Ph.D. degree in Systems Medicine

Curriculum in Human Genetics

European School of Molecular Medicine (SEMM)

University of Milan and University of Naples “Federico II”

Disciplinary sector: BIO/11

Integrative Genomics Unveils Transcription Factor Roles in Cellular Fate and Spatially Resolved Triple-Negative Breast Cancer Immunity Capacity

Shaked Slovin

Telethon Institute of Genetics and Medicine (TIGEM)

University ID No.: R12713

Supervisor: Prof. Davide Cacchiarelli

Internal advisor: Prof. Gennaro Gambardella

External advisor: Prof. Piero Carninci

Internal examiner: Prof. Leopoldo Staiano

External examiner: Prof. Matteo Cereda

Academic Year 2022-2023

“Tis not in mortals to command success, but we'll do more, Sempronius; we'll deserve it.”

Addison, J. (1713). Robert Falcon Scott quoted these lines in his diary during his ill-fated Terra Nova Expedition (1910-1913) to the South Pole.

למשפחה שלי, אמא, אבא, וסתיו שהייתם לצידי לאורך הדרך. אוהבת ומעריכה.

Table of Contents

Abstract

In this thesis, we delve into the realms of integrative genomics, employing advanced sequencing techniques across two distinct but interconnected projects. Integrative genomics combines transcriptomic, genomic, epigenomic, and proteomic data, offering a comprehensive perspective on the interplay within and between different functional layers to fully underscore biological systems and processes. This approach is key to understanding and elucidating the complex molecular mechanisms underlying the project presented in this dissertation.

An Integrated Screening to Infer Transcription Factor Regulatory Networks Governing Cell Fate Decisions

Recent advancements in cellular reprogramming have revolutionized our understanding of cellular differentiation, highlighting the crucial role of transcription factors (TFs) in shaping cellular fate. Despite these breakthroughs, current cellular conversion strategies remain inefficient, often yielding immature cellular phenotypes. This is primarily due to the complex regulatory landscape of TF activity, involving numerous constraints that heavily affect its modus operandi. Notably, a critical challenge in the field is the need for a systematic and comparative workflow for concurrently surveying TFs across multiple cellular systems, processes, and conditions to fully unlock their potential. This shortfall significantly impedes the comprehensive understanding of TFs' regulatory capabilities and the transcriptional and epigenetic barriers that modulate their activities.

In this context, we hypothesize that various TFs, yet uncharacterized, play crucial roles in cellular fate determination. In this project, we developed a comprehensive transcriptomic, epigenomic, and morphological screening of 130 developmental TFs to appraise their effect on cellular transdifferentiation. Our approach represents one, if not the only, case of side-by-side comparison of TF dosages within the same experimental setting. Essentially, by analyzing well-established and yet uncharacterized TFs, we identify novel pioneer factors that, either individually or in combination, play a pivotal role in regulating cellular identity. We anticipate that this strategy will ultimately lead to novel paradigms in deciphering crucial dynamics driving cell-fate decisions, which potentially allow us to unlock the full potential of the pioneer TF repertoire.

Spatial Transcriptomics Reveals Sub-Tumoral Identities and Novel Diagnostic markers in Triple Negative Breast Cancer with Immune Evasion Capacity

Triple-negative breast cancer (TNBC) is marked by its aggressiveness and inherent heterogeneity among diverse cell populations, presenting significant challenges to ineffective treatment. Despite extensive research in the field, efficient and comprehensive therapy for TNBC remains elusive. A critical aspect of managing TNBC involves an accurate histological diagnosis of the immune checkpoint protein, programmed death-ligand 1 (PD-L1), which is essential for the effectiveness of targeted immunotherapies. However, the lack of standardization across different PD-L1 diagnostic tests leads to inconsistent and often incomplete therapeutic outcomes. Nevertheless, recent advancements in spatial transcriptomics technologies have opened new opportunities to spatially resolve crucial aspects of tumor heterogeneity and, foremost, identify more novel diagnostic markers.

In light of these challenges, we hypothesize that integrating spatial context into the study of TNBC will significantly deepen our understanding of its intricate tumor architecture. This approach may reveal aspects often overlooked by traditional histological analysis, thereby facilitating the discovery of novel therapeutic diagnostic markers.

Therefore, in this study, we developed a cost-effective clinical workflow combining spatial transcriptomics, clinical-grade RNA sequencing (RNAseq), and Immunohistochemistry (IHC) to analyze TNBC at a spatial level focusing on PD-L1 status. Essentially, this method allowed us to explore sub-tumoral variations, understand tumor-microenvironment interactions, and identify LY6D as a novel diagnostic marker potentially complementing conventional PD-L1 test. Our findings underscore the importance of spatial transcriptomics in advancing personalized medicine for TNBC, offering new therapeutic avenues.

Thesis Outline

This thesis highlights the varied applications of integrative genomics through two projects, using advanced sequencing techniques. The first project examines transcription factors (TFs) activity in shaping cellular identity, employing Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq) and RNA sequencing (RNAseq) for a novel screening approach. The second one investigates triple-negative breast cancer (TNBC) using spatial transcriptomics and RNAseq, focusing on the role of Programmed Death-Ligand 1 (PD-L1) in tumor immunity. This study is currently under revision in the Journal of 'Modern Pathology'.

Each chapter in this thesis is divided into two sections dedicated to a specific project.

In the introduction (Chapter 1), we briefly review the scientific background and the current state-of-the-art. In the first project, we provide an overview of the essential properties of TFs, different approaches to studying them, and the forefront advances and challenges in shaping cellular identity in vitro. Subsequently, in the second project, we highlight the aggressive nature of TNBC, the role of PD-L1 in immune checkpoints, and spatial transcriptomics as a cutting-edge technology to decipher its highly heterogeneous landscape.

The materials and methods chapter (Chapter 2) then provides a detailed description of the sequencing techniques and computational strategies we developed. This includes the experimental workflows and approaches for data interpretation, tailored to the objectives of each project.

Then, in the results chapter (Chapter 3), the findings of the two projects are presented and analyzed in detail. The first project presents the outcomes of our novel screening approach of both known and uncharacterized TFs. The results highlight their effects on transcriptional, epigenetic, and phenotypic cellular levels during transdifferentiation and infer their functional interconnectivity. Subsequently, the discoveries of the second project centered on the use of spatial transcriptomics in TNBC, particularly in resolving tumor architecture, crosstalk dynamic between tumor and microenvironment, and identifying novel diagnostic markers.

The thesis concludes with a final discussion chapter (chapter 4) that presents the insights from both projects. The discussion for the first project delves into the implications of our TF study to understand their global role in re-wiring cellular identity. For the second project, we demonstrate that our TNBC study significantly contributes to a better understanding of tumor biology and identifies LY6D as a potential alternative diagnostic marker to PD-L1.

Overall, the insights gained from these studies illustrate the transformative potential of integrative genomics in addressing intricate biological questions and advancing cancer research.

Figures Index

1. Introduction

1.1. First Project - An Integrated Screening to Infer Transcription Factor Regulatory Networks Governing Cell Fate Decisions

1.1.1. Essential Properties of Transcription Factors

The term ‘transcription factors’ (TFs) is used to describe proteins that bind to DNA sequences and subsequently influence gene regulation[1,2]. Overall, TFs are categorized into two groups, ‘general TFs’ (GTFs) and ‘specific TFs’. GTFs identify core promoters and recruit RNA Polymerase II (Pol II), thereby forming the pre-initiation complex (PIC), which is essential for initiating transcription machinery[3] (Figure 1). Conversely, ‘specific TFs’ bind to unique genomic regulatory regions and modulate gene expression in a cell-type-specific manner[4] (Figure 1). Due to their restricted expression and targeted regulatory roles, ‘specific TFs’ are more likely to regulate cell fate identity than GTFs, which are ubiquitously expressed. Hence, in this thesis, the term 'TFs' exclusively denotes ‘specific TF’.

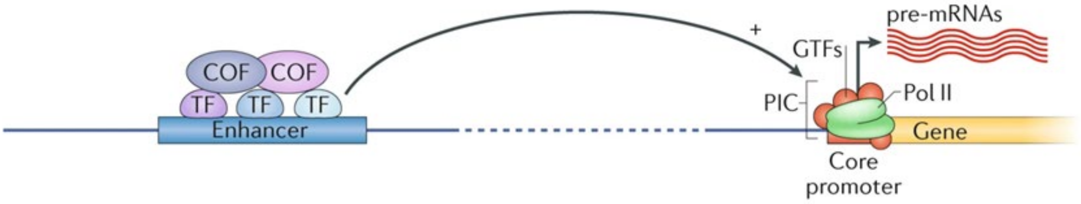


Figure 1. General and Specific Transcription Factors Role in Gene Expression Machinery. Schematic representation of the gene expression machinery. GTFs target core promoters to form the PIC, which initiates the transcription. Specific TFs uniquely bind to regulatory elements like distally located enhancers and recruit coactivators that transmit activation signals to the PIC at the core promoter. Pre-Initiation Complex (PIC); General transcription factors (GTFs); transcription factors (TFs). Adapted from Spitz & Furlong, Nature Reviews Genetics (2012)[4].

Extensive studies have established the integral role of TFs in re-wiring cellular identity[5,6], with a considerable majority being distinctly expressed across varied types of cells[2,7]. Through examining the functional multiplicity of TFs, this thesis underscores their importance in the regulatory networks orchestrating cellular identity. Subsequent sections will further explore the molecular mechanisms governing TF binding specificity, the regulatory variables impacting TF activity, and the far-reaching applications of their dynamic interplay in shaping cellular identity in vitro.

1.1.2. DNA Binding and Specificity

TFs have a modular structure, with a specific domain for DNA binding (DBD), through which they recognize and bind to definite consensus sequences on the genome, often referred to as 'motifs'[8]. Motif sequences are primarily organized into enhancer clusters, which are groups of binding sites situated far from the promoters they regulate, often tens to hundreds of kilobases apart. Only a minority of these motifs are found adjacent to the promoter regions[9]. While motifs contain both fixed and variable bases, a given TF can bind to these genomic sequences without necessarily changing its binding affinity[10]. However, the mere presence of TF binding motifs in a stretch of DNA is a poor predictor of its regulatory activity. This is primarily due to the various constraints that influence TF functionality[6]. For instance, chromatin structure imposes profound and ubiquitous effects on almost all DNA- related processes, serving as a primary determinant of TF occupancy and activity. Moreover, TFs frequently act in conjunction with other factors to fulfill their function. Also, TFs can be regulated by post-transcriptional mechanisms, which influence their activation state and their subcellular localization[6] (as detailed in section 1.1.3).

Upon DNA binding, TFs demonstrate significant variability in their effects on transcriptional dynamics (Figure 2). Certain TFs have the capability to directly engage RNA polymerase, while others recruit accessory factors, resulting in either transcriptional activation or repression[11]. Such coactivators and corepressors function as intermediary effectors, working in conjunction with TFs to regulate gene expression through various mechanisms. Typically, TFs comprise varied structural domains designated for chromatin engagement, \nucleosome restructuring, and the covalent adjustment of various proteins, including histones, other TFs, and RNA polymerase.

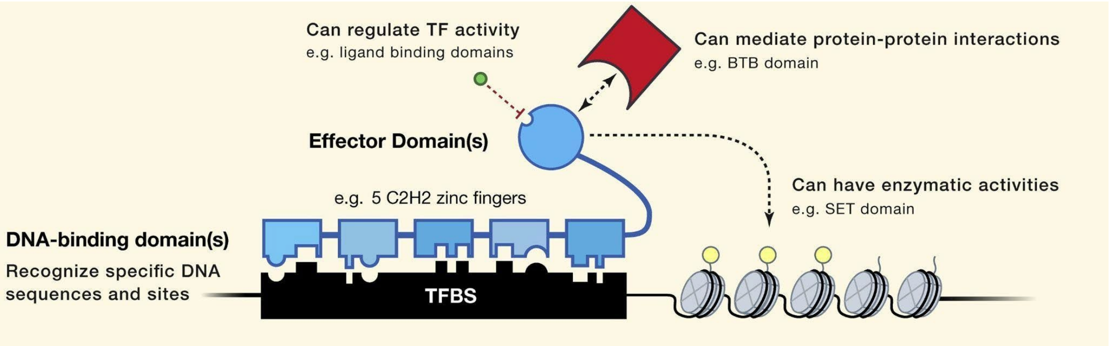

Can regulate TF activity
e.g. ligand binding domains
Can mediate protein-protein interactions
e.g. BTB domain
Effector Domain(s)
Can have enzymatic activities
e.g. SET domain
e.g. 5 C2H2 zinc fingers
DNA-binding domain(s)
Recognize specific DNA sequences and sites
TFBS

Figure 2. Transcription Factor Functionality Constrained by Diverse Regulatory Mechanisms representation of TF functionality. TF binds to specific sites on DNA, known as TFBS, using their, such as C2H2-ZF. The functional activity of a TF can be influenced by various effectors, including ligand binding, cooperative interactions with other proteins, and enzymatic modifications that can reshape the chromatin state. Transcription Factor (TF); TF binding sites (TFBS); DNA binding domains (DBD). Adapted from Lambert, Samuel A et al., Cell (2018)[6].

A recent comprehensive study has unravelled the DNA binding specificities of over 1,600 TFs across both mouse and human genomes[12]. These TFs have been systematically categorized into ten distinct superclasses based on their DBDs topology. Nine of these superclasses have well-defined characteristics, while the tenth consists of TFs with unresolved 3D structures and those that lack significant sequence similarity to other known

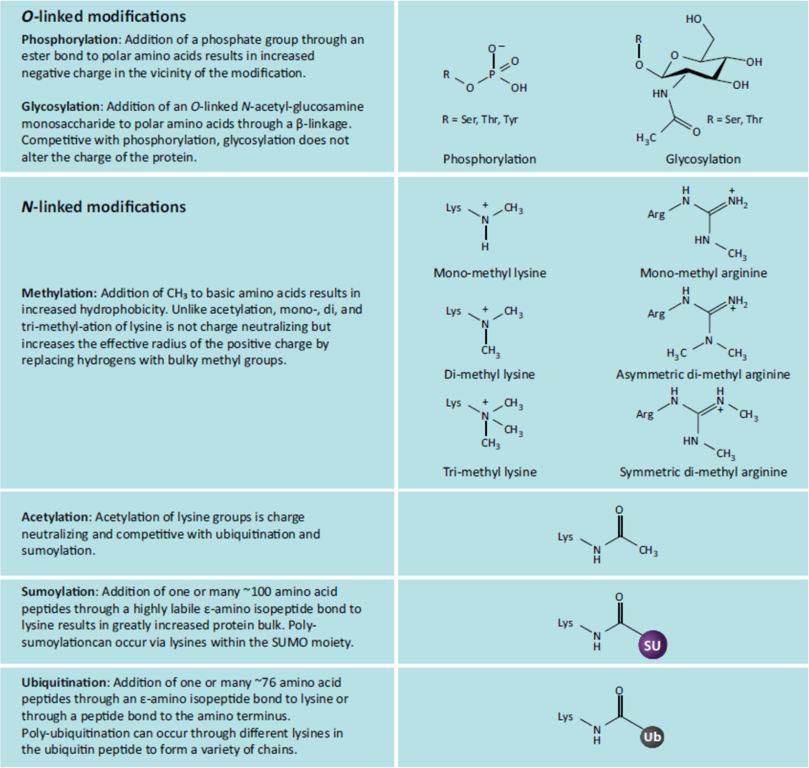
TFs. Such categorizations suggest that numerous TFs may exhibit shared DNA binding specificities due to analogous structural configurations of their DBDs. Furthermore, of the 1,639 human TFs, 93% are known or expected to associate with DNA through monomeric or homomultimeric interactions. While most TFs include repetitive units of an identical DBD type, only a minority, approximately 3%, incorporate multiple types of DBDs[6]. Overall, the varied binding properties of different TFs establish them as central components in achieving specificity for transcriptional regulation. This specificity allows TFs to regulate gene expression only under conditional genomic contexts, thus making them essential in guiding the process of cellular differentiation across various tissue types.

1.1.3. Regulation of Transcription Factor Activity

The diverse array of effectors influencing TF activity presents a significant challenge for studies attempting to comprehensively analyze TF functionality. A thorough investigation into TFs requires an evaluation of their DNA binding affinity, the nature of their interactions with other molecular entities, their distribution across various tissue types, and the extent to which external conditions modulate their functional capacity. Therefore, the complex interplay of these regulatory determinants increases the likelihood of overlooking certain aspects of TF functionality. In the subsequent sections, we will overview the various constraints that may impact TF activity.

1.1.3.1. Transcription Factors Post-Translational Modification

TFs, like many other protein classes, are subject to an array of post-translational modifications (PTMs). These PTMs occur after protein biosynthesis (translation) and involve the covalent addition of functional groups, cleavage of regulatory subunits, or the degradation of entire proteins. These modifications play a crucial role in orchestrating TFs functionality and binding capacity through a spectrum of underlying mechanisms. These include subcellular localization, protein-protein interactions, specificity to DNA sequences, transcription regulation, protein stability, and varied epigenetic regulations[13]. Among the varied types of PTMs, phosphorylation, acetylation, ubiquitination, methylation, SUMOylation, and O-GlcNAcylation stand out for their pronounced roles in modulating TF activity[13,14] (Figure 3). Phosphorylation, regulated by protein kinases and phosphatases, acts as a TF activity switch. Similarly, acetylation often enhances TF activity by increasing DNA binding affinity and stability. Ubiquitination mainly targets TFs for degradation but can also alter protein function or localization within the cell. Methylation, on the other hand, influences gene expression by altering TF interactions with chromatin and DNA, with histone methylation affecting TF dynamics. SUMOylation involves the addition of SUMO protein to lysine residues, a modification that can compete on the same lysine sites targeted with other PTMs, thereby influencing TF action. Lastly, glycosylation, the attachment of sugar units to amino acids, though less common, is crucial in dictating TF stability and their propensity for protein-protein and DNA interactions. In summary, the efficacy of TFs to carry out its functionality is extensively shaped by PTMs. Therefore, to comprehensively understand TF's regulatory roles, it is crucial to consider the PTMs that govern their activity.



***O*-linked modifications**

Phosphorylation: Addition of a phosphate group through an ester bond to polar amino acids results in increased negative charge in the vicinity of the modification.

Glycosylation: Addition of an *O*-linked *N*-acetyl-glucosamine monosaccharide to polar amino acids through a β-linkage. Competitive with phosphorylation, glycosylation does not alter the charge of the protein.

***N*-linked modifications**

Methylation: Addition of CH_3 to basic amino acids results in increased hydrophobicity. Unlike acetylation, mono-, di, and tri-methyl-ation of lysine is not charge neutralizing but increases the effective radius of the positive charge by replacing hydrogens with bulky methyl groups.

Acetylation: Acetylation of lysine groups is charge neutralizing and competitive with ubiquitination and sumoylation.

Sumoylation: Addition of one or many ~100 amino acid peptides through a highly labile ε-amino isopeptide bond to lysine results in greatly increased protein bulk. Poly-sumoylationcan occur via lysines within the SUMO moiety.

Ubiquitination: Addition of one or many ~76 amino acid peptides through an ε-amino isopeptide bond to lysine or through a peptide bond to the amino terminus. Poly-ubiquitination can occur through different lysines in the ubiquitin peptide to form a variety of chains.

Figure 3. Post-Translational Modifications Regulating Transcription Factor Activity

A table showing the six main PTMs of TFs, grouped into 'O-linked' (phosphorylation, glycosylation) affecting oxygen atoms, and 'N-linked' (methylation, acetylation, sumoylation, ubiquitination) targeting nitrogen atoms. Post-translational modifications (PTMs). Transcription Factors (TFs); Adapted from Filtz, Theresa M et al., Trends in pharmacological sciences (2014)[13].

1.1.3.2. Nucleosome Positioning and Chromatin Accessibility Influencing Transcription Factor Activity

Nucleosome positioning and chromatin structure dynamics are central to understanding how the physical organization of DNA in the nucleus influences TF activity and gene regulation. Nucleosomes are fundamental to the structure of chromatin and are composed of roughly 147 base pairs of DNA enfolded around a histone protein complex that includes matched sets of H2A, H2B, H3, and H4[15]. This structural foundation facilitates DNA compaction to control the exposure of gene regulatory regions, thereby enabling or hindering the binding of TFs and serving as a dynamic regulatory mechanism[16] (figure 4). Nucleosomes can be dynamically repositioned in response to cellular signals, altering the landscape for TF bindings. Techniques such as Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq) have become instrumental in exploring these dynamic elements of the genome, shedding light on the chromatin accessibility that marks active regulatory domains where TFs are likely to bind[16,17]. Such high-resolution insights into nucleosome organization and chromatin accessibility are integral to decoding the complexities of gene regulation.

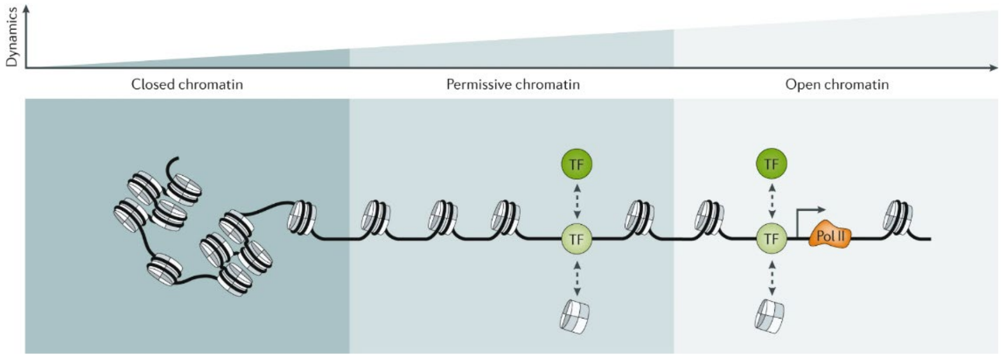


Figure 4. Chromatin Accessibility and Transcription Regulation

The diagram shows three chromatin states. Closed chromatin is compact, limiting TFs and RNA Pol II accessibility, signifying gene repression (left). Permissive chromatin has a more dynamic structure that permits TFs to bind and begin sequence-specific remodeling for accessibility, indicating a state primed for gene expression (middle). Finally, open chromatin is characterized by the active engagement of TFs and Pol II in transcription, denoting gene activation (right). Transcription Factor (TF); RNA polymerase II (Pol II). Adapted from Sandy L. Klemm, Zohar Shipony, & William J. Greenleaf, Nature Reviews Genetics (2019)[16].

Another epigenetic modification influencing TF activity is the addition of a methyl group to the cytosine bases within CpG dinucleotides. This epigenetic modification plays a crucial role in the regulation of gene expression. It is mainly known for its gene-silencing effects when it occurs in promoter regions, thereby blocking TF binding and altering the transcriptional landscape[18]. Some TFs are deterred by methylation at their binding sites, whereas others are attracted to methylated DNA, promoting chromatin changes that lead to gene repression[19]. This dynamic epigenetic mark influences cell differentiation and identity maintenance, and its dysregulation is implicated in disease pathogenesis, including cancer[20].

1.1.3.3. Physical Interactions with Cofactors and Chromatin Remodelers

Central to the functionality of TFs is their interaction with an accessory of cofactors, which modulate their binding affinity, specificity, and regulatory activity. These cofactors are composed of diverse entities, ranging from small molecule ligands, which can directly affect the TF's conformation and DNA binding affinity, to chromatin remodelers and histone-modifying enzymes, which alter the accessibility of DNA to TFs[6,21,22]. Specifically, cofactors may induce conformational changes to the TF structure, thereby enhancing or diminishing the protein's ability to recognize and bind to its target DNA sequence. For example, the steroid hormone receptor family of TFs, such as the glucocorticoid receptor (GR), requires the binding of a steroid hormone to induce a conformational change that permits dimerization and binding to specific DNA response elements[23]. Additionally, cofactors such as the nuclear receptor coactivators (NCOAs) can alter the functional state of TFs by promoting or stabilizing their dimerization, a condition often required for effective DNA binding and transcription initiation[24]. Cooperative TF-TF interactions can also function as cofactors by forming complexes that modulate transcriptional machinery at gene promoters. For instance, the TF NFAT cooperates with AP-1 (a dimer composed of Fos and Jun proteins) in T-cells to activate the expression of interleukin-2, a key cytokine in the immune response[25].

Furthermore, cofactors are instrumental in facilitating chromatin structure remodeling, thereby influencing TF access to DNA that is tightly wound in chromatin. An example of this is the SWI/SNF complex, which interacts with various TFs to remodel chromatin and regulate gene expression[26]. This chromatin remodeling is achieved by the recruitment of additional enzymatic activities that modify the histones around which DNA is wrapped, leading to either the exposure or occlusion of TF binding sites. The interaction between TFs and cofactors is also crucial for recruiting the transcriptional machinery to the promoter regions of genes, thereby dictating the rate and pattern of gene expression. The mediator complex, for instance, serves as a hub for signals from multiple TFs and transmits them to RNA Pol II, ultimately influencing transcriptional output[27]. TF-protein interactions mediated by cofactors can determine the assembly of this machinery, leading to either the activation or repression of transcriptional activity. Essentially, the multifaceted roles of cofactors are essential for both the fine-tuning of TF activity and the overall control of gene expression that defines cell functionality and identity.

1.1.3.4. Modulation of Transcription Factor Activity by Noncoding RNAs

RNA molecules such as microRNA (miRNA)[28], long noncoding RNA (lncRNA)[29], and enhancer RNAs (eRNAs)[30] also play crucial roles in regulating TF activities. miRNAs often target mRNAs for degradation or inhibit their translation, influencing TFs indirectly. An example is miR-34a, which is involved in the TP53 tumor suppressor network by modulating genes for cell cycle and apoptosis[31]. miRNAs also have a role in signal transduction, with the Let-7 family modulating RAS pathway signaling, which is essential for various TFs' functions[32]. On the other hand, lncRNAs like Gas5 and HOTAIR influence TFs by serving as molecular decoys, altering chromatin structure, or affecting transcription machinery assembly[33,34]. Additionally, lncRNAs such as MALAT1 are key in post-transcriptional regulation by influencing the splicing of pre-mRNAs, thereby altering the types and functions of TF isoforms[35]. eRNAs, on the other hand, are believed to be critical components of active enhancers, contributing in multiple ways[30]. Firstly, eRNAs interact with chromatin looping factors and bind to various TFs, aiding in enhancer-promoter loop formation, RNA Pol II loading, and histone modification. They also facilitate RNA Pol II pause release by activating the P-TEFb complex and function as decoys, sequestering cofactors to prevent transcription repression. These RNA-TF networks are crucial for adaptive cellular responses and have significant therapeutic potential.

1.1.4. The Fundamental Role of Transcription Factors in Development

The development and regenerative capacity of multi-cellular organisms are crucially dependent on the correct temporal and spatial control of gene expression. The selective use and distinctive interpretation of identical genetic material in each cell are orchestrated by various TFs. During developmental processes, TFs interpret cellular and environmental signals, ensuring that the cell's genetic information is read correctly during processes such as differentiation, morphogenesis, and growth[36]. In doing so, TFs initiate the commitment toward unique and irreversible cell fate. Specifically, TFs act by coordinating protein complexes at the associated promoter of developmental genes and distal enhancer elements, consequently modulating gene expression patterns[4,6,37]. Overall, TFs are crucial in embryonic development, guiding numerous cell fate decisions and cellular differentiation trajectories. This process initiates with a totipotent zygote cell that divides to form the blastocyst, featuring an inner cell mass (ICM) of pluripotent cells. These cells differentiate into the three germ layers (endoderm, ectoderm, and mesoderm), eventually leading to functionally differentiated cells (Figure 5).

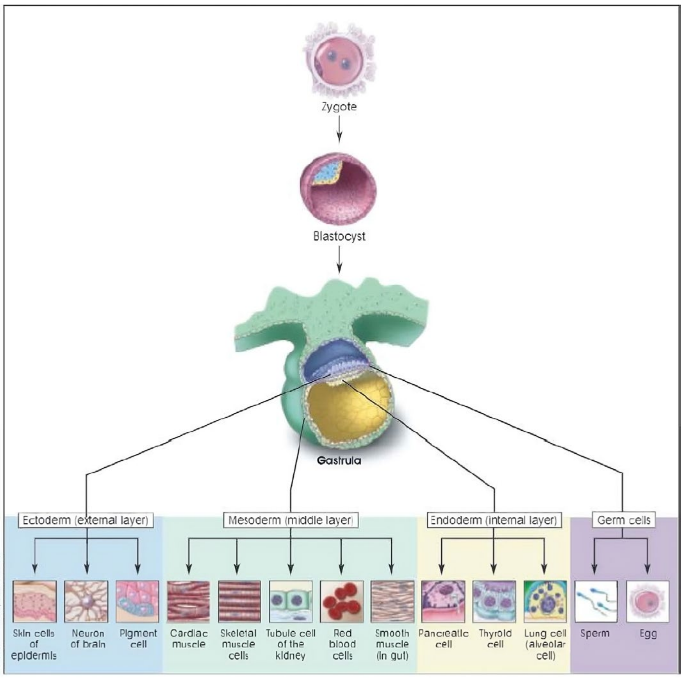


Figure 5. Embryonic Development Consists of Numerous Cell Fate Decisions Orchestrated by Transcription Factors

This schematic illustration depicts the unidirectional process of embryonic development from a totipotent zygote through terminally differentiated cells. The totipotent zygote divides into a blastocyst with pluripotent cells in the ICM. Then the gastrula is formed and generates the three germ layers, ectoderm, mesoderm, and endoderm, which eventually establish all specialized body cell types, and into gametocytes (germ cells), which will later form gametes. These numerous cell fate decisions are regulated by transcription factors (TFs), which guide the cells toward their specialized functions within the developing embryo. Inner cell mass (ICM); Transcription Factors (TFs). Adapted from National Institutes of Health (U.S.), June. 2001[39].

For instance, TFs like OCT3/4, CDX2, and TEAD4 guide early embryonic cell lineage differentiation, with complex interactions determining the eventual path to cell specialization. TFs such as NANOG, GATA6, SALL4, and SALL1 also play crucial roles in maintaining pluripotency and directing organ development[36]. Essentially, TFs operate within hierarchical cascades, with one TF's activity potentially affecting another, leading to sequential effects that either amplify or diversify the initial signal[4,38]. This, in turn, allows for a series of tightly regulated, stepwise events to guide cell identity trajectory toward a specific functional state. However, due to the intricate network of TFs, the precise mechanisms underlying their interplay and coordination with other regulatory molecules remain not yet fully understood. Thus, understanding the molecular mechanisms by which developmental TFs control cellular fate decisions is essential for accurate implementation to efficiently shape cellular fate in vitro.

1.1.5. Shaping Cellular Identity In Vitro and the Current State of the Art

In the past decades, the discoveries of myogenic reprogramming from somatic fibroblasts[40] and, more recently, the generation of induced pluripotent stem cells (IPSCs)[41,42], have shattered the long-held dogma that cellular differentiation was a unidirectional process (from pluripotency to terminally differentiated cell). These seminal findings have provided evidence that the ectopic expression of a critical set of TFs can effectively re-wire cell identity[43,44]. Indeed, current cellular conversion protocols enable the transformation of cell identities through several strategies: i) reprogramming to IPSCs, which involves reverting somatic cells back to a pluripotent state; ii) differentiation, also referred to as directed differentiation', where pluripotent cells are guided to become specific differentiated types; iii) transdifferentiation, the direct conversion of one type of somatic cell into another distinct somatic lineage[5] (Figure 6). In this context, recent studies have sought to decipher the functional mechanisms of TFs and how they affect genome conformation and activity. For instance, a recent study by Julia Joung et al.[45] established a barcoded library of more than 1600 human TF isoforms and assessed their influence on cellular identity in human embryonic stem cells (hESCs). Their research demonstrated that these TFs effectively induced distinct transcriptional signatures from all three germ layers and trophoblasts. Likewise, over a quarter of these TFs, when applied singly, had a substantial impact on the cellular state. Additional studies also proved that using a minimal set of key TFs can induce significant changes in gene regulatory networks, suggesting an additive or synergistic effect. These studies include not only the discovery of reprogramming to IPSCs, but also the conversion of somatic fibroblast cells into neurons, cardiomyocytes, hepatocytes, and blood progenitors[46–49]. Despite extensive research in the field, the functional role of numerous TFs has not been fully elucidated. With over 1,800 TF genes and more than 3,500 isoforms, the complexity of regulatory pathways is immense. Therefore, while significant advancements have been made, the thorough understanding of TF roles in gene regulation remains an ongoing and challenging area of exploration.

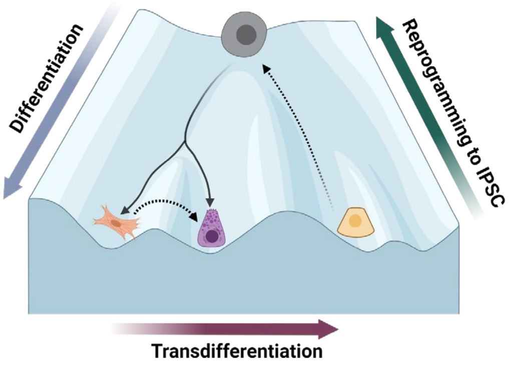


Figure 6. Strategies for Shaping Cellular Identity In Vitro
Schematic illustration of cell fate commitment using the Waddington landscape model. TF-mediated cellular conversion protocols include: i) Reprogramming to IPSC, reverting somatic cells to pluripotency; ii) Differentiation, transforming pluripotent cells into differentiated ones; iii) Transdifferentiation, directly converting somatic cells into a different somatic lineage. Transcription Factor (TF); Induced Pluripotent Stem Cell (IPSC). Created with BioRender.com

1.1.5.1. Interplay of Pioneer Transcription Factors and Cooperative Binding in Cellular Fate Reprogramming

Not all TFs possess the same functional capacity. Some factors, known as 'pioneer' TFs, demonstrate higher functional capacity to re-wire cellular fate. These TFs can interact with heterochromatin, thereby enabling a more permissive cellular state that facilitates the recruitment of other co-factors to further drive the developmental progression[50] (Figure 7). Essentially, pioneer factors are considered at the top of the TF hierarchy, playing a crucial role in initiating the commitment toward a unique cellular identity in vivo and in vitro. Therefore, TF-mediated cellular conversion protocols require at least one pioneer TF to initiate the commitment toward a target cell type[51].

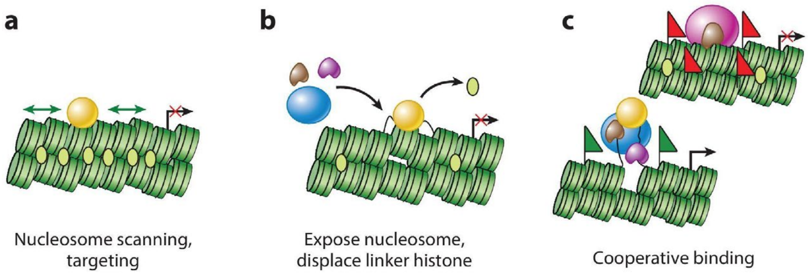

a
b
c
Nucleosome scanning, targeting
Expose nucleosome, displace linker histone
Cooperative binding

Figure 7. Functional Mechanisms of Pioneer Transcription Factors
a) A pioneer TF, represented as a gold sphere, laterally scans across chromatin, and targets a nucleosome. b) The pioneer TF reveals an underlying nucleosome in chromatin, resulting in the displacement of linker histone.

c) The pioneer TF facilitates the binding to other TFs, co-activators or co-repressors, and nucleosome remodelers. Green flags indicate activating histone modifications, while red flags denote repressive histone modifications. Transcription Factor (TFs). Adapted from Kenneth S Zaret, Annual review of genetics (2020)[50].

For instance, during the reprogramming of somatic cells to IPSCs using the classical Yamanaka factors (OCT4/POU5F1, SOX2, KLF4, and c-MYC), evidence of interactions between pioneer TFs and nucleosomes was observed[41]. OCT4, SOX2, and KLF4 were found to bind specific loci within condensed chromatin regions in the early reprogramming stages, thus functioning as pioneer TFs. In contrast, c-MYC exclusively exhibited binding capability to open chromatin regions[52].

1.1.5.2. Pre-Established Cellular Epigenetic State Affects Cellular Reprogramming Efficacy

The ability of TFs to affect cellular identities strongly depends on the cellular pre-established epigenetic state. In line with this, emerging evidence indicates that while pioneer TFs have the capability to remodel chromatin states, their effectiveness in re-wiring cellular identity is significantly influenced by the primary cellular context[53]. This is evident by the limitations of present cellular conversion techniques, marked by their inefficiency and the generation of immature target cell types[54,55], along with the ineffective translation of cellular reprogramming protocols from transgenic mouse models to human applications[56–58]. Additionally, in an atypical epigenetic environment, TFs may diverge from their canonical roles, leading to unforeseen results. This was exemplified in a paradoxical observation where SNAI1, a TF linked with epithelial-to-mesenchymal transition (EMT), unexpectedly improved the efficiency of reprogramming to IPSC instead of inhibiting it[59]. Typically, the process of reprogramming somatic cells to IPSC involves the opposite processes of mesenchymal-to-epithelial transition (MET). Essentially, these impediments highlight the difficulty in predicting the universal TF activities for practical applications.

1.1.6. Decoding the Intricacies of Transcription Factors Genomic Dynamics

At present, research on TF activity employs both experimental and computational approaches to better understand and refine TF-driven processes that determine cellular identity. This includes in vivo studies, which explore TFs activity in their native cellular environment, offering insights into their role in gene regulation amid various influencing factors[60,61]. In vitro studies simplify our understanding by isolating the direct interactions between TFs and gene regulation, excluding or controlling confounding effectors. This effector may introduce complexity into the regulatory network, thus potentially masking the global functionality of TFs [62,63]. Nevertheless, it is important to note that this approach may misrepresent essential TF interactions, such as promiscuous DNA binding, which can lead to non-canonical transcriptional effects.

1.1.6.1. Functional Assays to Study Transcription Factor Role in Cellular Differentiation

To dissect the multifaceted roles of TFs in cellular differentiation, functional assays often rely on overexpression techniques. Common methods involve vector-based, plasmid gene delivery, and viral transduction, noted for their efficiency crucial for studying TFs in cell fate determination[62]. The CRISPR/dCas9 system could also employed for its precision in upregulating specific TFs without off-target effects[63]. Alternatively, it can be used to induce transdifferentiation by permanently silencing specific genes. Another method, direct TF mRNA transfection, allows for accurate transient TF delivery to cells without leaving genomic scars[64]. Despite these advances, the results obtained by TF overexpression are subject to multiple constraints (such as pre-established epigenetic state, presence of varied effectors, experimental settings, etc), consequently leading to overestimation of TF functionalities and inconsistent outcomes across different cellular conversion protocols. Hence, while these techniques have advanced our knowledge, careful data interpretation is imperative to avoid misleading results and truly comprehend TFs' global interactions in varied cellular systems.

1.1.6.2. Omic Technologies and Computational Strategies to Study Transcription Factor Functionality

After conducting functional assays, the next critical step is further analyzing TF functionality assays from the retrieved data. Identifying a DNA-binding motif is an essential first step in unraveling the intricacies of genomic dynamics related to a TF of interest, as it sheds light on the TF's functional role. Specifically, facilitating the identification of TF’s putative binding site in differentially accessible chromatin regions enriches our understanding of the TF's prospective target genes and the biological processes it may regulate. In recent years, our ability to connect motif sequences to specific TF’s genomic binding sites has heavily relied on a curated database of eukaryotic TFs and their genomic binding sites, such as TRANSFAC[65], JASPAR[66], and HOCOMOCO[67]. Numerous in vitro and in vivo techniques are available for evaluating the sequence preferences and binding sites of a given TF. Chromatin Immunoprecipitation sequencing (CHIPseq)[68] is one such method that has revolutionized the field of ‘TF binding’ study by enabling the identification of genome regions occupied by TF of interest and thus inferring potential target genes. Specifically, following the cleavage and crosslinking of query chromatin into fragments, TF-bound DNA fragments are immunoprecipitated using antibodies specific to the TF of interest and subsequently sequenced. Then, by applying various computing methodologies, such as MEME tools[69], and drawing upon the previously described curated motif dataset, the attributes of TF binding can be identified by examining genomewide TF-bound DNA fragments. The motif enrichment data offers a limited foresight into the functional capabilities of its associated TF, since additional constraints (such as promiscuous binding due to high expression levels, PTMs, cofactor dynamics, and the epigenetic landscape) influence the function of a given TF. Hence, distinct tests are required to pinpoint the exact sites where a TF attaches to DNA and to assess if such attachment is evidently probable to influence the associated downstream gene activity.

An alternative approach is the ATACseq, which allows the profiling of both TF-induced chromatin accessibility and binding characteristics (Figure 8). The ATACseq protocol works by fragmenting and amplifying DNA sequences found in open chromatin areas using hyperactive Tn5 transposase that has preloaded adapters before proceeding with the sequencing step. The lack of a size-selection step allows us to determine the location of nucleosomes and identify accessible regions simultaneously. This is because DNA fragments larger than ~147 base pairs are where nucleosomes are located, whereas shorter fragments indicate accessible regions. Additionally, utilizing motif search tools mentioned earlier (such as the MEME tools)[69], it is possible to identify and quantify the probability of motifs present in an accessible chromatin region. These findings can then be evaluated against a random background or different conditions for comparative analysis to identify motifs in TF-induced open chromatin regions. Another way to decipher the TF binding regulation is to use footprint analysis[70]. In ATACseq, a footprint is an observable pattern denoting the interaction between an active TF and DNA, which effectively hinders Tn5 cleavage within the TF binding site. As a result, there is a discernible drop in the accessibility of the open chromatin area (Figure 8). Nevertheless, it is crucial to emphasize that TF footprint analysis's accuracy hinges on the quality and depth of the ATACseq data, as it necessitates the identification of subtle protective patterns at the sites where TFs bind[71]. In summary, ChIPseq and ATACseq are powerful epigenomic-based tools for identifying potential TF binding sites and chromatin accessibility. However, they do not provide direct evidence of TFs' functional gene expression regulation.

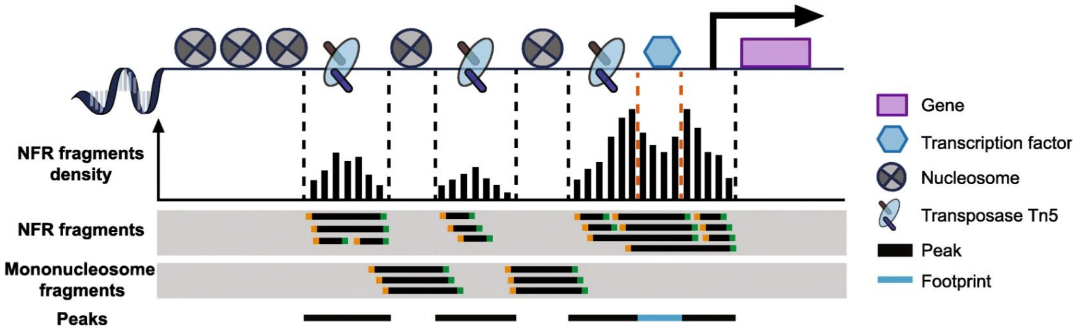


Figure 8. Schematic Diagram of ATACseq Technology for Studying Chromatin Accessibility

Schematic representation of ATACSeq experiment, Tn5 transposase targets open chromatin, integrating sequencing adapters. Sequencing identifies open chromatin (black) and transcription factor footprints (blue). NRFs indicate open chromatin that is more accessible for TF binding, while nucleosome-bound fragments (gray-shaded tracks) show nucleosome positions. Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq); Nucleosome-free regions (NFRs); Nucleosome-free regions (NFRs). Adapted from Yan, Feng et al., Genome biology (2020)[71].

Further experiments and analyses are usually required to determine whether a TF is actively regulating the expression of various target genes. These include gene expression assays, such as RNA Sequencing (RNAseq)[72], namely bulk RNAseq, to measure changes in transcriptional profile in response to TF-mediated cellular conversion assays. Specifically, RNAseq is a high-throughput method enabling a comparative analysis, namely differential expression analysis (DEA), of TF-induced transcriptional modifications against a given background or other biological condition of interest. Thus, this technique facilitates the discovery of target genes regulated by specific TFs of interest, offering insights into the molecular underpinnings of gene regulation. Additionally, it was demonstrated that cellular conversion protocols induce transient waves of developmental gene expression throughout the cellular differentiation process[53]. Consequently, time-series RNAseq studies are frequently employed to investigate TF functionality along the course of cell fate reprogramming. Tracking the temporal dynamics in gene expression enhances our understanding of whether these intermediate genetic signatures represent distinct cell populations or whether they are indicative of various transient cellular states during differentiation. Overall, despite the transformative potential of RNAseq, the success of this approach is limited. RNAseq data represent the sum of transcriptional activity and RNA molecular stability within the cell. This signifies that retrieved data encompasses both newly synthesized RNA molecules and pre-existing cellular ones. Hence, RNAseq results should be interpreted as an indirect measure of TF-induced transcription.

The main focus of this thesis is to investigate the properties of TFs by utilizing the advanced omic techniques of ATACseq and RNAseq. However, we acknowledge the extensive array of additional alternative experimental methods, which span from qualitative techniques like Electrophoretic Mobility Shift Assay (EMSA)[73], which visualizes TF-DNA interactions; to quantitative ones, including co-immunoprecipitation followed by mass spectrometry (IP-MS)[74,75] for detecting TF-related proteins, and additional sequencing approaches such as ‘DNase I hypersensitive sites sequencing’ (DNaseseq)[76], ‘Cap Analysis of Gene Expression sequencing’ (CAGEseq)[77], and ‘RNA And DNA Interacting Complexes Ligated and sequenced’ (RADICLseq)[78].

1.1.6.3. Computational Methodologies of Multiomic Data Integration and their Application in Transcription Factor Studies

As outlined above, each omic technology offers a unique insight into a distinct cellular layer (epigenetic, transcriptomic, proteomic, etc), bounded by its respective strengths and limitations. However, relying on a single omic data provides only a partial view of the inner complex functionality of a TF. Recently, the emergence of multiomic integration methods has facilitated a more precise dissection of various functional levels. Consequently, these strategies could be beneficial for gaining deeper insights into the complex TF regulatory network, which involves numerous interconnected cellular functional domains with multiple constraint[79]. Specifically, these methodologies integrate distinct omics data derived from various cellular levels, either sequentially or concurrently, thereby bridging the genotype-to-phenotype gap. Previously, Subramanian, Indhupriya et al. (2020)[79] broadly classified different multiomic data integration methodologies into six distinct approaches, each with unique applications and advantages[79] (Figure 9). These include the similarity-based methods, that exploit the inherent similarities across omics data to group or classify biological entities[80,81]. Correlation approaches, employ statistical tools to find correlations across omics datasets[82]. Network-based methods, involve constructing and analyzing biological networks to elucidate the relationships and interactions between various biological molecules, offering insights into the systemic organization and its perturbations[83,84]. Bayesian methods, introduce a probabilistic framework for integrating data, which is useful for incorporating prior knowledge and handling the uncertainty of biological data[85,86]. Fusion methods, integrate multiomics data at various stages, enabling a comprehensive analysis that leverages the strengths of each omic layer[87,88]. Lastly, multivariate methods, consider multiple variables and their interactions, facilitating a robust analysis of the interdependencies within multiomic data[89,90].

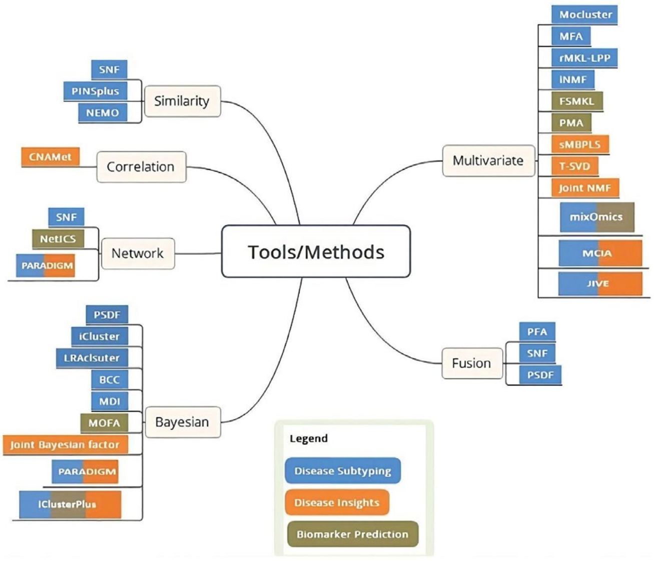


Figure 9. Categorization of Multiomic Data Integration Methods

Multiomic data integration methods are categorized into six groups based on their methodological approach. Color-coding indicates their applications: disease subtypes (blue), disease insights (orange), and biomarker prediction (green), providing a clear and concise overview of their functions. Several strategies, like PARADIGM and similarity network fusion (SNF), share several strategies across the categories. Adapted from Subramanian, Indhupriya et al., Bioinformatics and biology insights (2020)[79].

In this thesis, we repurposed the Similarity Network Fusion (SNF)[84] algorithm, a fusion and network-centric approach, to infer interconnectivity relationships among numerous TFs, determined by shared or distinct epigenetic and transcriptional traits. The integration approach used in SNF, known as "passage-based data integration," creates individual networks for each data type and then fuses these through iterative updates, enhancing similarity measures. Consequently, SNF offers more comprehensive and insightful biological interpretations, making it a powerful tool for understanding intricate biological systems. Eventually, this process leads to a single, integrated multiomic network that better captures the complexities of the data.

1.1.6.4. Advanced Morphological Assays in Elucidating Transcription Factor Roles in Phenotypic Determination

Morphological assays serve as an additional approach to studying TF activity in cellular conversion protocols. This includes a variety of high-resolution and high-throughput techniques that measure cellular properties such as volume, area, and shape. Consequently, these strategies allow for evaluating TFs' capacity to alter cellular identity at the phenotypic level and validate the agreement with transcriptional and epigenomic signatures. For instance, a common method used for capturing structural changes is immunofluorescence, which includes direct and indirect methods. The direct method involves a fluorophore-linked antibody binding directly to the antigen, while the indirect method uses a secondary antibody to amplify the signal. This allows for the precise visualization of proteins within cells and tissues under fluorescence microscopy. The utilization of specific antibodies such as DAPI (for nuclear shape), Tubulin (for microtubules, including the outer boundary of the cell and cytoplasmic structure), Fibronectin (for extracellular matrix organization and cell adhesion), and Cell Mask (for the cellular membrane) provides valuable insights into TF-induced changes in cellular morphology, highlighting their characteristics compared to reference cell types[91]. Moreover, staining of cellular-specific marker genes like ACTA2 (smooth muscle) or MAP2 (neuron) can help to assess the position of TF within a cell (nucleus - implies on its activation, cytoplasm - unactivated)[45,92]. An additional commonly used technique involves the introduction of fluorescently tagged proteins into live cells through exogenous overexpression. This is often accomplished by fusing the protein of interest with a fluorescent protein tag or by incorporating a fluorescent protein tag coding cDNA at the endogenous loci[93].

1.1.7. Major Challenges Addressed in This Dissertation

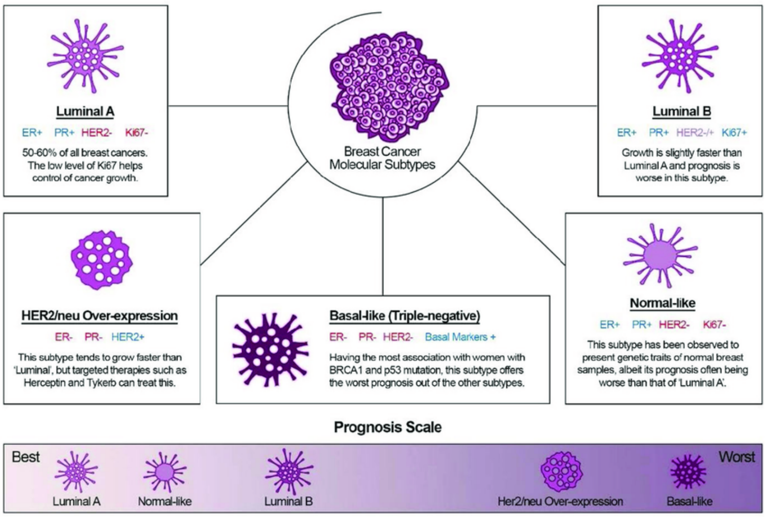
In the last decades, the primary dogma that cellular differentiation was a unidirectional process (naive to determined functional cellular fate) has been shattered by the paradigm-shifting discoveries of cellular reprogramming to IPSC and transdifferentiation of fibroblasts toward myogenic fate[40,41]. This has led to a new school of thought in which an ectopic expression of a crucial subset of TFs can re-wire cellular programmes[94]. Consequently, this has triggered an unprecedented boost in the study of TFs to precisely shape cellular fate in vitro[51]. These advancements are critical for addressing the growing needs of regenerative medicine, tissue engineering, and disease modeling. Nonetheless, implementing cellular conversion assays is a complex and challenging process. This complexity mainly arises from the intricate regulatory landscape governing TF activity. As a result, existing cellular conversion methods remain inefficient and yield phenotypically immature cells[54,55]. A significant limitation in this field is the influence of the initial cellular context, which can greatly affect the functioning of TFs. Indeed, TF-mediated cellular conversion protocols have shown varied and inconsistent outcomes, making it challenging to predict the overall activity and functionality of TFs. To date, research in this field has mainly focused on identifying the most effective TFs for generating specific target cell types of interest, such as hepatocytes, neurons, and IPSCs. While this approach is beneficial for producing distinct cell types, it tends to neglect comparative analyses across various TFsWhile this approach is beneficial for producing distinct cell types, it tends to neglect comparative analyses across various TFs, thus may overlook the TF global functionalities and co-interactions. Therefore, establishing a systematic workflow for side-by-side comparative analysis of TFs is crucial to gaining novel insights into their interconnectivity and overarching functionalities, extending beyond the constrained findings of focused cellular conversion protocols tailored toward the generation of unique target cell types. Such a breakthrough would not only unlock the full potential of TFs but also expand their utility across a wider range of cellular systems, enhancing the scope and effectiveness of cellular fate reprogramming studies.

To address the abovementioned challenges, we developed a comprehensive transcriptomic, epigenomic, and morphological screening of 130 developmental TFs to appraise their effect on cellular transdifferentiation. Our approach represents one, if not the only, case of side-by-side comparison of TF dosages within the same experimental setting. Essentially, by analyzing well-established and yet uncharacterized TFs, we identify novel developmental TFs that, either individually or in combination, play a pivotal role in regulating cellular identity. We anticipate that this strategy will ultimately lead to novel paradigms in deciphering crucial dynamics driving cell-fate decisions, which potentially allow us to unlock the full potential of pioneer TF repertoire.

1.2. Second Project - Spatial Transcriptomics Reveals Sub-Tumoral Identities and Novel Diagnostic markers in Triple Negative Breast Cancer with Immune Evasion Capacity

1.2.1. Overview of Breast Cancer

In the United States, breast cancer significantly impacts women's health, ranking just behind nonmelanoma skin cancers in frequency. The year 2023 was projected to see about 300,000 new cases of breast cancer, with the disease expected to claim the lives of approximately 43,000 women[95]. Despite advancements in detection through mammography and targeted therapies for hormone receptor-positive types, breast cancer continues to be a major cause of cancer-related deaths among women[96]. Crucially, this cancer is classified by gene expression patterns into five distinct molecular subtypes, luminal A, luminal B, HER2-enriched, basal-like, and normal breast-like-based[97–100] (Figure 10). Specifically, Luminal A cancers, the most prevalent, are characterized by lower aggressiveness and positive estrogen receptor (ER)/progesterone receptor (PR) expression, but are human epidermal growth factor receptor 2 (HER2) negative[101]. On the other hand, Luminal B cancers, comprising approximately 15-20% of cases, exhibit a heightened level of aggression. This is attributed to an elevated expression of genes related to cell proliferation and growth factor receptor signaling. Additionally, these cancers frequently express HER2 alongside ER and PR, further distinguishing their aggressive nature[101]. HER2-enriched cancers, which have a similar prevalence, exhibit minimal to no ER/PR expression and are known for their aggressiveness. This is largely due to the role of HER2 in regulating critical cellular processes such as proliferation, survival, angiogenesis, invasion, and metastasis[101]. Basal-like cancers, which constitute ~8-37% of breast cancer cases, are characterized by the absence of ER, PR, and HER2 receptors and are known for their notable aggressiveness. These tumors express high levels of basal myoepithelial markers, Caveolins 1 and 2, and the epidermal growth factor receptor (EGFR). Their aggressive nature is further amplified by frequent TP53 mutations and genomic instability, coupled with deregulated integrin expression, all contributing to their rapid progression and aggressive behavior[101]. Normal breast-like cancers, the least common subtype, are uniquely characterized by the absence of both standard markers (ER, PR, and HER2 receptors) and basal-like markers[101]. Essentially, although these subtypes are distinctly classified by their gene expression patterns, their application is primarily in research settings. The translation of these subtypes into clinical treatment strategies is not always direct or straightforward, highlighting the need for a nuanced approach to breast cancer treatment. A more clinically pertinent and prognostically insightful approach to classifying breast cancer subtypes relies on the available treatment options, dividing them into three distinct subgroups: hormone receptor-positive, HER2-positive, and triple-negative (i.e., TNBC)[102]. TNBC and basal-like breast cancer are often conflated due to a high incidence of overlap. The majority of TNBCs exhibit basal-like characteristics, and most basal-like cancers are categorized as TNBC. Nevertheless, their identification is based on distinct criteria, basal-like cancers are defined through gene expression profiling, whereas TNBC is characterized by the absence of hormone receptors as determined by immunohistochemical tests. Despite their frequent overlap, a notable discordance rate of 20-30% between these subtypes underscores the importance of precise diagnostics[99,100,102].



50-60% of all breast cancers. The low level of Ki67 helps control of cancer growth.

Growth is slightly faster than Luminal A and prognosis is worse in this subtype.

This subtype tends to grow faster than 'Luminal', but targeted therapies such as Herceptin and Tykerb can treat this.

Having the most association with women with BRCA1 and p53 mutation, this subtype offers the worst prognosis out of the other subtypes.

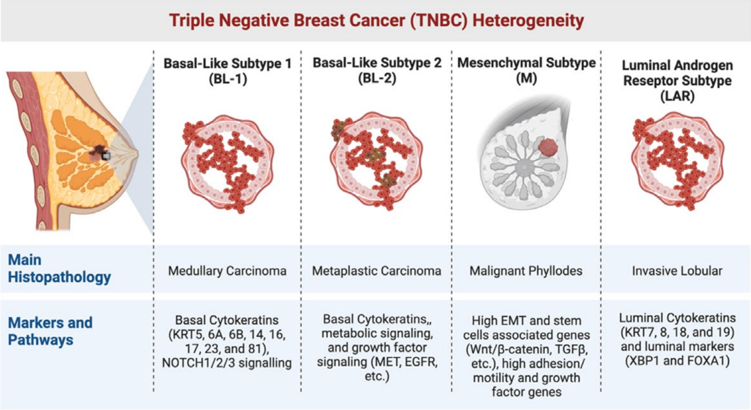
This subtype has been observed to present genetic traits of normal breast samples, albeit its prognosis often being worse than that of 'Luminal A'.

Figure 10. Molecular Subtypes of Breast Cancer

Schematic representation of breast cancer molecular subtype classification, categorized by distinct gene expression patterns. This includes an overview of their characteristics and levels of aggressiveness. The subtypes are arranged from left to right as follows: Luminal A, HER2-enriched, Basal-like, Normal Breast-like, and Luminal B. Adapted from Bobal, Pavel et al., International Journal of Molecular Sciences (2021)[103].

1.2.2. Molecular Characterization of TNBC

TNBC is an aggressive type of cancer that accounted for 10% of all breast cancer diagnoses in the US in 2019, demonstrating a higher prevalence among younger female populations, particularly those of African or Hispanic descent[96,104]. Unlike other breast cancer types, TNBC is characterized by the lack of expression of ER and PR receptors, and the absence of HER2, as defined by the ASCO/CAP criteria[105,106]. It is characterized by a notably poorer prognosis compared to other breast cancer subtypes, exhibiting lower overall survival rates and a higher recurrence frequency. Particularly, TNBC patients diagnosed at stages III and IV have an estimated 5-year survival rate of around 25%[107,108]. In the middle of the 2000s, the acronym TNBC emerged to classify a subset of breast cancers defined by the absence of ER, PR, and HER2[109]. However, subsequent studies have shown that TNBC comprises a constellation of malignancies that display distinct molecular characteristics, clinical outcomes, and responses to therapy. Due to this wide heterogeneity, it is necessary to conduct comprehensive molecular profiling of the various TNBC subtypes to accurately diagnose and plan a suitable and targeted therapeutic approach. Indeed, recent advancements in the molecular profiling of TNBC have led to the identification of four distinct subtypes at the genetic level[110]. These subtypes include basal-like 1 (BL1), basal-like 2 (BL2), mesenchymal (M), and luminal androgen receptor (LAR)[99,100,111,112] (Figure 11). Specifically, BL1 subtype is primarily characterized by a high expression of genes involved in cell cycle regulation, cell division, DNA damage response, and NOTCH signaling. The BL2 subtype exhibits a strong presence of genes associated with growth factor signaling. The M subtype is defined by genes linked to cell motility and differentiation, while LAR tumors, although estrogen receptor-negative, are enriched in hormonally regulated genes. Although there is no standardized method for classifying TNBC subtypes, the categorization into distinct molecular subtypes has been instrumental in pinpointing varied treatment responses. Basal-like TNBCs, in particular, have shown increased sensitivity to chemotherapy. Research also indicates that LAR tumors may respond well to anti-androgen therapy[111,113]. Therefore, a deeper understanding of TNBC's heterogeneous nature is vital for identifying effective therapeutic targets, predicting chemotherapy responses, and improving TNBC diagnosis tests and patient standard care.



	Basal-Like Subtype 1 (BL-1)	Basal-Like Subtype 2 (BL-2)	Mesenchymal Subtype (M)	Luminal Androgen Reseptor Subtype (LAR)
Main Histopathology	Medullary Carcinoma	Metaplastic Carcinoma	Malignant Phyllodes	Invasive Lobular
Markers and Pathways	Basal Cytokeratins (KRT5, 6A, 6B, 14, 16, 17, 23, and 81), NOTCH1/2/3 signalling	Basal Cytokeratins,, metabolic signaling, and growth factor signaling (MET, EGFR, etc.)	High EMT and stem cells associated genes (Wnt/β-catenin, TGFβ, etc.), high adhesion/ motility and growth factor genes	Luminal Cytokeratins (KRT7, 8, 18, and 19) and luminal markers (XBP1 and FOXA1)

Figure 11. Distinct Histological and Transcriptional Signatures of Triple Negative Breast Cancer Molecular Sub-Types

The various molecular subtypes of TNBC as they were refined by Lehmann et al. in 2021[114]. It encompasses basal-like 1 (BL1), basal-like 2 (BL2), mesenchymal (M), and luminal androgen receptor (LAR) subtypes. For each subt-ype, the figure highlights the principal histopathological characteristics, specific markers, and the key signaling pathways involved. Adapted from Mahmoud, Rinad et al. Cancers (2022)[115].

1.2.2.1. Current Approaches in Treatment of Triple Negative Breast Cancer

TNBC has traditionally been challenging to treat due to its aggressive nature and limited responsiveness to hormonal and targeted therapies (i.e. anti-HER2 therapy). Therefore, chemotherapy remains the cornerstone of treatment for most TNBC cases[116]. At present, the field faces a significant challenge with a limited range of predictive diagnostic markers for chemotherapy efficacy in recurrent or metastatic TNBC. Additionally, there is a noticeable shortfall in effective therapeutic approaches for these advanced stages of TNBC. A critical and promising direction in addressing these challenges involves enhanced molecular profiling of TNBC, which is anticipated to not only improve the prediction of chemotherapy response but also aid in the discovery of novel, targeted treatment options[117,118]. With that said, recent studies and clinical trials have shown that immunotherapy has an important role in the treatment of TNBC, particularly for patients with high Programmed Death-Ligand 1 (PD-L1) expression. Therefore, for TNBC patients, evaluating the PD-L1 Combined Positive Score (CPS) is a crucial aspect of the therapy management process[119]. Current standard care protocols, applicable across a range of cancer types in both early-stage (neoadjuvant/adjuvant) and metastatic settings, now increasingly incorporate the integration of chemotherapy with immunotherapy, specifically pembrolizumab or atezolizumab[120]. These drugs, which target the programmed death protein 1 (PD-1) and PD-L1 pathways, respectively, represent significant advances in TNBC treatment. In the following sections, I will provide a comprehensive overview of the canonical PD-1/PD-L1 pathway and how it contributes to tumor immune escape mechanisms in TNBC.

1.2.3. Immune Landscape and Targeted Therapies in Triple Negative Breast Cancer

Given the limited treatment options and the tendency to develop resistance to chemotherapy, managing patients with TNBC poses significant challenges. Recent studies on TNBC are increasingly focused on diagnostic marker approaches, especially the tumor immune landscape, due to the significant presence of tumor-infiltrating lymphocytes (TILs). This is often evidenced by the concurrent expression of PD-L1 on both neoplastic and immune cells within the tumor microenvironment[121,122]. To date, PD-L1 serves as a key biomarker to select patients with metastatic or locally advanced TNBC for immune checkpoint inhibitors (ICIs) treatment[123,124]. This often coincides with PD-L1 expression on both neoplastic and immune cells within the tumor microenvironment, a phenomenon that is key to understanding tumor immune mechanisms[125]. Canonically, PD-L1 and PD-1 regulate the immune response through the phosphorylation of PD-1, leading to the inactivation of CD28 and T cell receptor (TCR) function and signaling pathways[125]. This process attenuates the activation signal of T cells. When PD-L1 is expressed by neoplastic cells, it significantly contributes to the activation of the immune escape mechanism by binding to PD-1 receptors on T cells, thereby inactivating them[125]. This interaction suppresses the immune response against the tumor, facilitating immune evasion by the cancer cells. PD-L1 expression was suggested to be associated with cytokines within the tumor microenvironment (TME), particularly with interferon-γ (IFN-γ), indicating a potential crosstalk between the TME and the tumor (Figure 12)[126].

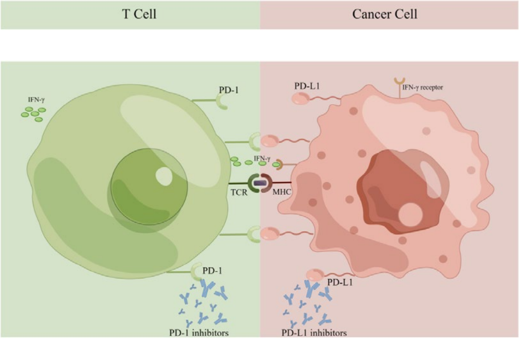


Figure 12. Tumor Immune Escape Mechanism Through PD-1/PD-L1 Pathway
Schematic representation of the PD-1/PD-L1 inhibitory pathway in cancer immunology. Lymphocyte proliferation, mediated by the TCR, is suppressed upon engagement of PD-1 with PD-L1, thus facilitating immune evasion. The greater the expression of PD-L1 in tumors, the more immunosuppressive the TME may become. Cytokines within the TME, such as IFN-γ, can induce the upregulation of PD-L1 expression in tumor cells. Therapeutic inhibitors targeting PD-1/PD-L1, including pembrolizumab (anti-PD-L1) (right) and atezolizumab (anti-PD-L1) (left), have demonstrated significant efficacy in treating various types of malignancies. T Cell Receptor (TCR); Programmed death-ligand 1 (PD-L1); Programmed cell death protein 1 (PD-1); Tumor Microenvironment (TME); Interferon-γ (IFN-γ). Adapted from Yang, Tinglin et al., Journal of personalized medicine (2023)[126].

Targeting the PD-1/PD-L1 pathway with ICIs has been developed to disrupt this mechanism, enhancing the host anti-tumor immune response to attack cancer cells, which is a pivotal strategy in treating various cancers, including TNBC. Indeed, the FDA's approval of atezolizumab (anti-PD-L1) and pembrolizumab (anti-PD-1), combined with chemotherapy for first-line therapy in PD-L1-positive metastatic TNBC, is based on the successful results of the IMpassion130 and KEYNOTE-355 clinical trials[127,128]. These trials, however, used different criteria for determining tumor's PD-L1 positivity. The addition of these drugs to chemotherapy significantly increased the overall response rates (ORR), thus marking a notable advancement in TNBC treatment. However, the efficacy of PD-1/PD-L1 therapy is highly varied[129] and influenced by a range of factors, notably the complexity of accurately determining PD-L1 expression levels. Histological diagnosis, which is crucial for identifying patients most likely to benefit from these treatments, represents a significant challenge. This complexity arises from variations in testing methodologies, the subjective interpretation of PD-L1 staining, and the dynamic nature of PD-L1 expression within the tumor microenvironment. indeed, the prevalence of PD-L1 positivity in TNBC varies widely, ranging from 17% to 59%, depending on the diagnostic methods and scoring used[130]. Additionally, while some TNBC patients with PD-L1 positive tumors respond well to ICIs, there are also cases where PD-L1 negative patients still benefit from these treatments[128,131,132]. ICIs represent a promising frontier in cancer therapy and are poised to transform cancer care in the near future. However, to enhance their applicability in TNBC, further detailed characterization and understanding of ICIs are required.

1.2.4. Advances in Triple Negative Breast Cancer Research

With the pressing need to unravel the biological complexity of TNBC, the emergence of next-generation sequencing (NGS) technologies has helped to discover new diagnostic, predictive, and prognostic markers, and enables the exploration of inter- and intra-tumor heterogeneity on the molecular level[110]. Specifically, bulk RNA sequencing (bulk RNAseq) has been used to identify transcriptional aberrations in TNBC. RNAseq has indeed brought significant advancements in classifying TNBC subtypes. As mentioned above, TNBC was primarily identified based on the absence of estrogen receptors, progesterone receptors, and HER2 amplification, which are phenotypic characteristics. However, this approach did not provide a comprehensive understanding of the diverse biology of TNBC. With the advent of RNAseq technology, researchers have been able to delve deeper into the molecular landscape of TNBC. This led to the discovery of more accurate molecular subtypes[133] that were first identified using microarray technology[110]. While bulk RNAseq offers valuable insights into gene expression patterns, it can obscure critical transcriptional trends within distinct subpopulations as it averages gene expression across subpopulations within a sample (also known as Simpson's Paradox[134]). This phenomenon has become particularly significant in cancer research as Tumors are inherently heterogeneous, consisting of various cell types each with unique transcriptional and epigenomic profiles. This heterogeneity, crucial for understanding the nuances of tumorigenesis and tumor progression, may be masked by the averaging effect of bulk RNAseq. The ability to discern these subtle yet vital differences in gene expression is pivotal, as it informs our understanding of cancer's complexity and guides the development of targeted therapies.

1.2.4.1. Single-Cell RNA Sequencing

The emergence of single-cell RNA sequencing (scRNAseq) effectively bridged this hurdle, offering a groundbreaking perspective for examining gene expression at an individual cellular level. Since its first introduction in 2009[135,136], this technique has been pivotal in revealing the intricate cellular heterogeneity within composite biological systems. Presently, efficient and cost-effective technologies enable standard labs to assemble sequencing libraries from thousands of cells, cementing scRNAseq role as a fundamental method in research[134]. These technological strides have facilitated the identification of previously novel cell types[137,138] and the in-depth analysis of cellular dynamics at a previously unattainable resolution[139,140]. Furthermore, this advancement has provided a new avenue for facilitating profound knowledge into subtle changes occurring in tumor biology by identifying distinctive clusters of cells, examining the tumor surrounding environment, and characterizing mutations in cellular genomics[141]. Focusing specifically on TNBC research, key studies have elucidated the intratumoral heterogeneity and distinct molecular subtypes within TNBC and the therapeutic implications of low PD-L1 expression in certain TNBC subtypes[114]. Essentially, these emphasize the importance of scRNAseq in developing TNBC treatments. However, accurately profiling tumor architecture through unlabeled scRNAseq without spatial context remains challenging due to the innate heterogeneity of tumors and their microenvironments[142]. ScRNAseq workflow is typically delineated into six fundamental stages: i) Dissociation of tissue samples to achieve a homogeneous single-cell suspension; ii) Optimization of input material quality through the assessment of cell viability; iii) Removal of lysed cells to ensure sample integrity; (IV) Barcoding of the transcriptome at the single-cell level; (V) Generation of complementary DNA (cDNA) from the processed cells; and vi) Construction of sequencing libraries followed by the sequencing process itself[134] (Figure 13). Consequently, the cell dissociation process inherent in scRNAseq impairs the ability to accurately determine the spatial organization and inter-cellular relationships within the original tissue architecture.

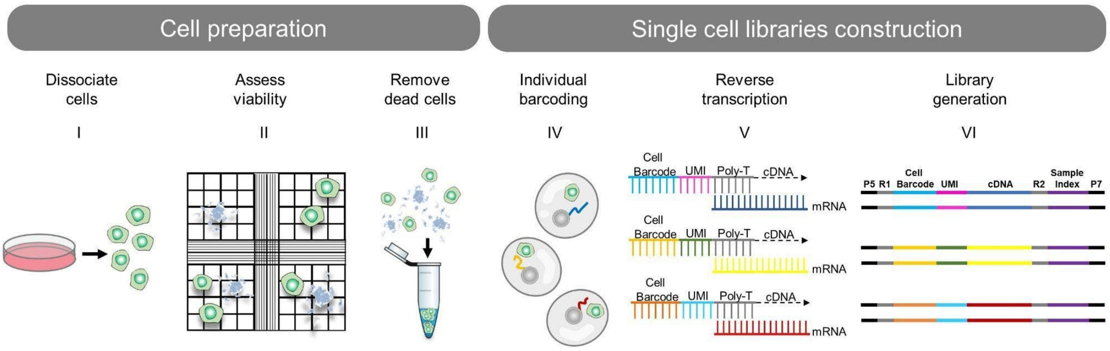


Figure 13. scRNA-Seq Workflow and the Inherent Loss of Spatial Information

Schematic representation of scRNAseq workflow steps. I) Cells are dissociated from the tissue, consequently leading to the loss of spatial information of cells. II) Optimization of input material quality. III) Removal of lysed cells. IV) Barcoding of captured RNA molecules at the single-cell level. V) Generation of cDNA. VI) Construction of sequencing libraries. Single cell RNA sequencing (scRNAseq); Complementary DNA (cDNA). Adapted from Slovin, Shaked, et al. Methods in molecular biology (2021)[134]

1.2.4.2. Spatial Transcriptomics and the State of the Art

With the advent of spatial transcriptomic methods, it is now possible to overcome these limitations by retrieving expression data while preserving cells' positional context, both in fresh and formalin-fixed paraffin-embedded (FFPE) tissues[143]. Spatial transcriptomics technologies can be broadly classified into two main categories[143] (Figure 14). The first category encompasses methods based on NGS. These methods encode positional information onto transcripts, thereby enabling the concurrent querying of the entire transcriptome. The second category comprises traditional approaches, primarily imaging-based, which include in situ sequencing-based methods and in situ hybridization-based methods. In the former method, transcripts are amplified and sequenced directly within the tissue, while in the latter, imaging probes are sequentially hybridized within the tissue[144]. While traditional image-based spatial transcriptomic methods excel in visualizing specific transcripts and their spatial distribution within tissues, NGS-based technologies offer more extensive coverage of the transcriptome, enabling a broader and more detailed analysis of gene expression patterns with quantitative precision. In this thesis, we will use the term "spatial transcriptomics" specifically to denote NGS-based technology.

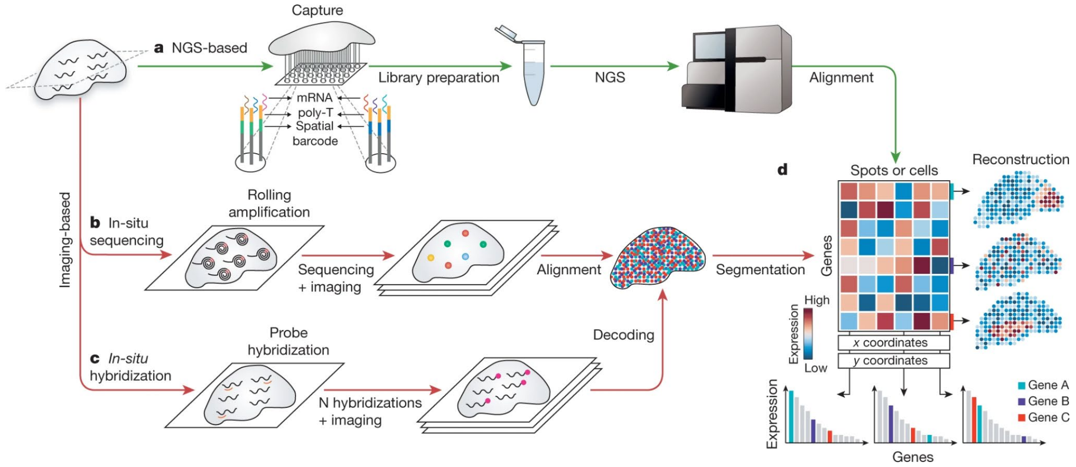


Figure 14. Next Generation Sequencing and Imaging Approaches in Spatial Transcriptomics Technologies

a) NGS-based methods barcode transcripts based on their spatial location in a grid of spots. (c-b) Imagine-based technologies. b) In situ sequencing reads transcript sequences directly from the tissue. c) In situ hybridization target sequences of interest using fluorescent probes. d) The outcome of all spatial transcriptomic technologies is a spatially resolved gene expression count matrix. Next Generation Sequencing (NGS). Adapted from Rao, Anjali et al. Nature (2021)[144].

The innovation in spatial transcriptomics technique is rooted in the use of capture probe slides featuring unique spatial barcodes (Figure 14). These barcodes allow the capture of poly-adenylated RNA molecules and subsequently spatially label them before the reverse transcription process[144]. Each capture probe slide has over a thousand spatial barcoded spots, offering spatial resolution ranging from ~0.5 to 100 micrometers[145]. Consequently, this breakthrough enables the precise mapping of gene expression data to its native spatial coordinates within the tissue of interest. NGS-based spatial transcriptomic technologies typically involve several key steps: 1) tissue block preparation and sectioning, 2) placement of sections on capture probe slides, 3) Hematoxylin and Eosin (H&E) staining and imaging, 4) barcoding cellular RNA based on their spatial positions, and 5) library construction followed by sequencing[144,145]. The resulting data can be used to create detailed spatial maps of gene expression patterns, which can be employed to infer crucial spatiotemporal dynamics of cancers, study the interactions between tumor and microenvironment cells, and, foremost, identify more accurately novel diagnostic markers[146].

Spatial transcriptomics analysis in the scope of TNBC has provided key insights. A recent study compared PD-L1-positive and PD-L1-negative tumors in TNBC, revealing differences in their tumor microenvironments and implications for immune therapy[147]. An additional study employed spatial transcriptomics to analyze 38,706 spatial features from a cohort of TNBC tissues, revealing nine distinct transcriptional clusters and race-associated differences in tumor characteristics, indicating a conserved spatial-transcriptional architecture in TNBC[148]. Nonetheless, the number of studies of TNBC using spatial transcriptomics is limited as it is a relatively new field. Indeed, until recently, the technologies required for such analysis were both complex and expensive. The early spatial transcriptomics strategies, such as the method introduced by Ståhl et al.[149] and Slide-seq[150,151], necessitated customized setups and specialized skills due to their Intricacy and labor-intensive nature. However, as spatial transcriptomics technologies have evolved, recent advancements have led to the commercialization and standardization of these methods, making them more accessible in any laboratory bench. Examples of this commercialization include the Visium technology by 10x Genomics, and the GeoMx platform by NanoString Technologies[152], which have significantly simplified the implementation of spatial transcriptomics techniques, reducing barriers to entry for researchers and clinicians alike. These advancements offer a promising avenue for advancing our understanding of TNBC's spatial and molecular heterogeneity, potentially leading to more effective treatments and enhanced patient care.

1.2.5. Tumor Microenvironment and its Role in Cancer Progression

TME plays a significant role in cancer biology, particularly in TNBC. It represents a complex and dynamic entity, encompassing several cellular components like lymphocytes, myeloid cells, fibroblasts, mesenchymal cells, endothelial cells, and non-cellular elements, including the extracellular matrix. Central to cancer biology, the transformation of somatic cells into malignant cells is closely linked with dynamic changes in the TME[153]. This transformation, signified by the onset of dysplasia and progression to cancer cells, is usually driven by genetic and epigenetic alterations that disrupt normal cellular functions[154]. Dysplasia, marked by abnormal cell growth, often precedes cancer, indicating a shift from regulated to uncontrolled cell proliferation. Concurrently, the TME undergoes significant changes through bidirectional cross-talk[153]. These changes denote a gradual shift from a state of homeostasis to one favoring tumor growth and survival. This includes immune response suppression, metabolic modification, angiogenesis, inflammation, and remodeling of the extracellular matrix, collectively promoting the survival and proliferation of cancer cells[153]. As a result, tumor cells and surrounding TME continuously adapt and interact to foster tumor growth, like PD1/PD-L1 biomarkers. Specifically, TME-resident growth factors, including Epidermal Growth Factor (EGF) and Vascular Endothelial Growth Factor (VEGF), along with the regulation of immune responses within the TME, which involves the production of cytokines such as Interferon-gamma (IFN-γ) and Tumor Necrosis Factor-alpha (TNF-α), promote Programmed Death-Ligand 1 (PD-L1) expression and T cell exhaustion, paving the way for tumor immune evasion[155].

1.2.5.1. Tumor Interactions with Microenvironment

Recent studies have significantly enhanced our understanding of the interaction between tumor cells and TME to promote tumor progression. These interactions occur through direct cell-to-cell contact, involving structures like gap junctions (GJs), or through the release of various soluble signaling molecules. Moreover, the communication can involve the deposition of ECM that affects the behavior of ECM-binding cells[156,157], thereby highlighting the complexity of the tumor microenvironment. In the realm of direct cell-cell communication, one key player is the presence of intercellular channels is the GJs which composed of connexin proteins. The alteration of connexins in tumor cells is linked to increased invasive behaviors[158], and tumor cells use these junctions to connect with stromal cells, promoting their own growth and survival[159]. Another cell-cell communication involves integrins, transmembrane receptors that mediate cell-ECM interactions[160]. These receptors are important for responding to environmental cues and facilitating cell adhesion and signaling. Integrins bind to various ECM components like fibronectin or collagen, and are capable of bidirectional signaling, thus acting as a bridge between intracellular and extracellular environments. Through integrin receptors, tumor cells are known to migrate toward areas with higher concentrations of ECM components like fibronectin, a phenomenon observed at tumor borders and near vascular structures[161,162]. Integrins can further induce ECM remodeling through the regulation of extracellular proteinases, such as metalloproteinases (MMPs)[163]. Consequently, this leads to changes in the ECM's composition and stiffness, thereby facilitating the formation of pre-metastatic niches and the process of metastasis[164]. Nonetheless, long-distance communication is facilitated by the dispersion of exosomes and small soluble molecules, such as cytokines, chemokines, and growth factors[165]. These substances function in both autocrine and paracrine ways by attaching to specific receptors. Consequently, this triggers a cascade of signaling pathways that influence cellular processes like survival, growth, and mobility, as well as the continued generation of soluble factors or the ECM[165].

1.2.5.2. Ligand-Receptor Analysis to Infer Tumor-Microenvironment Interactions

In the realm of oncology, the dynamic interplay between tumor cells and TME is pivotal in orchestrating the signaling cascades that drive tumor progression. This necessitates a comprehensive understanding of the cellular constituents within the TME and their interactive roles. Central to this understanding is the identification and quantification of specific ligand-receptor pairs expressed by each cell type to facilitate the tumor-TME crosstalk. These pairs serve as crucial communicative links, dictating the direction, intensity, and biological significance of interactions within the tumor landscape[166]. Different methods exist to uncover ligand-receptor interactions, such as Protein-Protein Interaction (PPI) assays[167]. Complementing these, proteomics and transcriptomics not only allow the detection of interactions but also confirm the presence of these proteins through evidence of their expression[166]. While proteomic-based methods offer the advantage of direct measurement, expression-based technologies often emerge as the preferred choice due to their comprehensiveness, availability, straightforward analysis, and adaptability to various sample types. With tools like bulk RNAseq, scRNAseq, and spatial transcriptomics, expression-based approaches provide a multifaceted view of ligand-receptor dynamics, offering insights at different levels of resolution. Emerging computational methods for deciphering cellular communication through transcriptional data are contingent upon the co-expression of ligand-receptor pairs, wherein each gene of a pair is expressed by one of the two interacting cells. These methods operate under the assumption that the levels of gene expression are reflective of protein abundance, thereby indicating the intensity of protein-protein interactions. However, this might overlook crucial factors such as post-translational modifications and the assembly of protein complexes, both of which play significant roles in protein interactions. Typically, the analysis of ligand-receptor interactions involves six steps, as was reviewed by Armingol et al.[166] (Figure 15); (i) measuring gene expression in cells or samples through NGS transcriptomics; (ii) preprocessing this data into a gene expression count matrix; (iii) compiling or sourcing from existing literature a list of known ligand-receptor pairs; (iv) refining the retrieved count matrix to focus on genes for these pairs; (v) computing an interaction score for each ligand-receptor pair based on expression values; and (vi) visualizing these scores using heatmaps and network diagrams to interpret intercellular communication.

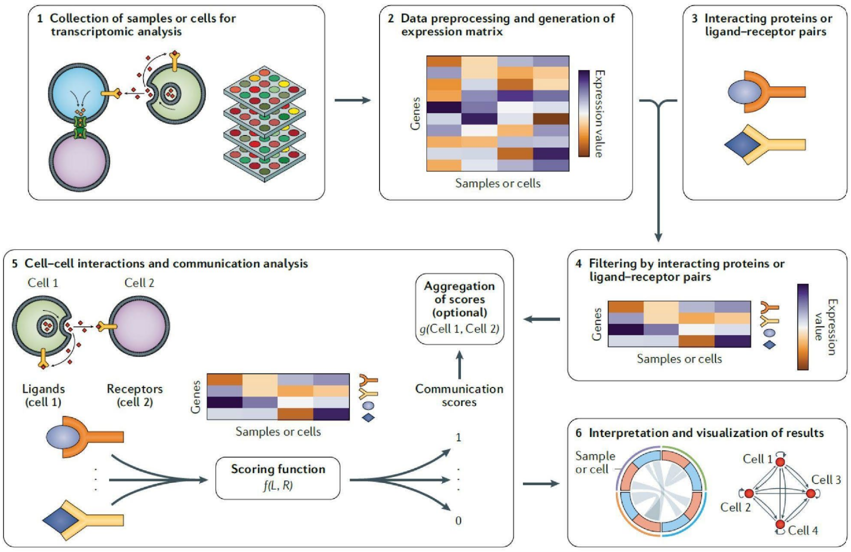

1 Collection of samples or cells for transcriptomic analysis
2 Data preprocessing and generation of expression matrix
Genes
Expression value
Samples or cells
3 Interacting proteins or ligand–receptor pairs
4 Filtering by interacting proteins or ligand–receptor pairs
Genes
Expression value
Samples or cells
5 Cell–cell interactions and communication analysis
Cell 1
Cell 2
Ligands (cell 1)
Receptors (cell 2)
Genes
Samples or cells
Aggregation of scores (optional)
g(Cell 1, Cell 2)
Communication scores
Scoring function
f(L, R)
1
0
6 Interpretation and visualization of results
Sample or cell
Cell 1
Cell 2
Cell 3
Cell 4

Figure 15. A step-by-Step Workflow of Ligand-Receptor Analyzing Workflow
1) Gene expression measurement using NGS transcriptomics; 2) Data preprocessing into a gene expression count matrix; 3) Compilation of a list of known ligand-receptor pairs; 4) Refinement of the matrix to focus on these pairs; 5) Calculation of interaction scores based on expression values; and 6) Visualization of these scores through heatmaps and network diagrams, providing insights into intercellular communication dynamics. Next generation Sequencing (NGS). Adapted from Armingol, Erick et al. Nature reviews (2021)[166].

Given the intricate nature of tumor-TME dynamics, the application of ligand-receptor analysis in cancer research has been pivotal not only to enrich the comprehension of tumor progression but also for the identification of potential diagnostic and therapeutic markers. Landmark studies, such as the Pan-Cancer Analysis of Ligand-Receptor Cross-talk[168], have shed light on these interactions across a spectrum of cancers. An additional study by Maffuid et al., utilized ligand-receptor analysis to elucidate the complex intercellular communications within the TME, aiming to understand how these interactions contribute to tumor progression and immune evasion[169]. In the context of colon adenocarcinoma, ligand-receptor interactions have shown promise in guiding treatment strategies[170]. Similarly, research in colorectal cancer has underscored the potential of ligand-receptor interactions in understanding cancer biology and developing treatment approaches[171]. Overall, ligand-receptor analysis represents a promising avenue for uncovering critical interactions within oncogenic pathways, potentially paving the way for improving diagnostic and therapeutic strategies. This holds particular significance for TNBC, given its limited current treatment options.

1.2.6. Hallmarks of Cancer

Despite the complex and genetically diverse nature of cancer, a universal characterization is achievable through the six hallmark traits identified by Hanahan and Weinberg in their seminal 2000 paper[172]. These hallmarks offer a conceptual framework for studying the intricate biological processes governing the development and progression of cancer. Significantly, the tumor-TME interactions play a pivotal role in facilitating the acquisition of these six hallmarks alongside epigenetics modifications and mutagenic events. i) Sustaining Proliferative Signaling. Cancer cells maintain unregulated growth through autocrine stimulation and external growth factors, activating pathways like BRAF/MAPK and PIK3/AKT. This is often triggered by mutations in genes like BRAF and PIK3CA[173]. ii) Evading Growth Suppressors. Cancer cells bypass growth suppression by impairing tumor suppressor genes such as RB1 and TP53, leading to unrestrained cell proliferation[174]. iii) Resisting Cell Death. Tumor cells avoid apoptosis by manipulating the BCL-2 family and inactivating TP53. They exploit autophagy for survival and use necrosis to promote tumor growth[175,176]. iv) Enabling Replicative Immortality. Cancer cells achieve limitless replication by upregulating telomerase to extend telomeres, contributing to genomic instability and influencing processes like WNT signaling[177,178]. v) Inducing Angiogenesis. Cancer promotes new blood vessel formation using angiogenic factors like FGF and VEGF, and can be influenced by oncogenes like RAS and MYC. vi) Activating Invasion and Metastasis. Alterations in cell adhesion and cytoskeletal dynamics, involving the loss of E-cadherin and changes in ECM adhesion, enable cancer cells to invade and metastasize. This is facilitated by the EMT process regulated by transcription factors and ECM proteins like fibronectin-1[179,180]. In addition to the six hallmark capabilities identified in their 2000 study, Hanahan and Weinberg in 2011[181] suggested two additional hallmarks, Reprogramming of Energy Metabolism and Evading Immune Destruction. These suggested hallmarks complement the original six by addressing the cancer cells' ability to alter metabolism for sustained growth and their strategies to evade detection and destruction by the immune system, respectively. Collectively, the comprehension of cancer hallmarks has undergone substantial growth, facilitating the investigation into intricate genetic and molecular landscapes that drive cancer development and progression.

1.2.7. Major Challenges Addressed in This Dissertation

As discussed above, in the realm of TNBC, one of the foremost challenges is its intrinsic heterogeneity. TNBC is characterized by complex and dynamic interactions between diverse cell populations, which significantly affect its heterogeneity and present a formidable obstacle in understanding and effectively treating this disease[107,108]. Despite the extensive research dedicated to TNBC, the development of an efficient and comprehensive therapy remains elusive. This gap underscores the pressing need for continued exploration and innovation in treatment strategies, particularly in the face of TNBC's variable nature.

A pivotal aspect of managing TNBC involves the precise evaluation of PD-L1 expression status for an effective application of ICI therapies, which are now fundamental to TNBC's standard of care[120]. However, the variability in Immunohistochemistry (IHC)-based PD-L1 diagnostic tests and the wide dynamic range of PD-L1 expression levels obstruct patient selection[130], resulting in inconsistent therapeutic outcomes and fluctuating ORR[129]. This variability emphasizes the challenges of employing PD-L1 as a reliable diagnostic marker and highlights the need for standardized clinical approaches to enhance the sensitivity and reproducibility of diagnostic tests. Establishing such a uniform framework is essential for refining patient selection and enhancing the precision of patient selection for anti-PD-1/PD-L1 therapies, consequently enabling the development of more effective treatment strategies that improve patient outcomes

Addressing these challenges necessitates the adoption of innovative strategies, such as spatial transcriptomics. This technology offers precise mapping of tumor environments, thus directly addressing the limitations of current PD-L1 diagnostic methods. Unlike traditional transcriptomic analyses like scRNAseq, which lose spatial context during the cell dissociation procedure, spatial transcriptomics retains crucial locational information within tissues. This capability not only enhances the resolution of transcriptional data but also facilitates a deeper understanding of the complex interactions between tumor cells, immune cells, and their milieu. By enabling the identification of novel diagnostic markers and therapeutic targets, spatial transcriptomics holds the promise of refining PD-L1 testing. This refinement could be achieved through the integration of new diagnostic markers in a clinical workflow, thereby fully leveraging spatial transcriptomics' potential to navigate the complexities of TNBC.

Collectively, this project addresses a spectrum of challenges in TNBC research, ranging from the inherent heterogeneity of tumors to the intricacies of PD-L1 diagnostic tests and the emerging role of spatial transcriptomics in enhancing our understanding of this complex disease.

To effectively navigate these challenges, we have developed an efficient and cost-effective clinical workflow that harnesses the innovative potential of spatial transcriptomics to dissect the multifaceted landscape of TNBC, with a particular focus on understanding PD-L1 expression. Our approach seamlessly integrates spatial transcriptomics, bulk RNAseq, and immunohistochemistry (IHC). Notably, it offers a distinct advantage by requiring minimal spatial transcriptomic input samples and being compatible with standard laboratory equipment. This methodology has enabled us to provide systems-level insights into the transcriptomic and cellular architecture of tumors and demonstrated how spatial expression data augments traditional histological annotation. Consequently, this allowed us to intricately delineate sub-tumoral variation and identify unique ligand-receptor interactions between tumors and TME that corresponded to the immunogenic capacity of TNBC based on PD-L1 status. At the core of our research is the identification of lymphocyte antigen 6 family member D (LY6D) as a potential novel diagnostic marker that could enhance PD-L1 testing in TNBC. Essentially, our research underscores the transformative impact of spatial transcriptomics on decoding TNBC architecture and molecular variability, paving the way for novel therapeutic strategies.

2. Materials and Methods

2.1. First Project - An Integrated Screening to Infer Transcription Factor Regulatory Networks Governing Cell Fate Decisions

2.1.1. Candidate Transcription Factor

We identified promising candidate TFs for our integrative screening through a de novo discovery process (Figure 17). Our multi-tiered approach prioritized TFs influencing cellular plasticity, starting with those showing ≥20 reads per kilobase per million mapped reads (RPKM) in fibroblast cell lines, Human Lung Fibroblasts (HLF), retinal pigment epithelium (RPE), human foreskin fibroblasts (BJ). We aimed to select pioneer TFs, considering their expression in different germ layers. The top 250 TFs with human-specific expression were shortlisted after comparing mouse (FANTOM5)[182] and human (ROADMAP/ENCODE)[183,184] epigenetic datasets. These TFs were evaluated using the Transcription Factor Epigenetic Remodelling Activity (TERA) score[185] for their epigenetic impact on cellular fate. Additionally, we used Mogrify[186], a gene expression-based network algorithm, to predict TFs essential for cellular fate transformation in HLF, RPE, and BJ human primary cell lines, by identifying TFs influencing 95% of genes required for 132 different target cell types. The resulting TF list was used to purchase plasmids containing the gene of interest from the ORF Library Clones of the Broad Institute, and using the following specific filters: 1) vector: pLX317 2) % of insert sequenced (any): ≥30%, 3) Intended mutant: NO, 4) Best Match is mutant: NO, 5) Best Match Taxon ID: 9606, 6) Best Match % (Nucl): ≥90%, 7) Best Match % (Prot): ≥90%. To broaden our study's scope, we included an additional 54 TFs that, while initially not meeting our strictest criteria, held significant scientific value. This comprehensive method yielded a final list of 277 candidate TFs. At present, we successfully screened a subset of 130 TFs that completed the cellular conversion process and stand the sequencing quality control (QC) thresholds.

2.1.2. Plasmids Preparation

The ORFs of all TFs were integrated into the pLX-317 lentiviral vector backbone. A total of 223 TFs were sourced from the Broad Institute ORFeome collection, while an additional 54 TFs were cloned starting with a green fluorescent protein (GFP) containing pLX317 vector. The cloning process was initiated by digesting the vector with NheI and EcoRV restriction enzymes. Subsequently, the insert (comprising the TFs ORFs) was efficiently cloned using the NEBuilder® HiFi DNA Assembly Master Mix (Catalog #E2621L).

2.1.3. Lentiviral Production and Transduction

Lentiviral production was conducted in X293T cells using a refined protocol developed in our laboratory, which enhances the efficiency and consistency of the lentiviral preparations. Cells, at 95-99% sub-confluency, were transfected in 6-well plates with Lentiviral and packaging plasmids. The next day, the medium was switched to DMEM/F12 with 20% heat-inactivated FBS and 1x Glutamax. This modification from the standard protocol, employing 20% FBS, enhances cell nutrition under stress, and the inactivated medium prevents complementary protein activity against virions, increasing production efficiency. After 48 hours, the supernatant was harvested.

For the transduction process, BJ fibroblast telomerase (BJ-T) cells were exposed to 10uL of the lentiviral preparation in a medium enriched with heat-inactivated serum and Polybrene, to facilitate efficient viral entry into the cells.

2.1.4. Cell Culture

To enhance the cellular lifespan and optimize growing conditions, BJ cells underwent immortalization via human telomerase reverse transcriptase (hTERT) lentiviral transduction. The cells were cultivated in DMEM/F12 (Euroclone, Catalog # ECM0095), enriched with a supplement mix. This mix included 10% FBS (Euroclone, Catalog # ECS0186L), 1x Glutamax (Thermo Fisher Scientific, Catalog #35050061), 1x Non-Essential Amino Acids (NEAA) (Thermo Fisher Scientific, Catalog #11140035), 2-Mercaptoethanol (Thermo Fisher Scientific, Catalog #21985023), and Hygromycin B (Invitrogen, Catalog # 10687010), providing a nutrient-rich environment conducive to robust cell growth.

2.1.5. Cellular Conversion Protocol

The cells underwent transduction with lentiviral vectors carrying V5-tagged TFs ORF, in the presence of 4 μg/ml polybrene (Sigma, Catalog #TR-1003-G) to enhance viral entry. Two days post-transduction, the culture medium was supplemented with 1 μg/ml puromycin (Thermo Fisher Scientific, Catalog #A1113803) for an additional two days to select successfully transduced cells. Following the antibiotic selection phase, the cells were maintained and allowed to grow and differentiate until day 9, with the medium being refreshed every two days. On day 9, the serum concentration in the medium was reduced to 2%, and the cells continued to grow until day 11. At the completion of this process, cells were either harvested or prepared for further downstream analysis, depending on the specific requirements of the subsequent experiments.

2.1.6. Immunofluorescence

Upon completing the cell conversion protocols, the cells were fixed with 4% paraformaldehyde (PFA) for 10 minutes at room temperature. This was followed by permeabilization using 100% Ethanol for another 10 minutes. Blocking was then performed for 40 minutes using a solution composed of 5% Bovine Serum Albumin (BSA) and 0.5% Triton X-100 in phosphate buffered saline (PBS). The cells were subsequently stained with the appropriate primary antibody (details provided below) for 2 hours at room temperature or overnight at 4°C. Post-primary antibody incubation, the samples underwent two washes with the blocking solution, followed by staining with the corresponding secondary antibody at a dilution of 1:500, DAPI at a dilution of 1:1000, or with a conjugated primary antibody, based on the primary antibody used, for 1 hour. Finally, the samples were washed twice with PBS and stored at 4°C until further analysis. The samples were then qualitatively analyzed using the Opera Phenix Plus High-Content Screening System.

Primary antibodies in used: i) Anti-V5 (clone 1H6, MBL, Catalog # M167-3), Anti-Tubulin Alexa Fluor 488 (clone 22833, Thermo Fisher Scientific, Catalog #MA3-22600-A488)Cell Mask (Abcam, Catalog #C10046)

Secondary Antibodies Used:

- Alexa Fluor 568 anti-mouse (Catalog #A-11004)
- Alexa Fluor 488 anti-mouse (Catalog #A28175)

2.1.7. RNA Sequencing

2.1.7.1. Sample Preparation and Library Construction

Before library preparation, RNA samples were extracted with RNadvance cell V2 kit (Beckman #A47943), quantified with Qubit™ RNA High Sensitivity (HS) (Q32852) and diluted to 50 ng/uL. The libraries were generated with QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen) according to the manufacturer's specifications or by halving the originally recommended volumes without compromising library quality. One or two sets of 96 library pools were sequenced on a SE100 cycles SP Novaseq flow-cell (Illumina).

2.1.7.2. RNA Sequencing Preprocessing

Sequencing data were preprocessed by Next Generation Diagnostic srl proprietary NEGEDIA Digital mRNAseq pipeline (v2.0). This pipeline encompasses several key steps, i) quality filtering and ii) trimming of reads, iii) alignment to the reference genome, and iv) gene counting. To enhance the analysis, samples were further filtered based on specific criteria, i) TF-transgene expression below 5 counts per million (CPM), ii) fewer than 10,000 detected genes, iii) a uniquely aligned reads ratio below 80%, and iv) a minimum of three samples per condition. Additionally, genes that did not demonstrate more than 3 CPM in at least one condition were omitted from the analysis. For normalizing the raw transcript counts, the DESeq2 package (version 1.38.3)[187] was utilized, employing the 'median of ratios' method for normalization. This approach ensures a robust and accurate adjustment of the transcriptomic data, facilitating reliable downstream analysis.

2.1.7.3. Infection Noise Inference

n the 'dose-dependent Multiplicity of Infection (MOI) experiment', we conducted a controlled study to model the noise structure derived from the infection procedure in our data. This was achieved by infecting BJ cells with varying levels of MOI, specifically at ratios x0.05, x0.25, x0.5, x1, x2, x4, and x8 of a GFP-viral vector. These levels were chosen relative to the original MOI used in the TF vector transduction protocol, as (detailed in section 2.1.3). After a cultivation period of 11 days, we collected samples from these cultures for RNAseq analysis, aiming to understand the dose-dependent responses of the cells to the viral vector infection.

To identify genes associated with varying infection intensities, we compared gene expression in cells infected at MOI x8 versus x0.05 using DESeq2 (version 1.38.3)[187]. Genes showing significant upregulation (log fold change (LFC) > 2, false discovery rate (FDR) < 0.05) in the MOI x8 condition were grouped into eight clusters through C-means clustering[188]. This clustering approach, typically employed in time-point studies, was adapted to analyze gene expression across different MOI levels.

Notably, cluster number 3, containing 64 genes, exhibited a correlation between increased gene expression and rising MOI levels (Figure 21). To ascertain the relevance of these genes in the infection process, we conducted gene set enrichment analysis using EnrichR package (version 3.2)[189], referencing the 2023 Gene Ontology (GO) Biological Process (BP) database[190,191].

We subsequently employed these identified genes to assess the infection level (i.e., infection score) in each TF sample (Figures 23). This involved modifying the script of the AddModuleScore function in the Seurat package (version 4.3.0)[192], a technique based on the strategy described by Tirosh et al.[193]. This adaptation enabled a more precise assessment of infection scores in our experimental samples.

2.1.7.4. Differential Expression Analysis and Pathways Enrichment

Prior to the DEA, we merged all RNAseq data from both the TF-driven transdifferentiation screening and dose-dependent MOI exploratory and performed batch effect correction (BEC) analysis using the Combat-seq function in the sva package (version 3.46.0)[194]. Notably, Combat-seq algorithm[194] distinguishes itself from many BEC methods by ensuring that the adjusted data remain as integer counts. This aspect is critical as it maintains the compatibility of the data for downstream DEA software like edgeR[195] and DESeq2[187]. For BEC using ComBat-seq, we input RNAseq count matrices from two experimental batches: i) TF-driven transdifferentiation screening, and ii) dose-dependent MOI exploratory, along with a corresponding batch separation vector. GFP samples from both experiments were utilized as a reference in the BEC model, forming the intercept column to normalize the overall mean and variance across batches. Notably, TF samples were not included in the intercept column to avoid confounding influences in the batch effect estimation, as they are not comparable due to their prompt different biological outputs (Figure 16).

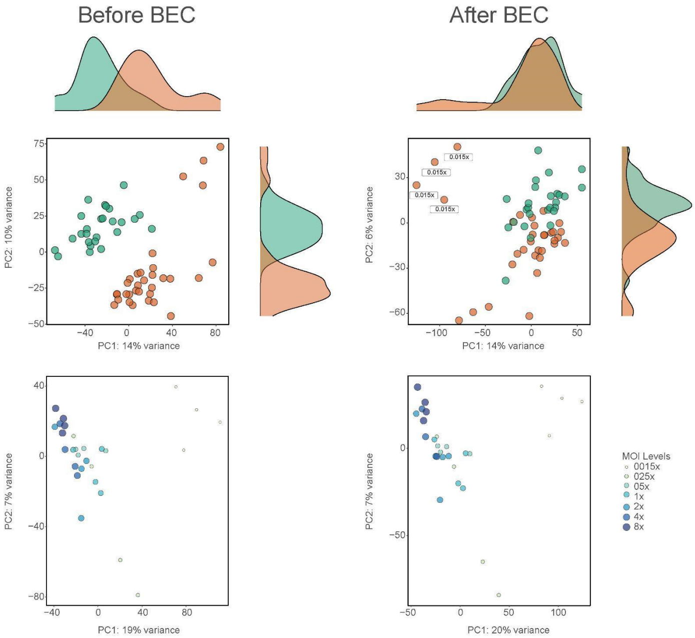


Figure 16. Batch Effect Correction Reduces Technical Artefacts While Preserving Biological Variability PCA plots demonstrate the variability of data before (left) and after (right) BEC. At the top, each dot represents GFP-infected samples from the screening assay (in green) and samples from the Dose-Dependent MOI Experiment (in orange), illustrating a reduction in batch effect variability. At the bottom, dots representing different samples infected at various MOI levels, revealing a clear directionality aligned with the MOI levels both before and after batch effect correction. Principal component analysis (PCA); batch effect correction (BEC); green fluorescent protein (GFP); Multiplicity of infection (MOI).

The DEA was conducted using the DESeq2 package (version 1.38.3)[187], comparing pairs of TF-samples to GFP-control with corresponding infection scores (Figure 23). For each TF condition, we averaged the infection scores across all replicates. These averages were then paired with GFP replicates at specific MOI concentrations. Pairing was based on minimizing the difference (delta) between the mean infection scores of the TF condition replicates and those of the GFP replicates at each MOI concentration. Likewise, the 64 infection-associated genes were excluded from downstream analyses.

The DEA was then conducted using the DESeq2 package (version 1.38.3)[187]. More specifically, we compare pairs of TF samples to GFP-control exhibiting corresponding infection scores (Figure 10 300tf results). More specifically, using the inferred 64 genes linked to the infection procedure, we scored the infection efficacy for each sample in the TF-driven transdifferentiation screening and GFP-controls exploratory from the dose-dependent MOI exploratory and averaged them for each condition (a given TF or GFP-control). GFP samples were categorized into eight classes based on their respective MOI levels. TF samples were paired with specific MOI classes that showed the smallest difference in average infection scores at each MOI concentration. This meticulous pairing was essential for ensuring the accuracy of the differential gene analysis. This careful matching was key to preventing the misattribution of differentially expressed (DE) gene changes to TF activity. Likewise, the 64 infection-associated genes were excluded from downstream analyses.

Gene set enrichment analysis was done using the fgsea R package (version 1.24.0)[196], referencing the implemented C5 Molecular Signatures Database (GO). For each TF condition, we ranked genes by the product of LFC-(log10(FDR)). Focusing on NEUROD6, MYOD1, and TP63, we identified and visualized the top 20 pathways, exhibiting the highest Normalized Enrichment Score (NES) coupled with the lowest p-value (NES-(log10(FDR)).

Gene set enrichment analysis using the fgsea R package (version 1.24.0)[196], referencing the implemented C5 Molecular Signatures Database (GO). For each TF condition, we ranked genes by the product of $LFC \times -(log10(FDR))$. Focusing on NEUROD6, MYOD1, and TP63, we identified and visualized the top 20 pathways, exhibiting the highest normalized enrichment Score (NES) coupled with the lowest p-value $(NES \times -(log10(FDR))$

2.1.7.5. Assessing Similarity Scores Based on Enrichment of Mutual Target Genes

To infer the similarity between screened TFs without setting an arbitrary threshold on differential gene expression values, we proposed the computational strategy from 'Mode of Action by NeTwoRk Analysis' (MANTRA)[197] to our data requisites (figure 13 results 300TF). We first created a matrix of genes LFC values for all TFs. Then, we scaled these values across all TFs using z-scoring. This step reduced the influence of genes that are up or downregulated universally across all TFS, which could distort our MANTRA analysis by indicating false TF similarities. Next, we identified an 'optimal' gene signature for each TF, selecting the top 250 overexpressed and bottom 250 most down-regulated genes based on z-scored LFC values. Using MANTRA's gene set enrichment approach, we analyzed if the 'optimal' gene signature of one TF consistently appeared at the extreme ends in the other TF's differentially expressed genes list and vice versa. We averaged the enrichment scores between each TF pair, creating a singular metric to quantify their relationship. This metric was then applied to construct an Euclidean distance matrix, serving as the basis for hierarchical clustering analysis.

2.1.8. ATACseq

2.1.8.1. Sample Preparation and Library Construction

ATACseq experiments were conducted by combining the methodology established by Bao et al. in 2015[198] with the OmniATAC approach developed by Coerces et al. in 2017. This allowed adapting the omniATAC protocol to high-throughput screening experiments.

Briefly, cells were permeabilized in situ prior to detachment from the 96-well plate. Permeabilization was accomplished by treating the cells with an ice-cold solution consisting of 0.1% NP40, 0.1% Tween 20, and 0.01% digitonin. The plates were left on ice for 3 minutes during the permeabilization process. Subsequently, a solution of 0.1% Tween 20 was used to rinse the cells and remove mitochondria. The supernatant was then discarded, and the cells were treated with 25 µL of tagmentation solution, following the methodology outlined by Coerces et al., and incubated with shaking at 37°C for 1.5 hours.

Tagmented nuclei were lysed using the Zymoclean DNA binding solution and transferred to a PCR 96-well plate. The tagmented DNA was subsequently extracted using 1.8 volumes of Ampure DNA clean beads. Finally, the ATAC-seq library was generated according to the protocol described by Coerces et al. in 2015 (omniATAC). And purified using ampure beads.

Once pulled samples were sequenced on a SE100 cycles SP Novaseq flow-cell (Illumina) with the following specifics: Read-1 50 bp, Read-2 50 bp, Index-i5 8 bp, Index-i7 8 bp

2.1.8.2. ATACseq Preprocessing Pipeline

Preprocessing ATACseq data lacks a standardized, universally accepted pipeline and instead involves a mix of stand-alone tools, many of which were initially developed for techniques like ChIPseq and DNaseseq. While there are emerging best practices and consensus on certain steps, such as alignment and peak calling, the field is dynamic, with continuous development leading to new methodologies and tools, contributing to the diversity of approaches in data preprocessing. Thus, herein, we developed a comprehensive ATACseq preprocessing pipeline by combining the best practice tools and approaches from various established pipelines and research studies. In the development of this pipeline, the primary framework was based on the established NextFlow pipeline[199]. Subsequent refinements and enhancements were systematically incorporated following an extensive review of the literature, notably integrating key methodologies outlined by Ou et al. (2018)[200], Yan et al. (2020)[71], and ENCODE's pipeline[201], which were selected for their demonstrated efficacy in improving ATACseq data quality assessment. The ATACseq preprocessing pipeline encompasses 18 steps as delineated in (see Table S1). Briefly, first the sequenced data undergoes demultiplexing, followed by the removal of adaptors. Subsequently, the data is aligned to the human reference genome. After alignment, the reads are filtered according to specific ATACseq quality criteria. The final stages involve Irreproducible Discovery Rate (IDR) peak calling, peak counting, and peak annotation, as referenced in (see Table S1).

2.1.8.3. Peak Differential Analysis

Quality Control (QC)

In our ATACseq analysis, samples underwent systematic QC filtering based on established ATACseq quality thresholds. While adhering to general guidelines[71,202], we adopted slightly more lenient criteria due to the extensive nature of our study, which involved dual replicates for over 130 TFs and limited sequencing depth. Key quality metrics included: i) fraction of reads in peaks (FRiP): Samples with FRiP score below 0.3 were excluded. This threshold, recognized by ENCODE standards, ensures minimal background noise in our dataset, with a FRiP score above 0.3 indicating acceptable quality; ii) number of peaks in IDR Files: According to ENCODE standards, a minimum of 50,000 peaks within an IDR peak file is generally required. However, due to our lower sequencing depth, we halved this threshold to 25,000 peaks; iii) number of reads for open chromatin detection: For mammalian species, a standard minimum of 50 million reads is recommended for effective open chromatin detection and differential analysis. In our study, this threshold was adjusted to 20 million due to the sequencing depth limitation. For consistency and reliability, all conditions were tested with two replicates. If one replicate failed to meet the QC criteria, its paired replicate was also excluded from further analysis.

Normalization and Peak Differential Analysis

Initially, a filtering criterion was applied to the peaks, requiring a minimum of 10 reads in at least one condition to ensure no peak had zero reads. Subsequently, we utilized the DESeq2 package in R (version 1.38.3)[187] to normalize the count peak matrix. It is noteworthy that many ATACseq peak differential analyses presuppose a negative binomial (NB) distribution and necessitate biological replicates for dispersion estimation. Consequently, these analyses often rely on RNAseq DE analysis packages such as

DESeq2. Thereafter, to discern regions where chromatin accessibility was significantly augmented in samples treated with the TF of interest, in comparison to control samples introduced with GFP, we conducted a pairwise peak differential analysis. This analysis employed a one-sided alternative hypothesis test using the altHypothesis="greater" option in DESeq2. Peaks were designated as differentially open chromatin regions only if they exhibited an $LFC > 3$ and $Benjamini - Hochberg\ p - adjusted\ value\ < 0.01.$

2.1.8.4 Differential Motif Enrichment Analysis

Peaks over 2500 base pairs in differential open chromatin regions (OCRs) were filtered out. To normalize predictive TF binding site distribution, we evaluated size distribution, with the median at ~500bp. Peak lengths were standardized to 500 bp, extending 250 bp from the center. We matched these regions with a standard model for TF DNA binding specificities using a comprehensive positional weight matrix (PWM). We retrieved human TF binding motifs from JASPAR 2020[66], HOCOMOCO-v11[67], and TRANSFAC 2014[65] databases, using MEME file formats and the transfac2meme tool (version 5.1.1). TF motif matching was performed with the FIMO tool (version 4.10.2)[203] using my PWM as a reference and kept motif alignments to retrieved OCRs below a significance of 1e-04. We then log-transformed the motif alignments p-value, and aggregated r summed them for each stand-alone peak, resulting in the motif enrichment map. This was necessary due to multiple motif occurrences per peak. To rank the motif enrichment for each TF distinctively, we applied 'Sparse Partial Least Squares regression' (sPLS)[204] analysis (spls library in R version 2.2-3). More specifically, we compared the motif binding enrichment matrix, which served as the predictor, with the normalized peak count matrix, including only the differential OCRs. This resulted in the following regression model:

$y=\beta 0+\beta 1x$

The term x represents the predictor variable - a given motif enrichment score. The dependent variable, y, is the peak count of a given OCR. The coefficient β1 indicates the degree of influence each predictor has on y.

The delta between the estimated beta coefficients of each motif in a given TF and the GFP-control was used as a weighting factor to assess its importance in the activity of the examined TF:

$\Delta\beta 1=\beta 1_{TF}-\beta 1_{GFP}$

2.1.8.5. Motif Reshape Analysis

To assess pairwise motif similarities, I utilized the TOMTOM tool (version 5.4.1)[205], employing the parameter "-thresh=1" to collect comprehensive alignment results, using the merged PWM as a reference. Subsequently, for each TF, motifs with an alignment q-value less than 0.05 were classified as “Expected motifs,” while those not reaching this threshold were labeled as “Observed motifs.” Additionally, for each motif, I assigned beta coefficient scores derived from motif differential analysis, categorizing them under “Expected motifs” and “Observed motifs.” Ultimately, TFs were ranked based on their motif reshape score, which was calculated by combining the delta of the mean motif enrichment scores between 'Expected' and 'Observed motifs' and the statistical significance of this difference.

2.1.9. Multiomics Data Integration

2.1.9.1. Transcription Factor Activity Scores

The normalized count matrices from RNAseq and ATACseq were refined to include only relevant genes or peaks. In the RNAseq count matrix, we maintained DE genes with an LFC>1 and a p-value<0.05. Similarly, in the ATACseq count matrix, we focused on peaks exhibiting an LFC>2 and a p-value<0.05. Subsequently, we calculated the mean expression of each gene or peak across replicates for each TF and GFP-control. Following this, principal Component Analysis (PCA) analysis was performed, and we calculated the Euclidean distance between the TFs and GFP-control within the latent space defined by the first two principal components (PC1 and PC2). Finally, to rank the overall TF activity, we scaled the Euclidean distances derived from both RNAseq and ATACseq data, combining them to yield a consolidated score.

2.1.9.2. Similarity Network Fusen

To adeptly navigate the complexities inherent in multiomic data integration, we adopted the SNF method[84] Unlike the conventional approach of using count matrices, we input enrichment matrices from both RNAseq and ATACseq to highlight biologically significant aspects of TF similarities. Specifically, we utilized the target gene enrichment matrix for the RNAseq omic level and the motif enrichment matrix for the ATACseq omic level. We constructed individual weighted network graphs for each omic level and identified the minimal number of edges necessary to maintain all nodes intact. We set thresholds for edge weights below which they were disregarded: less than 0.09 for RNAseq-based data and less than 0.24 for ATACseq-based data. Consequently, all values in the similarity matrices falling below these thresholds were set to zero. These refined matrices were then employed in the SNF process, which constructed standalone networks from RNAseq and ATACseq data. These networks were then fused through a message-passing process, a technique that maintains strong connections between factors while filtering out weaker ones. This selective fusion underscores the most pertinent interactions among TFs. Finally, in the resulting fused weighted network, we again applied the principle of retaining the minimal number of edges necessary to preserve all nodes, setting a threshold of less than 0.05. This approach ensured a focus on the most significant and robust connections in our multiomic data analysis.

2.2. Second Project - Spatial Transcriptomics Reveals Sub-Tumoral Identities and Novel Diagnostic markers in Triple Negative Breast Cancer With Immune Evasion Capacity

2.2.1. Samples Selection

Clinical samples for spatial, bulk RNAseq, and IHC were collected from the files of the Division of Pathology, European Institute of Oncology (IEO), Milan, Italy, under project registration number UID 2886, which includes appropriate informed consent and approval from the local ethical board. Samples were selected based on PD-L1 CPS gold standard thresholds[206,116], tumoral content, and, in the case of spatial profiling, a capture area at the edge of tumor and normal breast tissue.

Clinical samples for bulk RNAseq (LY6D/CD274 expression levels in a larger patient cohort - at least 50% cellularity TNBCs) are provided under a research agreement with Next Generation Diagnostic Srl (NEGEDIA), which holds appropriate ethical approval and/or MTA. For confidentiality and ethical restrictions, NEGEDIA does not hold the authorization to share such sequencing data.

2.2.2. Spatial Transcriptomics Library Preparations And Data Processing

For spatial transcriptomics, FFPE tissue blocks of sufficient quality obtained from 2 TNBC tumors positive and negative for PD-L1 were sectioned to a thickness of 5 μm and processed through 10X Visium technology (10X Genomics Inc). Briefly, samples were mounted on Visium Spatial capture probe slides, then underwent deparaffinization, staining with HE, imaged, and de-crosslinked as per the given protocol. Subsequent library preparation was carried out using the Visium Spatial Gene Expression Reagent Kits according to manufacturers' specifications. The final purified libraries were sequenced on a NovaSeq 6000 sequencing system (Illumina Inc.) according to Visium sequencing strategy.

Sequencing data preprocessing was done by using Spaceranger (version 1.3.1 - 10X Genomics Inc) to perform sample demultiplexing, alignment, tissue detection, fiducial frame detection, and unique molecular identifiers (UMIs) counting. The obtained data from Spaceranger were analyzed, visualized and integrated with histological annotation using Loupe Browser (version 6.0.0 - 10X Genomics Inc).

The output of the spaceranger pipeline was imported into R (version 4.2.0) and analyzed using Seurat R package (version 4.3.0)[207]. Utilizing the Load10X_Spatial function, both spot-level expression data and the corresponding tissue slice images for each sample were retrieved. The data underwent normalization via the SCTransform method, followed by dimensionality reduction using the UMAP (Uniform Manifold Approximation and Projection) approach, adopting Seurat's default parameters. For initial clustering, the FindClusters function was employed, setting the resolution parameter at 0.6 for PD-L1 negative samples and 0.8 for PD-L1 positive samples. Subsequently, the histological annotation CSV file from Loupe Browser was imported and integrated with the Seurat object using AddMetaData function.

2.2.3. Spatial Transcriptomics Data Analysis

Deduce tumor and stromal signatures from spatial data. The ESTIMATE R package (version 1.0.13)[208] was used to leverage gene signatures obtained from stromal cells to assess the relative enrichment of the Stromal Score. The computation of Tumor Purity followed the methodology outlined by Yoshihara et al[208]. The application of ESTIMATE was conducted on normalized expression data extracted from individual spatial spots. The output of the analysis yielded corresponding Stromal Scores and Tumor Purity values for each spatial spot, facilitating their representation through spatial visualization or in the form of a violin plot.

Gene expression clustering. Clustering was performed using the FindAllMarkers function in Seurat package (version 4.3.0)[207] with parameters set at only.pos=TRUE, min.pct=0.25, and logfc.threshold=0.25. For each cluster we reported the top 10 marker genes, prioritizing them based on the product of - log10 of Bonferroni-corrected p-values and log-fold change (LFC).

We then identified the top 10 marker genes for each cluster, ranking them by the product of LFC*-log10(adjusted-p-value) and conducted text mining to relate each cluster to a specific biological context.

Pathway Enrichment analysis. Differential gene expression analysis was performed using the FindMarkers function in Seurat package. Marker genes were identified for individual clusters annotated as tumor in comparison to all other surrounding clusters, and vice versa. After isolating the differentially expressed genes, we proceeded with the Gene Set Enrichment Analysis (GSEA) utilizing the fgsea R package (version 1.24.0) using the hallmark gene sets from the Molecular Signatures Database (MSigDB)[209]. For each distinct cluster, the foremost five pathways were chosen based on their rank, determined by the product of the -log10 of Bonferroni-corrected p-values and NES values.

Spatial visualization of biological pathways enrichment scores. For all pathways identified as enriched through the GSEA analysis, we further assessed their spatial enrichment profiles to discern whether they were predominantly enriched in the tumor or in its surrounding environment. The pathway score was computed for each spatial spot and for every pathway using the Seurat function 'AddModuleScores', which employed the gene set specific to each pathway[193].

Ligand Receptor analysis. From a list of ligand-receptor pairs[210], we filtered out those couples that were not experimentally validated and were not detected in our data. Cluster-to-cluster interaction scores were computed as shown in Panariello et al., 2023[211]. Interactors that demonstrated significance within the same clusters' class, either Tumor or Surrounding, were disregarded. Briefly, the average gene expression value of a ligand in a certain cluster was multiplied by the average value of its related receptor in another one. Significance was assessed with empirical p-value, generating a null distribution of 1000 permutations on the association between spots and clusters. Through the calculation of mean interaction scores for each cluster (see materials and methods), we succeeded in classifying ligand-receptor pairs into discrete Interaction Modules (IMs) through hierarchical clustering analysis.

2.2.4. Bulk RNA Sequencing

Total RNA was extracted from FFPE slides using the Maxwell RNA FFPE kit (Promega Corp.) and quantified using the Qubit 4.0 fluorimetric Assay (Thermo Fisher Scientific). Libraries were prepared from 250 ng of total RNA using the NEGEDIA Digital mRNA-seq clinical grade sequencing service (Next Generation Diagnostic srl)[212], which included library preparation, quality assessment and sequencing on a NovaSeq 6000 sequencing system using a single-end, 100 cycle strategy (Illumina Inc.).

Sequencing data were analyzed by Next Generation Diagnostic srl proprietary NEGEDIA Digital mRNA-seq pipeline (v2.0), which involves a cleaning step by quality filtering and trimming, alignment to the reference genome and counting by gene. The DESeq2 package[187] (version 1.38.3) was used to normalize raw transcript counts and perform differential expression (only on genes >5 CPM). Genes with an adjusted p-value less than 0.05 (using the Benjamini-Hochberg procedure) and LFC>2 were considered significantly differentially expressed.

2.2.5. Integration Of Spatial Transcriptomics With Bulk RNA Sequencing For The Identification And Ranking Of Candidate Diagnostic markers

Spatial differential gene expression analysis was performed using the FindMarkers function from the Seurat package (version 4.3.0)[207] with parameters set as min.pct=0, logfc.threshold=0, min.cells.feature=0 , min.cells.group=0. Marker genes were identified for spatial spots annotated as tumor in comparison to all other spatial spots annotated as surrounding. These analyses were executed separately for both PD-L1 positive and PD-L1 negative samples. Subsequently, the outcomes of the spatial differential analyses from both sample sets were integrated with the DEA results from a comparison between the bulk RNAseq of PD-L1 positive and PD-L1 negative samples. 31 genes resulted as being upregulated in PD-L1 positive tumor areas and bulk FFPE samples. Of these, only those genes that exhibited downregulation in the tumor regions of the PD-L1 negative spatial transcriptomics sample or were not expressed at all were earmarked as potential diagnostic marker candidates. This led to the identification of eight key genes out of which two discovered as outliers (see Table S2). In order to further prioritize the eight candidate diagnostic markers, we assessed their LFC values within each of the PD-L1 positive sub-tumor clusters and examined their variance across the different clusters. The comparison between the average LFC values and their variance across sub-tumor clusters enabled us to gain a better understanding of not only the diagnostic marker candidate specificity but also the consistency across the tumor area. This allowed us to select the top diagnostic marker gene that shows the best ratio between sub-tumor LFC variance and averaged LFC.

2.2.6. Immunohistochemistry Analysis And PD-L1 Expression Quantification

For clinical PD-L1 evaluation, four-microns thick FFPE sections of each block were subjected to PD-L1 IHC using the 22C3 pharmDx CE-IVD assay on a Dako Autostainer Link 48, according to the manufacturer’s instructions, as previously described[213]. Briefly, we used the 22C3 PharmDx assay (mouse monoclonal primary anti-PD-L1 antibody, prediluted, clone 22C3; Dako, Carpinteria, CA, USA) on the Dako Autostainer Link 48 with the EnVision 3,3′-Diaminobenzidine (DAB) Detection System (Agilent Technologies, Santa Clara, CA, USA). All the evaluations were performed on whole slides. The CPS was determined as the number of PD-L1-positive infiltrating tumor cells, lymphocytes, and macrophages divided by the total number of viable infiltrating tumor cells, multiplied by 100. Any perceptible and convincing partial or complete linear membranous staining of viable infiltrating tumor cells that were perceived as distinct from cytoplasmic staining was considered to be positive PD-L1 staining and was included in the scoring. Likewise, any membranous and/or cytoplasmic staining of mononuclear inflammatory cells within tumor nests and/or adjacent supporting stroma was considered to be positive PD-L1 staining and was included in the CPS numerator. Neutrophils, eosinophils, plasma cells, and inflammatory cells associated with in situ components, benign structures, or ulcers were excluded from the CPS. For the purpose of this study, cases with CPS>1 were considered PD-L1(+).

For LY6D and PD-L1 automated staining, seven-micron FFPE sections were processed with VENTANA BenchMark Ultra automated staining instrument (Ventana Medical Systems, Roche), using VENTANA reagents except as noted, according to the manufacturer's instructions. Slides were then counterstained with hematoxylin II followed by Bluing reagent. Bright-field sections were scanned with ZEISS Axio Scan.Z1. The whole digital slides were viewed using Zen Blue software. LY6D antibody (Sigma-Aldrich HPA024755) was used for immunostaining. DAB positive signals for PD-L1 and LY6D markers were quantified with QuPath software. The same 3 tumoral regions areas were selected for both PD-L1 and LY6D in each sample, and the results were expressed as a percentage.

3. Results

3.1. First Project - An Integrated Screening to Infer Transcription Factor Regulatory Networks Governing Cell Fate Decisions

3.1.1. Multiomic Screening Approach to Study the Key Role of Transcription Factors in Transdifferentiation

The understanding of TFs in defining cellular identity and their transcriptional and epigenetic barriers has been hindered by the variability of cellular engineering platforms and the lack of a systematic approach. Consequently, this has led to a substantial gap in our knowledge of molecular networks governed by different TFs, leaving many potentially key ones yet uncharacterized. In this study, we performed an unbiased survey of 130 key developmental TFs, examining their effect on gene expression, chromatin architecture, and cellular morphology. Our objective was to identify novel pioneer factors that, either individually or in combination, play a pivotal role in shaping cellular identity.

3.1.1.1. Transcription Factor Selection and Prioritization

The human genome contains over 1,800 TF loci, resulting in more than 3,500 isoforms that offer a wide range of possible regulatory effects[45]. Therefore, to identify the most promising candidates for our integrative screening, we used a multi-tiered approach to create a focused list of TFs potentially involved in developmental processes (Figure 17). To this end, we initially identified TFs exhibiting minimal expression levels in different human primary fibroblast cell lines (HLF, RPE, BJ), each derived from a distinct germ layer. This was implemented to address the effect of cellular context, thereby focusing on TFs that are more likely to act as pioneer factors. We further refined this selection by keeping the top 250 TFs that demonstrated the highest levels of human-specific expression. These TFs were then assessed using the "Transcription Factor Epigenetic Remodelling Activity" (TERA) score[185] which predicted their epigenetic functional capacity by integrating genomic and epigenetic data.

To complement our epigenetic approach, we utilized Mogrify[186], a network-based algorithm that leverages gene expression data. This method predicts TFs that are likely to regulate a designated target cell type, considering the gene expression profiles of the predetermined initial cellular state. In our application, we used the HLF, RPE, and BJ primary fibroblast cell lines as the starting cellular states and effectively predicted the optimal set of TFs that can regulate 95% of the genes within each of the 132 target cell types examined implemented in Mogrify platform.

We then merged the two candidate collections obtained from epigenomic and transcriptional approaches to create a comprehensive list of candidate TFs. From this set, we excluded those that did not exhibit a 'TF' or 'DNA binding' characteristic, ultimately yielding a list of 223 candidate TFs. To broaden our study's scope, we used a literature-based data extraction approach to include an additional 54 TFs that hold scientific value, despite not meeting our strictest selection criteria.

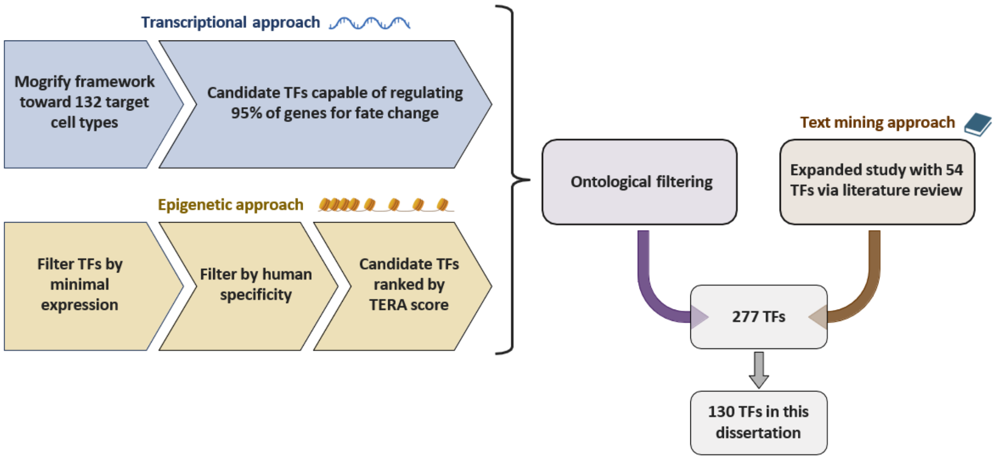

Transcriptional approach
Mogrify framework toward 132 target cell types
Candidate TFs capable of regulating 95% of genes for fate change
Epigenetic approach
Filter TFs by minimal expression
Filter by human specificity
Candidate TFs ranked by TERA score
Ontological filtering
Text mining approach
Expanded study with 54 TFs via literature review
277 TFs
130 TFs in this dissertation

Figure 17. Representative Scheme of Candidate Transcription Factors Selection Process

First, TFs were prioritized based on their gene expression (Mogrify algorithm) and epigenetic activity (TERA score). The resulting lists were then combined and filtered for specific gene ontology terms. An additional 54 TFs were identified from a systematic literature review, resulting in a total of 277 potential TFs. Of these, 130 have been analyzed and reviewed in this dissertation. Transcription Factors (TFs).

Through our integrative methodology, we have selected a comprehensive list of 277 TFs, categorized into eight distinct TF superclasses, out of a possible total of 10 as proposed by

Wingender E et al. (Fig 18).[12]. However, in this thesis we have successfully screened and analyzed a subset of 130 TFs from the original set, which includes six superclasses, ‘Alpha Exposed by Beta Structures’ (2.3%), ‘Basic Domain’ (35.4%), ‘Helix Turn Helix’ (36.2%), ‘Immunoglobulin Fold’ (6.2%), ‘Other All Alpha Helical DBD’ (1.5%), and ‘Zinc Coordinating DBD’ (18.5%). All 130 TFs successfully passed the screening procedure and met the data QC standards, as detailed in the Materials and Methods chapter. Although the entire scope of our screening study is not yet complete, this focused dataset of 130 TFs continues to underscore the complexity of their varied structural domains, thus offering insights into their diverse roles in regulating cell fate decisions (Figure 18).

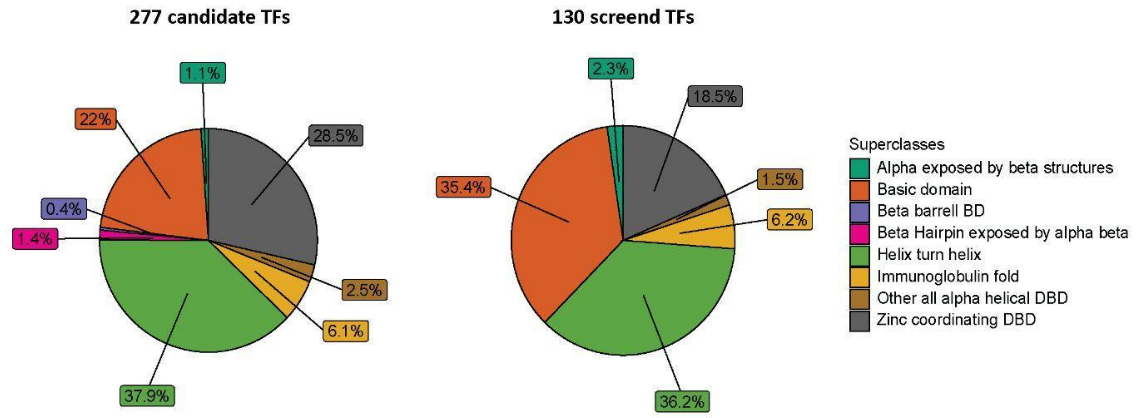


Figure 18. Superclass Representation in Candidate Transcription Factors Portfolio

A diagram illustrates the proportion of each superclass in the initial TFs collection (277), and the screened set in this dissertation (130 TFs). Transcription Factors (TFs).

3.1.1.2. Experimental Design

Next, the selected TFs were subsequently tested through unbiased transcriptomic, epigenomic, and morphological assays. To this end, we developed a systematic approach for surveying TFs under the same experimental settings. This enabled us to assess the impact of ectopic TF expression on cellular fate, using human primary fibroblasts as a terminally differentiated model (Figure 19). First, TFs were cloned into pLX317 vectors, each equipped with dedicated barcodes for identifying the inserted ORF and quantifying its expression level. The TFs were regulated by the EF-1a promoter and in-frame with a V5-tag, which facilitated the validation of their expression and nuclear localization. The EF-1a promoter was chosen due to its high expression capacity and to prevent potential methylation events. We then infected BJ-T cells (human immortalized foreskin fibroblasts) with the lentiviral vectors and cultured them for 11 days in low serum conditions to highlight their capacity to introduce new cellular functionality. Next, we accurately monitor the TF-induced changes in i) cellular morphology through high-content imaging (section 3.1.6), ii) gene expression dynamics using RNAseq (section 3.1.2), and iii) chromatin accessibility via ATACseq (section 3.1.3). Later, the collected data was used to create an integrated network depicting TF similarities in both transcriptional and epigenomic regulations (Results section 3.1.5).

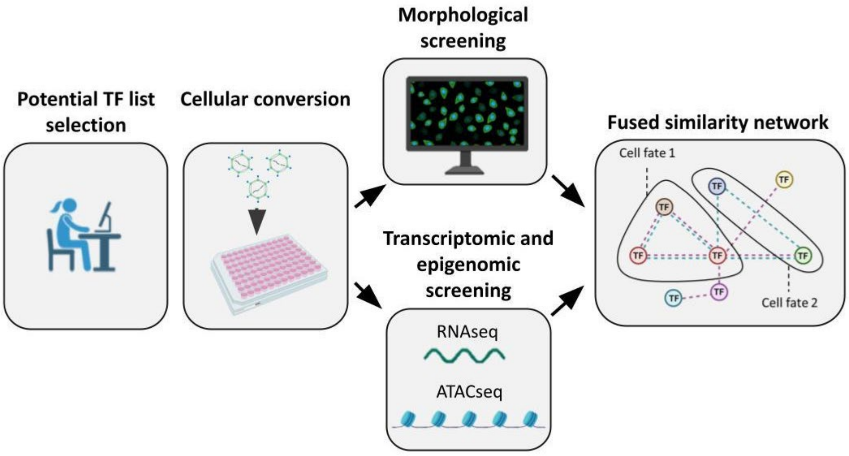


Figure 19. Multiomic Screening Approach to Deciphering Transcription Factors Role in Modulating Cellular Fate

A flowchart depicting the steps of our multiomic screening approach. Initially, potential candidate TFs were selected and subsequently integrated into lentiviral vectors to infect BJ fibroblast cells. Following 11 days of cellular conversion protocol, morphological assays, RNAseq, and ATACseq were conducted. Eventually, these individual datasets were then integrated to establish a comprehensive multiomic TFs connectivity network. Transcription Factors (TFs); RNA sequencing (RNAseq); Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq)

3.1.2. Dissecting Transcription Factors Regulation in Gene Expression Dynamics Revealed Significant Impact on Cell Fate Identity

In initiating our investigation into the impacts of the screened TFs on cellular fate, we first applied systematic analysis of their effects on the dynamics of gene expression. Upon completion of the cellular conversion process, we harvested cell samples and sequenced them using bulk RNAseq. This effectively captured data from over 130 quadruplicate TF conditions and GFP-infected cells as a control group (details in Materials and Methods). Given the large scale of the data, substandard analytical strategies were used to address associated challenges, such as considerable technical noise, data sparsity, and high variability.

3.1.2.1. Modeling Infection-Induced Variability for Effective Noise Reduction in Gene Expression Analysis

In our preliminary analysis, we delved into the biological variability present in our expression data. This examination revealed an unintended source of variation inherited from the infection procedure. Specifically, we identified a focused set of genes closely associated with the cellular stress response induced by this process (detailed below), exhibiting varying enrichment levels across all tested samples (Figure 20). The observed variation likely results from the inconsistent effectiveness of cellular infection and the extent to which the viral vector, containing the transgene, effectively integrates into the cells. Such technicality is expected, especially considering the wide-ranging scope of our screening strategy.

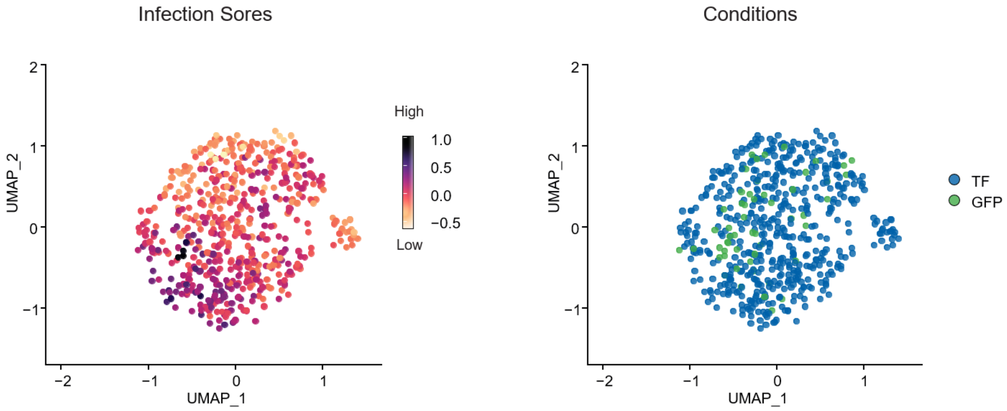


Figure 20. Infection-Inherited Variability Introduces Technical Noise in Gene Expression Data
A UMAP plot of RNAseq data visualizes each sample's transcriptional profile, with the infection scores color-mapped, indicating a directional effect on gene expression variability (left). And color-coded condition-specific data with TF-infected samples are in blue, and GFP-controsl are in green (right). Uniform Manifold Approximation and Projection (UMAP); RNA sequencing (RNAseq).

Notably, this variability is likely to introduce a technical bias in downstream analysis results, consequently, masking the true biological variation introduced by TF regulatory activity on gene expression dynamics. Therefore, to model the noise structure in our data, we performed a dose-dependent MOI control assay. In this experiment, we infected BJ-T cells with an increasing MOI concentration range (from 0.05x to x8 MOIs with respect to the original experiment) of the GFP-containing viral vector (Figure 21). To quantitatively measure the infection level observed in each sample, we first performed pairwise DE analysis between 8x and 0.05x MOI infected cells. Genes whose expression significantly changed (>2 LFC) were subdivided into eight groups by C-means clustering[214]. Hereof, we focused on a single cluster that demonstrated gene expression kinetics compatible with the incline MOI levels (Figure 21).

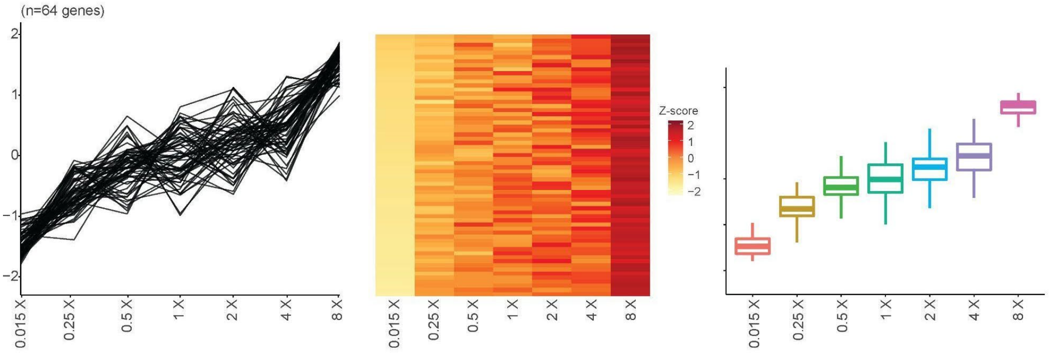


Figure 21. Selected Gene Cluster for Modeling Infection Score
Z-scored normalized expression (Y-axis) of 64 genes from the selected cluster, aligned with MOI concentrations (X-axis), shown as a line plot (left), heatmap (middle), and boxplot (right). Multiplicity of infection (MOI).

Notably, the 64 genes within the selected cluster demonstrated enrichment in pathways predominantly involved in immune and inflammatory cellular responses using GO enrichment analysis[190,191] (Figure 22). These findings highlight that these genes are indeed associated with the cellular response to the infection procedure.

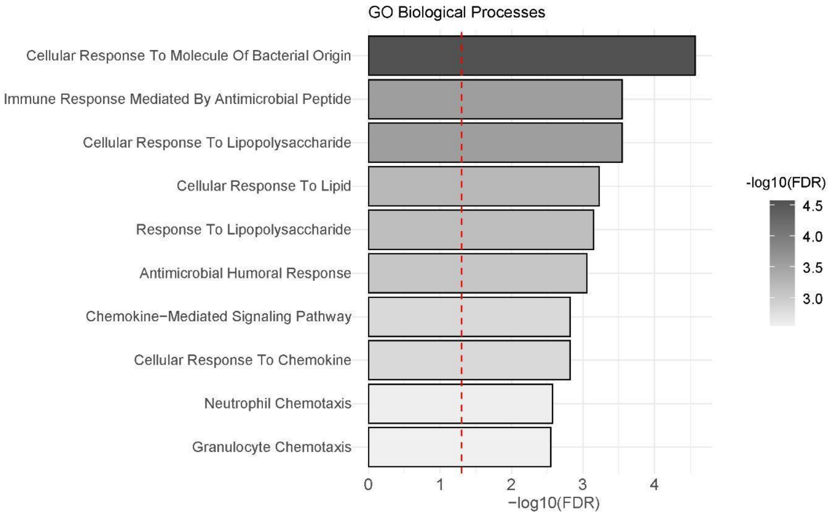


Figure 22. Cluster Selected 64 Genes Demonstrate an Enrichment in Infection-Related Gene Ontology Terms

A bar plot illustrating the top 10 most enriched biological processes in the selected genes, ranked by -log10 of their resulting FDR. False Discovery Rate (FDR); Gene Ontology (GO).

Building upon these findings, we employed the inferred genes to score each sample for its infection level using a computational strategy described by Tirosh et al[193]. Briefly, for each sample, we calculated the average expression level of the selected 64 genes and then subtracted the aggregated expression of control feature sets with a similar expression level (see Materials and Methods), thus generating unitless scores that can be compared across all samples. The resulting infection scores exhibited marked consistency with the dose-dependent MOI incline, thereby providing further evidence for the effectiveness application of our computational approach for modelling the infection-derived variability (Figure 23, left).

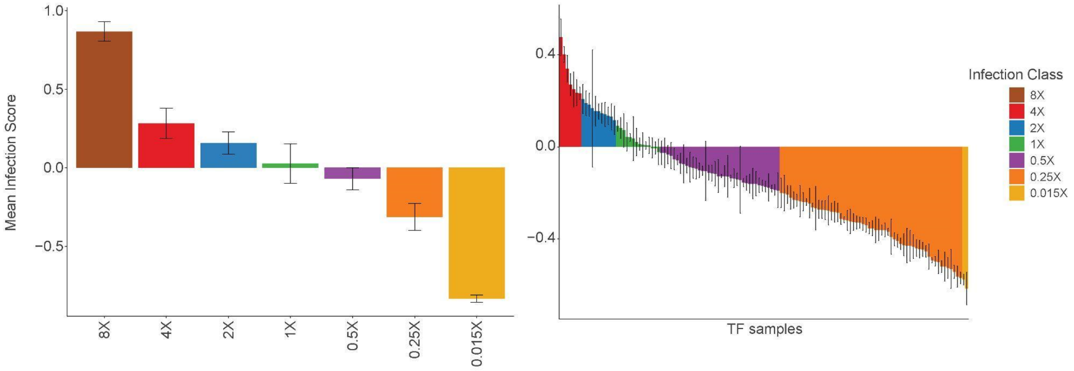


Figure 23. Stratification of Infection Scores for Differential Gene Expression Analysis

Three heatmaps depicting the 20 ranked motifs for NEUROD6 (left), TFAP2B (middle), and MYOD1 (right). The color gradient from light red to dark red corresponds to the enrichment score of each motif, denoting a unitless measurement scales from low to high enrichment.

Then, to mitigate the impact of infection-derived variability in subsequent analyses, we applied DEA by systematically pairing TF samples with GFP controls based on their corresponded infection levels (Figure 23, right). Initially, we calculated the average infection scores ($I_{\{TF\}}$) for each set of TF replicates (n number of replicates).

$$I_{\{TF\}} = \left(\frac{\sum_{i=1}^{n_{\{TF\}}} I_{\{TF_i\}}}{n_{\{TF\}}}\right)$$

These scores were then leveraged to categorize TFs into designated MOI concentration groups ($I_{\{GFP\}}(MOI)$), GFP control infection scores at a distinct MOI concentrations), prioritizing the ones that exhibited the smallest differences in infection levels. This can be represented as followed:

$$\Delta I_{\{TF,GFP\}} = \left|I_{\{TF\}} - I_{\{GFP\}}(MOI)\right|$$

Essentially, this approach enabled the normalization of inherent infection variability, by enhancing the accuracy of gene expression comparisons across TF and control samples, which are likely to exhibit similar transcriptional responses to the infection procedure.

3.1.2.2 The Majority of Transcription Factors Induce Profound Transcriptional Changes in Transcriptional Cell State

DEA results revealed a substantial impact on gene expression dynamics, this underscoring the potential role of these TFs in shaping cellular identity (Figure 24). Specifically, we found that a significant portion of the 130 TFs exhibited either a large effect 26.9% (> 400 DE genes), or a moderate impact 40.8% on gene expression (>100 and < 400 DE genes). Notably, a smaller fraction, 32.3%, showed a minimal or negligible effect (<100 DE genes).

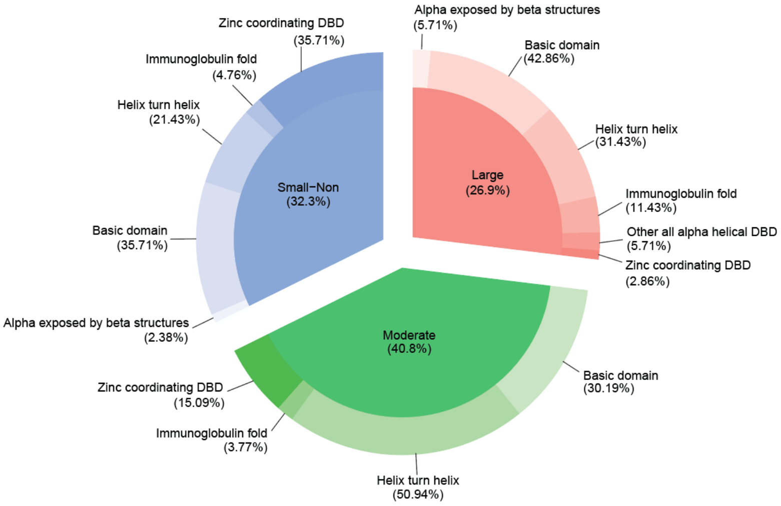


Figure 24. Categorization of Transcription Factors Impact on Gene Expression

A three-tiered pie chart illustrating the proportion of screened TFs categorized by their extent to affect gene expression, 'Small-Non' (blue), 'Moderate' (green), and 'Large' (pink). Each tier further delineates the distribution of TF superclasses, indicating the variety within each category. Transcription Factors (TFs).

Key pioneer factors such as FOXA2 (helix turn helix), TP63 (immunoglobulin fold), MEF2B (alpha-helices exposed by beta-structure), and MYOD1 (basic domain) significantly influenced gene expression, whereas other well-known TFs like HIF1A (basic domain), SNAI1 (zinc coordinating DBD), HOXA1 (helix turn helix), and LMX1B (helix turn helix) exhibited minimal impact. Additionally, when examining the distribution of the superclasses across these effect categories, it becomes clear that TFs belonging to Helix-Turn-Helix and Basic Domain are particularly influential (Figure 24). This trend may reflect the high representation of Helix-Turn-Helix and Basic Domain superclasses among the 130 examined TFs, rather than indicating a true intrinsic influence on gene expression dynamics. Such distinction in superclass representation highlights the need for further investigation into the unique characteristics of different TF structural domains in gene expression modulation.

Next, to validate our methodological approach, we conducted a pathway enrichment analysis using the transcriptional profiles of cells infected with well-characterized TFs as a reliable positive control (Figure 25). To this end, we focused on the gene expression data of MYOD1, NEUROD6, and TP63, aiming to assess their consistency with published data. Indeed, our results confirmed that the enriched signatures of each TF align with documented biological processes (Figure 25). Cells infected with MYOD1 revealed significant enrichment of varied myogenic pathways[215], NEUROD6 cells predominantly expressed gene signatures correlated with neuronal fate[216,217], while TP63 unveiled expressional dynamic linked with epithelial cell fate (in the contexts of skin development)[218].

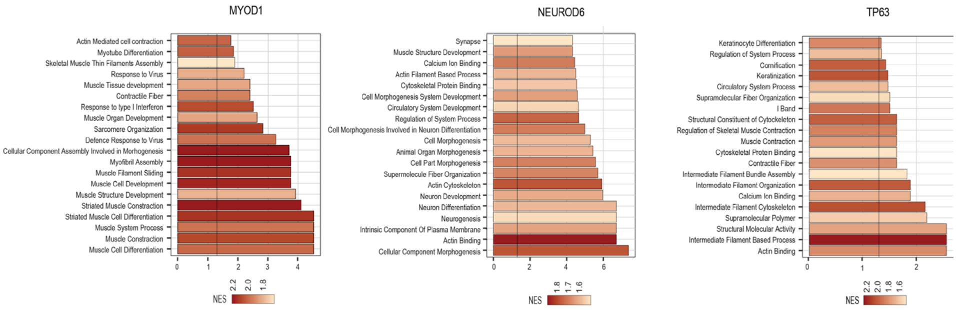


Figure 25. Key Transcription Factors Demonstrating Consistent Gene Signature with Published Literature

The top 20 most enriched pathways resulted from Gene set enrichment analysis using the C5 GO curated database colored by their NES for MYOD1 (left), NEUROD6 (middle), and TP63 (right). Gene Ontology (GO); Normalized Enrichment Scores (NES).

3.1.2.3. Analyzing Transcription Factors Target Genes Reveals Similarities in Gene Regulation Activity

Then, to unravel relationships between groups of TFs, we developed a gene set enrichment strategy that allowed us to evaluate to which extent candidate TFs impinge on common sets of target genes (Figure 26). Specifically, we repurposed the computational approach from the 'Mode of Action by NeTwoRk Analysis' (MANTRA) tool[197], and refined it to adjust to our data requisites. This approach allowed us to analyze TF transcriptional changes without setting an arbitrary cutoff in expression changes that may overlook the nuanced functional capacities of some TFs. These factors, while not significantly altering transcription dynamics on their own, can potentially re-wire cellular identity in combination with other TFs and additional effectors such as chromatin state, post-transcriptional mechanisms, and cofactors. First, for each TF we created a list of genes ranked according to their differential expression changes with respect to the GFP-control cells. We then extracted an "optimal" gene signature for each TF by selecting the top 250 genes that were overexpressed and the last 250 genes that were the most down-regulated. Using the gene set enrichment approach implemented in MANTRA, we then checked if the "optimal" gene signature of the first TF was graded consistently at the top or bottom of the entire ranked list of the second TF, and vice versa. By combining the enrichment score of the first TF in the second one, and reciprocally, we obtained a single value quantifying the similarity between a pair of TFs (Figure 26). In this dissertation, we refer to these similarity values as "Mantra Scores".

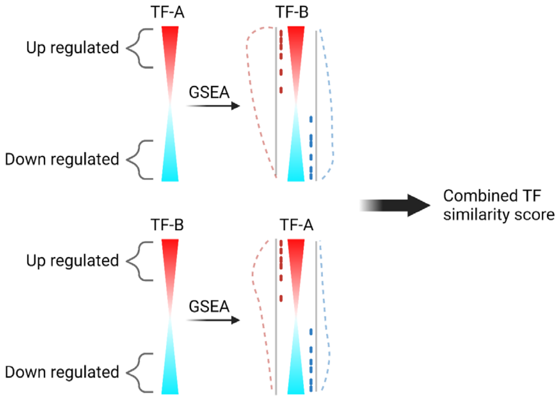


Figure 26. Cross-Enrichment Analysis of Transcription Factors' Target Genes

Gene set enrichment strategy for evaluating TF-A and TF-B similarity. Optimal gene signatures, derived from the top and bottom 250 differentially expressed genes from each TF, are cross-analyzed. The combined enrichment scores produce a single similarity score. Transcription Factor (TF).

Using the retrieved Mantra Scores, we performed hierarchical clustering, which successfully identified 37 distinct clusters (Figure 27). The dendrogram revealed that TFs sharing co-functional activities and those belonging to the same family tend to group closely together. For example, the clustering of TFs like MYOD1, MYF5, and MYOG represented myogenic fate[219], while the grouping of ATOH1, NEUROD and, NEUROG genes highlighted neurogenic fate[220,221]. Our analysis also sheds light on closely grouped TFs within the same family, such as the Hepatocyte Nuclear Factors (HNFs), including HNF1A, HNF1B, Forkhead box (FOX), with members like FOXA2, FOXA3, and homeobox genes (HOX) like HOXA6, HOXB6, and HOXD4. Interestingly, we observed uncharacterized TF interconnections. For instance, CREB5 an uncharacterized TF was closely associated with

TBX20[222], a key factor in cardiac development. Additionally, ASCL1, known to play a role in neuronal development[223] was closely grouped with myogenic TFs. Essentially, the clustering patterns not only confirmed the known functions and tissue-specific roles of the screened TFs but also highlighted the potency of our methodological approach to uncover previously unidentified TF Interconnections.

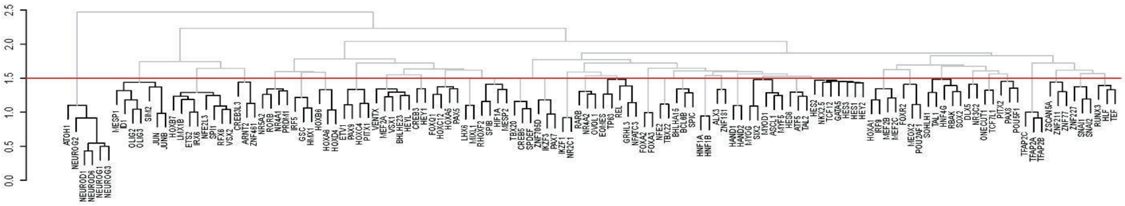


Figure 27. Transcription Factors Clustering Identifies Traditional and Novel Interconnections Based on Their Similarity in Downstream Target Genes

A dendrogram demonstrates the hierarchical clustering of 130 screened TFs into 37 distinct clusters, based on their respective target gene similarity (Mantra scores). Clusters below a threshold height of 1.5 are marked in black. Transcription Factors (TFs).

3.1.3. Transcription Factor Regulation on Chromatin Accessibility and Cellular State

As a further validation step, we added epigenomic data through ATACseq, aiding in the completion of our TF-induced transcriptional profiling. This integration offered insights into chromatin remodeling events and the identification of enriched binding sites. Given the lack of a standardized ATACseq data preprocessing protocol, we have developed a comprehensive pipeline that combines best practices from diverse sources, including NextFlow framework[199], ENCODE pipeline[224], and additional strategies from pivotal studies[71,200] (see Materials and Methods). Then, to identify TF-induced OCR, we performed differential peak analysis. For each query TF in our experiment, we created a list of significant differential OCR with respect to the GFP-control (LFC>2 P-adj<0.05). Systemic analysis showed that the majority of TFs demonstrated significant influence, with 22% having a large or 32% moderate effects. However, 46% of the TFs had a small to no impact (Figure 28). Further analysis into the superclasses proportion in each category, matched the findings from the RNAseq data (Figure 29). Specifically, the Helix-turn-helix TFs were particularly notable in the moderate effect group, while the Basic domains were most dominant in the large effect category.

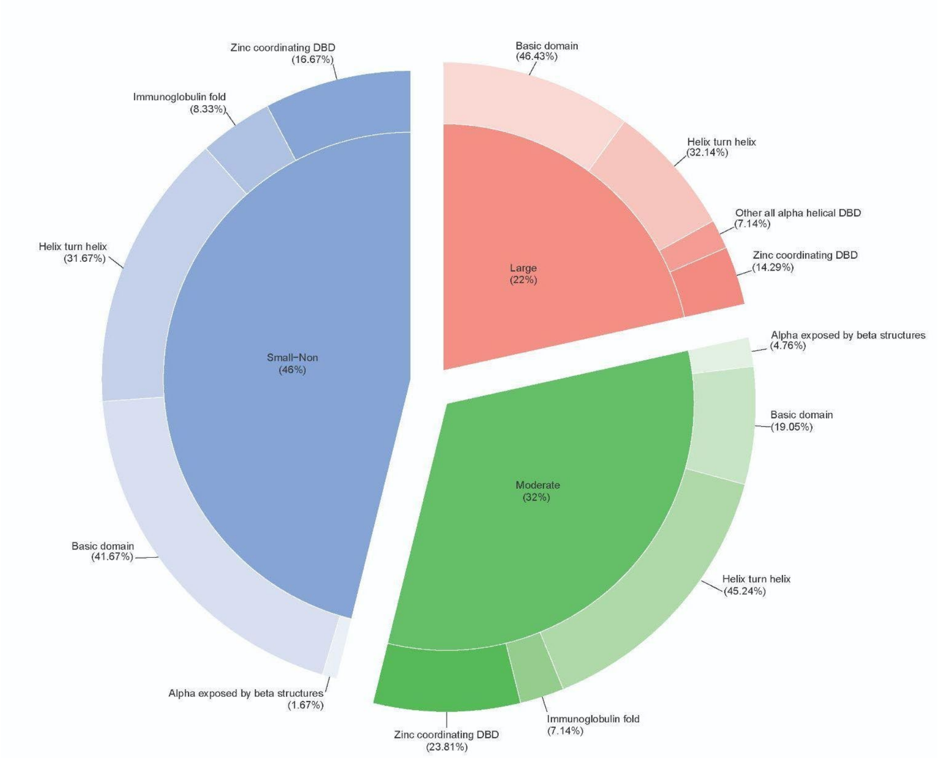


Figure 28. The Majority of Screened Transcription Factors Exhibit Moderate to Large Impacts on Chromatin Architecture

A three-tiered pie chart illustrating the proportion of TFs categorized by their extent to open new chromatin regions, 'Small-Non' (blue), 'Moderate' (green), and 'Large' (pink). Each tier further delineates the distribution of TF superclasses, indicating the variety within each category. Transcription Factors (TFs).

Overall, at both transcriptional and epigenetic levels, the majority of the 130 screened TFs exhibited a major to large impact on cellular processes. However, not all TFs that induced substantial changes in gene expression correspondingly affected chromatin architecture, and vice versa, indicating a partial overlap (Figure 29). This observation suggests that TFs not only vary in their efficiency to alter cellular state but also demonstrate distinct capacities in regulating transcriptional and epigenetic mechanisms.

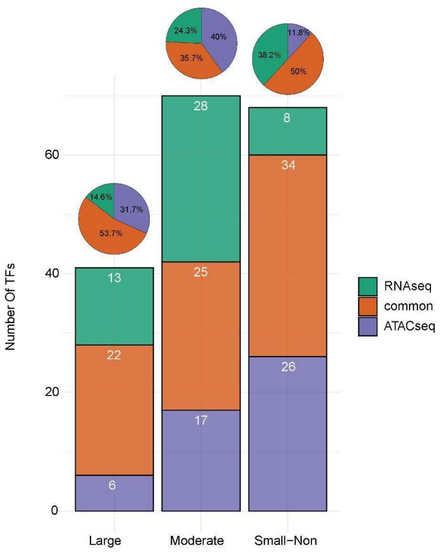


Figure 29. Transcription Factors Demonstrate Partial Consistency of Transcriptional and Epigenetic Mechanisms.

A bar plot (bottom) and accompanying pie charts (top) represent the varying impacts of TFs on transcriptional and epigenetic mechanisms. Both the bar plot and the pie charts are color-coded to signify the influence of TFs specifically at the transcriptional level (green), epigenetic level (purple), and on both levels concurrently (orange). Atop each bar, numerical values denote the count of TFs within each respective category and the pie charts detail their percentage. Transcription Factors (TFs).

3.1.3.1. Differential Motif Enrichment Analysis Reveals Distinct and Shared Transcription Factors Regulatory Modules

Next, we developed computational approach designated to evaluated the extent to which TFs can induce new OCRs that contain shared motifs with other factors, suggesting potential collaborative roles (Figure 30). Initially, we computed the likelihood of each motif sequence appearing in differential OCRs. Utilizing the FIMO tool[203], we computationally detected motif occurrences within uniformly sized peaks, drawing on a PWM matrix constructed from the TRANSFAC[65], JASPAR[66], and HOCOMOCO[67] databases as reference points (see Materials and Methods). Next, to prioritize the enriched motifs for each screened TF and effectively address background considerations, we developed differential motif enrichment analysis using Sparse Partial Least Squares regression (sPLS)[204] analysis. More specifically, we compared the motif binding enrichment matrix, which served as the predictor, with the normalized peak count matrix including only the differential OCRs. This resulted in the following sPLS regression equation

$$\mathbf{y=\beta0+\beta1x}$$

In this representative regression formula $\mathbf{y}$ represents the peak count for the differential OCRs of a given TF; $\mathbf{X}$ the motif enrichment; $\mathbf{\beta0}$ represents the baseline level of the peak count when there is no motif enrichmen; $\mathbf{\beta1}$ is the beta coefficient showing the connection between TFs impact on OCRs and the motif enrichment.

The ranking of each motif in relation to a specific TF was determined using a differential motif enrichment analysis method we developed. This method involves calculating the difference in beta coefficients (β1) between a given TF and a GFP-control. We express this calculation as:

$$\Delta\beta 1 = \beta 1_{TF} + \beta 1_{GFP\text{-}control}$$

$\beta 1_{TF}$ represents the beta coefficient for a given motif under the influence of the a given TF. $\beta 1_{GFP\text{-}control}$ is the beta coefficient for the same motif under the GFP-control.

The resulting Δ**β1**, is then used as a weight factor to rank the likelihood of a given motif sequence being influenced by the TF, with higher Δ**β1** values indicating a stronger influence of the TF on the motif.

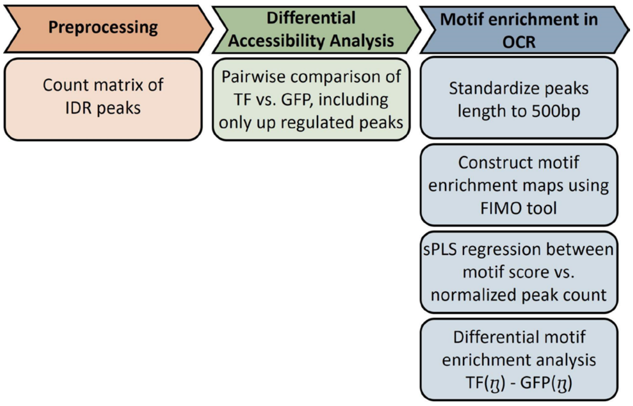

Preprocessing
Differential Accessibility Analysis
Motif enrichment in OCR
Count matrix of IDR peaks
Pairwise comparison of TF vs. GFP, including only up regulated peaks
Standardize peaks length to 500bp
Construct motif enrichment maps using FIMO tool
sPLS regression between motif score vs. normalized peak count
Differential motif enrichment analysis TF(ʒ) - GFP(ʒ)

Figure 30. . Differential Motif Enrichment Analysis Workflow

Flowchart overviewing the different steps in our differential motif enrichment analysis. Initially, preprocessing of ATACseq data produces an IDR peak count matrix (orange), followed by differential accessibility analysis to identify upregulated peaks (green). Motif enrichment analysis (blue) is then conducted by standardizing upregulated peaks to 500bp for FIMO tool analysis. Lastly, sPLS regression compares peak counts to motif enrichment, followed by the ranking motifs in TFs based on beta coefficient differences from GFP controls. Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq); Irreproducible Discovery Rate (IDR); Sparse Partial Least Squares regression (sPLS); Transcription Factors (TFs); Green Fluorecent Protein (GFP).

To validate our computational epigenetic approach, we assessed the accuracy of our differential motif analysis by testing well-characterized TFs as reliable positive controls. To this end, we focused on NEUROD6, MYOD1, and TFAP2B and compared their motif enrichment scores with the expected target motifs (Figure 31). Indeed, the results underscore a notable coherence between the top-ranked motifs and their respective anticipated ones (Figure 31). OCRs induced by NEUROD6 were found to be enriched in motifs associated with TFs crucial for neuronal development. This includes OLIG1, ZIC1, and ATOH1. Additionally, TFAP2B preferentially leads to OCRs enriched with motifs recognized by other members of the TFAP family like TFAP2A, and TFAP2C. MYOD1, on the other hand, was found to enrich motifs that are characteristic of myogenic TFs such as MYF5, MYF6, and MYOG[219]. Importantly, these TFs also demonstrate enrichment beyond their anticipated scope. For instance, NEUROD6 shows enrichment with motifs of cardiac TFs such as HAND2, MEF2C, and TBX20[222,225,226]. Meanwhile, TFAP2B induces motifs characteristic of ZNF genes, and MYOD1 enhances motifs associated with neuronal TFs like ASCL1, ATOH1, and OLIG2[221,223,227].

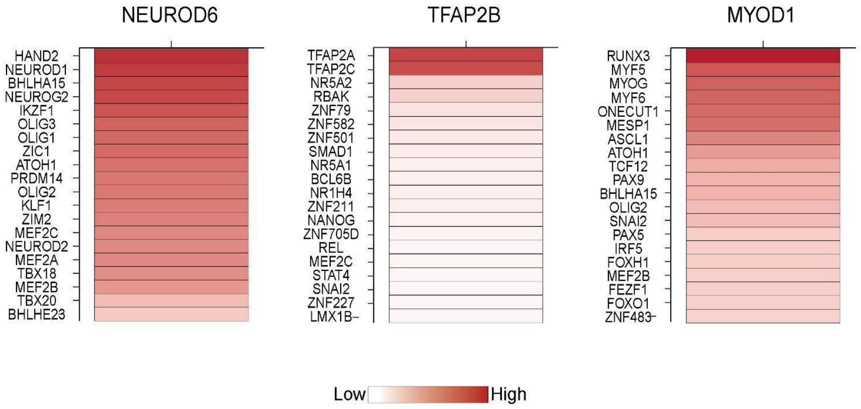


Figure 31. Well-Studied Transcription Factors Demonstrate Consistent Motif Enrichment and Reveal Extended Scope Beyond Published Data

Heatmaps illustrate the 20 top-ranked motifs for NEUROD6 (left), TFAP2B (middle), and MYOD1 (right), with a color gradient reflecting each motif's enrichment score, depicting both the anticipated and non-canonical results. The color range corresponds to the motif enrichment level for each TF, ranging from low (light red) to high (dark red) enrichment, referring the unitless characteristics of the scoring method. Transcription factor (TF)

Next, we inferred TFs that potentially exhibit co-functionality by evaluating their consistency in motif enrichment. To this end, we assessed the Euclidean distances among TFs based on their differential motif enrichment scores and applied hierarchical clustering, leading to the identification of 34 unique clusters (Figure 32). Similarly to RNAseq clustering results, we found that TFs with similar functions or belonging to the same family tend to cluster together. This includes the MESP, HOX, and PAX genes, and tissue-specific TFs like EOMES and NKX2-5 for cardiac induction[228,229], MYOG, MYOD1, and MYF5 for myogenic differentiation[219], and NEUROG/D genes[220], and ATOH1[221] for neurogenic fate regulation. Furthermore, our analysis revealed unexpected clusters of TFs that are closely grouped with a set of factors known to co-regulate similar cellular fate. This includes observations aligned with the clustering results derived from RNAseq data, which showed ASCL1 neuronal TF[223], clustered alongside myogenic factors, and CREB5 grouped with TFs regulating cardiomyocyte differentiation.

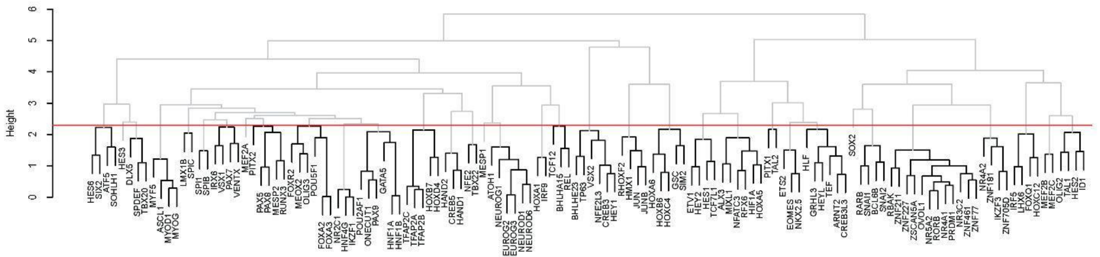


Figure 32. Transcription Factors Clustering Identifies Traditional and Novel Interactions Based on Their Similarity in Motif Enrichment

A dendrogram demonstrates the hierarchical clustering of 130 screened TFs into 34 distinct clusters, based on their similarity in differential motif enrichment. Clusters below a threshold height of 2.3 are marked in black. Transcription Factors (TFs).

We then conducted a comparative analysis of the dendrograms obtained from ATACseq and RNAseq analyses, depicting the clustering of TFs based on motif similarities or target gene enrichment, respectively. This comparison demonstrated a complex interplay between chromatin states and transcriptional activity (Figure 33). Notably, 11 clusters in the ATACseq dendrogram directly aligned with those in the RNAseq, particularly among TFs of the same family, such as TFAP, NEUROD, and HNF genes. Although not perfectly aligned, TFs like SNAI1 and SNAI2 were adjacent at both epigenetic and transcriptional levels. Interestingly, uncharacterized TF interactions were consistently observed at both levels. For instance, CREB5 in the ATACseq dendrogram clustered with early heart development TFs HAND1 and HAND2, while in the RNAseq it clustered with TBX20, indicating its potential role in cardiomyogenesis. Similarly, ASCL1 was grouped with myogenic TFs MYOG, MYOD1, and MYF5. Moreover, PRDM1 was consistently found alongside nuclear receptor family members NR5A2, RORB, and NR4A1 in both dendrograms, suggesting potential regulatory connections.

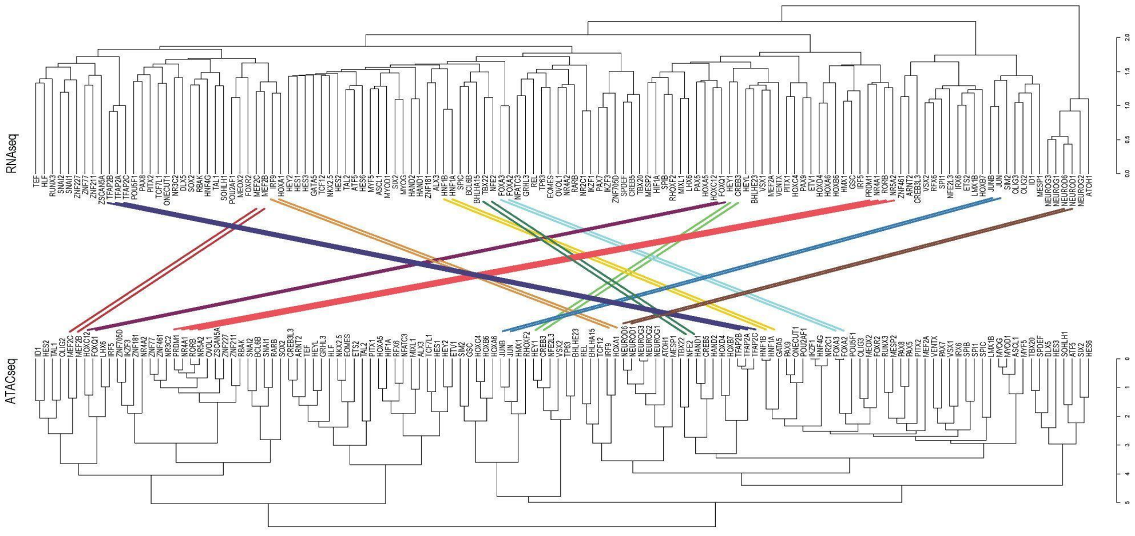


Figure 33. . Comparative Analysis of Transcription Factor Clustering from ATACseq and RNAseq Data

Visual comparison of two distinct clustering approaches. The bottom dendrogram shows TF clustering based on motif similarity (ATACseq), and the top shows clustering by target gene enrichment (RNAseq). Lines connect

11 corresponding TF clusters. Transcription Factors (TFs); Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq); RNA Sequencing (RNAseq).

Then by employing Baker's Gamma Index under the null hypothesis (which assumes fixed tree topologies), the results indicated a low level of statistical similarity between the two dendrograms (index=0.012, p-value=0.22). Therefore, the results imply that the TF-induced regulatory landscapes, as captured by ATACseq and RNAseq represent different aspects of regulatory mechanisms.

3.1.3.2. Motif Reshape Analysis to Highlight Transcription Factors with Greater Potential as Chromatin Remodelers

Next, we developed a 'motif reshape' analysis that allows for the identification of TFs with high potential to act as chromatin remodelers. Specifically, our goal was to determine whether a TF is more likely to target chromatin regions enriched with its expected motifs, or if it tends to regulate a broader range of areas without specific targeting, potentially due to its high expression levels (Figure 34). To this end, each TF was linked with a collection of similar or indirectly related motifs. Using the TOMTOM tool[205], we aligned each TF's motif sequence against others, categorizing significantly aligned ones as 'Expected' (q-value < 0.05) and the rest as 'Observed'. Then, for each category, we assigned their respective motif enrichment scores. Finally, TFs were ranked based on the delta of the two categories' enrichment scores and the statistical significance of these variances. Our results highlighted that while many key influential TFs (such as TFAP2B, FOXA2, MYOD1, OLIG2, and ATOH1) were highly ranked, other known TFs like NEUROG1, TP63, and SNAI2 demonstrated low epigenetic rankings.

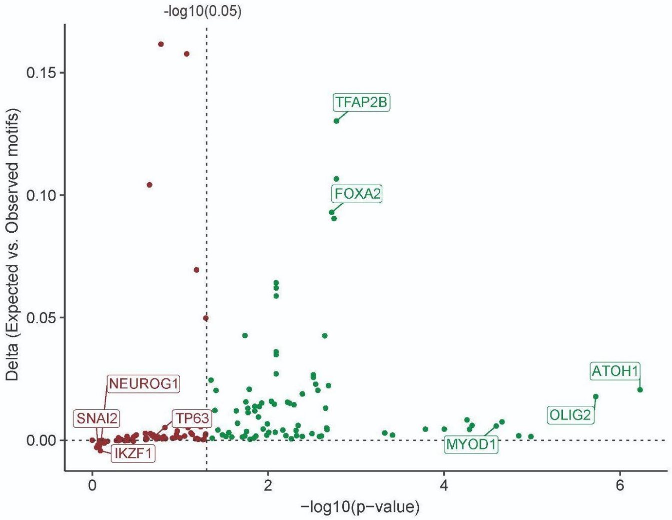


Figure 34. Variability in Transcription Factors Ability to Reshape Chromatin Regions with Target Regulatory Elements

scatter plot illustrates the ranking of TFs based on their capacity to reshape target regulatory elements. The Y-axis represents the delta between the expected and observed sets of motifs, while the X-axis indicates the significance of this difference. TFs are marked in green if they significantly and positively reshape target regulatory regions, and in red if the effect is otherwise. Transcription Factors (TFs).

3.1.4. Multiomics Data Analysis Enables the Inference of Transcription Factor Potential Activity Levels

As not all TFs hold the same functional capacity to alternate cell fate decisions, we prioritized them based on their impact on gene expression and chromatin architecture. Specifically, PCA analysis was applied separately to the mean-normalized count matrices from RNAseq and ATACseq data, focusing on genes or peaks exhibiting differential regulation (see Materials and Methods). Considering the linear nature of PCA, we quantified the Euclidean distance of each TF from the GFP-control within the latent space defined by the first two principal components (PC1 and PC2). This Euclidean distance was then used as a measurement to rank the activity of the TFs, with a larger distance from GFP-control corresponding to a higher activity rank (Figure 35).

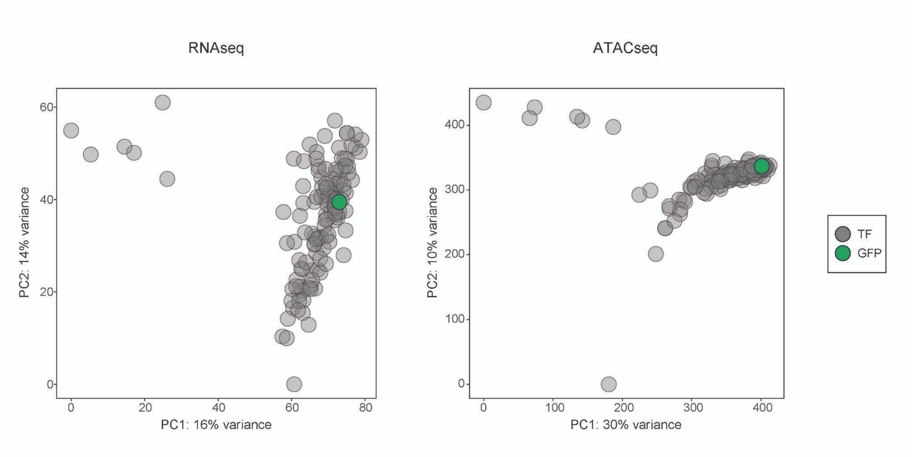


Figure 35. . Quantifying Transcription Factors Activity by Euclidean Distance from GFP-Control
PCA plots from RNAseq (left) and ATACseq (right) show mean-normalized expressions of cells infected with TFs (grey) versus GFP control (green). Distance from GFP indicates TF activity level. Principal Component Analysis (PCA); RNA sequencing (RNAseq); Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq); Green Fluorescent Protein (GFP).

Results showed a notable consistency in the TF activity scores across both RNAseq and ATACseq analyses, as evidenced by a 0.44 Pearson correlation with a p-value < 0.05 (Figuer 36 top and middle). This consistency was particularly prominent among the top-ranked TFs, such as the neuronal TFs (NEUROD and NEUROG genes, ATOH1, and ONECUT1)[221,230] and the HNF genes hepatocyte afate associated TFs (HNF1A, HNF1B)[231]. Interestingly, despite their canonical roles in ectodermal and endodermal fates, respectively, these TFs exhibited high activity scores in BJ fibroblast cells derived from the mesoderm germ layer. On the other hand, certain TFs exhibited substantial activity scores at the transcriptional level, independent of their lower epigenetic activities and contrariwise. This includes IRX6, MESP2, and SNAI1, all TFs involved in early embryonic development[232–234]. Then, to get the overall activity score we scaled the Euclidean distances from RNAseq and ATACseq and combined them (Figure 36 bottom). In this way, we prioritized TFs with potentially higher capacity to convert cell fate, thus focusing our study on potential pioneer factors. This will be taken into consideration in the next step of this project, where we will test non-canonical combinations of TFs to enhance the efficacy of cellular conversion protocols toward target cell identity.

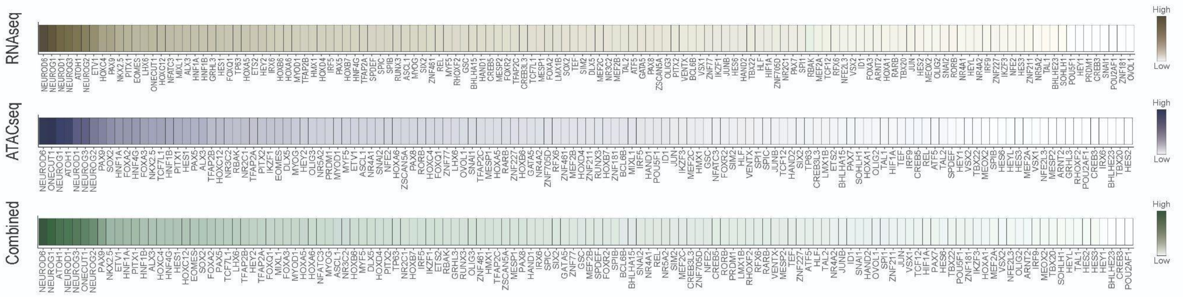


Figure 36. Integrated Transcription Factor Activity Scores at Transcriptional and Epigenetic Levels
TFs activity rankings from RNAseq (top) and ATACseq (middle) analyses, along with their combined scores (bottom). Each TF is color-coded based on its relative activity score. Transcription Factors (TFs).

3.1.5. Integrated Similarity Network Reveals Novel Co-functional Transcription Factor Modules

Next, we evaluated the similarity between TFs at both transcriptional and epigenomic levels, with the objective of identifying more comprehensive co-functional TF modules than those obtained from individual omic-level. To this end, we adopted the Similarity Network Fusion (SNF)[84] methodology, to effectively address the complexities of multiomic data integration. Specifically, this method allowed us to effectively retain strong connections between TFs while filtering out weaker ones, thereby emphasizing the most significant interactions. In contrast to the original study, which takes as input the count matrices from each omic level, our approach employs the TF similarity matrices tailored to highlight biologically meaningful aspects. Specifically, we focus on TF similarity from RNAseq data, which underscores their connectivity based on enriched target genes (Figure 27), and from ATACseq data, which highlights relatedness in terms of motif enrichment consistency (Figure 32). The retrieved fused TF similarity network effectively captured both shared and complementary information from epigenetic and transcriptional datasets, thus enabling us to infer TF modules that potentially co-function across both regulatory levels (Figure 37). The results obtained allowed us to generate several critical biological observations. First, we showed that TFs are grouped in closer clusters when they induce similar signatures, either transcriptional (green lines), epigenetic (yellow), or both (blue). Second, it is also possible to appreciate that many TFs belonging to the same family, induced the same transcriptional and epigenetic outputs (i.e., SNAI genes, NEUROG/D genes, JUN genes). At other times, the SNF analysis generated only similar epigenetic states (i.e., PTX genes, PAX genes, MESP genes). Additionally, TFs were closely grouped into a single module by exerting similar biological functions, such as cardiac fate (TBX20, HAND1, HAND2)[222,225,226], myogenic fate (MYOG, MYOD1, MYF5), and neuronal fate (ATOH1 and NEUROG/D genes)[221,230]. However, we identified unexpected connections of TFs that were closely grouped within modules of known co-regulator TFs. Such cases include ASCL1, which is known to regulate neurogenic development pathways[223], however it was clustered with pure myogenic TFs; and CREB5, which was grouped with pure cardiomyocyte developmental TFs, although its developmental role is not thoroughly studied.

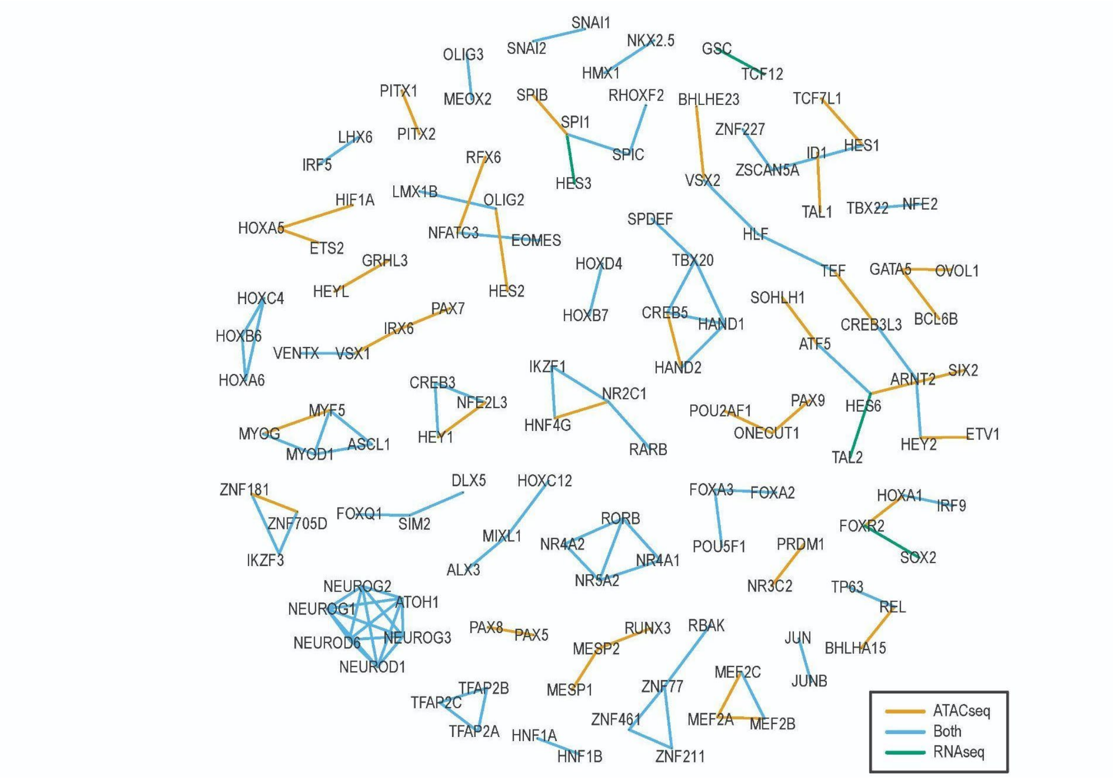


Figure 37. Integrated Connectivity Network Demonstrating Known and Uncharacterized Transcription Factors Modules

An integrated network represents similarities among TFs derived from RNAseq and ATACseq analyses. Color-coded lines indicate TFs exhibiting similar signatures at transcriptional (green) and epigenetic (yellow) levels, or both (blue). Transcription Factors (TFs); Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq); RNA sequencing (RNAseq).

3.1.6. Resolving Transcription Factors Induced Cellular Morphology Changes During Cellular Fate Conversion

To bolster the accuracy of our multiomic TF similarity network, we also assessed the morphological alterations resulting from TF modulation. This was accomplished through a high-throughput immunofluorescence staining, utilizing morphological markers such as TUBULIN for cytoskeletal visualization, CELL MASK to delineate membranal contours, DAPI for nuclear identification, and a V5-tag to confirm TF expression and its nuclear localization (Figure 38).

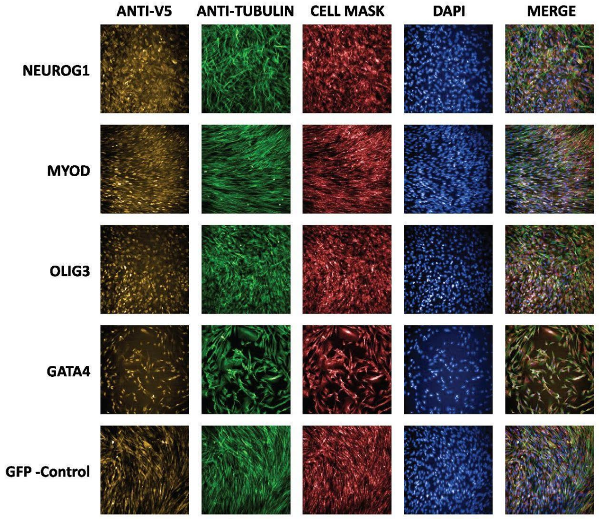


Figure 38. A High-Throughput Immunofluorescence Assay to Capture Induced Morphological Changes. Representative imaging results from a high-throughput immunofluorescence assay, employing morphological markers TUBULIN, CELL MASK, DAPI, and V5-tag, to capture structural changes.

Analyzing cellular morphology in transdifferentiation assays is particularly challenging, as this highly stressful procedure often triggers apoptosis in a considerable subset of the cells. Despite this, we were able to qualitatively identify substantial changes in cellular morphology in 15.4% of the 130 screened TFs, in contrast to cells infected with a GFP-control (Figure 39, left). Notably, only TFs that are part of the Basic domain, Helix-turn-helix, and other all-alpha helical DBD superclasses exhibited these morphological alterations (Figure 39, right). These observations might be attributed to the fact that the first two superclasses are the most prevalent in our survey (Basic Domain - 35.4%, and Helix Turn Helix - 36.2%, as shown in Figure 18).

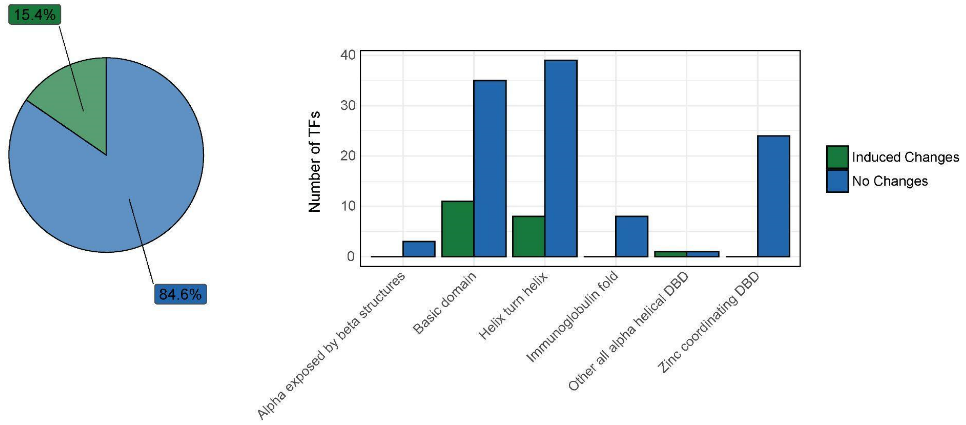


Figure 39. Differential Impact of Transcription Factors on Cell Morphology

A pie chart (left) and a bar plot (right) illustrate the effect of TFs on cellular morphology. The pie chart shows the proportion of TFs inducing morphological changes, while the bar plot compares the distribution of TF superclasses. Transcription Factors (TFs).

Additionally, we found that TFs successfully induced various types of morphological changes. For example, MYOD1[219], induced changes such as increased cell length and nuclei fusion, aligning with characteristics typical of myogenic-like cell types. Furthermore, TFAP2B, a key regulator of various developmental processes[235], led to changes in cell polarization and flatteningSimilarly, NEUROG1 promoted the formation of neuronal-like shapes[220], marked by the emergence of stellate cells and multilayering (Figure 40). Overall, these findings underscore a disparity between the induced morphological changes at the phenotypic level and the ones observed at the transcriptional-epigenomic scale.

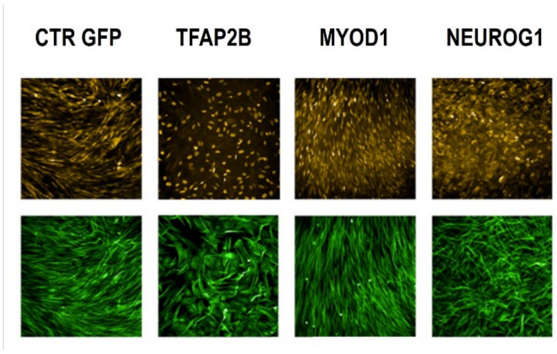


Figure 40. Differential Morphological Changes Induced by Key Transcription Factors Relative to GFP Control

The top row exhibits V5-tag immunostaining, while the bottom row presents anti-tubulin immunostaining. The control, GFP, maintains a standard fibroblast morphology.

3.2. Second Project - Spatial Transcriptomics Reveals Sub-Tumoral Identities and Novel Diagnostic markers in Triple Negative Breast Cancer With Immune Evasion Capacity

3.2.1. Cost-Effective Clinical Workflow for Accurately Identifying Novel Spatial Transcriptomics Signatures

To date, PD-L1 has been established as a crucial diagnostic marker for selecting patients with metastatic or locally advanced TNBC and for immune checkpoint inhibitor treatment[123,124]. While it achieved notable successes, the overall efficacy of PD-1/PD-L1 still needs to be fully realized[116]. Recent advances in spatial transcriptomics platforms might offer a way to identify surrogate and complementary diagnostic markers to augment the PD-L1 test and targeted therapies. Therefore, in this study, we have developed an efficient and cost-effective clinical workflow to spatially resolve the heterogeneous nature of TNBC architecture, focusing on the expression status of PD-L1. Our workflow is specifically designed to benefit from a minimal requirement of input samples and compatibility with standard laboratory equipment. This methodology involves three critical steps (Figure 41): 1) Spatial transcriptomic sequencing to delineate variations in gene expression across different tumor regions. 2) Clinical-grade RNAseq to evaluate the viability of potential diagnostic markers identified in the preliminary step. 3) IHC of the selected therapeutic diagnostic marker candidates to verify their protein expression levels and precisely map their locations within the tissue sections.

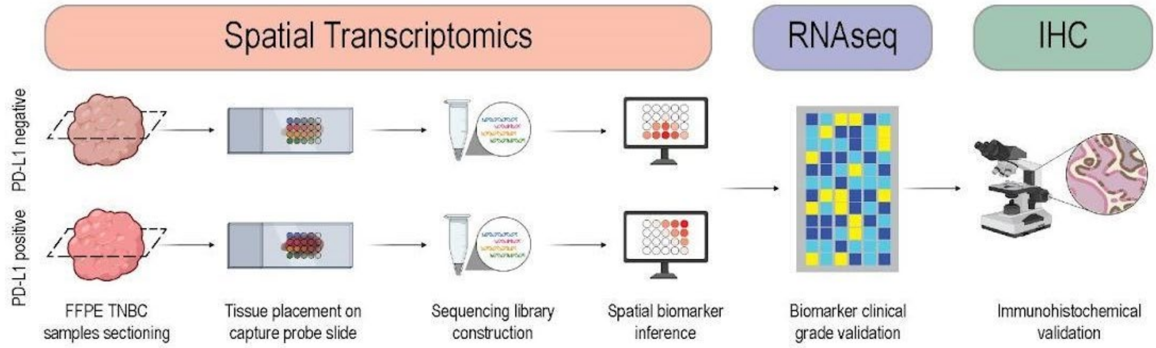

Spatial Transcriptomics
RNAseq
IHC
PD-L1 negative
PD-L1 positive
FFPE TNBC samples sectioning
Tissue placement on capture probe slide
Sequencing library construction
Spatial biomarker inference
Biomarker clinical grade validation
Immunohistochemical validation

Figure 41.Schematic Representation of Our Three-Tier Spatial Transcriptomic Workflow.
Our strategy involves spatial transcriptome sequencing, clinical-grade RNAseq for diagnostic marker feasibility, and IHC validation of protein expression and localization. RNA sequencing (RNAseq); Immunohistochemistry (IHC).

Initially, we acquired two FFPE tissue blocks of primary TNBC samples. These samples were rigorously confirmed as PD-L1 positive and negative in alignment with established clinical practices[116] (see Materials and Methods). Subsequently, sections of 5μm thickness were carefully positioned onto the Visium (10X Genomics Inc) spatial probe slide, successfully acquiring both the tumor edges and surrounding tissue. This aspect is essential, given that the Visium capture area is confined to a limited size of only 6.5x6.5mm. Capturing both the tumor and its adjacent surrounding tissue is crucial for analyzing the dynamics and interactions between these areas, especially in the context of PD-L1's role in tumor immune escape mechanisms. Next, the slides were stained with H&E, followed by a detailed histopathological analysis (Figure 42). This analysis involved annotating various regions within the tissues, including tumor, fat, stroma, and blood vessels. In certain instances, regions were designated as "mixed" annotations as the pathologists were unable to assign them to a singular, specific tissue compartment (Figure 42, right). Finally, the tissues were sequenced using Visium capture probe slides, each containing 5000 spots with unique spatial barcodes and a resolution of 100 µm. Essentially, this allowed for the accurate mapping of retrieved mRNA reads back to their original locations in the tissue section.

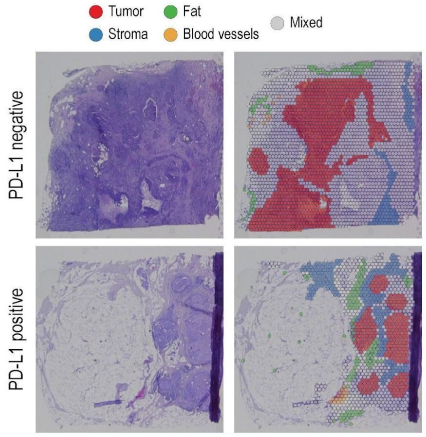


Figure 42. Hematoxylin and Eosin Stained Triple Negative Breast Cancer Samples and Corresponding Histological Annotations on Visium Spatial Capture Probe Slides

H&E stained TNBC samples placed on Visium spatial capture probe slides (left) and their associated histological annotation, delimiting areas identified as tumor, stroma, fat, blood vessels, or mixed regions (right). Hematoxylin and Eosin (H&E); Triple Negative Breast Cancer (TNBC).

3.2.2. Spatial Transcriptomics Reveals Alignment of Histological Features with Tumor Architecture and Heterogeneity in Triple Negative Breast Cancer Biopsies

Then, we evaluated the spatial integrity of the sequenced data. To this end, we examined the spatial gene expression profiles across the various histological compartments, including tumor, stroma, fat, and mixed histological regions. However, we omitted spatial spots identified as blood vessels due to their scarcity (less than 1.5% in both samples). Although results indicated an average of 7,842 mRNA counts per spatial spot in PD-L1 positive samples and 7,222 in negative samples, a notable increase in mRNA counts was observed specifically within tumor areas (Figure 43). This suggests a marked increase in transcriptional activity within the tumor region, consistent with prior studies showing a significant induction of gene expression in cancer cells[236].

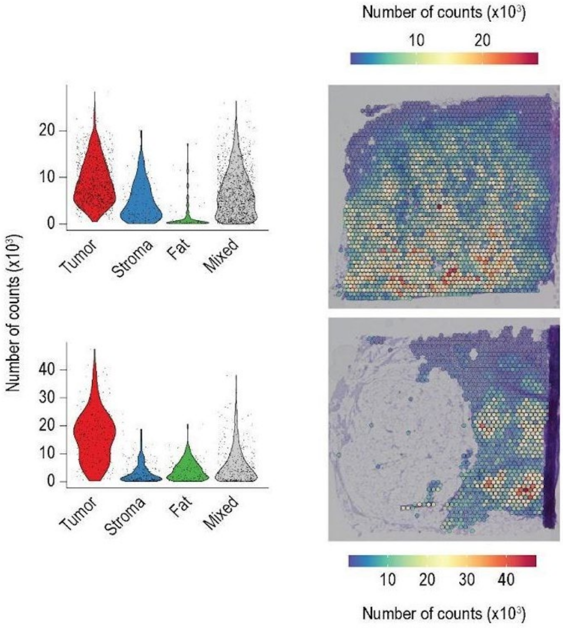


Figure 43. Tumor Regions Demonstrated Higher Transcriptional Activity

. Distribution of transcriptional activity represented by the read counts for each histological compartment (left) and their spatial arrangement across tissue sections (right).

This observation was further supported by the distinct and predominant expression of cancer-associated genes in spatial spots within tumor areas, specifically AKT1, BRCA1, E2F2, E2F4, EGFR, MYC, PIK3CA, STAT3, and TP53 (

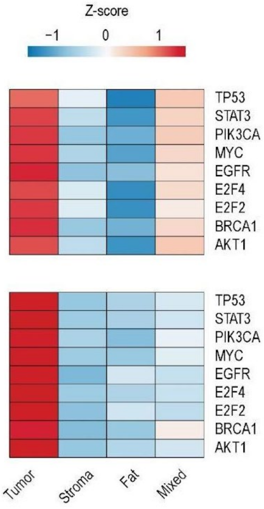


Figure 44. Cancer-Associated Genes Demonstrate Predominant Expression in Tumor Areas

Heatmaps presenting the z-score normalized expression levels of well-established cancer-associated genes across the different histological compartments.

Subsequently, we assessed PD-L1 expression as a sample-specific marker to determine its consistency with the initial evaluation. Indeed, the results revealed a reduced number of spatial spots expressing CD274 (the gene coding for the PD-L1 protein) in the PD-L1 negative sample compared to the PD-L1 positive sample (Figure 45, left and middle). Furthermore, our examination of CD274 positive spatial spots distribution across different histological compartments revealed a pronounced specificity in tumor areas of the PD-L1 positive sample and, to a lesser degree, in stroma/fat regions of the negative sample. The finding is consistent with previous studies indicating that PD-L1 can be expressed by either infiltrating immune cells within the microenvironment or by tumor cells[237,238]. As we evaluated the Euclidean distance distribution between CD274 expressing spots, our results showed a significantly higher spatial density (i.e., smaller mean Euclidean distance) within tumor regions of the PD-L1 positive sample (Figure 45, right). Overall, these results suggest that in the PD-L1 positive sample, CD274 positive spots are closely linked to neighboring cancer cells, resulting in a higher spatial density. Conversely, in the PD-L1 negative sample, positive spots are more likely associated with infiltrating immune cells, resulting in a more heterogeneous spatial density distribution throughout the entire tissue area.

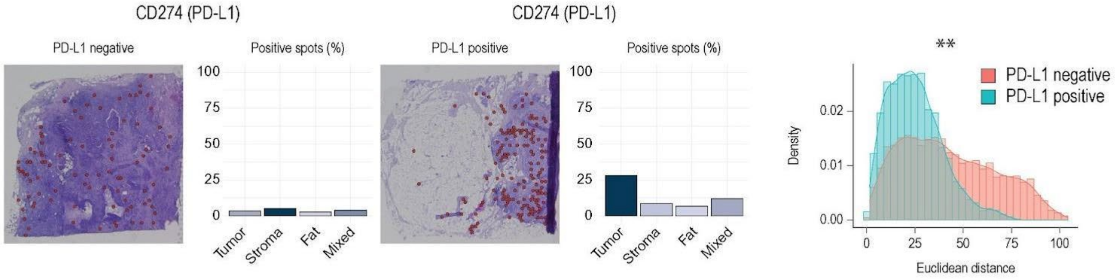


Figure 45. Comparative Spatial Expression Analysis of PD-L1 Among the Two Samples

Spatial spots expressing PD-L1 projected over the tissue space and their distribution among the different histological compartments (left), followed by the distributions of the Euclidean distances between PD-L1 expressing spots, **p-value < 2.2e-16 (right). Programmed death-ligand 1 (PD-L1).

Finally, we examined the agreement between labeled histological annotations and established diagnostic markers for breast cancer (KRT7), fat (FABP4), and stroma (FN1)[111,239,240] (Figure 46). These findings confirmed that these markers' expression levels were enriched and exhibited high spatial specificity for the corresponding tissue type. Specifically, KRT7-positive spots were enriched within tumor areas, irrespective of PD-L1 status. Similarly, FABP4 exhibited higher specificity in spots assigned as fat tissue. FN1 exhibited high expression in spots assigned to mixed and stromal regions but demonstrated slightly higher specificity in cancer cells. This is expected, given the dual role of FN1 in both tumor cells and the microenvironment[162]. Essentially, these findings suggest that the sequencing data effectively maintains spatial integrity, thus serving as an additional layer of information for discovery analyses.

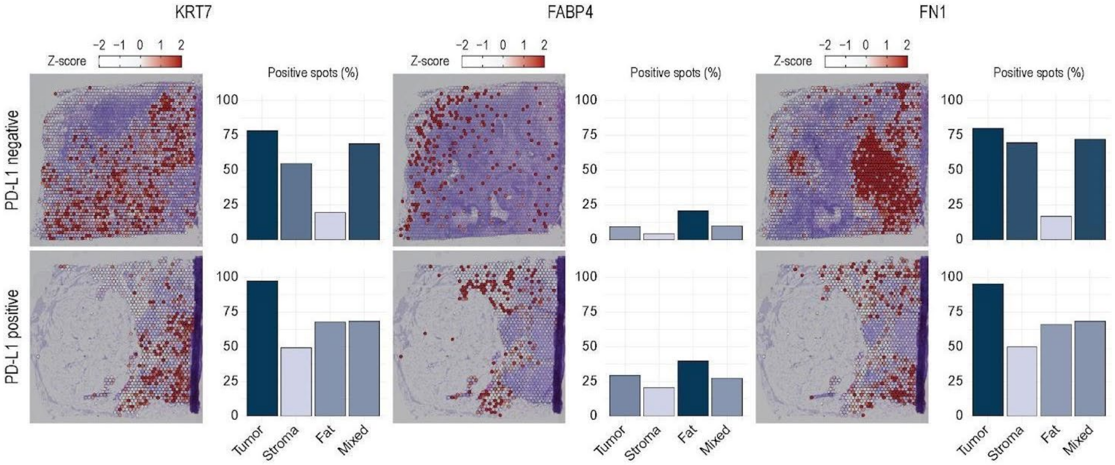


Figure 46. Tissue Marker Genes Exhibited High Spatial Specificity for Their Corresponding Histological Annotation

Spatial spots that are positive for KRT7 (tumor diagnostic marker), FABP4 (fat diagnostic marker), and FN1 (stroma/transitioning cells diagnostic marker) projected onto the tissue space and their distribution among the different histological compartments (see materials and methods).

3.2.3. Spatially Resolved Transcriptomics Refines Histological Annotations of Tumoral and Stromal Areas

Next, to gain a deeper understanding of the tumor and microenvironment architectures beyond standard histological analysis, we analyzed non-annotated spatial-expression data. Specifically, we summarized the transcriptional data into a lower dimensional space (UMAP) to infer data variability driven by biological differences. Then, we applied clustering analysis with the aim of revealing overlooked spatial characteristics in the histological analysis. This resulted in 10 clusters for the PD-L1 negative sample and 5 for the PD-L1 positive sample (Figure 47, left). Notably, these clusters exhibited a distinct separation without intermixing, evident both in the UMAP latent space and upon their projection onto the tissue space (Figure 47, middle). Additionally, the projection of histological annotations onto UMAP embeddings showed a general agreement between clusters and histological classes.

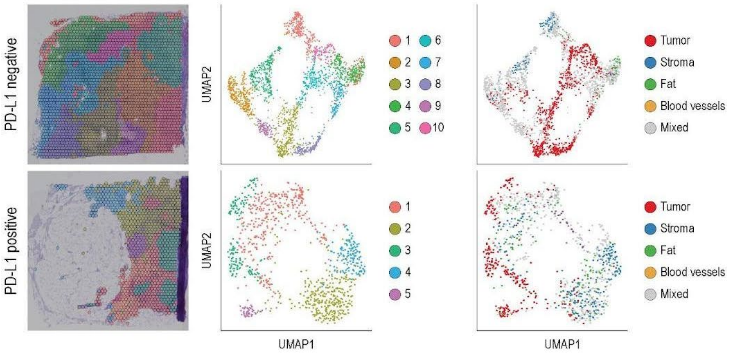


Figure 47. Clustering Non-Annotated Expression Data Revealed Overlooked Spatial Characteristics

Overlay of transcriptional-based clustering data on top of tissue histology (left). UMAP embeddings illustrate the distribution of transcriptional-based clusters (middle) and histological compartments within the tissue coordinates (right) for both PD-L1 negative (top) and positive (bottom) samples. Uniform Manifold Approximation and Projection (UMAP); Programmed death-ligand 1 (PD-L1).

However, a detailed quantitative evaluation of histological compartment compositions within each cluster, both in PD-L1 negative and positive, revealed a complex spatial pattern, with most clusters not being exclusively associated with a single histological class (Figure 48).

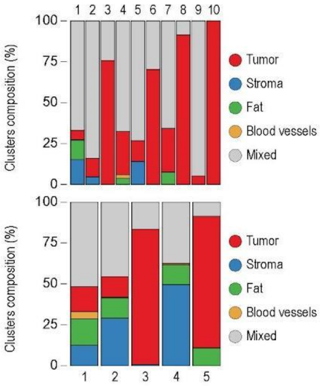


Figure 48. Histological Compartment Compositions Within Each Cluster Revealed Complex Spatial Patterns

Bar plots displaying the histological compartment compositions of each cluster for both PD-L1 negative (top) and positive (bottom) samples. Programmed death-ligand 1 (PD-L1).

Therefore, to gain a more comprehensive understanding of bonafide tumoral and stromal architecture, we used ESTIMATE (Estimation of Stromal and Immune cells in MAlignant Tumor tissues using Expression data)[208], a gene signature-based computational method. This approach allowed us to infer stromal and tumor signatures for each cluster, exclusively using the retrieved normalized expression data (see Materials and Methods). Notably, the identified signatures corresponded closely with their respective histological compartments (Figure 49). The stromal score was primarily found in spatial spots within tissue regions annotated as stroma and fat cells, while the Tumor Purity score was mainly enriched in spatial spots designated as tumor cells. Importantly, both Stromal and Tumor Purity scores were also attributed to spatial spots labeled as 'mixed' (i.e., those lacking distinct histological identification), thereby offering an additional quantitative annotation for these tissue compartments (Figure 49). Furthermore, the ESTIMATE’s signature annotations within clusters exclusively linked to a singular histological class (i.e., Tumor or Stroma) allowed us to refine the histological analysis. For instance, even though histopathological analysis identified clusters ‘8’ and ‘10’ in the PD-L1 negative sample as tumor regions (Figure 48), they exhibited distinct transcriptional characteristics (Figure 49). Cluster ‘8’ showed a high Tumor Purity score and low stromal signature, which aligned with annotated tumor regions. On the other hand, cluster ‘10’, despite being categorized as a tumor, exhibited a high stromal signature and a low tumor purity score.

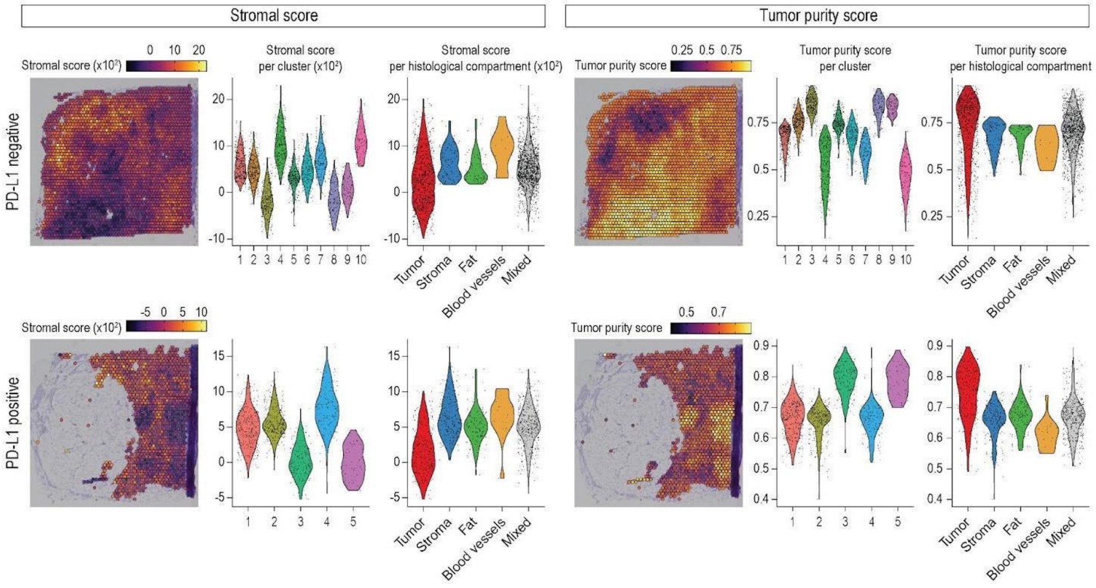


Figure 49. ESTIMATE Algorithm Inferred Stromal and Tumor signatures form Unlabaled Expression Data

ESTIMATE's Stromal and Tumor Purity scores are visualized in three ways: spatial representation in tissue samples (left), distribution among identified clusters (middle), and across diverse histological compartments (right), for both PD-L1 negative (top) and positive (bottom) samples.

Then, we utilized the ESTIMATE algorithm results to independently annotate the identified clusters. Notably, clusters 3, 8, and 9 in the PD-L1 negative sample, along with clusters 3 and 5 from the PD-L1 positive sample, exhibited high tumor purity and low stromal scores (Figure 49). Therefore, we classified these as tumor clusters, while the others were designated as surrounding (Figure 50, left). This categorization was then integrated with the initial histological analysis, to enhance the comprehension of the spatial annotations within the tissue. Consequently, by reannotating the identified clusters as either tumor (T) or surrounding (S) classes, we effectively distinguished between the tumor and its surrounding area (Figure 50, left), while also preserving the granularity of sub-tumor and microenvironment clusters (Figure 50, right). Specifically, we reassigned "Mixed" spots to either T or S, based on their classification in transcriptional data. Spots labeled as tumors in the expression data but identified as stroma in histological annotations were excluded. Similarly, spots recognized as surrounding areas in expression data but marked as tumor regions by pathologists were also removed. Then, to direct our downstream analysis toward tumor-microenvironment interactions, we eliminated all spots identified as fat tissue.

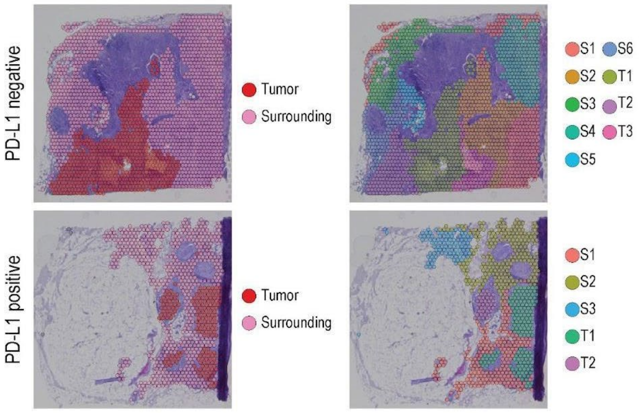


Figure 50. Spatial Spots Reannotation by Integrating Histological Analysis and Inferred Transcriptional Signatures

Spatial visualization of tumor and surrounding tissue annotations based on ESTIMATE scores (left) and transcriptional-based clustering projections (right) after integrating ESTIMATE-score with histological annotations (see materials and methods).

Notably, the reannotated clusters demonstrate more consistent annotation with the preliminary labeled histological compartments (Figure 51). Overall, by integrating spatial spot annotations and removing inconclusive ones, we obtained a more accurate molecular makeup of tissue architecture to better examine the tumor and its microenvironment (section 3.2.4) and predict their molecular interactions (section 3.2.5).

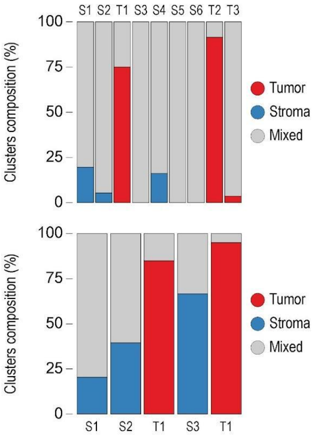


Figure 51. . Histological Compartment Compositions Within Each Cluster Following Spatial Annotation Refinement.

A bar plot displaying the histological compartments composition of each cluster, for both PD-L1 negative (top) and positive (bottom) samples.

3.2.4. Spatial Profiling of Gene Expression Identifies Distinct Transcriptional Signatures in Sub-Tumor and Microenvironment Regions

The re-annotation of spatial spots has not only successfully subdivided them into two major categories, Tumor (T) and Surrounding Environment (S), but has also effectively maintained subclusters within each category, characterized by distinct transcriptional profiles (Figure 52). To better understand the biological context of cells represented by individual sub-clusters, we initially conducted a spatial gene expression analysis, comparing each cluster with all others. Subsequently, we focused on the top 10 spatially variable genes (SVGs) within each sub-cluster to gain more detailed insights (Figure 52). Through a systematic literature review, we discovered that the genes predominantly expressed in the tumor sub-clusters were primarily linked with breast cancer and cancer hallmarks. For instance, SNCG, ATP1B1, and PGGHG in the PD-L1 positive sample, and EEF2, CHI3L2, and CRYAB in the PD-L1 negative sample. In contrast, the top genes expressed in the surrounding sub-clusters were mostly related to immune response (e.g., IGKV4-1), activities associated with cancer-associated fibroblasts (CAFs) (e.g., FN1), and factors contributing to cellular transformation and tumor invasiveness (e.g., FOS). Interestingly, the intratumor clusters in each sample exhibited both shared and unique SVGs, implying a degree of intratumor variability. This suggests that each sub-cluster may influence distinct key signaling pathways involved in regulating cancer progression.

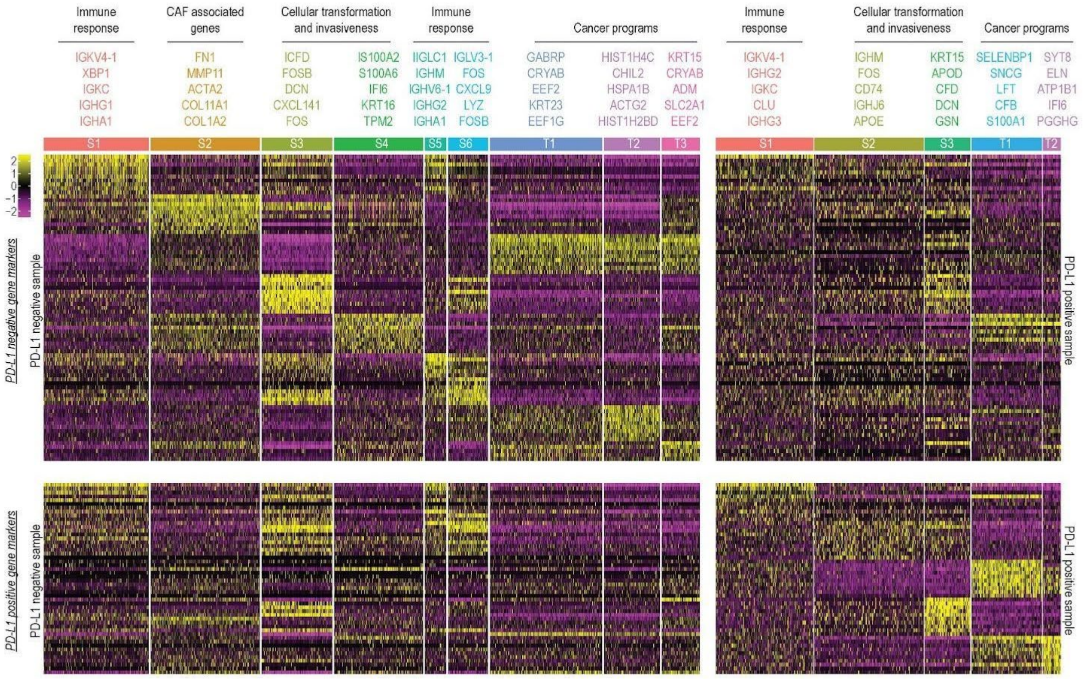


Figure 52. . Histological Compartment Compositions Within Each Cluster Following Spatial Annotation Refinement.

Top: Heatmaps of z-score normalized gene expression values for the top 10 up-regulated spatial-marker genes from the PD-L1 negative sample, showcased within individual sub-clusters from both PD-L1 negative (left) and positive (right) samples. Each cluster is indicated with the associated biological program and the top 5 genes with the highest rank score. Bottom: Heatmaps of z-score normalized gene expression values for the top 10 up-regulated spatial-marker genes from PD-L1 positive sample, showcased within individual sub-clusters from both PD-L1 negative (left) and positive (right) samples.

To further investigate these findings, we conducted a gene set enrichment analysis based on the spatial variation in gene expression (Figure 53). The results indicated that subclusters identified as tumor (T) exhibited a greater enrichment in pathways typically prevalent in cancer cells. These include MYC target pathways, WNT and NOTCH signaling, oxidative phosphorylation, and others. Conversely, the surrounding clusters (S) predominantly displayed enrichment in pathways related to tumor immune response, invasiveness, and metastasis. The distinct spatial enrichment patterns observed in tumors and their surrounding clusters suggest a distinguished functional specialization within the tumor landscape. While tumor cells are focused on growth and metabolism, the surrounding tissue aids in invasion and immune response.

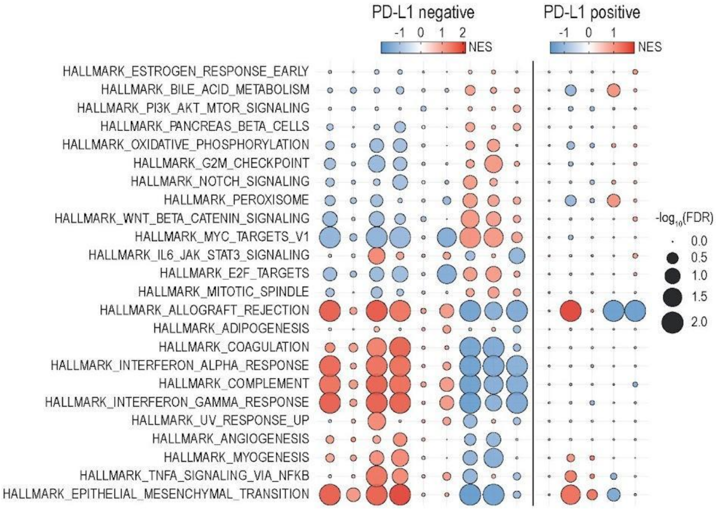


Figure 53. Gene Set Enrichment Analysis Demonstrate Distinguished Functional Specialization.

bubble plot illustrates the results of the GSEA by clusters utilizing curated Hallmark pathway datasets as a referene, for both PD-L1 negative (left) and positive (right) samples. Bubbles are color-coded based on the normalized enrichment score (NES). Bubbles size reports the significance of the enrichment, reported as -$\log_{10}$(adjusted p-value). Gene Set Enrichment Analysis (GSEA); Normalized Enrichment Scores (NES).

Indeed, the spatial visualization of pathway enrichment scores depicted a distinct spatial separation between the tumor and its surrounding regions in most enriched pathways in PD-L1 positive samples and, to a lesser extent, in PD-L1 negative ones, consistent with the GSEA results (Figure 54). More specifically, we computed for each spatial spot their relative pathways enrichment score (with respect to all other spatial spots) employing the approach proposed by Tirosh et al[193]. For the two most significant pathways in either tumor subclusters (MYC targets, Oxidative phosphorylation) or surrounding subclusters (TNFa signaling via NF-kB, epithelial-mesenchymal transition), we calculated the average expression level of the pathway's "leading genes" (key genes that are most strongly associated with the pathway) and then subtracted the aggregated expression of control feature sets with a similar expression level.

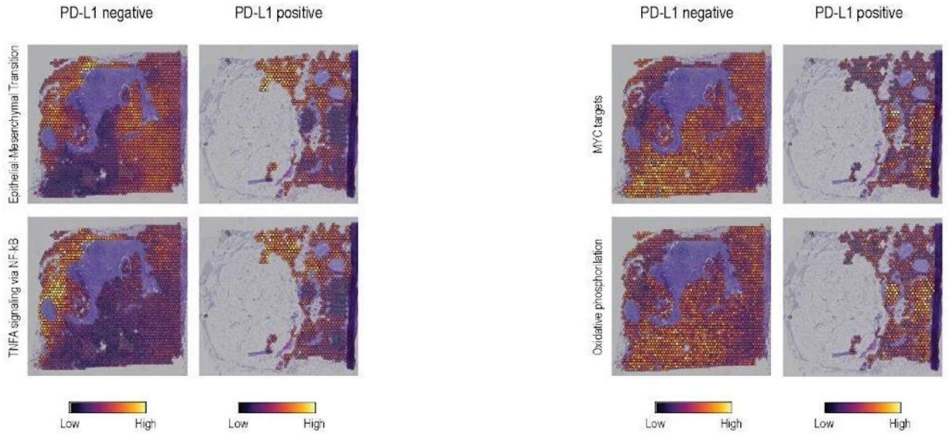


Figure 54. Pathway Enrichment Scores Underscore Distinct Spatial Separation Between the Tumor and its Surrounding Regions

Spatial visualization of pathway enrichment scores derived from the expression levels of pathway's leading genes, with two showing enrichment in the tumor regions, MYC targets and Oxidative phosphorylation (left), and two in the surrounding areas, TNFa signaling via NF-kB and epithelial-mesenchymal transition (right).

Then, to assess the commonalities and uniqueness of PD-L1 expression status in primary TNBC tumors, we compared the upregulated and downregulated SVGs in both samples (Figure 55). Specifically, we performed pairwise differential analysis between all spatial spots within tumor areas versus all the spatial spots in the surrounding areas. Most upregulated SVGs in the tumor areas, 65%, were common to both PD-L1 positive and negative samples, while 11.1% were downregulated in both. On the other hand, 16.6% and 7.3% of SVGs were uniquely upregulated in PD-L1 positive and negative samples tumor areas, respectively. Interestingly, genes upregulated in the tumor area of both samples were linked to known TNBC and general cancer-associated markers, including CTNNB1, DDR1, EPCAM, and KRT7. Conversely, genes downregulated in tumor areas and thus upregulated in the surrounding ones in both samples were predominantly involved in the extracellular matrix's structure, organization, angiogenesis, and metastasis, including DCN, MMP2, FN1, and COL3A1. Finally, a specific inflammatory signature can be identified only in the PD-L1 positive tumor area and is represented by immune hallmarks like STAT2, IFI6, ISG15, and BGN. Interestingly, IFI6, ISG15, and STAT2 are involved in interferon signaling pathways that can play a pivotal role in the upregulation of PD-L1 expression[241]. Additionally, while BGN does not have a direct connection to the PD-L1/PD-1 pathway, its role in immune regulation and inflammatory processes[242] suggests a potential indirect influence.

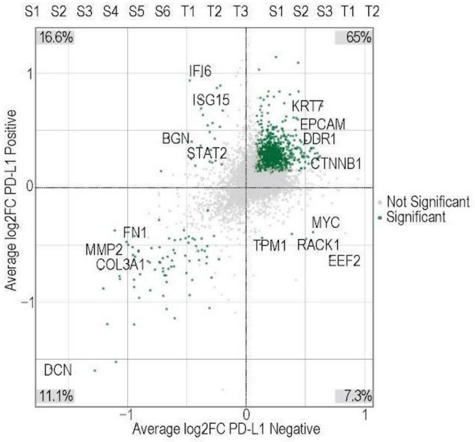


Figure 55. Triple Negative Breast Cancer Positive and Negative to PD-L1 Show Uniq and Shared Spatial Gene Expression Profiles

Scatter plot of genes expressed in tumor- against surrounding-associated clusters, displayed as Log2 Fold Change in both negative (x-axis) and positive (y-axis) samples. Green dots represent differentially expressed genes that are significant (FDR < 0.5). Genes mentioned in the text are reported.

3.2.5. Spatial Transcriptomics Depict Ligand-Receptor Crosstalk in Tumor and Adjacent Tissue Regions

After delineating the distinct functionalities of spatially categorial areas within the tumor and adjacent tissues, we examined potential crosstalks among their subclusters. This is particularly important in the context of PD-L1 to better understand the underlying mechanisms of tumor immune evasion. To this end, we inferred ad hoc interaction between ligands and receptors expressed along the tumoral assigned clusters and the surrounding tissue (Figure 56). Specifically, we adapted our previous ligand-receptor interaction analysis[211] using 2557 experimentally validated pairs[210]. To enhance the precision of our data analysis, we filtered out ligand-receptor pairs that exhibited no expression in the count matrix, resulting in sets of 802 and 769 interactors for PD-L1 negative and positive samples, respectively. We then computed the interaction score by multiplying the average gene expression of a ligand in a specific subcluster with the average value of its corresponding receptor in another cluster. Subsequently, we evaluated the significance of the results under a null hypothesis of spatial spots and cluster annotations. Only significant ligand-receptor pairs were selected, ensuring their enrichment in distinct cluster categories, either within the tumor or the surrounding tissue.

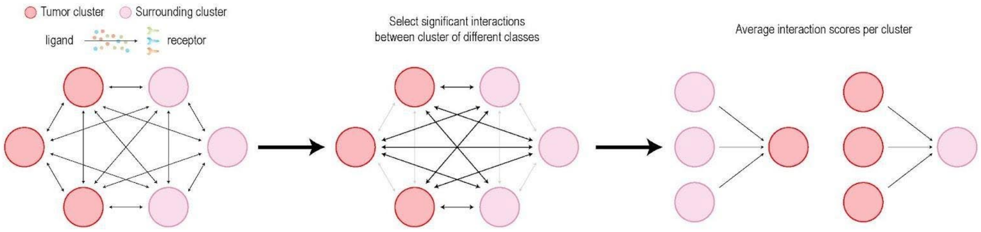

Tumor cluster
Surrounding cluster
ligand
receptor
Select significant interactions between cluster of different classes
Average interaction scores per cluster

Figure 56. Schematic Representation of Our Ligand-Receptor Analysis

The Analysis was applied to preselected interactors (left), emphasizing those with statistical significance across distinct cluster classes, Tumor or Surrounding (middle), while excluding interactions that were significant within the same class (right).

A hierarchical clustering of the interaction scores resulted in 11 Interaction Modules (IMs) exemplifying the intercommunication between distinct clusters within the tumor and the surrounding areas. Among these, 7 IMs were exclusively present in PD-L1 negative samples (Figure 57, left), while 4 were specific to PD-L1 positive samples (Figure 57, right). Although certain IMs demonstrated exclusive communication towards a particular cluster, only three IMs—IM2 in PD-L1 negative and IM11 and IM8 in PD-L1 positive—consistently exhibited robust interaction patterns across all clusters of the same category (either tumor or surrounding), thereby shedding light on the consistent dynamics between tumor and

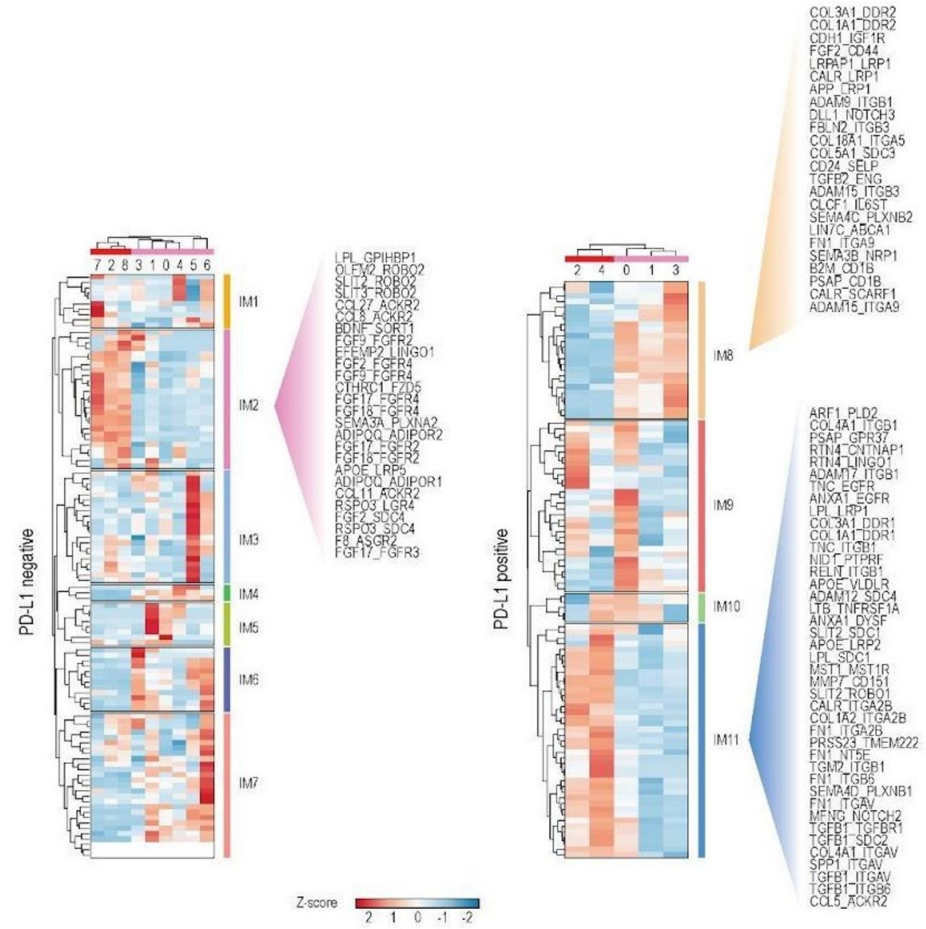


Figure 57. Hierarchical Clustering Analysis Identified Distinct Interaction Modules

Heatmap of mean-centered interaction scores for each cluster in PD-L1 negative (left) and positive (right) samples, with detailed Ligand-receptor couples for selected IMs (IM2, IM8, and IM11). Interaction Modules (IMs).

Then, to characterize the type of interactions within each IM, we performed GO analysis of their respective receptor sets. Specifically, IM2 is characterized by the induction of cancer pathways in the tumoral area of the PD-L1 negative samples, sustained by the secretion of ligands from the surrounding environment. This includes the activation of the Wnt signaling pathway, the PI3K-Akt signaling pathway, the regulation of the bile acid biosynthetic process, and the NOTCH1-mediated regulation of endothelial cell calcification. A prominent example of a ligand-receptor pair involved in this process is APOE and LRP5, which are known to have a role in tumor development through the Wnt/β-catenin pathway and have been recently prioritized in ovarian cancer[243] (Figure 58, top). Likewise, IM11 in the PD-L1 positive sample, demonstrated the impact of the surrounding tissue on the induction of Epithelial-Mesenchymal Transition (EMT) in tumor-associated clusters. This process involves key pathways such as the MAPK signaling pathway, focal adhesion, and TNFR1-mediated signaling, among others. Specifically, the spatial visualization of FN1 and COL3A1 ligands and ITGAV and DDR1 receptors, respectively, distinctly shows the compartmentalization of interactors within the tumor areas and its adjacent tissue (Figure 58, bottom). Conversely, exclusively in the PD-L1 positive sample, IM8 demonstrated that the tumor tissue itself may also secrete ligands. These ligands influence the surrounding extra-tumoral region, by inducing transcriptional programs associated with inflammatory responses. This includes pathways such as Inflammatory Response, TNF-alpha Signaling via NF-kB, Interleukin-6 Signaling, and cytokine-mediated signaling pathways. In this context, the projection of CALR-LRP1 and APP-LRP1 ligand-receptor pair expressions onto the tissue embeddings reveals a pronounced spatial dichotomy, with ligands predominantly located in the tumor region and receptors in the adjacent areas[244,245] (Figure 58, middle). This result is particularly interesting as PD-L1 expression is known to be linked to cellular responses to inflammation and immune signaling[246,247]. Altogether, ligand-receptor interactions allow us to highlight dynamic crosstalk networks occurring in TNBC and how they are specifically re-shaped in an immunomodulatory environment.

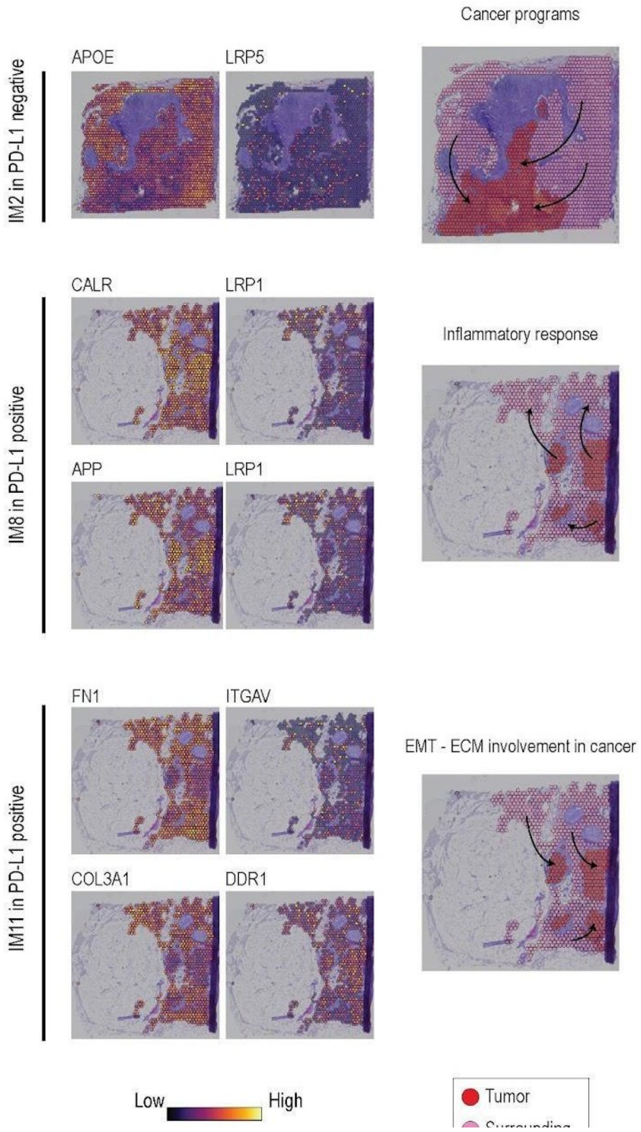


Figure 58. Ligand-Receptor Interactions Highlighted Dynamic Crosstalk between Tumor and Surrounding Tissue

Top: Spatial representation of ligand-receptor interactions between surrounding and tumor areas projected over tissue space for the selected IMs. Black arrows show the directionality of the examined interaction. Bottom: Gene expression of selected ligand and receptor couples for each IM projected over tissue space. Interaction Modules (IMs).

3.2.6. LY6D Unveiled as a Complementary Diagnostic marker to PD-L1 Positivity

Next, in order to explore new potential diagnostic markers that could improve current PD-L1 testing, we combined expression analysis of spatial transcriptomic and clinical grade RNAseq. Firstly, we conducted spatial expression analysis on all tumor and surrounding subclusters in both PD-L1 positive and negative samples. We selected genes that were significantly up-regulated in PD-L1 positive tumor areas but were downregulated or not expressed in the tumor area of PD-L1 negative samples. Then, to assess the suitability of these genes as alternative diagnostic markers for PD-L1, we performed clinical-grade

RNAseq on PD-L1 positive and negative FFPE samples (Table S2). Bulk RNAseq was used for this analysis as it provides a more standardized and consistent sequencing method, with reduced technical artifacts, making it a reliable baseline for confirming the insights gained from spatial transcriptomic analysis. Specifically, we conducted a differential expression analysis using the bulk RNAseq data and compared the results with the spatial expression patterns of the selected genes to determine their concordance. We retained eight candidate genes, ISG15, IFI27, TAP1, OASL, LY6D, CLIC3, RSAD2 (Table S2), which showed higher expression levels in PD-L1 positive tumor areas compared to PD-L1 negative ones (Figure 59).

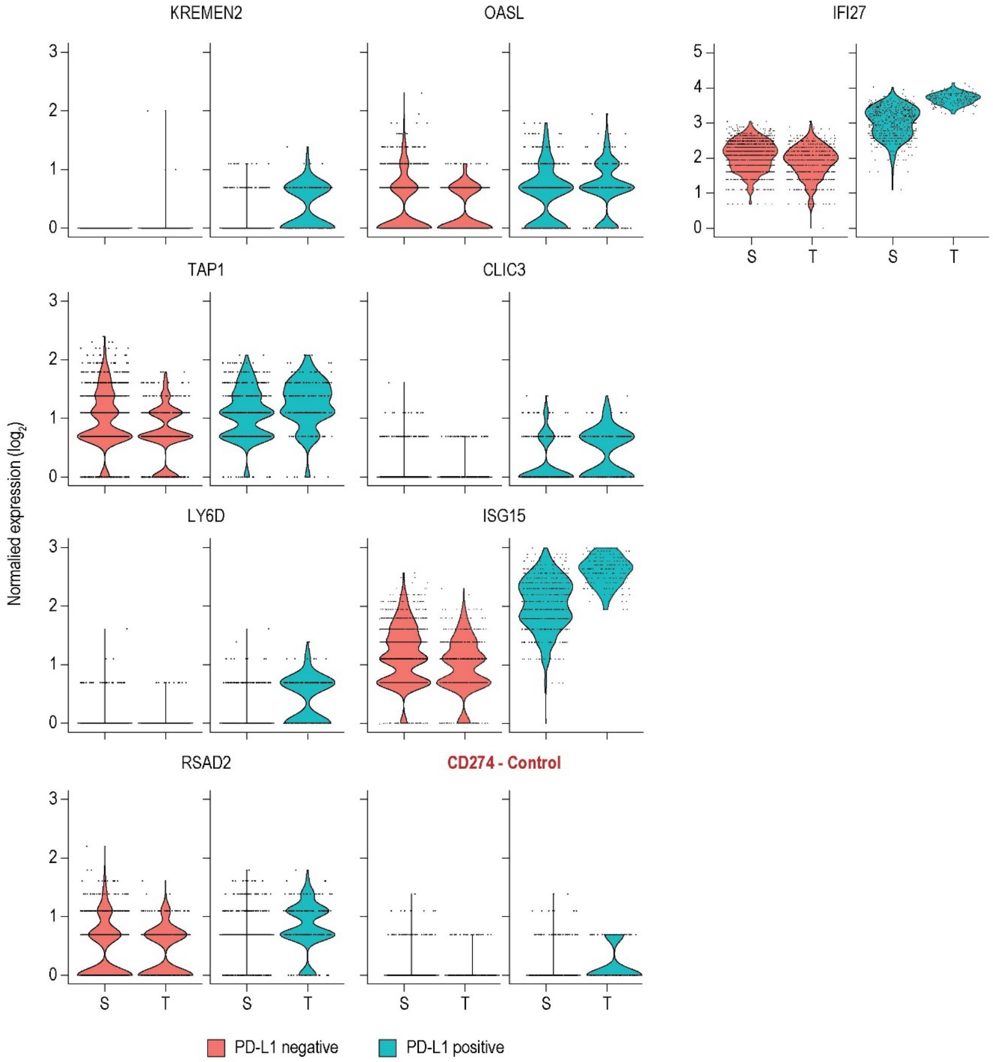


Figure 59. Candidate diagnostic Markers Demonstrated Substantial Specificity in PD-L1 Positive Tumor Areas.

Violin plot representing the normalized expression reads from spatial transcriptomic data of the eight selected candidate diagnostic markers and C274 as a control (mark in red) in PD-L1 positive and negative samples.

As previously shown, the sub-tumor clusters demonstrated an elevated transcriptional heterogeneity. Therefore, we prioritized candidate genes with the highest and most consistent expression across all PD-L1 tumoral clusters. This was determined by assessing the ratio between their average LFC values (tumor vs. surrounding areas) and their respective variance across all sub-tumor clusters (Figure 60). Notably, LY6D (Lymphocyte Antigen 6 Family Member D) emerged as the most prominent diagnostic marker candidate, demonstrating the highest LFC to variance ratio across all tumor clusters in the PD-L1 positive sample (Figure 60). Notably, ISG15 and IFI27 were identified as outliers and subsequently excluded from the pool of candidate diagnostic markers. Despite their high average LFC values, indicative of elevated expression levels in PD-L1 positive tumor areas, the considerable variance in their expression across tumor clusters revealed a lack of consistent expression (Figure S1). This inconsistency within different tumor subpopulations could compromise the reliability of diagnostic results. Therefore, these genes were considered unsuitable for further analysis.

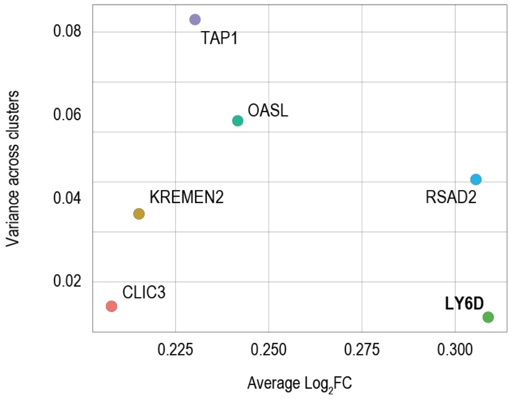


Figure 60. Diagnostic Markers Prioritization Highlight LY6D as the Most Promising Candidate
Scatter plot of the selected diagnostic marker candidates, representing the ratio between their respective LFC variance among tumor clusters in PD-L1 positive sample (y-axis) and their average LFC (x-axis). Notably, ISG15 and IFI27 have been excluded due to their outlier status (see Figure S1). Log2 Fold Change (LFC)

This observation was further corroborated as evident from both spatial transcriptomics and bulk RNAseq data (Figures 59, 61, and Table S2). More specifically, mapping LY6D expression onto tissue samples of PD-L1 positive and negative embeddings revealed an increased number of spatial spots positive for LY6D, particularly concentrated in tumor areas of the PD-L1 positive samples (Figure 61, left). Correspondingly, in the bulk RNAseq analysis, LY6D showed higher expression levels in PD-L1 positive samples compared to the negative ones (Figure 61, right). It was also evident that LY6D exhibited higher expression levels compared to CD274 within the same PD-L1 positive samples, indicating that it may serve as a more detectable diagnostic marker than PD-L1.

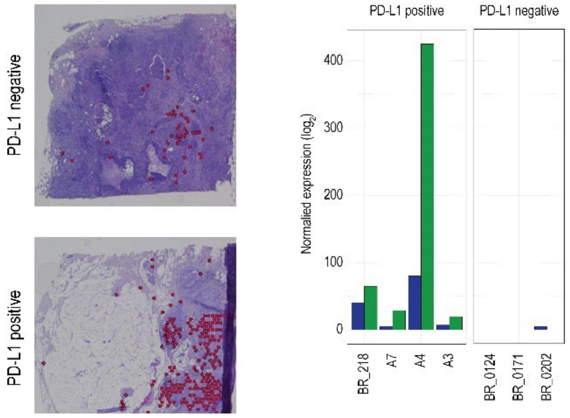


Figure 61. LY6D Exhibits Higher Expression Levels Compared to CD274 in both Spatial Transcriptomics and Bulk RNAseq Data

Comparative spatial expression analysis of LY6D among the two samples. Spatial spots expressing PD-L1 projected over the tissue space (left). A bar plot represents the normalized expression levels of PD-L1 and LY6D in each query sample (right).

To assess the feasibility of our findings, we used an automated IHC workflow to detect LY6D at the proteomic level (Figure 62). We used the same samples that were employed in the bulk RNAseq validation (Figure 61, right). Each sample was divided into three consecutive slices and subjected to separate IHC staining for PD-L1 (as control), LY6D, and H&E staining for histological annotation. To determine the percentage of the tumor area positive for PD-L1 or LY6D, we calculated the results for each slide (see Materials and Methods and Table S3). Consistent with both bulk RNAseq and spatial transcriptomics, IHC analysis showed minimal to no PD-L1 and LY6D expression in negative samples, while demonstrating pronounced signal in PD-L1 positive tumor areas (Figure 62 and Table S3). Furthermore, PD-L1 and LY6D consistently co-localized within the same PD-L1 positive tumor regions, with LY6D exhibiting significantly higher signal intensity compared to PD-L1 (Figure 62, right panels). Overall, the alignment of these findings across diverse analytical methods, probing distinct regulatory layers (proteomic and transcriptional), suggests a complex interplay between PD-L1 positivity and LY6D expression in the intricate and dynamic landscape of TNBC. Essentially, this reinforces the effectiveness of our clinical workflow in precisely mapping complex biological systems.

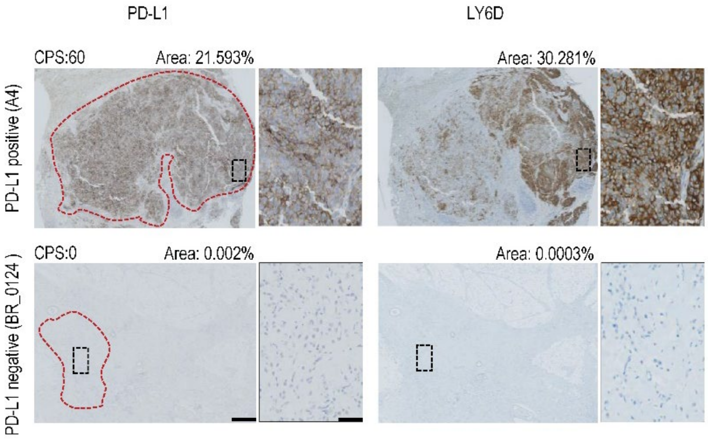


Figure 62. Immunohistochemistry Shows a High and Consistent Presence of LY6D within PD-L1 Positive Sub-tumor Areas

Representative images display IHC staining for PD-L1 (left) and LY6D (right) in two sequential sections from a pair of TNBC samples, PD-L1 positive (top – sample A4) and PD-L1 negative (bottom – sample BR_0124). The left panels for PD-L1 and LY6D depict large tissue sections at a 500 µm scale with tumor regions marked by dashed red lines. Insets in these regions (dashed black lines), magnified to a 50 µm scale on the right panels. The 'Area' metric quantifies signal robustness by averaging the ratio of the stained area to the total area in three tumor regions per sample. The tumor areas, delineated in the PD-L1 panels, correspond to identical regions in the LY6D staining due to the sequential processing of slides 4 µm apart. The CPS score clarifies the PD-L1 status. Immunohistochemistry (IHC); Triple-Negative Breast Cancer (TNBC); Combined Positive Score (CPS)

Next, we validated the expression of LY6D in a cohort of 23 TNBC characterized by high cellularity levels (≥50%), establishing that LY6D exhibits a broader dynamic expression range compared to CD274 (Figure 63). Essentially, the elevated expression levels of LY6D, coupled with its distinguishable and robust IHC signals, provide further evidence of its potential as a complementary marker for enhancing PD-L1 diagnostic assays.

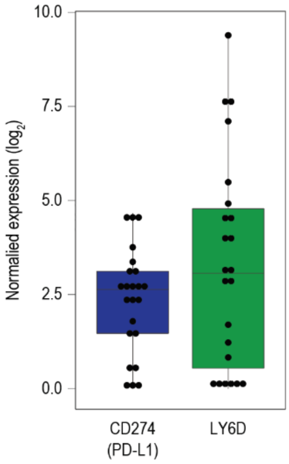


Figure 63. A Cohort of 23 Clinical Triple Negative Breast Cancer Samples Confirming LY6D Broad Expression Dynamic Range

A box plot displays the normalized expression (log2 scale) of CD274 (PD-L1) and LY6D across 23 clinical samples of TNBC with cellularity above 50%. The plot indicates the relative expression levels of each marker, without correlation to PD-L1 immunostatus.

4. Discussion

4.1. First Project - An Integrated Screening to Infer Transcription Factor Regulatory Networks Governing Cell Fate Decisions

Despite recent advances in studying TFs for accurately shaping cellular identity in vitro[51], current cellular conversion methods remain inefficient and yield phenotypically immature target cell types[54,55]. This is due to the complex regulatory landscape of TF activity, which includes multiple constraints that impact their activity[16]. Therefore, when studying TFs functionality it is crucial to consider the pre-established cellular state, which can significantly affect its modus operandi. Correspondingly, cellular conversion protocols necessitate at least one pioneer TF which initiates the commitment toward a unique cell fate by engaging silent and unmarked chromatin[50]. This, in turn, facilitates a permissive epigenetic state that allows additional factors to further specify cell identity[51]. With that said, evidence suggests that even though pioneer TFs possess the ability to reshape the chromatin state, their efficiency in shaping cellular fate hinges on a pre-existing epigenetic state[248]. Indeed, different TF-mediated cellular conversion protocols demonstrate highly variable outcomes, as any changes in the experimental settings can affect the TF functionality, making it challenging to predict its global activity and functionality. Up to date, there is a lack of a systematic workflow to survey TF activity agonistically of the cellular conversion platform. This significantly hampers a comprehensive understanding of molecular pathways regulated by each TF and how they might be interconnected.

In light of this, we hypothesize that various TFs, yet uncharacterized, play crucial roles in shaping cellular fate. Therefore, in this study, we developed a comprehensive transcriptomic, epigenomic, and morphological screening to assess the effect of numerous developmental TFs on cellular transdifferentiation. Our approach represents one, if not the only, case of side-by-side comparison of TF dosages within the same experimental setting. Consequently, this allowed us to identify novel pioneer factors that, either individually or in combination, play a pivotal role in regulating cellular identity.

A recent study, similar in focus to ours, systematically screened 3,548 TF splice isoforms to identify TFs that trigger changes in cellular state[45]. However, the study was conducted on human embryonic stem cells (hESCs), which have a relatively more open overall chromatin structure as compared to differentiated cells. This, in turn, makes their DNA more accessible, thus allowing easier genetic manipulation. Consequently, it may lead to overestimating TF capacity and poorly predicting their global activity in different cellular contexts. Our study complements these observations by accounting for pre-existing epigenetic states, thus offering more applicable insights for in vitro cellular conversion processes.

In this dissertation, I demonstrate a proof-of-principle for our systematic screening approach, focusing on a subset of 130 TFs. This subset is part of a more comprehensive list of 277 TFs, carefully selected for their ability to influence cellular fate at both transcriptional and epigenetic levels (Figure 17). Although our screening study is incomplete, these 130 TFs continue to elucidate the complexity of various structural domains across six superclasses (Figure 18) and shed light on their roles in regulating cell fate decisions.

To impartially compare these TFs and predict their global functionality, we performed the transdifferentiation assays under unfavorable conditions that did not intrinsically promote specific cellular fates. This approach allowed us to highlight TFs that are likely to act as pioneer factors by introducing new cellular functionality despite the impediments derived from the pre-existing epigenetic state.

Indeed, our results indicate that a majority of TFs, particularly those in the Helix-Turn-Helix and Basic Domain categories, significantly impact cellular changes at both transcriptional (67.7%) and epigenetic (54%) levels (Figures 24 and 28). This observation may be influenced by their prevalence among the 130 TFs studied, rather than an inherent effect on gene expression dynamics. Further insights will be gained by completing the screening of all 277 TFs in our study. Interestingly, we observed partial overlap between TFs impacting gene expression and those altering chromatin accessibility, implying varied TF capacities in regulating transcriptional and epigenetic levels (Figure 29).

While the mere presence of TF binding motifs and dynamic gene expression can offer hints of potential regulatory activity, it remains a poor predictor in practice. This is primarily due to the various constraints that influence TF functionality. To address this limitation, we have developed novel computational approaches designed to extract the crucial biological context of TFs from each omic data. Consequently, this has enabled us to go beyond raw count matrices, which otherwise yield primary information on TF functionality, and highlight TFs interconnectivity by focusing on shared signatures of cell fate alterations. Notably, we successfully delineated clusters of TFs based on their enrichment in target gene expression (RNAseq) and motif enrichment similarities (ATACseq) (Figures 27 and 32). We observed that TFs with similar functions or belonging to the same family tend to group closely together. However, despite this resemblance, the clustering results from the different omic levels were not significantly similar, suggesting distinct regulatory mechanisms at the transcriptional and epigenetic levels (Figures 33 and 34).

Nevertheless, our analysis yielded critical observations demonstrating consistent patterns of uncharacterized TF interactions across both cellular levels. For instance, CREB5, a previously uncharacterized TF, was associated with cardiac fate. Similarly, ASCL1, known for its role in neuronal development, was linked to myogenic factors. These results were further corroborated by the multiomic connectivity network that captures both shared and complementary information from epigenetic and transcriptional datasets (Figure 37).

Additionally, by utilizing multiomic data, we developed a computational method to prioritize TFs based on their predominant effect on gene expression and chromatin accessibility (Figure 36). Interestingly, despite displaying different clustering results dependent on their biological context, we found a significant consistency in the TF activity scores across both omic levels. In line with this, our observations revealed that key TFs, pivotal in regulating cellular differentiation across all three germ layers, exhibited high functional activity. These include NEUROD/G genes (ectodermal fate)[230], HNF genes (endodermal fate)[249], and MYOD/G genes (mesodermal fate)[215]. This was observed, although they were introduced into BJ fibroblast cells, which originate from the mesodermal germ layer. Consequently, this suggests that the differences between the inferred clusters of TFs at the transcriptional and epigenetic levels stem from real differences in TFs' functional mechanisms at each omic level and not due to artifacts of the analytical methods or biases in the data.

Overall, in our initial comparative analysis of 130 TFs, we successfully identified uncharacterized TF interactions and effectively distinguished paralog genes, which have evolved to diversify their functions (examples include MYF/MYO genes), from those maintaining identical biological functions, where differences primarily manifest in their spatial and temporal regulation (such as TBX20 and HAND genes). Consequently, this demonstrates the effectiveness of our screening approach in characterizing TF functionalities and revealing novel interconnections, thus providing a reliable approach for studying diverse sets of TFs in various biological contexts.

4.1.1. Perspectives: Enhancing In Vitro Myogenic and Cardiogenic Reprogramming Protocols Leveraging Insights from Our Transcription Factor Atlas Study

After completing the exploratory phase across all 277 selected TFs, they will be further tested in combination to improve conversion assay toward a myogenic and cardiac fate. We specifically selected these two cellular identities due to their unique yet interconnected nature, offering insights into the complexities of cell fate determination and potential advancements in regenerative medicine.

The myogenic pathway, centered on the pioneering role of MYOD1, offers a classic model for understanding how pioneer factors can initiate and direct cell fate conversion[40]. This pathway has historically been a cornerstone in cellular reprogramming, demonstrating the profound potential of specific TFs in cell fate determination. However, to date, the process is challenged by the limited responsiveness of certain cell types, such as ADSCs and MSCs, for clinical application of myocyte fate conversion protocols[250]. Additionally, the inability to achieve full maturation of muscle fibers underlines the need to investigate and overcome the influence of the chromatin epigenetic landscape on TF activity.

In contrast, cardiac differentiation, while sharing some mechanistic similarities with myogenic differentiation, presents further challenges. The complexity of inducing cardiomyocyte fate in human cells, requiring a composite orchestration of multiple TFs and microRNAs, contrasts with the simpler processes observed in murine models[251,252]. The fact that the most advanced human induced cardiomyocytes (iCM) generation protocols achieve only partial success and result in cells with immature properties, further emphasizes the need for comprehensive studies in this area[253].

To find potential candidate TFs that, in combination with the traditional ones can induce myogenic or cardiac fate, we will create multiomic network of 277 TFs to identify the nearest neighbors of known myogenic/cardiac reprogramming TFs and compute their connectivity level in the network. Using this information, we will compile a list of potential candidate TFs. These candidates will be ranked based on the product of their connectivity level and the activity score as we demonstrated in this dissertation. Additionally, we will further analyze the morphological changes to provide additional impetus to precise the concluded multiomic TF connectivity network.

Eventually, the expected data will represent a state-of-the-art encyclopedia of pioneer TF landscapes, including binding sites, downstream target genes, and enriched functional pathways.

4.2. Second Project - Spatial Transcriptomics Reveals Sub-Tumoral Identities and Novel Diagnostic markers in Triple Negative Breast Cancer With Immune Evasion Capacity

Recent advancements in spatial transcriptomics have shed light on novel approaches for investigating breast tumor heterogeneity and microenvironmental composition[254,255]. In this study, we have developed a novel clinical workflow that seamlessly integrates spatial transcriptomics with detailed histological examination, bulk RNAseq and IHC. This methodology was used to explore TNBC biological landscape, focusing on PD-L1 expression status and its role in tumor immune escape mechanisms (Figure 41). Our methodological approach was specifically designed to be a cost-efficient and easy-to-implement strategy in any lab bench, requiring minimal input sample sizes yet ensuring the preservation of data integrity.

A recent study, which shares a similar scope to ours, delved into the spatial features of TNBC while focusing on immune checkpoint pathways[256]. Specifically, they utilized imaging mass cytometry (IMC) to analyze 43 proteins and bulk RNAseq of 101 genes for subtyping purposes. In contrast, our study completes these observations with a more comprehensive approach by employing a non-targeted experimental approach using spatial transcriptomics and RNAseq alongside IHC validation.

While conventional histopathology remains the benchmark in tumor diagnostics, its accuracy can be variable and often necessitates specialized expertise that may not be readily available in pathology laboratories. This challenge is especially pronounced in the detection of PD-L1 in TNBC, necessitating the collaboration of skilled pathologists, state-of-the-art instrumentation, and digital pathology technologies[116]. Indeed, Due to this challenge there is a lack of standardization in PD-L1 diagnostic tests, leading to complexity in selecting patients immunotherapies[257]. Accordingly, the variability of the overall efficacy of PD-1/PD-L1 targeted therapy in TNBC is also a significant concern[258,259].

To address this, we applied an unbiased spatial transcriptomics analysis, which not only enabled a precise determination of PD-L1 (CD274) expression status in-situ, but also facilitated detailed gene expression mapping. This approach was crucial for the in-depth analysis of tumor architecture and the identification of potential alternative diagnostic markers, thereby expanding the scope of current oncological investigations. These results were further corroborated by clinical-grade RNAseq and IHC, which are globally accepted and standardized experimental approaches. This demonstrates consistent results at both the transcriptomic and proteomic cellular levels (Figures 61 and 62).

We first confirmed the spatial integrity of our sequencing data by conducting a thorough validation analysis using the histological annotations as our benchmark. Aligned with previous studies[236], our findings indicated an elevated transcriptional activity in tumor regions compared to the adjacent tissues (Figures 43 and 44). Additionally, we observed that known tissue markers accurately correspond to their respective histological compartments (Figure 46). Our analysis also showed that the expression profiles of CD274 were in line with the initial assessment of PD-L1 status (positive or negative) (Figure 45). These findings were in agreement with previous studies indicating that PD-L1 can be expressed by either infiltrating immune cells within the microenvironment or by tumor cells[237,238]. Overall, these results demonstrate the effectiveness of our methodological approach in maintaining tissue architecture while providing a comprehensive expression profile.

Therefore, we employed spatial transcriptomic data to enhance the precision of tumor and stromal annotations, addressing potential oversights inherent in traditional histological methods (Figure 49). Subsequent downstream analysis identified spatial signatures within the reclassified tumor clusters that were linked to various cancer hallmarks, with a specific focus on breast cancer. Conversely, marker SVGs in the surrounding tissue correlated with well-known characteristics of the tumor microenvironment[260] (Figures 52 and 53).

By refining our previously developed ligand-receptor analysis[211], we successfully pinpointed crucial interaction pairs that are essential for the communication between the tumor and adjacent tissue clusters (Figure 56). This highlights the importance of spatial interactions in the dynamics of the tumor microenvironment. Notably, among the three identified interaction modules, IM8, which was prevalent in PD-L1 positive samples, shed light on the potential role of tumor-secreted ligands in modulating inflammatory responses (Figure 57). This finding is in line with earlier research on the involvement of inflammasomes in immune checkpoint responses, including the expression of PD-L1[261,262].

The in-situ detection of mRNA in both tumor and stromal areas of PD-L1 positive and negative patients enabled us to prioritize alternative diagnostic marker candidates for PD-L1. During this investigation, LY6D emerged as a promising diagnostic marker with the highest and most consistent occurrence in PD-L1 positive sub-tumor regions (Figure 60). This was further corroborated in a broader sequencing cohort and at the protein level (Figures 61-63). Canonically, LY6D is involved in lymphocyte differentiation[263], and recent gene expression atlases allowed us to track its expression in several epithelial tissues, including certain glandular breast subtypes[264]. This finding suggests a connection between LY6D expression and both immune and glandular myo/epithelial components, which are notably altered in TNBC. Moreover, a previous study indicated that proteins in the LY6 family can enhance cytokine-induced PD-L1 activation, leading to immune evasion[265]. These insights imply a potential functional or regulatory link between LY6D and PD-L1.

PD-L1 serves a dual role in TNBC therapy, acting as both a predictive diagnostic marker and a therapeutic target. However, its low expression levels present significant challenges in diagnosis and antigen targeting[266,267], complicating patient management for targeted therapies. This factor has been hypothesized as a key reason behind the varied responses observed in anti-PD-1/PD-L1 therapies[257,268]. Given these considerations and in light of our research findings, we propose LY6D as a complementary diagnostic marker to PD-L1 in immune checkpoint-positive TNBC. This suggestion is based on the heightened expression dynamic range of LY6D (Figures 59, 61, and 63) and its more distinguishable and robust IHC signals, even when using research-grade antibodies.

In conclusion, our research highlights the critical importance of integrating spatial transcriptomics and histopathology to decode the intricate structure and ligand-receptor dynamics in tumor environments. This approach has shed light on new therapeutic possibilities for treating both PD-L1 positive and negative TNBC. The findings from our study are particularly promising for improving diagnostic methods and advancing therapeutic trials, especially with the use of emerging diagnostic markers such as LY6D. This integrated method marks a significant stride towards more effective and personalized cancer treatments.

Abbreviations

Assay for Transposase-Accessible Chromatin with Sequencing (ATACseq)

Basal-Like 1 (BL1)

Basal-Like 2 (BL2)

Batch Effect Correction (BEC)

BJ fibroblast telomerase (BJ-T)

Bovine Serum Albumin (BSA)

Chromatin Immunoprecipitation Sequencing (CHIPseq)

Combined Positive Score (CPS)

Combining Chromatin Immunoprecipitation (ChIP)

Complementary DNA (cDNA)

Counts Per Million (CPM)

Differential Expression Analysis (DEA)

Differentially Expressed (DE)

DNA Binding Domain (DBD)

DNase I Hypersensitive Sites Sequencing (DNaseseq)

Electrophoretic Mobility Shift Assay (EMSA)

Enhancer RNAs (eRNAs)

Epidermal Growth Factor Receptor (EGFR)

Epithelial-to-Mesenchymal Transition (EMT)

Estimation of Stromal and Immune cells in MAlignant Tumor tissues using Expression data (ESTIMATE)

Estrogen Receptor (ER)

European Institute of Oncology (IEO)

False Discovery Rate (FDR)

Formalin-Fixed Paraffin-Embedded (FFPE)

Fraction of Reads in Peaks (FRiP)

gap junctions (GJs)

Gene Expression sequencing (CAGEseq)

Gene ontology (GO)

Gene Set Enrichment Analysis (GSEA)

General Transcription Factors (GTFs)

Glucocorticoid Receptor (GR)

Green fluorescent protein (GFP)

Hematoxylin and Eosin (H&E)

Human Embryonic Stem Cell (hESCs)

Human Epidermal Growth Factor Receptor 2 (HER2)

Human Foreskin Fibroblasts (BJ)

Human Lung Fibroblasts (HLF)

Human Telomerase Reverse Transcriptase (hTERT)

Imaging Mass Cytometry (IMC)

Immune Checkpoint Inhibitors (ICIs)

Immunohistochemistry (IHC)

Immunohistochemistry (IHC)

Immunoprecipitation-Mass Spectrometry (IP-MS)

Induced Cardiomyocytes (iCM)

Induced Pluripotent Stem Cells (IPSCs)

Inner Cell Mass (ICM)

Interaction Modules (IMs)

Interferon-γ (IFN-γ)

Irreproducible Discovery Rate (IDR)

Irreproducible Discovery Rate (IDR)

Log fold change (LFC)

Long Noncoding RNA (lncRNA)

Luminal Androgen Receptor (LAR)

Mesenchymal (M)

Mesenchymal-to-Epithelial transition (MET)

MicroRNA (miRNA)

Mode of Action by NeTwoRk Analysis (MANTRA)

Multiplicity of infection (MOI)

Negative Binomial (NB)

Next-Generation Sequencing (NGS)

Normalized Enrichment Score (NES)

Nuclear Receptor Coactivators (NCOAs)

Nucleosome-Free Regions (NFRs)

Nucleosome-Free Regions (NFRs)

open chromatin regions (OCR)

Open Reading Frame (ORF)

Overall Response Rates (ORR)

Paraformaldehyde (PFA)

Positional Weight Matrix (PWM)

Post-Translation Modifications (PTM)

Pre-Initiation Complex (PIC)

Principal Component Analysis (PCA)

Progesterone Receptor (PR)

Programmed cell death protein 1 (PD-1)

Programmed death-ligand 1 (PD-L1)

Protein-Protein Interaction (PPI)

Quality Control (QC)

Reads Per Kilobase per Million mapped reads (RPKM)

Retinal Pigment Epithelium (RPE)

RNA And DNA Interacting Complexes Ligated and sequenced (RADICLseq)

RNA polymerase II (Pol II)

RNA Sequencing (RNAseq)

Similarity Network Fusion (SNF)

Single-Cell RNA Sequencing (scRNAseq)

Sparse Partial Least Squares regression (sPLS)

Spatially Variable Genes (SVGs)

T Cell Receptor (TCR)

Transcription Factor (TF)

Transcription Factor Binding Sites (TFBS)

Triple Negative Breast Cancer (TNBC)

Tumor Microenvironment (TME)

Tumor-Infiltrating Lymphocytes (TILs)

Uniform Manifold Approximation and Projection (UMAP)

Unique Molecular Identifiers (UMIs)

Supplementary Data

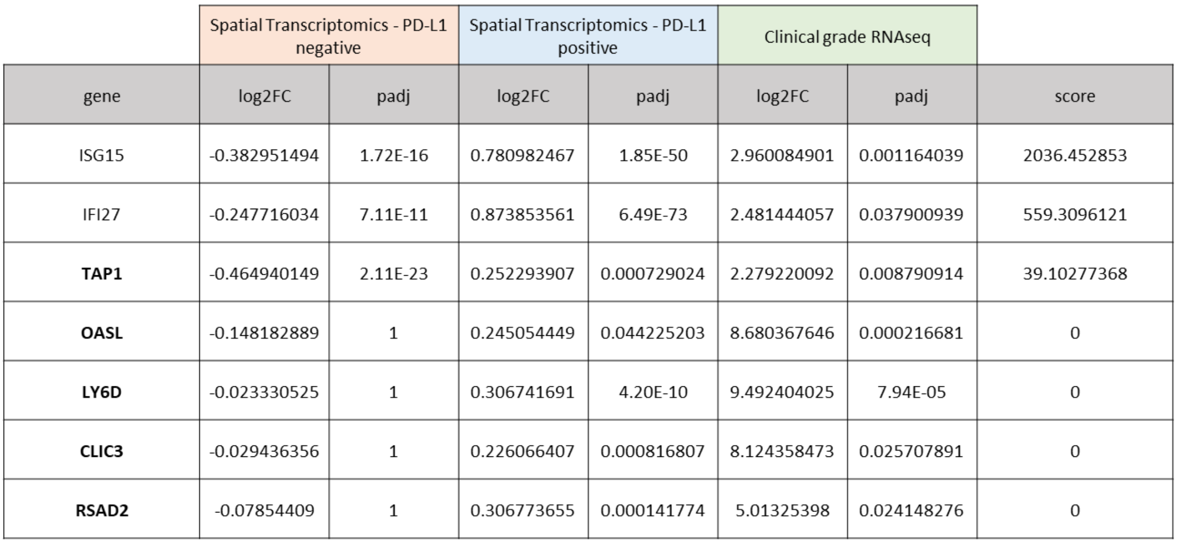
Table S1

Step	Description	Special Argument	Programs	Inputs	Outputs	Main Commands
1) Demultiplexing	Conversion of BCL to FASTQ files using BCL2fastq v2.20.0.422	--use-bases-mask Y51I8I8Y51	BCL2fastq	BCL format data	FASTQ files	BCL2fastq conversion
2) Adaptor Trimming and Pre-Alignment QC	Adaptors trimming with TrimGalore v0.6.3	--nextera, --length 30, --fastqc	TrimGalore	Raw fastqs	Trimmed fastq files	--fastqc --paired commands
3) Alignment and Preliminary filtering	Mapping reads to the human reference genome hg38 using BWA-MEM (v0.7.10) and SAMTools (v1.3)	-M, -F 0x0100	BWA-MEM, SAMTools	Trimmed fastq files	Sorted-indexed-BAM files	bwa mem and samtools commands
4) Evaluate library complexity	Assessing library complexity with Preseq (v2.0.0)	N/A	preseq	Raw aligned reads	TXT file for MultiQC	Preseq lc_extrap command
5) Remove Duplicates	Using Picard tools (v2.9.2) for removing duplicates	REMOVE_DUPLICATES=T	Picard MarkDuplicates	Primary BAM file	Semi filtered BAM file	Picard MarkDuplicates command
6) Secondary filtering	Refining the read dataset with SAMTools (v1.3) and BAMTools (v2.2.2)	Various, including tag and cigar options	samtools, bamtools	Sorted semi-filtered BAM	Semi-filtered BAM	SAMTools and BAMTools filtering commands
7) Third filtering	Filtering orphan reads and mates with a custom Python script	N/A	bampe_rm_orphan.py	Sorted semi-filtered BAM	Final filtered BAM	bampe_rm_orphan.py command
8) Assessing Replicate Correlation	Using Deeptools' tools (v3.5.1) for replicate BAM files correlation	N/A	Deeptools	Final filtered BAM file	Correlation matrix, plots	Deeptools multiBamSummary and plotCorrelation commands
9) Conversion to BEDPE and Alignment Shifting	Converting BAMPE to BEDPE and adjusting alignments using BEDTools (v2.29.1) and an awk script	N/A	BEDTools	Final filtered BAM file	BEDPE file	BEDTools and awk commands

10) Peak calling and FRiP QC	Peak calling with MACS2 (v2.2.7.1) adapted for ATAC-seq	--shift -100, --extsize 200	MACS2	BEDPE	Peak calling files	MACS2 callpeak command
11) Assessing Replicate Similarity	Using BEDTools (v2.29.1) for evaluating replicate peak similarity	Jaccard coefficient threshold >= 0.5	BEDTools	Peak files	Jaccard coefficient plot	BEDTools jaccard command
12) Peak Calling QC Plots	Generating QC plots for peak calling using an R script from NextFlow	N/A	Rscript, NextFlow script	Peak files	PDF with QC plots	R script commands for QC plots
13) Identifying Consistent Peaks with IDR	Using IDR software (v2.0.3) for assessing peak reproducibility	-log10(p-value) > 1.30103	IDR	Peak files	IDR intersected peaks	IDR commands
14) ATAQV	QC of peaks using ATAQV software (v1.2.1) for ATACseq-specific visualizations	N/A	ataqv	Various, including peak and BAM files	Json, html files	ataqv command
15) Create bigwig files	Creating normalized bigwig files using BEDTools genomecov (v2.29.1) and bedGraphToBigWig (v2.9)	N/A	bedtools, bedGraphToBigWig	Final filtered BAM files	Bigwig, bedgraph files	BEDTools and bedGraphToBigWig commands
16) Create a consensus peak- set	Integrating individual peak sets into a consensus set using BEDTools merge (v2.29.1)	N/A	BEDTools	IDR peak files	Consensus peak set	BEDTools merge command
17) Peaks Annotation	Annotating peaks using HOMER's annotatePeaks.pl	N/A	annotatePeaks.pl	Consensus peak set	Annotated peak file	HOMER annotatePeaks command
18) Create SAF file and Count Reads	Counting reads with featureCounts (v2.0.3) using SAF files	N/A	featureCounts	Annotated peak file, BAM files	Peak count matrix	featureCounts command

Table S1 ATACseq Preprocessing Pipeline Steps. This table outlines the 18 steps of the ATACseq analysis pipeline, designed to process sequencing data from initial raw BCL files to the final peak count matrix. Our approach integrates best practices and methodologies from a variety of established pipelines and research findings, as described in the materials and methods chapter. For each step, the table specifies its objective, the software and commands utilized, the parameters applied, and the formats of the input and output data.

Table S2



	Spatial Transcriptomics - PD-L1 negative		Spatial Transcriptomics - PD-L1 positive		Clinical grade RNAseq		
gene	log2FC	padj	log2FC	padj	log2FC	padj	score
ISG15	-0.382951494	1.72E-16	0.780982467	1.85E-50	2.960084901	0.001164039	2036.452853
IFI27	-0.247716034	7.11E-11	0.873853561	6.49E-73	2.481444057	0.037900939	559.3096121
TAP1	-0.464940149	2.11E-23	0.252293907	0.000729024	2.279220092	0.008790914	39.10277368
OASL	-0.148182889	1	0.245054449	0.044225203	8.680367646	0.000216681	0
LY6D	-0.023330525	1	0.306741691	4.20E-10	9.492404025	7.94E-05	0
CLIC3	-0.029436356	1	0.226066407	0.000816807	8.124358473	0.025707891	0
RSAD2	-0.07854409	1	0.306773655	0.000141774	5.01325398	0.024148276	0

Table S2. Differential Expression Analysis of PD-L1 Candidate Genes Across Spatial Transcriptomics and Clinical Bulk RNAseq. This table compiles differential expression analysis for PD-L1 positive and negative samples, integrating data from spatial transcriptomics and clinical-grade bulk RNA sequencing. For spatial transcriptomics, analysis compared tumor areas with surrounding spots in PD-L1 positive (middle column) and negative (right column) samples. In bulk RNAseq, differential expression between PD-L1 positive and negative samples is evaluated (left column). Filters are applied to select genes with a log fold change (LFC) > 0 and p-value < 0.05, specific to PD-L1 positive samples in both spatial transcriptomics and clinical RNAseq, while in spatial transcriptomic data of PD-L1 negative sample, we applied log fold change (LFC) < 0. Gene rankings are determined by a composite score, multiplying LFC by the negative logarithm of adjusted p-values across datasets, highlighting the top 8 candidates highlighted. The final six candidates are emphasized in bold text, while the two genes excluded due to being outliers are not highlighted.

Table S3: Immunohistochemistry test results

Condition	Sample	gene	position_1	position_2	position_3
PDL1_pos	BR218	PD-L1	0.0181	0.0181	0.0181
		LY6D	0.1869	0.1568	0.3504
	A7	PD-L1	0.4845	0.7678	0.1025
		LY6D	8.2205	3.1233	8.9423
	A4	PD-L1	6.3081	30.9757	27.3345
		LY6D	14.3203	31.7115	44.8106

	BR593	PD-L1	0.0002	0.0002	0.0001
		LY6D	3.9638	0.025	4.65
	A3	PD-L1	0.5778	0.1133	0.2034
		LY6D	0	0.0001	0.0005
PDL1_neg	BR118	PD-L1	0.0021	0.0009	0.0041
		LY6D	2.2945	0.0632	0.0525
	BR0124	PD-L1	0	0.0004	0
		LY6D	0.0002	0	0
	BR171	PD-L1	0.0022	0.0115	0.0004
		LY6D	0.0014	0.0029	0.0005
	BR202	PD-L1	0.0013	0.0039	0.0013
		LY6D	0	0	0.0008
	BR0164	PD-L1	0.0012	0.003	0.0068
		LY6D	0.0013	0.0067	0.0003

Figure S1

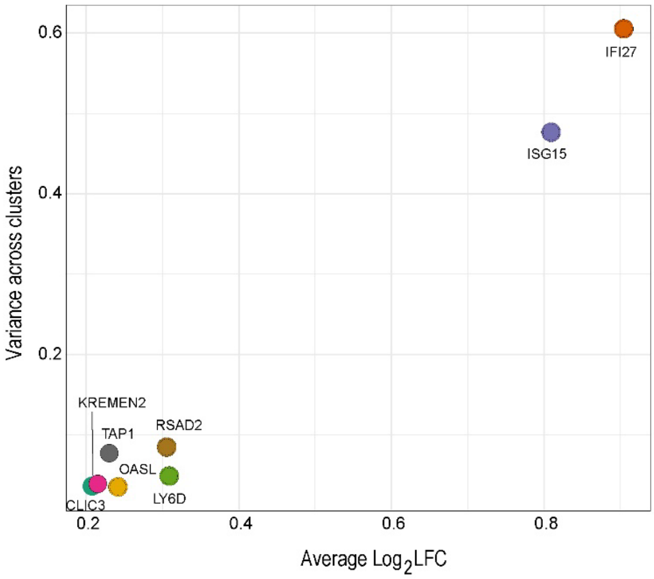


Figure S1. Prioritization of Candidate Diagnostic Markers Highlight ISG15 and IFI27 as Outliers. Scatter plot of the all 8 selected diagnostic marker candidates, representing the ratio between their respective LFC variance among tumor clusters in PD-L1 positive sample (y-axis) and their average LFC (x-axis). ISG15 and IFI27 are highlighted as outliers due to their marked variance, indicating substantial inconsistency in expression across different tumor clusters.

References

1. Fulton, D. L. *et al.* TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* **10**, R29 (2009).
2. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
3. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).
4. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
5. Takahashi, K. & Yamanaka, S. A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.* **17**, 183–193 (2016).
6. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
7. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
8. Latchman, D. S. Transcription factors: an overview. *Int. J. Biochem. Cell Biol.* **29**, 1305–1312 (1997).
9. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390.e19 (2019).
10. Nagy, G. & Nagy, L. Motif grammar: The basis of the language of gene expression. *Comput. Struct. Biotechnol. J.* **18**, 2026–2032 (2020).
11. Frietze, S. & Farnham, P. J. Transcription factor effector domains. *Subcell. Biochem.* **52**, 261–277 (2011).
12. Wingender, E., Schoeps, T., Haubrock, M. & Dönitz, J. TFClass: a classification of

human transcription factors and their rodent orthologs. *Nucleic Acids Res.* **43**, D97–102 (2015).

13. Filtz, T. M., Vogel, W. K. & Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol. Sci.* **35**, 76–85 (2014).
14. Kim, H. K., Jeong, M. G. & Hwang, E. S. Post-Translational Modifications in Transcription Factors that Determine T Helper Cell Differentiation. *Mol. Cells* **44**, 318–327 (2021).
15. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
16. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
17. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
18. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
19. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
20. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **17**, 284–299 (2016).
21. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol. Cell* **76**, 306–319 (2019).
22. Li, M. *et al.* Dynamic regulation of transcription factors by nucleosome remodeling. *Elife* **4**, (2015).
23. Weikum, E. R., Knuesel, M. T., Ortlund, E. A. & Yamamoto, K. R. Glucocorticoid receptor control of transcription: precision and plasticity via allostery. *Nat. Rev. Mol.*

Cell Biol. **18**, 159–174 (2017).

24. Lodrini, M. *et al.* P160/SRC/NCoA coactivators form complexes via specific interaction of their PAS-B domain with the CID/AD1 domain. *Nucleic Acids Res.* **36**, 1847–1860 (2008).
25. Macian, F. NFAT proteins: key regulators of T-cell development and function. *Nat. Rev. Immunol.* **5**, 472–484 (2005).
26. Centore, R. C., Sandoval, G. J., Soares, L. M. M., Kadoch, C. & Chan, H. M. Mammalian SWI/SNF Chromatin Remodeling Complexes: Emerging Mechanisms and Therapeutic Strategies. *Trends Genet.* **36**, 936–950 (2020).
27. Richter, W. F., Nayak, S., Iwasa, J. & Taatjes, D. J. The Mediator complex as a master regulator of transcription by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **23**, 732–749 (2022).
28. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **8**, 93–103 (2007).
29. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
30. Melgar, M. F., Collins, F. S. & Sethupathy, P. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.* **12**, R113 (2011).
31. Chang, T.-C. *et al.* Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell* **26**, 745–752 (2007).
32. Johnson, S. M. *et al.* RAS is regulated by the let-7 microRNA family. *Cell* **120**, 635–647 (2005).
33. Kino, T., Hurt, D. E., Ichijo, T., Nader, N. & Chrousos, G. P. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.* **3**, ra8 (2010).
34. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
35. Gutschner, T., Hämmerle, M. & Diederichs, S. MALAT1 -- a paradigm for long noncoding RNA function in cancer. *J. Mol. Med.* **91**, 791–801 (2013).

36. Islam, Z. *et al.* Transcription Factors: The Fulcrum Between Cell Development and Carcinogenesis. *Front. Oncol.* **11**, 681377 (2021).

37. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970–985 (2011).

38. Guo, X., Zhang, Y., Huang, H. & Xi, R. A hierarchical transcription factor cascade regulates enteroendocrine cell diversity and plasticity in Drosophila. *Nat. Commun.* **13**, 6525 (2022).

39. of Health, N. I. & Others. Stem cells: scientific progress and future research directions. *(No Title)* (2001).

40. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).

41. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

42. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

43. Parekh, U. *et al.* Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. *Cell Syst* **7**, 548–555.e8 (2018).

44. Ng, A. H. M. *et al.* A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol.* **39**, 510–519 (2021).

45. Joung, J. *et al.* A transcription factor atlas of directed differentiation. *Cell* **186**, 209–229.e26 (2023).

46. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).

47. Ieda, M. *et al.* Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**, 375–386 (2010).

48. Huang, P. *et al.* Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475**, 386–389 (2011).

49. Szabo, E. *et al.* Direct conversion of human fibroblasts to multilineage blood

progenitors. *Nature* **468**, 521–526 (2010).

50. Zaret, K. S. Pioneer Transcription Factors Initiating Gene Network Changes. *Annu. Rev. Genet.* **54**, 367–385 (2020).
51. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* **28**, 2679–2692 (2014).
52. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994–1004 (2012).
53. Cacchiarelli, D. *et al.* Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. *Cell* **162**, 412–424 (2015).
54. Jopling, C., Boue, S. & Izpisua Belmonte, J. C. Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nat. Rev. Mol. Cell Biol.* **12**, 79–89 (2011).
55. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: approaches, mechanisms and progress. *Nat. Rev. Mol. Cell Biol.* **22**, 410–424 (2021).
56. Tonge, P. D. *et al.* Divergent reprogramming routes lead to alternative stem-cell states. *Nature* **516**, 192–197 (2014).
57. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55 (2008).
58. Polo, J. M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617–1632 (2012).
59. Unternaehrer, J. J. *et al.* The epithelial-mesenchymal transition factor SNAIL paradoxically enhances reprogramming. *Stem Cell Reports* **3**, 691–698 (2014).
60. Odom, D. T. Identification of Transcription Factor-DNA Interactions In Vivo. *Subcell. Biochem.* **52**, 175–191 (2011).
61. Guertin, M. J., Martins, A. L., Siepel, A. & Lis, J. T. Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet.* **8**, e1002610 (2012).
62. Grath, A. & Dai, G. Direct cell reprogramming for tissue engineering and regenerative medicine. *J. Biol. Eng.* **13**, 14 (2019).

63. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

64. Gagliano, O. *et al.* Microfluidic reprogramming to pluripotency of human somatic cells. *Nat. Protoc.* **14**, 722–737 (2019).

65. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).

66. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–5 (2016).

67. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

68. Spencer, V. A., Sun, J.-M., Li, L. & Davie, J. R. Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. *Methods* **31**, 67–75 (2003).

69. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–73 (2006).

70. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).

71. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).

72. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

73. Woo, A. J., Dods, J. S., Susanto, E., Ulgiati, D. & Abraham, L. J. A proteomics approach for the identification of DNA binding activities observed in the electrophoretic mobility shift assay. *Mol. Cell. Proteomics* **1**, 472–478 (2002).

74. Mehmood, S., Allison, T. M. & Robinson, C. V. Mass spectrometry of protein complexes: from origins to applications. *Annu. Rev. Phys. Chem.* **66**, 453–474 (2015).

75. Trinkle-Mulcahy, L. *et al.* Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J. Cell Biol.* **183**, 223–239 (2008).

76. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, db.prot5384 (2010).

77. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).

78. Bonetti, A. *et al.* RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. *Nat. Commun.* **11**, 1018 (2020).

79. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).

80. Nguyen, H., Shrestha, S., Draghici, S. & Nguyen, T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* **35**, 2843–2846 (2019).

81. Rappoport, N. & Shamir, R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* **35**, 3348–3356 (2019).

82. Louhimo, R. & Hautaniemi, S. CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* **27**, 887–888 (2011).

83. Dimitrakopoulos, C. *et al.* Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018).

84. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).

85. Mo, Q. *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19**, 71–86 (2018).

86. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

87. Shi, Q. *et al.* Pattern fusion analysis by adaptive alignment of multiple heterogeneous

omics data. *Bioinformatics* **33**, 2706–2714 (2017).

88. Yuan, Y., Savage, R. S. & Markowetz, F. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.* **7**, e1002227 (2011).

89. Csala, A., Zwinderman, A. H. & Hof, M. H. Multiset sparse partial least squares path modeling for high dimensional omics data analysis. *BMC Bioinformatics* **21**, 9 (2020).

90. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15**, 162 (2014).

91. Im, K., Mareninov, S., Diaz, M. F. P. & Yong, W. H. An Introduction to Performing Immunofluorescence Staining. *Methods Mol. Biol.* **1897**, 299–311 (2019).

92. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

93. Conic, S. *et al.* Imaging of native transcription factors and histone phosphorylation at high resolution in live cells. *J. Cell Biol.* **217**, 1537–1552 (2018).

94. Morris, S. A. & Daley, G. Q. A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell Res.* **23**, 33–48 (2013).

95. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **73**, 17–48 (2023).

96. Giaquinto, A. N. *et al.* Breast Cancer Statistics, 2022. *CA Cancer J. Clin.* **72**, 524–541 (2022).

97. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).

98. Prat, A. *et al.* Predicting response and survival in chemotherapy-treated triple-negative breast cancer. *Br. J. Cancer* **111**, 1532–1541 (2014).

99. Garrido-Castro, A. C., Lin, N. U. & Polyak, K. Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment. *Cancer Discov.* **9**, 176–198 (2019).

100. Prat, A. *et al.* Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *Oncologist* **18**, 123–133 (2013).

101. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and

therapeutic implications. *World J. Clin. Oncol.* **5**, 412–424 (2014).

102. Malorni, L. *et al.* Clinical and biologic features of triple-negative breast cancers in a large cohort of patients with long-term follow-up. *Breast Cancer Res. Treat.* **136**, 795–804 (2012).

103. Bobal, P., Lastovickova, M. & Bobalova, J. The Role of ATRA, Natural Ligand of Retinoic Acid Receptors, on EMT-Related Proteins in Breast Cancer: Minireview. *Int. J. Mol. Sci.* **22**, (2021).

104. Nwagu, G. C., Bhattarai, S., Swahn, M., Ahmed, S. & Aneja, R. Prevalence and Mortality of Triple-Negative Breast Cancer in West Africa: Biologic and Sociocultural Factors. *JCO Glob Oncol* **7**, 1129–1140 (2021).

105. Wolff, A. C. *et al.* Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer. *Arch. Pathol. Lab. Med.* **147**, 993–1000 (2023).

106. Cardoso, F. *et al.* Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 1674 (2019).

107. Bergin, A. R. T. & Loi, S. Triple-negative breast cancer: recent treatment advances. *F1000Res.* **8**, (2019).

108. Borri, F. & Granaglia, A. Pathology of triple negative breast cancer. *Semin. Cancer Biol.* **72**, 136–145 (2021).

109. Brenton, J. D., Carey, L. A., Ahmed, A. A. & Caldas, C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J. Clin. Oncol.* **23**, 7350–7360 (2005).

110. Burstein, M. D. *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res.* **21**, 1688–1698 (2015).

111. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).

112. Lehmann, B. D. *et al.* Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* **11**, e0157368 (2016).

113. Lehmann, B. D. & Pietenpol, J. A. Identification and use of diagnostic markers in treatment strategies for triple-negative breast cancer subtypes. *J. Pathol.* **232**, 142–150 (2014).

114. Lehmann, B. D. *et al.* Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. *Nat. Commun.* **12**, 6276 (2021).

115. Mahmoud, R., Ordóñez-Morán, P. & Allegrucci, C. Challenges for Triple Negative Breast Cancer Treatment: Defeating Heterogeneity and Cancer Stemness. *Cancers* **14**, (2022).

116. Fusco, N. *et al.* Advancing the PD-L1 CPS test in metastatic TNBC: Insights from pathologists and findings from a nationwide survey. *Crit. Rev. Oncol. Hematol.* **190**, 104103 (2023).

117. Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E. & Gianni, L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat. Rev. Clin. Oncol.* **13**, 674–690 (2016).

118. Bianchini, G., De Angelis, C., Licata, L. & Gianni, L. Treatment landscape of triple-negative breast cancer - expanded options, evolving needs. *Nat. Rev. Clin. Oncol.* **19**, 91–113 (2022).

119. Gennari, A. *et al.* ESMO Clinical Practice Guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. *Ann. Oncol.* **32**, 1475–1495 (2021).

120. Valenza, C. *et al.* Evolving treatment landscape of immunotherapy in breast cancer: current issues and future perspectives. *Ther. Adv. Med. Oncol.* **15**, 17588359221146129 (2023).

121. Fife, B. T. & Bluestone, J. A. Control of peripheral T-cell tolerance and autoimmunity via the CTLA-4 and PD-1 pathways. *Immunol. Rev.* **224**, 166–182 (2008).

122. Fang, J. *et al.* Prognostic value of immune checkpoint molecules in breast cancer. *Biosci. Rep.* **40**, (2020).

123. Schmid, P. *et al.* Pembrolizumab for Early Triple-Negative Breast Cancer. *N. Engl. J. Med.* **382**, 810–821 (2020).

124. Criscitiello, C. *et al.* Immunotherapy in Breast Cancer Patients: A Focus on the Use of the Currently Available Diagnostic markers in Oncology. *Anticancer Agents Med. Chem.* **22**, 787–800 (2022).

125. Yamaguchi, H., Hsu, J.-M., Yang, W.-H. & Hung, M.-C. Mechanisms regulating PD-L1 expression in cancers and associated opportunities for novel small-molecule therapeutics. *Nat. Rev. Clin. Oncol.* **19**, 287–305 (2022).

126. Yang, T., Li, W., Huang, T. & Zhou, J. Immunotherapy Targeting PD-1/PD-L1 in Early-Stage Triple-Negative Breast Cancer. *J Pers Med* **13**, (2023).

127. Schmid, P. *et al.* Atezolizumab plus nab-paclitaxel as first-line treatment for unresectable, locally advanced or metastatic triple-negative breast cancer (IMpassion130): updated efficacy results from a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol.* **21**, 44–59 (2020).

128. Cortes, J. *et al.* Pembrolizumab plus chemotherapy versus placebo plus chemotherapy for previously untreated locally recurrent inoperable or metastatic triple-negative breast cancer (KEYNOTE-355): a randomised, placebo-controlled, double-blind, phase 3 clinical trial. *Lancet* **396**, 1817–1828 (2020).

129. Giugliano, F. *et al.* Harmonizing PD-L1 testing in metastatic triple negative breast cancer. *Expert Opin. Biol. Ther.* **22**, 345–348 (2022).

130. Noske, A. *et al.* Interassay and interobserver comparability study of four programmed death-ligand 1 (PD-L1) immunohistochemistry assays in triple-negative breast cancer. *Breast* **60**, 238–244 (2021).

131. Emens, L. A. *et al.* Atezolizumab and nab-Paclitaxel in Advanced Triple-Negative Breast Cancer: Diagnostic marker Evaluation of the IMpassion130 Study. *J. Natl. Cancer Inst.* **113**, 1005–1016 (2021).

132. Emens, L. A. *et al.* First-line atezolizumab plus nab-paclitaxel for unresectable, locally advanced, or metastatic triple-negative breast cancer: IMpassion130 final overall survival analysis. *Ann. Oncol.* **32**, 983–993 (2021).

133. Cieślik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.* **19**, 93–109 (2018).

134. Slovin, S. *et al.* Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. *Methods Mol. Biol.* **2284**, 343–365 (2021).

135. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and biological applications. *Mol. Aspects Med.* **59**, 36–46 (2018).

136. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 96 (2018).

137. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

138. Suo, S. *et al.* Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas. *Cell Rep.* **25**, 1436–1445.e3 (2018).

139. Velasco, S. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523–527 (2019).

140. Fischer, D. S. *et al.* Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* **37**, 461–468 (2019).

141. Huang, D. *et al.* Advances in single-cell RNA sequencing and its applications in cancer research. *J. Hematol. Oncol.* **16**, 98 (2023).

142. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).

143. Zhuang, X. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nat. Methods* **18**, 18–22 (2021).

144. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).

145. Wang, Y. *et al.* Spatial transcriptomics: Technologies, applications and experimental considerations. *Genomics* **115**, 110671 (2023).

146. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).

147. Tashireva, L. A. *et al.* Spatial Profile of Tumor Microenvironment in PD-L1-Negative and PD-L1-Positive Triple-Negative Breast Cancer. *Int. J. Mol. Sci.* **24**, (2023).

148. Bassiouni, R. *et al.* Spatial Transcriptomic Analysis of a Diverse Patient Cohort Reveals a Conserved Architecture in Triple-Negative Breast Cancer. *Cancer Res.* **83**, 34–48 (2023).

149. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

150. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

151. Marshall, J. L. *et al.* High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *iScience* **25**, 104097 (2022).

152. Kiessling, P. & Kuppe, C. Spatial multi-omics: novel tools to study the complexity of cardiovascular diseases. *Genome Med.* **16**, 14 (2024).

153. Jin, M.-Z. & Jin, W.-L. The updated landscape of tumor microenvironment and drug repurposing. *Signal Transduct Target Ther* **5**, 166 (2020).

154. Terekhanova, N. V. *et al.* Epigenetic regulation during cancer transitions across 11 tumour types. *Nature* **623**, 432–441 (2023).

155. Jiang, X. *et al.* Role of the tumor microenvironment in PD-L1/PD-1-mediated tumor immune escape. *Mol. Cancer* **18**, 10 (2019).

156. Brücher, B. L. D. M. & Jamall, I. S. Cell-cell communication in the tumor microenvironment, carcinogenesis, and anticancer treatment. *Cell. Physiol. Biochem.* **34**, 213–243 (2014).

157. Liu, C.-M. *et al.* Exosomes from the tumor microenvironment as reciprocal regulators that enhance prostate cancer progression. *Int. J. Urol.* **23**, 734–744 (2016).

158. Ableser, M. J., Penuela, S., Lee, J., Shao, Q. & Laird, D. W. Connexin43 reduces melanoma growth within a keratinocyte microenvironment and during tumorigenesis in vivo. *J. Biol. Chem.* **289**, 1592–1603 (2014).

159. Stoletov, K. *et al.* Role of connexins in metastatic breast cancer and melanoma brain colonization. *J. Cell Sci.* **126**, 904–913 (2013).

160. Kim, S.-H., Turnbull, J. & Guimond, S. Extracellular matrix and cell signalling: the

dynamic cooperation of integrin, proteoglycan and growth factor receptor. *J. Endocrinol.* **209**, 139–151 (2011).

161. Valdembri, D. & Serini, G. The roles of integrins in cancer. *Fac Rev* **10**, 45 (2021).

162. Spada, S., Tocci, A., Di Modugno, F. & Nisticò, P. Fibronectin as a multiregulatory molecule crucial in tumor matrisome: from structural and functional features to clinical practice in oncology. *J. Exp. Clin. Cancer Res.* **40**, 102 (2021).

163. Yue, J., Zhang, K. & Chen, J. Role of integrins in regulating proteases to mediate extracellular matrix remodeling. *Cancer Microenviron.* **5**, 275–283 (2012).

164. Psaila, B. & Lyden, D. The metastatic niche: adapting the foreign soil. *Nat. Rev. Cancer* **9**, 285–293 (2009).

165. Dominiak, A., Chełstowska, B., Olejarz, W. & Nowicka, G. Communication in the Cancer Microenvironment as a Target for Therapeutic Interventions. *Cancers* **12**, (2020).

166. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).

167. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-protein interaction detection: methods and analysis. *Int. J. Proteomics* **2014**, 147648 (2014).

168. Ghoshdastider, U. *et al.* Pan-Cancer Analysis of Ligand-Receptor Cross-talk in the Tumor Microenvironment. *Cancer Res.* **81**, 1802–1812 (2021).

169. Maffuid, K. & Cao, Y. Decoding the Complexity of Immune-Cancer Cell Interactions: Empowering the Future of Cancer Immunotherapy. *Cancers* **15**, (2023).

170. Hu, J. *et al.* Comprehensive analysis of ligand-receptor interactions in colon adenocarcinoma to identify of tumor microenvironment oxidative stress and prognosis model. *Curr. Med. Chem.* (2023) doi:10.2174/0929867331666230821092346.

171. Lin, H. *et al.* Delineation of colorectal cancer ligand-receptor interactions and their roles in the tumor microenvironment and prognosis. *J. Transl. Med.* **19**, 497 (2021).

172. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

173. Davies, M. A. & Samuels, Y. Analysis of the genome to personalize therapy for

melanoma. *Oncogene* **29**, 5545–5555 (2010).

174. Knudsen, E. S. *et al.* Pan-cancer molecular analysis of the RB tumor suppressor pathway. *Commun Biol* **3**, 158 (2020).

175. Warren, C. F. A., Wong-Brown, M. W. & Bowden, N. A. BCL-2 family isoforms in apoptosis and cancer. *Cell Death Dis.* **10**, 177 (2019).

176. Chen, X. *et al.* Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death Dis.* **13**, 974 (2022).

177. Robinson, N. J. & Schiemann, W. P. Telomerase in Cancer: Function, Regulation, and Clinical Translation. *Cancers* **14**, (2022).

178. Park, J.-I. *et al.* Telomerase modulates Wnt signalling by association with target gene chromatin. *Nature* **460**, 66–72 (2009).

179. Debnath, P., Huirem, R. S., Dutta, P. & Palchaudhuri, S. Epithelial-mesenchymal transition and its transcription factors. *Biosci. Rep.* **42**, (2022).

180. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).

181. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

182. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

183. Lu, C. & Verbridge, S. S. *Microfluidic Methods for Molecular Biology*. (Springer, 2016).

184. ENCODE. https://www.encodeproject.org/.

185. Ziller, M. J. *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518**, 355–359 (2015).

186. Rackham, O. J. L. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* **48**, 331–335 (2016).

187. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

188. Futschik, M. E. & Carlisle, B. Noise-robust soft clustering of gene expression time-course data. *J. Bioinform. Comput. Biol.* **3**, 965–988 (2005).

189. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–7 (2016).

190. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

191. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

192. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

193. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

194. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, lqaa078 (2020).

195. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

196. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

197. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14621–14626 (2010).

198. Bao, X. *et al.* A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. *Genome Biol.* **16**, 284 (2015).

199. Creators Harshil Patel1 Phil Ewels2 Alexander Peltzer3 Drew Behrens Gisela Gabernet4 Mingda Jin5 mashehu Maxime Garcia6 Show affiliations 1. The Francis Crick Institute 2. Science for Life Laboratory 3. Boehringer Ingelheim 4. @qbicsoftware 5. Zymo Research Corp. 6. @SciLifeLab | Karolinska Institutet. *nf-*

core/atacseq: nf-core/atacseq v1.2.1 - Iron Centipede. doi:10.5281/zenodo.3965985.

200. Ou, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).

201. *atac-seq-pipeline: ENCODE ATAC-seq pipeline*. (Github).

202. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

203. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

204. Chun, H. & Keleş, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 3–25 (2010).

205. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

206. Marletta, S. *et al.* Atlas of PD-L1 for Pathologists: Indications, Scores, Diagnostic Platforms and Reporting Systems. *J Pers Med* **12**, (2022).

207. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).

208. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

209. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

210. Ramilowski, J. A. *et al.* A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).

211. Panariello, F. *et al.* Cellular population dynamics shape the route to human pluripotency. *Nat. Commun.* **14**, 2829 (2023).

212. Xiong, Y. *et al.* A Comparison of mRNA Sequencing with Random Primed and 3'-Directed Libraries. *Sci. Rep.* **7**, 14626 (2017).

213. Sajjadi, E. *et al.* Immune microenvironment dynamics in breast cancer during pregnancy: impact of gestational age on tumor-infiltrating lymphocytes and prognosis.

Front. Oncol. **13**, 1116569 (2023).

214. Kumar, L. & E Futschik, M. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* **2**, 5–7 (2007).

215. Wardle, F. C. Master control: transcriptional regulation of mammalian Myod. *J. Muscle Res. Cell Motil.* **40**, 211–226 (2019).

216. Sun, Y. *et al.* Neurogenin promotes neurogenesis and inhibits glial differentiation by independent mechanisms. *Cell* **104**, 365–376 (2001).

217. Nayak, P., Colas, A., Mercola, M., Varghese, S. & Subramaniam, S. Temporal mechanisms of myogenic specification in human induced pluripotent stem cells. *Sci Adv* **7**, (2021).

218. Yoh, K. & Prywes, R. Pathway Regulation of p63, a Director of Epithelial Cell Fate. *Front. Endocrinol.* **6**, 51 (2015).

219. Zammit, P. S. Function of the myogenic regulatory factors Myf5, MyoD, Myogenin and MRF4 in skeletal muscle, satellite cells and regenerative myogenesis. *Semin. Cell Dev. Biol.* **72**, 19–32 (2017).

220. Dixit, R. *et al.* Neurog1 and Neurog2 control two waves of neuronal differentiation in the piriform cortex. *J. Neurosci.* **34**, 539–553 (2014).

221. Mulvaney, J. & Dabdoub, A. Atoh1, an essential transcription factor in neurogenesis and intestinal and inner ear development: function, regulation, and context dependency. *J. Assoc. Res. Otolaryngol.* **13**, 281–293 (2012).

222. Chen, Y. *et al.* The Role of Tbx20 in Cardiovascular Development and Function. *Front Cell Dev Biol* **9**, 638542 (2021).

223. Memic, F. *et al.* Ascl1 Is Required for the Development of Specific Neuronal Subtypes in the Enteric Nervous System. *J. Neurosci.* **36**, 4339–4350 (2016).

224. *atac-seq-pipeline: ENCODE ATAC-seq pipeline*. (Github).

225. George, R. M. & Firulli, A. B. Hand Factors in Cardiac Development. *Anat. Rec.* **302**, 101–107 (2019).

226. Desjardins, C. A. & Naya, F. J. The Function of the MEF2 Family of Transcription Factors in Cardiac Development, Cardiogenomics, and Direct Reprogramming. *J*

Cardiovasc Dev Dis **3**, (2016).

227. Del Águila, Á. *et al.* Olig2 defines a subset of neural stem cells that produce specific olfactory bulb interneuron subtypes in the subventricular zone of adult mice. *Development* **149**, (2022).

228. Stennard, F. A. *et al.* Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. *Dev. Biol.* **262**, 206–224 (2003).

229. Pfeiffer, M. J. *et al.* Cardiogenic programming of human pluripotent stem cells by dose-controlled activation of EOMES. *Nat. Commun.* **9**, 440 (2018).

230. Tutukova, S., Tarabykin, V. & Hernandez-Miranda, L. R. The Role of Neurod Genes in Brain Development, Function, and Disease. *Front. Mol. Neurosci.* **14**, 662774 (2021).

231. Nagaki, M. & Moriwaki, H. Transcription factor HNF and hepatocyte differentiation. *Hepatol. Res.* **38**, 961–969 (2008).

232. Mummenhoff, J., Houweling, A. C., Peters, T., Christoffels, V. M. & Rüther, U. Expression of Irx6 during mouse morphogenesis. *Mech. Dev.* **103**, 193–195 (2001).

233. Ajima, R., Sakakibara, Y., Sakurai-Yamatani, N., Muraoka, M. & Saga, Y. Formal proof of the requirement of MESP1 and MESP2 in mesoderm specification and their transcriptional control via specific enhancers in mice. *Development* **148**, (2021).

234. Dong, B. & Wu, Y. Epigenetic Regulation and Post-Translational Modifications of SNAI1 in Cancer Metastasis. *Int. J. Mol. Sci.* **22**, (2021).

235. Jin, C. *et al.* Crucial role of the transcription factors family activator protein 2 in cancer: current clue and views. *J. Transl. Med.* **21**, 371 (2023).

236. Villicaña, C., Cruz, G. & Zurita, M. The basal transcription machinery as a target for cancer therapy. *Cancer Cell Int.* **14**, 18 (2014).

237. Wong, J. J. W. & Selbo, P. K. Light-controlled elimination of PD-L1+ cells. *J. Photochem. Photobiol. B* **225**, 112355 (2021).

238. Lotfinejad, P. *et al.* PD-1/PD-L1 axis importance and tumor microenvironment immune cells. *Life Sci.* **259**, 118297 (2020).

239. Qiao, P. & Lu, Z.-R. Fibronectin in the Tumor Microenvironment. *Adv. Exp. Med. Biol.* **1245**, 85–96 (2020).

240. Shan, T., Liu, W. & Kuang, S. Fatty acid binding protein 4 expression marks a population of adipocyte progenitors in white and brown adipose tissues. *FASEB J.* **27**, 277–287 (2013).

241. Jacquelot, N. *et al.* Sustained Type I interferon signaling as a mechanism of resistance to PD-1 blockade. *Cell Res.* **29**, 846–861 (2019).

242. Zhang, S. *et al.* BGN May be a Potential Prognostic Diagnostic marker and Associated With Immune Cell Enrichment of Gastric Cancer. *Front. Genet.* **13**, 765569 (2022).

243. Ferri-Borgogno, S. *et al.* Spatial Transcriptomics Depict Ligand–Receptor Cross-talk Heterogeneity at the Tumor-Stroma Interface in Long-Term Ovarian Cancer Survivors. *Cancer Res.* **83**, 1503–1516 (2023).

244. Yang, L. *et al.* LRP1 modulates the microglial immune response via regulation of JNK and NF-κB signaling pathways. *J. Neuroinflammation* **13**, 304 (2016).

245. He, Y. *et al.* Silencing of LRP1 Exacerbates Inflammatory Response Via TLR4/NF-κB/MAPKs Signaling Pathways in APP/PS1 Transgenic Mice. *Mol. Neurobiol.* **57**, 3727–3743 (2020).

246. Liu, Z., Yu, X., Xu, L., Li, Y. & Zeng, C. Current insight into the regulation of PD-L1 in cancer. *Exp. Hematol. Oncol.* **11**, 44 (2022).

247. Antonangeli, F. *et al.* Regulation of PD-L1 Expression by NF-κB in Cancer. *Front. Immunol.* **11**, 584626 (2020).

248. Balsalobre, A. & Drouin, J. Pioneer factors as master regulators of the epigenome and cell fate. *Nat. Rev. Mol. Cell Biol.* **23**, 449–464 (2022).

249. Hiemisch, H., Schütz, G. & Kaestner, K. H. Transcriptional regulation in endoderm development: characterization of an enhancer controlling Hnf3g expression by transgenesis and targeted mutagenesis. *EMBO J.* **16**, 3995–4006 (1997).

250. Cai, A. *et al.* Myogenic differentiation of primary myoblasts and mesenchymal stromal cells under serum-free conditions on PCL-collagen I-nanoscaffolds. *BMC Biotechnol.*

18, 75 (2018).

251. Pawlowski, M. *et al.* Inducible and Deterministic Forward Programming of Human Pluripotent Stem Cells into Neurons, Skeletal Myocytes, and Oligodendrocytes. *Stem Cell Reports* **8**, 803–812 (2017).

252. Fu, J.-D. *et al.* Direct reprogramming of human fibroblasts toward a cardiomyocyte-like state. *Stem Cell Reports* **1**, 235–247 (2013).

253. Zhou, Y. *et al.* Single-Cell Transcriptomic Analyses of Cell Fate Transitions during Human Cardiac Reprogramming. *Cell Stem Cell* **25**, 149–164.e9 (2019).

254. Monjo, T., Koido, M., Nagasawa, S., Suzuki, Y. & Kamatani, Y. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Sci. Rep.* **12**, 4133 (2022).

255. Romanens, L. *et al.* Clonal expansion of intra-epithelial T cells in breast cancer revealed by spatial transcriptomics. *Int. J. Cancer* **153**, 1568–1578 (2023).

256. Wang, X. Q. *et al.* Spatial predictors of immunotherapy response in triple-negative breast cancer. *Nature* **621**, 868–876 (2023).

257. Emens, L. A. Breast Cancer Immunotherapy: Facts and Hopes. *Clin. Cancer Res.* **24**, 511–520 (2018).

258. Cortes, J. *et al.* Pembrolizumab plus Chemotherapy in Advanced Triple-Negative Breast Cancer. *N. Engl. J. Med.* **387**, 217–226 (2022).

259. Schmid, P. *et al.* Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. *N. Engl. J. Med.* **379**, 2108–2121 (2018).

260. Capobianco, E. Overview of triple negative breast cancer prognostic signatures in the context of data science-driven clinico-genomics research. *Annals of translational medicine* vol. 10 1300 (2022).

261. Perrichet, A., Ghiringhelli, F. & Rébé, C. Understanding Inflammasomes and PD-1/PD-L1 Crosstalk to Improve Cancer Treatment Efficiency. *Cancers* **12**, (2020).

262. Munir, S. *et al.* Inflammation induced PD-L1-specific T cells. *Cell Stress Chaperones* **3**, 319–327 (2019).

263. Inlay, M. A. *et al.* Ly6d marks the earliest stage of B-cell specification and identifies

the branchpoint between B-cell and T-cell development. *Genes Dev.* **23**, 2376–2381 (2009).

264. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* **7**, (2021).

265. AlHossiny, M. *et al.* Ly6E/K Signaling to TGFβ Promotes Breast Cancer Progression, Immune Escape, and Drug Resistance. *Cancer Res.* **76**, 3376–3386 (2016).

266. Sholl, L. M. Diagnostic markers of response to checkpoint inhibitors beyond PD-L1 in lung cancer. *Mod. Pathol.* **35**, 66–74 (2022).

267. Wang, Y. *et al.* NADPH Selective Depletion Nanomedicine-Mediated Radio-Immunometabolism Regulation for Strengthening Anti-PDL1 Therapy against TNBC. *Adv. Sci.* **10**, e2203788 (2023).

268. Nanda, R. *et al.* Pembrolizumab in Patients With Advanced Triple-Negative Breast Cancer: Phase Ib KEYNOTE-012 Study. *J. Clin. Oncol.* **34**, 2460–2467 (2016).