of the method to assign the correct label to each cell line shown in first column (green, correct; black, incorrect). Darker color indicates higher vote confidence across the views for either the non-sensitive (blue) or sensitive (red) class.

high confidence threshold for label learning and by using many information-independent views. In our experiments, LLV consistently helps identify optimal parameters.

Without missing data, PLATYPUS is equivalent to a classic ensemble classifier and often outperforms any single-view model. In order to understand the benefits of using additional unlabeled data, I compared the 'ensemble' (first) iteration of PLATYPUS to the final and the 'best' iterations. We define 'best' as the iteration with the highest AUC. In almost all cases, the PLATYPUS AUC is higher than the ensemble AUC (Fig. 2.4). The use of more samples by PLATYPUS helps ensure a more generalized model. For the experiments in this paper, we intentionally set a high number for maximum iterations to show how label learning can degrade over time, and therefore the final iteration often scores poorly. Label learning degradation is avoidable by using high label learning thresholds and an appropriate number of iterations.

### 2.3.2 Predicting drug sensitivity in cell lines

Our analysis focuses on the full CCLE data set, composed of 36 tumor types. For most drugs, the Sample- and Patient-Specific (SPS) view has the highest starting view performance with AUCs ranging from 0.6 to 0.8, and expression baseline views often performed similarly (Fig. 2.5). The mutation view is effective for some drugs, e.g. MEK inhibitors. Three of the four baseline views are top performers for predicting cancer cell line sensitivity to PD-0325901 (Fig. 2.2(b)), a MEK1/2 inhibitor. CNV view performance was never high enough to warrant inclusion in PLATYPUS models except as the 'aggregated copy number changes' feature in the SPS view.

Interpreted views often outperform the SPS view (Fig. 2.5). We found several ex-

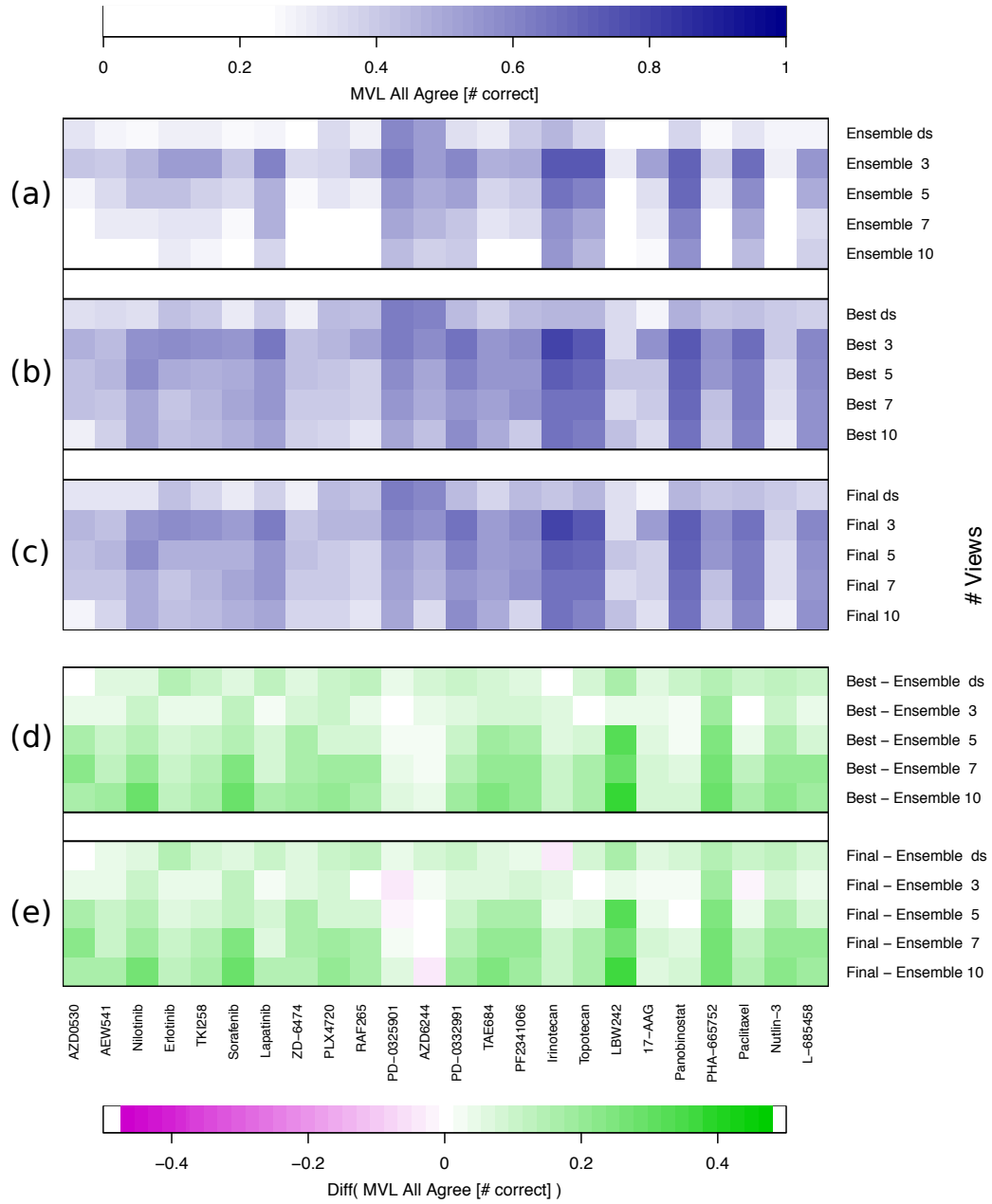


Figure 2.4: **Comparison to Ensemble.** Cross-validated number of correctly predicted samples (balanced accuracy × coverage) of PLATYPUS for 100% agreement predictions. **(a)** 'Ensemble' represents the first iteration of the MVL algorithm, in which no inferred labels have been added. **(b)** 'Best' is PLATYPUS iteration with the highest AUC. **(c)** 'Last' is the model from the final PLATYPUS iteration. Inferred labels were added until 75% agreement. **(d-e)** Comparison between the different MVL models by subtracting the Ensemble performance from the (d) Best and (e) Final performances. Each compound was predicted with the data-specific (ds) views (SPS, Mutation, Expression), and with the 3, 5, 7, and 10 most accurate interpreted single views.

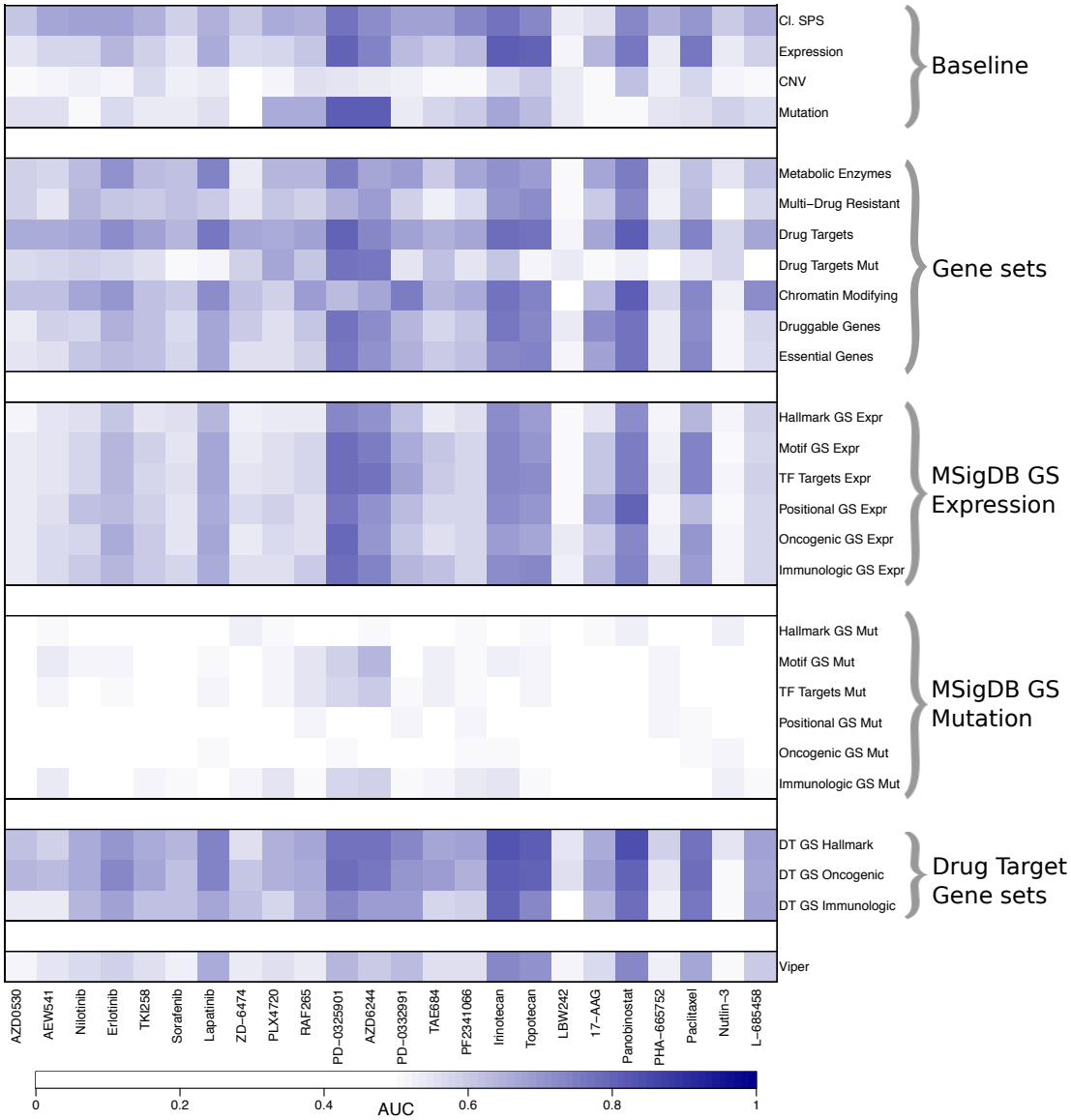


Figure 2.5: **Single View Performance.** Cross-validated area under the receiver operating characteristic curve (AUC) for each view when predicting sensitivity to each drug in CCLE, grouped by data type. All values $\leq$ 0.5 (AUC of a random predictor) are shown in white. GS = Gene Set; DT = Drug Target; Expr = Expression; Mut = Mutation.

amples in which a biological prior view outperformed the data-specific view, e.g. Metabolic Enzymes, Drug Targets, and Chromatin Modifying Enzymes are better at predicting Lapatinib sensitivity than the baseline expression predictor. The Drug Target Gene Set Hallmark view outperforms data-specific views in Irinotecan and Panobinostat sensitivity predictions. Such examples can be found for all compounds except for the MEK inhibitors, for which the baseline mutations view is always the top performer.

In general, views incorporating expression data have high accuracy (Fig. 2.5), whereas mutation views are comparable to a random prediction for most drugs. This could be due to the presence of many passenger mutations that have little bearing on cell fitness and drug response. In one notable exception, AZD6244, the Drug Target Mutation view is more accurate than the Drug Target Expression view. However, overall, mutation views have low accuracy despite mutations being key to drug sensitivity, indicating that other representations that increase the signal-to-noise ratio of this data should be explored in future work.

The Drug Target Gene Set views created from Molecular Signatures Database (MSigDB) gene set collections perform well overall, especially on Irinotecan, Topotecan, and Panobinostat (Fig. 2.5). For most compounds, the Drug Target Gene Set Hallmark is more accurate than the Oncogenic and Immunologic. A possible reason is that these gene sets are from the Hallmark collection, which are re-occurring, highly reliable gene sets built from combinations of other gene set collections. Their similar performance could also be due to a high overlap in the gene sets of the different collections. We recommend that users test for and subsequently remove highly correlated views before running label learning. One approach to handling correlated views is to extend the ensemble vote step to use stacked learning instead of the current agree-

ment formula. By training a model on the predictions from each view, PLATYPUS may be better able to handle correlated views by treating them with less weight than more independent views.

In addition to the MSigDB gene set views, master regulator-based predictors via Virtual Inference of Protein activity by Enriched Regulon analysis (VIPER) [8] were tested but are not among the top-performing ones for any drug. This could be due to the use of a generic regulon as VIPER input rather than tissue-specific versions for each cell line [8].

The PLATYPUS model for the drug PD-0325901 achieved the highest accuracy of all experiments, with a near-perfect AUC. We therefore chose to further investigate the results of this drug to identify the nature by which the MVL approach finds an improved classification. PD-0325901 was initially tested in papillary thyroid carcinoma cell lines and is known to be especially effective in cell lines with BRAF mutations [35]. Since these are frequent in the CCLE data, the high accuracy of the single-view models is expected. Fig. 2.6 shows changes from the ensemble to the 'best' PLATYPUS PD-0325901 models. Single view AUCs mostly increase after several iterations and feature weights within the models also shift to varying degrees. In the baseline mutations view, RAS gene mutations have higher Gini coefficient changes in the PLATYPUS model than in the ensemble (Fig. 2.6(c)), indicating increased model importance of those genes. Past studies of the CCLE data [12] and our analysis have found RAS and BRAF mutations in the data tend to be mutually exclusive, both of which are linked to PD-0325901 sensitivity. Thus, PLATYPUS is better able to identify the dual importance of RAS/BRAF mutations compared to the single view and the ensemble model.

We also chose to look at a case where PLATYPUS failed to achieve an improvement.

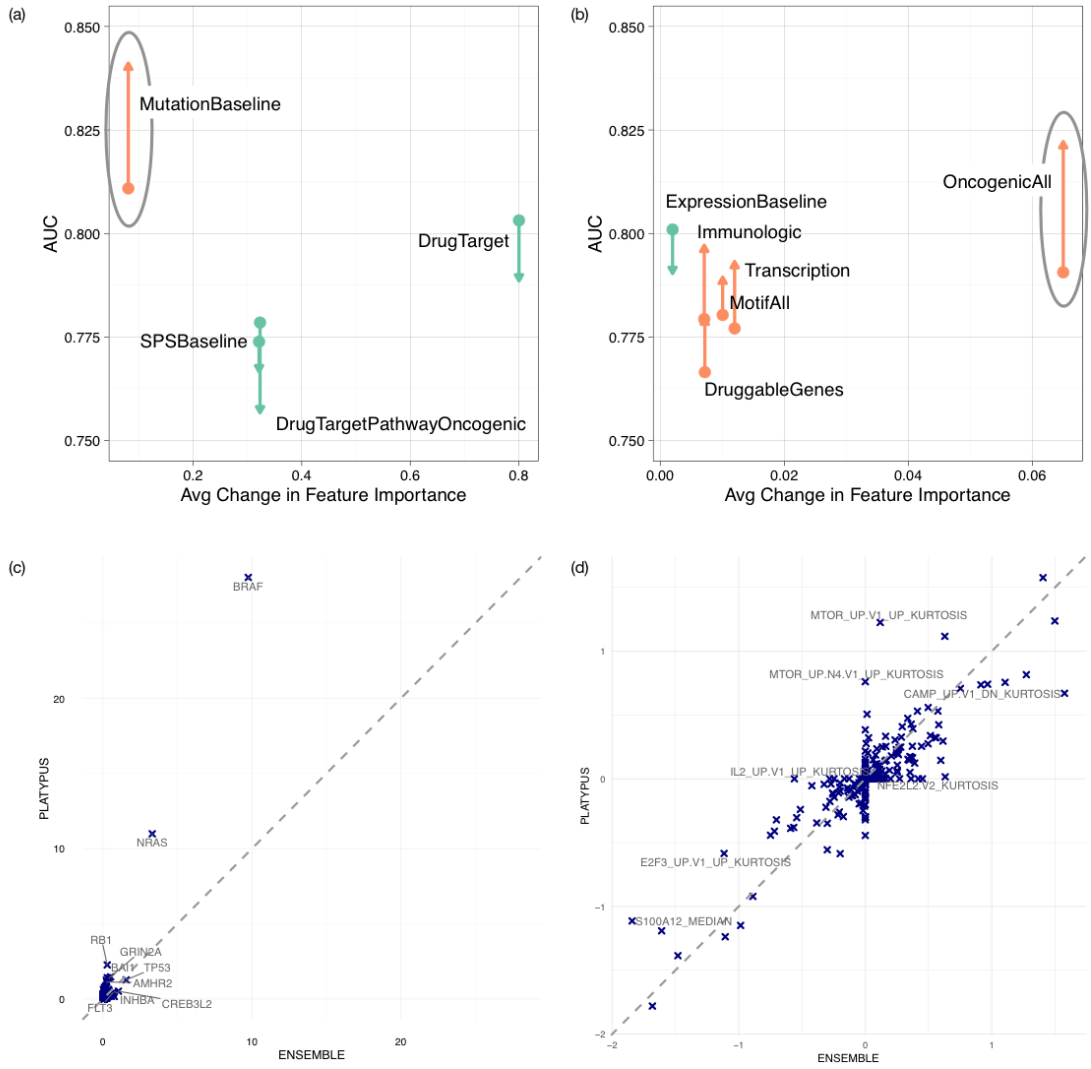


Figure 2.6: **Feature Weight Changes in PLATYPUS.** Performance and feature weight changes for single views between ensemble and PLATYPUS in predicting sensitivity to PD-0325901. **(a)** For each random forest view, the average Gini change for all features between the ensemble and the best PLATYPUS iteration, plotted against the view AUC for the ensemble (arrow tail) and PLATYPUS (arrowhead). Circled view is shown in detail in (c). **(b)** Same as (a), but showing the elastic net views and their average change in feature weights. **(c)** Scatter plot where each point is a feature in the Baseline Mutations view. Plot shows the ensemble feature weight versus the PLATYPUS feature weight. **(d)** Same as (c), but showing feature weight changes in the Oncogenic (OncogenicAll in (b)) view.

LBW242 is one such case. The single views for this drug all have near-random scores. However, instead of identifying an improvement through view combination as is the usual case in our experiments (e.g. PHA-665752 and Nutlin-3), the PLATYPUS models also achieved near-random performance (Fig. 2.2(a)). Further investigation reveals that the performance may not be the fault of PLATYPUS. Instead, little signal may be available in the drug sensitivity labels for this case due to our quantization strategy (i.e. using the upper and lower quartiles for the resistant and sensitive classes). The dose-response curve for LBW242 shows very few of the CCLE cell lines may be truly sensitive. While our approach creates balanced class sizes and ensures continuity between experiments, finding a more nuanced per-drug cutoff would likely improve model performance. Suboptimal label cutoffs lead to a low signal-to-noise ratio in the labels for a few of the drugs, which in general leads to low classifier performance [27]. It is also possible that the metric for drug sensitivity for some drugs is ineffective. Traditional methods to quantify sensitivity are dependent on population growth and thus slow-growing cell lines may appear to be resistant to all drugs [33].

These results are consistent with previous findings that have shown sensitivity to some compounds is easier to predict than others [22]. For example, the two MEK inhibitors (PD-0325901, AZD6244) and Panobinostat have higher overall accuracy in the single-view models (Fig. 2.5). Interestingly, in the case of Panobinostat, the 'Chromatin Modifiers' and 'Positional Gene Set' PLATYPUS views have higher single view accuracy than the baseline expression view, which could indicate that there is an epigenetic effect from chromatin modifiers. We postulate that a small region of the genome has been unwound, lending sensitivity to Panobinostat. PLATYPUS captures this interaction, whereas single-view models do not.

### 2.3.3 Key features from PLATYPUS models

Each machine-learning algorithm used by a view has its own internal feature selection. We extracted features from these models to evaluate the most informative features. Fig. 2.6(a-b) show changes in single-view model performance and average feature importance within those models, before and after PLATYPUS training. Fig. 2.6(c-d) show feature changes and enrichment of those features within one of the views. Fig. 2.6(c) highlights how PLATYPUS is able to remove feature weights of spurious correlations between cell line mutations and the true mutation features of importance: NRAS and BRAF. While the overall feature weights in the single-view model do not have large changes from the ensemble to PLATYPUS frameworks, there is a large shift in 2 key features which are known to be significantly associated with sensitivity to this particular drug. PLATYPUS is able to avoid overfitting the model whereas the ensemble is unable to draw from external information. In Fig. 2.6(d), the model has significantly changed both in AUC and in feature weights between the ensemble and PLATYPUS experiments.

Fig. 2.7 shows a closeup of the changes within the Fig. 2.6(d) view between PLATYPUS and a general ensemble. It focuses on one feature from the view, MTOR_up_V1_up kurtosis, which had the biggest increase in feature weight from ensemble to PLATYPUS. At a glance, this gene set is not associated with cancer - it describes genes that are regulated by an inhibitor used to prevent graft rejection by blocking cell proliferation signals via mTOR. However, the gene set kurtosis correlates with ActArea and with our binary drug sensitivity labels (Fig. 2.7(a-b)). A closer look shows that this is because of gene-gene correlations within the

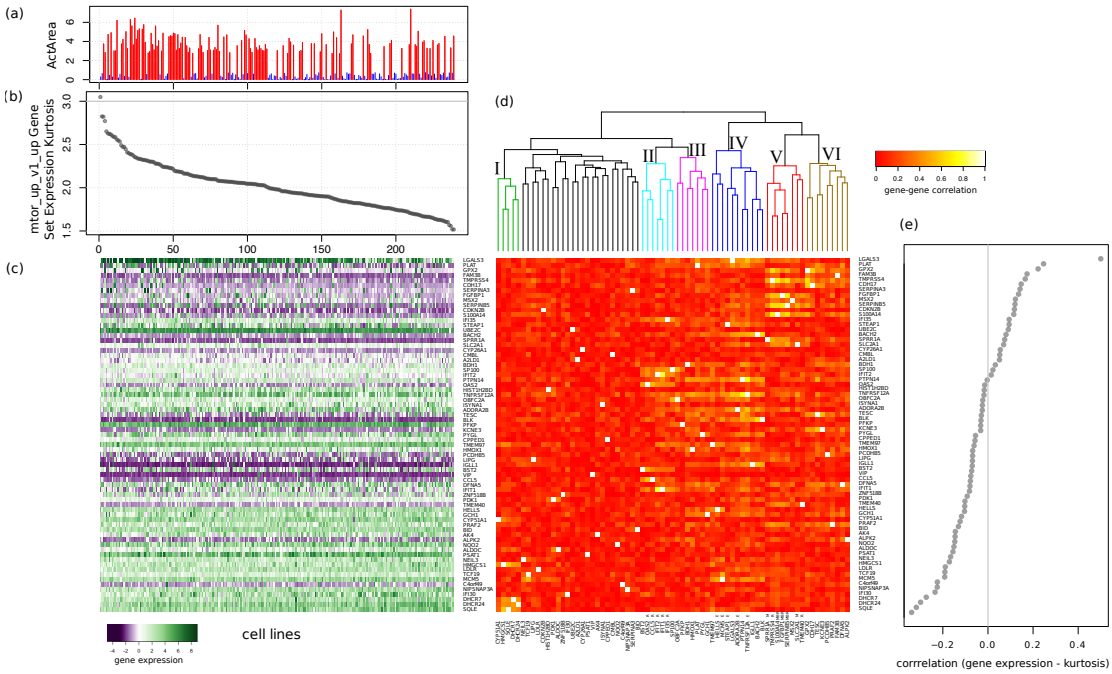


Figure 2.7: **MTOR_up_V1_up gene set feature and its relationship to expression and outcome.** **(a)** ActArea for each cell line, sorted by the MTOR gene set kurtosis value; a higher proportion of resistant cells (red lines) are associated with higher kurtosis values of this gene set (left side) compared to sensitive cells (blue lines). **(b)** MTOR gene set kurtosis value for each cell line. **(c)** RNA Expression of the genes within the gene set. **(d)** Same genes as in (c), now showing gene-gene expression correlation. Tree shows hierarchical clustering of the genes and highlights groups of similar genes. Genes involved in EGFR signaling are marked with E, metastasis with M, basal vs mesenchymal BRCA with B, and resistance to several cancer drugs with R. **(e)** Correlation between gene expression and MTOR gene set kurtosis value determines sorting of genes in (c) and (d).

gene set. Kurtosis features are intended to capture large changes within the gene set. Mean and median gene set correlation values do not capture cell line differences in the co-correlated gene clusters, whereas kurtosis highlights extreme values. No one gene expression correlates strongly with the kurtosis of the whole set (Fig. 2.7(c,e)), and so the set cannot be replaced with a single gene expression value. Clusters within the gene set are linked to EGFR signaling (cluster IV, genes marked E), metastasis and Basal vs Mesenchymal BRCA (cluster V, genes marked M and B respectively), and resistance to several cancer drugs (clusters II and V, genes marked R). Gene-gene correlations shown in Fig. 2.7(d) combine to form the overall kurtosis score. As shown in Fig. 2.7(e), many genes related to cancer processes are the driving force in the gene set kurtosis score. This highlights how small overall changes combine to improve PLATYPUS accuracy over the ensemble.

Many of the highly ranked features from other models are known oncogenes. Fig. 2.8 shows the expression baseline view for PD-0325901, other data is not shown. ETV4 for example was previously found to be correlated with MEK inhibitor sensitivity [45]. SPRY2, a kinase inhibitor, correlates with BRAF mutation status, both of which are predictive of sensitivity to PD-0325901, AZD6244, and PLX4720. DUSP6 has been named as a marker of FGFR inhibitor sensitivity [50] and a previous study shows a weak inverse correlation between DUSP6 expression and sensitivity to MEK1/2 inhibitors [31]. Thus, PLATYPUS recapitulates several known markers of drug sensitivity.

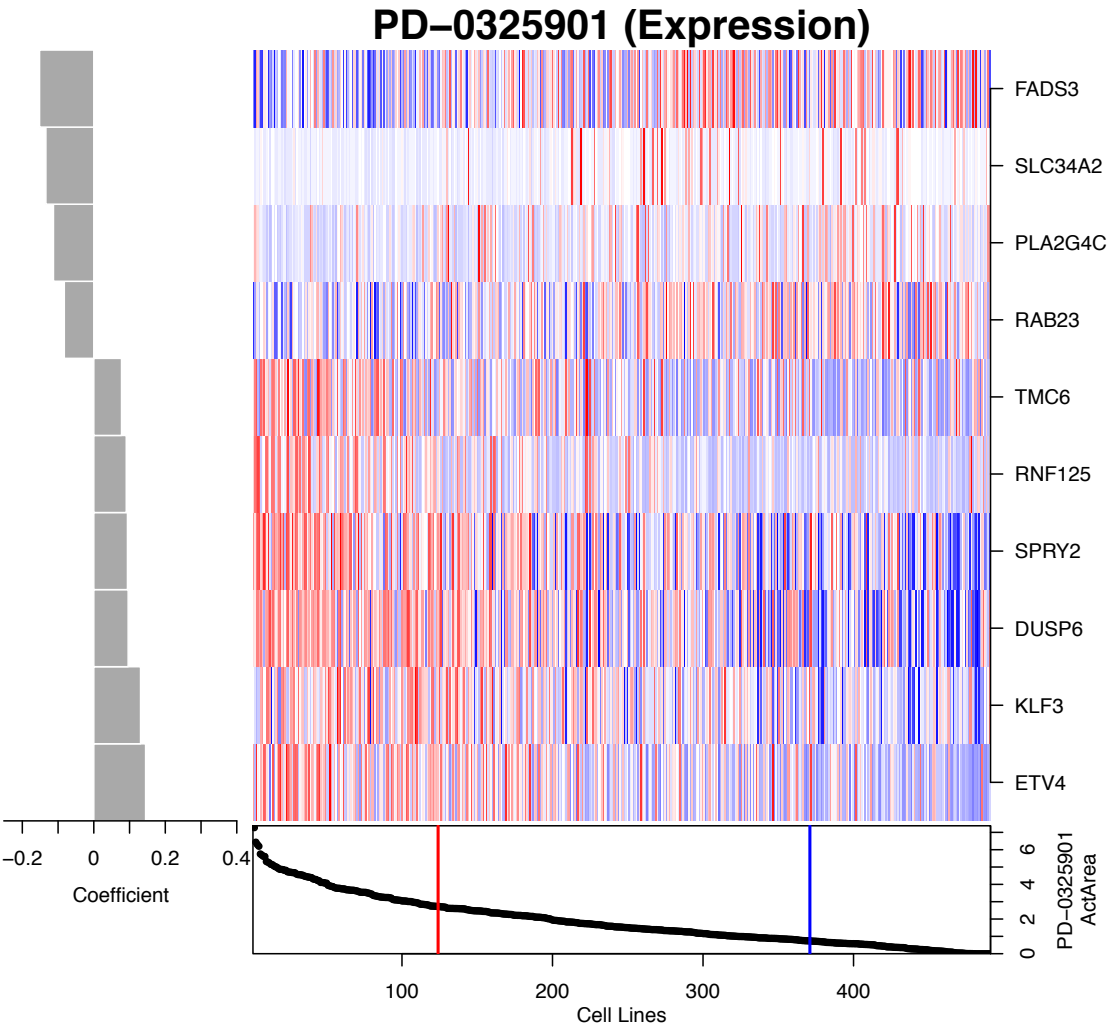


Figure 2.8: **Most important features in the PLATYPUS baseline expression view for PD-0325901 sensitivity prediction.** The ten features with the highest absolute coefficient in the trained elastic net model are shown (left panel). The gene expression values are normalized and shown in continuous colors from blue (low expression) to red (high expression). The drug response values in the lower panel determine the sample sorting but were used as binary labels as implied by the vertical lines: samples left of the red line are defined as sensitive, right of the blue line as non-sensitive.

## 2.4 Discussion

When compared to a traditional ensemble and to single view predictors, PLATYPUS performs better in most cases (Fig. 2.2). The multi-view approach uses the set of unlabeled samples as links between different views to find agreement in the different feature spaces. Since label learning validation shows that labels are learned correctly in most cases, the improved model performance may be due to doubling the number of samples that can be considered while training. In 96% of our experiments, PLATYPUS outperforms an ensemble (Fig. 2.4). Furthermore, PLATYPUS outperforms 85% of the single views and has a higher AUC than all of the single views for 17 of the 24 drugs. No one single view consistently outperforms any of the PLATYPUS models. Important features from PLATYPUS views, both baseline and interpreted, have previously been linked to drug sensitivity. The approach generally improves accuracy while incorporating significantly more data and allowing uncertainty - a necessity in medical research.

The PLATYPUS co-training approach has several important advantages. First, it is ideal when samples have missing data, a common scenario in computational biology. Imagine a new patient entering a clinic for whom not all of the same data is available as was collected for a large drug trial. A PLATYPUS model trained on the drug trial data is able to predict drug response for this patient without retraining, simply by restricting to views for which there is patient data. Second, co-training allows for the use of different classification methods for each data type, capturing the strengths of each data type and increasing flexibility in the framework. Third, PLATYPUS is effective when using information-divergent views. Fourth, co-training

combines predictions at a later stage in the algorithm, so that views are trained independently. This is ideal for ensemble learning, which has shown to be highly effective when models are independent, even with low individual model accuracy [10, 60].

It is worth mentioning some distinct limitations of the approach as a pointer toward future work. First, if missing data correspond to cases that are more difficult to classify, rather than missing at random, the poorer performance of individual views may result in a notably lower agreement, and thus little benefit in combining views. Second, combining multiple views introduces the need for setting additional parameters (e.g. the agreement threshold). This requires a user to gain familiarity with the performance of newly incorporated views in test runs before final results can be obtained. Finally, highly correlated views can inflate the agreement voting and down-weigh other, uncorrelated views. A future adjustment could incorporate prediction correlation on the labeled samples for the voting of unlabeled samples.

The application of PLATYPUS to the CCLE collection of cancer cell lines is an example of a controlled cancer model system showing that different cell lines respond differently to drugs and how it is possible to predict this response using multi-omics data types. The lower panel in Fig. 2.8 shows how differently cell lines respond to treatment with the MEK inhibitor drug PD-0325901. The responses vary from no response at all to a very strong response in a few cell lines. For this drug, we suspect that the response depends heavily on the genetic background of the cancer cell line, i.e. cancer driver mutations in genes like BRAF and RAS that create vulnerabilities that can be targeted by a drug. For other drugs, e.g. Panobinostat, we were able to find signals for an epigenetic effect on drug sensitivity. In most cases however, gene expression data was the best predictor for drug response, even though cancer is generally

thought of as resulting from genetic changes. Mutation data was not predictive for most drugs, and we suspect that this is due to a high signal-to-noise ratio due to many passenger mutations that have no influence on the cancer cells and only few driver events. Additionally, different mutations can have a similar effect on gene expression and subsequently on cell state, and this is not captured in the mutation data. In many cases, we were able to further enhance the signal in the gene expression data by incorporating prior biological knowledge in the form of gene sets.

These results show that the transcriptional signature is often able to better approximate the overall state of a cell line and can be used as a proxy for different genetic backgrounds that have similar effects on the cells. In chapter 3, such transcriptional signatures are used to find new genetic driver events in cancer samples that previously did not have any driver events annotated.

# Chapter 3

# Finding Cancer Driver Events with Transcriptional Signatures

The project described in this chapter was conceptualized and led by David Haan and was published in [32]. My main contributions were the visualization and analysis of the TCGA Pan-Cancer results shown in Figures 3.1 and 3.3.

## 3.1 Introduction

Cancer is a genetic disease caused by DNA mutations. Randomly occurring mutations are normally either repaired or the cells with these mutations die. In cancer cells, however, the involved biological pathways are not functioning properly and so mutations accumulate, which leads to uncontrolled cell growth. Mutations in cancer cells are classified into two groups: driver mutations that cause and accelerate cancer, and passenger mutations that are just remnants of non-functional DNA repair mechanisms. Each cancer is estimated to contain about 2-5 driver

mutations and about 10-200 passenger mutations [74]. Because these genetic driver events are what make the tumor grow, they offer a chance for therapeutic intervention. The cancer cells and their growth are dependent on those genetic changes, and therefore a therapy can either target those changes and try to reverse them, effectively slowing the growth of the tumor, or it can target pathways that keep the cancer cells alive but are synthetic lethal within the cancer's genetic background.

However, the identification of driver events among the much higher amount of passengers remains a challenge. In The Cancer Genome Atlas (TCGA) project, a publicly available data set of about 11,000 cancer patient samples across 33 different cancer types [36], the identification of driver events was the subject of many analyses. The analysis of papillary thyroid carcinoma identified two subtypes that are separated by different driver events: mutation of the BRAF gene and mutation of genes in the RAS family [5]. About 95% of samples had at least one of these driver mutations. The remaining 5% of samples were left without annotation of a driver event. Such situations are typical for many tumor types with some samples harboring known, well-studied driver events, but others are left unannotated because they have rare driver events.

To tackle this problem, we present LURE (Learning UnRealized Events), a machine learning tool that aims to find new driver events that share transcriptional signatures with known driver events. LURE has an iterative process that uses gene expression data to identify genetic events in samples that show a similar transcriptional pattern to known driver events. It was shown previously that many driver events can be identified through transcriptional signatures, e.g. TP53 [67, 77] or BRCA [42]. LURE uses such signatures to identify new events that are

related to known drivers but are too rare to be found by themselves.

## 3.2 Methods

LURE associates genetic events by finding similar molecular signatures among the samples in which the events occur. A genetic event here is a particular type of alteration that affects a single gene, such as a deletion, a missense mutation, a truncating mutation, or a gene fusion. In this analysis, mRNA sequencing-based expression data is used to generate signatures, although other data choices are possible, e.g. microRNA expression or DNA methylation. LURE finds related events by training a classifier using the samples containing a known driver mutation (the 'bait'). The method retains only those baits that yield accurate models determined by cross-validation. It then applies the classifier to find 'target' samples defined as high-scoring samples lacking bait alterations. In classic machine learning, the target samples would be considered false positives. However, in this case, they offer an opportunity to find drivers because the comprehensive collection of TCGA mutation calls can be searched to find other events that significantly coincide with the target samples. Any such event (the 'catch') is associated with the starting bait and provides a new set of labels for retraining a more accurate classifier in subsequent rounds. The iterative process comes to an end when either the classifier is not improving anymore, or no further catch events can be found. In a final step, LURE computes a minimal set of nearly mutually exclusive events in the catch that best span all the catch samples, the 'Catch Cover'. See the original publication for the full, more detailed method description [32].

## 3.3 New cancer driver associations for 33 TCGA tumor types

LURE was applied to each TCGA tumor type [36], restricting both baits and catches to mutation events involving one of the 723 COSMIC genes [69]. Tumor type-specific classification models were necessary to avoid any confounding effects due to tissue-specific expression and mutation patterns. Bait events were created for each gene for each distinct mutation type including missense mutation, truncating mutation, homozygous focal point copy number deletion, and gene fusion. In order to limit the number of putative false-positive results, we restricted the number of bait classifiers by considering only those with at least 10 positive samples. In order to produce a confident set of results, the baits were also limited to those with an area under the precision-recall curve (PR AUC) $> 0.5$, precision $> 0.4$, and recall $> 0.75$. Among the bait classifiers passing these thresholds, the most common bait gene across all tumor types was TP53, and the tumor types with the highest number of passing bait classifiers were Lower Grade Glioma (LGG), Thyroid Carcinoma (THCA), and Prostate Adenocarcinoma (PRAD).

LURE was run with the 81 remaining classifiers as baits, using missense mutations, truncating mutations, splice site mutations, gene fusions, focal point copy number amplifications, and homozygous deletions of COSMIC genes as possible catches. LURE found significant bait-catch associations for 35 of the 81 baits tested. Fig. 3.1 shows the resulting bait-catch associations for the 18 baits with the most accurate final classifier. Among these high-confidence results, 14 of the 59 bait-catch associations were between events involving the same gene, such as TP53 truncating, splice site, and missense mutations. There were four associations within the same gene families, e.g. IDH1/2 or the RAS protein family. In addition, we

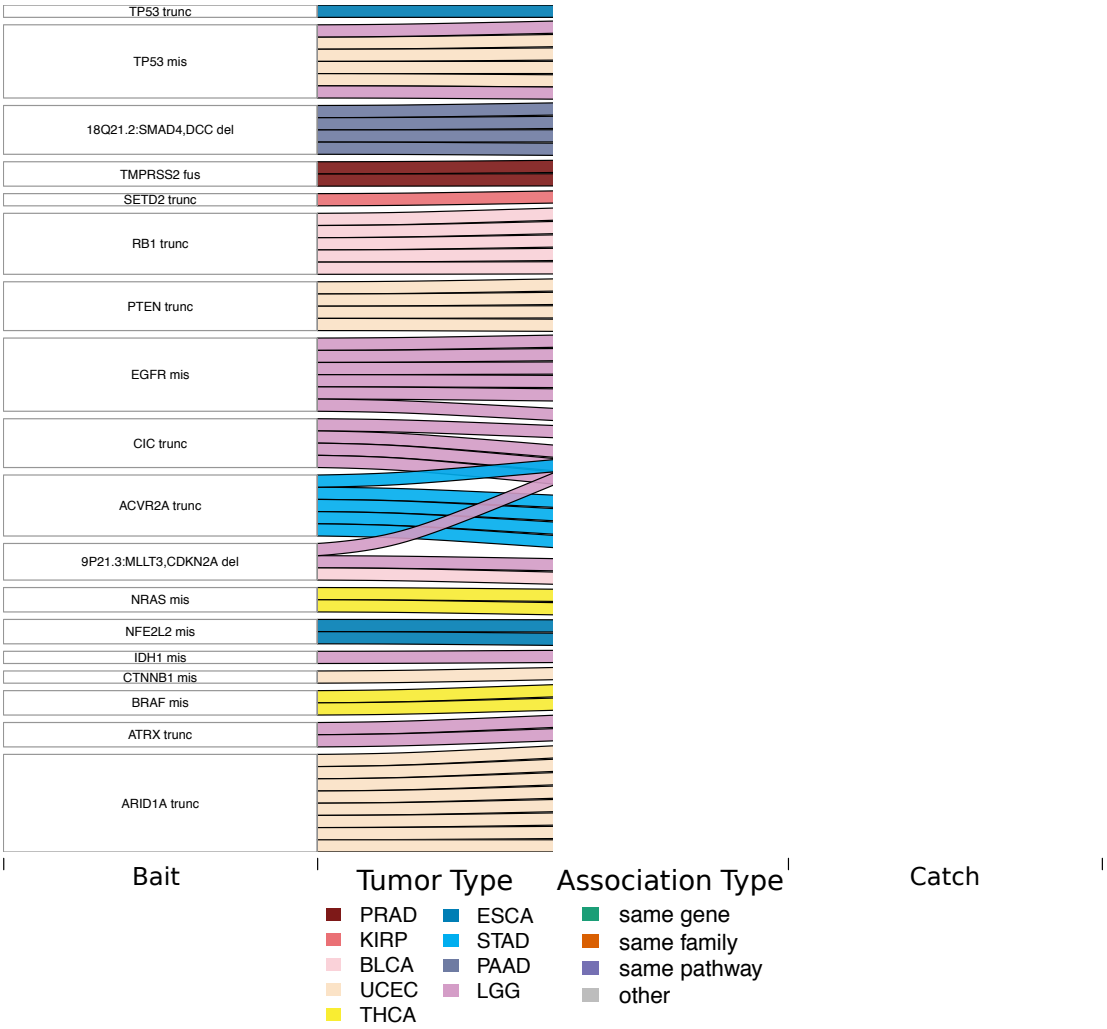


Figure 3.1: **High-Confidence Bait-Catch Associations from LURE on the Pan-Cancer Atlas Data Set.** Sankey diagram shows the high confidence bait-catch associations for the 18 baits with a final classifier PR AUC $> 0.8$. Bait gene and mutation type are shown on the left; catch gene and mutation type on the right. Each horizontal flow connection represents an association between the bait and catch event found by LURE. The left half of each flow bar is colored by tumor type in which the association was found. The right half of each flow bar is colored by the association type.

identified gene fusion event partners in BRAF and RET that associated with a BRAF missense mutation in THCA. For 20 of the 59 high confidence bait-catch associations, both genes were members of the same signaling pathway (excluding pathway gene sets with $> 1000$ genes) [49].

Superimposing all the bait-catch associations obtained within each tissue type into a compendium of bait-catch associations, referred to as the 'Event Net', reveals several pathway-oriented findings (Fig. 3.2). Some of them have support in multiple tumor types, most notably again the associations with TP53. LURE identified interesting associations for PTEN, in particular between PTEN and CTNNB1. This connection is supported by earlier research suggesting PTEN plays a role in regulating the subcellular localization of β-catenin [58]. Another striking LURE association is between PTEN and EGFR, which is consistent with other findings suggesting that PTEN regulates EGFR signaling [64]. These LURE associations for PTEN reveal crosstalk between pathways and provide further evidence that alterations in PTEN influence EGFR and β-catenin signaling.

LURE's TCGA Pan-Cancer Atlas analysis associated a total of 1,216 samples with 35 bait events through 2,845 connections to 53 catch events. Because baits were created for different alteration types in the same gene, as opposed to one bait for any alteration in a gene, the method was able to identify associations between different alteration types within the same gene as well as infer the functional impact of alterations. For example, α-thalassemia mental retardation X-linked (ATRX), a gene recurrently mutated in LGG, only has an oncogenic effect with a loss-of-function mutation - either a truncating mutation or copy number loss - whereas a missense mutation may not have an oncogenic effect [18].

I attempted to measure the degree to which LURE expands the list of known drivers in

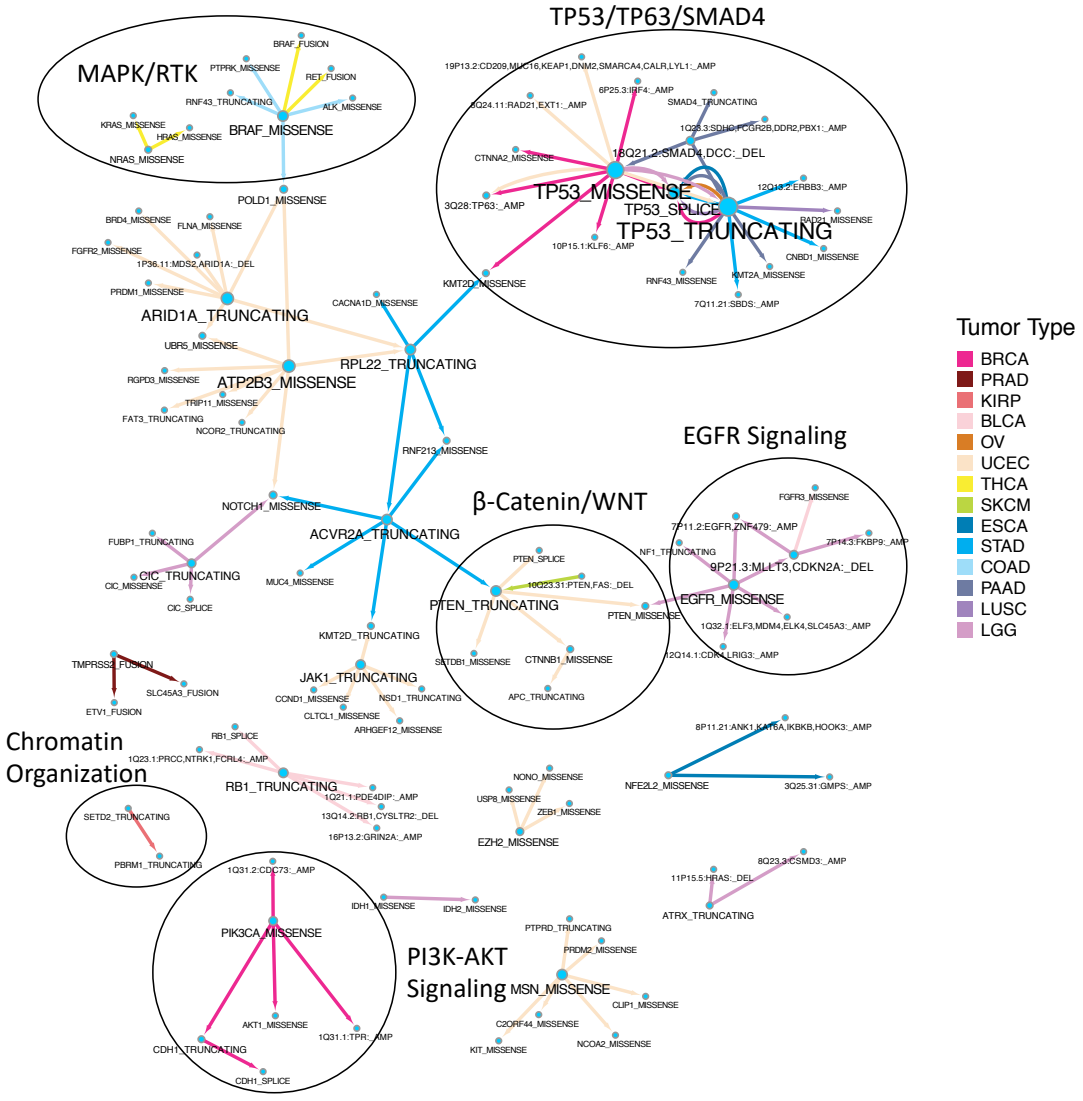


Figure 3.2: **LURE Event Net for the TCGA Pan-Cancer Atlas Data Set.** Cytoscape [63] visualization of all associations found in the TCGA data set. Each node represents an event. Directed edges represent an association and the direction of the LURE discovery (bait to catch). The color of each edge represents the tumor type in which the association was found. Pathway findings are circled and annotated by pathway name.

the TCGA data set. While estimating this quantity is difficult, I approximated it by counting up the frequency with which LURE connects catch events to baits between genes without any prior functional relationship. To this end, I found that 6.9% (10,271) of the total events considered were associated with a transcriptional signature either as a bait, catch, or both (Fig. 3.3A). Over a third of these events (34%; 3497) are from catches that were not themselves used as a bait event (Fig. 3.3B). We categorized these catch events according to whether they connect the same gene (20%, 683), family (2%; 77), pathway (25%; 890), or no known relationship (53%; 1847) with the original bait (Fig. 3.3C). Thus, the numbers demonstrate that among these 6.9% of events with a transcriptional signature, 18% (1847) are newly implicated by LURE linking one signature to another.

## 3.4 Discussion

Using molecular signatures to associate events will extend the list of known drivers and may ultimately identify targets for precision medicine. LURE identified an intriguing set of new candidate drivers based on the TCGA Pan-Cancer Atlas data set. Most of the found associations were between events of the same gene (e.g. TP53), gene family (e.g. IDH, RTK), or biological pathway (e.g. PI3K). In addition to associating known events, LURE identified 1847 events of unknown relationship with their associated bait event. The collection of these events provides possible actionable clues for cancer patients. For example, LURE found that deletions in 2q23.3 in head-and-neck cancers are strongly associated with RTK signaling. Whether such a focal deletion could be used as a novel biomarker for treating patients with an RTK inhibitor,

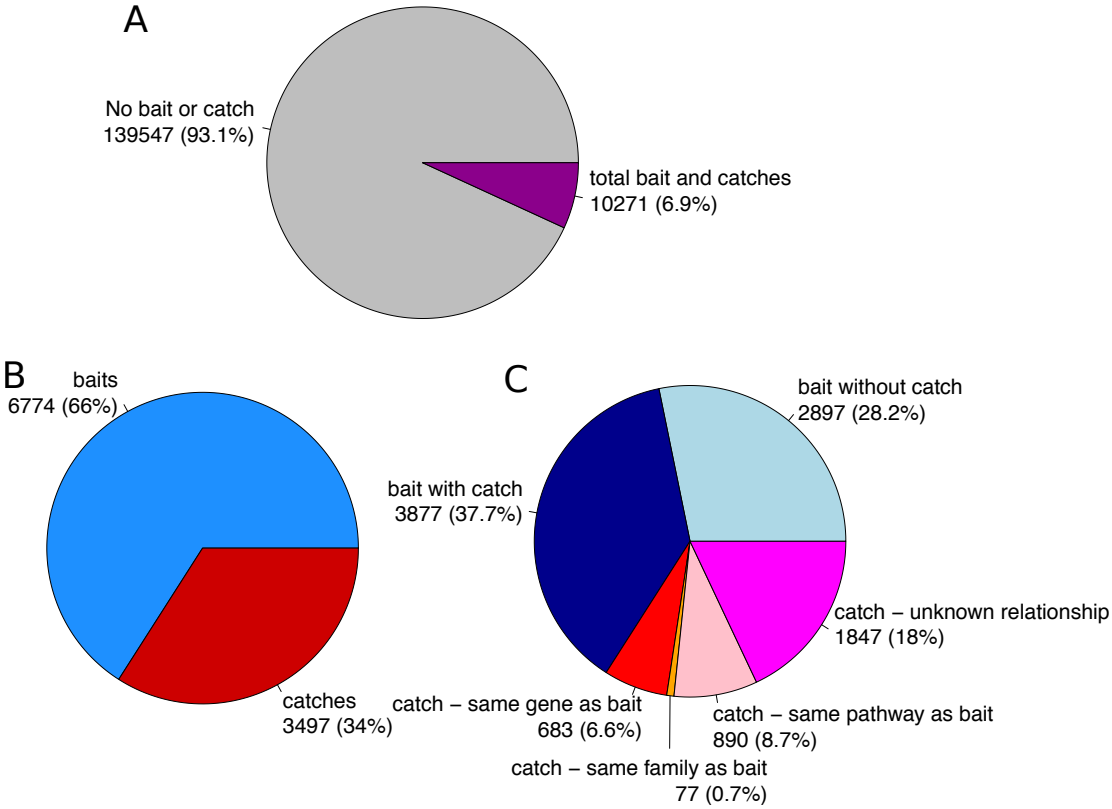


Figure 3.3: **LURE genomic events in TCGA Pan-Cancer Samples.** Pie charts showing the mutational events being present in TCGA samples and how they were used in LURE. **(A)** All genomic events used in LURE: missense mutations, truncating mutations, homozygous focal point copy number deletions, and gene fusions in 732 COSMIC genes for bait events, and additionally splice site mutations and focal point copy number amplifications in catch events. The pie chart is divided into events that were neither bait or catch in LURE (no transcriptional signal detected), and the events that were either bait or catch or both (transcriptional signal detected). **(B)** Pie chart including all events from 'total bait and catches' in (A), divided into events that were used as bait by LURE and events that were used as catch and never as bait. **(C)** Same as (B), but baits are further divided by if LURE was successful in associating the bait event with at least one catch (bait with catch), or not (bait without catch). Catches are further divided by their relationship with the bait they were associated with by LURE: bait and catch are from the same gene, bait and catch genes are in the same gene family, bait and catch genes are in the same biological pathway (excluding pathway gene sets with $> 1000$ genes) [49], or the bait and catch genes did not have any of those connections (unknown relationship).

such as gefitinib for EGFR, remains to be seen.

The application of LURE to the TCGA data set demonstrates that transcriptional signatures can be used within a tumor type to predict its genetic driver events. All associations between a bait and a catch event are examples of how a different genetic background can have the same phenotypic effect. However, almost all associations between driver events were found in only one tumor type. Only the associations around TP53 are supported by multiple tumor types (Fig. 3.2). The relationship between transcriptional state and genetic events does not seem to transfer well across different cancer types. The cell-of-origin of a tumor and the tumor microenvironment seem to play an important role in influencing the transcriptional state of a cancer sample, both of which I investigate further in the following chapters.

# Chapter 4

# Pan-Cancer Meta-Analysis of Molecular Subtype Classifiers

The work described in this chapter is a contribution to the TCGA Tumor Molecular Pathology Analysis Working Group (TMP AWG). The project is still subject to active research and will be published together with The Cancer Genome Atlas Research Network under the project lead of Peter W. Laird and Andrew D. Cherniak. My main contribution is the creation of meta-features that describe the prediction task and the analysis of their relationship to classifier performance.

## 4.1 Introduction

Cancer types are usually based on the organ or tissue of cancer origin. Cancer types are often further divided into subtypes, traditionally based on visual differences, e.g. histology defines categories based on how cancer cells look under a microscope. Since molecular data

like gene expression or genetic mutations for cancer biopsy samples are becoming more and more readily available, the definition and use of molecular cancer subtypes are increasing, and in some cases is already routinely used in the clinic, e.g. for breast cancer [56].

The Cancer Genome Atlas (TCGA) project collected multiple molecular data types for about 10,000 samples of different cancer types. This large-scale collection of molecular data provides a useful resource to define cancer subtypes on a molecular level. As part of the TCGA project, many disease experts were involved in the careful analysis of that molecular data for each cancer type in the study, e.g. breast cancer [41], colon and rectal cancer [51], squamous cell lung cancers [52], lung adenocarcinoma [21], and many more. In order to classify new cancer samples into the molecular subtypes defined by TCGA disease experts, we have to define what data points have to be measured and how to classify samples using these data points. The Tumor Molecular Pathology (TMP) Analysis Working Group did this in a systematic and comprehensive analysis for almost all TCGA tumor types.

## 4.2 Building accurate molecular subtype classifiers with minimal feature sets

The TMP group set out to create molecular cancer subtype classifiers for 26 different cancer type cohorts covering a total of 106 molecular subtypes. Five different state-of-the-art machine learning algorithms for biological data were used to find a minimal set of molecular features that can categorize cancer samples into their molecular subtype with high accuracy. The input features originated from five data platforms: gene expression measured by mRNA

sequencing (GEXP), microRNA expression (MIR), somatic DNA mutations (MUTA), somatic DNA copy number variations (CNVR), DNA methylation (METH). For each cancer cohort, the top-performing algorithm was chosen as the final model.

The classifiers overall reach a high level of accuracy: all final models have a weighted F1-score between 0.83 and 1.0, and the mean weighted F1-score over all cohorts is 0.93. In general, the data types that were used to define the molecular subtypes for a given cancer cohort, are also the most useful in predicting the subtypes for that cohort. However, it is not feasible to expect the large amount of data generated for TCGA samples to become the standard for every cancer sample in the near future. Many cancer cohorts reached a comparable level of classification accuracy using only gene expression data as input features, indicating that in most cases transcriptional data can be sufficient for accurate molecular subtype classification.

The comprehensive set of subtype classifiers over 26 cancer cohorts can give us insights for future cancer subtyping studies. What samples and data are needed for a successful subtyping study? To answer this question, I conducted a meta-analysis of the TMP subtype classifiers.

## 4.3 Meta-features describing the prediction task are indicative of classifier performance

The goal of this meta-analysis was to assess which aspects of the input samples, input data, and trained models lead to better or worse subtype classifier performance. I created a set of 55 meta-features that describe different aspects of the input data and the subtype prediction

task for each cancer cohort. I created meta-features describing the sample set for each cohort, e.g. the total number of samples in the cohort, the number of subtypes, the number of samples in the rarest subtype, the percentage of samples in the rarest subtype, etc. The largest set of meta-features describes the input feature data, e.g. the percentage of input features each data type provides to the classifier, or the average variance in each data type. Similarly, I also collect the number of principal components needed to capture 70% of the input data variance as a meta-feature. To measure how well the input features are suited to distinguish between the subtypes of a cohort before any model training, I calculate the average silhouette score for each data type. The silhouette score describes how well the subtypes are separated on a data platform and therefore indicates how difficult it might be to distinguish between the subtypes using that data type. I also created a few meta-features describing the final trained classification model, e.g. how many features are used in the model, and the percentage of model features for each data type.

In a first step, I compared all meta-features to each other using hierarchical clustering on pairwise spearman correlation across the 26 cancer cohorts (Fig. 4.1).

This analysis revealed seven major meta-feature groups (MFG) of related meta-features. The MFGs were assigned labels based on a common theme among the majority of contained meta-features.

Next, I correlated each meta-feature to the final model performance for each cohort, measured in weighted F1-score (Fig. 4.2). Two MFGs show the overall strongest correlation with F1 accuracy. MFG1, subtype cohesiveness, has the strongest positive correlation. This MFG contains the silhouette scores for all data types, and all of them show a positive correlation

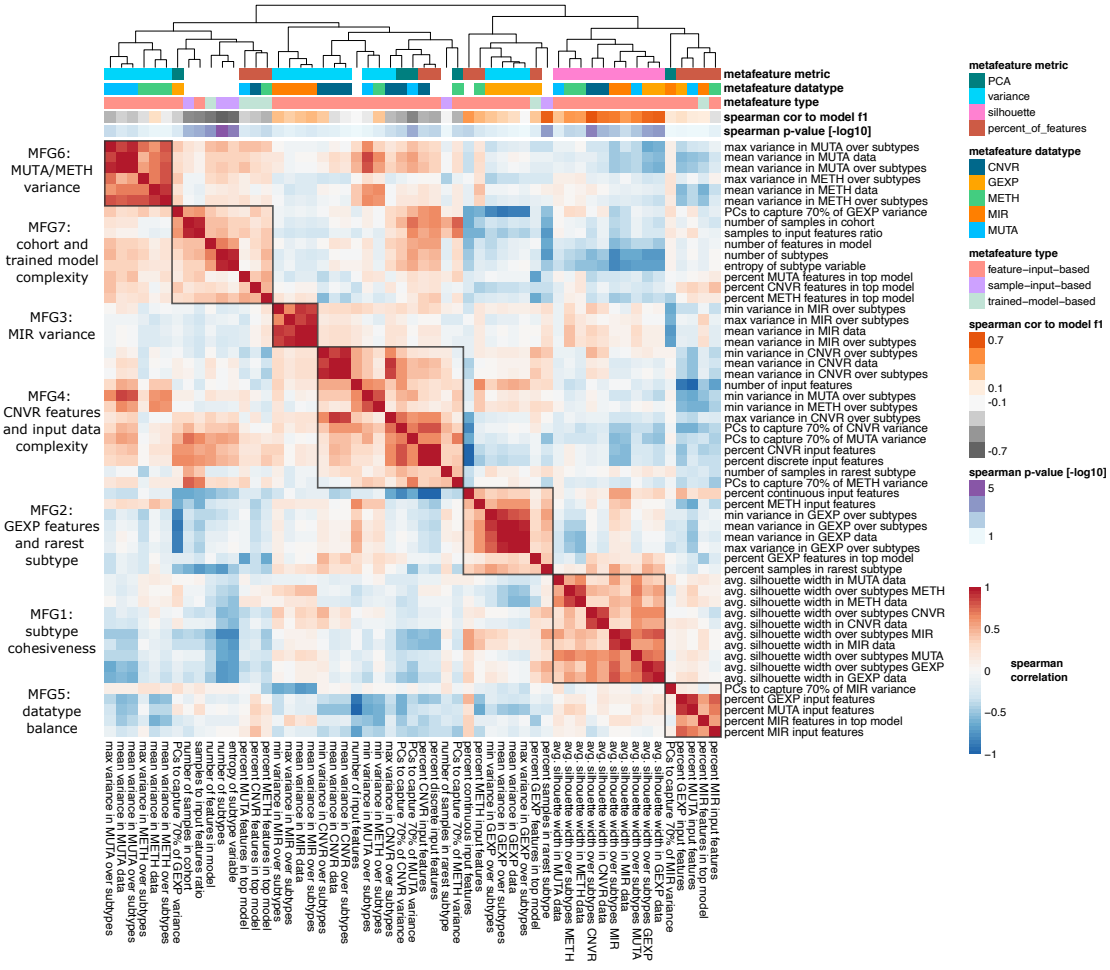


Figure 4.1: **Meta-feature correlation matrix and clustering.** Correlation heatmap of meta-features in 26 TCGA cohorts. Annotation on top of heatmap shows spearman correlation to model performance (measured in weighted F1 score), meta-feature categories (data type, metric, and meta-feature type), and a hierarchical clustering dendrogram. The hierarchical clustering solution is cut into 7 meta-feature groups (MFG). (CNVR = copy number variation, GEXP = gene expression, MIR = microRNA, MUTA = mutation, METH = methylation, PCs = principal components, PCA = principal component analysis).

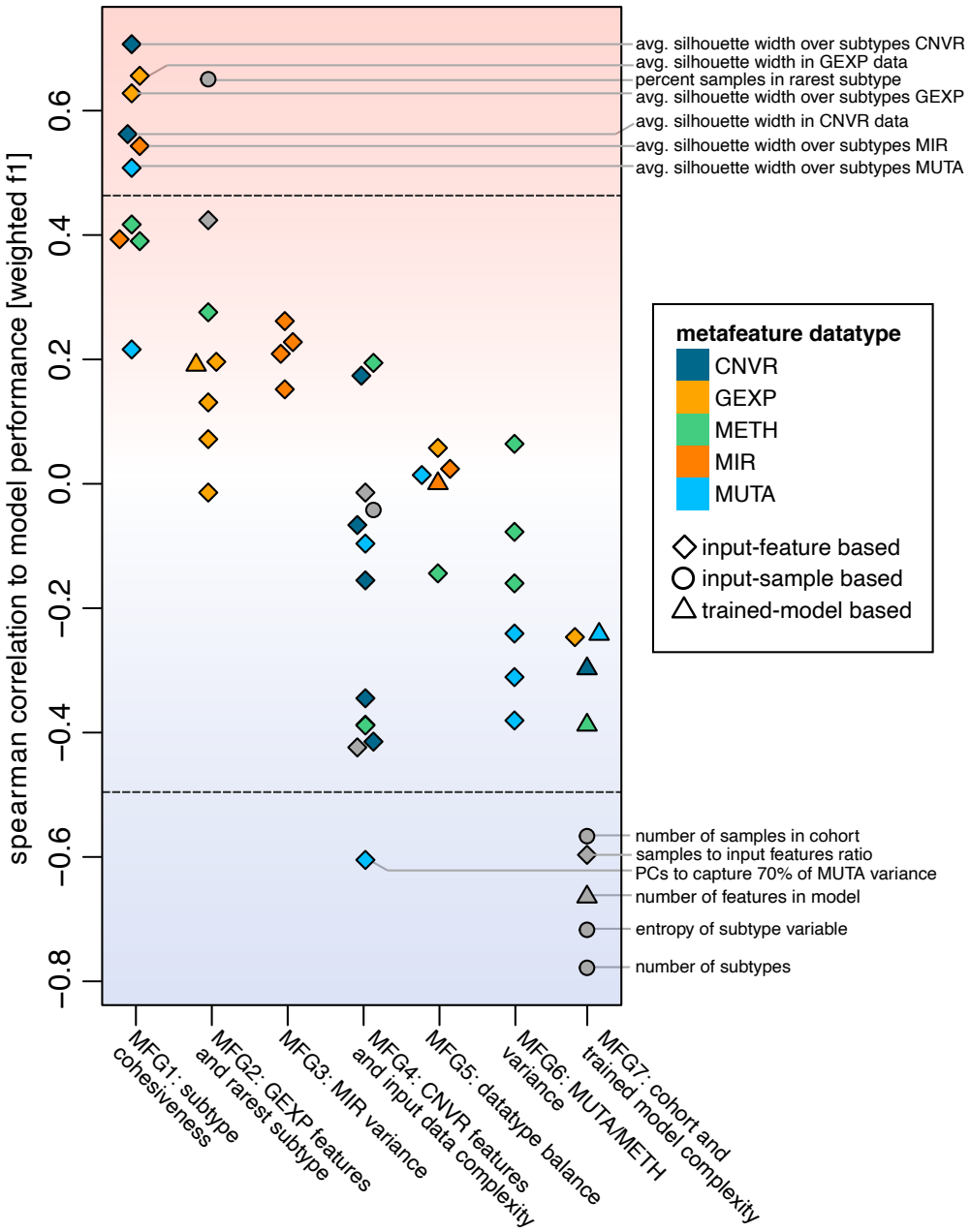


Figure 4.2: **Correlation of meta-features to subtype classifier performance.** Correlation between meta-features and model performance in 26 TCGA cohorts. Meta-features are grouped in 7 meta-feature groups (MFG1-7) by hierarchical clustering (see Fig. 4.1). Dashed horizontal lines mark significance thresholds at FDR-corrected spearman correlation p-values $\leq 0.05$ (above top line and below bottom line). Every significantly correlated meta-feature is listed on the right. (CNVR = copy number variation, GEXP = gene expression, MIR = microRNA, MUTA = mutation, METH = methylation, PCs = principal components).

with model accuracy. For six out of the ten meta-features in MFG1 this correlation passes the significance threshold, with an FDR-corrected p-value $< 0.05$.

One additional single meta-feature from MFG2 is significantly positively correlated with model performance: percent of samples in the rarest subtype. A low value in this meta-feature points to a very imbalanced distribution of samples over the subtypes in the cohort, whereas a high value shows a more balanced distribution. Notice also that the meta-feature that measures the absolute number of samples in the rarest subtype is uncorrelated to model accuracy (spearman correlation of -0.04). Therefore, it appears that the total number of samples for the rarest subtype is not indicative of model performance, but the relative distribution is.

MFG7, cohort and trained model complexity, has the strongest negative correlation to model accuracy. The five meta-features that pass the significance threshold deliver three major points. 1) The total number of subtypes and the entropy of the subtype variable describe the complexity of the prediction task at hand. In general, the more classes a classifier has to distinguish between, the harder the prediction task and the lower the expected model accuracy. 2) The total number of samples in the cohort and the ratio between samples and input features are also both significantly negatively correlated with F-1 score. These two meta-features might be confounded by the number of subtypes in a cohort, where cohorts with many samples are divided into more subtypes, or alternatively more samples were collected for very diverse cancer types. 3) The last significant meta-feature in MFG7 is the number of features that the machine learning algorithm chose to perform the prediction task for each cohort. During training, one goal was to minimize the number of features used in the model. Therefore, easier prediction tasks, or prediction tasks that had a few very distinct features available, did better. In compari-

son, prediction tasks that would have benefited from a larger set of features were restricted to a smaller set and might have had to compromise for a lower model accuracy.

Another single meta-feature from MFG4 is significantly negatively correlated with model performance: the number of principal components to capture 70% of variance in the mutation data. This is another measure of the complexity of the data set and sample cohort and a less complex cohort is indicative of better model performance.

I also want to mention which meta-features are not correlated to model performance in order to show that these seem to not heavily influence the expected performance of a subtype classifier. The input data variance is overall not strongly correlated to the model F1-score, as well as the number of input features and the data type balance in the input features. As mentioned above, the total number of samples in the rarest subtype is not correlated to model performance, instead a balanced sample distribution over subtypes seems to be important.

I was able to confirm some of the cohort-level meta-analysis results on a subtype-level. 17 of the cohort-level meta-features were suitable to be calculated on a per-subtype basis, e.g. the average silhouette width in a subtype, the variance of various data types, or the number of samples in a subtype. I used the same measure of model performance, weighted F1 score, on a per subtype basis and correlated it to the subtype-level meta-features. For the rest of the cohort-level meta-features it was not appropriate to calculate them per subtype, e.g. the entropy of the subtype variable or the percent samples in the rarest subtype. The subtype-level meta-features have a very similar relationship to the model performance compared to their respective cohort versions. The spearman correlation between cohort-level meta-features to F1 and subtype-level meta-features to F1 is 0.76 with a p-value of $3e{-}4$).

Overall, many results from this meta-feature analysis meet our prior expectations. However, a quantitative analysis like this one provides us with a way to test the assumptions we might have about the classification tasks at hand, as well as a framework to prioritize aspects of sample collections in the future. For example, one might try to create a very balanced cancer cohort instead of just trying to increase the number of samples in general while ignoring the cancer subtypes. The high positive correlation of silhouette scores with model performance indicates that it is very important to use data from data platforms that provide a very clear distinction between subtypes. Additionally, increasing the number of subtypes, while in some cases beneficial to capture more subtle differences between cancer samples, generally complicates the classification into these subtypes. For such cohorts, it might be beneficial to implement a step-wise hierarchical classification, with a few broad subtypes first and a more fine-grained classification in a second step.

The TMP study again indicated transcriptional data to be the most informative data type to use for most cancer types. The average silhouette score for gene expression data had, together with copy number variation data, the strongest positive correlation to model accuracy. And gene expression data alone was sufficient for accurate molecular subtype classification in most cohorts. In the following chapters in my thesis, I will continue to explore the different uses of transcriptional signals in cancer research.

# Chapter 5

# Defining a High-Risk Prostate Cancer Subtype Emerging from Cancer Treatment

The project described in this chapter was part of the West Coast Dream Team (WCDT) collaboration under the Stand-Up-To-Cancer Prostate Cancer Foundation. It was published in [2]. My main contributions are the training and application of the AR signature, the unsupervised sample analysis, and the training and application of the t-SCNC subtype signature.

## 5.1 Introduction

Prostate cancer is the most common cancer diagnosed in men in developed countries. Although overall survival rates are high, some cases are very aggressive. Small cell neuroendocrine prostate cancer (SCNC) is a highly aggressive subtype of prostate cancer, which constitutes less than 1% of newly diagnosed cases [4]. The treatment of metastatic castration-resistant prostate cancer (mCRPC) with androgen-receptor (AR)-targeting therapies has brought

significant clinical benefit. However, in some patients that acquire therapeutic resistance to AR-targeting therapy, a histologic subtype that resembles *de novo* small-cell prostate cancer emerges.

I was involved in the analysis of a cohort of mCRPC, an advanced form of prostate cancer, in which cases were previously treated with and acquired resistance to AR-targeting therapies. In this context, we showed that the incidence of SCNC increases to around 17%, and termed these cases treatment-emergent SCNC (t-SCNC). The following analyses define and characterize the t-SCNC subtype within the mCRPC disease.

The analysis is based on the mCRPC cohort, which consists of 202 patients of which 119 biopsies have mRNA-Seq expression data available. The samples underwent independent pathological review which resulted in a consensus pathological call for each evaluable sample, classifying the samples into pure adenocarcinoma, pure t-SCNC, or mixed subtypes. The samples have mutation calls available for a selected set of genes and were followed to collect survival outcome information.

## 5.2 Androgen receptor activity in mCRPC

The activity of the androgen receptor (AR) was measured in two different ways: (1) AR protein expression was analyzed using immunohistochemical (IHC) analysis. (2) In order to measure AR transcriptional activity, I trained an AR expression signature based on AR-positive cell lines in the presence of androgen vs. the absence of androgen [40]. The concordance of this classifier with a previously described AR activity signature [48] was $> 90\%$. Additionally,

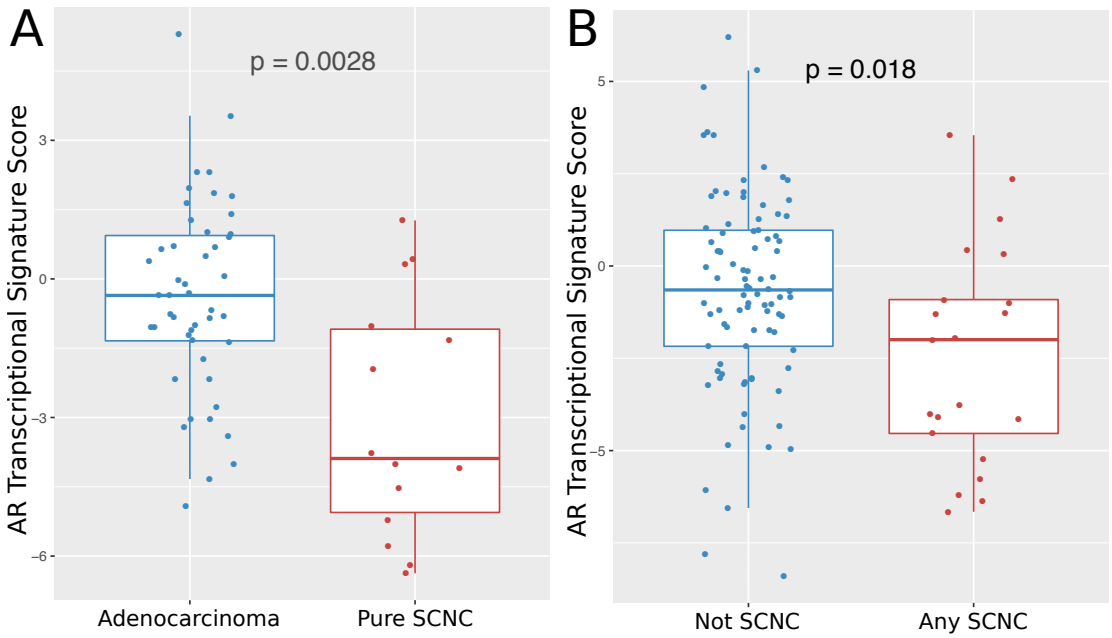


Figure 5.1: **Androgen Receptor (AR) activity measured by AR expression signature. (A)** Comparison of AR transcriptional signature scores between pure adenocarcinoma and pure SCNC samples. **(B)** Comparison of AR transcriptional signature scores between samples that presented with no SCNC pathology (Not SCNC) and samples that presented with either a pure or mixed SCNC pathology (Any SCNC). (p = t-test p-value).

2+/3+ nuclear AR expression by IHC and AR amplification were both positively associated with higher AR transcriptional signature scores. AR transcriptional activity was lower in tumors with t-SCNC histology (Fig. 5.1). The difference is strongest when comparing pure SCNC to pure adenocarcinoma samples (Fig. 5.1A). When all samples, pure and mixed, are considered, the samples with any SCNC pathology still have a significantly lower AR transcriptional signature score, but the difference is smaller (Fig. 5.1B).

I used the AR expression signature in two additional studies centering around mCRPC: (1) In [3], a subset of mCRPC samples was defined that has low serum Prostate-Specific Antigen (PSA) levels. This subset comprised 8% of all evaluable mCRPC samples. The incidence

of t-SCNC was elevated and AR transcriptional signature scores were lower in the low PSA secreting samples compared to normal secretors. The study overall concluded that the measurement of PSA secretion may be a readily available clinical selection tool for de-differentiated mCRPC cases with molecular features consistent with t-SCNC.

(2) In [7], AR activity was analyzed in relation to response to enzalutamide, an AR antagonist that is one of the principal treatments for men with CRPC. While AR genomic alterations and AR expression were similar in enzalutamide responders vs. non-responders, AR transcriptional activity was significantly lower in non-responders. In addition to a stemness-related program enriched in non-responders, the AR transcriptional activity could be used to enroll patients into clinical trials that try to overcome de novo enzalutamide resistance.

## 5.3 Unsupervised analysis of mCRPC defines a high-risk small-cell enriched molecular patient group

The gene expression profiles of all 119 mCRPC samples with RNA-Seq data available were used for unsupervised sample analysis. I performed hierarchical, complete-linkage clustering on the 5000 most varying protein-coding genes. The resulting sample tree was cut into five clusters and I performed an analysis of variance (ANOVA) over the 5 sample clusters to select genes for visualization and gene set testing (Fig. 5.2). The resulting 528 genes with an ANOVA FDR-corrected p-value $< 0.05$ were clustered using k-means with k = 5. Hypergeometric testing for gene sets showed the gene clusters are involved in (i) androgen response and metabolic processes; (ii) androgen response, AR activity and targets, and the FOXA1 network;

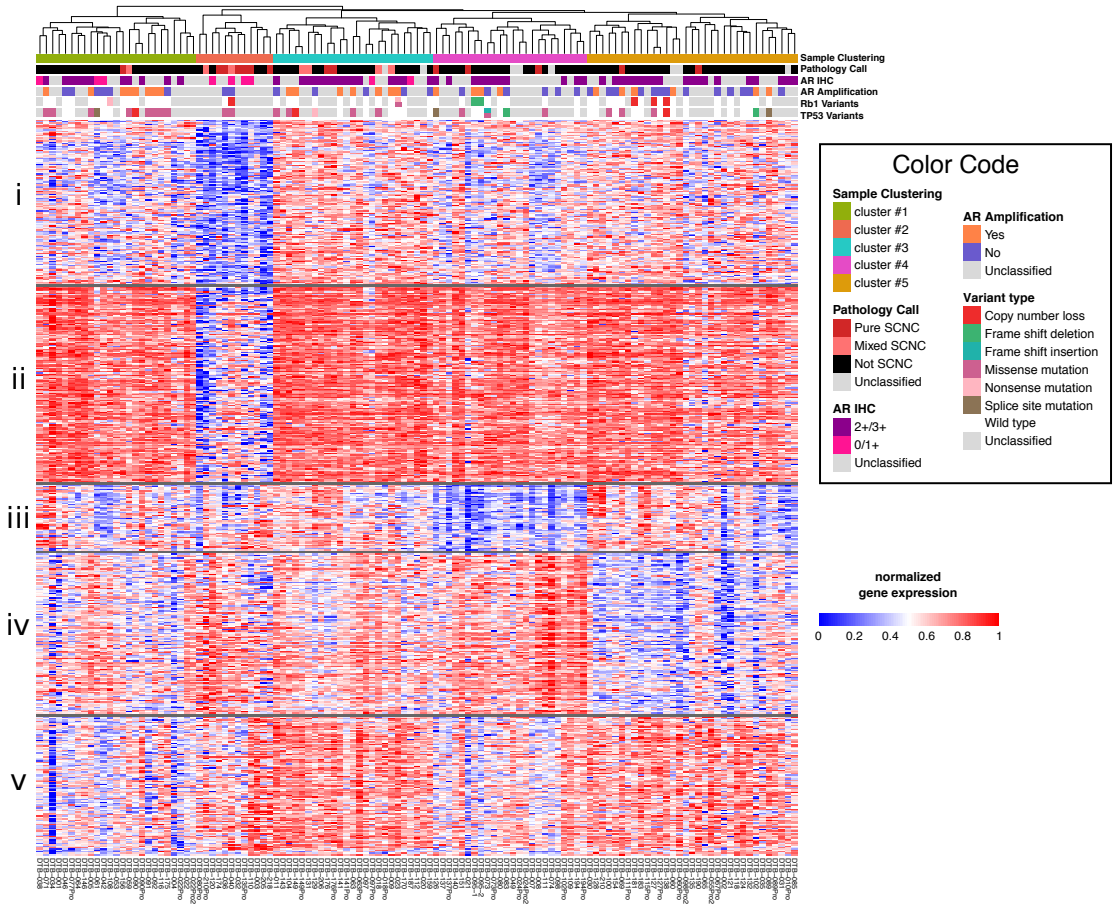


Figure 5.2: **Gene expression analysis identifies t-SCNC cases in unsupervised analysis.** Unsupervised hierarchical clustering of transcriptional profile of mCRPC biopsy specimens (n = 119). Sample cluster 2 is enriched for presence of t-SCNC histology. The rows show the normalized gene expression of 528 genes with false discovery rate (FDR)-corrected p-value $< 0.05$ for analysis of variance of gene expression in the five sample clusters. Genes are k-means clustered with k = 5 (labeled i-v). Pathology call, AR immunohistochemistry (IHC), and variant calls of TP53 and RB1 shown in annotation bars on top.

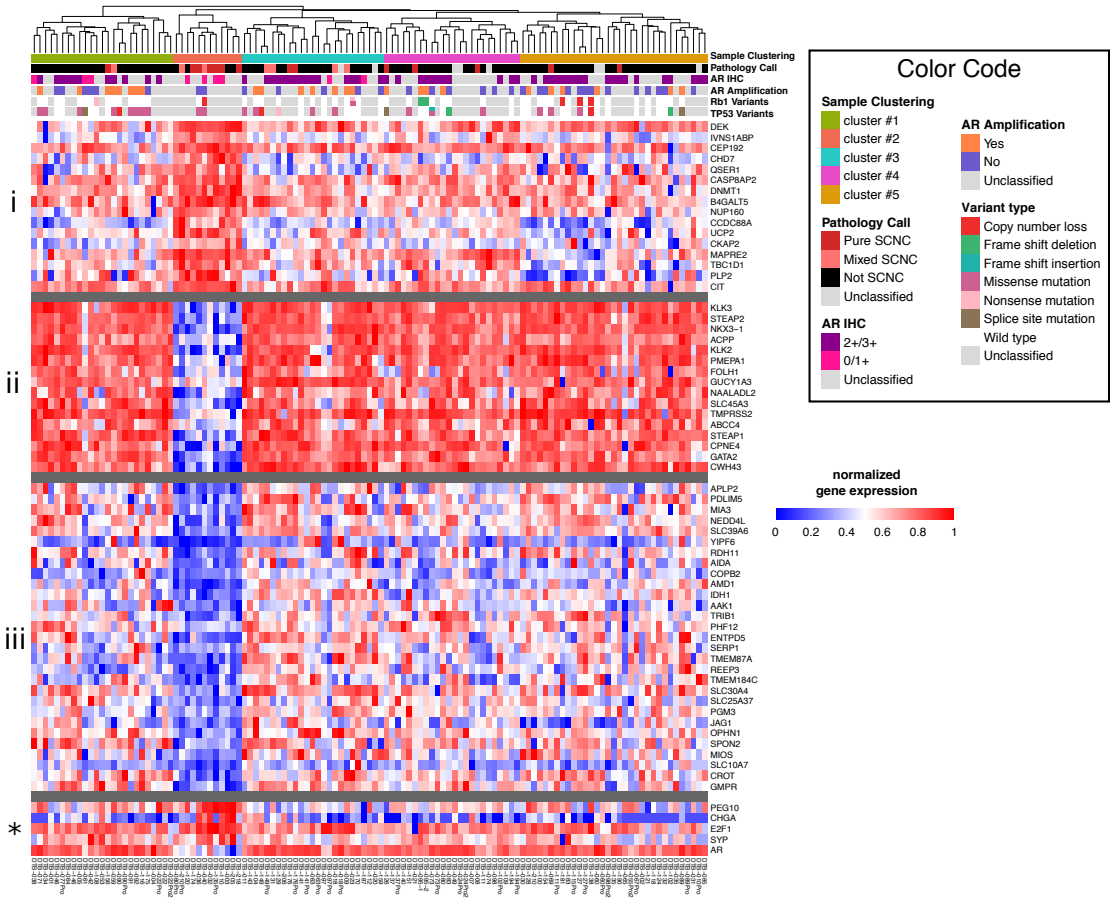


Figure 5.3: **Gene sets enriched in t-SCNC-enriched sample cluster.** Heatmap showing 61 genes with FDR-corrected p-value $< 0.05$ for t-test of gene expression in the t-SCNC–enriched cluster 2 versus all other samples. Genes are k-means clustered, with k = 3 (i-iii), in addition showing genes of interest PEG10, CHGA, E2F1, SYP, and AR (*).

(iii) translation; (iv) extracellular organization; and (v) cell cycle and transport.

Sample cluster 2 from this unsupervised analysis is enriched in t-SCNC cases: It consists of 12 samples of which 6 (out of 14, 43%) are pure t-SCNC samples, 2 (out of 7, 26%) are mixed t-SCNC samples, and only 4 (out of 90, 4%) are tumors with no t-SCNC pathology ($p < 0.001$). To further characterize the small-cell enriched cluster 2, I found the 61 differentially expressed genes between cluster 2 and all other samples (Fig. 5.3). The genes

are k-means clustered with k = 3. Hypergeometric testing for gene set enrichment showed gene cluster i, which is high in cluster2, contains genes of the Hallmark E2F Targets gene set. Cluster ii is dominated by genes related to androgen response and AR activity, and cluster iii contains genes of the Notch signaling pathway, with both gene clusters being low in cluster 2. The topmost overexpressed genes in cluster 2 are transcriptional targets of E2F Transcription Factor 1 (E2F1), which is negatively regulated by Retinoblastoma 1 (RB1). Further analysis with two different RB1 loss transcriptional signatures showed that t-SCNC–enriched cluster 2 indeed scores highest for RB1 loss.

Overall survival time in the t-SCNC samples is significantly shorter compared to samples without t-SCNC pathology, within the preplanned cohort of patients with prior AR targeting treatment (HR=2.02, Fig. 5.4A), as well as when post hoc including AR treatment-naive patients (log-rank p-value = 0.048). Patient samples in t-SCNC-enriched cluster 2 had a similarly worse survival (HR=2.20, Fig. 5.4B). When combining the pathology call and the transcriptomic data, patients with t-SCNC tumors falling within cluster 2 had significantly shorter survival than the patients without t-SCNC histology falling outside of cluster 2, with a greater separation of survival curves than either pathologic or genomic analysis alone (HR = 3.00, Fig. 5.4C).

This stronger trend observable from the transcriptional classification of samples, rather than the pathological classification, is seen again in a study in which we analyzed β2-adrenergic receptor (ADRB2) expression in the mCRPC patient cohort [43]. ADRB2 expression is slightly lower for t-SCNC samples compared to other samples, but the difference is not significant (p-value = 0.1). However, when comparing samples from the t-SCNC-enriched transcriptional cluster 2 to other samples, the difference in ADRB2 expression is highly significant

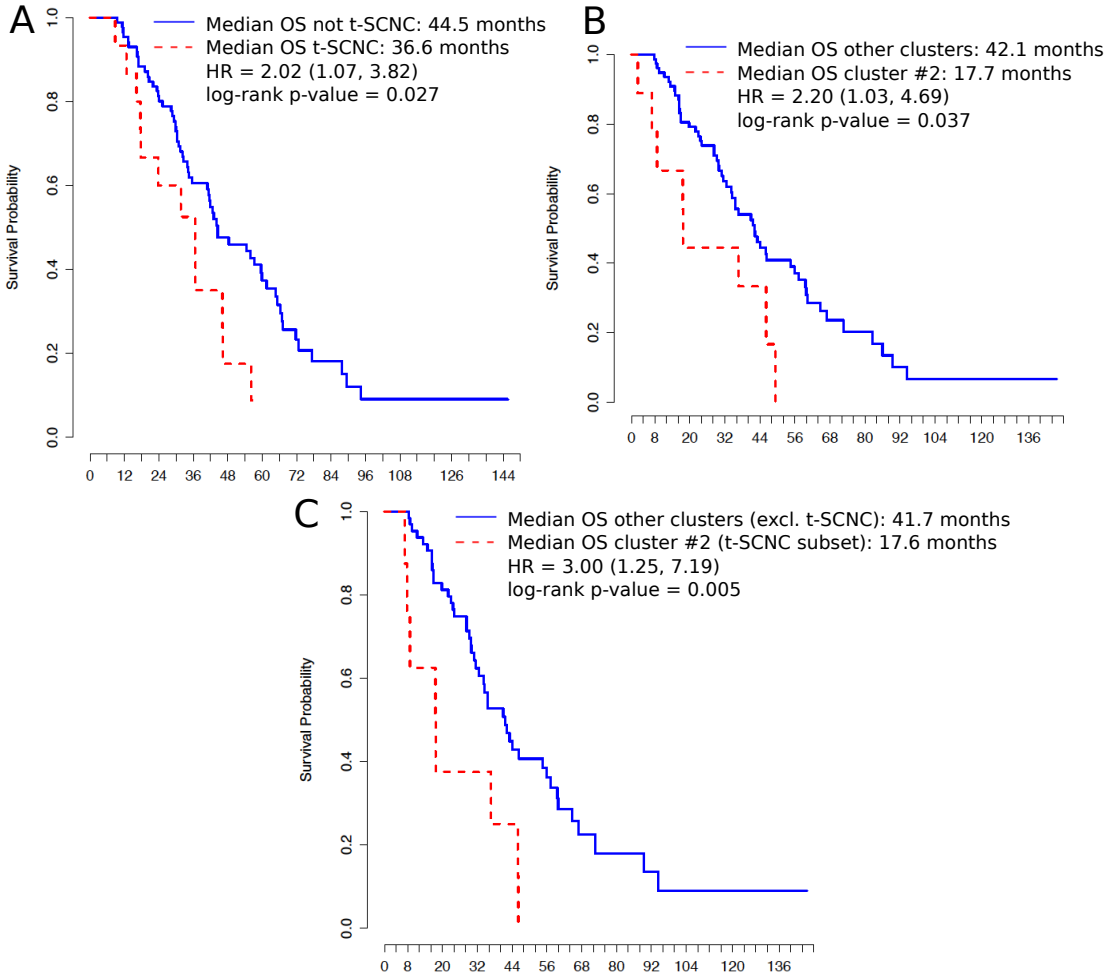


Figure 5.4: **Survival outcome by pathology and transcriptional sample clusters.** **(A)** Overall survival (OS) by pathology in the preplanned evaluable cohort of patients with prior AR targeting therapies. Comparing pure or mixed t-SCNC samples (red dashed line) to all samples not showing any t-SCNC pathology (blue solid line). **(B)** Survival by unsupervised transcriptional analysis (see Fig. 5.2 and 5.3), comparing samples in cluster 2 vs all other samples. **(C)** Comparison of survival between samples in cluster 2 that have t-SCNC histology vs samples in other clusters that do not have t-SCNC histology. (HR=hazard ratio, HR shown with 95% confidence interval).

(p-value $< 0.001$). A similar, but reverse result is seen for EZH2 expression, an upstream regulator of ADRB2. Overall, we concluded that the β-adrenergic signaling pathway including ADRB2 and EZH2 genes could serve as a therapeutic target in progressive mCRPC. Other studies have shown that ADRB2 down-regulation may be a marker of transdifferentiation to t-SCNC [17, 78]. Therefore EZH2 inhibitors should be investigated as a means to improve outcomes in mCRPC, potentially by reducing the risk of transdifferentiation into t-SCNC.

## 5.4 t-SCNC signature reliably classifies prostate cancer samples

Pathological review of tumor samples is a time-consuming and expensive task. We therefore aimed to develop a classifier for t-SCNC samples. I trained a t-SCNC signature based on transcriptional data using l2 penalized regression. RNA-Seq data from 18,538 protein-coding HGNC genes was used to distinguish t-SCNC samples from adenocarcinoma samples. Mixed pathology samples were excluded from training. A full leave-pair-out cross-validation was performed [6], testing all possible pairs of t-SCNC and adenocarcinoma samples used in training. In 91% of cases, the left-out samples were classified correctly.

I then applied the fully trained signature to mixed pathology samples (Fig. 5.5A), with mixed t-SCNC samples falling between pure t-SCNC and pure adenocarcinoma samples. Additionally, I applied the signature to different external data sets for validation. Signature scores for three external mCRPC data sets [14, 13, 59] are shown in Fig. 5.5B-D. In most cases, the samples are correctly classified based on the provided sample annotation, therefore proving the signature did not overfit to the training data and is applicable to external data sets. When

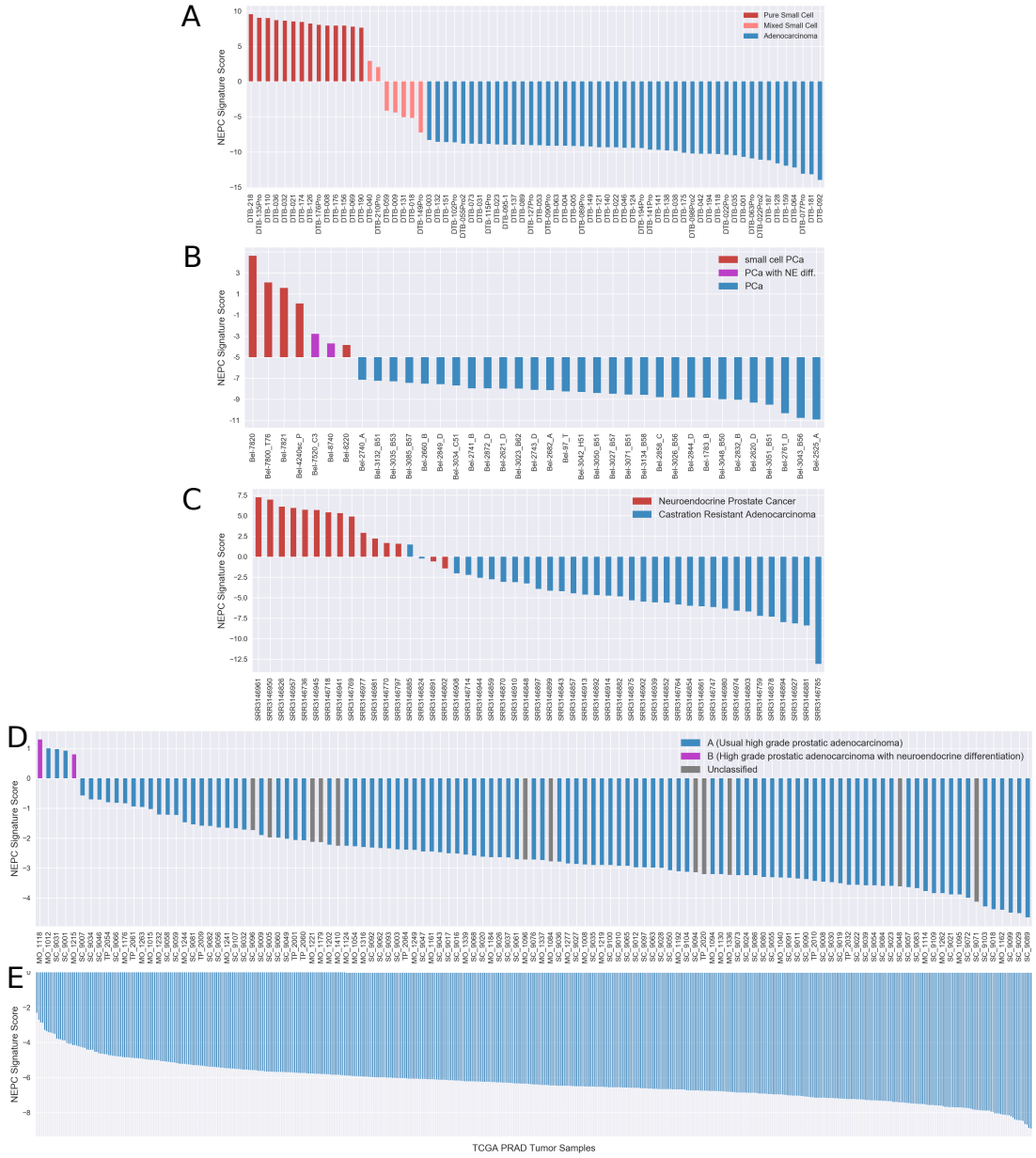


Figure 5.5: **Dense (l2 penalty) t-SCNC signature applied to different prostate cancer data sets. (A)** Samples from this study including training data (pure t-SCNC and adenocarcinoma) and samples with mixed t-SCNC pathology. **(B)** External mCRPC data set [14]. **(C)** External mCRPC data set [13] **(D)** External mCRPC data set [59] **(E)** TCGA collection of primary prostate cancers [1].

applying the t-SCNC signature to primary prostate cancer samples in the TCGA collection [1], all samples are correctly classified as adenocarcinoma.

In addition to the dense signature using all available genes, I trained a sparse signature using regression with l1 penalty in order to create a short list of signature genes for visualization and interpretation. Fig. 5.6 shows the l1 signature scores on training samples on the top and a gene expression heatmap of all 106 signature genes below. The signature genes are enriched for genes involved in neural development, especially in the signature genes with positive coefficients that mark t-SCNC samples.

## 5.5 Discussion

In this study, we characterized the mCRPC subtype of treatment-emergent small cell neuroendocrine prostate cancer (t-SCNC). It shows poor survival outcomes, with an even stronger separation of survival when the pathology call is combined with unsupervised sample clustering of transcriptional data. The t-SCNC subtype shows a distinct transcriptional signal which can be used for highly accurate classification.

But the classification of all mCRPC samples into two classes also has limitations. Samples with mixed t-SCNC pathology are scoring in the middle between pure t-SCNC and adenocarcinoma samples (Fig. 5.5A). We also showed that ADRB2 down-regulation is associated with t-SCNC pathology and it was suggested previously that it may be a marker of transdifferentiation to t-SCNC. This all points toward that a classification into two categories might not be an appropriate approach, and quantification on a continuum between established pure sub-

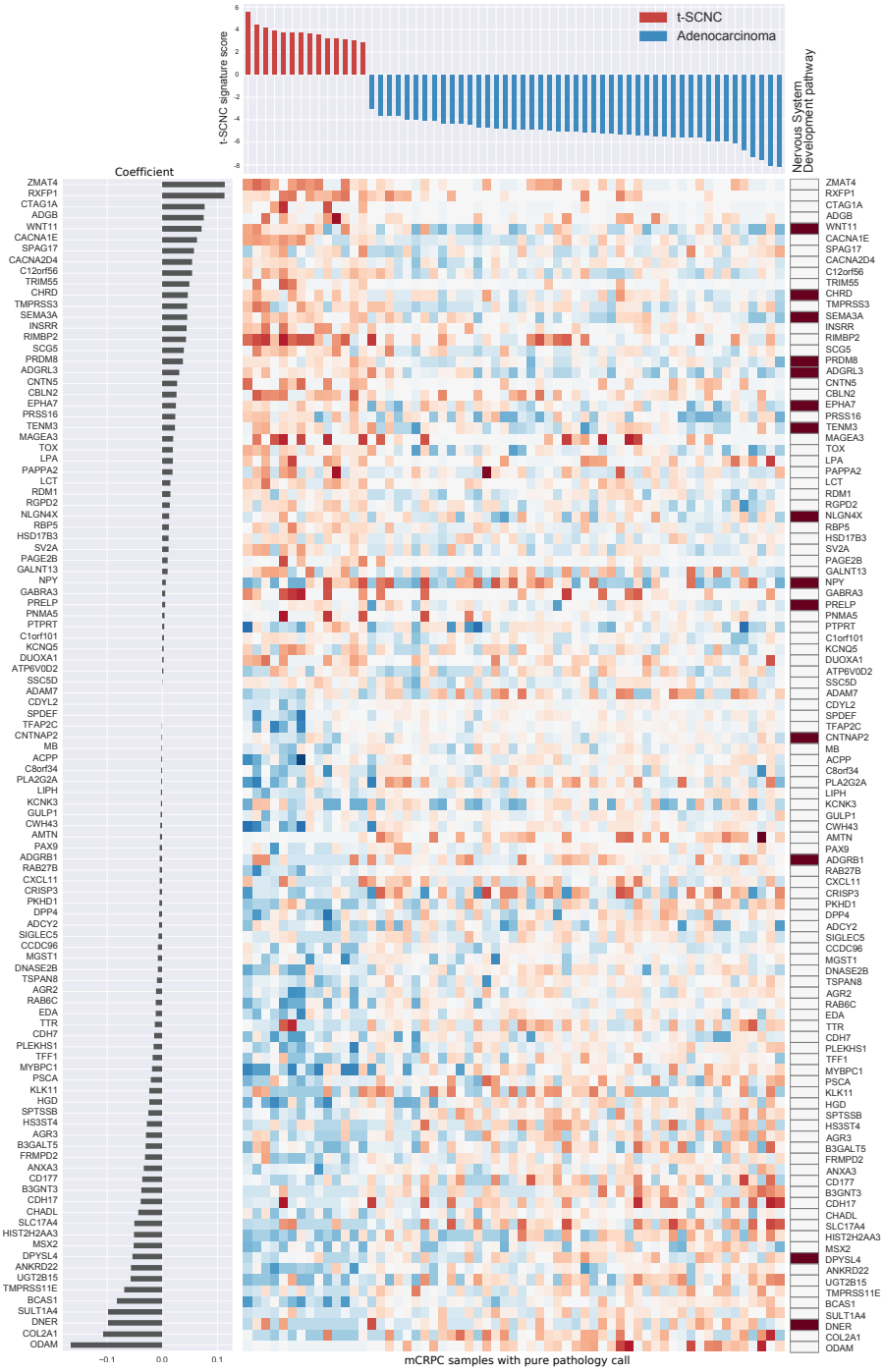


Figure 5.6: **Newly developed t-SCNC gene expression signature.** Waterfall plot on top shows sparse (l1 penalty) t-SCNC signature applied to the training data (pure t-SCNC and adenocarcinoma samples). The heatmap on the bottom shows median-centered gene expression of the 106 signature genes (high expression in red, low expression in blue). The membership of signature genes in neural system development pathways is annotated on the right, the model coefficient on the left.

types is necessary to capture samples that are either a mixture of subtypes or that have aspects of more than one subtype.

In the following section 5.6, I describe another cancer type, lung cancer, that shows a similar pattern of samples that lie between the more common subtype adenocarcinoma, and the more aggressive and rare small cell subtype.

## 5.6 Small cell-like appearance of lung adenocarcinoma samples

In a study of lung adenocarcinoma (LUAD) samples [19] within The Cancer Genome Atlas (TCGA) project, a cluster of samples was detected in which the samples harbor some characteristics of small cell lung cancer samples (SC-like). In order to determine if the samples in this cluster are in fact small cell cancers, I compared the samples to an existing data set of lung and prostate cancer samples ranging from normal to adenocarcinoma to small cell neuroendocrine samples for both cancer types [11]. In Fig. 5.7A, all samples are projected onto the first two principle components (PCs). PC1 represents the joint transition towards small cell cancer shared between lung and prostate cancer. Samples from the updated LUAD cohort (Carrot-Zhang2021) overlap well with the older version of a subset of those samples that were used in Balanis2019, showing that a joint analysis of these two data sets is possible. The SC-like LUAD samples tend to score higher on PC1, towards the small cell lung cancer samples (SCLC), but they still overlap mostly with other LUAD samples and do not form a separate cluster.

I also applied the t-SCNC signatures trained in section 5.4 to the updated LUAD

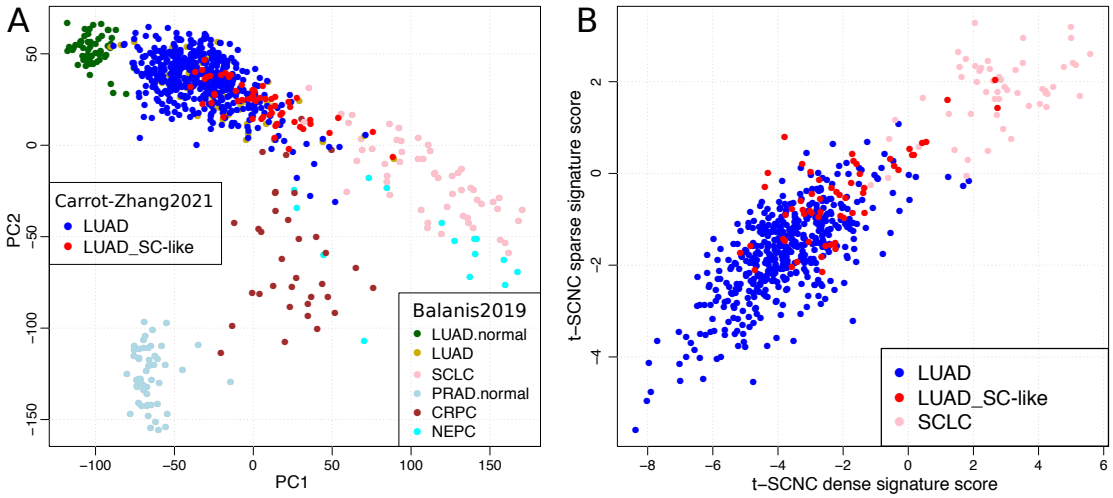


Figure 5.7: **Small cell signal in lung adenocarcinoma (LUAD) samples. (A)** Pan-cancer small cell convergence visualized on a principle component (PC) analysis 2D projection. Lung cancer and prostate cancer samples from Balanis 2019 study [11]. Updated and additional LUAD samples from Carrot-Zhang 2021 study [19] projected onto the same PCs. New LUAD samples are divided into two groups: small cell (SC)-like LUAD samples and rest of LUAD sample not showing SC signal. **(B)** Application of treatment-emergent small cell neuroendocrine prostate cancer (t-SCNC) signatures [2] (dense and sparse, see section 5.4) to lung cancer samples from A. (SCLC = Small cell lung cancer, PRAD = Prostate Adenocarcinoma, CRPC = Castration-resistant prostate cancer, NEPC = Neuroendocrine prostate cancer).

cohort, as well as the SCLC samples from Balanis2019 (Fig. 5.7B). The signature scores show a similar picture: the SC-like LUAD samples tend to score higher than the average LUAD sample, but they do not reach scores as high as true SCLC samples. The exception to this is three SC-like LUAD outliers (TCGA-86-8358-01, TCGA-55-6968-01, and TCGA-44-7667-01) that clearly score above 0 for both t-SCNC signature versions. These three SC-like LUAD samples score as high as true SCLC samples.

Because the SC-like LUAD samples did not show a difference in survival outcome compared to other LUAD samples, this subset of samples was not further investigated, but it is another example of a cancer type in which a subset of samples seems to exists in a mixture or transitional state in between well-defined subtypes.

# Chapter 6

# Defining the Transitional Spectrum between Prostate Cancer Subtypes

The project described in this chapter is part of the West Coast Dream Team (WCDT) collaboration under the Stand-Up-To-Cancer Prostate Cancer Foundation. The work is still subject to active research and will be published under the project lead of Eric J. Small and Joshua M. Stuart. My main contributions are the application of the AR signature, the intermediateness analysis of the mCRPC subtypes, the training of pairwise subtype classifiers, and the trajectory inference with gene set enrichment analysis in relation to trajectory pseudotime.

## 6.1 Introduction

As seen in chapter 5, prostate cancer has two established subtypes: the more common adenocarcinoma and the rare and highly aggressive small cell neuroendocrine carcinoma (SCNC). Under androgen deprivation therapy (ADT) pressure through treatments targeting the

androgen receptor (AR), metastatic castration-resistant prostate cancer (mCRPC) is believed to transition from adenocarcinoma to SCNC, a state we defined as treatment-emergent SCNC, or t-SCNC, which is more common than *de novo* SCNC (17% vs. 1%).

However, some samples in the mCRPC cohort can not be easily categorized into either adenocarcinoma or t-SCNC, because they present with a mixed or different pathological picture. In an effort to characterize these intermediate samples, our collaborators have defined a new set of pathological characteristics that describe a third subtype of mCRPC: intermediate atypical carcinoma (IAC). Samples in this subtype display a distinct set of pathological characteristics, which are a mix of adenocarcinoma characteristics and SCNC characteristics, but not a mixture of pure cells from either subtype. We hypothesize that the process of transition from adenocarcinoma towards t-SCNC during ADT involves IAC as an intermediate disease state. The analyses described in this chapter are based on an updated mCRPC cohort of 88 adenocarcinoma, 53 IAC, and 25 SCNC samples defined by consensus pathology call using the newly established IAC criteria.

## 6.2 AR transcriptional signature in mCRPC subtypes

In order to characterize the newly established IAC subtype, I applied the AR transcriptional signature described in section 5.2 to the updated mCRPC cohort. Adenocarcinoma samples have the highest level of AR activity, followed by IAC samples with a lower median AR score, but the difference is not significant (t-test p-value = 0.076, Fig. 6.1). The SCNC samples have a significantly lower median AR score compared to adenocarcinoma and IAC subtypes.

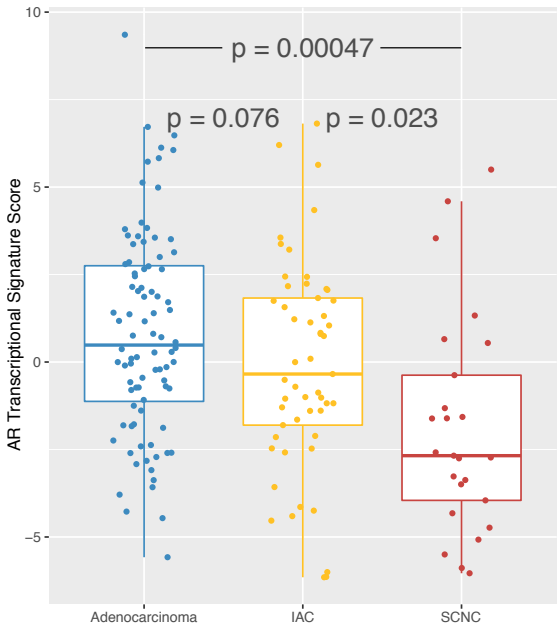


Figure 6.1: **Androgen receptor (AR) transcriptional signature** as described in section 5.2 applied to the three mCRPC subtypes. (p = p-value of pairwise differential t-test).

Analyzing results like this AR transcriptional signature score over the three mCRPC subtypes can hint at how the subtypes might relate to each other. However, the connection between the subtypes has to be analyzed more directly in order to understand if IAC really is an intermediate step between adenocarcinoma and t-SCNC. In the following sections, I attempt to find out if such a subtype ordering exists and derive its topology.

## 6.3 Distinguishing an intermediate cancer subtype from an independent subtype

There are two main options of how the IAC subtype could be connected to adenocarcinoma and SCNC: (1) IAC is an intermediate subtype between adenocarcinoma and SCNC, and

(2) IAC is an independent third subtype existing next to adenocarcinoma and SCNC (Fig 6.2A). In order to distinguish between these two scenarios, I applied a method based on Linear Discriminant Analysis (LDA) to the samples in the three subtypes. Other possible scenarios, i.e. a linear connection with a different ordering of the subtypes, are also implicitly tested by this approach.

LDA finds the direction of most inter-class variance while minimizing intra-class variance. I supplied LDA with RNA-Seq data and labels for the three subtypes: 88 adenocarcinoma, 53 IAC, 25 SCNC samples. For the full data set, IAC falls in the middle between adenocarcinoma and SCNC on the most important direction, the first linear discriminant (LD1, Fig. 6.2B). To verify the robustness of this observation, I analyzed 100 bootstrap samples of the data. With each of those bootstrap samples, I repeated the LDA approach, and in 94 out of 100 cases, IAC is again in the middle between adenocarcinoma and SCNC on LD1 (Fig. 6.2C). Compared to results for a random process, i.e. 33 out of 100 times, this result is highly significant (p-value $< 10^{-17}$).

## 6.4 Pairwise mCRPC subtype classification

Using LDA, I was able to show that IAC is an intermediate subtype between adenocarcinoma and SCNC. However, a distinction between either one of the outside classes and the intermediate IAC subtype is difficult. I built pairwise classifiers between the three subtypes, but only the adenocarcinoma vs. SCNC classifier reaches an acceptable accuracy level: The linear regression classifier using l1 regularization between adenocarcinoma and SCNC has a

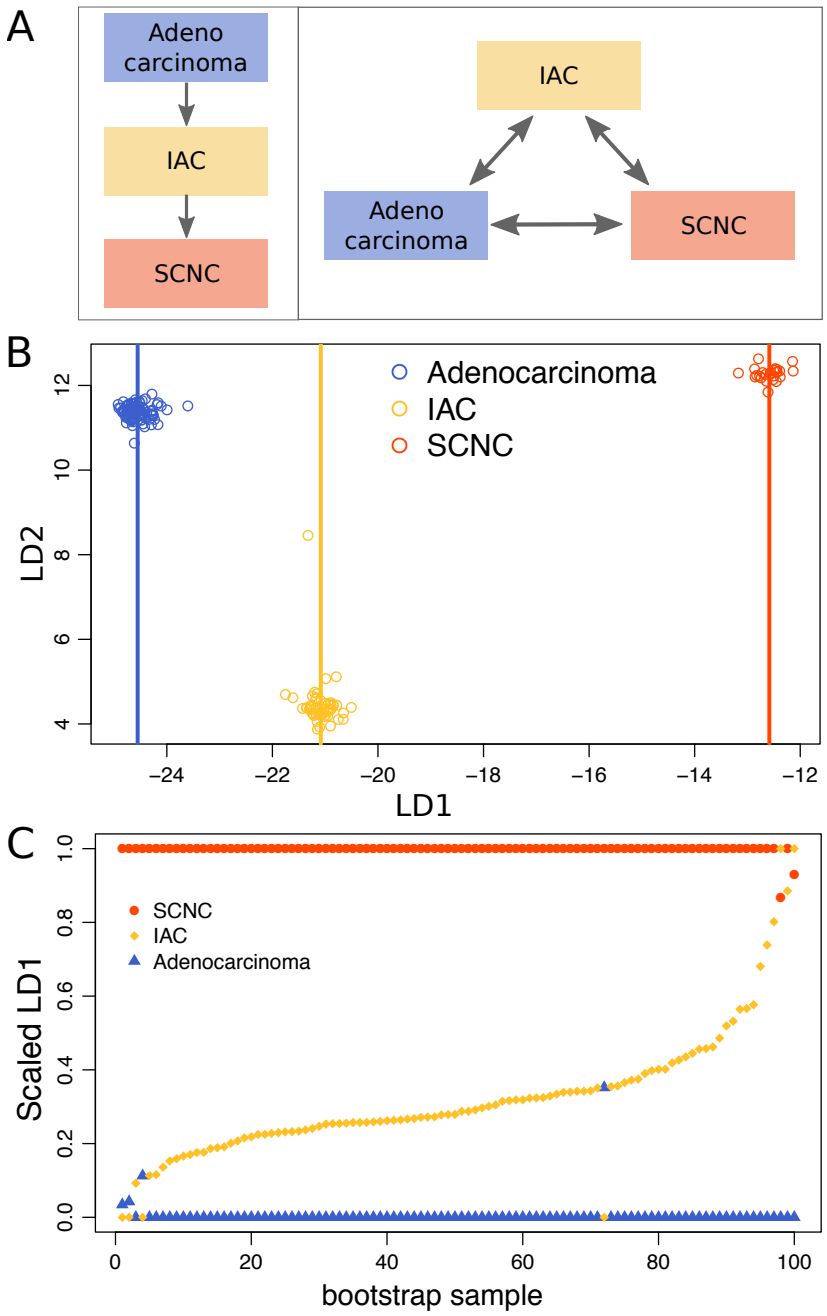


Figure 6.2: **IAC placement as additional mCRPC subtype.** **(A)** Two possible scenarios for the relationship between the three mCRPC subtypes. Right: IAC is an intermediate subtype between adenocarcinoma and SCNC. Left: IAC is an independent subtype next to adenocarcinoma and SCNC. **(B)** Linear discriminant analysis (LDA) of the three subtypes. **(C)** 100 bootstrap replicates of the analysis in B. LD1 is scaled between 0 to 1 and ordered so that adenocarcinoma is closer to 0. 94 out of 100 times IAC lands in the middle ($p < 10^{-17}$). (IAC = intermediate atypical carcinoma; SCNC = small cell neuroendocrine prostate cancer.)

| classes | LPOCV performance (l1) | LPOCV performance (l2) |
|---|---|---|
| SCNC vs. Adenocarcinoma | 84.9% | 89.5% |
| IAC vs. SCNC | 75.3% | 74.4% |
| IAC vs. Adenocarcinoma | 71.9% | 69.6% |

Table 6.1: **Binary classification performance between the three subtypes of metastatic castration-resistant prostate cancer.** L1 regularization producing a sparse signature, l2 regularization producing a dense signature (see section 5.4). (IAC = Intermediate Atypical Carcinoma, SCNC = Small Cell Neuroendocrine Carcinoma, LPOCV = Leave-pair-out cross-validation [6]).

leave-pair-out cross-validation (LPOCV) accuracy of 85%, whereas classifiers between IAC and adenocarcinoma, and IAC and SCNC only reach an LPOCV accuracy of 75% and 72%, respectively. The discrepancy for a l2-penalized regression is even more pronounced: the SCNC vs. adenocarcinoma classifier has an accuracy of 90%, but IAC vs. adenocarcinoma, and IAC vs. SCNC classifiers only have 74% and 70%, respectively (see Table 6.1). A possible reason for the failure of these classifiers is that the samples are lying on a continuous spectrum and do not form distinct groups. Samples that are already closer to a neighboring subtype can possibly confuse the classifiers and this would be consistent with the reduced classifier performance between adenocarcinoma and IAC as well as between IAC and SCNC.

## 6.5 A supervised trajectory approach defines the mCRPC subtype spectrum

I applied a trajectory inference method from single-cell RNA-Seq analysis to the bulk RNA-Seq WCDT data to find the trajectory connecting the three mCRPC subtypes and derive the ordering of the samples along this trajectory. The strongest signal in the WCDT RNA-Seq

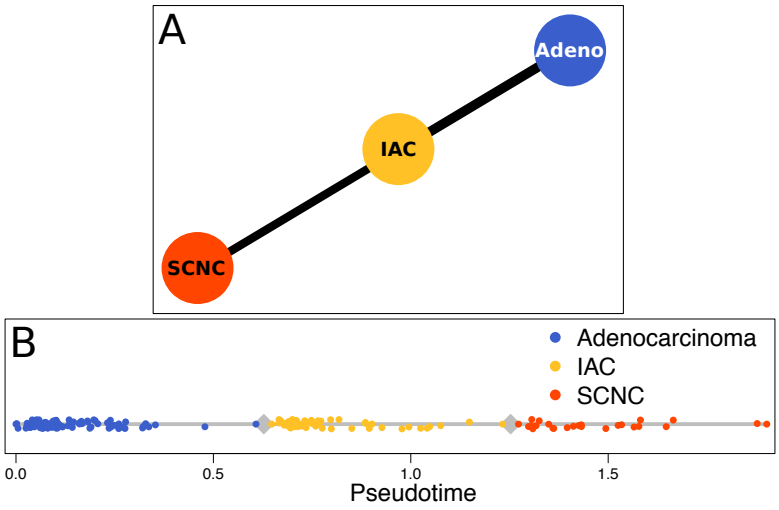


Figure 6.3: **Supervised PAGA Tree trajectory inference results.** Bulk RNA-Seq data was used as input with mCRPC subtype labels as prior information. **(A)** The trajectory topology between the three mCRPC subtypes as inferred by PAGA Tree. **(B)** Pseudotime ordering of mCRPC samples after inferring direction of the trajectory by providing a root sample (mean pseudotime over all orderings after providing each adenocarcinoma sample as root once).

data are not the mCRPC subtypes (see hierarchical clustering results in section 5.3 and [2]). Therefore, in order to infer a trajectory between the subtypes, I provided the subtype annotation as prior information to the trajectory inference method, i.e. I performed a supervised trajectory inference. I used the PAGA Tree trajectory inference method [76], because it showed good results in benchmarking experiments [61], it does not assume a specific trajectory topology, and it offers a manual input of sample groups in order to supervise the trajectory inference.

I provided PAGA Tree with the subtype labels as sample groups. The method then builds a trajectory branch for every group in the prior. It finds connections between branches and the ordering of the samples on the branches by a nearest neighbor approach. Similar to the results from the LDA method, the trajectory found by PAGA Tree has IAC as the intermediate class between adenocarcinoma and SCNC (Fig. 6.3A).

By default, a trajectory inferred by PAGA Tree does not have a direction. In order to infer a pseudotime ordering of the samples along the trajectory, a root sample can be provided, which is then used to define the starting point. I provided each Adenocarcinoma sample as a root sample once and averaged over all pseudotime results. The resulting sample ordering is shown in Fig. 6.3B.

In order to find biological processes that change along the trajectory, I correlated each gene's expression with the pseudotime value for each sample. I used spearman correlation for a rank-based, non-parametric approach. I then ordered the genes based on their correlation value and performed a pre-ranked gene set enrichment analysis (GSEA) [66] on the resulting gene list, using a comprehensive collection of gene sets [49] (downloaded March 1, 2021). The most enriched gene set in the positively correlated genes, i.e. genes that are low in Adenocarcinoma and increase along the full trajectory, was the Hallmark E2F Targets gene set (Figure 6.4, very top). The most enriched gene set in negatively correlated genes, i.e. genes that are high in Adenocarcinoma and decrease along the trajectory, was the Hallmark Androgen Response gene set (Figure 6.4, very bottom). Both of these gene sets were seen before in section 5.3 in relation to the SCNC-enriched cluster from unsupervised analysis.

It is possible that these global trajectory results are confounded by subtype groups rather than varying along the trajectory, e.g. if E2F targets are highly expressed in SCNC and lowly expressed in Adenocarcinoma samples the correlations along the trajectory would be positive, even though it is just a difference between two subtypes. In order to find processes that do not just vary between the different subtypes, but within the subtypes along the trajectory, I also correlated each gene only within the trajectory sections for each subtype. Figure 6.4

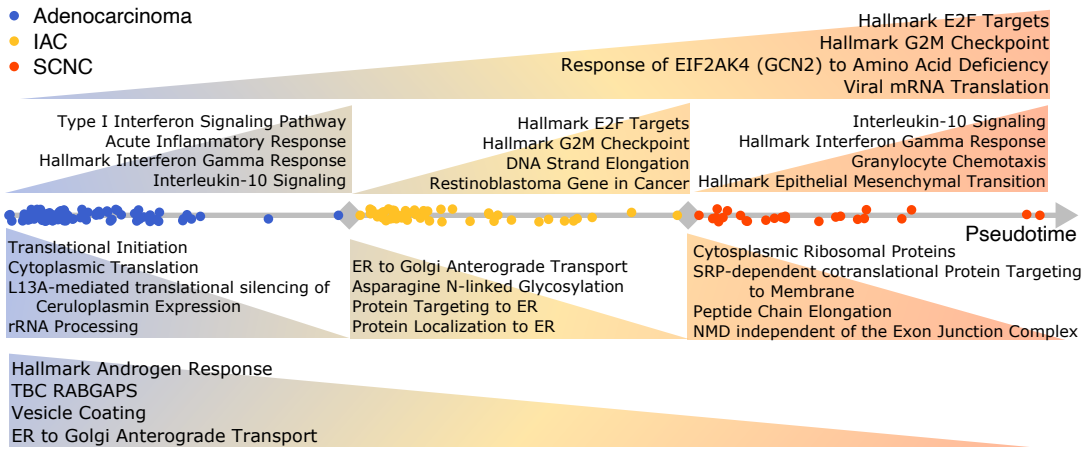


Figure 6.4: **GSEA along the mCRPC subtype trajectory.** Pathways enriched in genes that are positively (top) and negatively (bottom) correlated to the pseudotime along the full mCRPC subtype trajectory, as well as the single section within each subtype. The four most enriched pathways are shown for each section and direction.

shows the most enriched gene sets in both directions. The E2F targets are again the most enriched gene set in the positively correlated genes along the IAC trajectory section, therefore confirming that this is not just a trend seen between subtype groups. Additionally, gene sets that are involved in immune response and cell growth, e.g. interleukin signaling and interferon signaling, are enriched along the SCNC and adenocarcinoma trajectory sections, respectively. Pathways reducing activity along the Adenocarcinoma to IAC to SCNC trajectory are involved in translation and protein transport.

## 6.6 Discussion

Two independent methods applied to gene expression data from patient biopsy samples, i.e. linear discriminant analysis and supervised trajectory inference, showed that IAC is an intermediate mCRPC subtype between adenocarcinoma and SCNC. The results from train-

ing pairwise subtype classifiers have shown that the distinction between neighboring subtypes is challenging because the samples lie on a continuous spectrum rather than forming distinct groups. The trajectory inferred from a supervised trajectory inference approach was able to formalize this continuum and place the samples on it, assigning a pseudotime value to each sample. Biological pathways that increase activity along the Adenocarcinoma to IAC to SCNC trajectory are cell growth and immune response-related. Protein synthesis and transport-related pathways decrease in activity along the trajectory.

The cancer types analyzed above, advanced prostate cancer and lung cancer (see section 5.6), are both examples for cancers that show patterns of transition between subtypes, and the categorization in distinct groups seems not appropriate for all samples. This suggests developing reliable classifiers or signatures from pure examples to then apply to all samples in order to score those tumors that do not fit well into the established classes. There are different possible reasons for such transitional patterns to exist between tumor samples. A tumor might progress from one disease subtype to another as it is continually changing its genetic makeup. Additionally, the tumor microenvironment, the normal cells surrounding and mixing with the tumor cells, can change. For example, a tumor's immune infiltration might change, whether it is the amount or the type of immune cells that are present in the tumor microenvironment. In the following chapter, I present results that show the diverse microenvironment that is present in tumor biopsy samples and how this can affect patient outcomes.

# Chapter 7

# Dissecting Signals in the Cancer Microenvironment

## 7.1 Building a tumor immune infiltration map from consensus deconvolution estimates

The work described in this section is a contribution to the TCGA Tumor Deconvolution and Immunogenicity Analysis Working Group (TDI AWG). The project is still subject to active research and will be published together with The Cancer Genome Atlas Research Network under the project lead of Bhavneet Bhinder, Paul Spellman, and Olivier Elemento. My main contributions are the creation of *in silico* mixture samples for method validation (not shown) and the creation of the Tumor Immune Infiltration map described below.

Cancer is a disease involving the interplay of many different types of cells. Cancer cells are surrounded by the tumor microenvironment involving various stromal and blood cells,

including immune cells. Immune cells are usually programmed to kill abnormal cells, but cancers trick them into sparing tumor cells by disguising the abnormal cells or by deactivating the immune cells. This mechanism is called immune evasion. Recently developed immunotherapies have shown great promise in cancer treatment by recruiting immune cells to the tumor or activating immune cells present in the tumor, or both. It is crucial to determine how many and what types of immune cells are present within a tumor in order to foresee the effectiveness of such therapies.

The TDI AWG analyzes tumor immunogenicity through deconvolution in about 10,000 samples in the 33 cancer types of the TCGA project. The group used five established deconvolution tools that deconvolve bulk RNA-Seq gene expression data using marker gene sets or expression profiles. The estimates of each deconvolution algorithm were integrated for 42 different cell types, mostly immune cells and their progenitors, and a few high-level stromal cell types. Each cell type estimate was standardized across all samples. The standardized estimates for identical cell types were then averaged over all deconvolution algorithms that had the cell type in their reference. This resulted in 42 cell-type-specific integrated scores (iScores) for each tumor sample. Additionally, the specific cell type iScores were aggregated over all leukocytes to measure overall immune infiltration (leukocyte iScore).

Many of the estimated immune cell types show variable levels within and across different cancer types. Some immune cell types correlate with faster disease progression in certain tumor types, others correlate with slower disease progression. The following analysis results focus on taking a comprehensive pan-cancer look at the immune infiltration estimates. I projected the TCGA samples onto a two-dimensional landscape, using the 42 immune cell type iScores as

input to the Tumor Map tool [54]. The interactive Tumor Immune Infiltration (TII) map is available online (bit.ly/TIImap). Fig. 7.1A shows a snapshot of the full TII map, colored by TCGA cancer type. As expected, most samples cluster by their cancer type, and similar tumor types and tumors of related organs group together, e.g. squamous cell cancer types (CESC, HNSC, ESCA) and brain cancer types (GBM, LGG). The cell-of-origin signal in cancer molecular data is strong [36] and the deconvolution estimates are based on mRNA-Seq data.

The TII map also shows unexpected results, for example as shown in Fig. 7.1B-D. Breast invasive carcinoma (BRCA) and bladder urothelial carcinoma (BLCA) both separate into two clusters, and the two clusters of each tumor type are grouping together. This is interesting because BRCA and BLCA samples do not generally group together when looking at TCGA gene expression data [36]. Fig. 7.1B and C show the TII map colored by cancer subtype for BRCA and BLCA samples respectively. For both tumor types, the subtypes are not equally distributed between the upper and the lower sample cluster. For BRCA, the upper cluster is enriched in luminal A breast cancers, and the lower cluster is enriched in basal breast cancers. In BLCA, the upper cluster is highly enriched in luminal papillary samples, and the lower cluster is enriched in basal/squamous bladder cancers. One of the most differential iScores between the upper and lower clusters for both cancer types is the leukocyte iScore (Fig. 7.1D). The lower cluster shows a higher leukocyte iScore and therefore these samples seem to have a higher immune infiltration. The most differential cell-type-specific iScores for both BRCA and BLCA are CD8+ T cells, M1 macrophages, and Multi-Potent Progenitor (MPP) cells.

The BRCA and BLCA clusters show an imbalance in cancer subtypes, but they do not exclusively represent certain subtypes. This imbalance has to be considered for analyses

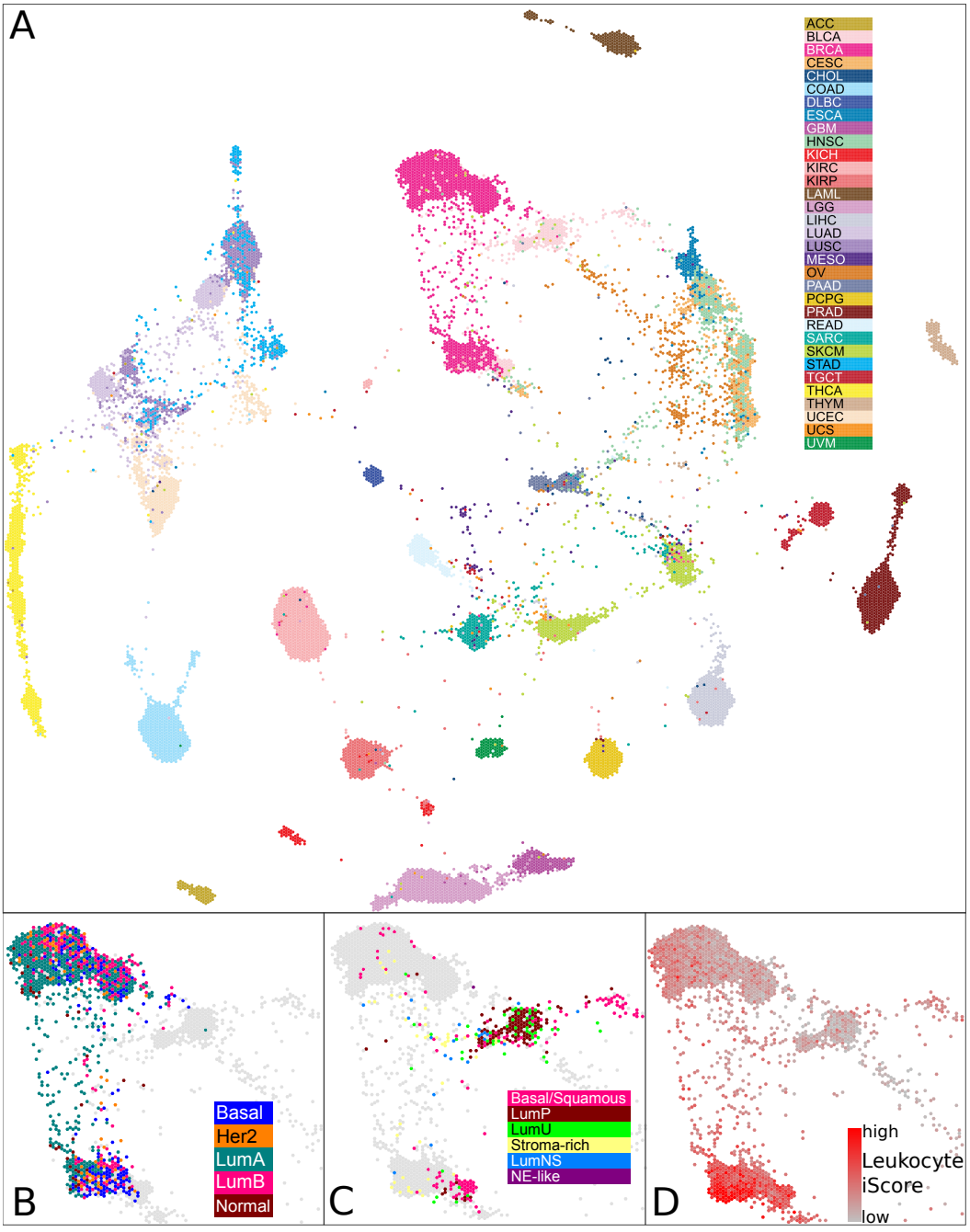


Figure 7.1: **Tumor Immune Infiltration (TII) Map (A)** Full TII map colored by TCGA cancer type. **(B)** Part of the TII map in A with breast invasive carcinoma (BRCA) samples colored with PAM50 breast cancer subtype. **(C)** Like B but bladder urothelial carcinoma (BLCA) samples colored with cancer subtype. **(D)** Like B-C but samples colored by the Leukocyte iScore, a composite score for leukocyte infiltration. Link to the interactive TII map: bit.ly/TIImap (Lum = luminal, LumP = luminal papillary, LumNS = luminal non-specified, LumU = luminal unstable, NE-like = Neuroendocrine-like)

following these results. Additionally, estimated tumor purity is lower in the BRCA/BLCA cluster with high leukocyte infiltration. This could be a confounding factor in the analysis and should be further investigated.

The estimation of immune infiltration using established marker genes and references was successfully applied to the TCGA cancer samples in this analysis. However, the cell types estimated in this analysis do not represent the full spectrum of cells in the tumor microenvironment. In the following section 7.2, I describe the definition of a comprehensive cell type library from single-cell data and their deconvolution results in TCGA cancer samples.

## 7.2 Using a comprehensive, single-cell sequencing derived cell type library for tumor deconvolution

The work described in this section is a joint effort between Yuanqing (Bianca) Xue and myself. The project is currently in preparation for publication.

### 7.2.1 Introduction

Single-cell RNA sequencing (scRNA-Seq), first described in 2009 [68], has since transformed biological research. For the first time, it is now possible to determine gene expression separately for each cell in a biological sample. The technology offers the opportunity for a more detailed and more accurate definition of cell types and cell states. In contrast to other technologies available to define cell types and their transcriptional profile, it does not rely on surface markers but directly offers the complete transcriptional information. scRNA-Seq has

already expanded our catalog of known immune cell types [73]. The work in this chapter leverages scRNA-Seq data to improve our understanding of cancer and help identify at-risk patient groups. We use scRNA-Seq data to detect newly defined cell types, represented by cell clusters, in cancer samples. We create transcriptional signatures for these clusters, evaluate the signatures with simulated mixture data, and then use the signatures to detect the clusters in cancer bulk RNA-Seq samples. We then analyze if the presence of these cell types in tumors influences the survival probability of patients.

### 7.2.2 Methods and validation

**scBeacon workflow** With the rapid development of scRNA-Seq technology, the amount of generated data is enormous. It is essential to build a computationally efficient pipeline to process large amounts of scRNA-Seq data while preserving the biological information within transcriptomics data. Motivated by this idea, we built scBeacon, a scRNA-Seq processing pipeline that rapidly clusters and integrates scRNA-Seq data sets.

The scBeacon workflow is shown in Fig. 7.2A. Starting from a library of scRNA-Seq data sets, each data set is clustered using the louvain algorithm [16]. We found louvain to have the best performance regarding speed and memory efficiency in different programming environments. For each cell cluster, the centroid is computed. We evaluated that centroids are accurately and robustly representing single-cell clusters containing at least 50 cells.

The integration of different data sets is not trivial due to batch effects caused by differences in sample handling, sequencing technology, etc. We use cell-wise rank normalization to reduce such batch effects. To validate this approach, we used 12 scRNA-Seq PBMC (periph-

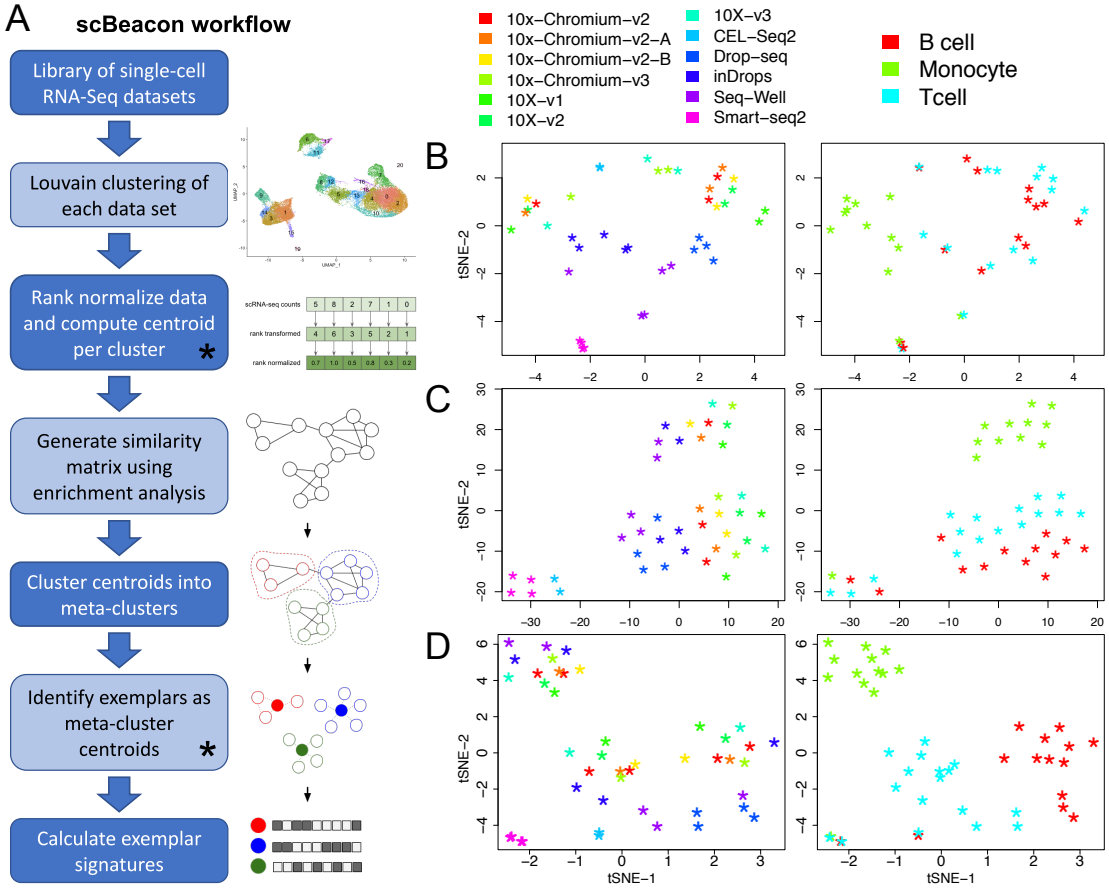


Figure 7.2: **scBeacon Workflow and Validation.** **(A)** scBeacon workflow. The asterisks mark steps with noise-filtering. **(B)** t-SNE plot of PBMC scRNA-Seq centroids from count-based data, tpm ( transcripts per million reads) transformed. Left plot is colored by single-cell sequencing technology, right plot by inferred cell type. **(C)** Same as B, but centroids are rank-normalized. **(D)** Same as C, but using scBeacon similarity matrix as input to t-SNE.

eral blood mononuclear cell) data sets from different sequencing platforms [25, 28]. PBMCs consist mainly of monocytes, B cells, and T cells, with other minor fractions of dendritic cells, NK cells, and macrophages [15]. We clustered each data set individually and assigned the three main clusters to monocytes, B cells, and T cells using the expression of established marker genes (CD3E for T cells, MS4A1 for B cells, and CD14 for monocytes). Fig. 7.2B shows a t-SNE plot of the cluster centroids based on gene expression counts. There is some clustering of centroids from the same cell types, but mostly centroids from the same data set cluster together. In Fig. 7.2C, the same centroids are shown, but the t-SNE plot is based on rank-transformed gene expression data. Apart from two exceptions, the centroids now strongly cluster by cell type.

In order to group centroids by unique cell types, we first transform centroids from gene expression space to pathway space using the Biological Process Activity (BPA) method [24]. It has been proven that this gene set-based transformation of scRNA-Seq data provides a robust description of cellular states and that it facilitates the alignment of cell types across different organisms. Here we use the same transformation to minimize the batch effect and align the same cell types across different tissues and technologies. After rank-normalization, the top 10 percent most active pathways in each centroid are used to perform pairwise enrichment analysis and similarity scores are computed between all centroids. Fig. 7.2D shows the PBMC centroid t-SNE projection based on this similarity measure. The three main cell types from nearly all data sets now form separate clusters. scBeacon then connects two centroids if their similarity score is in the 0.01 upper quantile of the empirical distribution of all similarity scores, and uses a graph-based clustering algorithm (iGraph [23]) to find meta-clusters within the centroid con-

nection graph. Each meta-cluster is considered to represent a cell type, and it can consist of multiple centroids from one or multiple data sets or be a unique cluster from just one data set.

The meta-cluster centroids, also called exemplars, are used as cell-type signatures. The exemplar cell-type signatures are used for deconvolution of cancer bulk RNA-Seq data, in which each signature's contribution to the mixture is estimated. We use the Cibersort deconvolution method [53], a widely used deconvolution method that recently placed first and second in the two sub-challenges of the Dream deconvolution competition [75]. Cibersort takes a signature matrix of cell types as input. We select signature genes by comparing the highest expressing exemplar to the second-highest expressing exemplar for each gene. The 20% genes with the largest difference are included in the signature matrix. This strategy ensures that only genes that are uniquely describing one cell type signature are included in the signature matrix.

**Deconvolution validation experiments using *in silico* mixtures** I created different types of *in silico* cell type mixtures simulating immune infiltration in cancer tissue in order to validate using the scBeacon exemplar signatures for deconvolution. I made 200 *in silico* mixtures from bulk RNA-Seq data for each of 7 different melanoma cell lines [57]. The cell line represents the pure cancer cell reference without microenvironment cells. The cell line is randomly assigned a mixture percentage of 50-90%. The rest of the mixture is randomly distributed between 6 immune cell types: B cells, dendritic cells, monocytes, NK cells, neutrophils, and T cells. The immune cell types were purified from blood using marker genes in a vaccination study [37]. I take the average of the 2 patients at time point t0 (before vaccination) to represent pure cell type references. For both data sets, the expression data were reduced to the overlapping genes

between the two data sets and quantile normalized to remove batch effects and enable mixing.

First, Cibersort was run with its default signature matrix, LM22, and more specific deconvolution results were summed for T cells, B cells, and monocytes. Fig. 7.3A shows the correlation between the mixed in cell-type proportions and the estimated cell-type proportions for a set of 200 *in silico* mixtures. While the correlation between true and estimated cell-type proportions is high, the method overestimates the immune cell contributions in all cases resulting in a high Root Mean Square Error (RMSE).

We then created cell-type signatures from scRNA-Seq PBMC data sets [25, 28]. Each PBMC data set was clustered using Louvain, clusters were assigned to cell types using marker genes, and a signature matrix was created as described above. First, only one data set and its count-based expression data were used for deconvolution of the *in silico* cancer mixtures (tpm-normalized, Fig. 7.3B). Then, we used the rank-normalized expression data of the same single scRNA-Seq data set (Fig. 7.3C). Finally, we used the rank-normalized PBMC meta-centroids from multiple scRNA-Seq data sets as shown in Fig 7.2D for deconvolution of the *in silico* mixtures (Fig. 7.3D). While there is not a consistent trend over all three immune cell types, the deconvolution estimate is closer to the mixed in proportion using the scRNA-Seq signature matrices resulting in a lower RMSE. While the correlation for T cells drops when using rank-normalized data, it is still on an acceptable level and the T cell proportions are not overestimated. The estimation of B cells seems to be an easier task, with the rank-normalized combined signature matrix giving the best estimate. The estimation of monocytes however seems to be more difficult in general, but the rank-normalized combined signature matrix reaches a comparable result to the default Cibersort run with a slightly lower correlation, but also a slightly lower

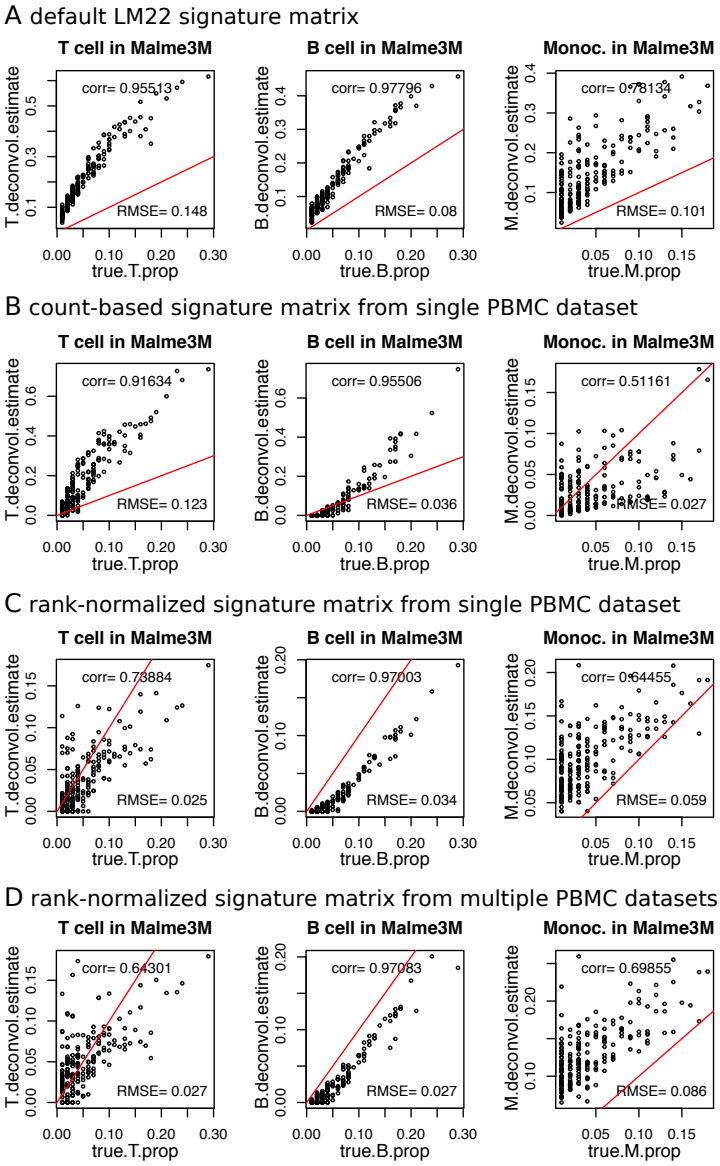


Figure 7.3: **Validation of deconvolution using single-cell derived signatures in *in silico* mixtures from sorted bulk RNA-Seq data.** **(A)** Correlation between the true mixture proportion of *in silico* mixtures from bulk RNA-Seq and the deconvolution results of using Cibersort's default signature matrix, LM22, summing more specific deconvolution results into T cells (T), B cells (B), and Monocytes (M). The synthetic mixtures are built from bulk RNA-Seq of the cancer cell line Malme3M and bulk RNA-Seq of purified immune cells. Red line marks the correct estimate (x=y). **(B)** Same as A, but using a count-based signature matrix from a single PBMC scRNA-Seq data set (10X-v2). **(C)** Same as A, but using a rank-normalized signature matrix from a single PBMC scRNA-Seq data set (10X-v2). **(D)** Same as A, but using a rank-normalized signature matrix from the combination of multiple PBMC scRNA-Seq data sets (10X chemistry v1-v3, CEL-Seq2, Drop-Seq, inDrops, Seq-Well). (RMSE = Root Mean Square Error, corr = pearson correlation)

RMSE.

Overall, we conclude that while the scRNA-Seq data does not show an overall better deconvolution result, we showed that it is possible to use scRNA-Seq derived cell-type signatures for the deconvolution of bulk RNA-Seq data with comparable results. The use of scRNA-Seq data, in comparison to sorted bulk RNA-Seq cell type references, has multiple independent advantages, e.g. it is independent of cell surface markers used for sorting, the whole transcriptional profile of cells can be used to define cell types, and it is able to detect rare cell types. Additionally, the rank-normalization and combination of multiple data sets do not hurt, and sometimes even help the deconvolution process. This is important because the integration and combination of multiple data sets are inevitable in order to build a comprehensive cell type signature library and make use of the many scRNA-Seq data sets that are already publicly available.

### 7.2.3 Building a comprehensive single-cell derived cell type signature library

The Single Cell Expression Atlas, a part of EMBL-EBI's Expression Atlas, is a public single-cell RNA sequencing data consortium that hosts data sets from published studies for six different species [55]. For this analysis, we downloaded the 62 homo sapiens scRNA-Seq data sets available in February 2020. The data sets come from a wide range of healthy and diseased tissues, consisting of numerous cell types in the human body, and were processed with different single-cell sequencing technologies. The EBI Single Cell Expression Atlas serves as the fundamental data resource for this project to build a comprehensive collection of human cell-type signatures, which are then used for bulk tumor deconvolution.

Fig. 7.4 shows the scBeacon workflow described in section 7.2.2 applied to the 62 human scRNA-Seq data sets. The centroids from all data sets are combined into 117 cell type exemplars (Fig. 7.4A). There are a large number of unique centroids that stay disconnected from others and build their own meta-cluster. However, there are also a few large meta-clusters combining many centroids from different data sets. These centroids either come from very specialized data sets that are first clustered, but later merged back together when considering that in the overall picture of the variety of cell types the clusters are very similar to each other. Alternatively, these centroids might also represent a common cell type that is found in many tissues, as is the case with immune cell types for example.

11 (9.4%) cell type signatures were implicated by two or more data sets, meaning clusters in the original data were merged into a meta-cluster. An interesting example of a meta-cluster being able to combine the same cell type from multiple data sets is shown in Fig. 7.4B-C. Meta-cluster X10 is a combination of centroids from five data sets (Fig. 7.4C) that all sequenced different states of pancreatic tissue. All centroids show an elevated expression of the insulin gene (Fig. 7.4B).

The meta-clusters from the human EBI Single Cell Expression Atlas are further processed with the scBeacon workflow described in section 7.2.2 to create a signature matrix to use in deconvolution (Fig. 7.4D). We use the cancer bulk RNA-Seq data available for 33 different tumor types from The Cancer Genome Atlas (TCGA) [36]. The HTSeq count data from NCI's Genomic Data Commons per tumor type was downloaded from Xena [29] and TPM (transcripts per million reads) normalized. The signature matrix is used with the Cibersort deconvolution method for each tumor type. Fig. 7.4D shows a general overview of the deconvolution results.

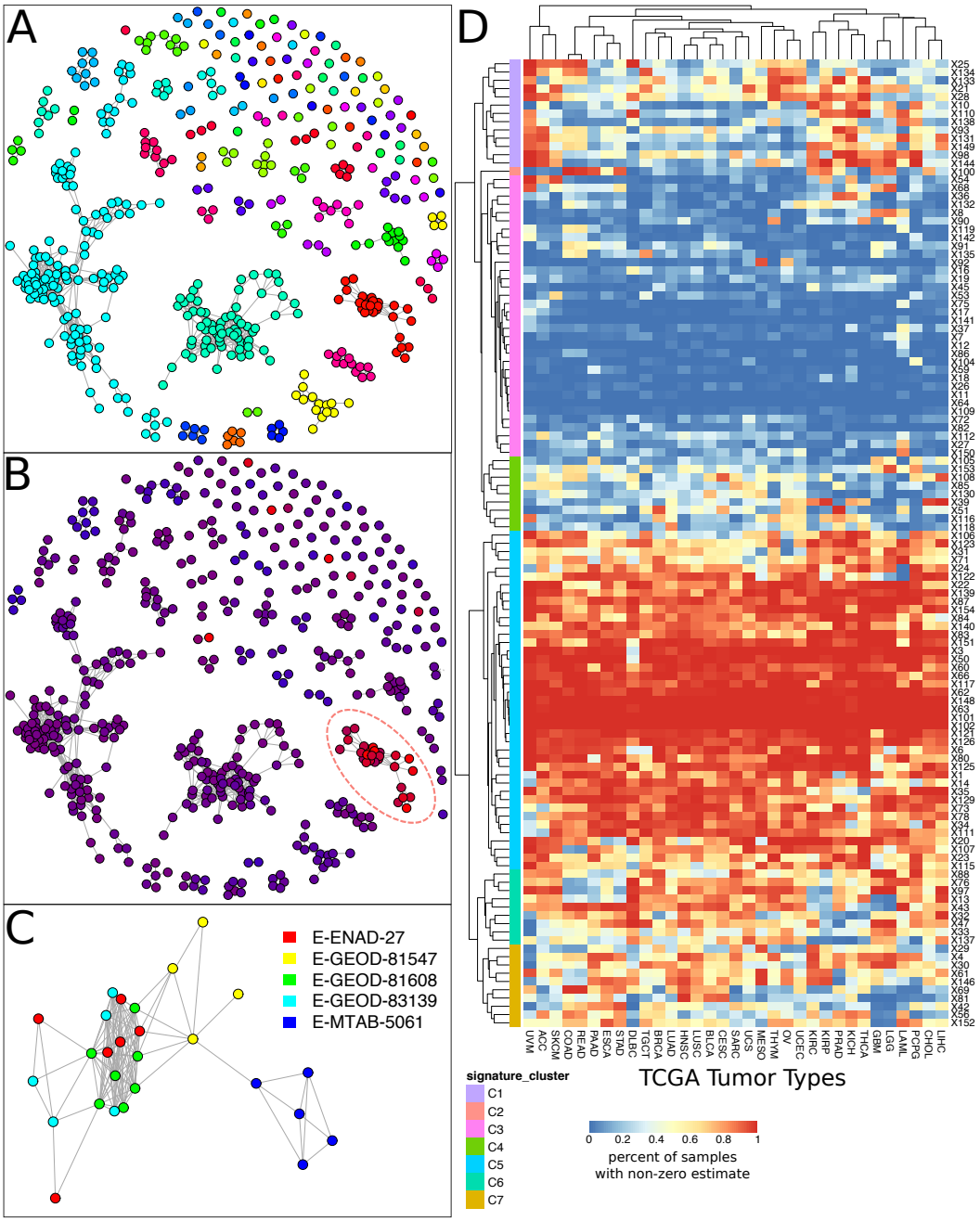


Figure 7.4: **scBeacon applied to human EBI Single Cell Expression Atlas. (A)** iGraph plot of meta-clusters: nodes represent centroids, centroids are only connected if their similarity score is within the top 0.01 quantile of all similarities, colored by meta-cluster grouping of centroids. **(B)** Same as A, but colored by the expression of the INS (insulin) gene from blue for low expression to red for high expression. Circled is meta-cluster X10. **(C)** Detailed iGraph plot for meta-cluster X10, colored by data set. **(D)** Deconvolution estimates of meta-cluster signatures in 33 TCGA cancer types. Heatmap shows the percent of samples in each cancer type that have a deconvolution estimate $> 0$.

As expected, many cell-type signatures are estimated to not be present in most tumor samples, and if they are detected, the estimated level is low. Therefore, I am showing the percentage of samples in each tumor type that has a cell type estimate greater than zero.

As expected, similar cancer types or cancer types from related organs show a similar profile of estimated cell-type signatures. The estimated cell-type profile for COAD (colon adenocarcinoma) is most similar to the estimated cell-type profile of READ (rectum adenocarcinoma). Similarly, LIHC (liver hepatocellular carcinoma) and CHOL (cholangiocarcinoma) are clustering together, as well as GBM (glioblastoma multiforme) and LGG (brain lower grade glioma), and a group of squamous cell carcinomas (HNSC = head and neck squamous cell carcinoma, LUSC = lung squamous cell carcinoma, BLCA = bladder urothelial carcinoma), CESC = cervical squamous cell carcinoma and endocervical adenocarcinoma). I have created a two-dimensional projection of the cell type signature estimates using the tool Tumor Map [54]. It is available for exploration on the Tumor Map web interface (bit.ly/TMEmap) and the results from a detailed analysis on this projection are described in section 7.2.5

A few cell types are detected in very few samples, e.g. X64, which is never detected. The cells in X64 stem from glioblastoma tissue (E-GEOD-84465). It is difficult to interpret such negative results because the deconvolution estimate can be zero due to technical reasons, e.g. too much noise or poor representation of the cell type in the signature matrix, or it could be due to actual biology, and the distinction between these possibilities is difficult. Another small set of signatures is detected, to some degree, in almost every tumor sample. Most cell-type signatures though are detected in some samples or tumor types, and not in others. These signatures might be more interesting to investigate, as they potentially mark samples from more

or less aggressive diseases. In the next section, I describe a methodical analysis investigating all cell-type signatures' influence on patient outcomes in each TCGA tumor type.

### 7.2.4 Cell-type signatures that correlate with patient outcomes in single tumor types

As seen in Fig. 7.4D, all 33 tumor types show a variety of different estimated levels of many cell-type signatures. In order to find a set of cell-type signatures that show different levels across samples of the same tumor type, we try to fit each cell type deconvolution estimate in each tumor type to a bimodal distribution. If the deconvolution estimates do not fit a bimodal distribution, the signature is not further analyzed for survival differences. If a fit to a bimodal distribution is found, the two modes of that distribution are used to separate the samples of the tumor type into two groups: an 'up' and a 'down' group, representing samples with little content of that signature vs. samples with a lot of content of that signature, with respect to the tumor type in question (Fig. 7.5B).

Survival curves are fit to the 'up' and 'down' sample groups using the R package survminer. We use progression-free interval (PFI) to measure survival for all TCGA tumor types, except for the Acute Myeloid Leukemia (LAML) cohort, which only has overall survival (OS) available [46]. To measure the separation between the two sample groups, we use the Cox Proportional-Hazards (CoxPH) Model and report the hazard ratio between the two groups with a 95% confidence interval and the p-value of the log-rank test. We report the results for a 'naive' signature outcome separation (SOS), which is a univariate CoxPH model. We also curated subtype annotations for all tumor types, mostly from the TCGA PanCanAtlas project [36] and

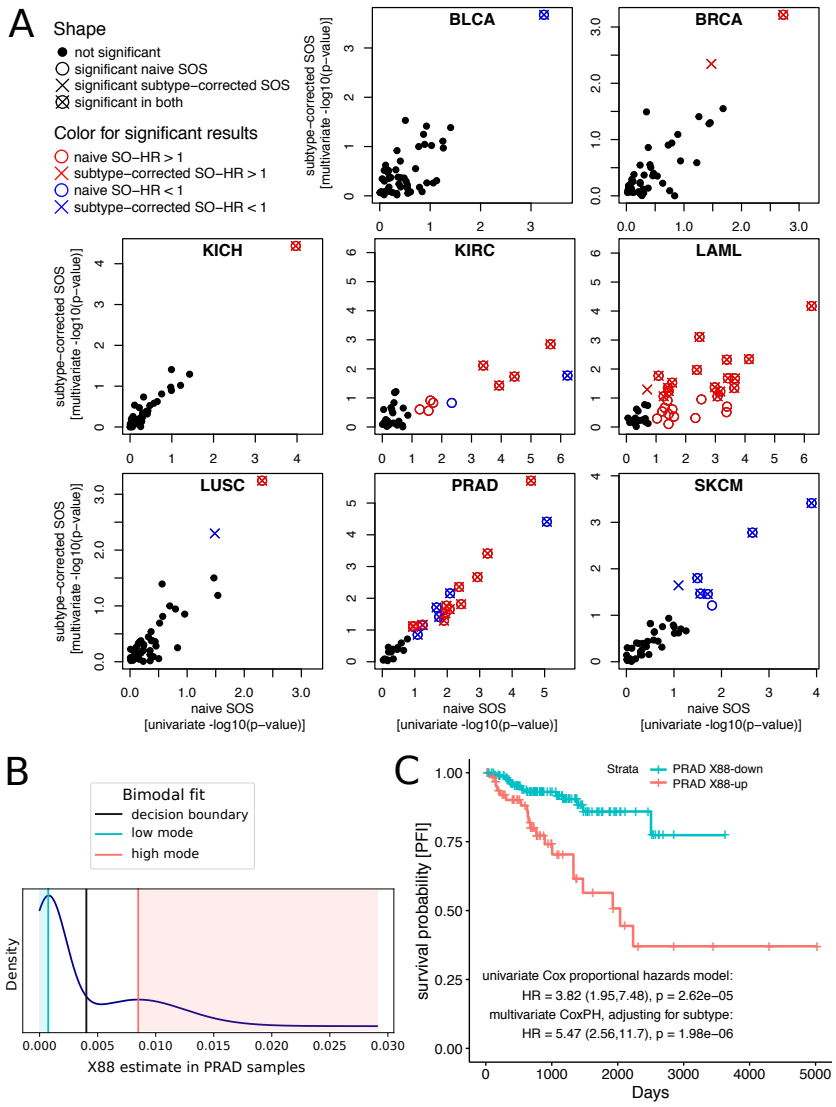


Figure 7.5: **(A)** Signature outcome separation (SOS) in selected TCGA tumor types. Plotted are log10(p-value) (not multiple-testing corrected) from log-rank test in a CoxPH survival model between 'up' and 'down' sample groups for each signature that passes the bimodal distribution test. The results from univariate CoxPH models are on x, the results from multivariate CoxPH models, correcting for unbalanced cancer subtypes, are on y. Results are considered significant with an FDR-corrected p-value $< 0.25$. Survival models which show significant separation with a signature outcome hazard ratio (SO-HR) above 1 are colored red ('up'-group has worse chance of survival). Survival models with significant separation with SO-HR below 1 are colored blue ('up'-group has a better chance of survival). Outcome is measured in progression-free interval for all tumor types, except for LAML which only has overall survival available. **(B)** Bimodal distribution fit of signature X88 estimate in PRAD samples. The samples in the shaded area below the low mode are defined as the X88-down group, samples in the shaded area above the high mode as X88-up group. **(C)** Kaplan-Meier survival plot for X88-down vs. X88-up group in PRAD.

| tumor type | signature | subtype-cor. SOS [fdr-adj pval] | subtype-cor. SO-HR | naive SOS [fdr-adj pval] | naive SO-HR | subtype OS [fdr-adj pval] |
|---|---|---|---|---|---|---|
| BLCA | X102 | 0.00998 | 0.36 | 0.02599 | 0.49 | 0.0136830 |
| BRCA | X82 | 0.02854 | 2.45 | 0.08853 | 2.23 | 0.1172773 |
| KICH | X101 | 0.00176 | 16.37 | 0.00513 | 11.15 | 0.0942052 |
| KIRC | X117 | 0.04684 | 2.47 | 0.00004 | 3.42 | 0.0000002 |
| LAML | X131 | 0.00303 | 4.02 | 0.00003 | 4.28 | 0.0001183 |
| | X123 | 0.01762 | 5.25 | 0.01186 | 2.45 | |
| LUSC | X116 | 0.02630 | 5.73 | 0.22363 | 4.02 | 0.7260900 |
| PRAD | X88 | 0.00006 | 5.47 | 0.00042 | 3.82 | 0.2334339 |
| | X98 | 0.00062 | 0.23 | 0.00027 | 0.22 | |
| | X62 | 0.00414 | 3.84 | 0.00609 | 3.19 | |
| | X13 | 0.01723 | 3.13 | 0.00934 | 3.04 | |
| | X148 | 0.02810 | 2.91 | 0.02310 | 2.82 | |
| | X100 | 0.03706 | 0.37 | 0.03363 | 0.39 | |
| SKCM | X22 | 0.01585 | 0.44 | 0.00522 | 0.42 | 0.1172773 |
| | X84 | 0.03415 | 0.56 | 0.04572 | 0.59 | |

Table 7.1: **High-confidence signature outcome separation (SOS) results.** All results with a subtype-corrected SOS (multivariate CoxPH model with covariate tumor subtype) FDR-adjusted p-value $\leq$ 0.05. SO-HR = Signature Outcome Hazard Ratio, OS = outcome separation. A SO-HR $>$ 1 indicates that the signature 'up'-group has a worse survival outcome, a SO-HR $<$ 1 indicates that samples in the 'up'-group have less risk on average.

TCGAbiolinks [20], except for DLBC (diffuse large B-cell lymphoma), which has no subtype information available. This subtype information was used as a covariate in a multivariate CoxPH model per tumor type in order to measure SOS between the cell type 'up' and 'down' groups while correcting for a potential imbalance in subtypes.

Table 7.1 lists all tumor type – signature pairs that have a subtype-corrected signature outcome separation (SOS) with an FDR-adjusted p-value $\leq$ 0.05. A total of eight tumor types has at least one signature with a significant SOS that could not be explained by the subtyping

of that tumor type.

The overall strongest SOS is measured for signature X88 in Prostate adenocarcinoma (PRAD) with a highly significant subtype-corrected SOS p-value and a SO-HR (signature outcome hazard ratio) of 5.5. Fig. 7.5B shows the bimodal distribution fit to X88 deconvolution estimates in PRAD samples. Fig. 7.5C shows Kaplan-Meier survival curves for the 'up' and 'down' sample groups defined from Fig. 7.5B. Signature X88 is derived from organoid cells (E-HCAD-5). All cells in the experiment are annotated with the same cell type: neural cell from skin, induced pluripotent stem cell (iPS cell line wibj_2), grown in organoid culture. A biological pathway activity [24] analysis of the X88 signature reveals that the most enriched pathways are neural development related, e.g. 'GO Forebrain Generation of Neurons'. This might be an indication that the prostate adenocarcinoma samples that have a higher estimate of signature X88 (X88-up group) are heading in the direction of the neuroendocrine subtype of prostate cancer. TCGA specifically only included primary prostate cancer samples of adenocarcinoma pathology, but, as described above in chapters 5 and especially 6, there seems to be a continuum between different prostate cancer subtypes. The neuroendocrine subtype is a more aggressive subtype. Therefore TCGA PRAD samples that exhibit neurodevelopmental signals might be on the far end of this spectrum, and the worse patient outcome for the X88-up group would be in line with that assumption.

Fig. 7.5A shows the correlation between the naive outcome analysis for the cell-type signatures with the subtype-corrected analysis. The outcome analysis of some of the cancer types shown seems to be heavily subtype-dependent. For example, many signature outcome separations (SOS) for KIRC (kidney renal clear cell carcinoma) are strong in the naive anal-

ysis, but after correcting for the subtype imbalance, the SOS are much weaker, even though still significant in a few cases ($p \leq 0.25$). Table 7.1 lists the outcome separation (OS) between subtypes and KIRC has a very strong separation. LAML (acute myeloid leukemia) is another example with a strong OS by subtype, but the influence on the SOS is mixed, with some signatures having a similar OS when comparing the naive to the subtype-corrected analysis. For other cancer types, like PRAD, LUSC (lung squamous cell carcinoma), and SKCM (skin cutaneous melanoma), the subtype has very little influence on the SOS analysis results. In some instances, the SOS is stronger in the subtype-corrected survival model, e.g. X82 in BRCA (breast invasive carcinoma) (see Table 7.1). This means that subtypes are imbalanced between the signature 'up' and 'down' groups, but contrary to the survival separation, and the SOS in the naive model is masked by imbalanced subtypes, instead of exaggerated.

The cells in signature X82 stem from a data set of CD45+ human fetal kidney samples (E-HCAD-10). Most of the cells in the cluster that built signature X82 are unannotated by the authors, and the few that are annotated are very mixed. We suspect these cells might be non-differentiated cells and are therefore difficult to annotate with a cell type. The signature with the second strongest SOS in BRCA seems to support this suspicion. Signature X107 stems from a data set of stem-cell-derived retinal ganglion cells (E-MTAB-6108) and consists of embryonic stem cells and retinal ganglion cells. The SOS for X107 in BRCA has an FDR-adjusted p-value of 0.12 and is included in Fig. 7.6). The pathway analysis in Fig. 7.6C does not show the shared non-differentiated signal that we are suspecting, the most enriched pathway for both signatures relates to their respective organ (kidney and eye development). However, the pathway results for signatures X82 and X107 contain many developmental-related processes that are expected


Figure 7.6: **(A)** The heatmap shows the Signature Hazard Score (SHS) which is defined as log10(FDR-corrected p-value) from the multivariate CoxPH model (correcting for covariate subtype) between 'up' and 'down' sample groups as defined by the bimodal distribution test. Negative values in blue represent a low SHS for which the 'up' group has a significantly better chance of survival than the 'down'-group (low hazard if signature is detected). Positive values in red mark cases in which the 'up'-group has a worse chance of survival (high hazard if signature is detected). All results with an FDR-adjusted p-value $> 0.25$ are grey. Signatures in white are ones that were not used for survival analysis because the deconvolution results did not pass the bimodality test. For each tumor type, the p-value for survival separation by cancer subtype is also annotated. **(B)** Data set origin for meta-clusters in A. **(C)** The most enriched GO pathway for each meta-cluster in A.

for the data sets they are coming from (fetal kidney and stem-cell-derived cells), but are not expected in an adult tissue sample, except for tumor cells that dedifferentiated and became more stem cell-like.

Fig. 7.6 shows an overview of all tumor type – signature pairs that have a subtype-corrected signature outcome separation (SOS) with an FDR-adjusted p-value $\leq 0.25$. Signature X22 shows a SOS in four tumor types, PRAD, LAML, KIRC, and SKCM, but in SKCM the detection of the signature correlates with better outcome (low Signature Hazard Score = SHS), whereas in the other tumor types the detection of the signature correlates with worse outcome (high SHS). The cells in signature X22 stem from neonatal testis tissues and are annotated as sertoli cells. The pathway activity analysis in Fig. 7.6C lists the 'Peripheral Nervous System Development' pathway as the most enriched in this signature. Other pathways that are positively enriched in the signature are related to the reproductive process.

Another signature that occurs in multiple tumor types is X13. Similar to X22, SKCM samples show a better outcome with high X13 estimates, PRAD and LAML show worse outcomes. Signature X13 represents all cells in a data set of precursors of human CD4+ cytotoxic T lymphocytes (E-GEOD-106540). The pathway activities for X13 show enrichment for immune response-related pathways, especially T cell activation and proliferation. T cells in the tumor microenvironment have been shown to influence patient outcome before, and cytotoxic CD4+ T cells are generally shown to prevent tumorigenesis [70]. The data set X13 stems from specifically analyzes T cell precursors, which might indicate why the signature correlates with worse patient outcomes for PRAD and LAML. For SKCM, the detection of cell-type signatures seems to generally correlate with better patient outcomes. This might be due to tumor purity, as it was

shown before that survival outcomes in SKCM are correlated to the purity of the samples [9].

### 7.2.5 The Tumor Microenvironment Map defines a high-risk pan-cancer patient group

In order to take a more comprehensive look at the cell type signature estimates in TCGA tumors, instead of analyzing each signature in each tumor type, we projected the TCGA samples onto a two-dimensional landscape, using the estimates of all 117 cell types as input to Tumor Map [54]. The interactive Tumor Microenvironment (TME) map is available online (bit.ly/TMEmap).

Fig. 7.7 shows snapshots of the TME map. The full TME map in Fig. 7.7A shows that most samples cluster by their tumor type. This is expected because the cell-of-origin signal in cancer molecular data is strong and the deconvolution estimates are based on mRNA-Seq data [36]. The major sample projection in the TME map is mostly recapitulating the results described in Fig. 7.4D, where similar tumor types and tumors of related organs are clustering with each other.

The samples on the TME map were clustered using hdbscan, a spatial hierarchical clustering method [47] (Fig. 7.8A). If a tumor type has at least ten samples in multiple clusters, an outcome analysis like described in section 7.2.4 was performed between the main cluster of that tumor type and the smaller minor cluster. During this analysis, a pattern occurred for multiple tumor types. Cluster c29 primarily contains BRCA samples (compare Fig. 7.8A to Fig. 7.7A), but also minor populations of other tumor types. For five tumor types, the minor sample group is at high risk of disease progression compared to the main group of samples.

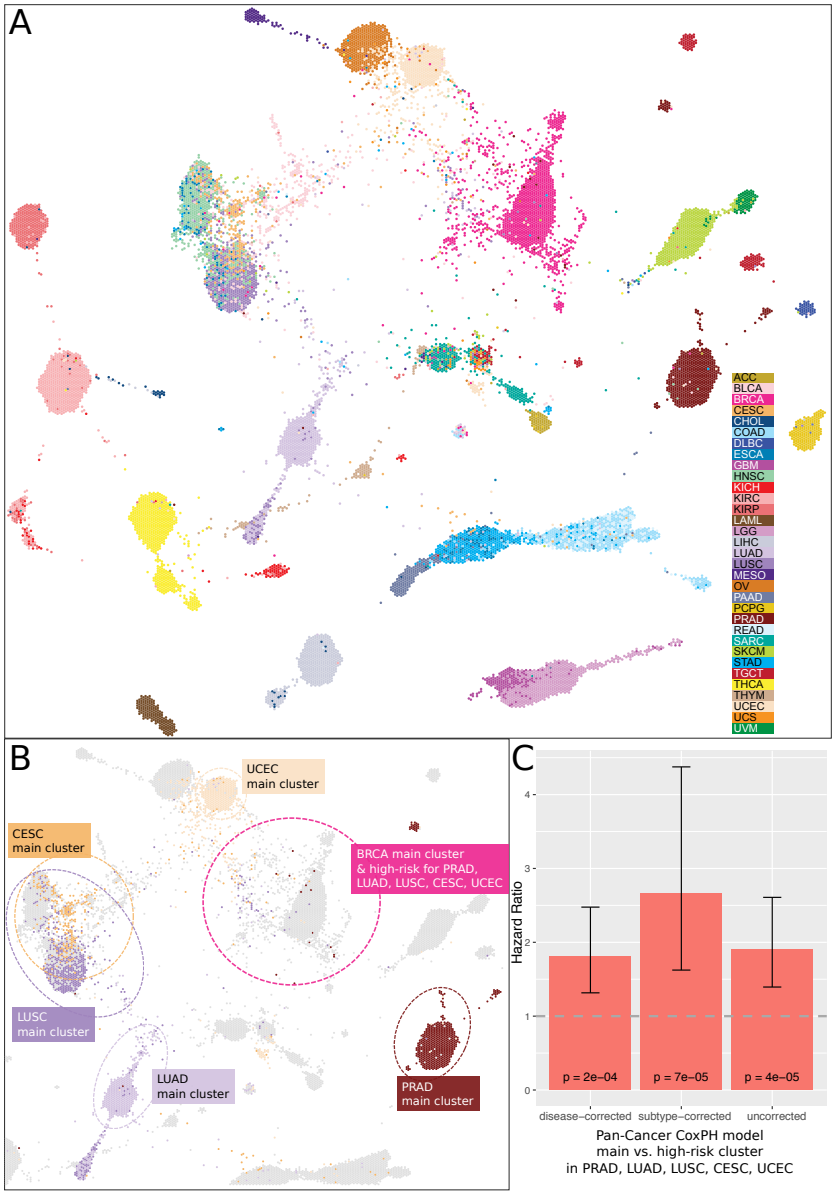


Figure 7.7: **Tumor Microenvironment (TME) Map (bit.ly/TMEmap). (A)** Full TME map colored by TCGA disease. **(B)** Part of the TME map in A, colored are only those tumor types that show a survival separation between the tumor types' main cluster (circled) and the tumor type samples that cluster with the BRCA samples on the TME map (high-risk cluster c29 circled in pink). **(C)** CoxPH model hazard ratio (HR) contrasting samples in the main cluster for each tumor type (c08 for PRAD, c16 for LUAD, c15 for LUSC and CESC, c30 for UCEC, circled in B) with samples in cluster c29 (pink circle in B). HR is reported for an uncorrected model, corrected for covariate disease, and corrected for covariate subtype. Vertical bars indicate the 95% confidence interval around the HR. Horizontal dashed line marks HR=1, with HRs above the line indicating a higher hazard for samples in c29. See Figure 7.8 for clustering solution and KM plot per tumor type.

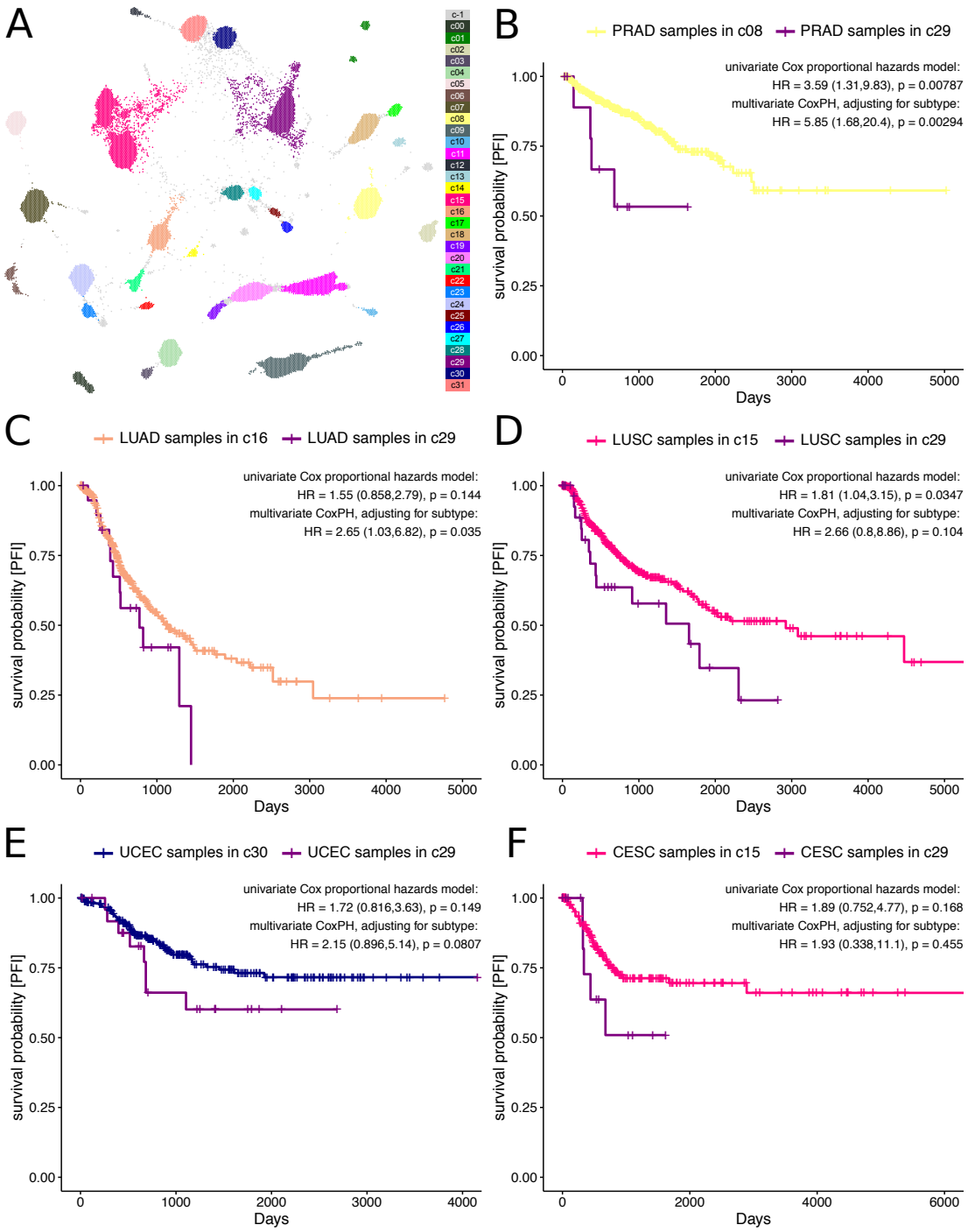


Figure 7.8: **Definition and survival analysis of pan-cancer high-risk sample group. (A)** Full Tumor Microenvironment (TME) Map colored by spatial clustering (hdbscan, minimal cluster size = 50). **(B-F)** Kaplan-Meier survival plot for 5 different tumor types contrasting samples in the main cluster for each tumor type vs. samples in high-risk cluster c29 (circles in Fig. 7.7B).

For two tumor types, this result is significant in the subtype-corrected multivariate survival analysis at a level of $p \leq 0.05$ (p=0.003 for PRAD, p=0.035 for LUAD). For three other tumor types, the separation does not meet this significance threshold (p=0.10 for LUSC, p=0.081 for UCSC, p=0.50 for CESC), but the direction of the separation is the same in all five tumor types (Fig. 7.8B-F).

I also performed a pan-cancer outcome analysis for these five tumor types. I combined the samples in all main clusters for the five tumor types and contrast them to the samples in high-risk cluster c29. The survival separation is highly significant, i.e. a multivariate CoxPH model that corrects for disease has a log-rank p-value of 2e-4 and a hazard ratio of 1.8, and a multivariate CoxPH model that corrects for tumor subtype, therefore indirectly also correcting for tumor type, has a log-rank p-value of 7e-5 and a hazard ratio of 2.7 (Fig. 7.7C).

When looking for cell-type signatures that have different deconvolution estimates in c29 compared to the main tumor type clusters, there are many different signatures that show a consistent difference. I took a closer look into the most differential signatures that had a common pattern across the five tumor types, i.e. consistently have either high or low estimates in c29 compared to the main tumor type cluster. Signature X21 shows the biggest difference between clusters, with c29 having a higher deconvolution estimate. It is derived from Spermatocytes and Spermatogonia cells from adult testis samples. Consistent with that, the most upregulated pathway in the X21 signature is 'male meiosis'. Two signatures from normal pancreas samples are also estimated in higher levels in c29: X60 from pancreatic A, B, and D cells, and X56 from acinar cells. Additionally, three signatures from stem cell and differentiation related data sets are estimated in higher levels in c29: X14 from human pluripotent stem cells differentiat-

ing towards definite endoderm, X88 from induced pluripotent stem cells (wibj_2) which is also described above in relation to outcome in PRAD, and X76 from H9 embryonic stem cells in organoid culture.

There are also signatures that are estimated on lower levels in c29: X118 from alveolar epithelial cells from lung carcinoma samples, X146 from extra thymic aire-expressing dendritic cells, and X81 from epithelial cells from esophagus samples. The signatures just described show the strongest difference between c29 and the main tumor type clusters, but are not the only ones that are different. There is not an obvious common theme between these cell types.

Additionally to the cell-type signatures, I also look at the TCGA gene expression data directly. The clustering of high-risk samples in PRAD, LUAD, LUSC, UCEC, and CESC is not recapitulated when looking at the overall gene expression data. However, with a supervised differential gene expression analysis between c29 and the main tumor type clusters, there is a set of genes that emerges with a similar pattern of differential expression for all five tumor types. A differential pathway activity analysis between c29 and the main tumor type clusters reveals that activity in c29 is enriched in chromosome organization, histone modification, and cell cycle pathways. The most up-regulated pathway in c29, averaged over five tumor types, is 'GO Negative Regulation of Chromosome Organization', which indicates a shared signal of increased chromosome instability.

### 7.2.6 Discussion

We presented scBeacon, a scRNA-Seq processing pipeline that rapidly clusters and integrates data sets to create a cell type signature matrix. We validated the deconvolution using

scBeacon-derived cell-type signatures in silico mixtures from single-cell and sorted bulk RNA-Seq data. We used the EBI single-cell atlas as a database to create a comprehensive set of cell-type signatures with the scBeacon method. We used the resulting 117 signatures for the deconvolution of 33 different cancer types from TCGA. Many of the cell-type signatures are found to be correlated to patient outcomes in single tumor types, some also over multiple tumor types.

In a pan-cancer, pan-cell-type approach, I used the Tumor Map tool to create a tumor microenvironment (TME) map. The Tumor Map considers the estimates for all 117 cell-type signatures simultaneously to calculate tumor sample similarities and the two-dimensional projection. A consistent pattern of a high-risk sample cluster emerges from the TME map in five different tumor types. Due to the small sample numbers in this high-risk cluster, the outcome separation is only significant for two out of five tumor types. Nevertheless, all five tumor types show an elevated hazard ratio for samples in the high-risk cluster and this consistent signal supports the validity of this finding. When the five tumor types are combined, the risk of disease progression for samples in the high-risk cluster is significantly elevated. Supervised differential gene expression analysis found genes that show consistent expression changes between the tumor types' main clusters and the high-risk cluster, with pathway activity analysis indicating differences in chromosome instability and epigenetic changes. However, the sample grouping was only found through the presented deconvolution analysis with our cell-type signatures and is not reflected in unsupervised analysis of the TCGA cancer samples.

The annotation of the established collection of cell-type signatures is difficult. Where possible, we used annotations given by the authors for each data set. But, many data sets

have no or incomplete cell-type annotations for their cells. There have been attempts to automate cell-type annotation for single-cell sequencing data, but the success varies greatly with the availability and quality of marker genes. Additionally, many cell types are defined by cell surface markers, and these markers might not transfer well into scRNA-Seq data. They also limit the annotation to known cell types, which negates one of the advantages of single-cell analysis.

The interpretation of the deconvolution results also has challenges. When a cell type is detected in a cancer sample, it can be due to the cell type being present in the tumor microenvironment. However, another possibility is that the tumor cells themselves have acquired certain characteristics of other cell types, which is then interpreted as the expression signature of that cell type by the deconvolution method. Yet another possibility is that the usage of an incomplete reference might influence the deconvolution estimate to detect a related cell type when the actual cell type is not included in the signature matrix.

We have also thought about the granularity of our cell type reference. Some data sets in our reference database are represented completely by just one cell type signature. This happens because all cells in the data set are from a specific cell type and are very similar to each other when compared to the complete set of data sets. Nevertheless, a more fine-grained cell type definition might be desirable in some cases, and a hierarchical definition of cell types and cell-type signatures might be a solution to this issue.

# Chapter 8

# Conclusion

Even though cancer is a disease originating in the genome, the cell phenotype is often better classified with transcriptional signatures. In chapter 2, I showed that, except for a few cases that mostly rely on mutational information, most drug sensitivity predictors work best when using gene expression data. Prior knowledge about biological pathways and gene sets further increases the signal in the transcriptional data. With the LURE method described in chapter 3, I showed that different cancer driver events can have related cell phenotypes and that they can be linked using transcriptional signatures. Of the 10,271 genomic driver events that we were able to define a transcriptional signal for, up to 1,847 were newly implicated by our analysis.

The meta-analysis of molecular subtype classifiers in chapter 4 showed that a balanced sample set and well distinguishable subtypes are the most important indicators for good classification performance. The TMP study overall came to the conclusion that gene expression was sufficient for many cancer types to reach an acceptable subtyping performance compared

to using all five data platforms available to the final classifiers.

In chapter 5, I described the characterization of a very aggressive subtype of prostate cancer emerging under treatment pressure. Again, this subtype can reliably be described by a gene expression signature, and the combination of transcriptional signature and pathological classification defines the most at-risk patients. However, in the case of metastatic prostate cancer, as well as in lung cancer, described in section 5.6, the strict categorization into exactly one subtype is not applicable to all samples. Instead of distinct groups, a continuum between subtypes is observed in both cancer types. In chapter 6, I formally described this continuum that exists between subtypes of advanced prostate cancer.

In the final chapter 7, I presented a tumor immune infiltration map derived with established immune cell type deconvolution methods from gene expression data. Taking this analysis a step further, I described the deconvolution of cancer samples into a more comprehensive set of cell types derived from single-cell RNA sequencing data. This dissection of the tumor and the tumor microenvironment revealed a high-risk patient group that is not detected by traditional differential expression analysis.

Altogether, I presented a variety of research projects that demonstrate the use of transcriptional signatures in the analysis of tumors and their microenvironment. I described the new insights gained from gene expression analysis about genetic cancer driver events, cancer subtypes, and ultimately patient risk and outcomes.

# Bibliography


[1] Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.

[2] Rahul Aggarwal, Jiaoti Huang, Joshi J. Alumkal, Li Zhang, Felix Y. Feng, George V. Thomas, Alana S. Weinstein, Verena Friedl, Can Zhang, Owen N. Witte, Paul Lloyd, Martin Gleave, Christopher P. Evans, Jack Youngren, Tomasz M. Beer, Matthew Rettig, Christopher K. Wong, Lawrence True, Adam Foye, Denise Playdle, Charles J. Ryan, Primo Lara, Kim N. Chi, Vlado Uzunangelov, Artem Sokolov, Yulia Newton, Himisha Beltran, Francesca Demichelis, Mark A. Rubin, Joshua M. Stuart, and Eric J. Small. Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study. *Journal of Clinical Oncology*, 36(24):2492, 2018.

[3] Rahul Aggarwal, Gustavo Rubio Romero, Verena Friedl, Alana Weinstein, Adam Foye, Jiaoti Huang, Felix Feng, Joshua M Stuart, and Eric J Small. Clinical and genomic characterization of Low PSA Secretors: a unique subset of metastatic castration resistant prostate cancer. *Prostate cancer and prostatic diseases*, pages 1–7, 2020.

[4] Rahul Aggarwal, Tian Zhang, Eric J Small, and Andrew J Armstrong. Neuroendocrine prostate cancer: subtypes, biology, and clinical outcomes. *Journal of the National Comprehensive Cancer Network*, 12(5):719–726, 2014.

[5] Nishant Agrawal, Rehan Akbani, B Arman Aksoy, Adrian Ally, Harindra Arachchi, Sylvia L Asa, J Todd Auman, Miruna Balasundaram, Saianand Balu, Stephen B Baylin, et al. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell*, 159(3):676, 2014.

[6] Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. A comparison of AUC estimators in small-sample studies. In *Machine learning in systems biology*, pages 3–13. PMLR, 2009.

[7] Joshi J. Alumkal, Duanchen Sun, Eric Lu, Tomasz M. Beer, George V. Thomas, Emile Latour, Rahul Aggarwal, Jeremy Cetnar, Charles J. Ryan, Shaadi Tabatabaei, Shawna Bailey, Claire B. Turina, David A. Quigley, Xiangnan Guan, Adam Foye, Jack F. Youngren, Joshua Urrutia, Jiaoti Huang, Alana S. Weinstein, Verena Friedl, Matthew Rettig,

Robert E. Reiter, Daniel E. Spratt, Martin Gleave, Christopher P. Evans, Joshua M. Stuart, Yiyi Chen, Felix Y. Feng, Eric J. Small, Owen N. Witte, and Zheng Xia. Transcriptional profiling identifies an androgen receptor activity-low, stemness program associated with enzalutamide resistance. *Proceedings of the National Academy of Sciences*, 117(22):12315–12323, 2020.

[8] Mariano J Alvarez, Yao Shen, Federico M Giorgi, Alexander Lachmann, B Belinda Ding, B Hilda Ye, and Andrea Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature genetics*, 48(8):838, 2016.

[9] Dvir Aran, Marina Sirota, and Atul J. Butte. Systematic pan-cancer analysis of tumour purity. *Nature communications*, 6(1):1–12, 2015.

[10] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.

[11] Nikolas G Balanis, Katherine M Sheu, Favour N Esedebe, Saahil J Patel, Bryan A Smith, Jung Wook Park, Salwan Alhani, Brigitte N Gomperts, Jiaoti Huang, Owen N Witte, et al. Pan-cancer convergence to a small-cell neuroendocrine phenotype that shares susceptibilities with hematological malignancies. *Cancer Cell*, 36(1):17–34, 2019.

[12] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[13] Himisha Beltran, Davide Prandi, Juan Miguel Mosquera, Matteo Benelli, Loredana Puca, Joanna Cyrta, Clarisse Marotz, Eugenia Giannopoulou, Balabhadrapatruni VSK Chakravarthi, Sooryanarayana Varambally, et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nature medicine*, 22(3):298–305, 2016.

[14] Himisha Beltran, David S Rickman, Kyung Park, Sung Suk Chae, Andrea Sboner, Theresa Y MacDonald, Yuwei Wang, Karen L Sheikh, Stéphane Terry, Scott T Tagawa, et al. Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. *Cancer discovery*, 1(6):487–495, 2011.

[15] Heike Bittersohl and Werner Steimer. Chapter 9 - intracellular concentrations of immunosuppressants. In *Personalized Immunosuppression in Transplantation*, pages 199 – 226. Elsevier, San Diego, 2016.

[16] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[17] Peder R Braadland, Håkon Ramberg, Helene Hartvedt Grytli, Alfonso Urbanucci, Heidi Kristin Nielsen, Ingrid Jenny Guldvik, Andreas Engedal, Kirsi Ketola, Wanzhong

Wang, Aud Svindland, et al. The β2-adrenergic receptor is a molecular switch for neuroendocrine transdifferentiation of prostate cancer cells. *Molecular Cancer Research*, 17(11):2154–2168, 2019.

[18] Cancer Genome Atlas Research Network. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine*, 372(26):2481, 2015.

[19] Jian Carrot-Zhang, Xiaotong Yao, Siddhartha Devarakonda, Aditya Deshpande, Jeffrey S. Damrauer, Tiago Chedraoui Silva, Christopher K. Wong, Hyo Young Choi, Ina Felau, Gordon A. Robertson, et al. Whole-genome characterization of lung adenocarcinomas lacking the RTK/RAS/RAF pathway. *Cell reports*, 34(5):108707, 2021.

[20] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, 44(8):e71–e71, 2016.

[21] Eric Collisson, Joshua Campbell, Angela Brooks, Alice Berger, William Lee, Juliann Chmielecki, David Beer, Leslie Cope, Chad Creighton, Ludmila Danilova, et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*, 511(7511):543–550, 2014.

[22] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

[23] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

[24] Hongxu Ding, Andrew Blair, Ying Yang, and Joshua M Stuart. Biological process activity transformation of single cell gene expression for cross-species alignment. *Nature communications*, 10(1):1–6, 2019.

[25] Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology*, 38(6):737–746, 2020.

[26] Alexander Drilon, Theodore W Laetsch, Shivaani Kummar, Steven G DuBois, Ulrik N Lassen, George D Demetri, Michael Nathenson, Robert C Doebele, Anna F Farago, Alberto S Pappo, et al. Efficacy of larotrectinib in TRK fusion–positive cancers in adults and children. *New England Journal of Medicine*, 378(8):731–739, 2018.

[27] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.

[28] 10X Genomics. http://go.10xgenomics.com/revolution Revolutionizing Gene Expression with 10x.

[29] Mary Goldman, Brian Craft, Jingchun Zhu, Teresa Swatloski, Melissa Cline, and David Haussler. The UCSC Xena system for integrating and visualizing functional genomics, 2016.

[30] Kiley Graim, Verena Friedl, Kathleen E. Houlahan, and Joshua M. Stuart. PLATYPUS: A Multiple-View Learning Predictive Framework for Cancer Drug Sensitivity Prediction. *Pacific Symposium on Biocomputing*, 24, 2019.

[31] Sudheer Gupta, Kumardeep Chaudhary, Rahul Kumar, Ankur Gautam, Jagpreet Singh Nanda, Sandeep Kumar Dhanda, Samir Kumar Brahmachari, and Gajendra PS Raghava. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. *Scientific reports*, 6:23857, 2016.

[32] David Haan, Ruikang Tao, Verena Friedl, Ioannis N. Anastopoulos, Christopher K. Wong, Alana S. Weinstein, and Joshua M. Stuart. Using Transcriptional Signatures to Find Cancer Drivers with LURE. *Pacific Symposium on Biocomputing*, pages 343–354, 2020.

[33] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature methods*, 13(6):521, 2016.

[34] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[35] Ying C Henderson, Yunyun Chen, Mitchell J Frederick, Stephen Y Lai, and Gary L Clayman. MEK inhibitor PD0325901 significantly reduces the growth of papillary thyroid carcinoma cells in vitro and in vivo. *Molecular cancer therapeutics*, pages 1535–7163, 2010.

[36] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

[37] Kristen L Hoek, Parimal Samir, Leigh M Howard, Xinnan Niu, Nripesh Prasad, Allison Galassie, Qi Liu, Tara M Allos, Kyle A Floyd, Yan Guo, et al. A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. *PloS one*, 10(2):e0118528, 2015.

[38] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[39] In Sock Jang, Rodrigo Dienstmann, Adam A Margolin, and Justin Guinney. Stepwise group sparse regression (SGSR): gene-set-based pharmacogenomic predictive models with stepwise selection of functional priors. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 32–43. World Scientific, 2014.

[40] Dmitri Kazmin, Tatiana Prytkova, C Edgar Cook, Russell Wolfinger, Tzu-Ming Chu, David Beratan, JD Norris, Ching-yi Chang, and Donald P McDonnell. Linking ligand-induced alterations in androgen receptor structure to differential gene expression: a first step in the rational design of selective androgen receptor modulators. *Molecular endocrinology*, 20(6):1201–1217, 2006.

[41] Daniel Koboldt, Robert Fulton, Michael McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua McMichael, Lucinda Fulton, David Dooling, Li Ding, Elaine Mardis, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[42] Panagiotis A. Konstantinopoulos, Dimitrios Spentzos, Beth Y. Karlan, Toshiyasu Taniguchi, Elena Fountzilas, Nancy Francoeur, Douglas A. Levine, and Stephen A. Cannistra. Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *Journal of Clinical Oncology*, 28(22):3555, 2010.

[43] Daniel H. Kwon, Li Zhang, David A. Quigley, Adam Foye, William S. Chen, Christopher K. Wong, Felix Y. Feng, Adina Bailey, Jiaoti Huang, Joshua M. Stuart, Verena Friedl, Alana S. Weinstein, Tomasz M. Beer, Joshi J. Alumkal, Matthew Rettig, Martin Gleave, Primo N. Lara, George V. Thomas, Patricia Li, Austin Lui, Eric J. Small, and Rahul R. Aggarwal. Down-regulation of ADRB2 expression is associated with small cell neuroendocrine prostate cancer and adverse clinical outcomes in castration-resistant prostate cancer. *Urologic Oncology: Seminars and Original Investigations*, 38(12):931.e9–931.e16, 2020.

[44] Dung T Le, Jennifer N Uram, Hao Wang, Bjarne R Bartlett, Holly Kemberling, Aleksandra D Eyring, Andrew D Skora, Brandon S Luber, Nilofer S Azad, Dan Laheru, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine*, 372(26):2509–2520, 2015.

[45] C Chang-Yew Leow, S Gerondakis, and A Spencer. MEK inhibitors as a chemotherapeutic intervention in multiple myeloma. *Blood cancer journal*, 3(3):e105, 2013.

[46] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

[47] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), March 2017.

[48] Prateek Mendiratta, Elahe Mostaghel, Justin Guinney, Alok K Tewari, Alessandro Porrello, William T Barry, Peter S Nelson, and Phillip G Febbo. Genomic strategy for targeting therapy in castration-resistant prostate cancer. *Journal of clinical oncology*, 27(12):2022–2029, 2009.

[49] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D. Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLOS One*, 5(11):e13984, 2010.

[50] Yoshito Nakanishi, Hideaki Mizuno, Hitoshi Sase, Toshihiko Fujii, Kiyoaki Sakata, Nukinori Akiyama, Yuko Aoki, Masahiro Aoki, and Nobuya Ishii. ERK signal suppression and sensitivity to CH5183284/Debio 1347, a selective FGFR inhibitor. *Molecular cancer therapeutics*, 14(12):2831–2839, 2015.

[51] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[52] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519, 2012.

[53] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453, 2015.

[54] Yulia Newton, Adam M Novak, Teresa Swatloski, Duncan C McColl, Sahil Chopra, Kiley Graim, Alana S Weinstein, Robert Baertsch, Sofie R Salama, Kyle Ellrott, Manu Chopra, Theodore C Goldstein, David Haussler, Olena Morozova, and Joshua M Stuart. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. *Cancer research*, 77(21):e111–e114, Nov 2017.

[55] Irene Papatheodorou, Pablo Moreno, Jonathan Manning, Alfonso Muñoz-Pomer Fuentes, Nancy George, Silvie Fexova, Nuno A Fonseca, Anja Füllgrabe, Matthew Green, Ni Huang, et al. Expression Atlas update: from tissues to single cells. *Nucleic acids research*, 48(D1):D77–D83, 2020.

[56] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

[57] Jeff S Pawlikowski, Tony McBryan, John van Tuyn, Mark E Drotar, Rachael N Hewitt, Andrea B Maier, Ayala King, Karen Blyth, Hong Wu, and Peter D Adams. Wnt signaling potentiates nevogenesis. *Proceedings of the National Academy of Sciences*, 110(40):16009–16014, 2013.

[58] Amit Persad, Geetha Venkateswaran, Li Hao, Maria E. Garcia, Jenny Yoon, Jaskiran Sidhu, and Sujata Persad. Active beta-catenin is regulated by the PTEN/PI3 kinase pathway: a role for protein phosphatase PP2a. *Genes & Cancer*, 7(11-12):368, 2016.

[59] Dan Robinson, Eliezer M Van Allen, Yi-Mi Wu, Nikolaus Schultz, Robert J Lonigro, Juan-Miguel Mosquera, Bruce Montgomery, Mary-Ellen Taplin, Colin C Pritchard, Gerhardt Attard, et al. Integrative clinical genomics of advanced prostate cancer. *Cell*, 161(5):1215–1228, 2015.

[60] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[61] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, 2018.

[62] José A Seoane, Ian NM Day, Tom R Gaunt, and Colin Campbell. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, 30(6):838–845, 2013.

[63] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, November 2003.

[64] Swapnil Rohidas Shinde and Subbareddy Maddika. PTEN modulates EGFR late endocytic trafficking and degradation by dephosphorylating Rab7. *Nature Communications*, 7:10689, 2016.

[65] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: a Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[66] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 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(43):15545–15550, 2005.

[67] Shin Takahashi, Takuya Moriya, Takanori Ishida, Hiroyuki Shibata, Hironobu Sasano, Noriaki Ohuchi, and Chikashi Ishioka. Prediction of breast cancer prognosis by gene expression profile of TP53 status. *Cancer Science*, 99(2):324, 2008.

[68] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, May 2009.

[69] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.

[70] Rong En Tay, Emma K Richardson, and Han Chong Toh. Revisiting the role of CD4+ T cells in cancer immunotherapy - new insights into old paradigms. *Cancer Gene Therapy*, pages 1–13, 2020.

[71] Turki Turki and Zhi Wei. A link prediction approach to cancer drug sensitivity prediction. *BMC systems biology*, 11(5):94, 2017.

[72] David Venet, Jacques E Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240, 2011.

[73] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), 2017.

[74] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546, 2013.

[75] Brian S White, Andrew J Gentles, Aurélien de Reyniès, Aaron M Newman, Andrew Lamb, Laura Heiser, Joshua J Waterfall, Thomas Yu, and Justin Guinney. A tumor deconvolution DREAM Challenge: Inferring immune infiltration from bulk gene expression data, 2019.

[76] F. Alexander Wolf, Fiona Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Gottgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv*, 2018.

[77] Shigeo Yamaguchi, Shin Takahashi, Kaoru Mogushi, Yuki Izumi, Yumi Nozaki, Tadashi Nomizu, Yoichiro Kakugawa, Takanori Ishida, Noriaki Ohuchi, Chikashi Ishioka, and Shunsuke Kato. Molecular and clinical features of the TP53 signature gene expression profile in early-stage breast cancer. *Oncotarget*, 9(18):14193, 2018.

[78] Yan Zhang, Dayong Zheng, Ting Zhou, Haiping Song, Mohit Hulsurkar, Ning Su, Ying Liu, Zheng Wang, Long Shao, Michael Ittmann, et al. Androgen deprivation promotes neuroendocrine differentiation and angiogenesis through CREB-EZH2-TSP1 pathway in prostate cancers. *Nature communications*, 9(1):1–17, 2018.

[79] Yuanyuan Zhang and Zemin Zhang. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular & molecular immunology*, 17(8):807–821, 2020.