



**Interpreting single cell transcriptomes in the pathway  
space and its applications in cancer**

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE  
DEGREE OF

**DOCTOR OF PHILOSOPHY**

BY

**SMRITI CHAWLA**  
**PhD17203**

Department of Computational Biology  
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

**September 2022**

## THESIS CERTIFICATE

This is to certify that the thesis titled Interpreting single cell transcriptomes in the pathway space and its applications in cancer is submitted by Smriti Chawla to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



**Dr. Debarka Sengupta**  
INDRAPRASTHA INSTITUTE OF  
INFORMATION TECHNOLOGY  
DELHI  
IIIT Delhi, 110020



**Dr. Vibhor Kumar**  
INDRAPRASTHA INSTITUTE OF  
INFORMATION  
TECHNOLOGY DELHI  
IIIT Delhi, 110020

Place: New Delhi

Date: September 2022

## **ACKNOWLEDGEMENTS**

I want to take this opportunity to express my heartfelt gratitude to my supervisor Dr. Debarka Sengupta for guiding me throughout my study, providing timely advice, and unwavering support. He encourages me to stay focused, excited, and creative about the research. He has been incredibly inspiring, understanding, and a guide in the true sense. There are many aspects of his personality that I had like to incorporate into my own, particularly his outlook on life. Every discussion with him motivates me to learn and grow more and contribute to the betterment of society. I want to commend him for maintaining a comfortable and friendly work environment leading to fruitful collaborations. I will always be indebted to him for all his support and patience.

I feel a deep sense of gratitude towards my co-supervisor, Dr. Vibhor Kumar. Without his invaluable guidance and support, this work would not have been possible. I appreciate all his mentoring throughout this work. His constant motivation and professional advice were indispensable throughout this journey of the doctoral thesis. His zeal for science kept me engaged in my research.

I sincerely thank Dr. Naveen Ramalingam, Dr. Stefanie Jeffrey, and Professor Colleen Nelson for research collaboration and co-mentoring me, and providing me with novel and valuable data to apply computational approaches to aid in societal benefits. I want to thank Anja Rockstroh, Melanie Lehman, Ellca Ratther, Bianca Troncarelli, and Ning Ma for their research collaboration and their valuable time, biological insights, and theoretical expertise.

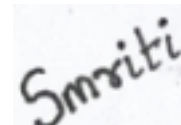
I am also grateful to all the faculty members of Department of Computational Biology for their unwavering constructive support and learnings they have shared in the form of courses they taught. I am also thankful to Ms. Priti Patel and other administrative staff members for all the administrative support they have lent whenever needed. I want to thank the IT department, especially Mr. Adarsh Agarwal, for being extremely helpful in solving our technical queries.

A special thanks to all my colleagues, collaborators, and friends without whom this

journey would not have been possible. In particular, I had like to thank my friends and colleagues, Chitrita, Omkar, Namrata, Vivek, Sumeet Patiyal, Shruti, Samriddhi, Krishan Gupta, Anjali lathwal, Chakit Arora, Dilraj Kaur, Shiju S for providing friendly and healthy work environment. I am delighted to appreciate the role of some of the great friends, Jyoti Maggu, Priyadarshini, Neetesh, Sarita, Shreya, Raghav, Madhu and Indra Prakash Jha. They stood as pillars of support during this journey. I am also thankful to Atishay Jain, Anuneet Anand, and Apoorva Gupta for helping me. My appreciation also extends to some special friends, Kavita Kochar and Bhumika Mehta, for being a great source of strength and support.

No words can describe the support those closest to me have provided. I am incredibly thankful to my parents, Mr. Deepak Chawla and Mrs. Vibha Chawla, for always having my back. I want to extend my thanks to my brother Dheeraj Chawla, sister-in-law Anshika Chawla and my niece Myra Chawla for encouraging me in every phase of life. Lastly, I would like to express my deep gratitude towards my extended family for their constant support and blessings.

Some of the icons used in figures are adapted from <https://www.flaticon.com/>.

A small, square logo with the word "Smriti" written in a stylized, handwritten font. The logo is tilted slightly to the right.



## ABSTRACT

Single-cell transcriptomics is a powerful technique that has revolutionized our approach to dissect cellular phenotypes and diversity in complex tissues at an unprecedented resolution. The emergence of this groundbreaking technology has dramatically enhanced our understanding of cellular heterogeneity, interactions, and cell fate decisions during the development and progression of cancer. These new technologies have shown to be promising in the field of cancer genomics. Despite all the goodness, many computational challenges remain.

Human cells express about 20,000 genes, which dynamically carry out a multitude of biophysical activities. Statistical and machine learning-based methods treat genes as independent variables in the process of characterizing intra-tumoral heterogeneity and developing insights into cancer progression, pathogenesis, and clinical outcomes. This approach is quite limiting since constantly accumulating somatic genomic alterations are often manifested through the dysregulation of molecular pathways or cancer-relevant gene signatures. Thus, exploiting gene set and pathway scores to decipher heterogeneity in the single-cell will aid in many applications in cancer genomics.

We propose a statistically robust method called UniPath to represent single cells in terms of pathway or gene set enrichment scores. UniPath projects gene expression readouts and single-cell ATAC-seq profiles into pathway scores while accounting for dropouts and sequencing depth. Further, it allows pseudotemporal ordering of single cells in pathway space. Visualization of gradients and distribution of pathways on a pseudotemporally ordered tree helps understand the lineage potency of cells. Another application of UniPath is that it helps enumerate differences in two cell populations through the exploitation of pathway co-occurrences. In a connected work, we introduce, Precily, deep learning framework that leverages pathway scores of gene expression profiles and drug descriptors for anti-cancer drug response predictions. We thoroughly validated our proposed approach using bulk and single-cell gene expression profiles. We also assessed the performance of our approach on several in-house generated prostate cancer datasets. Finally, we interrogated the transcriptomic profile of triple-negative

breast cancer tumor and Natural killer cell doublets and their physical distance captured at single-cell resolution. We discovered that physical distances are governed by activities of regulatory modules, pinpointing the presence of transcriptional memory. In addition, our investigation into ligand-protein pairs interactions that are responsible for conveying messages into cells by activating signaling pathways revealed inflated activities of some of the specific pairs in NK-immune cell doublets. We concluded that intercellular communications in tumors play an essential role in deciphering the underlying mechanism operating in cancer. Our approach of capturing and profiling single-cell doublets will aid in the understanding of complex tumor microenvironment and cellular interactions.

# List of publications

## Publications

1. **Chawla, Smriti**, Sudhagar Samydurai, Say Li Kong, Zhengwei Wu, Zhenxun Wang, Wai Leong Tam, Debarka Sengupta, and Vibhor Kumar. "UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles." *Nucleic acids research* 49, no. 3 (2021): e13-e13.

2. Goswami, Chitrita, **Smriti Chawla**, Deepshi Thakral, Himanshu Pant, Pramod Verma, Prabhat Singh Malik, Ritu Gupta, Gaurav Ahuja, and Debarka Sengupta. "Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of NSCLC." *BMC genomics* 21, no. 1 (2020): 1-12.

## Preprints

3. **Chawla, Smriti**, Anja Rockstroh, Melanie Lehman, Ellca Rather, Atishay Jain, Anuneet Anand, Apoorva Gupta et al. "Gene expression based inference of drug resistance in cancer." *bioRxiv* (2021).

4. Flores, Bianca, **Smriti Chawla**, Ning Ma, Chad Sanada, Praveen Kujur, Ludmilla Chinen, Kyle Hukari et al. "Microfluidic system-based time-course tracking of physical proximity between cells and its effect on gene expression for elucidating live single cancer-immune cell interactions." (2021).

5. Poonia, Sarita, Anurag Goel, **Smriti Chawla**, Namrata Bhattacharya, Priyadarshini Rai, Yi Fang Lee, Yoon Sim Yap et al. "Marker-free characterization of single live circulating tumor cell full-length transcriptomes." *bioRxiv* (2021).

# CHAPTER 1

## INTRODUCTION

Every year cancer causes several million deaths worldwide. The number of new incidences is increasing at staggering rates due to a lack of early detection methods and ineffective treatments for patients with advanced and metastatic cancer [128]. Cancer is a highly complex dynamic and heterogeneous disease evincing phenotypic and genetic diversity [57]. This cellular heterogeneity steered by anomalies at the genomics, transcriptomics, and proteomics level is characterized by mutations, aberrant gene expression, and transcriptional stochasticity. Cellular heterogeneity poses significant challenges to effective cancer treatments and is an essential factor contributing to therapeutic resistance and disease recurrence, eventually governing clinical outcomes. Cancer not only comprises discrete unambiguous pathologies but there is considerable heterogeneity among different cells within each tumor [167].

### 1.0.1 Hallmarks of Cancer

The enormous array of cancer genotypes is the embodiment of six key physiological changes in cell that collectively governs malignant growth and metastatic dissemination (Figure 1.1) [79][80].

#### **Self sufficiency in growth signals**

Plausibly the most fundamental characteristic of cancer cells is their capability to sustain persistent proliferation. The synthesis and release of growth-promoting signals are tightly regulated in normal tissues. This ensures proper balance of cell number and thus proper maintenance of normal tissue structure and function. On the other hand cancer cells gain control over their own fate by disrupting these growth signals [80].

## **Evading growth suppressors**

In addition to hallmark capability of sustaining proliferative signaling, cancer cells can evade growth suppression. Dozens of tumor suppressor genes function in diverse ways to limit cell proliferation and growth. Internal or external stimuli trigger the activation of these tumor suppressor genes, resulting in apoptosis, process of programmed cell death or cell cycle arrest. Therefore, cancer cells must circumvent the activation or expression of tumor suppressor genes [75] [80].

## **Resisting cell death**

Over the last two decades, functional studies have established the role of apoptosis as a natural obstruction to the development of cancer. Apoptosis is an autonomous process in which many genes are activated, expressed, and regulated, resulting in programmed cell death to clear abnormal cells for keeping a stable internal environment. Internal or external stimuli can trigger apoptosis. However, tumor cells can evade or attenuate apoptosis and become resistant to therapy [37] [75] [80].

## **Limitless replicative potential**

The three acquired hallmark characteristics of cancer cells—self-sufficiency in growth signals, insensitivity to antigrowth signals, and evading apoptosis—all lead to dysregulated cell proliferation programs [75] [80]. The resulting dysregulated cell proliferation is sufficient to impart cancer cells with limitless replicative potential that leads to the generation of macroscopic tumors. This property is in sharp contrast to the behavior of normal cells that undergoes a limited number of rounds of cell growth and cell divisions [80] [75].

## **Inducing Angiogenesis**

Angiogenesis is a process that involves new blood vessels formation from the existing ones and is responsible for supplying oxygen and nutrients to body tissues. This process is critical for tumor growth and metastasis. Tumor expansion and sustenance are dependent on nutrients and oxygen provided by new blood vessels formed by tumors. Under

normal physiological conditions, angiogenesis is tightly regulated and regulated by balancing various endogenous pro and anti-angiogenic factors. However, tumors activate an event known as the angiogenic switch to progress by disrupting the balance between pro and anti-angiogenic factors more toward pro-angiogenic outcome and ultimately resulting in the malignant phenotype of the dormant lesion [89][6].

## Tissue invasion and metastasis

The sequence of events resulting in the malignant transformation of cells is quite complex. Malignant cells possess a number of key distinguishing hallmarks, namely potential for uncontrollable cell growth and capabilities to spread into, invade nearby tissues and metastasize [103]. About 90% of human cancer fatalities are caused by the settlements of tumor cells in distant organs or tissues. The propensity of cancer cells to invade, migrate, and metastasize allows them to abscond from primary tumor and colonize new territories in the body having sufficient nutrients and space initially. The invasion and metastasis is a multi step process involving a sequence of distinct steps, often referred to as the invasion-metastasis cascade. The cascade envisages a series of cell-biologic alterations, starting with local invasion, then tumor cell intravasation, movement of cancer cells via the lymphatic system, followed by extravasation, micrometastases, and finally colonization [80].

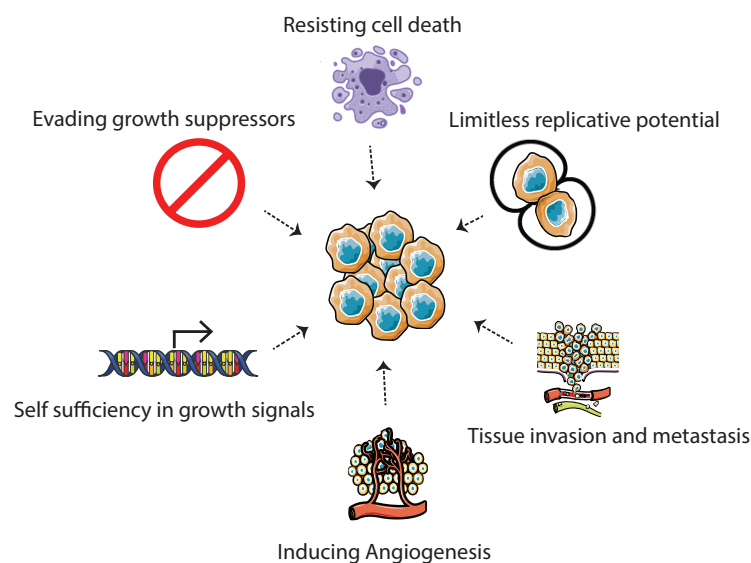


Figure 1.1: Hallmarks of cancer

## 1.0.2 Complex tumor ecosystem

A single cell is a basic unit of life activity, holding a blueprint for biological functions. Approximately 37.2 trillion cells reside in the human body, exhibiting intricate interplay between genetic mechanisms and the cellular environment, thus directing the genesis and functionalities of complex tissues and organs [145] [164]. In cancer, a single cell can collapse an entire organism. Aberrant genetic alterations in a single cell can drive the formation of a malignant tumor mass with distinct lineage and sub populations accompanied by intratumor heterogeneity. Genetic diversity in single tumors has been apparent for a long time, but with the advancement in high throughput sequencing methods, the full magnitude of intratumor heterogeneity is becoming noticeable [140]. Clonal diversity exhibited by cancer cells offers them selective advantages. Cancer being complex and dynamic in nature can be viewed as equivalent to a tumor ecosystem in which tumor cells and host cells cooperate and communicate with each other in the tumor microenvironment and can even adapt and evolve in diverse conditions, resulting in the invasion, metastasis, host system hijacking, and therapy resistance. In-depth discernment of cellular composition, cross talks and interactions, and dynamic behavior within the tumor ecosystem's tumor microenvironment is necessary to comprehend cancer biology and evolution [164].

## 1.0.3 Clonal evolution and diversity

Cancer evolution is a dynamic process that governs the emanation of cancer cell sub populations by Darwinian selection, giving rise to clonal diversity. The clonal evolution model is depicted in Figure 1.2. As a tumor grows, their evolutionary history is reflected by the catalog of somatic mutations accumulated over time, which confers survival advantages, thereby determining the clonal population's overall fittest [142] [180]. The tissue ecosystem serves as a habitat for the evolution of tumor clones. Their complex networks and framework evolved over a billion years is to augment and assimilate multicellular functions while confining renegade clonal expansion. Tissues within the tumor microenvironment serve as a context for the development and evolution of cancer. Limited resources and other microenvironmental constrictions result in the natural selection of tumors. The interplay of driver lesions, passenger lesions, deleterious lesions and tumor microenvironment drives the clonal evolution. Thus, understanding

clonal expansion, diversification and selection is critical in understanding the progression of cancer and therapeutic interventions [72].

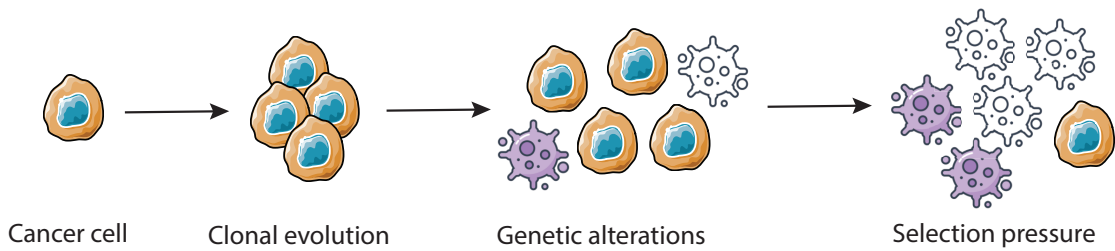


Figure 1.2: Clonal evolution model

### 1.0.4 Intra-tumor heterogeneity

In today's era, understanding and clinical evaluation of tumor heterogeneity are of utmost importance for improving clinical oncology [183]. In particular, intratumor heterogeneity (ITH) is regarded to be one of the most important predictors of therapeutic resistance and treatment failure. ITH is linked with cancer progression, recurrences, and poor survival outcomes in cancer patients with metastatic disease [218]. ITH is attributed to the coexistence of subclonal populations of tumor cells exhibiting remarkable variability in their genetic, phenotypic, or behavioral traits within the primary tumors [138]. ITH is a dynamic process that is detected at multiple levels and often follows the Darwinian type approach. The tumor progression is driven by unpredictable and frequently chaotic cellular processes triggered by oncogenic changes and environmental factors and these processes hold the key to understanding tumor development. Given the intricate and ever-changing nature of the tumor architecture, it is critical to comprehend that molecular alterations themselves evolve within the tumor during disease progression and metastasis. Both genetic and non-genetic subclonal alterations are responsible for enduing cancer with adequate phenotypic plasticity to adapt to microenvironmental forces and overcome the hurdles offered by anti-tumoral therapy. Thus, ITH poses significant challenges to personalized medicine since it can restrict treatment efficacy and contribute to drug resistance [95] [218].



### 1.0.5 Tumor microenvironment and its role in tumorigenesis

The tumor microenvironment within a solid tumor is represented by malignant and non-cancerous cells. These noncancerous cells include endothelial cells, fibroblasts, stromal and immune cells. The extracellular matrix is also a vital component of the tumor microenvironment. Apart from manifested heterogeneity in tumor subclones, heterogeneity among noncancerous cells in the microenvironment further adds a layer of complexity and is important in tumor growth, dissemination, and therapeutic responses [164] (Figure 1.3). This unique tumor microenvironment emerges as the tumor progresses due to its intercommunication with the host. It is created and influenced by the tumor, which governs key molecular events occurring in neighboring tissues. Furthermore, the tumor microenvironment directs anomalous tissue function and plays a vital role in developing aggressive and metastatic cancer. The tumor microenvironment presents a hurdle in the functioning of immune cells by inhibiting their anti-tumor activities. Other cells such as Cancer-associated fibroblast (CAFs), inflammatory cells, adipose cells, and neuroendocrine cells in the tumor microenvironment also impact immune cells. CAFs influence cancer progression through extracellular matrix modification, and induction of angiogenesis. They also directly impact cell proliferation via the secretion of growth factors. While the presence of immune-inflammatory cells in chronic inflammation sites is connected with varied tissue pathologies that include abnormal angiogenesis and neoplasia. Neuroendocrine and adipose cells are the accomplices of tumor formation. Tumors adopt different tactics and mechanisms to escape immune surveillance and drug therapies, which vary in various cancer types. Each tumor exhibits a unique signature that defines its microenvironment. Thus, understanding tumor microenvironment at the molecular and cellular level is paramount [202] [203] [207].

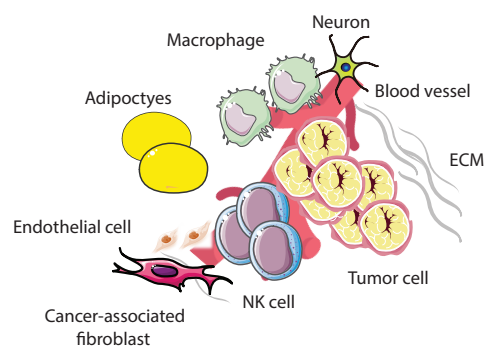


Figure 1.3: Complex tumor microenvironment comprising heterogenous population of tumor cells and variety of other non-malignant cells

## 1.0.6 Natural killer cells and cancer

The tumor microenvironment exhibits a high degree of complexity, and immune escape is now recognized as an essential cancer hallmark. Immune escape plays a significant role tumor development and metastatic dissemination. Natural killer (NK) cells are the principal effector cells in innate immunity and exhibit diverse states in the microenvironment. Most current tumor microenvironment therapeutic strategies rely on T cell immunity, either by stimulating activatory signals or repressing inhibitory signals. However, the limited success of immunotherapies involving T cells pinpoints the significance of developing novel immunotherapies, for instance, utilizing previously overlooked NK cells. NK cells are an essential aspect of tumor immunosurveillance, as indicated by greater cancer susceptibility and metastasis in mouse models and clinical studies with attenuated NK cell activity [133][212]. NK cells possess potent cytotoxic activity coordinated by a complex network of multiple inhibitory and activating signals [141] (Figure 1.4). NK cells inhibit tumor growth either by direct interactions with cancer cells or modulating the functionalities of other immune cells in the tumor microenvironment. They have the ability to distinguish aberrant populations of cells from normal ones, resulting in more targeted anti-tumor cytotoxic effects and fewer off-target complications. Given their critical role in cancer biology, NK cells have emerged as a probable target for cancer management [212].

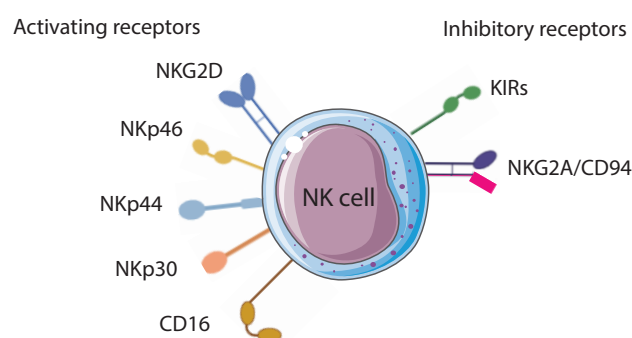


Figure 1.4: Figure depicts major activatory and regulatory receptors on the NK cell surface

## 1.0.7 Signaling pathways

Cancer is caused by a series of genetic and epigenetic changes that allow cells to bypass homeostatic control mechanisms that generally subdue inappropriate cell proliferation

and impede the survivability of aberrantly proliferating cells outside their usual niches. Recent advances in next-generation sequencing have paved the way for multi-omics analysis that has significantly refined our understanding of how an intricate network of signaling pathways within cells can be exploited to develop more targeted therapies. The progression of cancer is linked to dynamic and complex interactions of tumor cells with neighboring non-neoplastic cells and the extracellular matrix. The acquired hallmark capabilities of tumor cells are primarily attributed to the dysregulation of signal transduction pathways [178]. A wide variety of molecular hubs and signaling pathway nodes have been linked with cancer development, and many of these, such as receptor tyrosine kinase and downstream signaling pathways, are targets of drugs approved by various regulatory bodies [223]. PI3K signaling is one of the major signaling pathways downstream of receptor tyrosine kinase, which is often mutated or amplified in most solid cancers. Many of the PI3K inhibitors are currently under clinical trials [225]. Oncogenic mutations can result in the production of mutated proteins with dysregulated activities. Examples of such proteins involved in a multitude of signaling pathways are small GTPases (Ras), cytoplasmic tyrosine kinases (Src and Abl), nuclear receptors (estrogen receptor). Components of other signaling pathways for example Wnt, Hedgehog, and Notch can also be impacted. Further, hyperactivation of oncogenic pathways like PI3K-Akt and Ras-ERK can result in uncontrolled cell proliferation in cancer. These pathways are also involved in regulating cell death in multiple ways. Also, cancer cell metabolism is regulated by components of intracellular signaling pathways that are disrupted by mutations in tumor suppressor genes and oncogenes. Cancer cell metabolism is characterized by increased uptake of glucose and glycolysis, resulting in advancing cancer progression [94][178]. The PI3K-Akt pathway is incorporated in targeting a myriad of substrates to promote metabolic changes in tumors. Another signaling pathway Wnt is an important regulator of development and stemness, which has also been linked to cancer. Wnt signaling promotes epithelial to mesenchymal transition. The intricate network of signaling pathways in cancer poses significant challenges in the development of therapeutics [226].

### 1.0.8 Cancer management

Cancer remains one of the deadliest malignant diseases that jeopardizes human life as it is the most difficult disease to treat [91]. Cancer is a unique disease, and it can not be considered as one disease by physicians since each patient exhibits a specific disease. First, physicians identify the existence of cancerous tissue in patients. Then the next question is how cancer can be treated under such circumstances with optimal and safest drugs. The duration of treatment is also crucial, and patient progress must be tracked by clinicians to plan future therapeutic strategies. The initial stage of treatment involves categorizing the tumor as indolent or not, whether it is aggressive or has the potential to metastasize or not, and identifying tumor grade. The therapy must be planned while accounting for factors such as tumor location, metastatic region, tumor stage, or cell types present in heterogeneous tissue samples. The treatment success must strive for control of disease, prevention of the reoccurrence of metastasis, and improving overall survival and life quality of patients [222]. Some of the possible cancer treatment options are radiotherapy, surgery, immunotherapy, chemotherapy, targeted therapy, and Personalized therapy as shown in Figure 1.5. Currently, the most effective cancer treatment is surgery which involves the removal of tissues with cancer cells. Chemotherapy makes use of standard anti-cancer drugs to kill cancer cells. Meanwhile, radiotherapy involves using high-dose X-rays and gamma rays to treat a tumor at a post-surgery tumor location. These rays are extremely effective in killing tumor cells that may linger after surgery or recur where the tumor was excised. In recent years, immunotherapy has gained a lot of attention and is primarily defined as the use of components of the immune system in cancer treatment. Although chemotherapy has remained a mainstay for cancer management in many tumor types but suffers from limitations of limited response rate and side effects. Translational research has transformed the way we develop new cancer treatments. One of the most significant advancements in modern oncology is the shift from an organ-centric strategy to a patient-tailored approach guided by deep molecular analysis [5] [51] [87] [159] [179].

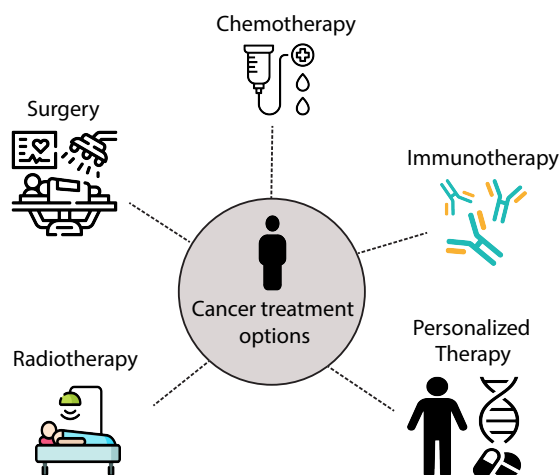


Figure 1.5: Different cancer treatment options

## 1.1 Overview of RNA Sequencing

Massively parallel next-generation sequencing development has revolutionized transcriptomics by allowing RNA analysis via complementary DNA (cDNA) sequencing. This method is known as RNA sequencing (RNA-seq). Over the past few decades, RNA sequencing has become an invaluable tool in refining our understanding of complicated and dynamic attributes of the transcriptome and genomic functions [112] [184]. Transcript identification and gene expression quantification have been indispensable components of molecular biology since the revelation of RNA's function as a pivotal intermedial between the proteome and genome [43]. The standard workflow begins with RNA extraction, mRNA enrichment, cDNA synthesis, and adaptor-ligated sequencing library preparation. The prepared library is sequenced using a high-throughput sequencing platform. Then ultimately, through computational approaches, sequencing reads are aligned or assembled to the transcriptome, and reads are quantified. Bulk RNA-seq experiments typically estimate the total gene expression level from a heterogeneous population of cells. The transformation and evolution of RNA-seq have been propelled forward by technological developments in wet-lab and computational domains. It has provided a clearer and unbiased view of RNA biology and transcriptomes, and bulk RNA-seq is extensively used to refine our understanding of cancer biology. Though bulk RNA-seq holds great potential in developing novel and efficient clinical applica-

tions, true signals driving biological processes can be blurred by mean gene expression from bulk RNA-seq profiles. This biological problem impelled the birth of new scRNA-seq technology [127][184].

## 1.2 Overview of Single-cell sequencing

Traditional methods like flow cytometry and fluorescence in situ hybridization (FISH) for measuring single cells were developed years ago and are conventionally used in laboratories. However, these approaches have limited applicability as the number of genes and proteins profiled is significantly less, restricting the amount of information drawn from single-cell samples. The emergence of groundbreaking technologies in single-cell isolation and genomics, transcriptomics, and proteomics profiling has imparted scalability to single-cell analysis in recent years. One of the critical obstructions in the single-cell investigation is amplifying a limited quantity of input nucleic acid material to capture the desired threshold level. As of late, considerable technical breakthroughs in whole-genome amplification (WGA) or whole-transcriptome amplification (WTA) have been accomplished to attain an ample amount of input material for preparing next-generation sequencing (NGS) libraries.

Single-cell sequencing technologies for profiling RNA transcriptomes face challenges in quantifying various RNA species. Several methods have emanated to amplify small quantities of mRNA in a single cell. The first method used in the first single-cell microarray and mRNA-seq study utilized the Poly-A tailing method. This method involves adding anchoring sequences to the 3' end of synthesized cDNA by terminal transferase. However, this method suffers from a strong 3' bias due to cDNA synthesis's inefficiency by reverse transcriptase. Another one-step protocol utilizing multiplexed RT-PCR for target amplification is that Sequence-specific amplification (SSA) is restricted to investigating a limited number of genes. The template switching-based protocol method called Smart-seq is a commonly used technology for sequencing full-length mRNA in single cells by using Moloney Murine Leukemia Virus (MMLV) reverse transcriptase. Although Smart-seq has improved the coverage of reads across transcripts, its ability to profile lowly expressed mRNAs is limited. CEL-seq is another robust and efficient method that involves the addition of T7 promoters to cDNA and uses *in vitro* transcription for mRNA amplification. However, a common shortcoming of

these methods is inadequate throughput due to the independent handling of single-cell samples and accidental human fallacy. High throughput molecular barcoding of single cells in microdroplets has been exploited to overcome the limitations of the methods mentioned above. The recent introduction of droplet-based single-cell transcriptomics techniques has facilitated the parallel screening of thousands of single cells. Such strategies have dramatically improved the throughput of single-cell transcriptomics. The development of microfluidic devices has also complemented single-cell transcriptomics. Microfluidic devices help increase the sensitivity and throughput of single-cell analysis by automating the processing and examination of biological materials [145] [167].

Although single-cell RNA sequencing (scRNA-seq) methods are rapidly evolving, single-cell epigenome profiling remains the most technically challenging task. Bisulfite sequencing is the gold standard method for detecting DNA methylation of cytosine (5mc) residues genome-wide at single-base resolution. However, interrogation of DNA methylation at single-cell resolution was not feasible until recently due to the degradation of a significant amount of DNA caused by bisulfite treatment. The first single-cell method, reduced representation bisulfite sequencing (scRBBS), was developed to quantify cytosine methylation modifications at CpG islands across the genome. This is a powerful technique as it allows the analysis of many promoters at a low cost. However, it has limited coverage.

Chromatin immunoprecipitation sequencing (Chip-seq) is another powerful tool to identify genome-wide histone marks, which play a crucial role in influencing transcriptional states. Due to background noise, performing Chip-seq at a single cell resolution is quite perplexing. Recently, a droplet-based approach has been used to overcome this limitation. In this technique, a pool of single cells that had already been subjected to micrococcal nuclease digestion and barcoding is used for immunoprecipitation on chromatin. Hi-C-based techniques have recently been proposed to capture chromosome interactions and conformations, but they are limited in their resolution [41].

Assay for transposase-accessible chromatin using sequencing (ATAC-seq) is used to map open chromatin in single cells. It uses the Tn5 transpose enzyme to simultaneously fragment DNA and tag open chromatin regions with adapter sequences. This process is referred to as tagmentation. The DNA fragments are PCR amplified and sequenced. At single-cell resolution, combinatorial indexing and microfluidics-based approaches have

been used to assess chromatin in thousands of single cells. Several single-cell platforms such as C1 and Chromium systems allow scATAC-seq [41] [104]. These technologies are cutting-edge tools in refining our overall understanding of cancer biology.

### **1.3 A brief overview of computational methods for analyzing scRNA-seq data**

With the technological revolution in single-cell technology, unprecedented amounts of high throughput single cell data are getting generated. Computational tools have emerged to handle such large datasets and have become a fundamental part of single cell analysis. The first step in single cell analysis is preprocessing, which ensures quality control and normalization while controlling for confounders.

High-dimensional single cell data often poses challenges in visualization. However, several dimensional reduction based visualization approaches is available like PCA, t-SNE, UMAP and diffusion maps. These methods help draw biological insights and investigate relationships among different cell types in low-dimensional space [167]. Unsupervised clustering in scRNA-seq is vital to identify previously unknown cell subpopulations. Apart from classical clustering methods, other clustering approaches have emerged to handle large datasets, such as local sensitivity hashing (LSH) based drop-Clust [181], the graph-based clustering strategy used by Seurat [173]. Differential gene expression analysis (DE) can be used to discriminate between different cell populations. In recent years several tools to perform DE analysis have been developed for analyzing scRNA-seq data. For instance, MAST, which uses the hurdle model and another tool SCDE (single-cell differential expression), uses a Bayesian approach to account for dropouts in single cell data while modeling gene expression [167],[61],[106]. Although clustering techniques can divulge the intrinsic group structure within data, they are insufficient to reveal cellular heterogeneity. Trajectory inference methods emerged to study dynamical biological processes including differentiation, cell cycle and cell state transitions.



## 1.4 Applications of single cell analysis in cancer genomics

Recent technical advances in single-cell technologies have transformed our overall understanding of biology and opened up new research avenues. This powerful technology has dramatically enriched our knowledge of cancer progression in terms of invasion, metastasis and therapeutic responses. One can envision numerous novel unexplored applications of this technology in the field of cancer management. Some of the applications discussed are in Figure 1.1.

### 1.4.1 Pseudotemporal analysis and RNA velocity

scRNA-seq can capture a high-resolution view of gene expression patterns in a heterogeneous cell population. Thus, scRNA-seq provides a more accurate way of scrutinizing dynamic and complex processes like cell cycle, cellular differentiation, and activation. One instrumental approach to exploiting scRNA-seq data to gain valuable biological insights is to order cells along a hypothetical time trajectory computationally. A computational approach called trajectory inference or pseudotemporal ordering analysis models such dynamic processes, which order cells based on their passage through the process or reflect the gradual transition of their transcriptomes. Over the last few years, a plethora of pseudotemporal ordering methods have been developed, and even new ones are emerging every month [99][168]. A monocle is an unsupervised approach for ordering single cells in pseudotime. To resolve complex biological problems, Monocle uses a machine learning approach such as Reversed Graph Embedding to learn the principal graph from single-cell datasets [195]. Another method, TSCAN that employs a cluster-based minimum spanning tree (MST) procedure to order single cells for studying dynamic changes in single gene expression profiles along the pseudotime [99]. SCUBA is another approach that utilizes bifurcation analysis to uncover lineage relationships from single-cell transcriptomes. Another approach RNA velocity incorporates mRNA dynamics to predict the future cell states on a timescale of hours to aid in the analysis of developmental processes and cellular dynamics [136]. Such methods could help study cancer development and cancer mechanisms, such as identifying cellular state transitions in cancer. RNA velocity and trajectory inference methods hold considerable potential in detecting underlying mechanisms of altered cell development

processes in cancer pathogenesis [58].

### **1.4.2 Characterization of Intratumor heterogeneity (ITH)**

Cancer is a highly heterogeneous disease displaying a high degree of phenotypic diversity impelled by molecular anomalies at various levels such as genetic, epigenetic, and transcriptomic in cells that communicates within lucid spatially assembled microenvironments. Solid tumors are composed of neoplastic cells and mesenchymal cells. Furthermore, in a single lesion, multiple subclones are present within the tumor cells, complicating tumor samples' analysis. Such heterogeneity poses significant challenges to the current standard of care by promoting metastasis and resistance to therapy, thus ultimately influencing clinical outcomes. Therefore precise delineation of heterogeneity is critical for characterizing underlying mechanisms of carcinogenesis, developing novel and effective treatment strategies, and drug development. Single-cell technology offers a powerful tool to differentiate intratumor heterogeneity (ITH) and provide a precise measure of genomic diversity in solid tumors. Interrogation of complex clonal genotypes is possible due to single-cell sequencing. It has enabled the detection of genetic diversity among different cancer types. In patient-derived xenograft models of lung adenocarcinoma, scRNA-seq identified drug-resistant tumor subclones. Further, scRNA-seq accentuated previously unappreciated intratumor heterogeneity in primary glioblastoma and deciphered that this heterogeneity is linked with potential prognostic implications [58] [128] [154]. A novel approach called Reference component analysis was developed, which helped in the unraveling of cellular heterogeneity in colorectal tumors through characterizing abnormal cell states within a tumor [125].

### **1.4.3 Distinguishing malignant cells from non-malignant cells**

Various computational strategies and methodologies have been developed to discriminate malignant cells from non-malignant cells. Malignant cells often display altered pathways and unique activated oncogenic programs illustrative of cancer. Their genetic makeup and transcriptional programs are quite pronounced from normal cells that they can be identified using clustering techniques. Different methods have been developed to identify cell clusters, but it is challenging to annotate them as malignant or non-

malignant. In some cancers, detecting specific marker genes or investigating gene-set enrichment can differentiate malignant and non-malignant cells. However, sparsity and prevalence of dropouts in scRNA-seq data may subject marker-based classification to false negatives. Aberrant upregulation of cancer-associated pathways and oncogenic signatures might help annotate neoplastic cell clusters. For instance, the scRNA-seq analysis of glioblastoma revealed the presence of neoplastic subpopulations displaying upregulated transcriptional programs associated with oncogenic signaling and proliferation. Another method to differentiate cancerous cells from non-cancerous cells involves the detection of large-scale copy number variations (CNVs) by exploring gene expression profiles of malignant cells compared to reference normal tissue [57].

#### **1.4.4 Inferring cell-cell communication with the tumor microenvironment**

Cell-to-cell communication mediated by ligand-receptors plays a vital role in the development and cancer progression. It allows cancer to reprogram the tumor microenvironment and cells at distant sites. The crosstalk between malignant cells and non-malignant cells in the intricate tumor microenvironment is crucial for the progression of tumor and dissemination, therapeutic resistance, immune infiltration, and inflammation [39] [113]. Given the relevance of receptor-ligand interactions on patient outcomes, therapeutic choices that target these interactions have become indispensable in the clinical management of cancer. For instance, ipilimumab (immune checkpoint inhibitor) blocks CD28 and CTLA4 interaction. These therapeutic agents have promising results in some tumor types. However, response rates are limited primarily due to the intricate network of cellular interactions operating in the tumor microenvironment, our comprehension of such networks is still limited. To uncover the interactions that could be targeted, it is necessary to comprehend the vast expanse of cellular interactions operating in tumor microenvironments and by what means such interactions influence patient outcomes [113]. scRNA-seq approaches allow characterizing many cell types within the tumor microenvironment. scRNA-seq methods to infer cell-cell communications generally rely on comparing gene expression levels of receptors in one cell type and their corresponding ligands in other cell types using existing catalogs of receptor-ligand pairs. CellPhoneDB is a novel repository of ligand-receptor interactions that computes the

mean expression of receptor and ligand genes in their respective cell types [53]. The statistical significance of the mean is assessed by comparing it to the null distribution. To assess statistical significance, a graph-based strategy for producing null distribution has been used as well [182]. The putative intercellular communications can also be assessed by estimating the correlation between the expression of receptor gene and ligand gene across single-cell datasets [57]. A computational method called NicheNet was developed that incorporates gene expression data with prior information on cell-cell signaling and gene regulatory networks to predict ligand-target links for interacting cells [25].

### 1.4.5 Therapy resistance and response

Since the past decade, technological advances in high-throughput biology are resulting in the generation of an increasing volume of biological data. Given the abundance of data, it is obvious to benefit from data-oriented recommendations in precision oncology. Precision oncology strives to provide tailored treatments based on the distinct characteristics of a patient's tumor. This aim is based on the notion that as the volume of data increases, better computational models are increasingly being utilized to predict drug response precisely by integrating data from multiple sources. This is important to aid clinicians in selecting the most effective treatment options available [215] [76]. This has been fueled chiefly in part by a paradigm shift in cancer classification from purely relying on histopathological characteristics of tumors to the interrogation of molecular features that indicate treatment responses [12]. Various machine learning algorithms have been created for improving drug sensitivity analysis.

Some of these include kernelized bayesian matrix factorization (KBMF) approach for predicting drug response through leveraging known pathway-response associations [7], matrix factorization for predicting drug response using cell line and drug structural similarity [201]. Another method CDRscan, a deep learning framework for drug response prediction that uses cancer genomic signatures, was proposed by Chang and colleagues [34]. Sakellaropoulos, Theodore, et al. reported that deep neural network-based drug response predictions perform better in comparison to ElasticNet and Random Forest. Liu, Chuanying, et al. suggested an ensemble learning approach that simultaneously incorporates a low-rank matrix completion and a ridge regression ma-

chine learning model for drug response prediction in cancer cell lines [131]. Initially bulk profiles have been used for training these models, but single-cell data-based techniques are beginning to show potential [1]. Though many cancers show preliminary responses to chemotherapy or targeted drugs, most tumors eventually develop therapy resistance. The evolution of therapy resistance in human cancers is often poorly understood. Key questions remain whether a drug-resistant subpopulation is already present in the tumor mass, i.e., adaptive resistance, or whether resistant phenotype evolves after therapy administration, i.e., acquired resistance.

Furthermore, epithelial to mesenchymal transition and cellular plasticity might confer drug resistance. The emergence of single-cell technologies has made it feasible to develop patient-tailored therapies. It has been used to predict drug sensitivity in multiple myeloma and optimize treatment strategies in renal cell carcinoma. Also, scRNA-seq identified several signaling pathways in lung adenocarcinoma cell lines associated with drug resistance [128] [145]. Integrating single-cell profiling with strategies that can quickly discover the effective and promising combinations of drugs will likely play a vital role in improving cancer care [1].

Apart from the applications discussed above, single-cell technologies also help interrogate and characterize circulating tumor cells (CTCs). The analysis of CTCs is a valuable tool to comprehend the biology of cancer metastasis, tracking disease progression, and clinical management of the condition. scRNA-seq of CTCs provides considerable details on their tumors of origin and is a powerful method allowing fair identification of CTCs. Recently, unCTC was developed to enable unbiased detection and characterization from single-cell gene expression profiles. unCTC provides many features, including standard and unique computational approaches and statistical modules for various analyses [158].

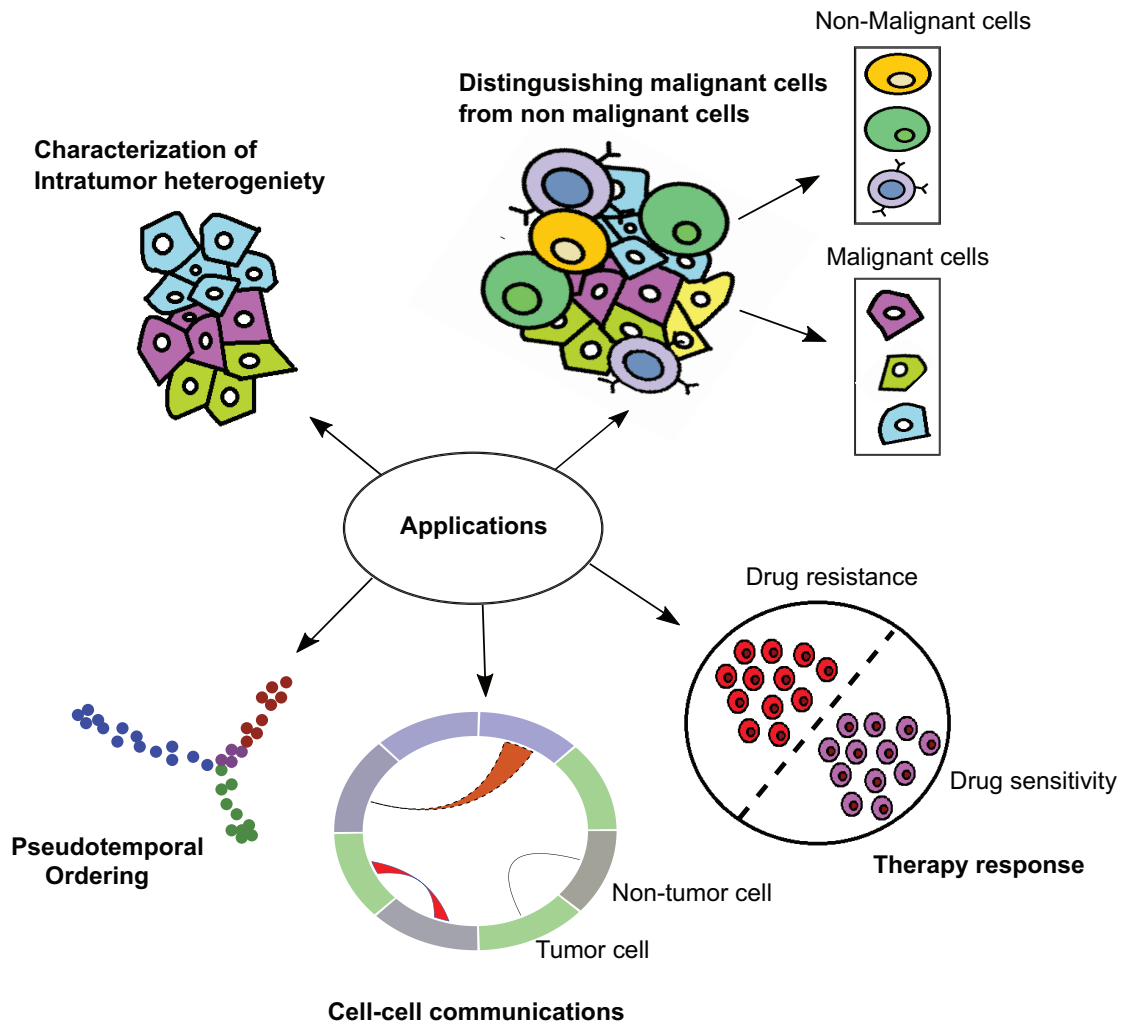


Figure 1.6: Applications of single cell sequencing in cancer

## 1.5 Scope of thesis

scRNA-seq technology has been widely employed to uncover intracellular heterogeneity at highly unprecedented resolution. scRNA-seq transcriptomic analysis has enabled researchers to unveil new and startling biological discoveries compared to typical bulk methods. Many computational tools were developed for cell clustering, lineage inference or pseudotemporal ordering, marker gene identification, cell type annotation, and visualization of the single-cell datasets. However, achieving greater biological and functional interpretability is often challenging.

Pathways are biological networks in which genes work in coordination but not independently to achieve specific cellular functions in discrete cell types. This is crucial in characterizing transcriptional diversity, disease subtype classification and drug discov-

ery and development. At the single cell level, pathway activation analysis, a powerful approach involves transforming gene expression data into meaningful gene sets to capture biologically relevant information and uncover the potential mechanism of cellular heterogeneity and irregularities in diseases [230] [231].

Inspired by the biological significance of pathways, we aim to develop a statistical and computational based algorithms to represent genes in meaningful gene sets and pathways at a single cell level and exploit pathway space for clinical applications. We have addressed three key issues in this context, outlined below.

### **1.5.1 Representation of single cells in terms of pathways**

Single-cell transcriptome and open chromatin data often suffer from high technical noise, dropouts, and sequencing depth issues. To decipher cellular heterogeneity from single-cell transcriptomes, we addressed the issue of depicting single cells in meaningful gene sets and pathways by utilizing their gene expression and open chromatin profiles while accounting for dropouts and sequencing depth. There are rarely any studies that have used single-cell open chromatin profiles for pathway analysis. Further, existing methods such as GSEA do not fully solve the objective of interrogating heterogeneity of pathway activity at the single-cell level. Then, other classes of methods such as SVA, RUV, scLVM, and f-scLVM do not provide gene set or pathway enrichment scores in every single cell. Although PAGODA provides pathway scores in every cell, however, it is computationally not fast and is not able to deal with single-cell profiles having a relatively less heterogeneous population of cells. On the other hand, AUCell is mainly used for cell-type identification but is not for clustering and pseudotemporal ordering of scRNA-seq data. The number of downstream analyses, including clustering, pseudotemporal ordering, and exploitation of pathway co-occurrences to differentiate two groups of cells using pathways estimated using our approach, have revealed novel biological insights that are difficult to achieve via gene expression profiles.

### **1.5.2 Drug response prediction using pathway activity scores**

Predicting tumor sensitivity to specific targeted therapies has been a challenge of utmost importance for personalized medicine. The majority of the current models to

predict drug responses use gene expression profiles, but genes' biological interpretability is limited. However, pathway-based strategies to predict drug response in cancer offer more useful biological insights as therapies work through the concerted action of genes within pathways. We demonstrated the utility of molecular drug descriptors and pathway activity scores for predictive drug response modeling. We assessed the efficiency of our model on single-cell and bulk RNA-seq profiles. A limited number of studies have leveraged subclones profiled at single-cell resolution for drug response prediction. We evaluated our model using several in-house generated prostate cancer (PCa) data including cell lines and xenografts exposed to different treatments. We further tested our approach on pan-cancer RNA-seq profiles from The Cancer Genome Atlas (TCGA) compendium. Our results revealed that pathway activity scores are indicative of drug resistance and sensitivity. Thus, integration of pathway activity scores with drug structure information to predict drug response will aid in developing patient-tailored therapies and clinical decision support systems.

### **1.5.3 Analysis of tumor-immune cell doublets at single cell level**

Studying intercellular communication and physical interactions between cells is essential in order to comprehend cancer initiation, progression, and immune responses. Although there have been significant refinements in high-throughput microscopy and single-cell technology, methods to quantify live cellular interactions are inadequate. Here, we interrogated the triple-negative breast cancer, and Natural killer cell doublets transcriptomes and physical distances captured utilizing a novel microfluidic integrated fluidic circuit (IFC) platform that revealed novel biological insights. This enabled the characterization of distinct molecular signatures operative in NK cells having the potential to kill tumor cells. Further, our results revealed that cell-cell interactions and physical distance between cells are governed by complex regulatory activities, pinpointing the existence of transcriptional memory as an essential governing strategy of cells. Additionally, we delineated increased correlation in some specific ligand-protein pairs in cancer-immune doublets. Cell-cell communications are manifested through specific ligand-protein pairs interactions that activate signaling pathways that might be involved in regulating cancer. Thus, this platform can help researchers investigate cell-cell interactions at single-cell doublet resolution to gain an understanding of tumor progression



and design and administer NK cell-based immunotherapies.

## CHAPTER 2

# Transformation of single cell transcriptomics and epigenomics data in pathway scores using UniPath and its evaluation

### 2.1 Introduction

Single-cell sequencing technologies in transcriptomics and epigenomics have emanated as a powerful tool to delineate complex and dynamical biological systems and unveil cellular heterogeneity at an unprecedented level. They have paved the way for numerous new opportunities and challenges. Despite the goodness, technical issues, i.e., dropouts and sequencing depth, are critical challenges in analyzing single-cell datasets. Therefore meaningful transformation of read counts is necessary for the comprehensive characterization of cellular heterogeneity. Studying single cells in terms of biological pathways has emerged as a powerful approach to uncover potential underlying cellular heterogeneity mechanisms and dissect complex diseases such as cancer.

Cellular heterogeneity and diversity among single cells are often used for defining a cellular composition of heterogeneous tissues, rare cell type detection, and interrogating the regulation of genes and transcription factors [100] [29]. However, novel questions and applications such as discovering co-occurrence among pathways, developmental potency, and lineage of cells and uncovering more specific targeted pathways for cancer therapy can be addressed and explored by representing single cells in meaningful and functional gene sets or pathway scores. There are tools for gene set enrichment analysis such as GSEA [186]. Still, their applicability is limited in deciphering heterogeneity of pathways at single-cell resolution since they utilize a differential gene expression-based approach between the group of cells.

There is another category of tools used for capturing cellular heterogeneity such as SVA [120], RUV [64], scLVM [29] and f-scLVM [30] furnishes relevance score for a group of single cells. While on the other hand, PAGODA [56] and AUCell [3] methods

provide relevance and enrichment scores in every single cell. PAGODA accounts for variability and high dropout rate in scRNA-seq data while computing scores for gene sets, but it is quite slow and is designed for handling a relatively more homogeneous population of cells. While AUCell is mainly used for identifying cells with one or two active gene signatures at a time. However, it isn't generally utilized for other scRNA-seq downstream analyses like clustering and pseudotemporal ordering. The main obstacle in modeling every single-cell gene expression profile into multiple pathways' enrichment scores is the default reliance on gene read count data. The vast proportion of zero in scRNA-seq data due to dropouts or true low gene expression and sequencing depth issues among single cells makes statistical modeling of single-cell open chromatin profiles and single-cell transcriptomics data challenging. Additionally, there has seldom been any attempt to utilize single-cell open chromatin profiles for transformation into pathways scores for clustering and pseudotemporal ordering. Therefore, there has been a requirement for a uniform approach to transforming single-cell RNA-seq and single-cell open chromatin profiles from homogeneous and heterogeneous cells or samples into pathway or gene set enrichment scores.

To this end, we developed UniPath to tackle the challenge of representing single-cell transcriptomes and open chromatin profiles in respect of pathways and gene set activity scores despite sequencing depth and variability in technical noise among cells. Instead of scaling or normalizing read counts using parametric distributions like negative binomial or Poisson distribution to avoid inadvertent artifacts, we use the common null model for adjusting pathways scores computing using scRNA-seq. The division of read counts by global accessibility scores highlighting the enhancers is utilized in single-cell open chromatin profiles (Figures 2.1). Comprehensive benchmarking of our methodology for estimating gene set enrichment scores and null models was performed using publicly available single-cell datasets. Further, we developed a method for pseudotemporal ordering of single cells in pathway space to avoid biases observed in the temporal order by using gene expression profiles directly. We have applied UniPath on several single-cell datasets to attain biologically relevant results that gene expression cannot achieve.

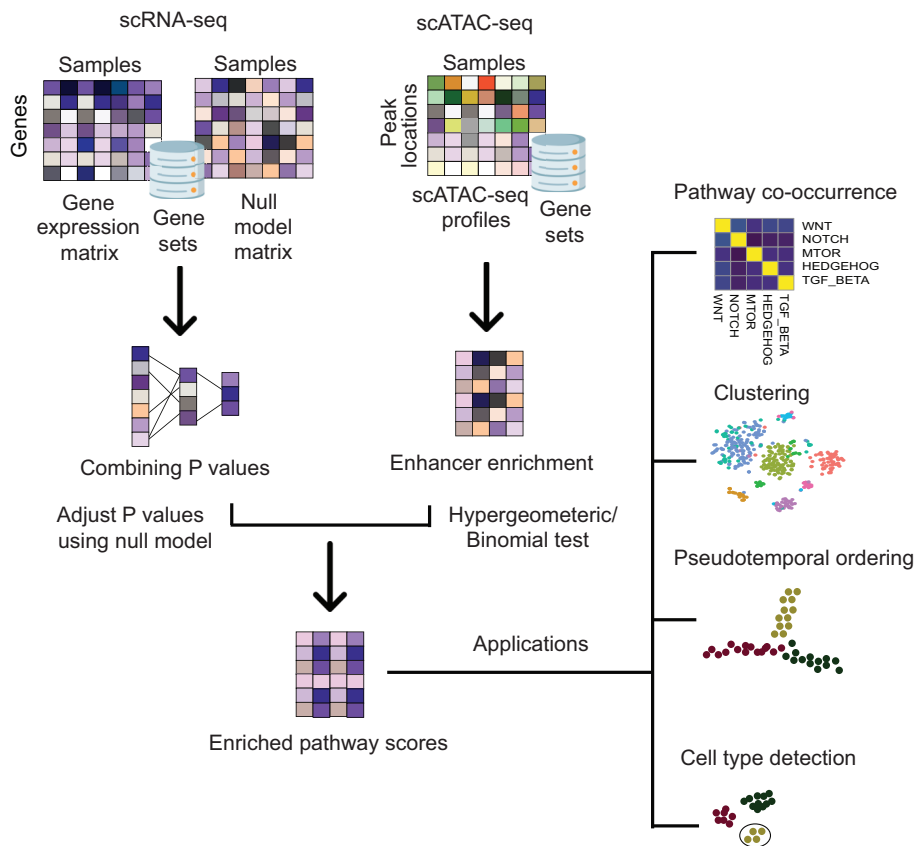


Figure 2.1: Overall schematics of UniPath workflow. For scRNA-seq, UniPath first transforms single-cell profiles into P-values that are combined utilizing Brown’s method for individual gene sets. Then, the null model is used to adjust the combined p-values. This yields the final adjusted pathway score matrix. For scATAC-seq profiles, UniPath converts scATAC-seq profiles to pathway or gene set enrichment scores by highlighting enhancers and using the Hypergeometric or Binomial test. For highlighting enhancers, global accessibility scores are used to normalize read count at a peak.

## 2.2 Methodology

### 2.2.1 Computing gene sets enrichment scores for single-cell open chromatin profiles

The majority of cell types harbor high open chromatin accessibility in their regulatory elements or sites such as insulators and active promoters. However, studying the sites having activities specific to particular cell type such as enhancers could help decipher differences in cells by leveraging single cell open chromatin profiles. Besides, enhancer profiles can give a clear view of pathways active in a cell. Thus, UniPath first highlights each cell’s enhancers by normalizing the scATAC-seq read count data by accessibility

scores. The accessibility scores are computed for the combined DNA-seq and ATAC-seq peak list sourced from ENCODE and IHEC consortiums [31]. The accessibility score is determined for a site as the fraction of cell types in which it is identified as an open chromatin peak. The tag count normalization is done as

$$t_{i,j} = \frac{p_{ij}}{a_i + \epsilon}$$

Here  $p_{ij}$  represents the tag count for peak  $i$  in a single cell  $j$ ,  $a_i$  denotes the global accessibility score for peak  $i$  and  $\epsilon$  is a pseudo count. Using this approach of global accessibility scores for highlighting enhancers doesn't require any tag count normalization between cells. Further, it enables the consistent conversion of the scATAC-seq profiles from multiple research groups without recomputing counts from the aligned DNA reads on a common peak list. Then, for each cell, peaks with high normalized read counts constitute a foreground set, and the background set consists of all the peaks. Generally, we have set a default cut off of 1.25 above global accessibility for selecting foreground peaks. However, the threshold can vary as per more stringent requirements. The set of foreground peaks is likely to represent regulatory sites and enhancers exhibiting cell type-specific activity. Then peaks having proximal genes within 1Mbp distance are retained. UniPath provides two statistical methods, Hypergeometric and Binomial tests, for computing gene set enrichment scores. For the Binomial test, we have used the formula:

$$\sum_{i=k_m}^n \frac{n!}{i!(n-i)!} p_m^i (1-p_m)^{n-i}$$

The test computes P-value or statistical significance for the gene set  $m$ . Here  $k_m$  denotes the number of occasions proximal genes out of  $n$  peaks of foreground set are from  $m$  gene set.  $p_m$  is the probability that genes belonging to gene set  $m$  appear to be proximal to peaks in the background set. In the case of the Hypergeometric test, statistical significance or P-value of gene set enrichment is calculated using the below equation:

$$\sum_{i=k_m}^{\min(n, K_m)} \frac{\binom{K_m}{i} \binom{N-K_m}{n-i}}{\binom{N}{n}}$$

Here  $K_m$  denotes the same as described above for the binomial test,  $m$  represents the number of times genes in the gene set  $m$  happens proximal to peaks in the background list and  $N$  represents the total number of peaks present in the background list.

### 2.2.2 Computing gene set enrichment scores for single-cell gene expression profiles without normalization

To estimate the importance of pathway enrichment in scRNA-seq profiles, we utilized logarithmic gene expression that could include FPKM, TPM, RPKM, and UMI counts values and treated each cell independently from the other. Instead of scaling and normalization across different cells, which might generate artifacts due to noise and gene dropouts, we use the generally recognized fact that non-zero gene expression values in a cell or sample follow an approximately log-normal distribution. UMI count data was treated as expression data as they are generally free of any biases due to gene length [156]. We modeled the gene expression (log) values as bimodal distribution. One mode represents genes having zero expression, and the other mode is the normal distribution of expressed genes. For expression value (log)  $x$  in a cell, the probability distribution function (pdf) can be defined as

$$f(x) = p_0 I(x = 0) + (1 - p_0) N(x; \mu, \sigma)$$

$N(x; \mu, \sigma)$  signifies Gaussian pdf for genes having non-zero expression, indicator function is represented by  $I(x = 0)$ , and  $p_0$  depicts a fraction of genes having no expression covering both those which are lowly expressed and those suffering from dropouts. The  $\mu$  and  $\sigma$  denote the mean and standard deviation of log values for non-zero expression only in a single cell. For each cell, it's on  $\mu$  and  $\sigma$  are used to transform log non-zero expression values into P-values making an assumption of Gaussian distribution. The advantage of converting gene expression values into P-values is that they can be combined using Brown's method. Thus, P-values of genes belonging to the gene set having dependence on each other are combined using Brown's method [157]. Hence, we combined P-values using Brown's method for a gene set having  $k$  genes with non-zero expression values is given by

$$P_{combined} = 1.0 - \phi_{2f}(\psi/c)$$

Here  $\psi = -2 \sum_{i=1}^k \log P_i$  and  $P_i$  represents the P-value of gene expression (log) of gene  $i$  in a cell or sample.  $\phi_{2f}$  represents the cumulative distribution function for chi-square distribution  $X_{2f}^2$ .  $f$  stands for the scaled degree distribution and is computed using the equation

$$f = E[\psi]^2 / \text{var}[\psi]$$

The value of  $c$  in the Pcombined equation is estimated as

$$C = \text{var}[\psi] / 2E[\psi]$$

where

$$E[\psi] = 2k$$

and

$$\text{var}[\psi] = 4k + 2 \sum_{i < j} \text{cov}(-2\log P_i - 2\log P_j)$$

This approach results in the computation of the combined P-value for every gene set in each cell independently. The covariance among log P-values of genes is computed by utilizing their values in all the cells from the same datasets. To have a robust estimate that is not influenced by one or two genes, a minimum cut off of 5 genes having some expression values is used to compute the combined P-value for a gene set. We use Brown's method for combining P-values and these P-values are corrected with the permutation-based approach that uses a null model as combined P-values might have been influenced by housekeeping genes, multiple hypothesis testing, and insignificant enriched gene-sets.

We downloaded multiple scRNA-seq datasets from the recount2 database [42] to create a null model. We randomly selected cells from these single cell studies to have a uniform representation of multiple cell types. We selected highly variable 500 genes using the coefficient of variation approach and conducted hierarchical clustering of the cells. Using the dynamic cut tree approach, we obtained clusters of cells. We selected 1000 pairs of cells so that every pair consists of cells belonging to different clusters to ensure that the null model has sufficient heterogeneity. For each pair, the mean expres-

sion value of all the genes was computed. Therefore, the null model comprised 1000 false cells or expression vectors, each corresponding to the mean of gene expression vectors of pair of cells. For the null model, combined P-values of each gene sets were computed for every false cell using the aforementioned method. Thus for every gene set or pathway, we acquired 1000 P-values equivalent to the number of false cells present in the null model. We consider the proportion of false cells in a null model to adjust or correct the P-value for a pathway in a specific cell with a lower combined P-value than the target cell.

### **2.2.3 Pseudotemporal ordering of single cells using UniPath computed pathway scores**

Almost all approaches of pseudotemporal ordering that leverage single cells directly exploit gene expression profiles. However, pathway or gene-set scores can elude covariate effects and impart weightage to biologically relevant pathway activity. Consequently, we devised a novel pseudotemporal ordering approach that can perform reliably on pathway enrichment scores of single cells. First, prior to detecting the order between the cell clusters, our approach conducts hierarchical clustering of cells. Then weighting of distance and finally learning the minimum spanning tree is performed. This strategy is based on Zhicheng and colleagues' findings that detect minimum spanning tree by considering direct distances between cells. For instance, Monocle 1 [195] can result in the incorrect links among cells owing to technical noise or other biases [99]. However, we do not follow the approach of Zhicheng and colleagues entirely as it does not provide reliable order of cells at the single-cell level. Therefore, we developed a method in which pathway scores based initial clustering of cells were followed by shrinking distances between every cell pair while accounting for their belongingness to the same class and based on the neighborhood index among their classes. We first compute every cell's top  $k$  nearest neighbor to compute the neighborhood index between classes. Then we count how many times each class's cells have top  $k$  neighbors in other classes. For instance, if cells belonging to class A have total of  $M$  neighbors in different classes with  $mb$  cells belonging to class B, then A's neighborhood index with B ( $A \rightarrow B$ ) is computed as  $mb/M$ . Distance between the cells in class A and B is shrunk by  $mb/M$ . Then this shrunk distance matrix is used for finding the minimum spanning tree, which is plotted



using the netbioV R library [196]. This approach of finding a minimum spanning tree has lower chances of being impacted by noise as the shrinking distance between cells is based on consensus information.

#### **2.2.4 Differential pathways co-occurrence analysis**

To determine the relevance of the different patterns of pathway co-occurrence in two kinds of cells permutation test was used. The difference in values of Spearman correlation of pathway scores, i.e., for adjusted p-value in two kinds of cells, was computed for every pathway pair. This difference is referred to as true difference. Firstly, group labels of the cells were subjected to random shuffling. This was followed by the computation of differences in the Spearman correlation values of adjusted p-values for the two shuffled groups. This resulted in the compilation of vectors of false differences in correlation for a pair of pathways utilizing shuffled groups. The P-value is determined as a fraction of false differences bigger than the true difference in terms of absolute value. Here, we have computed the difference in Spearman correlation for adjusted p-value, which increases the robustness of this approach as it turns into rank-based scores that assist in subduing effects because of only one or two genes. Thus, the correlation between two pathway pairs manifested through adjusted P-values is less likely to be influenced by one or two genes or outliers.

#### **2.2.5 Experimental methodology**

Wang et al. [206] reported the source and culture conditions for Tumour sphere (TS) and Adherent (Adh) cells. TS cell line obtained from lung cancer patient was maintained in the medium supplemented with DMEM/F12 (US Biomedical), Sigma Bovine Serum Albumin (4mg/ml), Non-essential amino acids, sodium pyruvate (Life Technologies), and Epidermal Growth Factor (20 ng/ml), bovine Fibroblast Growth Factor (4 ng/ml) and Insulin – Transferrin Selenium (Sigma). While for Adh, cells were cultured in similar conditions as described above but without the addition of bFGF, EGF, ITS, and BSA. Media was also supplemented with fetal bovine serum (10%).

## **RNA extraction, library preparation, and single-cell sequencing for NSCLC cells**

Non-small-cell lung carcinoma (NSCLC) single-cells in suspension were dissociated using trypsin followed by loading into C1 96 well-integrated microfluidic chip (IFC) as per the manufacturer's instruction. The Fluidigm-C1 system captured single cells on C1 96 (large size). Then, to identify viable single cells and discard doublets from the single cells captured, single cells were imaged utilizing an auto imaging fluorescent microscope. The SMART-seq2 protocol was utilized to prepare reagents for reverse transcription and cDNA pre-amplification to be loaded into the IFC. Then, reverse transcription and cDNA amplification were automatically performed through the SMART-seq2 script in the C1-Fluidigm machine. cDNA was harvested from the C1 chip, and picoGreen assay was used to quantify the samples and normalized to the range of 0.2–0.3 ng/ $\mu$ l. The cDNA product quality was assessed using an Agilent bio-analyzer machine. Utilizing the Illumina Nextera XT Library Prep Kit, the single-cell cDNA was barcoded on a 96-well plate. Single-cell libraries with unique barcodes were pooled and sequenced using an Illumina HiSeq-Hi-output-2500 sequencer. In total, we obtained 87 transcriptsomes of TS cells and 75 transcriptsomes of Adh cells.

### **2.2.6 Data availability**

The raw and processed single-cell RNA-seq lung cancer data is available from GEO id: GSE156138

### **2.2.7 Code availability**

<https://reggenlab.github.io/UniPathWeb> and <https://github.com/reggenlab/UniPath>

## **2.3 Results**

To transform scRNA-seq gene expression profiles into pathway or gene set scores, we treat each cell independently. The gene expression quantified in terms of RPKM (read per Kilobase per million) and FPKM (fragment per kilo per million) generally has a

bimodal distribution. Here, one mode represents genes having zero expression values, and the other mode corresponds to the genes having non-zero expression values. We have converted the non-zero gene expression values from different gene expression quantification strategies, including FPKM, RPKM, TPM, and UMI-counts, to P-values utilizing log-normal distribution based on the theoretically accepted assumption that non-zero expression values within the sample or cells follow a log-normal distribution [143]. In the case of UMI counts, we use log-transformed UMI counts to compute p-values. UMI counts don't require to be normalized by gene length to get gene expression [156]. The study by Furusawa et al. [63] also supported the assumption that non-zero expression values follow a log-normal distribution. Furthermore, skewed distributions frequently fit log-normal [130]. We use Brown's method which tends to lower the effect of co variation between genes to combine P-values of the genes in the gene set. The P values combined using Brown's method are adjusted through employing a null model created through the Monte Carlo technique (methods). The null model is used to pinpoint cell-type-specific pathway activity. The adjusted p-value of a gene set or pathway is referred to as its score in a single cell.

### **2.3.1 Assessment of UniPath's approach**

In the absence of standard gold standards, the evaluation of gene set enrichment methods for heterogeneous bulk samples is not trivial. However, to test methods like UniPath, single cells from well established cell lines, cell-type-specific marker gene sets can be directly used. We used cell type-specific marker gene sets to compare our method UniPath with the existing single-cell methods like PAGODA [56] and AUCell [3]. We also compared our approach with GSVA [81] which was designed for bulk RNA-seq. UniPath outperformed PAGODA, GSVA, and AUCell in computing gene-set enrichment for the right cell type as one of the top 5 enriched terms in a systematic assessment using scRNA-seq gene expression profiles from ten publicly available datasets (Figure 2.2A). We performed comprehensive evaluation of UniPath using 10 single cell studies (see Appendix A, Figure A.1). The aim of using the marker gene-set specific to multiple cell-types was to test the correctness of gene-set enrichment for the downstream analysis involving clustering and pseudotemporal ordering. To elaborate further, we created a set of non-immune-associated pathway terms, and as a spike in, we included

two known pathway terms related to B cells. Two T cells associated gene sets were also added to the same set. We looked at how many cells had these suitable terms in the top 5 enriched terms (Figure 2.2B).

UniPath was significantly more accurate than the other 3 methods namely, PAGODA, AUCell, and GSVA in identifying the correct respective pathways in the top 5 enriched pathway terms (Figure 2.2B) in this control experiment for B cell and T cell [96]. To further ensure unbiased results, we used GSEA [186] to build a reference list of substantially enriched gene sets considering FDR of 0.2 in T cells compared to others in the mouse cell atlas (MCA) dataset [78]. We evaluated UniPath and compared it to the other three methods regarding the appearance of gene sets from the reference list in the top ten terms in each single T cell. We have considered gene sets in the reference list as positives for T cells. On comparing to PAGODA, AUCell, and GSVA, we noted UniPath had a significantly higher level of existence of reference gene-sets, i.e., positives among the top ten term enriched terms (Figure 2.2B). We found similar findings when we repeated the experiments with B cells, demonstrating that UniPath better estimates gene set enrichment for specific cell types in single cells.

We also looked at the consistency of UniPath's pathway enrichment compared to the other three approaches—PAGODA, GSVA and AUCell. We investigated scRNA-seq gene expression profiles of B (GM12878) [124] cells while combining them with different cell types each time. The pathway or gene set scores were not consistent for PAGODA and GSVA, and for each cell, scores were relying on the cell type composition of the dataset. On the other hand, UniPath and AUCell computed gene set enrichment scores for a cell remain consistent and are unaffected by adjacent cells (Figure 2.2C). Thus, UniPath also solves the problem of consistently pinpointing correct gene sets and related pathways for every single cell, regardless of the degree of cellular heterogeneity in the scRNA-seq data.

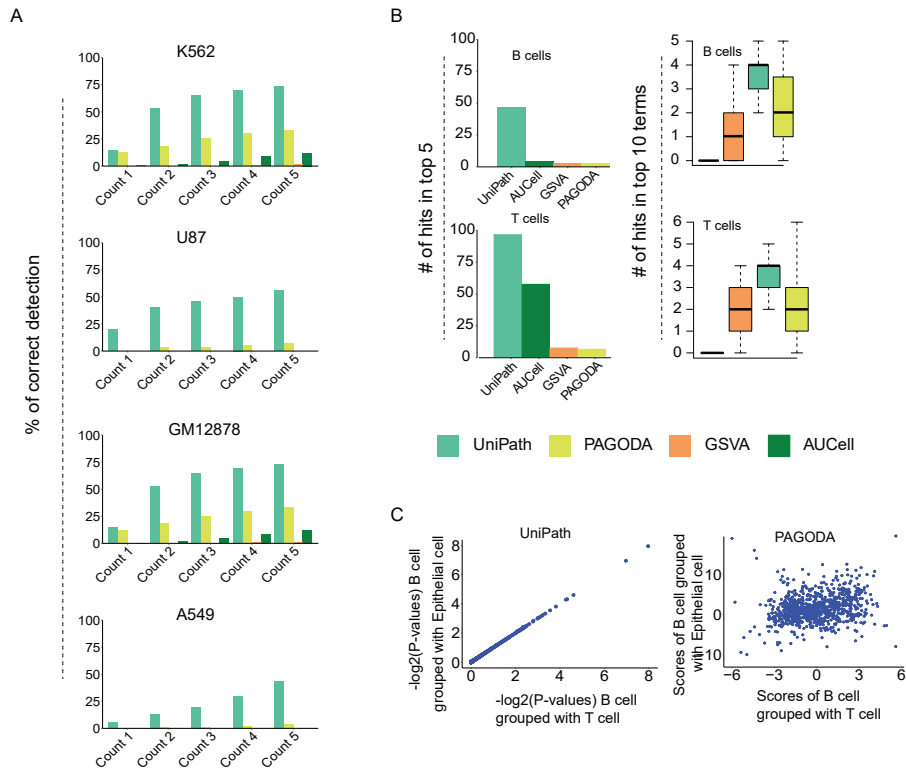


Figure 2.2: Assessment of UniPath utilizing scRNA-seq gene expression profiles. **(A)** Estimation of accuracy of correct detection of cell-type-specific gene sets among top enriched terms. **(B)** Estimation of accuracy of pathway enrichment for scRNA-seq gene expression profiles of B and T cells using UniPath and other three methods. The bar plots are created using a gene set list consisting of non-immune gene sets along with 2 relevant gene sets for B and T cells (positives). The boxplots are created using all gene sets but pathways or gene sets in the positive set were selected based on applying GSEA to B and T cells from the MCA data. The boxplots depict counts of positives in the top ten enriched pathway terms. **(C)** Scatter plot showing consistency in pathway scores of UniPath when T or epithelial cells are combined with B cells on comparing to PAGODA. When the similar cell is combined with other cells, PAGODA’s pathway enrichment scores vary. UniPath’s output, on the other hand, is consistent.

### 2.3.2 UniPath’s pathway or gene set enrichment scores as an alternative dimension-reduction approach for scATAC-seq profiles

To transform scATAC-seq profiles into pathway or gene set enrichment scores, UniPath underscores the enhancers first by using global accessibility scores to normalize read-count on peaks (see methods). It accomplishes this by intersecting a pre-collated list of genomic regions with pre-computed global accessibility scores with the peak list of an input scATAC-seq profile. The global accessibility score for a genomic site is

proportional to the number of times it was detected as a peak in bulk open chromatin profiles. The purpose of normalizing each peak's read count utilizing its global accessibility score is to maintain consistency while circumventing the fixing of sequencing depth variability and dropouts. For each cell, a foreground set exhibits genomic locations with a high normalized read count. UniPath then uses genes proximal to peaks in the foreground list to calculate the P-value or statistical significance of enrichment of gene-sets using the Hypergeometric or Binomial test for each cell. The P-value of the pathway or gene set enrichment is referred to as its score. We carried out a thorough evaluation of the bulk ATAC-seq of immune cells [45] and various single-cell ATAC-seq profiles using a cell-type-specific marker gene set [28] [27].

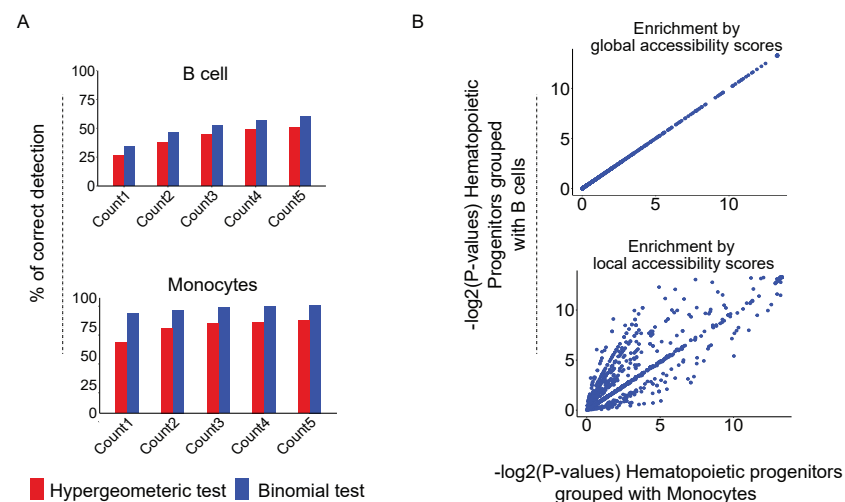


Figure 2.3: Evaluation of UniPath on scATAC-seq data. **(A)** Accuracy of UniPath in identifying precise gene-sets among top enriched terms for single-cell open chromatin profiles of B cell (GM12878) and Monocytes using global accessibility scores for highlighting enhancers. **(B)** The pathway scores of Hematopoietic progenitor cells combined with scores of B cells are more consistent when enhancers are highlighted using global accessibility scores in comparison to mean based normalization (local accessibility scores)

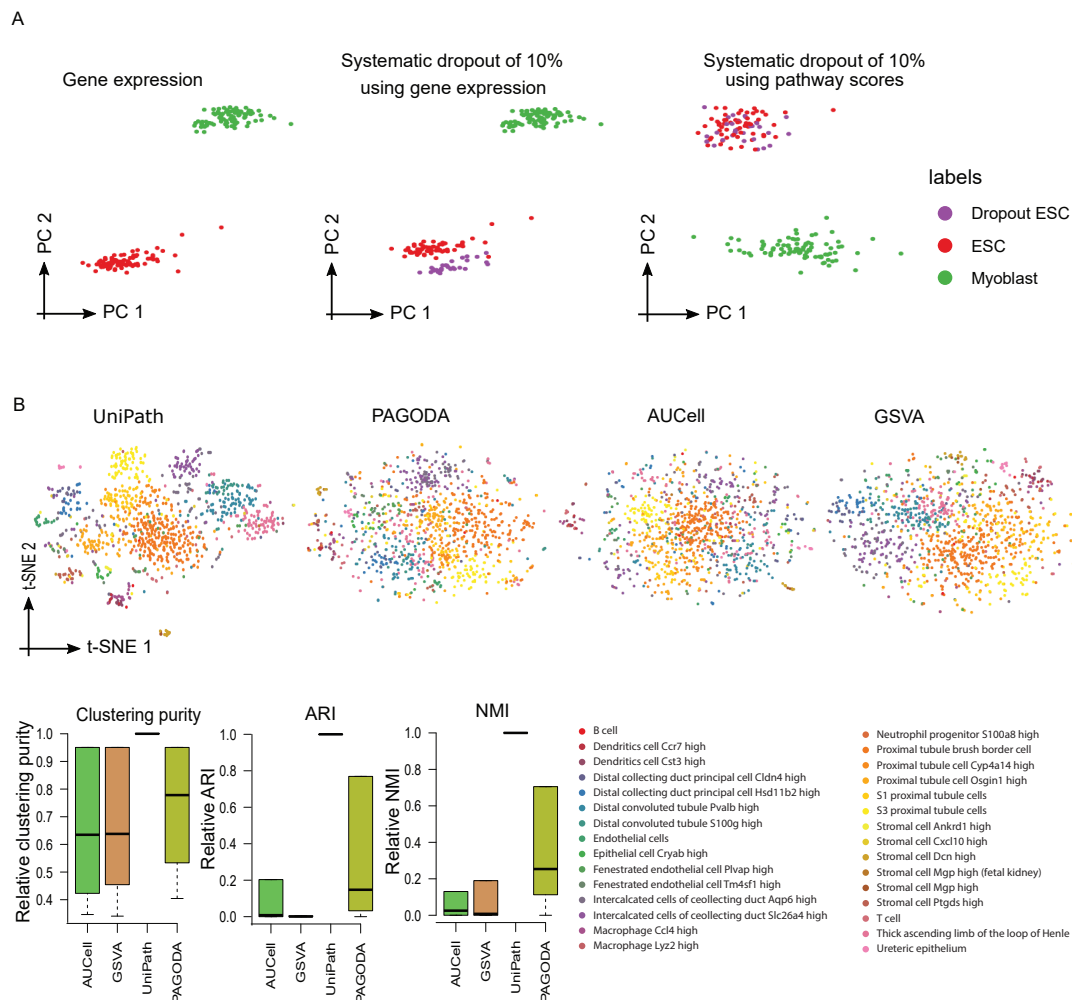
In bulk profiles and single-cell ATAC-seq profiles, UniPath highlights the right cell type among the top 5 enriched gene-set most of the time (Figure 2.3A). The availability of bulk open-chromatin profiles allows for creating a global peak list with accessibility scores. If there aren't sufficient publicly available open-chromatin profiles for a given species, then one can utilize local accessibility scores (mean/median) for normalization in UniPath for computing pathway scores. Local accessibility scores, on the other hand, are dependent on the cell type composition in the single-cell dataset, which may result in

inconsistency in computing gene-set or pathway enrichment scores (Figure 2.3B). Thus, the UniPath reports consistent and mostly accurate pathway or gene set enrichment scores for every cell independently using scATAC-seq profiles.

### **2.3.3 Dropout and batch effect handling and evaluation through visualization and clustering**

There is heterogeneity in the dropout rate among cells in single-cell scRNA-seq profiles, and dropouts in genes could either be systematic or random. Systematic dropouts are generally caused by variations in sequencing depth or degradation levels of RNA between different batches of samples. We tested the ability of UniPath to withstand variability in systematic dropout among cells by simulating systematic dropouts using publicly available scRNA-seq datasets. We noticed that using PCA on raw gene expression data resulted in the formation of the artifactual cluster due to systematic dropout. However, a similar type of cells stayed in the same cluster when using UniPath pathway scores, regardless of the dropout rate pattern (Figure 2.4A). UniPath pathway scores yielded robust results as the same types of cells grouped together when scRNA-seq profiles with 10 percent systematic simulated dropout. In contrast, other methods, PAGODA, AUCell, and GSVA, formed artifactual clusters for cells with 10 percent systematic dropout on the same dataset. We also used the publicly available dataset of microglia cells having systematic bias or dropout. This dataset comprised single cell profiles of fresh microglial cells and profiles of nuclei isolated from frozen tissues [68]. When gene expression profiles were subjected to t-SNE based visualization, frozen cells' expression profiles showed distinct clusters. Frozen cells formed their separate cluster in the t-SNE plot even when highly variant genes were used. However, fresh and frozen microglial cells were grouped together in a t-SNE scatter plot created using UniPath's pathway scores. In addition to being robust to the systematic dropouts, UniPath also facilitates correcting batch effect before computing adjusted P-values for pathway enrichment using existing tools. UniPath's framework avoids normalization artifacts caused by the sequencing depth issues and variable dropout rate, allowing it to be used for systematic clustering of single cells. UniPath-based pathway scores yielded equivalent clustering-purity to raw gene expression-based results during hierarchical clustering. Thus, UniPath and gene expression-based clustering and visualization re-

sults could be comparable but for scRNA-seq harboring systematic dropouts, UniPath based pathway scores are more valuable than raw gene expression to avoid artifactual clusters.



**Figure 2.4: UniPath’s pathway score based reduction of artefacts and clustering. (A)** Visualization of scRNA-seq profiles of human ESC and myoblasts in terms of PCA. PCA was performed using gene expression data from scRNA-seq. In PCA-based visualization of raw gene expression profiles, simulation of 10 percent systematic dropout in genes in some hESCs results in forming a distinct group of hESCs. Regardless of systematic dropout, PCA on UniPath’s pathway scores resulted in the clustering of ESCs in the same cluster. **(B)** scRNA-seq gene expression data of kidney cells from mouse cell atlas data were transformed to pathway scores. Pathways scores were subjected to t-SNE based visualization, and clustering efficiency was assessed using clustering purity, ARI, and NMI using different eps parameters of the db-scan method. Visualization and clustering results of scRNA-seq profiles transformed into pathway scores were compared for four different methods.

We also evaluated visualization and grouping for four different methods of scRNA-seq gene expression profiles transformed into pathway scores. We used t-SNE for vi-



sualization and dbSCAN with different eps values [194] to perform clustering of t-SNE coordinates. UniPath pathway scores provide better clustering purity for different values of the dbSCAN eps parameter. The adjusted rand Index (ARI) and normalized mutual information (NMI) values for UniPath-based clustering were also higher than the other three methods: PAGODA, AUCell, and GSVA (Figure 2.4B). Better visualization and clustering with UniPath enabled us to find biologically meaningful and relevant groups of stromal cells in the Uterus tissue in MCA dataset. Using pathway scores of imputed scATAC-seq profiles for hierarchical clustering resulted in reasonably good clustering purity (Figure 2.5).

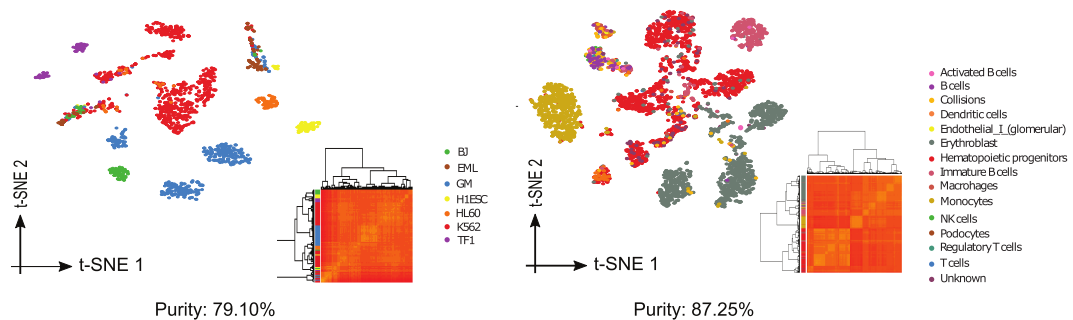


Figure 2.5: Evaluation of clustering purity in scATAC-seq profiles. t-SNE based visualization of scATAC-seq profiles transformed into pathway enrichment scores.

We further compared UniPath’s clustering and visualization of scATAC-seq profiles with the output of the other two methods, ChromVar [174] and SCALE [216]. However, these methods handle scATAC-seq count data and are not for computing gene sets or pathway enrichment scores. On comparing to output of ChromVar and SCALE, UniPath pathway scores of imputed scATAC-seq profiles were better or comparable in terms of visualization and clustering results. The high clustering purity achieved using gene set or pathway enrichment scores demonstrates that UniPath could be an efficient method for characterizing single-cell open chromatin and transcriptomics datasets in terms of pathway activities.

## 2.4 Discussion

UniPath bridges the gap between the need for consistent and uniform gene set enrichment methods for various applications and the availability of a wealth of single-cell

transcriptomics and open chromatin data. UniPath is scalable and provides consistent gene set enrichment scores due to its novel strategy of treating every cell independently and the usage of global null models. Further, utilizing the normalization-free approach for converting single-cell gene expression profiles into pathway scores helps handle noise in cell-to-cell variability, sequencing depth, and gene dropouts. UniPath can withstand systematic dropouts and can handle strong technical batch effects in scRNA-seq profiles.

UniPath pathway scores resulted in reasonably better visualization and clustering purity than other comparable methods. The pathway scores computed using UniPath improved visualization and clustering accuracy scRNA-seq gene expression profiles including UMIs and non-UMI datasets, compared to other comparable methods (PAGODA, AUCell, and GSVA). Downstream analysis of scRNA-seq and scATAC-seq profiles is similar after transformation into pathway scores. Therefore, UniPath provides a uniform and stable platform for investigating single-cell transcriptomics and open chromatin data in terms of pathways. It is recommended to perform imputation using other techniques to improve UniPath's performance on scATAC-seq profiles [70]. Using UniPath's approach, multiple scATAC-seq profiles can be transformed to the same feature set regardless of differences in peak list. Further, we demonstrated that scRNA-seq fresh and frozen samples were mixed as visualized through t-SNE instead of gene expression profiles in pathway space. Overall, UniPath is a robust tool for the conversion of scRNA-seq and scATAC-seq profiles into pathway scores. Further, UniPath can be used for other sequencing technologies, including single-cell spatial transcriptomics [214], cDNASE-seq [44], or SNARE-seq [38].

## CHAPTER 3

### Applications of UniPath

A wealth of high throughput sequencing data have become available that uncovers cell states in different diseases and normal conditions, thus facilitating understanding of the complex biological system. However, the challenge is in procuring reliable and predictive molecular biomarkers for the identification of a disease or biological state, defining personalized treatment regimens, and unveiling critical biological processes and underlying mechanisms [176]. Biological pathways are of particular interest as they drive many biological processes that are crucial for classifying complex diseases such as Cancer, functional characterization of cellular heterogeneity, and diversity. Thus, the use of pathway activity scores in the single-cell domain has emanated as a powerful tool to exploit cellular heterogeneity for extracting novel and relevant biological information for numerous applications. In this chapter, we will discuss the applications of UniPath in the pathway space.

#### 3.1 Pseudotemporal ordering, visualization of the lineage potency and pathway co-occurrence continuum

Representation of single cells in terms of pathway scores provides a new dimension in the context of the cell-to-cell similarity measure. It further helps in suppressing the effect of known covariates such as cell cycle-related pathways or terms and tissue microenvironment and culture conditions. Current methods [169] for pseudotemporal ordering of single cells are primarily intended to handle gene expression or read-count data. Therefore, we enhanced the utility of our method UniPath with a novel pseudotemporal ordering method for single cells that rely on pathway scores. For pseudotemporal ordering, before learning a minimum spanning tree (MST), two levels of distance shrinking among single cells are applied based on their prior clustering results and continuum among their classes. After initial clustering, we use a KNN-based approach to determine accurate temporal order between cell clusters in order to detect a contin-

uum between different classes. The MST did not reveal the true order of cells in some cases when two levels of distance shrinkage among cells were not applied. Overall, we discovered that UniPath could capture the approximate true order of single cells from single-cell RNA-seq gene expression and ATAC-seq profiles. We applied UniPath to scRNA-seq gene expression profiles of human embryonic stem cells and their differentiated states collected at different time points 0, 12, 36, 72, and 96 hours during differentiation to definitive endoderm (DE) to capture pseudo temporal order [40].

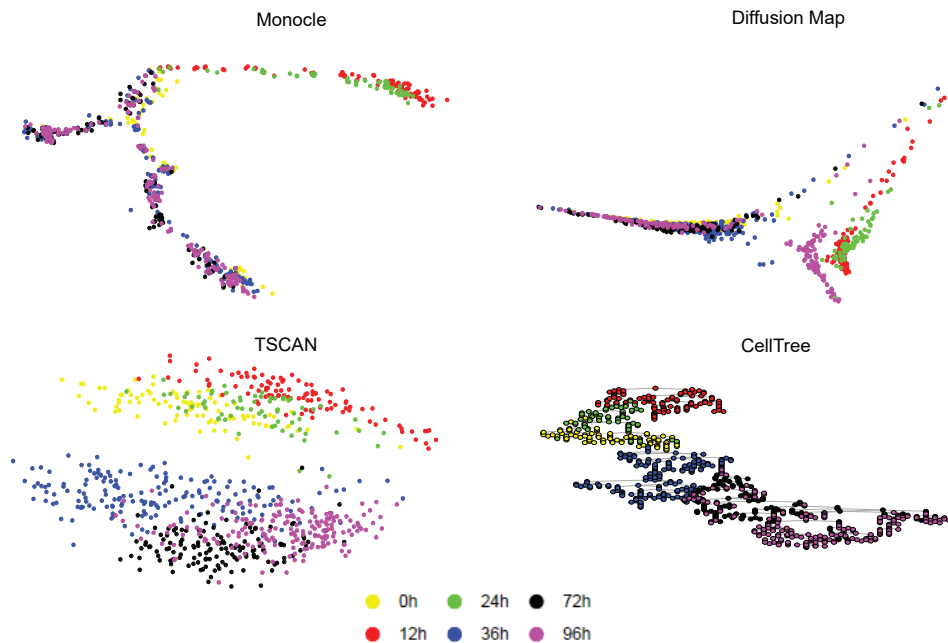


Figure 3.1: Comparison of different pseudotemporal ordering methods. Pseudotemporal ordering of single cell profiles of human embryonic stem cells harvested at multiple time points 0, 12, 24, 36, 72 and 96 hours during course of differentiation towards definitive endoderm revealed incorrect ordering of Monocle, CellTree and DiffusionMaps using gene expression matrix.

Other approaches including Monocle [195], TSCAN [99], DiffusionMap [77] and CellTree [52] for pseudotemporal ordering with gene expression matrix (Figure 3.1) predicted incorrect cell order for the same dataset [40]. But in the case of UniPath, when we removed the gene set or terms linked with the cell cycle, we attained the correct order of single cells (Figure 3.2A). We found that at 0 and 12 h, the gene set scores for the cell cycle S phase are higher, probably due to the high rate of proliferation (Figure 3.2B).

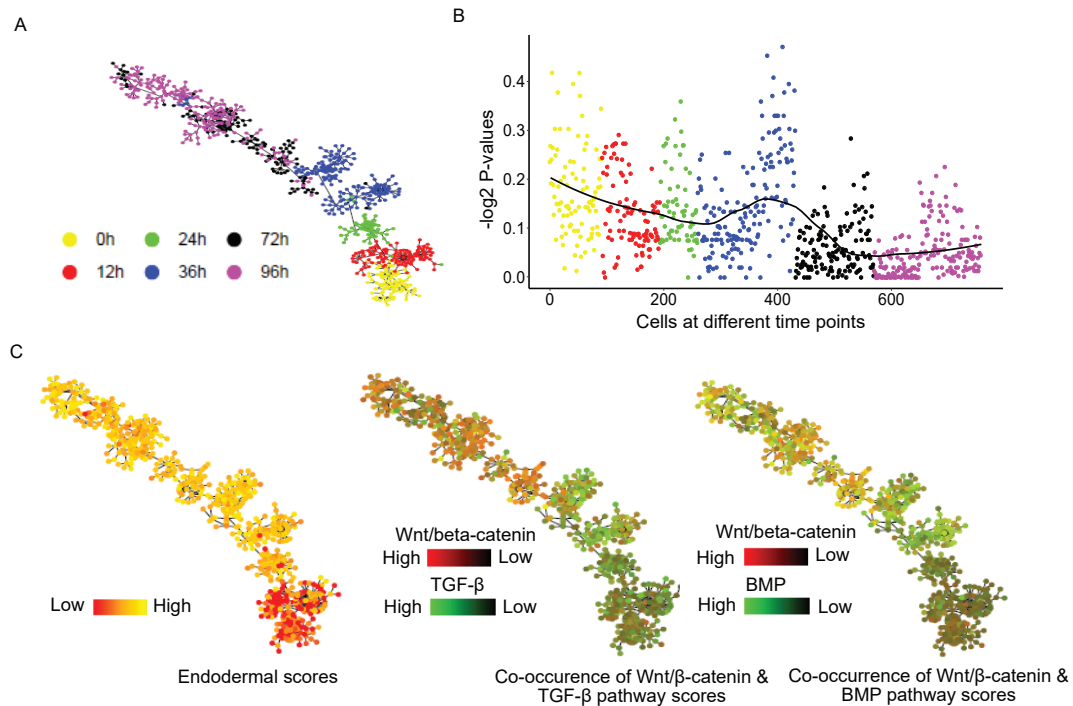


Figure 3.2: Pseudotemporal ordering of single cells using pathway scores and visualization of lineage potency and pathway co-occurrences on temporal tree using UniPath. (A) UniPath's pathway score based temporal ordering was correct as per the true time points. (B) Distribution of pathway scores of S-phase at different time points. (C) Gradient of endoderm lineage and pathway co-occurrence is shown on pseudo temporally ordered tree. Endodermal scores increase as cells proceed towards differentiation. Pattern of co-occurrence of Wnt/beta-catenin and BMP pathway scores on temporal tree revealed that BMP pathway is more enriched at 24 and 36 hours.

As the cells differentiated into endoderm, the gene-set score of the S phase decreased. At 36 h, we found two groups of cells. One group of cells exhibits a significantly lower cell cycle S phase gene set score than another group. The presence of such batches of cells suggests that the cell cycle effects act as a potential covariate in predicting the temporal ordering of cells. UniPath can be utilized for visualization of lineage potency continuity and co-occurrence of two gene sets or pathways on pseudotemporally ordered tree, in addition to treating known covariates (Figure 3.2C). UniPath allows finding clusters of pathways and interpretation of pathway co-occurrence patterns, which assist in uncovering context-specific regulations. UniPath is useful in predicting the correct temporal order of single cells represented in terms of pathways and analyzing patterns of co-occurrence of pathways at various stages during the course of the differentiation of cells.

## 3.2 Analysis of large scale mouse cell atlas scRNA-seq dataset

UniPath's use of global null models ensures consistency in calculating pathway enrichment scores for single-cell, allowing for horizontal scalability. We were able to transform UMI counts or gene expression profiles of over 61000 single-cells from the mouse cell atlas (MCA) [78] dataset by splitting it into small chunks of cells using UniPath's horizontal scalability, speed, and consistency. We chose 49507 cells with >800 genes (expressed genes). The transformed pathway scores from the scRNA-seq MCA dataset were subjected to t-SNE [197] based visualization and subsequently to dbscan [194] based clustering of t-SNE coordinates (Figure 3.3). This result showed that most cells were correctly grouped according to their tissue type. Some cells did not cluster with the cluster of their origin of tissue, as expected, but instead formed their own class. For example, immune cells from various organs were clustered together in cluster numbers 13, 14 and 15 (Figure 3.3). Unexpectedly, some non-immune cells belonging to different tissues clustered together, pointing towards convergence. This has only been reported infrequently by investigation of single-cell but is reinforced by scientific reports and literature. Pathway scores-based clustering resulted in biologically significant co-clustering of cell types belonging to different organs or tissues. For instance, cluster 40 harbored Afp+ fetal liver hepatocytes, a few Fabp1+ hepatocytes and Afp+ placental endodermal cells. Afp+ placental endodermal cells and Afp+ fetal liver hepatocytes were part of different groups in the original MCA study. It has previously been demonstrated that placenta-derived multipotent cells (PDMCs) having Afp (Figure 3.4A) expression have endodermal characteristics and can easily differentiate into hepatocyte-like cells [90]. Differential pathway analysis revealed that among the top 50 enriched pathways, 22 pathways were common for Afp+ placental endodermal cells and hepatocytes cells of cluster number 40. These pathways were linked to lipid metabolism. However, in t-SNE-based visualization, there was a clear distinction between hepatocytes and Afp+ placental endodermal cells of Cluster 40 (See Figure 3.4A). We also noticed convergence in cluster 3, which contained virgin mammary gland luminal-epithelial cells, alveoli cells, and uterus glandular epithelial cells. Another example of convergence was observed in cluster number 52 that harbored Col10a1+ and Cmnd+ bone marrow mesenchyme stromal, chondrocytes cells, and pre-osteoblast. Past studies have revealed

that bone marrow mesenchyme stromal or mesenchymal stem cells possess the ability to transform into pre-osteoblast and chondrocytes cell states [55] [11]. Cxcl1+ MSC, on the other hand, clustered with trophoblast stem cells in cluster number 21. It can be seen that when cell types from various organs converged in a major cluster, they didn't entirely overlap but instead formed their own sub-cluster within their main (Figure 3.4A). Nevertheless, convergence to major class indicates a lowering of covariates in pathway or gene set scores, which results in the grouping of cells with similar states together. Thus, UniPath added a new aspect to the clustering of single cells and revealed that although there are specific cells required for the functioning of the organ but it also harbors cells having regulatory states similar to other cell types in the body.

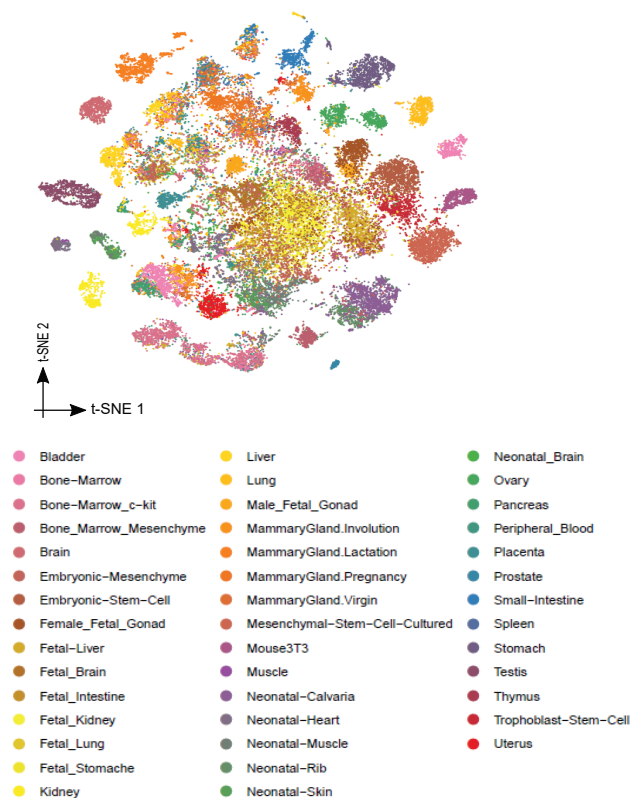


Figure 3.3: Clustering of large scale scRNA-seq mouse cell atlas data. t-SNE based visualization of mouse cell atlas data in terms of pathway score.

### **3.3 Utilization of pathway scores for revealing minor classes and annotating unlabelled cells**

Technical noise, the effect of a few covariates, sparsity and dropouts can be reduced with feature extraction in terms of pathway scores. As a result, it can assist in highlighting clusters of cells that would otherwise go undetected when using raw gene expression. For instance, interrogation of pathway scores of brain tissue revealed a new cluster of oligodendrocyte precursor cells. Notably, genes *Tuba1a*, *Sirt2*, *Cd9*, *Plp1*, and *Bcas1* [4] [134] [59] [98] exhibited higher expression in oligodendrocyte precursor cells from the new identified cluster (Figure 3.4B). These genes play a role in oligodendrocyte precursor differentiation into mature oligodendrocytes. We discovered two new sub clusters of unlabelled bladder cells in the MCA dataset (Figure 3.4C). Cells in one of the newly discovered bladder clusters were dendritic cells (*Cd74* high). Thus, UniPath based analysis enabled us to pinpoint a few new groups of cells not detected in the original study that employed read counts. We also attempted to annotate a few unlabelled cells.



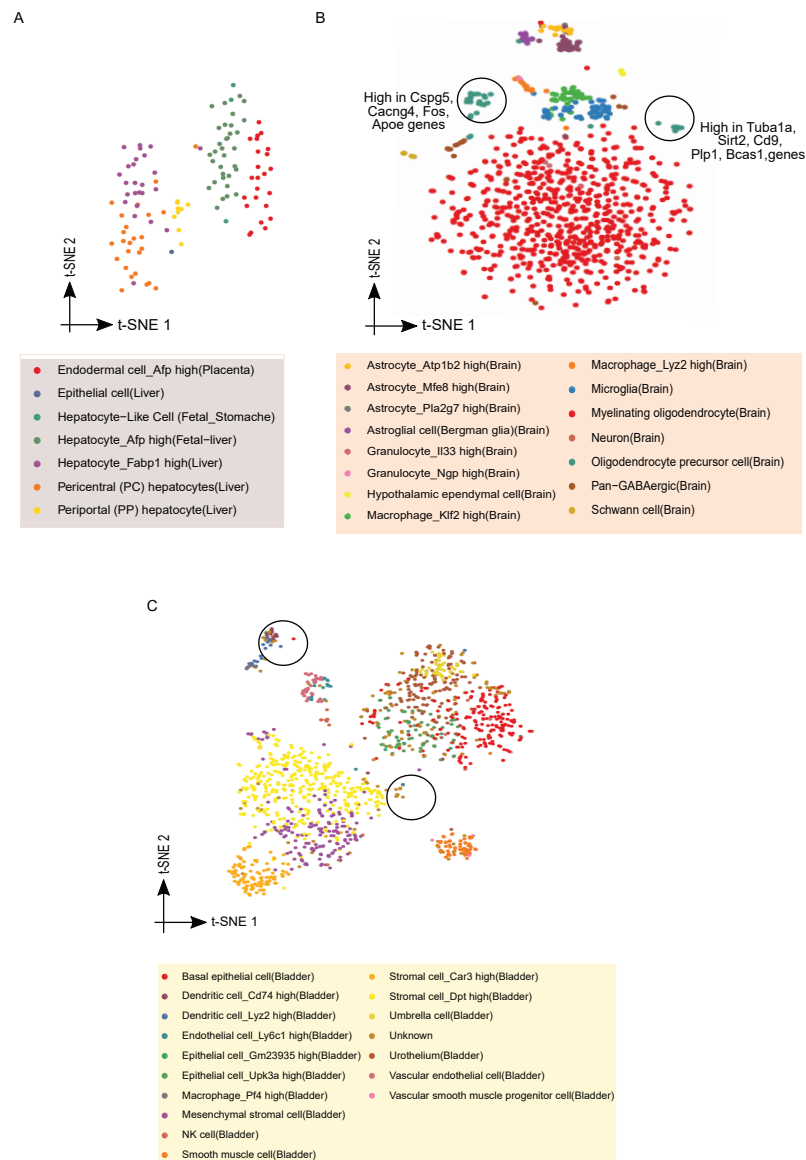


Figure 3.4: Analysis of large scale scRNA-seq mouse cell atlas data using pathway scores. **(A)** 2D-scatter plot of cluster 40 showing distinct clusters of Afp+ hepatocytes and AFP high placenta endodermal cells. **(B)** Visualization of brain cells revealing two distinct clusters of oligodendrocyte precursor cells as obtained using pathway scores. **(C)** Pathway scores of bladder cells subjected to t-SNE based visualization revealed two separate clusters of unknown cells. One of the unknown cluster was identified to be cd74 high dendritic cells.

### 3.4 Inference of context-specific regulation in cancer cells

We further scrutinized the utility of UniPath in interrogating context-specific regulations in cancer which are frequently needed in precision oncology and precision medicine. Wang et al [206] recently showed that two types of non-small cell lung cancer (NSCLC) cell lines, non-adherent tumorspheres (TS) and adherent (Adh) cells, have different metabolic profiles. Using mouse xenograft models, they showed that non-adherent TS cells have a higher tumorigenic potential than adherent TS cells. We did single-cell expression profiling on 162 NSCLC cells, about half of which were TS cells, and the rest were Adh cells. The pseudotemporal ordering of these cells is shown in Figure 3.5A.

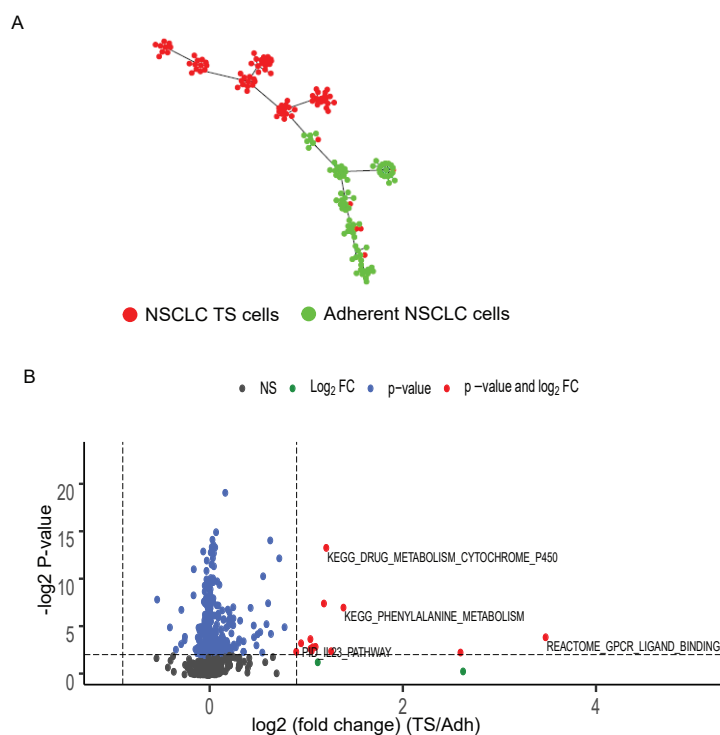


Figure 3.5: Pseudotemporal ordering and differential enrichment analysis of pathways in NSCLC dataset. This dataset consists of FPKM values for TS Adh cells which were transformed into pathway scores using UniPath based approach. (A) Pseudotemporal ordering of single cell RNA-seq of lung cancer transformed into pathway scores. (B) Volcano plot showing differential enrichment of pathways in TS and Adh cells.

After converting scRNA-seq gene expression profiles into pathway scores, we used Wilcoxon rank-sum test to perform differential pathway enrichment analysis. The results showed higher enrichment of the IL23 pathway, GPCR ligand binding, cytochrome P450 drug, and phenylalanine metabolism pathways in TS cells (Figure 3.5B). NSCLC

plasticity and proliferation are known to be linked to GPCR and IL23 signalling [116] [105] [14]. Cytochrome P450 is also implicated in the growth of tumors [152].

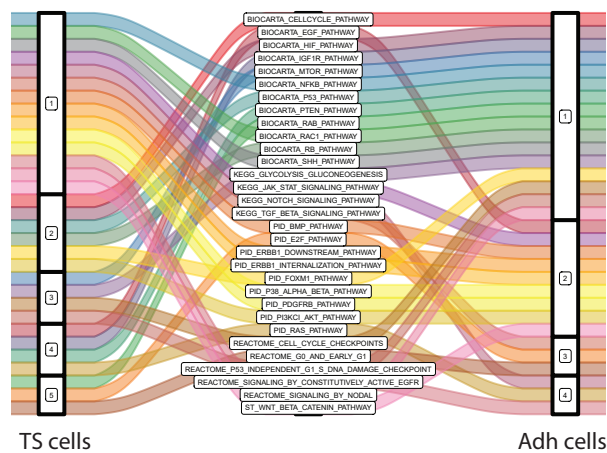


Figure 3.6: Alluvial plot depicting alterations in co-enrichment patterns of pathways in Adh and TS cells.

We used a technique that is seldom used for scRNA-seq. For pathway and gene-set pairs, we used co-occurrence and differential co-occurrence analysis. The Wnt pathway exhibited the highest correlation with the stemness associated gene set in the TS cells, but it was not in the top correlated pathways with stemness in Adh cells (Figure 3.8A). In contrast to Adh cells, the Wnt/beta-catenin pathway in TS had a markedly higher correlation with the TGF beta signaling pathway (P-value < 0.05, Jaccard index = 0) (Figure 3.8A). However, there was no significant difference in TGF-beta pathway enrichment between TS and Adh cells.

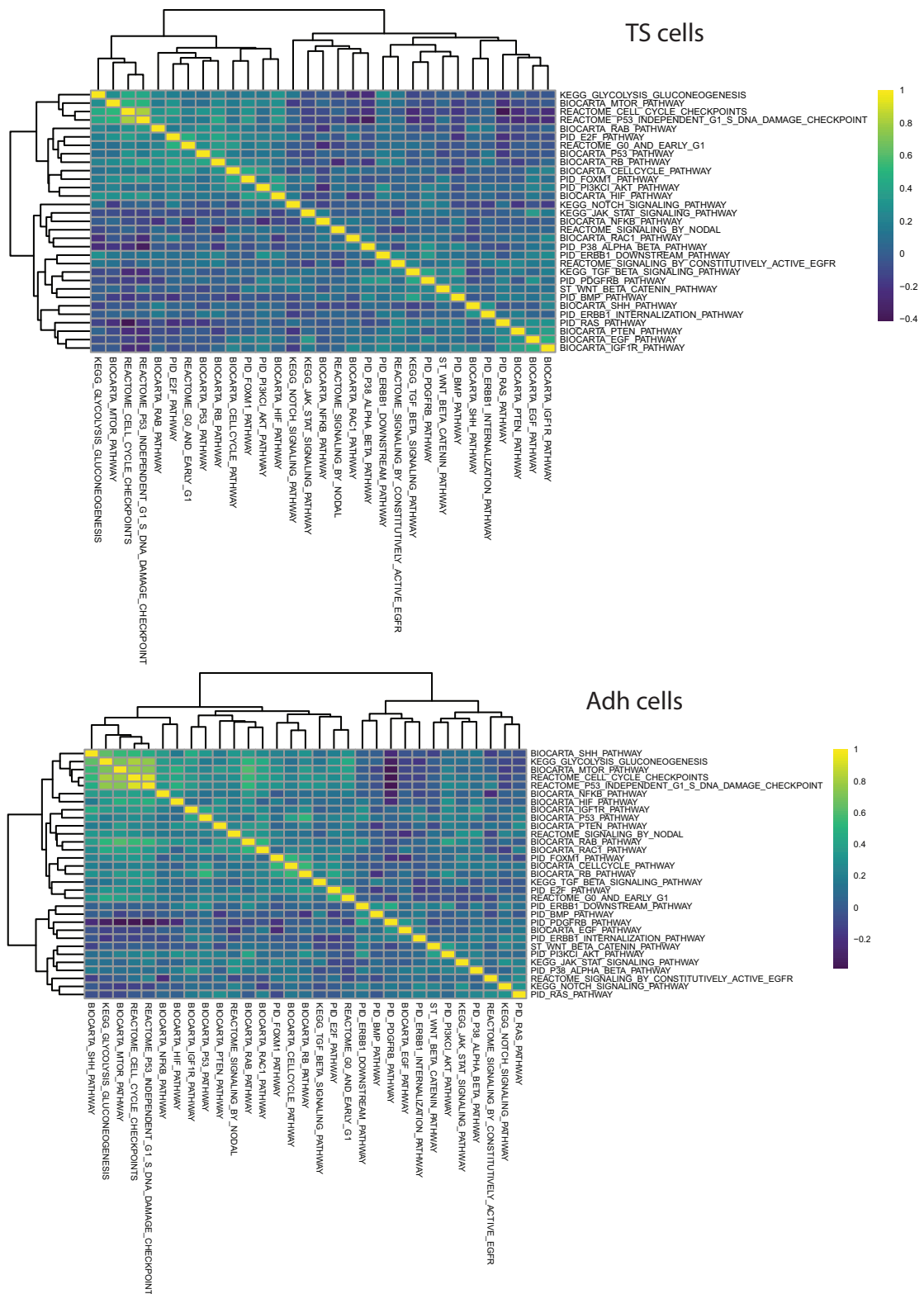


Figure 3.7: Heatmaps show alterations in enrichment patterns of pathways in TS and Adh cells.

In cancer cells, the Wnt/beta-catenin and TGF-beta pathways have been demonstrated to facilitate the epithelial to mesenchymal transition (EMT), which is linked with increased tumorigenic potential [85]. Furthermore, synchronous over-activation of the

Wnt/beta-catenin and TGF-beta pathways has been shown to promote tumorigenesis and therapy resistance in NSCLC cells [32]. In TS cells, pathways including TGF-beta, Wnt/beta-catenin, and PDGFRB grouped together, based on hierarchical clustering of 31 selected pathways. However, in Adh cells, the WNT/beta-catenin pathway was clustered with ERBB1 and PI3K1 signaling. The distinction in the co-occurrence pattern of Wnt/beta-catenin in the two lung cancer cell types, i.e., TS and Adh cells, and foreknowledge about the impact of its co-stimulation with TGF-beta in NSCLC point to a probable reason of higher tumorigenicity in TS cells (Figure 3.6, 3.7). Glycolytic intermediates were found to be more abundant in Adh cells, as reported by Wang et al.

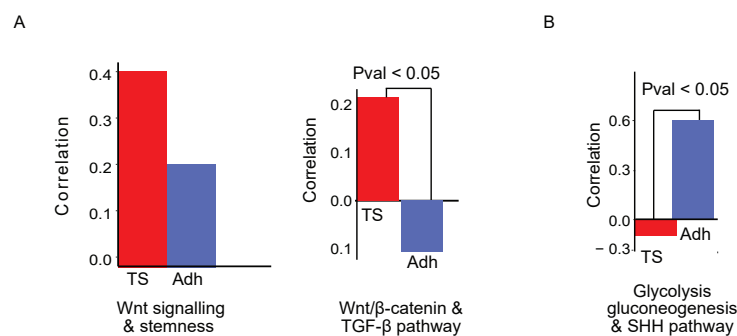


Figure 3.8: Analysis of differences and patterns in enrichment and co-occurrence of pathways in two kinds of cells of NSCLC. (A) Barplot is showing correlation between Wnt pathway with stemness gene set in TS and Adh cell. Other barplot is showing concurrence of WNT/beta catenin and TGF-beta pathway in two different lung cancer cells TS and Adh. (B) Barplot show correlation between Glycolysis Gluconeogenesis pathway and sonic hedgehog pathway (SHH) pathway in the two lung cancer cells TS and Adh cells.

According to our analysis, among non-metabolite gene sets, the sonic hedgehog (SHH) pathway had higher differential co-occurrence with the glycolysis gene set. The correlation values between SHH and glycolysis pathway in TS and Adh cells were 0.63 and -0.138, respectively (Figure 3.8B). The SHH pathway is known to promote glycolysis in a variety of cancers [66]. The SHH pathway appears to cluster with cell-cycle associated gene-sets as revealed through hierarchical clustering (Figure 3.6,3.7), implying that it is implicated in controlling proliferation in Adh cells. Our findings show that its function is context-dependent and may play a more dominant role in Adh-like NSCLC cells than in TS cells. Thus, UniPath will aid researchers in conducting studies in a context-dependent manner in cancer.

## 3.5 Discussion

Usage of pathways and gene-set enrichment to exploit single-cell heterogeneity can lead to a slew of new applications. Covariates such as cell cycle or tissue microenvironment can influence gene expression slightly, which could hamper the downstream analysis using raw gene expression profiles directly. However, with the utilization of pathway scores, covariance tend to have more weightage due to the grouping of genes that tend to ward off such covariate effects. Thus, UniPath based approach helps subside such covariate effects in clustering and pseudotemporal ordering. The pathway score-based clustering of the MCA dataset unveiled some clusters where one of the member cell types could be differentiated into another. For instance, one of the clusters harbored Afp+ placental endodermal cells and fetal liver hepatocytes. Our findings suggest that UniPath could assist biologists in finding convergence and applicability of conversion between different cell types.

We have demonstrated some results achieved by employing UniPath pathway scores that cannot be attained using gene expression profiles directly. 1) correct pseudotemporal ordering of differentiating hESC cells towards endoderm. 2) grouping of cells from different tissues. Further, pathway scores can help in deciphering underlying cellular heterogeneity. For instance, in the Chu et al. dataset, we observed a bimodal distribution of some pathway scores at a time point of 36 h. Notably, such bimodal distribution in the case of pathway scores could furnish valuable insights into regulatory and functional states implicated in bifurcation and cell fate decisions. Using UniPath for pathway score calculation opens up new windows for novel ways of analysis, which cannot be achieved through the usage of gene expression profiles directly. These analyses include computing pathway co-occurrences, detecting co-regulated pathways clusters, and estimating differential co-occurrence for pathway pairs to differentiate between two groups of cells. The co-occurrence pattern of the Nodal pathway with other pathways, including SMAD2, Wnt/beta-catenin, was revealed using UniPath based interrogation of dataset involving hESC differentiation towards DE. These findings align with previous findings. Our approach of using the null model to calculate the significance of the co-occurrence in pathway pair enables us to predict alterations in the co-enrichment pattern across different groups of cells. This kind of analysis could be beneficial in linking dysregulation of pathways to disease-associated genesets in target cells. Uni-

Path provides new dimensions to exploit cellular heterogeneity for several downstream applications.

# CHAPTER 4

## Inference of drug response sensitivity in cancer by leveraging gene expression data in pathway space

### 4.1 Introduction

Cancer is a multifaceted disease driven by a high degree of genetic and phenotypic heterogeneity. Even though cancer management through chemotherapy, immunotherapy, and targeted therapies has considerably enhanced the clinical efficacy over the past few years, some individuals show partial or no response. The inherent heterogeneity translating into differential drug responses of patients and complex tumor microenvironment are significant roadblocks in precise modeling and predicting drug response in individual patients [67] [60]. A one-size-fits-all approach for cancer treatment is obsolete since every patient responds differently to drug therapy. Unfortunately, problems become profound for the cancer types lacking appropriate genetic targets of anticancer drugs, for instance, EGFR and KRAS mutations or BCR-ABL fusions. However, drug targets or status of mutation alone are typically abysmal therapeutic indicators for targeted therapies [1] [135]. Further, using a targeted therapy without taking into account drug resistance may deteriorate the patient's condition. Consequently, early conjecturing of drug response, determined from pretreatment of cancer's molecular profiles, has become a prerequisite to guide personalized treatment regimens [60] [187].

In the last few years, the availability of large-scale pharmacogenomics databases aided in strengthening the precision oncology domain and our knowledge of drug response [60]. Cancer Cell Line Encyclopedia (CCLE) [18], Genomics of Drug Sensitivity in Cancer (GDSC) [221] and Cancer Therapeutics Response Portal v2 (CTRPv2) [177] are notable databases among these. These projects manifest a vast corpus of information that includes high throughput screening experiments encompassing hundreds of anticancer drugs screened on more than 1000 cell lines. The Cancer Genome Atlas (TCGA) [33] archives consist of RNA-seq data of mostly primary tumors across multiple cancer types furnishing another source of information to study pharmacogenomics



and unmask patient-drug response. This burgeoning amount of data has enabled the development of numerous machine learning predictive models for drug response prediction in cancer. Some of these methods include a deep variational autoencoder that is involved in the imputation of drug response by compressing multiple genes into latent vectors and these latent vectors in low dimensional latent space are used for training prediction models [102]. The kernelized bayesian matrix factorization approach for modeling drug response through conjecturing known pathway-drug associations was proposed by Ammad-Ud-Din, Muhammad, et al. [7]. Another method involved using matrix factorization with similarity regulation that incorporated drug and cell line similarity space for improving the prediction of anticancer in cell lines [201]. CDRscan is a deep learning neural net framework that utilizes cancer signatures for predicting drug response in cancer, was proposed by Chang and colleagues [35]. Sakellaropoulos, Theodore, et al. reported a deep neural network that leverages gene expression profiles for predicting drug response and outperforms state-of-the-art methods—ElasticNet and Random Forest [170].

Two significant scopes of improvement were identified by carefully inspecting the aforementioned methods. First, most past studies do not take into account the structural properties of drugs as features or explanatory variables for modeling drug response. Consequently, the machine learning models learn inefficiently and are incapable of making predictions on new drugs that aren't included in the training dataset. Second, gene expression values are assumed to be separate entities, overlooking the combinatorial effects of pathways. Past works have illustrated the usage of pathway activity scores for many analyses instead of gene expression profiles [81] [36]. Notably, our previous works demonstrated how pathway activity scores enable a reasonably better illustration of biological processes [36][125][21]. As an additional advantage, integrating data using pathway activity scores abates batch effects. While single-cell single-cell RNA-seq (scRNA-seq) facilitates unbiased delineation of cellular heterogeneity, there are minimal efforts to utilize this precise molecular information for drug response prediction at the sub-clonal level. This is mainly because most of the training datasets available are bulk gene expression profiles. Testing scRNA-seq profiles on models trained on bulk RNA-seq profiles is expected to result in inaccurate predictions. Pathway transformations of scRNA-seq and bulk RNA-seq gene expression profiles reasonably mitigate this issue. The only notable work in this context is by Suphailai, Chayaporn, et al.,

[188] which has primarily focused on scRNA-seq profiles of head and neck cancer for drug response prediction, ignoring the benefits of descriptors of molecular compounds to generalize the prediction models.

In this study, we developed a deep neural network based approach called Precily to model drug response in both *in vitro* and *in vivo* setups. We used open-source pharmacogenomics databases: CCLE, GDSC and CTRPv2 entailing multiple drug-cancer cell line combinations, and patient profiles from the TCGA database. We evaluated our CCLE cell line trained models model on various bulk and scRNA-seq profiles.

After being convinced by the plausibility and the cell line model's overall performance, we investigated the outcome of drug response predictions on our internally generated prostate cancer (PCa) cell lines and xenograft models exposed to different treatments. Even though PCa is the most frequently diagnosed malignancy in men, treatment options are limited for advanced-stage malignancy. Androgen deprivation therapy (ADT) is frequently used as an effective treatment strategy in clinical settings. It takes advantage of the reliance of PCa on androgen receptor (AR) signaling for tumor growth and progression. ADT is beneficial in the majority of patients. However, the effect is transient, and ultimately cancer cells exhibit resistant phenotypes with the appearance of metastatic castration-resistant prostate cancer (CRPC). Only a few anti-cancer therapies are efficient and clinically approved for the CRPC treatment, but their vested survival advantage is limited. Therefore, selecting appropriate drugs and combinations is critical in the dynamically developing landscape of cancer to gain maximal survival for the patients [191][108][190]. We used our in-house bulk RNA-seq gene expression profiles of baseline PCa cell lines exposed to various treatment conditions to further validate our cell line model. To verify the cross-sample applicability of the model, we interrogated our LNCaP cell line-derived xenograft data portraying *in vitro* treatments. We used LNCaP xenografts derived from a PCa tumor progression study in which tumors were collected at multiple phases, including precastration (PRE-CX), post castration (POST-CX), castration-resistant prostate cancer (CRPC), and while on treatment with an androgen inhibitor enzalutamide (ENZ) during the progression of the tumor. Our findings unveiled biologically and clinically meaningful relationships of drugs and pathways in the context of resistance and sensitivity. We also assessed the ability of Precily to predict responses to the drugs that the training models had never seen. To this end, we considered two drugs, metformin and orlistat, used for treating

type 2 diabetes [144] and obesity [16]. But these two drugs have also been discovered to hold therapeutic potential in PCa. Finally, we evaluated the model's efficiency, trained on patient RNA-seq profiles from TCGA compendium on RNA-seq melanoma cancer patient profiles of before treatment and post relapse matched samples. Our study links systematized prediction of drug response with multiple in vivo and in vitro evaluations encompassing cell lines, xenografts, and patients, which is vital for the clinical translation and implementation of such approaches.

## **4.2 Methodology**

### **4.2.1 PCa cell lines and culture**

The human PCa cell lines LNCaP, VCaP, DuCaP, DU145, and PC3 were maintained in Phenol-red free RPMI medium and fetal bovine serum (5%) supplement in a humidified atmosphere with the temperature of 37°C and 5% CO<sub>2</sub>. During the exponential growth phase, the cells were harvested for RNA extraction.

### **4.2.2 In vitro experiments**

The PCa cell line LNCaP (#CRL-1740™ clone FGC) was bought from the American Type Culture Collection (ATCC). The LNCaP cells were seeded into a growth media containing RPMI media without Phenol-red augmented with fetal bovine serum (5%) and cultured in a humidified atmosphere with a temperature of 37°C and 5% CO<sub>2</sub> for 72 hr. Thereafter cells were incubated in an androgen-depleted environment in medium + charcoal-stripped serum (5%) for 48 hours. Then in the presence or absence of dihydrotestosterone (DHT, dissolved in vehicle(EtOH)) (10 nM), LNCaP cells were treated with androgen inhibitors: enzalutamide (10uM), apalutamide (10uM), and bicalutamide (10uM) for 48 hrs.

### **4.2.3 In vivo studies**

In the in vivo study, LNCaP xenografts were developed by injecting 1e6 LNCaP cells subcutaneously into the flank area of NOD-SCID male mice. When the tumor size

reached 200 square mm, mice underwent mock surgery (mice are anaesthetised, and incision is made and skin is stapled together again but without removing testes) or were surgically castrated for the PRE-CX group. When tumor became 1000 square mm, tumors were harvested from the PRE-CX group. One week post castration, when serum PSA (Prostate-Specific Antigen) reached its nadir, tumors were harvested from the POST-CX (post castration) group. Tumors were harvested from the CRPC group when the size of the tumor became 1000 square mm after castration. Daily treatment with ENZ (10 mg/kg) began as serum PSA rose post castration for the ENZ groups. Tumors were collected either when PSA had reached the nadir i.e., Enzalutamide sensitivity while on treatment with enzalutamide or when the tumor had grown to 1000 square mm even with treatment with enzalutamide.

#### **4.2.4 RNA isolation and library preparation and bulk RNA sequencing**

The Norgen RNA Purification PLUS kit #48400 from Norgen Biotek Corp., Thorold, Canada was used to extract total cellular RNA for mRNAseq according to the instructions of the manufacturer, including DNase treatment. An Agilent 2100 Bioanalyzer and a Qubit®. 2.0 Fluorometer was used to ascertain the quality and quantity of RNA (Thermo Fisher Scientific Inc, Waltham, USA). RNA-seq library construction and sequencing were performed utilizing the Illumina TruSeq Stranded mRNA Sample Prep Kit (strand-specific, polyA enriched, Illumina, San Diego, USA) using an input of 500 ng - 1 ug total RNA and RIN>8. Then paired-end sequencing was performed with a 100-150 bp read length, and per sample, around 30-60 M read pairs were produced.

The raw RNA-seq data were processed using an in-house pipeline. The quality of raw reads was evaluated using FastQC tool [9] and trimmed through TrimGalore [109]. Then aligning of reads to GRCh38 / hg38 reference human genome and Ensembl.v.99 (Gencode version 33, Jan-2020) transcriptome was performed using STAR aligner [50]. Reads were quantified using RSEM software. In the case of xenograft samples, mouse: GRCm38 / mm10, Gencode.v.M24 / Ensembl version 99, Jan-2020 chimeric human+mouse reference was used for STAR alignment. Quantification of reads was performed using RSEM. RSEM generated TPM values were used for GSVA scoring.

#### 4.2.5 Cancer cell lines RNA-seq data

To predict and interrogate drug response measured as LN IC50 i.e. half-maximal inhibitory concentration, we utilized RSEM (RNA-Seq by Expectation-Maximization) software quantified bulk RNA-seq TPM (transcript per million) normalized data of cell lines (n=1019) from public project Cancer Cell Line Encyclopedia (CCLE). The related drug response data for the cell lines were acquired from another database Genomics of Drug Sensitivity in Cancer (GDSC). We have used the GDSC2 dataset from this database [220]. The GDSC2 dataset consisted of some cell line drug combinations with multiple LN IC50 values. In such scenarios, to avoid ambiguity, we took the mean of LN IC50 values. For training models, we used bulk RNA-seq dataset of CCLE cell lines (n=550) intersecting with the GDSC2 dataset cell lines. This matrix constituted 57820 Ensembl Gene IDs and these IDs were transformed into their official gene symbols using gencode.v19.genes.v7\_model.patched\_contigs GTF annotation file. As a result, multiple Ensembl gene IDs corresponded to the same gene ids. In such cases, we have averaged out the gene expression values. At this point, our gene expression matrix contained 54301 genes and 550 cell lines. This matrix was log2 transformed after the addition of a pseudo count of 1.

#### 4.2.6 Gene expression profiles of patients

Like cell lines, we used TCGA tumor mRNA sequencing data to model drug response as responder and non-responder. We downloaded TCGA RNA-seq data encompassing 33 tumor types from the Broad GDAC firehose database [24]. For our study, we have employed gene-level Illumina HiSeq RNA-seq v2 data obtained using RSEM software [122]. The NCI Genomic Data Commons portal [73] was used to source patient clinical drug response information. Drug names and corresponding patient response information were fetched from clinical metadata files and manually corrected to eliminate typing, misprint, and spelling mistakes to make drug names uniform across clinical metadata files. Patients who had a complete or partial response were defined as responders, whereas those who had clinically progressing or stable disease were classified as non-responders. We retained gene expression profiles for those cancer types which had drug response information for at least two patients. We were left with RNA-seq profiles of 29 cancer types at this stage. Next, for each cancer type, scaled estimate values from

RSEM files were converted into TPM by scaling with one million. Then TPM files were then converted into log<sub>2</sub> scale and 1 was added as pseudo count. Some patients have similar barcodes in the TCGA datasets. We have averaged out gene expression in such cases for further analysis.

#### **4.2.7 Molecular drug descriptor data**

We sourced information on the drug response of 192 compounds corresponding to 550 CCLE cell lines from the GDSC2 dataset. Additionally, we also obtained clinical drug response data for 215 drugs corresponding to 1517 patient samples in TCGA. The structural information of these drugs was obtained as a simplified molecular-input line-entry system (SMILES) using the Python PubChemPy package [189]. But SMILES notations were not known for all the compounds. Consequently, we were left with SMILES of 173 drugs for 550 cell lines, and for 1443 unique TCGA patients, we had 139 compounds. Then, the SMILEVec package was utilized to transform SMILES into numerical vector embeddings by leveraging embeddings data trained on Pubchem and embeddings of length 100 [153].

#### **4.2.8 Gene expression to pathway activity scores**

We used pathway enrichment scores to train models. We utilized the open-source Gene Set Variation Analysis (GSVA) [82] R package to estimate pathway activity scores. We supplied two inputs to GSVA: log<sub>2</sub> transformed TPM normalized matrix of gene expression and gene set file. We employed Cp.v.6.1 collection of canonical pathways entailing 1329 pathways from the Molecular Signatures Database (MSigDB) [129] and set the minimum gene set size (min.sz) as 5 for running GSVA. Then we combined the pathway activity score matrix with the numerical drug embeddings. The final processed CCLE cell line training data comprised 80056 drug-cell line pairs in rows and columns containing a total of 1429 variables involving 1329 pathway vectors and drug features in the form of vector embeddings. The vector embeddings for each drug corresponds to size 100. These pathway and drug features represent the explanatory variable and the response variable is LN IC<sub>50</sub>. In the case of TCGA data, bulk RNA-seq gene expression profiles of 29 cancer types were converted into pathway scores independent of

each other. Then we merged the GSVA scores of those samples for every cancer type having drug response data based on the common pathways. Our final TCGA training dataset contained 3108 drug-patient pairs with 1427 features entailing pathway features, numerical drug descriptors, and response variables as responder and non-responder labeled as 1 and 0, respectively.

#### 4.2.9 Model training using CCLE RNA-seq cell line dataset

We employed commonly used machine learning approaches for drug response prediction at the genes and pathways level. We devised drug response prediction as a regression task. We split the CCLE modeling data into 90% training and 10% test sets, ensuring no cell lines overlap. For hyperparameter tuning, we used k-fold cross-validation and partitioned the training dataset into five non-overlapping folds. The Random forest was trained using the ranger R package [210]. We conducted a grid search on each fold of the training set. For every fold, we used different values of mtry parameter (1,2,3,4,5,6,7,8,9,10), and for the number of trees, values varied from 100 to 1000 with a step size of 100. We chose the five best models with the lowest Mean Squared Error (MSE) for each training data subset. Finally, using parsnip R package [111] we trained 5 models on the complete training data based on pre-trained hyper-tuned models. ElasticNet was run using caret and glmnet R packages [83][110]. The caret runs bootstrapping 25 times for every training fold by default to identify the best model based on the minimum value of Root-Mean-Square Error (RMSE). The five optimal models obtained are used for training the entire training dataset.

We used the python-based Keras platform to develop a deep neural network (DNN). We modeled DNN using the RELU activation function and one input layer of the size of a number of features (pathway scores and drug embeddings) present in the training dataset. This input layer was followed by one hidden layer of size 512. The first two layers were kept fixed. The Keras Tuner library [149] was employed to find the best set of hyperparameters for the deep neural network. The Hypberband [126] approach with five-fold cross-validation on the training dataset was used to find optimal training parameters based on the validation loss. The different tuning hyperparameters used are as follows: the number of layers was kept between 2 and 6, and the number of neurons in the layers was defined as 128 (minimum) and 256 (maximum), with a step size of



4. The 30 epochs were used. We used drop-out layers between the layers whose values varied from 0.1 to 0.5 with a step size of 0.1. The ADAM optimizer with different learning rates: 1e-3, 1e-4 and 1e-5 and Mean Squared Error (MSE) as the loss function was used. Then, we trained five models on the complete training dataset based on the pre-trained hyper-tuned fold-specific models using a batch size of 128 and 50 epochs. We used these models to assess the performance of multiple independent test datasets.

For models considering genes as features, we employed the same DNN architecture as employed for the pathway-based model, except for the size of the input layer. In gene based model size of the input layer was kept at 600 entailing 500 top highly variable genes and drug embeddings of size 100. The same strategies were used in the case of ranger and ElasticNet as for pathway based modeling.

#### **4.2.10 Benchmarking Precily**

Precily was benchmarked against two previously published methods namely CaDRReS-Sc [188] and another method by Sakellaropoulos, Theodore, et al. [170]. We applied CaDRReS-Sc with default parameter settings to CCLE bulk RNA-seq profiles for which drug response information was sourced from the GDSC database. In the case of Sakellaropoulos, Theodore, et al method, we trained drugs specific models using default parameters except for the varcut parameter which was set to 10.

#### **4.2.11 CTRPv2 data Processing**

CTRPv2 database features a collection of small molecule probes, drugs and cancer therapeutics screened against well-known cancer cell lines. We retrieved SMILES for 377 compounds in CTRPv2, and cell line-specific IC50 values for these drugs were sourced from the PharmacGx R package. We eliminated IC50 outliers using the interquartile range (IQR) rule. At this point, our matrix contained 153899 drug-cell line combinations and 1429 features (pathway scores and drug embeddings). Precily, performance was evaluated in a similar manner as done for CCLE analysis.



#### **4.2.12 Model using TCGA RNA-seq patient data**

We used the R H2O AutoML [118] framework to train models on the TCGA dataset for predicting drug response as responder and non-responder, thus formulating a classification task. We split TCGA data into 90% train set and 10% test set. The 90% training dataset was used as an input to the `h2o.automl()` function. We have used five-fold cross-validation such that there is no overlap of patients among folds and `max_models` option as 20. As a result, various machine learning models ( $n=34$ ), such as Deep learning, DRF, XGboost, GBM, GLM, XRT, and stacked ensemble models, were trained automatically. The reason for the training of more models was to reach convergence. Due to scarcity of data, deep learning model performance was sub optimal. Also, we used Precily architecture with appropriate changes to make it suitable for classification. The sigmoid activation function was used in the last layer. We used binary cross-entropy as a loss function. We used the same strategy for data splitting as used for AutoML.

#### **4.2.13 Survival analysis**

After pre-processing, the TCGA dataset contained 3108 drug-patient pairs. We performed survival analysis on a TCGA test dataset (20%) comprising 293 drug/patient combinations. We used the median value of the predicted response probability to stratify the patients and compute survival.

#### **4.2.14 Performance measures**

In the regression task, we have used the coefficient of determination ( $R^2$ ) computed using the `caret` R package and Pearson correlation ( $\rho$ ) to measure the performance of our model. While in the case of TCGA dataset, several metrics were used: 1. AUC 2. AUC-PR 3. F1 score

#### **4.2.15 Imputation**

Missing pathway features in the test input dataset are imputed using `impute` package from R by employing the nearest neighbor-based averaging approach.

#### **4.2.16 Validation of model trained on cancer cell lines using scRNA-seq profiles of cancer cell lines**

The scRNA-seq raw UMI count dataset of well-established cancer cell lines consisting of genes (n=30314) in rows and cells (n=56982) in columns spanning multiple cell lines (n=207) was acquired from a previously published study by Kinker, G. S., and colleagues [107]. The UMI count matrix was subjected to quality check using data\_processing.R R script, from the Kinker, G. S. et al. study. At this point, our count matrix consisted of 53299 cells encompassing 198 cell lines. This matrix was converted into TPM by multiplying with 1e6 and dividing the UMI count of the gene by the sum of the UMI count of the cell. The UMI counts are not affected by gene length bias [156]. Then the TPM matrix was subjected to log2 transformation and 1 as a pseudo count. The gene expression levels of the same cell line were averaged, and this matrix was transformed into pathways using GSVA. For running GSVA, we have used default parameters. Our final validation dataset comprised drug-cell line pairs (n=17279) in rows involving 116 cell lines intersecting with the GDSC2 dataset cell lines and corresponding to 173 drugs.

#### **4.2.17 Predicting paclitaxel response in single MDA-MB-231 cells using model trained on cell lines**

To obtain Lee et al. dataset, sequencing short reads data from study id SRP040309 were downloaded using prefetch from SRA Toolkit. SRA files were converted to FASTQ files using fasterq-dump. Then we utilized STAR aligner [50] to align the FASTQ files with GRCh37/h19 reference genome and Ensembl GRCh37 GTF file, release 75 [88]. To quantify the gene expression, we employed HTSeq-count [8]. The counts obtained from HTSeq were converted into TPM by dividing the counts by gene length, then scaling with one million and dividing by sum total of the counts in the cell. We retained only those genes for which length was available from the R package EDASeq [166]. Then using the human GRCh37 (v75) GTF annotation file, we converted Ensembl gene IDs to their official gene symbols. We included those genes having a TPM value  $\geq 1$  in  $\geq 10\%$  of samples. Then, this matrix was log2 transformed with 1 as pseudo count. The gene expression values of five samples each of untreated cells, stressed cells, and paclitaxel

sensitive cell population was averaged out and transformed into pathway scores using GSVA. Then we paclitaxel drug response was predicted for treatment-naive cells and a population of cells that became more sensitive to paclitaxel.

#### **4.2.18 Evaluation of model trained on TCGA patient tumor data using bulk RNA-seq profiles of melanoma patients**

We used previously published bulk gene expression profiles of melanoma patients having BRAF mutation for the drug response prediction task. The dataset comprised of Reads Per Kilobase of transcript per million mapped reads (RPKM) RNA-seq gene expression for six patients. The profiling is performed prior to treatment and post-treatment with dabrafenib or dabrafenib and trametinib during the progression of the disease [199]. This dataset involves clinical response information as well. We have converted RPKM RNA-seq data into TPM data by dividing each RPKM value by the sum of all RPKM values for all genes in the sample or cell and multiplying with a factor of  $1e6$ . The transformed TPM normalized matrix was subjected to  $\log_2$  transformed with a pseudo count of 1 and transformed into pathway scores utilizing GSVA. We predicted response for the first line of drug for melanoma i.e., dabrafenib and trametinib in 3 patients. The details on mechanisms to acquired resistance to dabrafenib and trametinib is available from the original study for these 3 patients.

#### **4.2.19 Data and code availability**

<https://github.com/SmritiChawla/Precily>

### **4.3 Results**

#### **4.3.1 Overview of workflow**

In this work, we introduce Precily, a deep neural network based approach that leverages gene expression data to model drug response in both *in vitro* and *in vivo* setups. We employed bulk RNA-seq gene expression profiles of cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) database. First, we transformed bulk RNA-seq data of

cancer cell lines from the CCLE database into pathway activity scores using R package GSVA [82]. Second, we combined numerical descriptors of the drugs, i.e., simplified molecular-input line-entry system (SMILES) vector embeddings with pathway scores of CCLE cell lines. The Genomics of Drug Sensitivity in Cancer (GDSC) database was used to acquire drug response data for CCLE cell lines. Notably, 550 cell lines were overlapping in the CCLE and GDSC databases that were tested for 173 unique drugs for which SMILES chemical annotations were available. SMILES for 173 drugs were retrieved from the PubChemPy [189] python tool and transformed into embeddings using the SMILESVec python tool [153]. The final training data comprised 80056 cell line-drug pairs in rows and 1429 features containing 1329 pathways and molecular descriptors of size 100 for each drug in columns for a regression task. These 1429 features constituted the explanatory variable set for predicting drug response as acquired from the drug screening experimental datasets (Figure 4.1A). We utilized the Keras platform for constructing a suitable DNN architecture (Figure 4.1B).

We used cross-validation best practices to build models and reported results on an independent test dataset. We recognize that random splitting of data (cell line/drug combinations) into train, validation and test set causes data leak problems and do not correspond to practical applications. In such scenarios, training data becomes aware of gene expression profiles of cell lines along with the sensitivity to some drugs, thus making it relatively simple to predict sensitivity to a new drug. Therefore, we perform a cell line-wise split of the dataset such that there is no overlap of cell lines in train, validation and test datasets. We compared Precily with two other methods: Cancer Drug Response prediction using a Recommender System for single-cell RNA-seq (CaDRReS-Sc) [188] and another approach by Sakellaropoulos, Theodore, et al [170]. Both of these methods predict drug response by leveraging gene expression profiles. We compared the performance of Precily with state-of-the-art machine learning methods- ElasticNet and random forest (RF). Past studies have utilized these methods to predict drug response [165][93][49]. Further, as a baseline, we also assess the performance of the three ML models (Precily, RF and ElasticNet) using gene expression profiles instead of pathway scores. We used highly variable 500 genes selected using the squared coefficient of variation ( $CV^2$ ) approach for training models. We noted the highest correlation between Precily predictions and observed LN IC50 values on the independent test dataset. CaDRReS-Sc closely followed the Precily results. The distribution of Pearson's cor-

relation coefficients obtained for predicted and observed LN IC50 values across drugs for considered methods is shown in Figure 4.2A. One of the methods we have used for comparing Precily uses the H2O framework to train drug-specific models. This is the reason we reported correlations at the level of drugs. However, this approach is not optimal as it does not consider the structural features of drugs for predictions. To get a more comprehensive picture, we pooled predictions for cell line-drug pairs and obtained a coefficient of determination ( $R^2$ ) of 0.77 and Pearson's correlation coefficient of 0.88 with a statistical significance of  $p < 2.2e-16$  (Figure 4.2B)

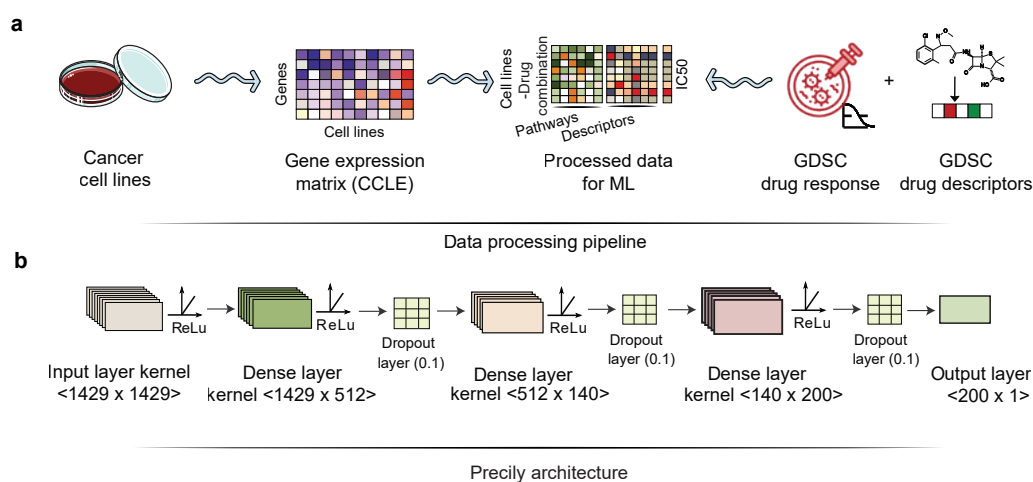


Figure 4.1: Overview of the Precily approach (A) Schematics of the processing of training data. The initial step involves the training dataset processing. The RSEM quantified bulk TPM normalized genes expression data of CCLE cancer cell lines were converted into pathway scores using GSVA. The pathway score matrix was combined with numerical drug descriptors of each molecular compound. This constituted our training data. Parts of the figure were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (<https://creativecommons.org/licenses/by/3.0/>). (B) Architecture of DNN model. The second step involved training of machine learning model on this dataset which contained pathway scores and drug features. The pathway score vector and drug features forms explanatory variable and LN IC50 obtained from the GDSC database is a response variable. A deep neural network (DNN) from the Keras platform performed a regression task to predict drug response.

GDSC database primarily features anti-cancer drugs, on the other hand, the Cancer Therapeutics Response Portal v2 (CTRPv2) presents a collection of small molecule probes, drugs and cancer therapeutics. We performed a similar analysis on CCLE/CTRPv2 data as CCLE/GDSC and obtained a coefficient of determination ( $R^2$ ) of 0.70 and Pear-

son's correlation coefficient of 0.84 with a statistical significance of  $p < 2.2e-16$  (Figure 4.2C). Our analyses point to reasonably accurate and reproducible susceptibilities to anti-cancer drugs in cancer cell lines.

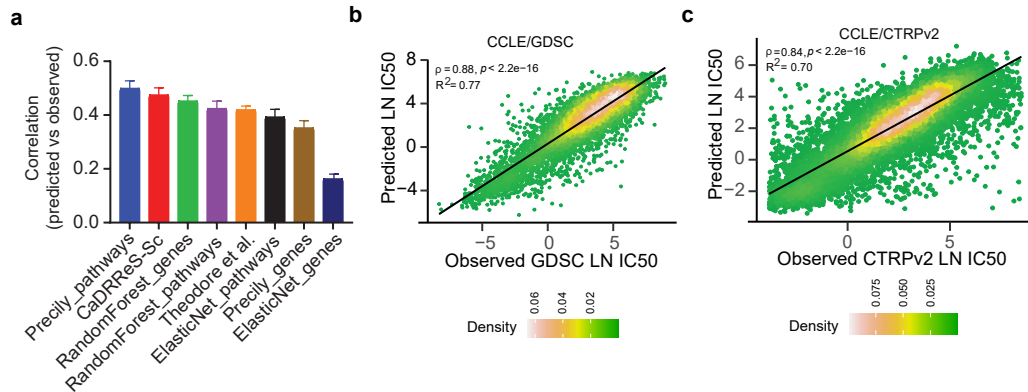


Figure 4.2: Evaluation of Precily **(A)** Comparison of different methods for drug response prediction. Barplots illustrating distribution of Pearson's correlation for observed vs predicted LN IC50 across individual drugs. **(B)** Scatter plot showing the efficiency of the Precily model on CCLE/GDSC held out dataset measured in terms of Pearson correlation between observed and predicted LN IC50. **(C)** Scatter plot showing the efficiency of the Precily model on CCLE/CTRPv2 held out dataset measured in terms of Pearson correlation between observed and predicted LN IC50.

### 4.3.2 Prediction of drug response using single-cell expression profiles

Single-cell RNA sequencing (scRNA-seq) technologies have greatly improved our understanding of intra and inter-tumoral heterogeneity. While many clinical studies have adopted single-cell technologies as a method of choice, we have not utilized the full potential of this technology for predicting drug response at subclonal resolution while accounting for intra-tumor heterogeneity. Therefore, we evaluated the ability of Precily for drug response prediction in single-cell studies. We utilized two single-cell studies. First, we used the Kinker, G. S. et al [107] scRNA-seq profiles of 207 cell lines, of which 116 were common with CCLE data. We removed Kinker, G. S. et al. cell lines and retrained CCLE/GDSC model and used this model on Kinker, G. S. et al. dataset. We obtained a coefficient of determination ( $R^2$ ) of 0.73 and Pearson's correlation coefficient of 0.85 with a P-value of  $p < 2.2e-16$  (Figure 4.3A). Furthermore, we also assessed the model's performance on another publicly available single-cell transcriptome by Lee

et al. [119]. This dataset comprised the treatment-naive population of MDA-MB-231 cells and the cells that had developed a vulnerability to paclitaxel drug. In this study, metastatic breast cancer cells MDA-MB-231 were treated with paclitaxel drug. After five days of treatment, the majority of cells died. On the other hand, some residual cells cultured in a medium without a drug begin to proliferate and establish clones. Surprisingly, when re-exposed to paclitaxel drug, these residual cells became more susceptible to paclitaxel. To substantiate that these cells have more susceptibility to paclitaxel, we used the Precily on this dataset. We observed that these cells were predicted to be more sensitive to paclitaxel drug than the treatment-naive population, affirming the actual findings (Figure 4.3B).

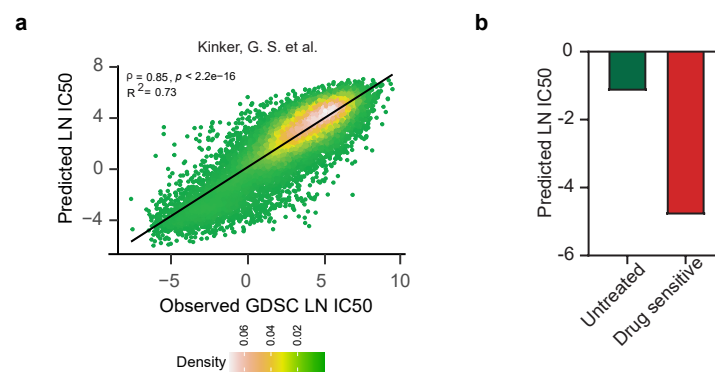


Figure 4.3: Evaluation of Precily on scRNA-seq profiles. (A) Scatter plot showing the efficiency of the Precily on scRNA-seq cancer cell lines Kinker, G. S. et al. dataset for observed and predicted LN IC50 estimated using Pearson correlation. (B) Barplots showing predicted paclitaxel drug response in single MDA-MB-231 untreated cells and cells susceptible to paclitaxel.

### 4.3.3 Verification of Precily in Prostate cancer cell lines

Despite significant therapeutic advancements in prostate cancer (PCa), current therapeutic choices are finite, and the development of resistance to treatment presents substantial hurdles to treatment selections [139] [146]. Consequently, it is crucial to make optimal drug choices to treat PCa. We independently verified our CCLE/GDSC model using our internal PCa datasets. The Precily model was applied to bulk RNA-seq gene expression profiles of five baseline PCa cell lines (LNCaP, DUCAp, VCAP, PC3 and DU145) having two biological replicates of each cell line. In each of these ten samples, we predicted drug response for 155 drugs. These 155 drugs are specific to PCa cell lines in the GDSC and target various cell signaling pathways. Overall, two DU145

and PC3 PCa cell lines, known to be Androgen Receptor (AR) negative, were predicted to be more invulnerable to these anticancer drugs by our model. On the other hand, AR-positive cells, namely LNCaP, DUCAP, and VCAP, were predicted to be sensitive (Figure 4.4A). Notably, the AR-positive LNCaP cells were predicted to be more susceptible to these drugs among these five cell lines (Figure 4.4B). In particular, LNCaP cells exhibited more sensitivity towards PI3K/mTOR pathway inhibitors such as AKT inhibitors, namely, ipatasertib, afuresertib, uprosertib and in particular AZD2014. Further, LNCaP cells displayed higher enrichment of GSVA scores for mTOR-associated signaling terms, pinpointing that this may play a role in the predicted sensitivity towards AKT inhibitors. We then compared the predicted LN IC<sub>50</sub> in the two biological repeats of LNCaP cells with the observed GDSC LN IC<sub>50</sub>. For the two biological repeats of LNCaP cell lines, we discovered a high correlation of 0.86 ( $p < 2.2e16$ ) (Figure 4.4C).



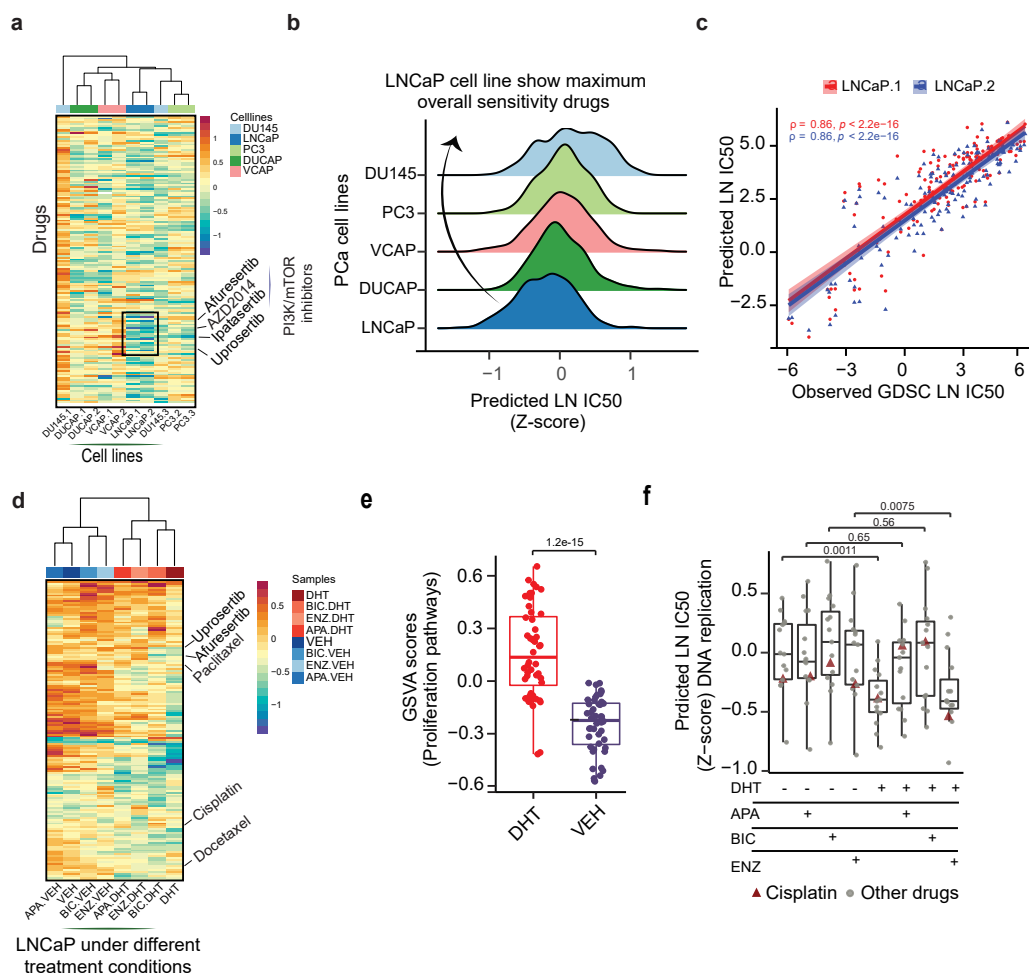


Figure 4.4: Analysis of prediction of drug response in PCa cell lines **(A)** Heatmap depicting the predicted LN IC50 (Z-score) across five untreated PCa cell lines for 155 drugs screened against PCa cell lines in the GDSC database pinpointing drugs targeting PI3K/mTOR signaling. A lower LN IC50 indicates the sample is predicted to be more sensitive for a particular drug. Color bars depict types of cell lines. **(B)** Ridgeplot depicting overall patterns of predicted LN IC50 (Z-score) across five untreated PCa cell lines. **(C)** Scatterplot displaying Pearson correlation ( $\rho$ ) for actual vs. predicted LN IC50 for the two biological repeats of the LNCaP cell line. The line color indicates two biological repeats of the LNCaP cell line. **(D)** Heatmap showing predicted LN IC50 (Z-score) across 155 drugs in LNCaP cells exposed to various treatment conditions, i.e., in the absence and presence of DHT and AR inhibitors—ENZ, BIC, and APA. **(E)** Boxplots showing enrichment of proliferation-associated pathways with and without the DHT (P-value is computing using Wilcoxon rank-sum). **(F)** Boxplots depicting predicted LN IC50 (Z-score) for DNA replication inhibitors specifically highlighting cisplatin drug across different treatment conditions (P-values estimated using Wilcoxon rank-sum test).

It is known that androgens play a role in the proliferation of prostate cancer cells [84][161]. We were further inquisitive in interrogating how predictions of drug re-

response change when LNCaP cells are exposed to androgen receptor (AR) agonist dihydrotestosterone (DHT) in comparison to vehicle control (VEH) in androgen depleted settings. Additionally, we wanted to investigate how exposure with Food and Drug Administration (FDA) approved AR antagonists, namely bicalutamide (BIC), enzalutamide (ENZ), and apalutamide (APA), under these settings influence drug response prediction patterns. This data comprised two biological repeats for each sample, and subsequently, we have utilized the mean of GSVA scores for biological repeats for further analysis. The overall trend indicates that cells cultured in the DHT settings appeared to be more susceptible to anticancer drugs as predicted by our model, on the other hand, cells exhibited resistance to these drugs in the absence of DHT (Figure 4.4D). In addition, we found that cells grown with DHT exhibited high GSVA scores for proliferation-related pathways. This substantiates the idea that actively proliferating cells are more vulnerable to particular anticancer drugs.

On the other hand, cells in a quiescent state appeared to be more resistant (Figure 4.4E). Notably, in the presence of DHT, treatment with androgen-targeted therapies or AR antagonists did not entirely invert the drug sensitivity observed with DHT treatment. However, Precily predicted cells to be sensitive to cisplatin even in the presence of AR antagonists, ENZ (Figure 4.4F). These results suggest using Precily in zeroing on new combinatorial therapies. Next, we demonstrated the ability of Precily to predict responses for the drugs that training data had never seen. For this, we looked into two drugs, metformin, used for treating type 2 diabetes and orlistat used for treating obesity. Lately, many studies have suggested that these two drugs might be effective in treating some cancer types. Notably, Precily based predictions and experimental IC50 values showed concordance at a relative scale (Figure 4.5A; Table 1).

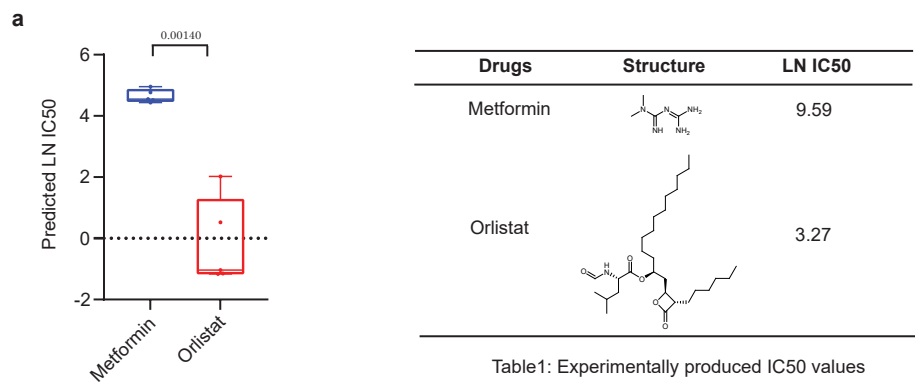


Figure 4.5: Evaluation of Precily on unseen drugs. **(A)** Boxplot demonstrating distribution of predicted drug response for two drugs namely metformin and orlistat across five pre trained models. As expected we Precily based predictions were concordant with the experimental IC50 values at a relative scale. P-value was estimated using t-test. Table 1 shows structure of these two drug along with experimentally determined IC50 values.

#### 4.3.4 Analysis of Precily based predictions in xenografts

Xenografts have emanated as powerful models in the clinical diagnostics domain for predicting anticancer drug response and assessing its clinical relevance. As such, we evaluated our potential to predict drug response in LNCaP derived xenografts. We used our internally generated LNCaP xenografts datasets. The LNCaP xenografts were established from a well-annotated large PCa progression study interrogating responsiveness and the eventual emergence of resistance to AR targeting therapies (see Methods). In male mice (PRE-CX), the establishment and initial growth of LNCaP xenograft tumors rely on androgens. Castration results in suppression of the activity of AR and tumor growth (POST-CX), but this susceptibility to castration is invariably followed by castration resistance (CRPC). Furthermore, EZN treatment of CRPC did confer initial therapeutic response, i.e., ENZ Sensitive (ENZS). However, resistance arises with due course of time, i.e., ENZ Resistance (ENZR) (Figure 4.6).

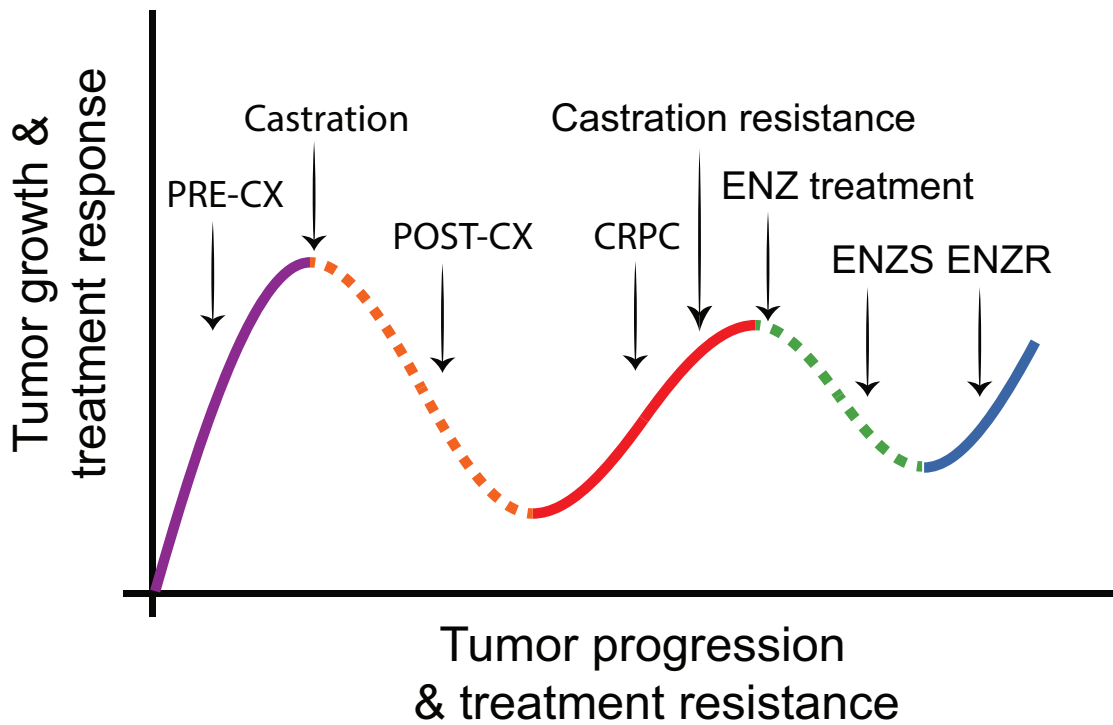


Figure 4.6: The experimental design for the LNCaP xenograft derived from large prostate cancer progression study illustrating treatments, drug responsiveness, and resistance stages. Tumor growth and treatment resistance are represented by solid lines, whereas treatment responsiveness is represented via dotted lines.

We predicted drug response for every 54 samples using Precily across this range of sequential therapeutics responsive and resistant states. Uniform Manifold Approximation and Projection (UMAP) based visualization of predicted response for the 155 drugs in LNCaP xenograft tumors revealed three distinct clusters (Figure 4.7A). We observed that Cluster 1 harbored the most resistant tumors, linked with their lower proliferation index. Cluster 1 samples were almost exclusively Enzalutamide-treated tumors, most of them belonging to the Enzalutamide sensitive/responsive (ENZS) group. It consisted of one CRPC sample, 10 ENZR out of 15, and 12 ENZS samples. On the other hand, samples in cluster 3 exhibited the highest overall predicted sensitivity towards 155 drugs, which might be connected to their high proliferation index, as pinpointed by the high GSVA scores for cell proliferation-related pathways (Figure 4.7B, C).

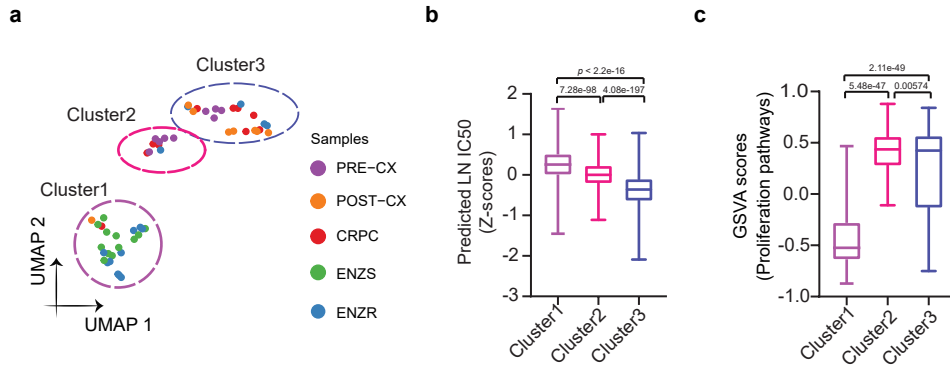


Figure 4.7: Analysis of prediction of drug response in LNCaP xenografts across three clusters (A) UMAP based visualization of predicted drug response as LN IC50 revealed three distinct clusters. We used principal component analysis (PCA) on predictions then used first 10 principle components for as an input for UMAP. (B) Box plots showing the changes in distribution of predicted LN IC50 (Z-score) in 3 clusters (P-values were estimated using Wilcoxon rank-sum test). (C) Boxplots showing the patterns of enrichment of GSVA scores of proliferation-associated pathways in each of 3 clusters (P-values were estimated by Wilcoxon rank-sum test).

ENZR tumors were present in all three clusters, implying that multiple underlying mechanisms mediate resistance and that more in-depth analysis may reveal differential vulnerability to specific drugs in ENZR. The multimodal distribution of drug response predictions in ENZR tumors, instead of the uniform distribution in ENZS tumors, bolstered the hypothesis of various ENZ resistance mechanisms (Figure 4.8A).

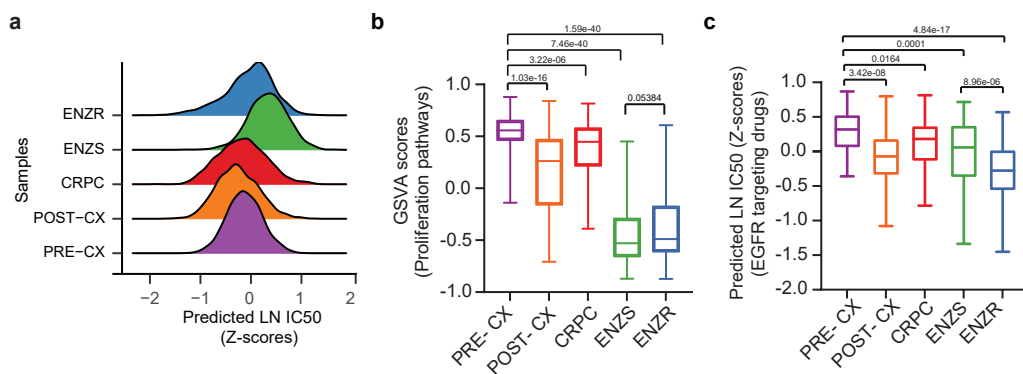


Figure 4.8: Analysis of prediction of drug response in LNCaP derived xenografts under treatment conditions (A) Ridgeplot depicting the overall pattern of predicted LN IC50 (Z-score) of 155 drugs across different types of tumor. (B) Boxplots depicting the distribution of GSVA scores of proliferation-associated pathways across different types of tumor (P-values were estimated using Wilcoxon rank-sum test). (C) Box plot displaying predicted LN IC50 (Z-score) of EGFR pathway inhibitors (P-values were estimated using Wilcoxon rank-sum test).

ENZR samples, in comparison to ENZS, were predicted to acquire some degree of susceptibility to a few drugs (Figure 4.7A). The scores of proliferation-related pathways as obtained through GSVA were higher in ENZR samples than in ENZS samples, however, statistical significance was not achieved (Figure 4.7B). Additionally, we noted ENZR tumors appear to be more responsive to EGFR inhibitors in comparison to other tumor types in this study (Figure 4.7C), with Sapatinib showing the greatest effect. While we attained encouraging results for the drugs constituting our training dataset, we could also predict informative and biologically meaningful responses for the drugs that were not part of our training dataset— APA, BIC, and ENZ.

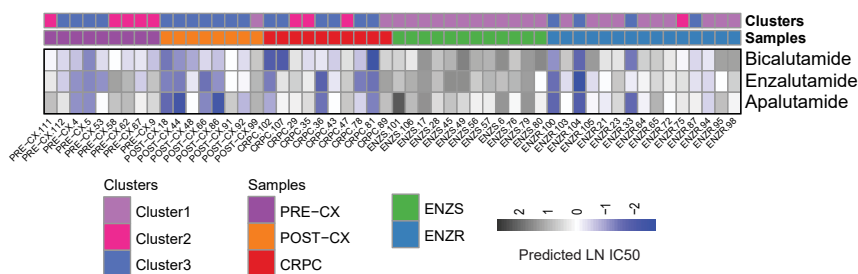


Figure 4.9: Heatmap depicting predicted LN IC50 for BIC, APA, and ENZ, These drugs are not present in the GDSC2 dataset. Color bars represent different types of tumors and clusters obtained through UMAP based 2D projections of Predicted LN IC50.

We noted that PRE-CX, POST-CX, and CRPC groups tend to be sensitive to AR antagonists. Notably, additional AR antagonists were predicted to not confer an additional advantage to ENZS tumors treated with ENZ (Figure 4.9).

### 4.3.5 Verification of models trained on TCGA patient tumor profiles

The Cancer Genome Atlas (TCGA) is an extensive collection of datasets that span different cancer types and include bulk RNA-seq profiles of patient tumors as well as clinical drug response data. The clinical drug response furnishes patient demographics and response information for the administered drug. We intend to perform classification by grouping patients into responders (Patients showing complete or partial response) and non-responders (patients showing clinically progressing or stable disease). To make most of this information entailing gene expression profiles of the patient and

patient drug response information, we build a classifier using open-source AutoML from H2O.ai [118] in R. Analogous to bulk RNA-seq gene expression profiles of the CCLE cell lines dataset, we combined embeddings of drugs (n=139) obtained from clinical drug response metadata files with processed TCGA GSVA score matrix (methods). This matrix contained 3108 patient/drug pairs in rows and 1427 feature set involving 1327 pathway vectors and embeddings of length 100 for each molecular compound. These vectors formed explanatory variables for predicting drug response as sourced from TCGA clinical metadata files. Due to the staggering diversity of cancer genomes, these numbers are not sufficient, therefore we conducted a pooled analysis of data, irrespective of cancer stage or type. Due to scarcity of data, we use AutoML, R library by H2O.ai (<https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html>) [118] to build a classifier for drug response prediction by leveraging bulk patient profiles from TCGA. AutoML framework allows users to train and optimize various models automatically by defining the maximum number of models that need to be trained. This results in the automation of machine learning workflows. We split final processed dataset of dimension  $3108 \times 1427$  into 90% training and 10% test set. 90% training set was subjected five-fold cross-validation and hyper parameter tuning. We ensured there was no overlapping patients in train-validation-test sets. AutoML resulted in 34 models spanning across different classes of models, namely, GBM, XGBoost, DRF, DeepLearning, XRT models, and two stacked ensemble models. Two automatically trained stacked ensemble models corresponded to one based on all the previously trained models, while the second one is based on each family's best model [118]. Extremely Randomized Trees (XRT) was yielded as the best model. We assessed the performance of trained models using independent test dataset. We noted that the XRT outperformed all other models having an AUC-PR of 0.85 (Figure 4.10A).

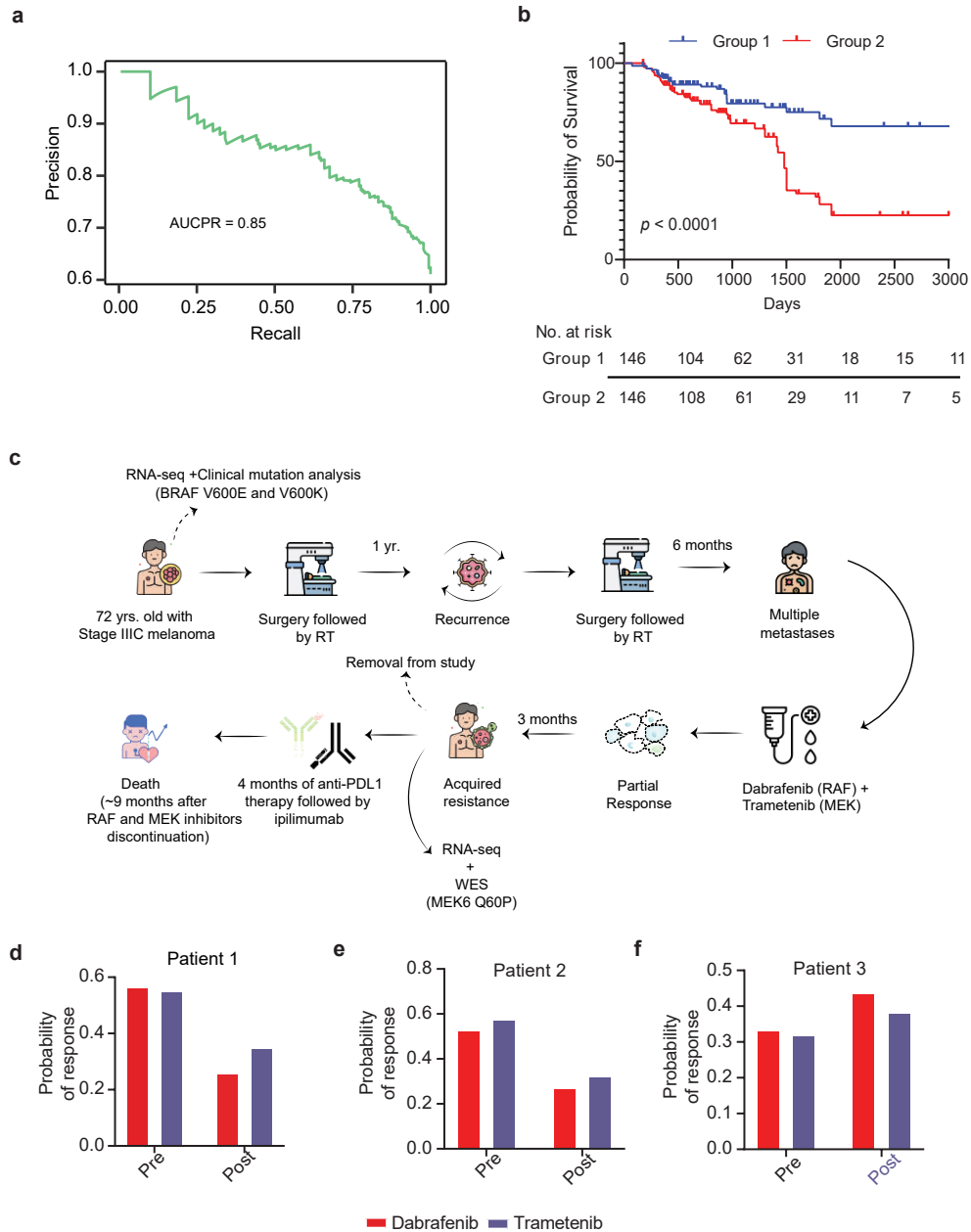


Figure 4.10: Assessment of efficiency of TCGA model (A) Precision-recall curve depicting performance of best model on TCGA test dataset. (B) Survival analysis on a TCGA test dataset encompassing different cancers. Patients were stratified into high and low-risk groups based on the predicted response probability's median value (0.63). A p-value (P-value < 0.0001) was obtained from the log-rank statistical test to compare survival between the two groups. (C) Schematics depicting a melanoma patient's progression from diagnosis to initial therapy and recruitment into dabrafenib and trametinib clinical phase I/II study resulting in initial relapse of lesions but eventually patient died nine months after stopping RAF and MEK inhibitors. Image credit: <https://www.flaticon.com>. (D)-(F) For patients 1, 2, and 3, bar plots showing the predicted probability of response to the first-line treatments dabrafenib and trametinib.



We also investigated whether including cancer stage information improves drug response prediction. Surprisingly, the model's performance declined with the inclusion of the stage, and we obtained an AUC-PR of 0.79. Furthermore, we also used Precily (classifier variant) on the TCGA dataset which yielded an AUC-PR of 0.77. This is expected due to paucity. Moreover, TCGA data furnishes survival data for patients. Utilizing this information, we performed survival analysis on the TCGA test dataset containing 293 patients by grouping patients using the median value of the classifier's estimated probability of response. The median value was 0.63. Notably, patients in group 1 had a predicted probability of response higher than the median value. As expected, this group had better survival ( $P < 0.0001$ ) (Figure 4.10B). Further, the 5-year survival probability for group 1 and group 2 was 0.72 and 0.28, respectively. Worth mentioning that due to the limited availability of patient data at the level of disease, cancer stage and treatment, it is challenging to assess therapeutic application of our patient model.

Next, we tested the efficiency of our model using an external dataset having bulk RNA-seq profiles of patients before and after treatment along with clinical response information. We predicted response for two drugs: dabrafenib, an RAF inhibitor, and trametinib, a MEK inhibitor in three matched prior treatment and post-relapse BRAF mutant melanoma patients. The journey of patient 1, who was detected with melanoma (Stage IIIC), is shown in (Figure 4.10C). Then after one year of primary treatment, this patient exhibited indications of recurrent disease. The patient was subjected to pre-treatment biopsy and radiotherapy which involved the usage of intensity-modulated radiation therapy (IMRT). This patient had BRAF V600E as well as V600K mutations as revealed through clinical mutational analysis of tumor and was enrolled in a phase I/II study of dabrafenib and trametinib drugs. However, the patient was withdrawn from the clinical study after three months of enrollment due to the emergence of resistance. Then for four months, this patient was given an anti-PDL1 antibody until the progression of the disease and thereafter with four rounds of ipilimumab. Unfortunately, the patient died about nine months after discontinuation of RAF and MEK inhibitors. The potential cause of therapy resistance was the MEK2Q60P mutation, as demonstrated through whole-exome sequencing (WES) and RNA-seq analysis. Patient 2 was also detected with melanoma (stage IB) who developed widely metastatic melanoma after five years of initial diagnosis. The existence of metastatic melanoma was confirmed through pleural biopsy i.e. pre-treatment biopsy and this patient exhibited BRAF V600E mutation.

The patient received dabrafenib and trametinib as the first line of treatment. However, only a partial response was observed, and after 3 months regular scans showed considerable progression of the disease. A plausible reason for therapy resistance was the existence of a BRAF splicing variant discovered through using RNA-seq and WES in post-treatment samples but splice variant was not detected in pre-treatment tumors. Unfortunately, the patient died approximately six months after the initial presentation of a metastatic condition. On the other hand, patient 3 was diagnosed with melanoma of the left thigh (stage IIIC). This patient also exhibited BRAF V600E mutation. After six months of surgery, this patient was recruited in a clinical study of dabrafenib and trametinib. However, patient 3 developed a progressive disease nearly after one year on a trial. The potential reason for therapeutic resistance is the existence of amplification of BRAF in the post-treatment sample as revealed by WES. The patient was treated with ipilimumab for a brief period but unfortunately died after four rounds and three months after stopping BRAF and MEK inhibitor. The presence of alterations such as BRAF amplification, MEK2Q60P mutation, and BRAF Splice Isoform in post-treatment tumors appears to be a probable reason for vesting antagonism to RAF and MEK inhibitors in these three patients [199]. Notably, our prediction results showed similar patterns, where the response probability for the two drugs trametinib and dabrafenib was higher in pre-treatment profiles of patient one and patient two than in matched post-treatment (Figure 4.10D, E). Additionally, we accurately predicted these two patients as responders which is in concordance with the actual annotations provided in the study where they are categorized in the partial response category based on Response Evaluation Criteria In Solid Tumors (RECIST) [54]. We predicted patient three to be resistant to dabrafenib in pre and post-treatment settings (Figure 4.9F). This also aligns with the actual study where this patient falls in a category of stable disease based on RECIST criteria.

## 4.4 Discussion

Prediction of the best treatment and drug response in cancer is of utmost priority for personalized cancer treatment. In this work, we developed Precily, a deep neural network (DNN) model for drug response prediction based on pathway scores of bulk gene expression profiles of cancer cell lines, and numerical embeddings of associated drugs.

Owing to the usage of pathway activity scores, our model pinpoints underlying biological mechanisms involved in inducing drug resistance. Further, the usage of pathway scores for modeling drug response enabled us to reliably conjecture the fate of cells upon treatment from single-cell gene expression profiles. This could pave the way for the prediction of drug response at the sub-clonal level using scRNA-seq tumor profiles.

We assessed Precily by demonstrating a suitable correlation between observed LN IC50 and Precily predictions for real-life LNCaP baseline transcriptome samples. Further, we observed that PTEN null LNCaP cell line was predicted to be susceptible to drugs targeting PI3K/mTOR signaling pathways. Moreover, past studies have proposed that enhanced PI3K/AKT/mTOR signaling is linked with susceptibility of PI3K targeting drugs in LNCaP cells and other PTEN negative cancer cell lines. Additionally, PTEN null tumors might show sensitivity to mTOR inhibitors [224]. We were able to spot the susceptibility of LNCaP cells to PI3K/mTOR inhibitors with our predictive model. These findings are concordant with earlier studies in the LNCaP cell line. With the belief attained from cell line-based verification, we investigated how drug response prediction alters with drug treatments, drug-resistant and responsive states in LNCaP cell line and xenografts. We endeavor to interrogate how our findings associate with existing biology encompassing these states and treatments. LNCaP cells were predicted to be more susceptible to drugs targeting highly proliferative cells in the presence of androgens. This is an anticipated outcome since androgens induce proliferation in androgen-positive PCa cell lines and xenografts [84][161]. AR inhibitors are standard treatment options for metastatic PCa. AR antagonists suppress AR signaling pathways at the molecular level. With AR antagonists involving ENZ, APA, and BIC treatments, LNCaP cells predicted patterns demonstrated to have prominent similarities and differences, reflecting the intricate biology and mechanism underlying the therapeutic responses. The substantial reversal of DHT endowed sensitivity was noted for PI3K/mTOR inhibitors, with ENZ showing the most significant effect. On the other hand, treatment with ENZ appeared to augment the DHT endowed sensitivity to some drugs such as cisplatin, paclitaxel and docetaxel. While BIC and APA may not for cisplatin. These differential effects must be taken into consideration for selecting optimal combinatorial therapy while accounting for therapeutic resistance.

To further test the applicability of Precily, we utilized our internally generated data of xenografts from a large biologically well-elucidated study. This study progressed

from the early androgen-responsiveness state to the advanced CRPC stage and then the aggressive ENZ treatment-resistant stage. The drug response predictions illustrate altering susceptibilities of the tumors as they progress through different stages and treatments. In particular, we noted that ENZR tumors were predicted to develop new vulnerabilities to novel therapies, opening up new opportunities for novel therapeutic options. For instance, in the case of treatment with ENZ, we could predict sensitivity to a small group of drugs specifically those involved in targeting of EGFR pathway. However, in the absence of androgens, prostate tumors were predicted to be highly insensitive to EGFR signaling inhibitors. This pinpoints the emergence of distinct vulnerability during passage to ENZR. As EGFR signaling targeting drugs have been approved for multiple cancer, this may help in the clinical evaluation of combinatorial therapy in PCa patients who have received ENZ treatment or have developed resistance to ENZ. Past studies have revealed that using a combination of EGFR inhibitor and ENZ might be a viable treatment option to overcome ENZ resistance. Further laboratory experiments and preclinical studies are required to investigate molecular mechanisms associated with predicted drug response patterns to corroborate our prediction patterns.

Assessment of the TCGA model using an external patient melanoma dataset resulted in clinically applicable predictions. As envisaged, the probability of response as predicted by the model for patients belonging to the partial response category was more than the post-treatment samples due to the emergence of resistance to standard treatments. This recommends that top drugs predicted by our model, such as cisplatin and cyclophosphamide, could be used as an alternative therapy to overcome acquired resistance in combination with other drugs. This requires further studies.

There are several advantages of our work. First, due to the usage of numerical drug embeddings, Precily models can be used to conjecture drug response for any sample-drug pair. Second, this approach enables model performance improvement by allowing the pooling of cell line-drug combinations across cancer types. Third, Precily can be used to predict responses for drugs that are not part of training data. Finally, Precily monotherapy predictions can be used to infer potential combinatorial therapies. Thus, Precily can assist in clinical decision making.

There are certain limitations of Precily. We could not obtain a good correlation between predicted and observed IC50 values. The relative sensitivities are, nevertheless,

quite accurately reflected between drugs. Although we achieved promising results on TCGA data, due to data scarcity of patient data and corresponding drug response information, we believe this model can be further improved by adding data from various clinical studies.

Overall, this is the first study linking computational drug response predictions to clinically explicable findings in both in vitro and in vivo setups. Furthermore, to the best of our knowledge, this is the first work interrogating the prospect of bulk tumor RNA-seq data profiles for drug response prediction in prostate cancer. Thus, this work will open up new avenues and help researchers and clinicians in clinical decision-making and assessing drug resistance and sensitivity in cell lines, xenografts, and patient tumors.

# CHAPTER 5

## Effect of physical proximity on gene expression, cell-cell interactions and signaling at single cell resolution

### 5.1 Introduction

The discovery of important molecular pathways has refined our understanding of tumor microenvironment, tumor progression and dissemination, and oncogenesis. The complexities of cell-cell communication and the possibilities for modulation open up new avenues for cancer treatment. Cell-cell interaction or cell-cell communication is crucial in orchestrating the development of multicellular organisms. It is a complex phenomenon where a single cell interacts with other cells through physical contact, ligand-receptor interactions, and paracrine signaling [148]. Thus, it is a critical process for morphogenesis, differentiation, and maintaining biological functions and microenvironmental homeostasis [10] [26]. Cancer is a complex global health issue. One of the significant hurdles to understanding this disease is cell communication in the tumor microenvironment. Tumor tissue comprises non-cancerous cells and non-cellular components, and essentially 50 percent of its make-up can come from non-neoplastic cells. Cancer cells alter host cells and impart tumor-supportive traits to them. The altered host cells further aid in tumor progression and modifications of other normal cells within the microenvironment [148]. However, complex cellular interaction networks between cancer and the host cell are poorly understood [10] [148]. With the development of Single-cell RNA sequencing (scRNA-seq), our understanding of functional heterogeneity of tissues and the functional cell states within the tumor microenvironment have been dramatically refined. It is critical to comprehend the underlying mechanisms of different cellular components interactions to uncover tumor growth emergent behavior. However, there is still a dearth of comprehension of how these relationships quantitatively connect to particular phenotypic effects of interest due to the finite number of methods to quantify live cell-cell interactions [114]. Recently, combinations of spatial-omics approaches have been utilized to characterize live cell-cell interactions [13] [22]. We

devised a microfluidic workflow that captures and co-incubates live single immune, single cancer cells, or doublets. This framework utilizes a single-cell dosing mRNA-seq integrated fluidic circuit (IFC) system (Fluidigm®) [171], allowing transcriptional and spatial cell-cell interactions. To illustrate the performance of our approach in quantifying cell-cell interactions, we applied novel microfluidic workflow to cancer-immune doublets (CIDs) of natural killer and triple-negative breast cancer cells.

TNBC was selected as a model system because it is an aggressive subtype of breast cancer and more challenging to treat in comparison to hormone-positive breast cancer and is associated with higher metastatic potential and poor prognostic outcomes [160] [234]. The standard treatment for TNBC is neoadjuvant chemotherapy [150]. However, chemotherapy responses are generally momentary. Cancer immunotherapies are transforming the cancer cell therapy landscape. It is conjectured that the immune responses elicited by immunotherapies are expected to target and eliminate tumor cells while sparing normal cells. Blocking immune checkpoints with neutralizing and blocking antibodies, cytotoxic T lymphocytes (CTLs) induction, and remodeling of the tumor microenvironment to increase CTL activity are among the immunotherapy approaches that have already been developed and tried [101]. Clinically promising findings take advantage of specific traits of cross-talk between immune-tumor cells entailing immunosuppression and anti-tumor responses. Among immune cells, Natural killer (NK) cells are the central effector cells of innate immunity and exhibit a high level of heterogeneity in the microenvironment. They are named for their capabilities to destroy target cells autonomously. The majority of existing treatment options employing tumor microenvironment rely on the immunity of T cells. However, little success is achieved with T cell immunotherapy. This highlights the need to develop new immunotherapies such as previously overlooked NK cells [213]. NK cells are major constituents of the innate immune system that play a crucial role in cancer control. NK cells' critical role in cancer immunity stems from their ability to identify malignant cells using a variety of receptors on their surface, allowing them to detect and destroy tumor cells rapidly through targeted cytotoxicity. It is hypothesized that heterogeneity in NK cells results in dynamic interaction between NK and tumor cells with divergent regulation of their cytotoxic effects, eliciting tumor death depending on the balance between activating and inhibitory receptor levels. It is conjectured that genetically manipulated NK cells can influence cancer immune surveillance and tumor progression. Further, they play an

essential role in orchestrating cancer immunity locally through communications with other cells in the tumor microenvironment via secretions of multiple chemokines and cytokines [15]. Therefore, it is critical to identify the molecular level signals stemming from single NK cells when they come in contact with tumor cells. Consequently, a single-cell framework to quantify interactions between NK and cancer cells is essential to study the role of NK cells in cancer immunotherapy.

To gain a better understanding of NK and tumor cells interactions, we propose a microfluidic approach that involves capturing and co-incubating single NK and cancer cells doublets (CIDs) by employing the Polaris™ Single-Cell Dosing mRNA Seq IFC. Time-lapse imaging was used to track physical distances between CIDs captured in the microfluidic chamber. Single-cell RNA sequencing (scRNA-seq) is performed on the cells after a 13-hours of incubation with growth media exchanged at a set interval of time, i.e. 5 hours. This yielded 290 transcriptomes comprising single NK, single cancer cells, and NK-cancer cell doublets (CIDs). Unsupervised clustering of single-cell transcriptomes indicated heterogeneity in TNBC cell lines. Furthermore, we could characterize gene signatures associated with the anti-tumor activity of NK cells. Only a few killing events were observed among the incubated CIDs where NK cells were involved in the lysis of cancer cells. We correlated hourly computed physical distances between NK cells and tumor cells with the terminally computed single-cell gene expression profiles of doublets. The results pointed towards the presence of transcriptomic memory, which is driven by explicit regulatory modules that are active in a time-dependent manner. Additionally, we interrogated ligand-protein interactions and found that few ligand-protein pairs, including CD24-SIGLEC10 and ANXA1-EGFR, had augmented activity in doublets and substantiated earlier reported interaction between CD24 and SIGLEC10 as a potential target for cancer immunotherapy in ovarian and TNBC [17].

## 5.2 Methods

### 5.2.1 NK cell activation

The NK cells activation was confirmed, in general, by carefully assessing the NK cell line NK-92MI and breast cancer cell line MDA-MB-231. After 24 and 48 hours of



incubation, activation was estimated in the proportion of 3 NK cells for each breast cancer MDA-MB-231 cell by assessing the known markers using flow cytometry (Sony SH800S). These markers include CD25, CD69 and CD314 which are expressed when one cell comes in contact with other.

## **5.2.2 Cell lines and culture**

TNBC MDA-MB-231 (ATCC® HTB-26™), was cultured in DMEM/high modified culture medium with inactivated fetal bovine serum (10%) and antibiotics penicillin/streptomycin (1%). The culture was replicated every three days, maintaining a confluence of 40% at the time of passage and maintained at a temperature of 37°C in a humidified atmosphere with 5% CO<sub>2</sub>. The adherent cells were detached with Gibco TrypLE reagent and then resuspended in the complete culture medium.

NK-92MI (ATCC® CRL-2408), genetically modified human NK cell cells to produce interleukin two were cultured in the medium supplemented with Alpha Minimum Essential medium (AMEM), inositol (0.2 mM), mercaptoethanol (0.1 mM), folic acid (0.02 mM), fetal bovine serum (12.5%), and horse serum (12.5%). The homogenization of NK cells was performed to get separate clusters before replication in a new vessel i.e. 25 cm<sup>2</sup> flask with previously cultured cells (1mL) + new culture medium (9mL). Culturing of the NK and MDA-MB-231 cells was performed individually for the subsequent cell-cell interactions experiments employing the Fluidigm Polaris system.

## **5.2.3 Fluidigm Polaris protocols for selection and incubation of cells**

The initial step involves priming of IFC with beads which aid in the cell adhesion that are to be incubated in the chambers. To differentiate cell types, cells and reagents were already labeled using specific markers after treatment. NK cells labeled with celltracker fluorescent dye far red, cancer cells with celltracker orange, and calcein AM for viability were pipetted into IFC. The Polaris system was set up to select NK and cancer cells that were positive for celltracker far red fluorescent dye and calcein AM.

In some IFC wells, only a single cell was maintained for comparing single-cell gene expression profiles with the gene expression profiles acquired from incubation of

cells as doublets (NK cell+ breast cancer cell). Following selection, Polaris equipment was configured to incubate cells for 16 hours, replacing the culture medium containing DMEM (20%) + AMEM (20%) every 5 hours and time-lapse imaging was performed every hour before and after culture image change. Then, after incubation of single cancer cells (n=71), single NK cells (n=77), and cancer-immune doublets (CIDs) (n=132), cells were subjected to lysis, reverse transcription, amplification of cDNA, and finally single-cell RNA sequencing using NextSeq Illumina 500.

#### **5.2.4 Distance estimation between cells**

In this study, the distance reported for the CIDs group is the shortest distance between the membrane of NK and MDA-MB-231 cells on IFC. The distance was quantified for 13-time points. We utilized the National Institutes of Health's ImageJ software which is open-source Java-based software for image processing to scrutinize videos of cells frame by frame to generate distance data [175].

#### **5.2.5 Preprocessing of dataset**

In total, we acquired single-cell gene expression data for 340 cells. The cells can be categorized into the following categories: 77 single NK cells, 71 single tumor cells, 132 CIDs throughout all time points, 10 CIDs that remained solely NK cells at the 13th time point. Out of a total of 340 cells, we discarded 50 cells, including— two bulk biological replicates each of MDA-MB231 and NK cell lines, four unoccupied chambers where none of the cells could be detected from beginning to end, eight unoccupied chambers which started with tumor cells, ten unoccupied chambers that started with NK cells, two CIDs that initially began as single NK cells and 22 tumor cells that began as CIDs. After discarding these cells, a total of 290 cells remained. Further, we screened the dataset for low-quality cells and retained those cells having >2000 expressed genes, i.e., non-zero TPM RNA-seq quantified by RSEM software [122]. Next, we kept genes having TPM expression value >5 in  $\geq 10$  cells. Notably, at this stage, our gene expression matrix contained 290 cells and 8907 protein-coding genes. This matrix was subjected to the Seurat pipeline and used for various other analyses.

### 5.2.6 Seurat workflow

After basic preprocessing steps, the filtered gene expression matrix was subjected to the single-cell Seurat pipeline from R and other downstream analyses [185]. The standard preprocessing steps of Seurat include gene filtering (second pass). We utilized genes present in at least five cells, followed by log-normalization and variance stabilizing transformation to detect highly variable genes using functions `NormalizeData()` and `FindVariableFeatures()` using default parameters. We have used Seurat's data integration workflow to integrate data originating from two independent runs. To ward off the batch effect and integrate data, we used the `FindIntegrationAnchors()` function with `k.filter` option as 100 to identify anchor cells that depict matching pairs of cells with similar biological states across two datasets to transform transcriptomes into shared space. Then, anchors were integrated using the `integrateData()` function, which involves Canonical Correlation Analysis (CCA). This results in the batch adjusted matrix. The 2D map of cells was plotted and visualized using the `RunUMAP()` function.

### 5.2.7 Differential genes

The R package Limma with Limma-voom [117] functionality to obtain differentially expressed genes in the cell groups. We used an adjusted p-value  $< 0.05$  and log 2 fold change cutoff of 1 to select differential genes.

### 5.2.8 Survival analysis using genes upregulated in NK cells exhibiting cytotoxic activity

We used PROGgeneV2 [71] combined gene signature analysis functionality to perform overall survival analysis using gene signature or gene list upregulated in NK cells displaying anti-tumor activity on the TCGA-BRCA dataset comprising survival data of 594 patients. Of 187 elevated genes in the NK cells exhibiting tumor-killing activity, 164 genes were mapped to the TCGA-BRCA dataset. Kaplan-Meier plot was generated based on the categorization of patients into a high and low-risk group based on the median value of combined expression of gene signature as a cut-off.

### 5.2.9 Regulation of intercellular distances and transcriptional memory

We used two matrices to access the connection between cell-cell distances captured at 13 time points and gene expression profiles that were profiled at the 13th time point. The first matrix is the gene expression matrix of dimension  $|G|*|C|$  where  $|G|$  corresponds to genes and  $|C|$  corresponds to the CIDs. The second matrix is of cell-cell distances of dimensions  $|C|*|T|$  where  $|C|$  represents the same cells as in the gene expression matrix and  $|T|$  is the time points at which cell-cell distances were captured. These two matrices were used to estimate the Pearson correlation matrix of dimension  $|G|*|T|$  where  $|G| = 2000$  and  $|T| = 13$ . This matrix contained Pearson's correlation coefficients  $\rho_{g,t}$  for each gene-time point pairs  $(g, t)|t \in T, g \in G$ . We retained 90 genes that had a correlation greater than 0.25 in at least one of the time points. The correlation matrix with these 90 genes was used for hierarchical clustering, and we obtained four gene modules using `cutree()`. All four modules were subjected to motif enrichment and TF activity analysis using `RcisTarget` [3] and `ShinyGO` [65], respectively. Utilizing the `hg19-tss-centered-10kb-7species.mc9nr.feather` database including genome-wide ranking for the motifs, `RcisTarget` identified enriched TF binding motifs and provided list of TFs for every module. Using `igraph` [46] R package regulatory networks were built.

### 5.2.10 CIDs and cell-cell signaling

We utilized the `iTALK` R package [205] containing 2,648 unique ligand-receptor interactions specific to cancer. Using our gene expression matrix that comprises 8907 protein-coding genes and 290 cells as an input to `rawParse()` function using `mean` as method of a `stats`, the top 50% of highly expressed genes were selected. This resulted in the identification of 230 ligand-receptor pairs from the `FindLR()` function. Then, Pearson's correlation coefficient was calculated between the gene expression vectors linked with chosen ligand-protein pairs in CIDs—TU-NK/TU-NK and TU-NK/NK. Those pairs showed Pearson's correlation of more than 0.4 qualified for downstream analysis. We also computed Pearson's correlation coefficient for NK/NK and TU/TU cases to preclude the possibility that observed co-expression is exclusively due to NK or tumor cells alone. Using these criteria, we chose twenty ligand-protein pairs, and

three of them were found to be directly implicated in signaling in breast cancer.

### 5.2.11 Data availability

All raw and processed sequencing data used in this study have been submitted to the NCBI Gene Expression Omnibus under accession number GSE181591.

### 5.2.12 Code availability

<https://github.com/SmritiChawla/NKCell>

### 5.2.13 Results

To investigate NK-TNBC cells interactions, single NK-92MI cells, single MDA-MB-231 cells, and cancer-immune doublets (CIDs) entailing one NK-92MI cell and one MDA-MB-231 were captured and incubated for 13 hours using the Fluidigm Polaris system [171][211][162][208] (Figure 5.1A). We used time-lapse imaging to capture snapshots of CIDs every hour to quantify the physical distance between them. Incubation was followed by processing single cells and cancer-immune doublets for single-cell RNA-sequencing. The integrated fluidic circuit (IFC) allows multiple assays to be performed parallelly involving capturing of cells, co-incubation, lysis of cell, reverse transcription, and cDNA amplification [162]. To discriminate between different cell types, i.e., single-cell and CIDs, we scrutinized the expression of known cell type-specific marker genes. Based on differential gene expression analysis between single NK and cancer cells, we identified known markers of these cells. Notably, we noted high expression of NK cell marker genes, namely KLRD1, CCR6, LAIR1 [229] and TNFRSF9 [198] in NK cells, and single cancer cells showed expression of TNBC marker genes HMGA1, ANKRD11 [172] and TACSTD2 [204] (Figure 5.1B) when visualized through SCANPY python package for single-cell analysis [209].

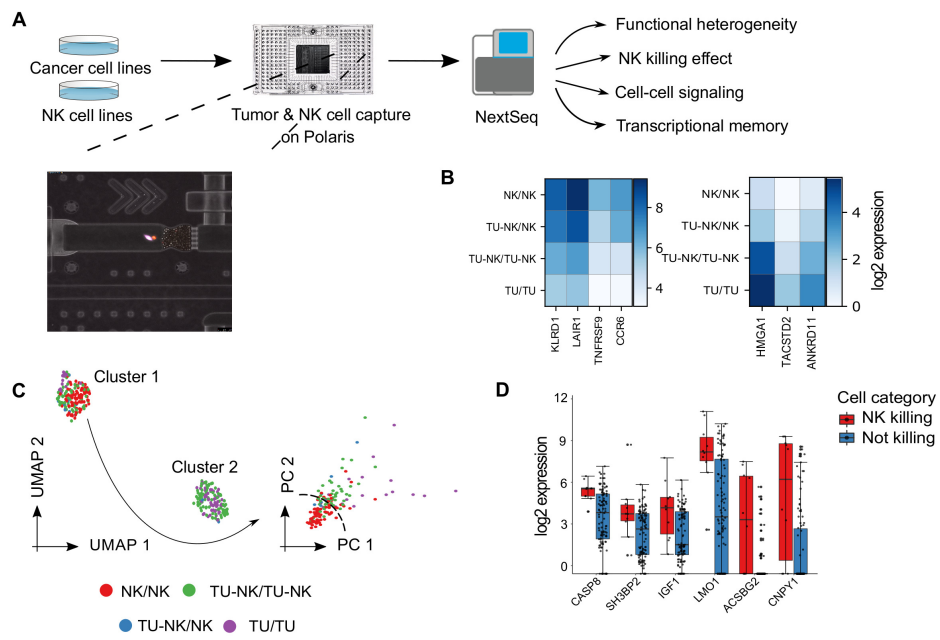


Figure 5.1: **(A)** Schematics of cell interaction studies. The method comprises propagation of NK and cancer cells in culture, staining them off-chip, and capturing them as live single cells and cancer-immune doublets. Cells are incubated and co-incubated and their images are captured over time. A microscopic view of a chamber on the microfluidic Integrated Fluidic Circuit (IFC) with an MDA-MB-231 (Blue) and NK cell (Red) doublet is shown by the inset. Subsequently, the cells are subjected to lysis followed by reverse transcription, and the cDNA is amplified in-situ within the chambers. Then preparation of library, barcoding of sample, and sequencing are done utilizing the Illumina NextSeq system off-chip. **(B)** Heatmap showing mean expression of canonical markers for single NK and cancer cells, substantiating their lineage identities. **(C)** UMAP based visualization shows two distinct clusters of cells highlighting heterogeneity in cancer cell lines and PCA-based visualization reveals spatial separation of NK cells. **(D)** Boxplots depicting differentially expressed genes in the cells belonging to the NK killing cell vs. non-killing group.

Furthermore, to ensure the identities of cell types, we exploit the Polaris system's imaging capability. The NK cells were labeled with CellTracker™ Deep Red fluorescent dye, and cancer cells were labeled with CellTracker™ Orange CMRA fluorescent dye before selecting cells on the microfluidic IFC. The 2D visualization of z-score normalized intensities of NK and cancer cell channels using scatterD3 R package revealed the clustering of cells according to their annotations based on cell labeling. Further, we subjected the expression profiles to Seurat v3 based unsupervised analysis of transcriptional heterogeneity within the single cells and the doublets (i.e., CIDs). Uniform

Manifold Approximation and Projection (UMAP) based 2D projections of gene expression profiles unveiled two distinct clusters [185], which were predominantly dominated by cancer cell line clonal heterogeneity (Figure 5.1C). The single NK cells were found to be part of cluster 1 that harbored cancer cells as well. We noticed spatial segregation of NK cells when transcriptomes of cluster 1 were subjected to principal component analysis (PCA) (Figure 5.1C). To further characterize the heterogeneity exhibited by the TNBC cancer cell line, we exclusively conducted unsupervised clustering of transcriptomes of single cancer cells. As expected, this also resulted in two separate clusters featuring exclusive arrays of differentially upregulated genes.

#### **5.2.14 Tracking of distance in cancer-immune cells over time reveals the existence of transcriptional memory**

Over 13 hours of incubation in the same microfluidic compartment, we monitored the CIDs for deciphering dynamic alterations in the physical proximity between the cancer cells and corresponding NK cells (Figure 5.1A). After 13 hours, the CIDs ( $n = 102$ ) were subjected to RNA sequencing. Consequently, we could decipher the relationship between terminally-estimated gene expressions with distances of cancer-immune cells quantified across multiple time points.

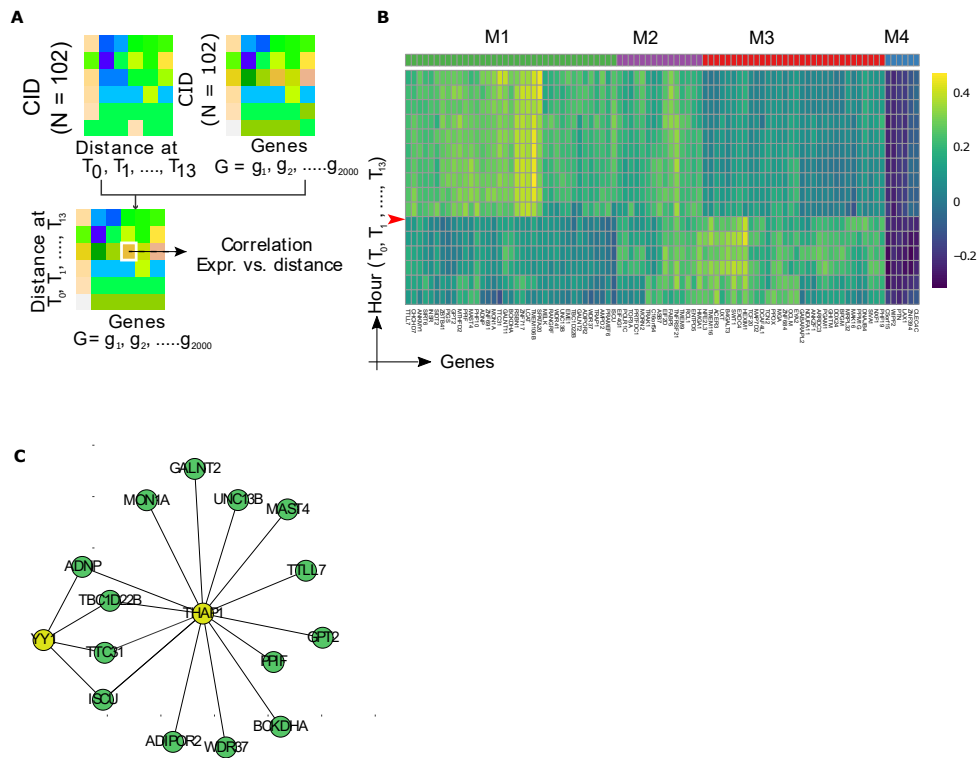


Figure 5.2: (A) Overall workflow of estimating correlation between gene expression profiles and cell-cell distances captured at 13 time points. (B) Heatmap showing a correlation between gene expression profiles and cell-cell distances along with the regulatory gene modules governing these associations. Genes influencing cellular distances between NK and cancer cells are clustered into four modules: M1 to M4. The red arrow indicates the time point at which of first culture medium exchange was performed. (C) Regulatory network for the genes belonging to module 1 (M1) which are potentially regulated by the putative Transcription Factors (TFs) THAP1 and YY1 as identified through RcisTarget based TF binding motif Enrichment Analysis.

The transcriptomes entailing transcripts whose profiling was performed at the culmination of the 13th hour correlated significantly with distances among CIDs across all the time points (Figure 5.2A, B). Our results revealed time-bound activities of at least three distinct gene regulatory modules governing cellular distances. We found changes in gene regulatory modules over time that correlated to the physical distances between CIDs (Figure 5.2B). At time point 6, i.e., after five-hour incubation and culture medium change, we observed a dramatic shift in distance modulation activities, with a new bunch of genes in module 2 (M2) taking over the control. Notably, a similar change in gene expression was not observed in later time points of culture medium change. We used ShinyGO [65] and RcisTarget [3] R packages to interrogate these gene modules and performed transcription factor analysis for each module. This resulted in inferring



the regulatory role of three putative transcription factors, including BRCA1, YY1, and THAP1. Among these, ShinyGO predicted BRCA1, while YY1 and THAP1 with their putative target genes were predicted by RcisTarget (Figure 5.2C).

### 5.2.15 Analysis of killing events in cancer-Immune cells

Our unique experiment design allowed us to track the killing events of cancer cells by NK cells and the associated gene signatures. Only ten cancer cell lysis events were spotted across 132 CIDs, signifying the rarity of the event. In order to identify gene signatures associated with cancer cell elimination, differential gene expression analysis was performed between two CIDs subgroups — the CIDs featuring cancer cell lysis (NK killing) and the remaining CIDs (non-killing). CASP8, SH3BP2, IGF-1, CNPY1, and LMO1 are noteworthy among the 187 genes found to be upregulated in the minor CID subgroup (NK-killing group) (Figure 5.1D).

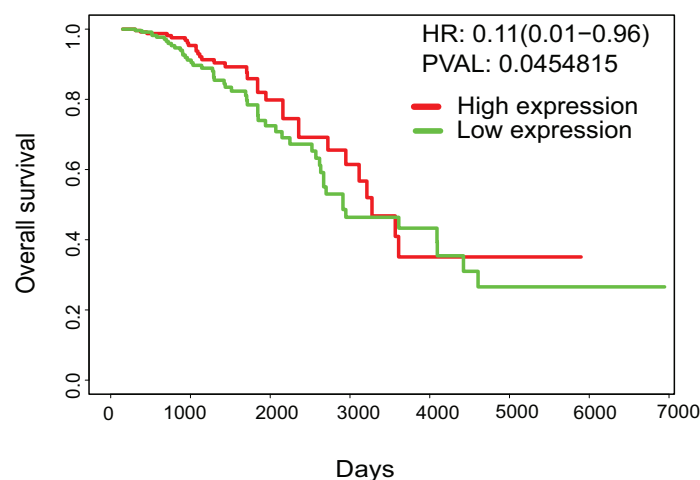


Figure 5.3: Kaplan-Meier survival curve created using PROGgeneV2 for TCGA-BRCA dataset using gene signature upregulated in NK killing group. Patients were grouped into high and low expression groups of the gene signature based on a median value (Log-rank test P-value of 0.045).

Furthermore, we analyzed the prognostic impact of combined expression of 164 genes upregulated in the NK killing event having antitumor activity on overall survival of The Cancer Genome Atlas-Breast Invasive Carcinoma (TCGA-BRCA) patients using PROGgene V2, a webserver to study the prognostic implication of genes of interest in different cancer types [71]. We found a survival benefit in the patients having a higher mean expression of the combined gene signature associated with NK cells= anti-

tumor activities with a Hazard ratio (HR) of 0.11 and a statistical significance of  $p < 0.05$  (Figure 5.3).

### 5.2.16 Analysis of cell-cell signaling in CIDs

Intercellular signaling is a vital element of interactions between cancer and immune cell. We employed gene expression profiles as a substitute for ligand-protein activity. We primarily concentrated on cancer-specific ligand-protein pairs presented in the iTALK database [205]. To discern the level of ligand-protein interaction, we estimated Pearson's correlation coefficient for ligand-protein pairs in CIDs and single cells (Figure 5.4). In CIDs, we noted a correlation between ANXA1 and EGFR. ANXA1 is associated with the endocytosis of the EGFR receptor ANXA1-S100A11 complex [92][23][48] and also plays a critical role in cellular communication via exosomal EGFR [163]. In the CIDs, we also noted a high correlation between CD24 and SIGLEC10. Further, an elevated correlation was also observed between EGFR and HSP90AA1.

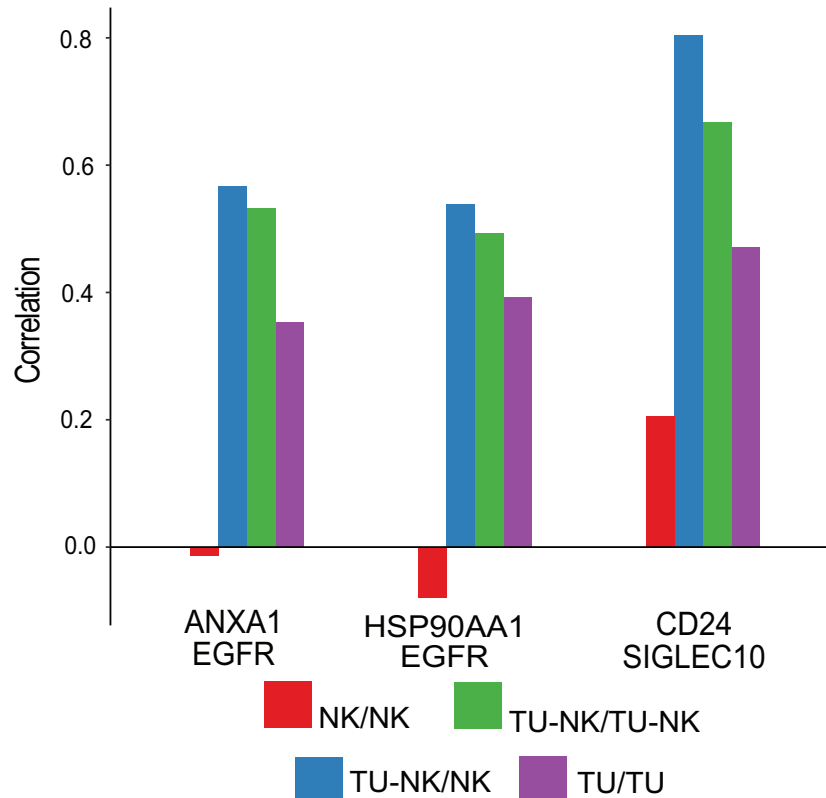


Figure 5.4: Barplots showing increased overall coordination or correlation for the three ligand-protein pairs namely ANXA1-EGFR, HSP90AA1-EGFR, and CD24-SIGLEC10.

### 5.3 Discussion

TNBC is one of the highly aggressive subtypes of breast cancer with very few therapeutic options. However, immunotherapy has lately emerged as a potential strategy for the clinical management of cancer. In addition to existing immunotherapy based on T cells that are routinely employed, now the investigation is also concentrated on exploiting NK cells that play significant roles in innate immune responses in cancer. Consequently, exhaustive delineation of cancer and NK cell interactions profiled at a single-cell level might untangle applicable biomarkers and pathways implicated in the development and progression of tumors.

In this study, we interrogated the gene expression profiles of cancer-NK cell doublets rendered by computing the interaction distance between CIDs over time. Computational and bioinformatics approaches to the data enabled us to discern gene expression signatures connected with NK cell killing events. We further verified that these gene signatures are associated with patient survival using TCGA gene expression profiles. Our single-cell analysis revealed that heterogeneity in two cancer subclones interfered dimensionality reduction approach by obscuring the identification of NK cells. Further, we observed the activation of gene regulatory modules in a timely manner directing cellular distances during the course of the interaction of NK-cancer cells under co-incubation. We also observed a link between cell-cell distances for CIDs and their gene expression profiles that were profiled at the last time point of co-incubation i.e., 13 hours in live cells. Gide, Tuba N., et. demonstrated prospective associations between the physical proximity of cancer-immune cells with anti-PD-1 therapeutic response in metastatic melanoma patients [69]. This highlights the significance of intercellular distance as a valuable measure for understanding cancer immunosurveillance and response. Intuitively, there's a possibility that the exposure of cancer and immune cells to each instigates an adaptive configuration of transcriptional memory. Recent work by Battich and colleagues has pinpointed that regulated mRNA transcripts synthesis and degradation is a critical regulatory approach impacting the fate of cells [19].

Transcriptional memory is the ability of cells to possess reversible memory to retaliate to previously encountered stimuli more robustly in the future [20]. At time point six, we observed a new group of genes in action. The cells might form a lytic immune synapse as reported by past studies during the initial hours of co-incubation; after that,

interactions might deviate [151][115]. In this study, we noted the regulatory role of three putative transcription factors in transcriptional memory. These TFs are BRCA1, YY1, and THAP1. BRCA1 is the best-studied tumor suppressor gene and is known to be involved in breast and ovarian cancer. YY1 is involved in promoting oncogenic programs and activities in breast cancer [200], whereas THAP1 plays a vital role in the repair of DNA and has also exhibited elevated levels in breast cancers. Overexpression of one of the genes TRAP1 belonging to module 1 promotes tumor growth in breast cancer. On the other side, TRAP1 also reduces metastasis by controlling the mitochondrial dynamics (process of mitochondrial fission and fusion) [227]. Another gene MELK from the same module 1 is involved in the proliferation of TNBC. MELK inhibition can result in the arrest of the cell cycle through reduction of cyclin B1 and increase of p27 and p-JNK [123]. Another gene, EYA2, is also implicated in breast cancer promotion. Elevated expression of this gene can increase proliferation markers such as cyclin E, PCNA, and EGFR [217].

Further, we investigated CIDs for cell killing events and noted transcriptomic signatures—CASP8, SH3BP2, IGF-1, and LMO1. Activation of CASP8 by FASLG leads to activation of the extrinsic apoptotic pathway in the target cells [235]. SH3BP2 [97] and IGF-1 [147], on the other hand, have been found to play a critical role in NK cell cytotoxicity and development. Notably, we observed that strong differential expression cues stemmed from certain genes that are mainly not documented for their role in NK cell cytotoxicity. These include upregulation of the LMO1 gene, which is overexpressed in T lymphocytes in lymphoblastic leukemia [132]. Other genes upregulated in the NK killing group are CNPY1 and ACSBG2. CNPY1 gene is known to regulate FGF signaling in zebrafish [86]. On the other hand, ACSBG2 is known to play role in fatty acid metabolism [155]. However, the role of these genes concerning NK cell-mediated cytotoxicity has not been documented in the literature. Thus, this warrants further investigation.

Additionally, since cell-cell signaling is an important element of cancer-immune interactions, we investigated coordination among specific ligand-protein pairs in CIDs. Notably, CD24 (receptor) - SIGLEC10 (ligand) transcripts exhibited a specifically higher correlation in doublets in comparison to single NK and cancer cells. The association among this pair in CIDs highlights the significance of the CD24-SIGLEC10 in NK and TNBC cells. A previous study reported the role of SIGLEC10 in hindering NK cell

functionality and is also linked with poor survival of patients with hepatocellular carcinoma (HCC) [228]. Another study looked into the possibility of CD24-SIGLEC10 interactions in TNBC in the presence of tumor-associated macrophages (TAMs). CD24 and SIGLEC10 are upregulated in several tumor types and TAMs [17], respectively. Targeting this interaction can be important from a therapeutic point of view. Then, in another ligand-protein pair, EGFR-HSP90AA1, HSP90AA1 is involved in keeping the stability and functionality of its receptor EGFR. This stability results in promoting pathogenesis in breast, head and neck cancers [2] [62]. This is accomplished through epithelial to mesenchymal transition and activating signaling pathways linked with tumor migration pathways in MDA-MB-231 cells [193].

Currently, significant challenges are involved with the co-incubation of single cells in regulated condition that concurrently enables the analysis of cell-cell interactions between live cells and their impact on gene expression [10]. Here, we report a study on cell-cell interactions by employing automated conditions that accurately regulate media exchange, temperature, the composition of gas, and humidity to scrutinize and quantify cellular distances constantly. By processing doublets within the same microfluidic chamber on IFC using microfluidic multi-step chemistry that involves lysis of cells, reverse transcription, and amplification of cDNA, physical proximity quantification can be directly associated with downstream transcriptomic changes.

The proposed microfluidic workflow enabled us to pinpoint unique molecular signatures specific to NK cells exhibiting cytotoxic activities. We identified highly coordinated activities of gene expression profiles driving and influencing the distances of interacting cancer and NK cells, substantiating transcriptional memory as a primary governing strategy of cells. We could also delineate elevated coordination in specific ligand-protein pairs, as demonstrated through gene expression profiles. In the future, this microfluidic approach might open up new windows to studying cellular interactions in an immuno-oncology context and further aid in development and administration of NK cell-based cancer immunotherapies.

# CHAPTER 6

## Conclusion

This thesis focuses on various statistical modeling and computational biology approaches for analyzing single-cell data. Our work incorporated essential features of integrative analysis of transcriptomics and genomics data. Further, we have demonstrated the applications of statistical modeling, and machine learning approaches in context-specific regulations and drug response prediction in cancer.

### 6.0.1 Summary of contribution

In this section, we give a brief summary of the chapters giving a comprehensive view of the thesis.

### 6.0.2 Transformation of single-cell transcriptomics and epigenomics data in pathway scores using UniPath and its evaluation

Recent breakthroughs in single-cell RNA-sequencing and ATAC-sequencing have opened up new doors of challenges and opportunities for investigating new applications with a relevant conversion of read counts that generally possess high technical noise and dropouts. We developed a novel method UniPath to represent single cells in pathway space. Our method transforms single-cell gene expression and open chromatin profiles into pathway or gene enrichment scores. The robust statistical framework of UniPath involving the use of the global null model results in high consistency, accuracy, and scalability in computing gene set or pathway enrichment scores in each cell. UniPath approach is such that it can handle easily systematic dropouts and batch effects in scRNA-seq gene expression profiles. Further, pathway scores obtained from UniPath provide improved accuracy of visualization and clustering for scRNA-seq profiles than other similar methods, including PAGODA, AUCell, and GSEA. UniPath also outperforms these methods in computing cell type-specific enrichment of genesets in single cells. UniPath framework also facilitates dimensions reduction of single-cell

open-chromatin profiles. After transformation into pathway scores, similar downstream analyses could be performed on scRNA-seq and scATAC-seq profiles. Therefore, the proposed method provides a uniform platform for interrogating single-cell gene expression and open-chromatin profiles at pathways resolution.

### **6.0.3 Applications of UniPath**

The utilization of pathway enrichment scores in the single-cell discipline has emanated as an effective tool to decipher cellular heterogeneity to procure novel and biologically relevant information for a multitude of applications. This chapter introduces the applications of UniPath transformed pathway scores. UniPath pathway scores can be used for pseudotemporal ordering of single cells while enabling for suppressing of covariate effects such as cell cycle, tissue microenvironment. We were able to capture the true order of cells differentiating into endoderm. However, other methods, namely Monocle, TSCAN, DiffusionMap, and CellTree that leverages gene expression profiles instead of pathway scores, resulted in the wrong order prediction. UniPath also enables visualization of the continuum of lineage potency and co-occurrence of pathways on pseudotemporally ordered tree. This might help visualize cancer cell fates. Further, our pathway-based clustering of large-scale mouse cell atlas data showed biologically relevant grouping of cell types from various distant organs and also revealed a new sub-cluster of cells. Such an approach can be extended to cluster tumor biopsies to reveal new cancer cell states and discover new and rare cell types. UniPath proved beneficial in discerning context-specific regulations in cancer that are often needed in precision oncology. Further, our approach could highlight patterns of pathway co-occurrence for distinguishing two groups of cells in NSCLC.

### **6.0.4 Inference of drug response sensitivity in cancer by leveraging gene expression data in pathway space**

Tumor heterogeneity is a significant hurdle in the cancer treatment. Recently, a considerable amount of drug screening datasets have become publicly available. These large-scale datasets provided an opportunity to apply machine learning that holds potential in predicting appropriate patient-tailored therapies. Numerous machine learning

methods have been developed for predicting drug response. However, after carefully scrutinizing these methods, we found two crucial areas for improvement. First, most previous studies do not account for molecular structure information, and secondly, most of the studies utilize gene expression profiles to predict drug response. It is increasingly becoming apparent that coordinated activities of multiple genes in a pathway could influence the drug response instead of a single gene. In this study, we developed Precily, deep neural network based framework to model drug response in *in vivo* and *in vitro* contexts by utilizing pathway enrichment scores and numerical drug descriptors. Our pathway-based approach enabled our model to highlight biological mechanisms associated with drug resistance and sensitivity. Further, using pathway scores to model drug response enabled us to reliably allude fate of cells under treatment from scRNA-seq gene expression data. This can open up avenues for drug response prediction at the subclonal level using cancer scRNA-seq data. Drug response predictions on in-house generated prostate cancer datasets, including cell lines, cell lines under differential treatment conditions and xenografts, revealed biologically significant results. We evaluated our approach on pan-cancer TCGA data and external melanoma dataset, resulting in clinically relevant predictions. This is the first study linking systematic drug response prediction to clinically relevant findings in *in vivo* and *in vitro* settings. Overall, our results suggest that patterns of pathway scores in cancer cells have the potential to highlight drug sensitivity in cancer cells and thus can be used for personalized treatment decisions.

### **6.0.5 Effect of physical proximity on gene expression, cell-cell interactions and signaling at single cell resolution**

Our study is the first to monitor time-dependent interactions between NK and triple-negative breast cancer cell doublets at single-cell resolution, which allowed us to identify gene signatures specific to NK cells that can potentially kill the cancer cells. We identified highly coordinated regulatory activities of gene expression profiles influencing active changes in the physical distance in doublets, supporting the transcriptional memory narrative as an essential regulatory programme of cells. We could also delineate inflated coordination among some specific selected ligand-receptor pairs using gene expression profiles. The proposed microfluidic workflow and our initial obser-



vations might provide new insights into studying cellular interactions and signaling in immuno-oncology contexts, which holds considerable potential in helping in designing NK-based immunotherapies.

## **6.1 Future work**

Some of the probable future extensions of the presented works are outlined below.

1. Owing to the advent of Single-cell RNA-seq technologies, our understanding of intratumor heterogeneity has been dramatically refined. Although single-cell transcriptomics provides unprecedented advantages in the clinical domain, they are yet fully explored for designing patient-specific therapies accounting for intratumor heterogeneity. We have limitedly exploited single-cell RNA-seq data for drug response prediction in our work. We want to exploit scRNA-seq profiles of the tumor microenvironment further to discern subclonal drug response and resistance.

2. Spatial transcriptomics is a powerful technique that has considerably improved our understanding of cellular interactions and the functional organization of tissues. This cutting-edge technology allows positional mapping of gene activity which is crucial in understanding tumor pathogenesis. However, currently, it is achieved by investigating the expression of one gene at a time. But this approach is quite limiting as cancer is a complex disease and the majority of well-known cancer phenotypes are manifested through the coordination of multiple genes. The spatial distribution of such transcriptomics gene signatures is not well understood. Therefore, linking clinically relevant gene signatures with spatial coordinates will help in a better understanding of diseases such as cancer as it is greatly influenced by the tumor microenvironment.

# APPENDIX A

## Supplementary Information

We evaluated UniPath using cell type markers and compared it with three other methods namely, PAGODA, AUCell and GSVA. We used 10 scRNA-seq studies including heterogeneous and homogeneous datasets to systematically assess UniPath in revealing correct terms among top enriched terms. The terms here refer to cell types. We estimated the percentage of cells with correct cell types among top enriched terms. In most of the cases UniPath performed better compared to PAGODA, AUCell and GSVA (Figure A.1)

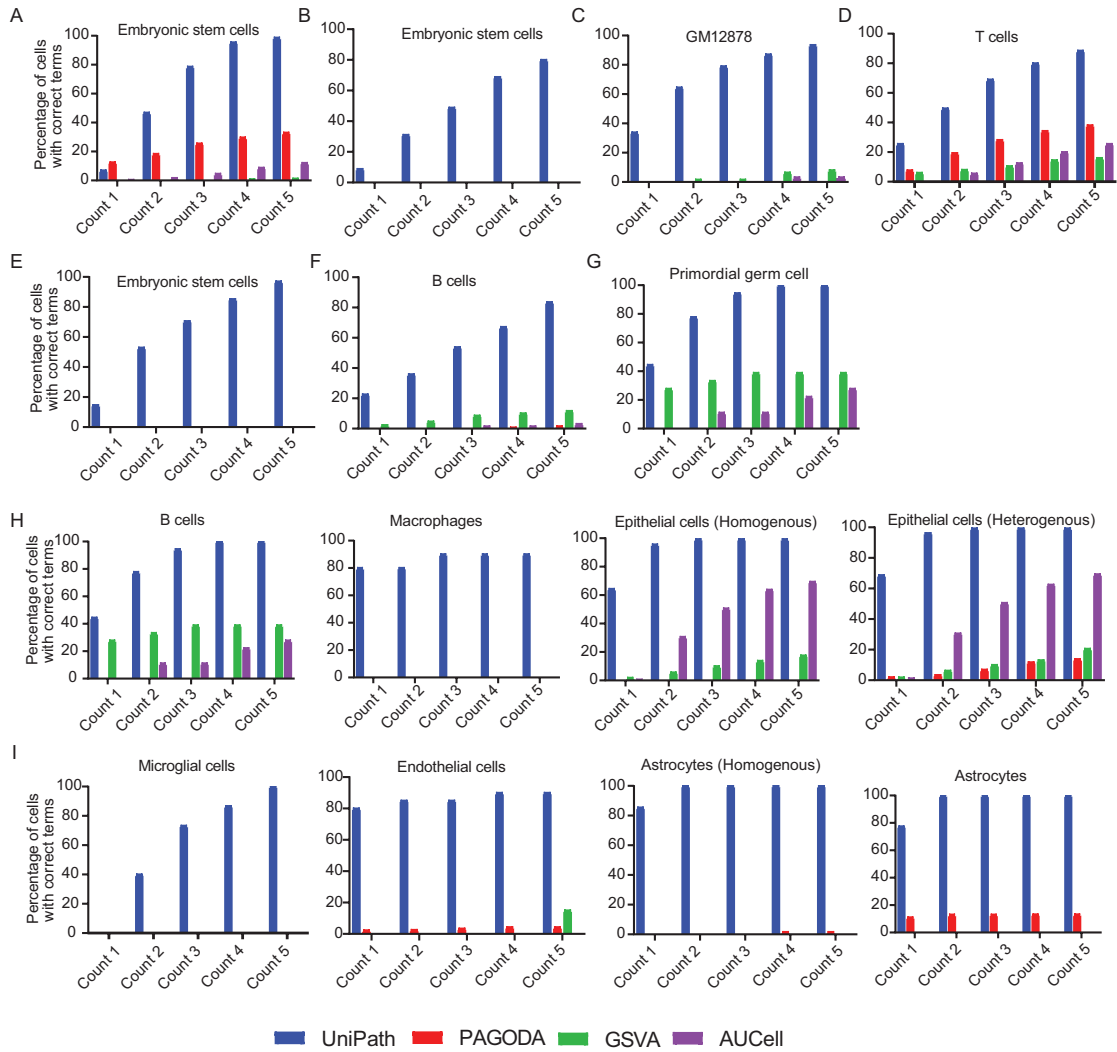


Figure A.1: Comprehensive assessment of UniPath on scRNA-seq profiles. (A) The bars depict the percentage of Embryonic stem cell (ESC) detection among the top enriched terms for the homogeneous dataset (GSE64016) [121]. Count1 represents the percentage of correct cell type as first enriched term and Count5 shows the percent of correct cell type in the top 5 enriched terms. (B) Estimating accuracy of correct detection of ESC from the homogeneous dataset (GSE71858) [192] among top enriched terms. (C) Accuracy of correct cell type detection in homogeneous dataset of GM12878 cells (GSE44618) [137]. (D) Accuracy of correct cell type detection in homogeneous dataset of T cells (GSE98638) [232]. (E) Accuracy of correct cell type detection in a heterogeneous dataset of ESC (GSE36552) [219]. (F) Accuracy of correct cell type detection in a homogeneous dataset of B cells [233]. (G) Accuracy of correct cell type detection in a heterogeneous dataset of Primordial germ cell (GSE63818) [74]. (H) Accuracy of correct cell type detection for B cells (heterogeneous), Macrophages (heterogeneous) and epithelial cells (both homogeneous and heterogeneous) (GSE81861) [125]. (I) Accuracy of correct cell type detection for Microglial cells (heterogeneous), Endothelial cells (heterogeneous) and Astrocytes (both homogeneous and heterogeneous) (GSE67835) [47].

## REFERENCES

- [1] **Adam, G., L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg** (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, **4**(1), 1–10.
- [2] **Ahsan, A., S. G. Ramanand, C. Whitehead, S. M. Hiniker, A. Rehemtulla, W. B. Pratt, S. Jolly, C. Gouveia, K. Truong, C. Van Waes, D. Ray, T. S. Lawrence, and M. K. Nyati** (2012). Wild-type EGFR is stabilized by direct interaction with HSP90 in cancer cells and tumors. *Neoplasia*, **14**(8), 670–677.
- [3] **Aibar, S., C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, and S. Aerts** (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**(11), 1083–1086.
- [4] **Aiken, J., G. Buscaglia, E. A. Bates, and J. K. Moore** (2017). The  $\alpha$ -Tubulin gene TUBA1A in brain development: A key ingredient in the neuronal isotype blend. *J Dev Biol*, **5**(3).
- [5] **Akkin, S., G. Varan, and E. Bilensoy** (2021). A review on cancer immunotherapy and applications of nanotechnology to chemoimmunotherapy of different cancers. *Molecules*, **26**(11).
- [6] **Al-Ostoot, F. H., S. Salah, H. A. Khamees, and S. A. Khanum** (2021). Tumor angiogenesis: Current challenges and therapeutic opportunities. *Cancer Treatment and Research Communications*, 100422.
- [7] **Ammad-Ud-Din, M., S. A. Khan, D. Malani, A. Murumägi, O. Kallioniemi, T. Aittokallio, and S. Kaski** (2016). Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, **32**(17), i455–i463.
- [8] **Anders, S., P. T. Pyl, and W. Huber** (2015). HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2), 166–169.
- [9] **Andrews, S. et al.** (2017). Fastqc: a quality control tool for high throughput sequence data. 2010.
- [10] **Armingol, E., A. Officer, O. Harismendy, and N. E. Lewis** (2020). Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.*, **22**(2), 71–88.
- [11] **Ashton, B. A., T. D. Allen, C. Howlett, C. Eaglesom, A. Hattori, and M. Owen** (1980). Formation of bone and cartilage by marrow stromal cells in diffusion chambers in vivo. *Clinical orthopaedics and related research*, (151), 294–307.
- [12] **Azuaje, F.** (2017). Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics*, **18**(5), 820–829.

- [13] **Baccin, C., J. Al-Sabah, L. Velten, P. M. Helbling, F. Grün-schläger, P. Hernández-Malmierca, C. Nombela-Arrieta, L. M. Steinmetz, A. Trumpp, and S. Haas** (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat. Cell Biol.*, **22**(1), 38–48.
- [14] **Baird, A.-M., J. Leonard, K. M. Naicker, L. Kilmartin, K. J. O’Byrne, and S. G. Gray** (2013). IL-23 is pro-proliferative, epigenetically regulated and modulated by chemotherapy in non-small cell lung cancer. *Lung Cancer*, **79**(1), 83–90.
- [15] **Bald, T., A.-M. Pedde, D. Corvino, and J. P. Böttcher** (2020). The role of NK cell as central communicators in cancer immunity. *Adv. Immunol.*, **147**, 61–88.
- [16] **Ballinger, A.** (2000). Orlistat in the treatment of obesity. *Expert opinion on pharmacotherapy*, **1**(4), 841–847.
- [17] **Barkal, A. A., R. E. Brewer, M. Markovic, M. Kowarsky, S. A. Barkal, B. W. Zaro, V. Krishnan, J. Hatakeyama, O. Dorigo, L. J. Barkal, and I. L. Weissman** (2019). CD24 signalling through macrophage siglec-10 is a target for cancer immunotherapy. *Nature*, **572**(7769), 392–396.
- [18] **Barretina, J., G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, Jr, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palesscandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway** (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391), 603–607.
- [19] **Battich, N., J. Beumer, B. de Barbanson, L. Krenning, C. S. Baron, M. E. Tanenbaum, H. Clevers, and A. van Oudenaarden** (2020). Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science*, **367**(6482), 1151–1156.
- [20] **Beiter, T., A. M. Nieß, and D. Moser** (2020). Transcriptional memory in skeletal muscle. don’t forget (to) exercise. *J. Cell. Physiol.*, **235**(7-8), 5476–5489.
- [21] **Ben-Hamo, R., A. J. Berger, N. Gavert, M. Miller, G. Pines, R. Oren, E. Pikarsky, C. H. Benes, T. Neuman, Y. Zwang, et al.** (2020). Predicting and affecting response to cancer therapy based on pathway-level biomarkers. *Nature communications*, **11**(1), 1–16.
- [22] **Berglund, E., J. Maaskola, N. Schultz, S. Friedrich, M. Marklund, J. Bergenstråhle, F. Tarish, A. Tanoglidli, S. Vickovic, L. Larsson, F. Salmén, C. Ogris, K. Wallenborg, J. Lagergren, P. Ståhl, E. Sonnhammer, T. Helleday, and J. Lundeberg** (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**(1), 2419.

- [23] **Boudhraa, Z., B. Bouchon, C. Viallard, M. D’Incan, and F. Degoul** (2016). Annexin A1 localization and its relevance to cancer. *Clin. Sci.*, **130**(4), 205–220.
- [24] **Broad Institute TCGA Genome Data Analysis Center** (2016). Analysis-ready standardized TCGA data from broad GDAC firehose 2016\_01\_28 run.
- [25] **Browaeys, R., W. Saelens, and Y. Saeys** (2020). Nichenet: modeling intercellular communication by linking ligands to target genes. *Nature methods*, **17**(2), 159–162.
- [26] **Brücher, B. L. D. M. and I. S. Jamall** (2014). Cell-cell communication in the tumor microenvironment, carcinogenesis, and anticancer treatment. *Cell. Physiol. Biochem.*, **34**(2), 213–243.
- [27] **Buenrostro, J. D., M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang, and W. J. Greenleaf** (2018). Integrated Single-Cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, **173**(6), 1535–1548.e16.
- [28] **Buenrostro, J. D., B. Wu, U. M. Litzénburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf** (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**(7561), 486–490.
- [29] **Buettner, F., K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle** (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**(2), 155–160.
- [30] **Buettner, F., N. Pratanwanich, D. J. McCarthy, J. C. Marioni, and O. Stegle** (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.*, **18**(1), 212.
- [31] **Bujold, D., D. A. d. L. Morais, C. Gauthier, C. Côté, M. Caron, T. Kwan, K. C. Chen, J. Laperle, A. N. Markovits, T. Pastinen, B. Caron, A. Veilleux, P.-É. Jacques, and G. Bourque** (2016). The international human epigenome consortium data portal. *Cell Syst*, **3**(5), 496–499.e2.
- [32] **Cai, J., L. Fang, Y. Huang, R. Li, X. Xu, Z. Hu, L. Zhang, Y. Yang, X. Zhu, H. Zhang, J. Wu, Y. Huang, J. Li, M. Zeng, E. Song, Y. He, L. Zhang, and M. Li** (2017). Simultaneous overactivation of Wnt/ $\beta$ -catenin and TGF $\beta$  signalling by mir-128-3p confers chemoresistance-associated metastasis in NSCLC. *Nat. Commun.*, **8**, 15870.
- [33] **Chang, K., C. J. Creighton, C. Davis, L. Donehower, J. Drummond, D. Wheeler, A. Ally, M. Balasundaram, I. Birol, Y. S. Butterfield, et al.** (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet*, **45**(10), 1113–1120.
- [34] **Chang, Y., H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T. S. Kim, J. Jung, and J.-M. Shin** (2018). Cancer drug response profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.*, **8**(1), 8857.

- [35] **Chang, Y., H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T. S. Kim, J. Jung, and J.-M. Shin** (2018). Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, **8**(1), 1–11.
- [36] **Chawla, S., S. Samydurai, S. L. Kong, Z. Wu, Z. Wang, W. L. Tam, D. Sen-gupta, and V. Kumar** (2021). Unipath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic acids research*, **49**(3), e13–e13.
- [37] **Chen, L., Y. Zeng, and S.-F. Zhou** (2018). Role of apoptosis in cancer resistance to chemotherapy. *Current understanding of apoptosis-programmed cell death*.
- [38] **Chen, S., B. B. Lake, and K. Zhang** (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, **37**(12), 1452–1457.
- [39] **Chiodoni, C., M. T. Di Martino, F. Zazzeroni, M. Caraglia, M. Donadelli, S. Meschini, C. Leonetti, and K. Scotlandi** (2019). Cell communication and signaling: how to turn bad language into positive one. *J. Exp. Clin. Cancer Res.*, **38**(1), 128.
- [40] **Chu, L.-F., N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendzioriski, R. Stewart, and J. A. Thomson** (2016). Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, **17**(1), 1–20.
- [41] **Clark, S. J., H. J. Lee, S. A. Smallwood, G. Kelsey, and W. Reik** (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.*, **17**, 72.
- [42] **Collado-Torres, L., A. Nellore, K. Kammers, S. E. Ellis, M. A. Taub, K. D. Hansen, A. E. Jaffe, B. Langmead, and J. T. Leek** (2017). Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, **35**(4), 319–321.
- [43] **Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al.** (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, **17**(1), 1–19.
- [44] **Cooper, J., Y. Ding, J. Song, and K. Zhao** (2017). Genome-wide mapping of dnase i hypersensitive sites in rare cell populations using single-cell dnase sequencing. *Nature protocols*, **12**(11), 2342–2354.
- [45] **Corces, M. R., J. D. Buenrostro, B. Wu, P. G. Greenside, S. M. Chan, J. L. Koenig, M. P. Snyder, J. K. Pritchard, A. Kundaje, W. J. Greenleaf, R. Majeti, and H. Y. Chang** (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48**(10), 1193–1203.
- [46] **Csardi, G., T. Nepusz, and Others** (2006). The igraph software package for complex network research. *InterJournal, complex systems*, **1695**(5), 1–9.

- [47] **Darmanis, S., S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Hayden Gephart, B. A. Barres, and S. R. Quake** (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, **112**(23), 7285–7290.
- [48] **de Graauw, M., L. Cao, L. Winkel, M. H. A. M. van Miltenburg, S. E. le Dévédec, M. Klop, K. Yan, C. Pont, V.-M. Rogkoti, A. Tijmsma, A. Chaudhuri, R. Lalai, L. Price, F. Verbeek, and B. van de Water** (2014). Annexin A2 depletion delays EGFR endocytic trafficking via cofilin activation and enhances EGFR signaling and metastasis formation. *Oncogene*, **33**(20), 2610–2619.
- [49] **Ding, M. Q., L. Chen, G. F. Cooper, J. D. Young, and X. Lu** (2018). Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol. Cancer Res.*, **16**(2), 269–278.
- [50] **Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras** (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**(1), 15–21.
- [51] **Dumbrava, E. I. and F. Meric-Bernstam** (2018). Personalized cancer therapy—leveraging a knowledge base for clinical decision-making. *Molecular Case*.
- [52] **duVerle, D. A., S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda** (2016). CellTree: an r/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics*, **17**(1), 363.
- [53] **Efremova, M., M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo** (2020). Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, **15**(4), 1484–1506.
- [54] **Eisenhauer, E. A., P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij** (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer*, **45**(2), 228–247.
- [55] **Elsafadi, M., M. Manikandan, M. Atteya, J. A. Hashmi, Z. Iqbal, A. Aldahmash, M. Alfayez, M. Kassem, and A. Mahmood** (2016). Characterization of cellular and molecular heterogeneity of bone marrow stromal cells. *Stem cells international*, **2016**.
- [56] **Fan, J., N. Salathia, R. Liu, G. E. Kaeser, Y. C. Yung, J. L. Herman, F. Kaper, J.-B. Fan, K. Zhang, J. Chun, and P. V. Kharchenko** (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods*, **13**(3), 241–244.
- [57] **Fan, J., K. Slowikowski, and F. Zhang** (2020). Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp. Mol. Med.*, **52**(9), 1452–1465.



- [58] **Fan, J., K. Slowikowski, and F. Zhang** (2020). Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine*, **52**(9), 1452–1465.
- [59] **Fard, M. K., F. van der Meer, P. Sánchez, L. Cantuti-Castelvetri, S. Mandad, S. Jäkel, E. F. Fornasiero, S. Schmitt, M. Ehrlich, L. Starost, T. Kuhlmann, C. Sergiou, V. Schultz, C. Wrzos, W. Brück, H. Urlaub, L. Dimou, C. Stadelmann, and M. Simons** (2017). BCAS1 expression defines a population of early myelinating oligodendrocytes in multiple sclerosis lesions. *Sci. Transl. Med.*, **9**(419).
- [60] **Feng, F., B. Shen, X. Mou, Y. Li, and H. Li** (2021). Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine. *Journal of Genetics and Genomics*.
- [61] **Finak, G., A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo** (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
- [62] **Friedland, J. C., D. L. Smith, J. Sang, J. Acquaviva, S. He, C. Zhang, and D. A. Proia** (2014). Targeted inhibition of hsp90 by ganetespib is effective across a broad spectrum of breast cancer subtypes. *Invest. New Drugs*, **32**(1), 14–24.
- [63] **Furusawa, C., T. Suzuki, A. Kashiwagi, T. Yomo, and K. Kaneko** (2005). Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *Biophysics*, **1**, 25–31.
- [64] **Gagnon-Bartsch, J. A. and T. P. Speed** (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**(3), 539–552.
- [65] **Ge, S. X., D. Jung, and R. Yao** (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, **36**(8), 2628–2629.
- [66] **Ge, X., P. Lyu, Y. Gu, L. Li, J. Li, Y. Wang, L. Zhang, C. Fu, and Z. Cao** (2015). Sonic hedgehog stimulates glycolysis and proliferation of breast cancer cells: Modulation of PFKFB3 activation. *Biochem. Biophys. Res. Commun.*, **464**(3), 862–868.
- [67] **Gerdes, H., P. Casado, A. Dokal, M. Hijazi, N. Akhtar, R. Osuntola, V. Rajeeve, J. Fitzgibbon, J. Travers, D. Britton, et al.** (2021). Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nature communications*, **12**(1), 1–15.
- [68] **Gerrits, E., Y. Heng, E. W. G. M. Boddeke, and B. J. L. Eggen** (2020). Transcriptional profiling of microglia; current state of the art and future perspectives. *Glia*, **68**(4), 740–755.
- [69] **Gide, T. N., I. P. Silva, C. Quek, T. Ahmed, A. M. Menzies, M. S. Carlino, R. P. M. Saw, J. F. Thompson, M. Batten, G. V. Long, R. A. Scolyer, and J. S. Wilmott** (2020). Close proximity of immune and tumor cells underlies response to anti-PD-1 based therapies in metastatic melanoma patients. *Oncoimmunology*, **9**(1), 1659093.

- [70] **Gong, W., I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry** (2018). Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, **19**(1), 1–10.
- [71] **Goswami, C. P. and H. Nakshatri** (2014). PROGgeneV2: enhancements on the existing database. *BMC Cancer*, **14**, 970.
- [72] **Greaves, M. and C. C. Maley** (2012). Clonal evolution in cancer. *Nature*, **481**(7381), 306–313.
- [73] **Grossman, R. L., A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt** (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**(12), 1109–1112.
- [74] **Guo, F., L. Yan, H. Guo, L. Li, B. Hu, Y. Zhao, J. Yong, Y. Hu, X. Wang, Y. Wei, et al.** (2015). The transcriptome and dna methylome landscapes of human primordial germ cells. *Cell*, **161**(6), 1437–1452.
- [75] **Gutschner, T. and S. Diederichs** (2012). The hallmarks of cancer: a long non-coding rna point of view. *RNA biology*, **9**(6), 703–719.
- [76] **Güvenç Paltun, B., H. Mamitsuka, and S. Kaski** (2021). Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Briefings in bioinformatics*, **22**(1), 346–359.
- [77] **Haghverdi, L., F. Buettner, and F. J. Theis** (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, **31**(18), 2989–2998.
- [78] **Han, X., R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang, Y. Xu, W. Huang, M. Jiang, X. Jiang, J. Mao, Y. Chen, C. Lu, J. Xie, Q. Fang, Y. Wang, R. Yue, T. Li, H. Huang, S. H. Orkin, G.-C. Yuan, M. Chen, and G. Guo** (2018). Mapping the mouse cell atlas by Microwell-Seq. *Cell*, **173**(5), 1307.
- [79] **Hanahan, D. and R. A. Weinberg** (2000). The hallmarks of cancer. *cell*, **100**(1), 57–70.
- [80] **Hanahan, D. and R. A. Weinberg** (2011). Hallmarks of cancer: the next generation. *cell*, **144**(5), 646–674.
- [81] **Hänzelmann, S., R. Castelo, and J. Guinney** (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
- [82] **Hänzelmann, S., R. Castelo, and J. Guinney** (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
- [83] **Hastie, T. and J. Qian** (2014). Glmnet vignette. *Retrieved June*, **9**(2016), 1–30.
- [84] **Heinlein, C. A. and C. Chang** (2004). Androgen receptor in prostate cancer. *Endocr. Rev.*, **25**(2), 276–308.
- [85] **Heldin, C.-H., M. Vanlandewijck, and A. Moustakas** (2012). Regulation of EMT by TGF $\beta$  in cancer. *FEBS Lett.*, **586**(14), 1959–1970.

- [86] **Hirate, Y.** and **H. Okamoto** (2006). Canopy1, a novel regulator of fgf signaling around the midbrain-hindbrain boundary in zebrafish. *Current Biology*, **16**(4), 421–427.
- [87] **Hoeben, A., E. A. J. Joosten,** and **M. H. J. van den Beuken-van Everdingen** (2021). Personalized medicine: Recent progress in cancer therapy.
- [88] **Howe, K. L., P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. Garcia Giron, T. Grego, C. Gujjarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O. G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. Gonzalez Martinez, J. C. Marugán, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D. N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. De Silva, B. Flint, A. Frankish, S. E. Hunt, G. R. Hsley, N. Langridge, J. E. Loveland, F. J. Martin, J. M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, F. Cunningham, A. D. Yates, D. R. Zerbino,** and **P. Flicek** (2021). Ensembl 2021. *Nucleic Acids Res.*, **49**(D1), D884–D891.
- [89] **Hu, Z., S. A. Brooks, V. Dormoy, C.-W. Hsu, H.-Y. Hsu, L.-T. Lin, T. Massfelder, W. K. Rathmell, M. Xia, F. Al-Mulla,** *et al.* (2015). Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: focus on the cancer hallmark of tumor angiogenesis. *Carcinogenesis*, **36**(Suppl\_1), S184–S202.
- [90] **Huang, H.** (2007). Isolation of human Placenta-Derived multipotent cells and in vitro differentiation into Hepatocyte-Like cells.
- [91] **Hwang, B., J. H. Lee,** and **D. Bang** (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, **50**(8), 1–14.
- [92] **Jaiswal, J. K., S. P. Lauritzen, L. Scheffer, M. Sakaguchi, J. Bunkenborg, S. M. Simon, T. Kallunki, M. Jäättelä,** and **J. Nylandsted** (2014). S100A11 is required for efficient plasma membrane repair and survival of invasive cancer cells. *Nat. Commun.*, **5**, 3795.
- [93] **Jang, I. S., E. C. Neto, J. Guinney, S. H. Friend,** and **A. A. Margolin** (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac. Symp. Biocomput.*, 63–74.
- [94] **Jang, M., S. S. Kim,** and **J. Lee** (2013). Cancer cell metabolism: implications for therapeutic targets. *Experimental & molecular medicine*, **45**(10), e45–e45.
- [95] **Janku, F.** (2014). Tumor heterogeneity in the clinic: is it a real problem? *Therapeutic advances in medical oncology*, **6**(2), 43–51.
- [96] **Jerby-Arnon, L., P. Shah, M. S. Cuoco, C. Rodman, M.-J. Su, J. C. Melms, R. Leeson, A. Kanodia, S. Mei, J.-R. Lin, S. Wang, B. Rabasha, D. Liu, G. Zhang, C. Margolais, O. Ashenberg, P. A. Ott, E. I. Buchbinder, R. Haq,**

- F. S. Hodi, G. M. Boland, R. J. Sullivan, D. T. Frederick, B. Miao, T. Moll, K. T. Flaherty, M. Herlyn, R. W. Jenkins, R. Thummalapalli, M. S. Kowalczyk, I. Cañadas, B. Schilling, A. N. R. Cartwright, A. M. Luoma, S. Malu, P. Hwu, C. Bernatchez, M.-A. Forget, D. A. Barbie, A. K. Shalek, I. Tirosh, P. K. Sorger, K. Wucherpfennig, E. M. Van Allen, D. Schadendorf, B. E. Johnson, A. Rotem, O. Rozenblatt-Rosen, L. A. Garraway, C. H. Yoon, B. Izar, and A. Regev** (2018). A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, **175**(4), 984–997.e24.
- [97] **Jevremovic, D., D. D. Billadeau, R. A. Schoon, C. J. Dick, and P. J. Leibson** (2001). Regulation of NK cell-mediated cytotoxicity by the adaptor protein 3BP2. *J. Immunol.*, **166**(12), 7219–7228.
- [98] **Ji, S., J. R. Doucette, and A. J. Nazarali** (2011). Sirt2 is a novel in vivo downstream target of nkx2. 2 and enhances oligodendroglial cell differentiation. *Journal of molecular cell biology*, **3**(6), 351–359.
- [99] **Ji, Z. and H. Ji** (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, **44**(13), e117.
- [100] **Jia, G., J. Preussner, X. Chen, S. Guenther, X. Yuan, M. Yekelchyk, C. Kuenne, M. Looso, Y. Zhou, S. Teichmann, and T. Braun** (2018). Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. Commun.*, **9**(1), 4877.
- [101] **Jia, H., C. I. Truica, B. Wang, Y. Wang, X. Ren, H. A. Harvey, J. Song, and J.-M. Yang** (2017). Immunotherapy for triple-negative breast cancer: Existing challenges and exciting prospects. *Drug Resist. Updat.*, **32**, 1–15.
- [102] **Jia, P., R. Hu, G. Pei, Y. Dai, Y.-Y. Wang, and Z. Zhao** (2021). Deep generative neural network for accurate drug response imputation. *Nature communications*, **12**(1), 1–16.
- [103] **Jiang, W. G., A. J. Sanders, M. Katoh, H. Ungefroren, F. Gieseler, M. Prince, S. Thompson, M. Zollo, D. Spano, P. Dhawan, et al.**, Tissue invasion and metastasis: Molecular, biological and clinical perspectives. *In Seminars in cancer biology*, volume 35. Elsevier, 2015.
- [104] **Kashima, Y., Y. Sakamoto, K. Kaneko, M. Seki, Y. Suzuki, and A. Suzuki** (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.*, **52**(9), 1419–1427.
- [105] **Kastner, S., T. Voss, S. Keuerleber, C. Glöckel, and others** (2012). Expression of g protein-coupled receptor 19 in human lung cancer cells is triggered by entry into s-phase and supports g2–m cell-cycle progression. *Mol. Cancer*.
- [106] **Kharchenko, P. V., L. Silberstein, and D. T. Scadden** (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**(7), 740–742.
- [107] **Kinker, G. S., A. C. Greenwald, R. Tal, Z. Orlova, M. S. Cuoco, J. M. McFarland, A. Warren, C. Rodman, J. A. Roth, S. A. Bender, B. Kumar, J. W. Rocco, P. A. C. M. Fernandes, C. C. Mader, H. Keren-Shaul, A. Plotnikov,**

- H. Barr, A. Tsherniak, O. Rozenblatt-Rosen, V. Krizhanovsky, S. V. Puram, A. Regev, and I. Tirosh** (2020). Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.*, **52**(11), 1208–1218.
- [108] **Komura, K., C. J. Sweeney, T. Inamoto, N. Ibuki, H. Azuma, and P. W. Kantoff** (2018). Current treatment strategies for advanced prostate cancer. *International Journal of Urology*, **25**(3), 220–231.
- [109] **Krueger, F.** (2015). Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*, **516**, 517.
- [110] **Kuhn, M.** (2015). caret: Classification and regression training.
- [111] **Kuhn, M. and D. Vaughan** (2020). parsnip: A common api to modeling and analysis functions. *R package version 0.0*, **5**.
- [112] **Kukurba, K. R. and S. B. Montgomery** (2015). Rna sequencing and analysis. *Cold Spring Harbor Protocols*, **2015**(11), pdb-top084970.
- [113] **Kumar, M. P., J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. C. Drummond, D. A. Lauffenburger, and A. Raue** (2018). Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics. *Cell reports*, **25**(6), 1458–1468.
- [114] **Kumar, M. P., J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. C. Drummond, D. A. Lauffenburger, and A. Raue** (2018). Analysis of Single-Cell RNA-Seq identifies Cell-Cell communication associated with tumor characteristics. *Cell Rep.*, **25**(6), 1458–1468.e4.
- [115] **Kumar, S.** (2018). Natural killer cell cytotoxicity and its regulation by inhibitory receptors. *Immunology*, **154**(3), 383–393.
- [116] **Kuzumaki, N., A. Suzuki, M. Narita, T. Hosoya, A. Nagasawa, S. Imai, K. Yamamizu, H. Morita, T. Suzuki, Y. Okada, H. J. Okano, J. K. Yamashita, H. Okano, and M. Narita** (2012). Multiple analyses of G-Protein coupled receptor (GPCR) expression in the development of Gefitinib-Resistance in transforming Non-Small-Cell lung cancer.
- [117] **Law, C. W., Y. Chen, W. Shi, and G. K. Smyth** (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**(2), R29.
- [118] **LeDell, E. and S. Poirier**, H2o automl: Scalable automatic machine learning. *In Proceedings of the AutoML Workshop at ICML*, volume 2020. 2020.
- [119] **Lee, M.-C. W., F. J. Lopez-Diaz, S. Y. Khan, M. A. Tariq, Y. Dayn, C. J. Vaske, A. J. Radenbaugh, H. J. Kim, B. M. Emerson, and N. Pourmand** (2014). Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **111**(44), E4726–35.
- [120] **Leek, J. T. and J. D. Storey** (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**(9), 1724–1735.

- [121] **Leng, N.** and **C. Kendzioriski** (2015). Oscope: a statistical pipeline for identifying oscillatory genes in unsynchronized single cell rna-seq experiments. *gene*, **1**(1), 1.
- [122] **Li, B.** and **C. N. Dewey** (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
- [123] **Li, G., M. Yang, L. Zuo,** and **M.-X. Wang** (2018). MELK as a potential target to control cell proliferation in triple-negative breast cancer MDA-MB-231 cells. *Oncol. Lett.*, **15**(6), 9934–9940.
- [124] **Li, H., E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan, M. Wong, P. J. Choi, L. J. K. Wee, A. M. Hillmer, I. B. Tan, P. Robson,** and **S. Prabhakar** (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**(5), 708–718.
- [125] **Li, H., E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan,** *et al.* (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, **49**(5), 708–718.
- [126] **Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh,** and **A. Talwalkar** (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, **18**(1), 6765–6816.
- [127] **Li, X.** and **C.-Y. Wang** (2021). From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, **13**(1), 1–6.
- [128] **Liang, S.-B.** and **L.-W. Fu** (2017). Application of single-cell technology in cancer research. *Biotechnol. Adv.*, **35**(4), 443–449.
- [129] **Liberzon, A., C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov,** and **P. Tamayo** (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**(6), 417–425.
- [130] **Limpert, E., W. A. Stahel,** and **M. Abbt** (2001). Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. *BioScience*, **51**(5), 341–352.
- [131] **Liu, C., D. Wei, J. Xiang, F. Ren, L. Huang, J. Lang, G. Tian, Y. Li,** and **J. Yang** (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Molecular Therapy-Nucleic Acids*, **21**, 676–686.
- [132] **Liu, H., M. Liu, H. You, X. Li,** and **X. Li** (2020). Oncogenic network and hub genes for natural Killer/T-Cell lymphoma utilizing WGCNA. *Front. Oncol.*, **10**, 223.

- [133] **Liu, S., V. Galat, Y. Galat, Y. K. A. Lee, D. Wainwright, and J. Wu** (2021). Nk cell-based cancer immunotherapy: From basic biology to clinical development. *Journal of Hematology & Oncology*, **14**(1), 1–17.
- [134] **Lourenço, T., J. Paes de Faria, C. A. Bippes, J. Maia, J. A. Lopes-da Silva, J. B. Relvas, and M. Grãos** (2016). Modulation of oligodendrocyte differentiation and maturation by combined biochemical and mechanical cues. *Sci. Rep.*, **6**, 21563.
- [135] **Maeda, H. and M. Khatami** (2018). Analyses of repeated failures in cancer therapy for solid tumors: poor tumor-selective drug delivery, low therapeutic efficacy and unsustainable costs. *Clinical and translational medicine*, **7**(1), 1–20.
- [136] **Marco, E., R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan** (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, **111**(52), E5643–E5650.
- [137] **Marinov, G. K., B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold** (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and rna splicing. *Genome research*, **24**(3), 496–510.
- [138] **Martelotto, L. G., C. K. Ng, S. Piscuoglio, B. Weigelt, and J. S. Reis-Filho** (2014). Breast cancer intra-tumor heterogeneity. *Breast Cancer Research*, **16**(3), 1–11.
- [139] **Mateo, J., R. McKay, W. Abida, R. Aggarwal, J. Alumkal, A. Alva, F. Feng, X. Gao, J. Graff, M. Hussain, F. Karzai, B. Montgomery, W. Oh, V. Patel, D. Rathkopf, M. Rettig, N. Schultz, M. Smith, D. Solit, C. Sternberg, E. Van Allen, D. VanderWeele, J. Vinson, H. R. Soule, A. Chinnaiyan, E. Small, J. W. Simons, W. Dahut, A. K. Miyahira, and H. Beltran** (2020). Accelerating precision medicine in metastatic prostate cancer. *Nat Cancer*, **1**(11), 1041–1053.
- [140] **McGranahan, N. and C. Swanton** (2017). Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, **168**(4), 613–628.
- [141] **Melaiu, O., V. Lucarini, L. Cifaldi, and D. Fruci** (2020). Influence of the tumor microenvironment on nk cell function in solid tumors. *Frontiers in immunology*, **10**, 3038.
- [142] **Morita, K., F. Wang, K. Jahn, T. Hu, T. Tanaka, Y. Sasaki, J. Kuipers, S. Loghavi, S. A. Wang, Y. Yan, K. Furudate, J. Matthews, L. Little, C. Gumbs, J. Zhang, X. Song, E. Thompson, K. P. Patel, C. E. Bueso-Ramos, C. D. DiNardo, F. Ravandi, E. Jabbour, M. Andreeff, J. Cortes, K. Bhalla, G. Garcia-Manero, H. Kantarjian, M. Konopleva, D. Nakada, N. Navin, N. Beerenwinkel, P. A. Futreal, and K. Takahashi** (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.*, **11**(1), 5327.

- [143] **Mukherjee, S., Y. Zhang, J. Fan, G. Seelig, and S. Kannan** (2018). Scalable preprocessing for sparse scRNA-seq data exploiting prior knowledge. *Bioinformatics*, **34**(13), i124–i132.
- [144] **Nasri, H. and M. Rafieian-Kopaei** (2014). Metformin: current knowledge. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, **19**(7), 658.
- [145] **Navin, N. E.** (2015). The first five years of single-cell cancer genomics and beyond. *Genome Res.*, **25**(10), 1499–1507.
- [146] **Nevedomskaya, E., S. J. Baumgart, and B. Haendler** (2018). Recent advances in prostate cancer treatment and drug discovery. *Int. J. Mol. Sci.*, **19**(5).
- [147] **Ni, F., R. Sun, B. Fu, F. Wang, C. Guo, Z. Tian, and H. Wei** (2013). IGF-1 promotes the development and cytotoxic activity of human NK cells. *Nat. Commun.*, **4**, 1479.
- [148] **Nishida-Aoki, N. and T. S. Gujral** (2019). Emerging approaches to study cell-cell interactions in tumor microenvironment. *Oncotarget*, **10**(7), 785–797.
- [149] **O’Malley, T., E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, et al.** (2019). Kerastuner. <https://github.com/keras-team/keras-tuner>.
- [150] **Omarini, C., G. Guaitoli, S. Pipitone, L. Moscetti, L. Cortesi, S. Cascinu, and F. Piacentini** (2018). Neoadjuvant treatments in triple-negative breast cancer patients: where we are now and where we are going. *Cancer management and research*, **10**, 91.
- [151] **Orange, J. S.** (2008). Formation and function of the lytic NK-cell immunological synapse. *Nat. Rev. Immunol.*, **8**(9), 713–725.
- [152] **Oyama, T., K. Sugio, H. Uramoto, T. Onizuka, T. Iwata, T. Nozoe, M. Takenoyama, T. Hanagiri, T. Isse, T. Kawamoto, and K. Yasumoto** (2007). P2-049: Cytochrome P450 expression in non-small cell lung cancer. *J. Thorac. Oncol.*, **2**(8), S509–S510.
- [153] **Öztürk, H., E. Ozkirimli, and A. Özgür** (2018). A novel methodology on distributed representations of proteins using their interacting ligands. *Bioinformatics*, **34**(13), i295–i303.
- [154] **Patel, A. P., I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein** (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**(6190), 1396–1401.
- [155] **Pei, Z., Z. Jia, and P. A. Watkins** (2006). The second member of the human and murine “bubblegum” family is a testis- and brainstem-specific acyl-coa synthetase. *Journal of Biological Chemistry*, **281**(10), 6632–6641.
- [156] **Phipson, B., L. Zappia, and A. Oshlack** (2017). Gene length and detection bias in single cell rna sequencing protocols. *F1000Research*, **6**.



- [157] **Poole, W., D. L. Gibbs, I. Shmulevich, B. Bernard, and T. A. Knijnenburg** (2016). Combining dependent p-values with an empirical adaptation of brown's method.
- [158] **Poonia, S., A. Goel, S. Chawla, N. Bhattacharya, P. Rai, Y. F. Lee, Y. S. Yap, J. West, A. A. Bhagat, J. Tayal, et al.** (2021). Marker-free characterization of single live circulating tumor cell full-length transcriptomes. *bioRxiv*.
- [159] **Pucci, C., C. Martinelli, and G. Ciofani** (2019). Innovative approaches for cancer treatment: current perspectives and new challenges. *Ecancermedicalscience*, **13**, 961.
- [160] **Qiu, S., R. Hong, Z. Zhuang, Y. Li, L. Zhu, X. Lin, Q. Zheng, and others** (2019). A single-cell immune atlas of triple negative breast cancer reveals novel immune cell subsets. *bioRxiv*.
- [161] **Rajamahanty, S., C. Alonzo, S. Aynehchi, M. Choudhury, and S. Konno** (2010). Growth inhibition of androgen-responsive prostate cancer cells with brefeldin a targeting cell cycle and androgen receptor. *J. Biomed. Sci.*, **17**, 5.
- [162] **Ramalingam, N., B. Fowler, L. Szpankowski, A. A. Leyrat, K. Hukari, M. T. Maung, W. Yorza, M. Norris, C. Cesar, J. Shuga, M. L. Gonzales, C. D. Sanada, X. Wang, R. Yeung, W. Hwang, J. Axsom, N. S. G. Devaraju, N. D. Angeles, C. Greene, M.-F. Zhou, E.-S. Ong, C.-C. Poh, M. Lam, H. Choi, Z. Htoo, L. Lee, C.-S. Chin, Z.-W. Shen, C. T. Lu, I. Holcomb, A. Ooi, C. Stolarczyk, T. Shuga, K. J. Livak, C. Larsen, M. Unger, and J. A. A. West** (2017). Corrigendum: Fluidic logic used in a systems approach to enable integrated Single-Cell functional analysis.
- [163] **Raulf, N., P. Lucarelli, S. Thavaraj, S. Brown, J. M. Vicencio, T. Sauter, and M. Tavassoli** (2018). Annexin A1 regulates EGFR activity and alters EGFR-containing tumour-derived exosomes in head and neck cancers. *Eur. J. Cancer*, **102**, 52–68.
- [164] **Ren, X., B. Kang, and Z. Zhang** (2018). Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol.*, **19**(1), 211.
- [165] **Riddick, G., H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine** (2011). Predicting in vitro drug sensitivity using random forests. *Bioinformatics*, **27**(2), 220–224.
- [166] **Risso, D., K. Schwartz, G. Sherlock, and S. Dudoit** (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 480.
- [167] **Saadatpour, A., S. Lai, G. Guo, and G.-C. Yuan** (2015). Single-Cell analysis in cancer genomics. *Trends Genet.*, **31**(10), 576–586.
- [168] **Saelens, W., R. Cannoodt, H. Todorov, and Y. Saeys** (2019). A comparison of single-cell trajectory inference methods. *Nature biotechnology*, **37**(5), 547–554.
- [169] **Saelens, W., R. Cannoodt, H. Todorov, and Y. Saeys** (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**(5), 547–554.

- [170] **Sakellaropoulos, T., K. Vougas, S. Narang, F. Koinis, A. Kotsinas, A. Polyzos, T. J. Moss, S. Piha-Paul, H. Zhou, E. Kardala, et al.** (2019). A deep learning framework for predicting response to therapy in cancer. *Cell reports*, **29**(11), 3367–3373.
- [171] **Sanada, C. D. and A. T. Ooi** (2019). Single-Cell dosing and mRNA sequencing of suspension and adherent cells using the Polaris™ system. *Methods Mol. Biol.*, **1979**, 185–195.
- [172] **Santuario-Facio, S. K., S. Cardona-Huerta, Y. X. Perez-Paramo, V. Trevino, F. Hernandez-Cabrera, A. Rojas-Martinez, G. Uscanga-Perales, J. L. Martinez-Rodriguez, L. Martinez-Jacobo, G. Padilla-Rivas, et al.** (2017). A new gene expression signature for triple-negative breast cancer using frozen fresh tissue before neoadjuvant chemotherapy. *Molecular Medicine*, **23**(1), 101–111.
- [173] **Satija, R., J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev** (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**(5), 495–502.
- [174] **Schep, A. N., B. Wu, J. D. Buenrostro, and W. J. Greenleaf** (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, **14**(10), 975–978.
- [175] **Schneider, C. A., W. S. Rasband, and K. W. Eliceiri** (2012). NIH image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**(7), 671–675.
- [176] **Schubert, M., B. Klinger, M. Klünemann, A. Sieber, F. Uhlitz, S. Sauer, M. J. Garnett, N. Blüthgen, and J. Saez-Rodriguez** (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.*, **9**(1), 20.
- [177] **Seashore-Ludlow, B., M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, et al.** (2015). Harnessing connectivity in a large-scale small-molecule sensitivity dataset harnessing connectivity in a sensitivity dataset. *Cancer discovery*, **5**(11), 1210–1223.
- [178] **Sever, R. and J. S. Brugge** (2015). Signal transduction in cancer. *Cold Spring Harbor perspectives in medicine*, **5**(4), a006098.
- [179] **Sharma, G. N., R. Dave, J. Sanadya, P. Sharma, and K. K. Sharma** (2010). Various types and management of breast cancer: an overview. *J. Adv. Pharm. Technol. Res.*, **1**(2), 109–126.
- [180] **Shi, X., P. Chakraborty, and A. Chaudhuri** (2018). Unmasking tumor heterogeneity and clonal evolution by single-cell analysis. *Journal of Cancer Metastasis*.
- [181] **Sinha, D., P. Sinha, R. Saha, S. Bandyopadhyay, and D. Sengupta** (2019). Improved dropclust R package with integrative analysis support for scRNA-seq data. *Bioinformatics*.
- [182] **Smillie, C. S., M. Biton, J. Ordovas-Montanes, K. M. Sullivan, G. Burgin, D. B. Graham, R. H. Herbst, N. Rogel, M. Slyper, J. Waldman, et al.** (2019). Intra-and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, **178**(3), 714–730.

- [183] **Stanta, G.** and **S. Bonin** (2018). Overview on clinical relevance of intra-tumor heterogeneity. *Frontiers in medicine*, **5**, 85.
- [184] **Stark, R., M. Grzelak,** and **J. Hadfield** (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, **20**(11), 631–656.
- [185] **Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert,** and **R. Satija** (2019). Comprehensive integration of single-cell data. *Cell*, **177**(7), 1888–1902.
- [186] **Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander,** and **J. P. Mesirov** (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**(43), 15545–15550.
- [187] **Suphavilai, C., D. Bertrand,** and **N. Nagarajan** (2018). Predicting cancer drug response using a recommender system. *Bioinformatics*, **34**(22), 3907–3914.
- [188] **Suphavilai, C., S. Chia, A. Sharma, L. Tu, R. P. Da Silva, A. Mongia, R. Das-Gupta,** and **N. Nagarajan** (2021). Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Medicine*, **13**(1), 1–14.
- [189] **Swain, M.** (2014). PubChemPy: A way to interact with PubChem in python.
- [190] **Swami, U., T. R. McFarland, R. Nussenzweig,** and **N. Agarwal** (2020). Advanced prostate cancer: treatment advances and future directions. *Trends in Cancer*, **6**(8), 702–715.
- [191] **Teo, M. Y., D. E. Rathkopf,** and **P. Kantoff** (2019). Treatment of advanced prostate cancer. *Annual review of medicine*, **70**, 479–499.
- [192] **Thomsen, E. R., J. K. Mich, Z. Yao, R. D. Hodge, A. M. Doyle, S. Jang, S. I. Shehata, A. M. Nelson, N. V. Shapovalova, B. P. Levi,** *et al.* (2016). Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nature methods*, **13**(1), 87–93.
- [193] **Tian, Y., C. Wang, S. Chen, J. Liu, Y. Fu,** and **Y. Luo** (2019). Extracellular Hsp90 $\alpha$  and clusterin synergistically promote breast cancer epithelial-to-mesenchymal transition and metastasis via LRP1. *J. Cell Sci.*, **132**(15).
- [194] **Tran, T. N., K. Drab,** and **M. Daszykowski** (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics Intellig. Lab. Syst.*, **120**, 92–96.
- [195] **Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen,** and **J. L. Rinn** (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**(4), 381–386.
- [196] **Tripathi, S., M. Dehmer,** and **F. Emmert-Streib** (2014). NetBioV: an R package for visualizing large network data in biology and medicine. *Bioinformatics*, **30**(19), 2834–2836.

- [197] **Van der Maaten, L. and G. Hinton** (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(11).
- [198] **Vinay, D. S. and B. S. Kwon** (2011). 4-1bb signaling beyond t cells. *Cellular & molecular immunology*, **8**(4), 281–284.
- [199] **Wagle, N., E. M. Van Allen, D. J. Treacy, D. T. Frederick, Z. A. Cooper, A. Taylor-Weiner, M. Rosenberg, E. M. Goetz, R. J. Sullivan, D. N. Farlow, D. C. Friedrich, K. Anderka, D. Perrin, C. M. Johannessen, A. McKenna, K. Cibulskis, G. Kryukov, E. Hodis, D. P. Lawrence, S. Fisher, G. Getz, S. B. Gabriel, S. L. Carter, K. T. Flaherty, J. A. Wargo, and L. A. Garraway** (2014). MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. *Cancer Discov.*, **4**(1), 61–68.
- [200] **Wan, M., W. Huang, T. E. Kute, L. D. Miller, Q. Zhang, H. Hatcher, J. Wang, D. B. Stovall, G. B. Russell, P. D. Cao, Z. Deng, W. Wang, Q. Zhang, M. Lei, S. V. Torti, S. A. Akman, and G. Sui** (2012). Yin yang 1 plays an essential role in breast cancer and negatively regulates p27. *Am. J. Pathol.*, **180**(5), 2120–2133.
- [201] **Wang, L., X. Li, L. Zhang, and Q. Gao** (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer*, **17**(1), 1–12.
- [202] **Wang, M., J. Zhao, L. Zhang, F. Wei, Y. Lian, Y. Wu, Z. Gong, S. Zhang, J. Zhou, K. Cao, X. Li, W. Xiong, G. Li, Z. Zeng, and C. Guo** (2017). Role of tumor microenvironment in tumorigenesis. *J. Cancer*, **8**(5), 761–773.
- [203] **Wang, M., J. Zhao, L. Zhang, F. Wei, Y. Lian, Y. Wu, Z. Gong, S. Zhang, J. Zhou, K. Cao, et al.** (2017). Role of tumor microenvironment in tumorigenesis. *Journal of Cancer*, **8**(5), 761.
- [204] **Wang, Q., M. Xu, Y. Sun, J. Chen, C. Chen, C. Qian, Y. Chen, L. Cao, Q. Xu, X. Du, et al.** (2019). Gene expression profiling for diagnosis of triple-negative breast cancer: a multicenter, retrospective cohort study. *Frontiers in Oncology*, **9**, 354.
- [205] **Wang, Y., R. Wang, S. Zhang, S. Song, C. Jiang, G. Han, M. Wang, J. Ajani, A. Futreal, and L. Wang** (). iTALK: an R package to characterize and illustrate intercellular communication.
- [206] **Wang, Z., L. Y. Yip, J. H. J. Lee, Z. Wu, H. Y. Chew, P. K. W. Chong, C. C. Teo, H. Y.-K. Ang, K. L. E. Peh, J. Yuan, S. Ma, L. S. K. Choo, N. Basri, X. Jiang, Q. Yu, A. M. Hillmer, W. T. Lim, T. K. H. Lim, A. Takano, E. H. Tan, D. S. W. Tan, Y. S. Ho, B. Lim, and W. L. Tam** (2019). Methionine is a metabolic dependency of tumor-initiating cells. *Nat. Med.*, **25**(5), 825–837.
- [207] **Whiteside, T. L.** (2008). The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, **27**(45), 5904–5912.
- [208] **Wills, Q. F., E. Mellado-Gomez, R. Nolan, D. Warner, E. Sharma, J. Broxholme, B. Wright, H. Lockstone, W. James, M. Lynch, M. Gonzales, J. West, A. Leyrat, S. Padilla-Parra, S. Filippi, C. Holmes, M. D. Moore, and R. Bowden** (2017). The nature and nurture of cell heterogeneity: accounting for

- macrophage gene-environment interactions with single-cell RNA-Seq. *BMC Genomics*, **18**(1), 53.
- [209] **Wolf, F. A., P. Angerer, and F. J. Theis** (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**(1), 15.
- [210] **Wright, M. N. and A. Ziegler** (2015). ranger: A fast implementation of random forests for high dimensional data in c++ and R.
- [211] **Wu, S., H. Zhang, S. Fouladdel, H. Li, E. Keller, M. S. Wicha, G. S. Omenn, E. Azizi, and Y. Guan** (2020). Cellular, transcriptomic and isoform heterogeneity of breast cancer cell line revealed by full-length single-cell RNA sequencing. *Comput. Struct. Biotechnol. J.*, **18**, 676–685.
- [212] **Wu, S.-Y., T. Fu, Y.-Z. Jiang, and Z.-M. Shao** (2020). Natural killer cells in cancer biology and therapy. *Molecular Cancer*, **19**(1), 1–26.
- [213] **Wu, S.-Y., T. Fu, Y.-Z. Jiang, and Z.-M. Shao** (2020). Natural killer cells in cancer biology and therapy. *Mol. Cancer*, **19**(1), 120.
- [214] **Xia, C., J. Fan, G. Emanuel, J. Hao, and X. Zhuang** (2019). Spatial transcriptome profiling by merfish reveals subcellular rna compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, **116**(39), 19490–19499.
- [215] **Xia, F., J. Allen, P. Balaprakash, T. Brettin, C. Garcia-Cardona, A. Clyde, J. Cohn, J. Doroshow, X. Duan, V. Dubinkina, et al.** (2021). A cross-study analysis of drug response prediction in cancer cell lines. *arXiv preprint arXiv:2104.08961*.
- [216] **Xiong, L., K. Xu, K. Tian, Y. Shao, L. Tang, G. Gao, M. Zhang, T. Jiang, and Q. C. Zhang** (2019). SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.*, **10**(1), 4576.
- [217] **Xu, H., Y. Jiao, M. Yi, W. Zhao, and K. Wu** (2019). EYA2 correlates with Clinico-Pathological features of breast cancer, promotes tumor proliferation, and predicts poor survival. *Front. Oncol.*, **9**, 26.
- [218] **y Cajal, S. R., M. Sesé, C. Capdevila, T. Aasen, L. De Mattos-Arruda, S. J. Diaz-Cano, J. Hernández-Losa, and J. Castellví** (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine*, **98**(2), 161–177.
- [219] **Yan, L., M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, et al.** (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, **20**(9), 1131–1139.
- [220] **Yang, W., J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett** (2013). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**(Database issue), D955–61.

- [221] **Yang, W., J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, et al.** (2012). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, **41**(D1), D955–D961.
- [222] **Yildizhan, H., N. P. Barkan, S. Karahisar Turan, Ö. Demiralp, F. D. Özel Demiralp, B. Uslu, and S. A. Özkan**, Chapter 1 - treatment strategies in cancer from past to present. In **A. M. Grumezescu** (ed.), *Drug Targeting and Stimuli Sensitive Drug Delivery Systems*. William Andrew Publishing, 2018, 1–37.
- [223] **Yip, H. Y. K. and A. Papa** (2021). Signaling pathways in cancer: Therapeutic targets, combinatorial treatments, and new developments. *Cells*, **10**(3), 659.
- [224] **Yu, K., L. Toral-Barza, C. Shi, W.-G. Zhang, and A. Zask** (2008). Response and determinants of cancer cell susceptibility to PI3K inhibitors: combined targeting of PI3K and mek1 as an effective anticancer strategy. *Cancer Biol. Ther.*, **7**(2), 307–315.
- [225] **Yuan, T. and L. Cantley** (2008). Pi3k pathway alterations in cancer: variations on a theme. *Oncogene*, **27**(41), 5497–5510.
- [226] **Zhan, T., N. Rindtorff, and M. Boutros** (2017). Wnt signaling in cancer. *Oncogene*, **36**(11), 1461–1473.
- [227] **Zhang, B., J. Wang, Z. Huang, P. Wei, Y. Liu, J. Hao, L. Zhao, F. Zhang, Y. Tu, and T. Wei** (2015). Aberrantly upregulated TRAP1 is required for tumorigenesis of breast cancer. *Oncotarget*, **6**(42), 44495–44508.
- [228] **Zhang, P., X. Lu, K. Tao, L. Shi, W. Li, G. Wang, and K. Wu** (2015). Siglec-10 is associated with survival and natural killer cell dysfunction in hepatocellular carcinoma. *J. Surg. Res.*, **194**(1), 107–113.
- [229] **Zhang, X., Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, L. Xu, G. Liao, M. Yan, et al.** (2019). Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, **47**(D1), D721–D728.
- [230] **Zhang, Y., Y. Ma, Y. Huang, Y. Zhang, Q. Jiang, M. Zhou, and J. Su** (2020). Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.*, **18**, 2953–2961.
- [231] **Zhang, Y., Y. Zhang, J. Hu, J. Zhang, F. Guo, M. Zhou, G. Zhang, F. Yu, and J. Su** (2020). scTPA: a web tool for single-cell transcriptome analysis of pathway activation signatures. *Bioinformatics*, **36**(14), 4217–4219.
- [232] **Zheng, C., L. Zheng, J.-K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J. Y. Huang, Q. Zhang, et al.** (2017). Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing. *Cell*, **169**(7), 1342–1356.
- [233] **Zheng, G. X., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al.** (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**(1), 1–12.

- [234] **Zhou, S., Y.-E. Huang, H. Liu, X. Zhou, M. Yuan, F. Hou, L. Wang, and W. Jiang** (2021). Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks. *Mol. Ther. Nucleic Acids*, **23**, 682–690.
- [235] **Zhu, Y., B. Huang, and J. Shi** (2016). Fas ligand and lytic granule differentially control cytotoxic dynamics of natural killer cell against cancer target. *Oncotarget*, **7**(30), 47163–47172.