Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering, and Natural Sciences

of the

Ruprecht-Karls-University

Heidelberg, Germany

Presented by

M. Sc. Aryan Kamal

Born in: Tehran, 1992

Oral examination: 6th June, 2024

Identify and study essential transcription factors
in a celltype/condition-specific manner

Referees: Prof. Dr. Benedikt Brors

Dr. Georg Zeller

Abstract

Cells, despite having identical genetic information, show various functions and structures, which are determined by their unique gene expression profiles. Key elements such as chromatin, transcription factors (TF), and genes play an important role in regulating these profiles, together forming complex gene regulatory networks (GRNs). Understanding GRNs is crucial for interpreting how cellular identity is established and maintained, and how it can be disrupted in diseases. However, developing computational methods to reconstruct GRNs presents significant challenges, particularly in evaluation due to the absence of a definitive gold standard or ground truth. Relying only on experimentally validated connections for evaluation could lead to a bias toward well-established TFs. To address this, I have developed GRaNPA, a novel method that evaluates networks unbiasedly based on their ability to predict gene expression perturbations.

GRaNPA has a dual purpose: as a benchmarking tool, it compares different GRNs based on the hypothesis that true connections between TFs and genes can, to some extent, predict gene expression perturbations. In addition, it identifies key TFs essential for predicting these variations in gene expression. This functionality of GRaNPA is particularly beneficial for unraveling the underlying biological mechanisms.

I applied GRaNPA to assess an enhancer-based GRN (eGRN) that I constructed using GRaNIE for an iPSC-derived macrophage dataset. GRaNIE is a method that reconstructs GRNs based on co-variation across individuals, establishing connections both from TFs to enhancers and from enhancers to genes. For the macrophage eGRNs, I initially demonstrated their ability to accurately predict differential expression values between naive macrophages and those infected with Salmonella. Subsequently, by comparing their predictive accuracy with networks derived from AML and CD4+ T cells and demonstrating that these networks are predictive only for differential expression values specific to their cell type, I confirmed the cell type specificity of these eGRNs. Additionally, I showed that eGRNs from different cell types contain TFs with almost entirely distinct regulons, highlighting the various roles of TFs in different cell types. I also utilized GRaNPA's second function to identify important TFs under various conditions, such as infection with Salmonella, breast cancer, and tuberculosis disease. Beyond identifying well-known TFs like NFKB, which is influential in M1 macrophages exposed to INF-γ, I discovered lesser-known TFs like PURA, potentially playing a proinflammatory role in macrophages. Overall,

this demonstrates GRaNPA's utility in evaluating and identifying key TFs in various conditions, helping our understanding of the underlying biology.

Understanding and identifying important TFs could be beneficial for unraveling gene regulation in the context of cell fate determination. A few TFs, such as terminal selectors or master regulators, can drive specific cell lineages. Additionally, safeguard repressors actively repress alternative cell fates to induce and maintain cell identity. Following a computational screening for potential safeguard candidates, my collaborators and I identified Prox1 as a possible safeguard repressor for hepatocytes. I then explored Prox1's role as a potential tumor suppressor in a human hepatocyte carcinoma cell line and found that it restrains proliferation by targeting the key TF MYC. Further, I confirmed Prox1's role in hepatocyte fate induction at the single-cell level and utilized GRaNPA to identify its important targets, Pparg and Prrx1, which are key regulators in adipocyte and fibroblast fates. We also demonstrated that Prox1 can prevent transdifferentiation, suggesting that its absence might lead to an identity shift from hepatocyte carcinoma to cholangiocarcinoma. In conclusion, Prox1 promotes hepatocyte fate by targeting alternative fates and acts as a tumor suppressor in cancer.

In summary, understanding GRNs is key to interpreting the gene expression profiles of distinct cells. I developed GRaNPA, a method for assessing GRNs and identifying important TFs that explain specific variations, thereby enhancing our understanding of complex scenarios, including diseases.

Zusammenfassung

Zellen weisen trotz identischer genetischer Information unterschiedliche Funktionen und Strukturen auf, die durch ihre einzigartigen Genexpressionsprofile bestimmt werden. Schlüsselelemente wie Chromatin, Transkriptionsfaktoren (TF) und Gene spielen eine wichtige Rolle bei der Regulierung dieser Profile und bilden zusammen komplexe genregulatorische Netzwerke (GRN). Das Verständnis der GRN ist von entscheidender Bedeutung, um zu verstehen, wie die zelluläre Identität aufgebaut und erhalten wird und wie sie bei Krankheiten gestört werden kann. Die Entwicklung computergestützter Methoden zur Rekonstruktion von GRNs stellt jedoch eine große Herausforderung dar, insbesondere bei der Bewertung, da es keinen endgültigen Goldstandard oder eine Grundwahrheit gibt. Wenn man sich bei der Bewertung nur auf experimentell validierte Verbindungen verlässt, könnte dies zu einer Verzerrung in Richtung gut etablierter TFs führen. Um dieses Problem anzugehen, habe ich GRaNPA entwickelt, eine neuartige Methode, die Netzwerke unvoreingenommen auf der Grundlage ihrer Fähigkeit, Störungen der Genexpression vorherzusagen, bewertet.

GRaNPA verfolgt einen doppelten Zweck: Als Benchmarking-Tool vergleicht es verschiedene GRNs auf der Grundlage der Hypothese, dass echte Verbindungen zwischen TFs und Genen in gewissem Maße Störungen der Genexpression vorhersagen können. Darüber hinaus identifiziert es Schlüssel-TFs, die für die Vorhersage dieser Veränderungen in der Genexpression wesentlich sind. Diese Funktionalität von GRaNPA ist besonders nützlich, um die zugrunde liegenden biologischen Mechanismen zu entschlüsseln.

Ich habe GRaNPA angewandt, um ein Enhancer-basiertes GRN (eGRN) zu bewerten, das ich mit GRaNIE für einen iPSC-abgeleiteten Makrophagen-Datensatz erstellt habe. GRaNIE ist eine Methode, die GRNs auf der Grundlage von Kovariation über Individuen hinweg rekonstruiert und Verbindungen sowohl von TFs zu Enhancern als auch von Enhancern zu Genen herstellt. Für die Makrophagen-eGRNs habe ich zunächst gezeigt, dass sie in der Lage sind, unterschiedliche Expressionswerte zwischen naiven und mit Salmonellen infizierten Makrophagen genau vorherzusagen. Anschließend habe ich ihre Vorhersagegenauigkeit mit Netzwerken verglichen, die von AML- und CD4+ T-Zellen abgeleitet wurden, und gezeigt, dass diese Netzwerke nur unterschiedliche Expressionswerte vorhersagen können, die für ihren Zelltyp spezifisch sind. Außerdem habe ich gezeigt, dass eGRNs aus verschiedenen Zelltypen TFs mit fast völlig unterschiedlichen Regulonen enthalten, was die verschiedenen Rollen von TFs in verschiedenen Zelltypen verdeutlicht. Ich

nutzte auch die zweite Funktion von GRaNPA, um wichtige TFs unter verschiedenen Bedingungen zu identifizieren, z. B. bei einer Infektion mit Salmonellen, Brustkrebs und Tuberkuloseerkrankungen. Neben bekannten TFs wie NFKB, die in M1-Makrophagen, die INF-γ ausgesetzt sind, eine wichtige Rolle spielen, entdeckte ich weniger bekannte TFs wie PURA, die möglicherweise eine proinflammatorische Rolle in Makrophagen spielen. Insgesamt zeigt dies den Nutzen von GRaNPA bei der Bewertung und Identifizierung von Schlüssel-TFs unter verschiedenen Bedingungen und trägt daher zu unserem Verständnis der zugrunde liegenden Biologie bei.

Das Verständnis und die Identifizierung wichtiger TFs könnte dazu beitragen, die Genregulation im Zusammenhang mit der Bestimmung des Zellschicksals zu entschlüsseln. Einige TFs, wie terminale Selektoren oder Hauptregulatoren, können bestimmte Zelllinien steuern. Darüber hinaus unterdrücken Schutzrepressoren aktiv alternative Zellschicksale, um die Zellidentität zu induzieren und zu erhalten. Nach einem computergestützten Screening nach potenziellen Schutzkandidaten haben meine Mitarbeiter und ich Prox1 als möglichen Schutzrepressor für Hepatozyten identifiziert. Anschließend untersuchte ich die Rolle von Prox1 als potenzieller Tumorsuppressor in einer menschlichen Hepatozyten-Karzinom-Zelllinie und stellte fest, dass es die Proliferation durch Angreifen der Schlüssel-TF MYC einschränkt. Darüber hinaus bestätigte ich die Rolle von Prox1 bei der Induktion des Hepatozytenschicksals auf Einzelzellebene und nutzte GRaNPA, um seine wichtigen Zielstrukturen, Pparg und Prrx1, zu identifizieren, die Schlüsselregulatoren für das Adipozyten- und Fibroblastenschicksal sind. Wir haben auch gezeigt, dass Prox1 die Transdifferenzierung verhindern kann, was darauf hindeutet, dass sein Fehlen zu einer Identitätsverschiebung vom Hepatozytenkarzinom zum Cholangiokarzinom führen könnte. Zusammenfassend lässt sich sagen, dass Prox1 das Hepatozytenschicksal fördert, indem es auf alternative Schicksale abzielt, und als Tumorsuppressor bei Krebs wirkt.

Zusammenfassend lässt sich sagen, dass das Verständnis von GRNs der Schlüssel zur Interpretation der Genexpressionsprofile verschiedener Zellen ist. Ich habe GRaNPA entwickelt, eine Methode zur Bewertung von GRNs und zur Identifizierung wichtiger TFs, die spezifische Variationen erklären und dadurch unser Verständnis komplexer Szenarien, einschließlich Krankheiten, verbessern.

Acknowledgements

My PhD has been a rewarding and enjoyable journey, during which I met and enjoyed the company of many wonderful people. Among my friends outside the lab, I often felt like an outsider during their discussions about PhD struggles, as my experience was quite different, thanks to the wonderful environment in our lab. This unique experience is largely because of incredible Dr. Judith Zaugg. I have great memories from our very first interview and from each of our many meetings thereafter. In every interaction, I learned and grew, and I cannot imagine having a better supervisor. Thank you, Judith, for all your help and guidance.

I consider myself lucky to have lab members whom I can call friends. Thank you Christian, for all your support and everything you taught me. Thank you Anna and Max, for always backing me up and being there for me. And to Annique, Gwen, Nila, Ivan, Ignacio, Karin, Sophia, Victor, Charles, Sara, Mikael, Evi, Frosina, Neha, Daria, Brian, Guido, Jonas, Kristy, Josephine, Daniel, Olga, Dirk, Jupa, Rim, thank you, you all have been amazing lab mates.

I want to give a special thanks to my collaborators from DKFZ. Dr. Moritz Mall, your help and advice have been invaluable – thank you. Bryce, you've been much more than just a collaborator to me, making every aspect of our project so much easier. I deeply appreciate all your help.

I would also like to express my gratitude to all my TAC members: Prof. Dr. Benedikt Brors, Dr. Georg Zeller, and Dr. Oliver Stegle. Their support, advice, and guidance have been very important throughout my PhD journey. And a special thanks to Prof. Dr. Henrik Kaessmann for being part of my thesis defense committee.

Finally, I want to thank Setareh. You always being there for me and offering your support. Going through this journey without your companionship, kindness, and patience is unimaginable to me. Thank you so much.

Heidelberg, December 2023

Contents

List of Figures

Chapter 1

Introduction

The text in this chapter was originally written by me, and has been proofread by large language model-based tools.

1.1 Cell identity and differentiation

The progression from unicellular to multicellular life forms, characterized by cellular proliferation and differentiation, underlies the diversity of species, each characterized by unique biological attributes derived from their cellular contribution. This progression emphasizes the significance of studying the cellular organization and functionality in understanding the range of whole-organism diversity and interspecies functional variations (Mazzarello, 1999; Zeng, 2022). Over the past century, studies have revealed that cells within an organism can be grouped into distinct types based on their subcellular and molecular structure and function, which simplifies the investigation of cellular organization and function in complex organisms like mammals (Arendt, 2008).

Despite the diverse functions and structures of cells in complex organisms, every cell contains essentially the same genetic information. Furthermore, all of these cells originate from a single cell called the zygote. Through processes of cell division and differentiation, the zygote gives rise to the complex multicellular organism. while having the same genetic background, the differentiation process is regulated and maintained by complex networks of transcriptional and epigenetic mechanisms (Bernstein et al., 2007). Waddington's classic 'epigenetic landscape' model, illustrating cell differentiation as a marble rolling down a landscape into distinct channels, demonstrates how the undifferentiated, pluripotent state naturally restricts and guides cell fate towards various differentiated somatic states during development Figure 2.1.1a (Ladewig et al., 2013).

Recent studies demonstrate that induced pluripotent stem cells (iPSCs) can be created by reprogramming from differentiated cells, by overexpression of a few transcription factors, proving that the journey from pluripotent stem cells to differentiated somatic states is reversible (Takahashi

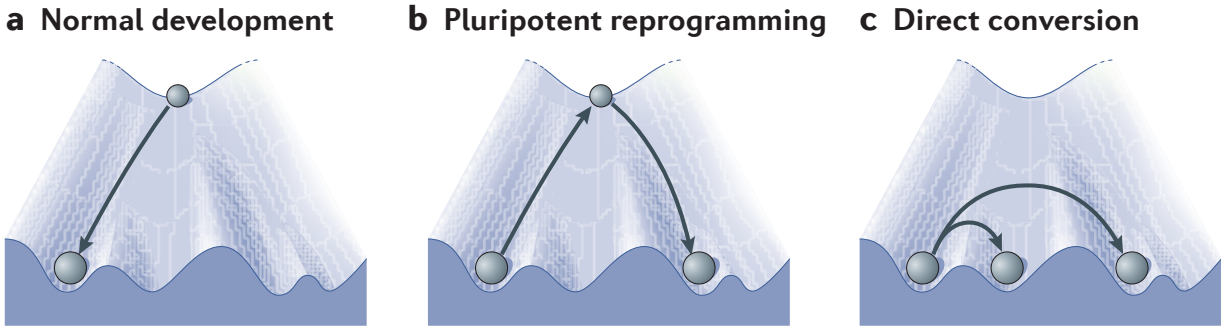

Figure 1.1.1. Waddington's model of the epigenetic landscape
(a) In normal development, a pluripotent cell, Shown at the hill's summit, rolls through distinct channels, determining its tissue-specific fate.
(b) Reprogramming a differentiated cell to pluripotency is shown as a marble ascending from the hill's base to the top, allowing for potential re-differentiation.
(c) Direct conversion is illustrated by a marble either leaping over a small hill to transform into a related tissue type or a larger hill for conversion into a different germ layer cell type.
This figure is adapted from (Ladewig et al., 2013). Reproduced with permission from Springer Nature with license number "5676490929904".

and Yamanaka, 2006), which establishes the pluripotent state as a key 'hub' in Waddington's landscape, raising questions about its essential role in transitioning between various cell fates or germ layers Figure 2.1.1b (Ladewig et al., 2013).

In 1987, Lassar and colleagues first demonstrated transdifferentiation by showing that the MyoD transcription factor could induce muscle-specific markers in fibroblasts, with later studies confirming similar fate conversions within the same germ layer (Davis et al., 1987). However, it wasn't until 2010 that Wernig and colleagues proved that cells from one germ layer could be converted into cell types of another, as seen in fibroblasts transforming into functional neurons using a cocktail of three transcription factors Figure 2.1.1c (Ladewig et al., 2013; Vierbuchen et al., 2010).

1.1.1 Cell identity regulation

The concept of transdifferentiation between different somatic cell fates, established over 30 years ago, began with experiments that used chromatin remodeling agents to induce distinct cellular identities from fibroblasts (Ladewig et al., 2013). The key role of transcription factors in regulating cell identity soon became clear, as these DNA-binding proteins were found to be essential for the induction and stabilization of gene regulatory networks (Kamimoto et al., 2023). As a result, certain transcription factors have been demonstrated to be important in reprogramming cells from mesodermal, ectodermal, and endodermal lineages into various somatic cell types (Ladewig et al., 2013). In addition, other epigenetic mechanisms play an important role in defining cell-type identity and function, through cellular reprogramming, as it involves a fundamental modifying of the epigenome, whether it's converting somatic cells to iPSCs or through transdifferentiation

(Basu and Tiwari, 2021). Furthermore, HOX genes, a group of transcription factors, that are repressed in undifferentiated pluripotent stem cells, have appeared as master regulators of cell fate determination during embryogenesis, while maintaining cellular identity throughout life (Steens and Klein, 2022). All together, cell fate transformation, controlled by transcription factors and epigenetic remodeling, highlights the complex process of differentiation and altering cell identity, from pluripotent states to specialized tissue functions.

1.2 Transcriptional regulation

The complex gene expression programs that define specific cell states are regulated by an extensive network of transcription factors, cofactors, and chromatin regulators (Lee and Young, 2013). These elements all together, control the gene expression of individual cells through direct genetic interactions and epigenetic effects. Transcription factors directly or indirectly influence RNA polymerase activity, while epigenetic modifications like DNA methylation and histone alterations change gene accessibility. In recent years, significant advances have been made in understanding mammalian gene regulatory elements and the transcriptional and chromatin regulators operating at these sites, particularly in embryonic stem cells and some of the differentiated cell types, it has been found that a small subset of transcription factors can be responsible for activating the gene expression program (Graf, 2011; Lee and Young, 2013; Ng and Surani, 2011; Orkin and Hochedlinger, 2011). The complexity of the transcriptional network is further regulated by the involvement of various non-coding RNAs (ncRNAs), including microRNAs, small nucleolar RNAs, long ncRNAs, circular RNAs, and enhancer RNAs, each contributing to a larger RNA communication network that controls transcription as well as protein translation (Casamassimi et al., 2017; Casamassimi and Ciccodicola, 2019). Notably, defects in transcriptional regulation have been observed early in diseases such as Huntington's Disease, as demonstrated in multiple cellular and animal models (Cha, 2007). However, transcriptional regulation works at two primary levels: one involves chromatin structure and its regulatory elements, and the second involves the transcription factors and the transcription mechanism (Lee and Young, 2013).

1.2.1 Chromatin regulators

Chromatin regulators are central in gene transcription and the establishment of cell identities, dynamically altering the chromatin state in response to developmental changes (Shu et al., 2012; Chen and Dent, 2014a). The nucleosome, a key structural unit of chromosomes, is mainly subjected to this complex regulation. ATP-dependent complexes, like the SWI/SNF family, reposition nucleosomes during gene activation, while an array of chromatin remodeling enzymes introduces chemical modifications to nucleosomes, facilitating transcriptional control (Lee and Young, 2013).

In contrast, gene repression features induce a different form of chromatin modifications, indicated by various repression mechanisms. For instance, chromatin regions marked by the Polycomb protein complex are set for later inactivation, while other types of repressors, found in silenced genomic regions, combine specific forms of nucleosome modification with DNA methylation (Lee and Young, 2013; Feng et al., 2010; Lejeune and Allshire, 2011). Ultimately, the broader chromatin landscape is shaped by a combination of mechanisms including transcriptional repression, DNA modifications, histone variants incorporation, and non-coding RNA-mediated regulation, all together playing a crucial role in cellular differentiation and fate determination (Chen and Dent, 2014b).

Furthermore, the process of cellular differentiation illustrates the epigenetic nature of chromatin structure, in which a transition from a more open to a compact chromatin state accrues, as observed in pluripotent towards differentiated cells. The epigenetic remodeling is not just a passive reflection of gene expression changes, but actively influences the key developmental stages and cell fate decisions, as evidenced in recent emerging research (Chen and Dent, 2014b). This perspective highlights the dynamic and multifunctional role of chromatin in determining cellular identity and fate.

1.2.2 Transcription factors

Transcription factors (TFs), the most extensive protein family in mammalians, play an important role in regulating important cellular processes like development and differentiation. Each cell type selectively expresses a subset of TFs, chosen from the nearly 1,600 TFs present in the human genome, which control the gene expression program of the cell. These TFs bind specifically to particular DNA regions, such as enhancers and promoter (proximal) sites Figure 1.2.1, coordinating gene expression through those DNA-binding domains. The activity of TFs after binding to the DNA plays a key role in several cellular processes. For example, induction of Yamanaka factors (Pool of four main TFs) in reprogramming fibroblasts to iPSC, highlighting their importance in both intrinsic and extrinsic cellular functions and responses (Oksuz et al., 2023; Weidemüller et al., 2021; Takahashi and Yamanaka, 2006).

The detection and functional assessment of TFs is challenging due to their low abundance in cells, and in addition their reliance on post-translational modifications and interactions with other proteins. Although techniques such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) offer genome-wide insights into TF binding activity, researchers often struggle to differentiate between functional and non-specific binding events. This complexity is compounded by the fact that the presence of a TF on a DNA region does not guarantee its functional activity, while it depends on its binding to chromatin at key regulatory sites such as enhancers and promoters (Weidemüller et al., 2021).

In the meantime, computational approaches have also become increasingly essential in identifying TF binding sites and understanding transcriptional regulation. By analyzing different omics data such as gene expression, chromatin accessibility, ChIP-seq, etc., various computational methods help us to model the relationships between gene expression and TF binding motifs, particularly in promoter and enhancer regions. The advanced computational techniques have significantly improved the accuracy of predicting transcription factor binding sites and activity, offering a deeper understanding of the transcriptional regulatory networks (Wang et al., 2015). The integration of experimental and computational methodologies is thus vital in unraveling the complex mechanisms by which TFs regulate gene expression and cellular behavior.

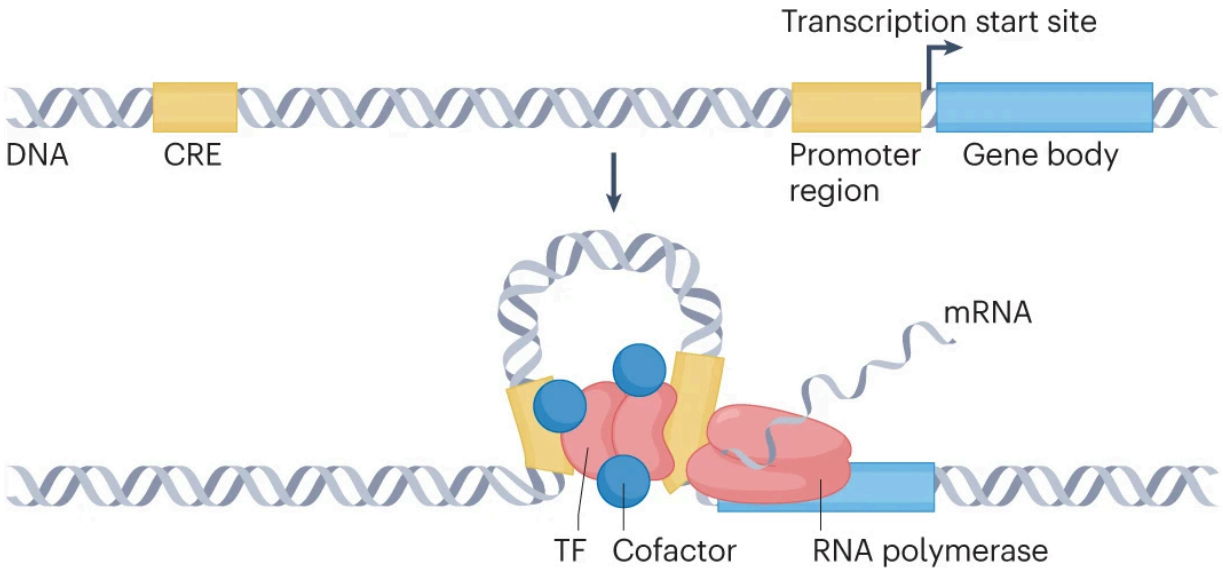

Figure 1.2.1. The role of transcription factors and cis-regulatory elements in gene regulation
Transcription factors play a central role by attaching to promoter regions and cis-regulatory elements (CREs) in gene regulation. This binding action displaces nucleosomes, thereby rendering the transcription start site accessible. The collaborative interplay among TFs, cofactors, and other various proteins is crucial for recruiting and stabilizing the RNA polymerase complex at the transcription start site, which is responsible for transcribing the gene body's DNA into mRNA.
This figure is adapted from (Badia-I-Mompel et al., 2023), reproduced with permission from Springer Nature with license number "5675430281503".

1.2.3 Enhancers

The concept of "enhancers" was originally identified in studies on Simian virus 40, defining them as short DNA elements that control gene expression from a distance (Panigrahi and O'Malley, 2021). Enhancers are essential cis-regulatory elements throughout DNA, carrying epigenetic information through specific histone modifications. Their function is mainly defined by the binding of certain transcription factors that interpret developmental and differentiation regulatory signals in a highly context-specific manner (Panigrahi and O'Malley, 2021; Zaugg et al., 2022). Their disruption, through mechanisms such as chromosomal rearrangements, genetic variations, or epigenetic modifications, is increasingly implicated in a variety of diseases. Dysregulation of enhancers can lead to unusual gene expression, highlighting the significance of enhancers in disease causation, particularly in enhanceropathies. In addition, enhancers are important for maintaining cell identity and lineage determination, as their cell-type-specific patterns play a key role in cellular development. Alterations of enhancer signatures can lead to incorrect lineage formation and various diseases, again emphasizing the importance of understanding enhancer regulation in both healthy and diseased condition management (Claringbould and Zaugg, 2021; Zaugg et al., 2022; Panigrahi and O'Malley, 2021).

Enhancers function is primarily recruiting TFs and forming a chromatin loop with gene promoters, a process which is critical for the spatial organization of the genome before gene

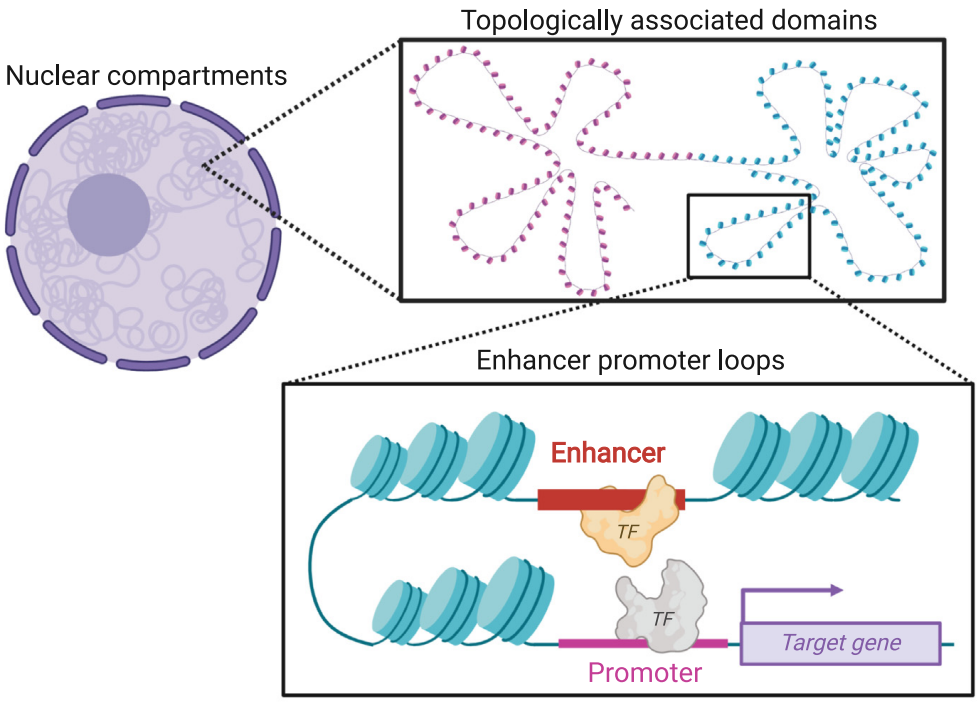

Figure 1.2.2. Illustration of enhancer-promoter loop formation
Chromatin forms a loop that bridges the enhancer and promoter regions, enabling the binding of TFs, which in turn triggers the expression of the associated target gene.
This figure is adapted from (Claringbould and Zaugg, 2021), under the Creative Commons Attribution License.

expression. This organization includes topologically associating domains (TADs) and smaller enhancer–promoter loops, which are largely influenced by the cohesin complex Figure 1.2.2 (Claringbould and Zaugg, 2021; Zaugg et al., 2022).

Recent advances in CRISPR technology have been important in investigating enhancer activity within their natural genomic context. Using this technique, by selectively activating or repressing enhancer elements, researchers could observe direct effects of enhancer dysregulation on gene expression. Additionally, the biochemical properties of enhancers, characterized by specific chromatin modifications such as H3K4me1 and H3K27ac, helped researchers in predicting endogenous enhancers and study them throughout the genome. Notably, active enhancers often show bidirectional capped RNAs, indicating their ability to initiate transcription in both directions (Claringbould and Zaugg, 2021).

Furthermore, super-enhancers, recognized as clusters of highly active enhancers, have been identified by having a shared function in cell-type-specific gene regulation. However, it's still unclear whether each enhancer within these clusters contributes uniquely or if their functions largely overlap. These clusters of enhancers are notably vulnerable to changes in chromatin structure, like decreased levels of cohesin, which can impact how effectively they operate. (Claringbould and Zaugg, 2021). Therefore, novel technologies as well, as further computational developments are necessary to fully unravel the complexity of gene regulation through different regulatory elements such as enhancers.

One of the recent methodologies to investigate gene regulation at a complex multi-cellular resolution is gene regulatory networks. Given its nature, gene regulatory networks are crucial for systematically explaining the role of chromatin remodeling, transcription factors, and promoter-enhancer interactions in different forms of gene regulation. These networks reveal how all these elements interact to control gene expression, providing essential insights into their functional mechanisms and implications in cellular processes and diseases.

1.3 Gene regulatory networks

Gene Regulatory Networks (GRNs) are complex computational models that represent the regulation of gene expression as networks or graphs. These models incorporate diverse elements of gene regulation such as transcription factors, splicing factors, long non-coding RNAs, microRNAs, and metabolites, along with genes which are often represented as nodes in the graph. While edges usually describe regulatory interactions between those elements (Badia-I-Mompel et al., 2023). The analysis of GRNs is fundamental in understanding the establishment and maintenance of cellular identity, providing crucial insights into cell differentiation, cell fate determination (Su et al., 2022), as well as preventing the diseased state (Claringbould and Zaugg, 2021; Badia-I-Mompel et al., 2023).

The evolution of GRNs research has transitioned from primarily curated or experimentally derived methods to more advanced computational approaches in recent years. Historically, the study of GRNs can be traced back to multiple pioneering works, for instance, the analysis of the bacterial lactose operon in the 1960s, where the initial networks were often assembled from experimentally validated regulatory events. However, the advent of high-throughput data marked a significant improvement in GRN reconstruction and analysis. Computational methods began to play an essential role since then, allowing for the creation of GRNs that are more contextually relevant to a specific biological question, as opposed to the generalist nature of database-extracted GRNs (Badia-I-Mompel et al., 2023).

One of the simplest forms of GRN, relying only on transcriptomics data, is known as Weighted Gene Co-expression Network Analysis (WGCNA) (Badia-I-Mompel et al., 2023; Langfelder and Horvath, 2008) Figure 1.3.1a. This approach focuses on the variations between gene expression by performing pairwise correlations across the whole transcriptome. WGCNA identifies clusters of co-expressed genes and tries to summarize these clusters using metrics such as the module eigengene or intramodular hub genes. Despite its utility in identifying gene modules, WGCNA's major limitation is the lack of causal regulatory connections, leading to potentially high false positive discoveries and therefore, limited interpretability (Langfelder and Horvath, 2008; Badia-I-Mompel et al., 2023).

To overcome these limitations, more advanced methods such as GENIE3 have been developed (Huynh-Thu et al., 2010a). GENIE3 differentiates between TFs and target genes, using tree-based ensemble methods such as Random Forests to predict the expression pattern of each gene Figure 1.3.1a. This method significantly reduces the number of interactions to consider and introduces directed connections into the network, thereby suggesting putative causal relationships (Huynh-Thu et al., 2010a). GENIE3's versatility allows it to handle complex regulatory scenarios, including combinatorial and non-linear interactions, making it a robust tool for GRN inference from transcriptomic data (Badia-I-Mompel et al., 2023; Huynh-Thu et al., 2010a).

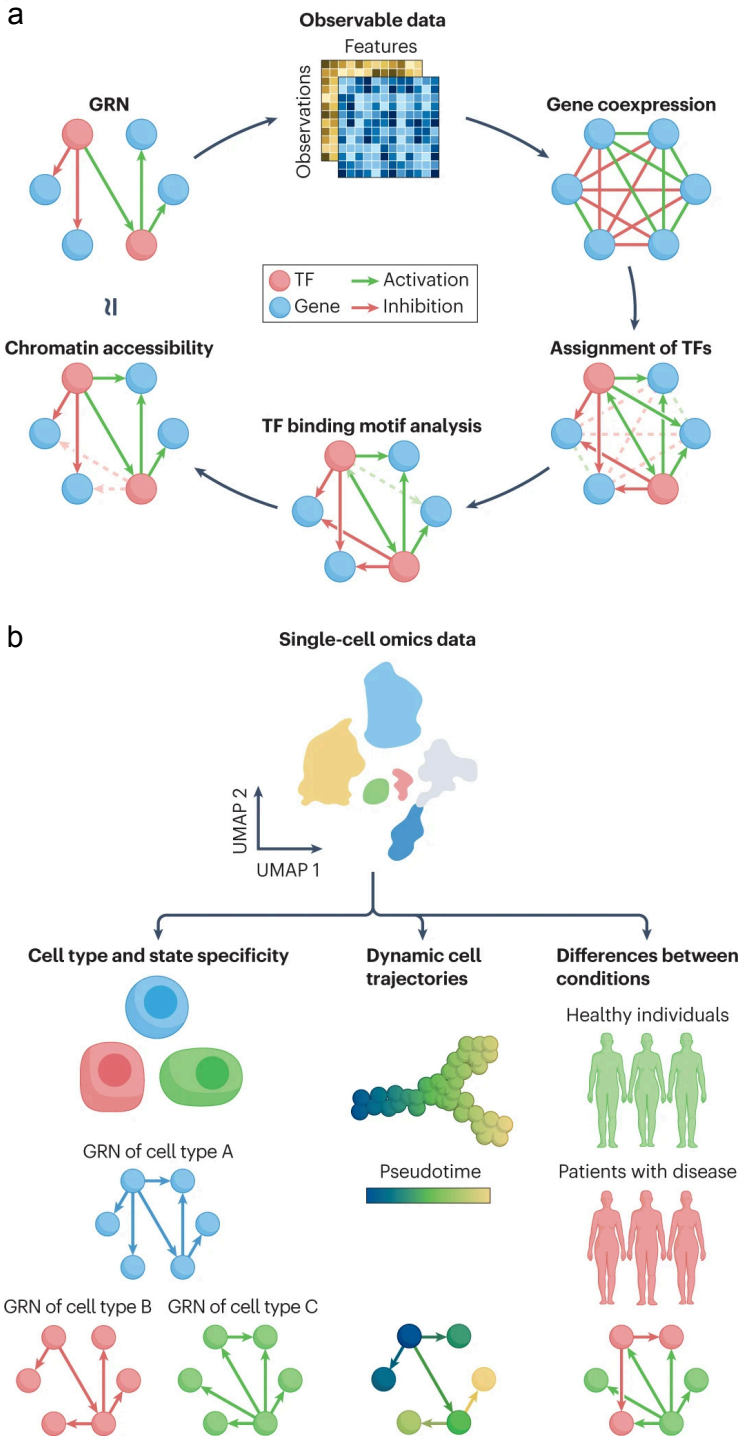

Figure 1.3.1. Gene regulatory network analysis
(a) Inference of GRNs from omics data, showing TFs and target genes as nodes with edges representing regulatory interactions. Enhanced accuracy is achieved by incorporating TF binding predictions and chromatin accessibility. (b) The application of GRNs derived from single-cell omics data to unravel cell type specificity, dynamic cell state changes, and condition-dependent variations.
This figure is adapted from (Badia-I-Mompel et al., 2023). Reproduced with permission from Springer Nature with license number "5675430281503".

However, GRNs based only on transcriptomics data miss fundamental information about the role of DNA regulatory regions such as enhancers. This gap is addressed by incorporating data

from assays like ChIP-seq and ATAC-seq, which measure TF binding and chromatin accessibility, respectively. By integrating this data into GRN reconstruction, the resulting network accounts for the interactions between TFs and regulatory elements like enhancers and promoters (Badia-I-Mompel et al., 2023).

Enhancer-based networks, which include RNA and chromatin accessibility data, offer a more comprehensive understanding of gene regulation. These networks are particularly useful as they consider the role of accessible chromatin regions, or CREs, in gene regulation (Badia-I-Mompel et al., 2023). For instance, ANANSE, a tool developed for network-based analysis, utilizes enhancer-encoded regulatory information from ATACseq data to identify key transcription factors in cell fate determination. By predicting genome-wide binding profiles of TFs in different cell types using enhancer activity, ANANSE reconstructs cell type-specific gene regulatory networks, improving the understanding of transcription factor-mediated regulatory mechanisms in various cell types (Xu et al., 2021).

Furthermore, the progression from bulk to single-cell transcriptome profiling has significantly improved our understanding of GRNs (Badia-I-Mompel et al., 2023). Unlike bulk data, single-cell RNA-sequencing (scRNA-seq) doesn't average biological signals over all cells in a tissue sample but rather differentiates signals from the particular cell types, allowing for a more precise view of cellular heterogeneity and lineage differentiation (Pratapa et al., 2020) Figure 1.3.1b. However, single-cell data come with their own challenges, such as high variability in sequencing depth, and sparsity due to dropouts. Despite these challenges, single-cell data are particularly promising for GRN reconstruction, offering high-resolution insights into cellular differentiation and transitions between different cell states (Pratapa et al., 2020) Figure 1.3.1b.

SCENIC (Single-Cell Regulatory Network Inference and Clustering) is a significant advancement in this field (Aibar et al., 2017). It leverages scRNA-seq data to simultaneously reconstruct gene regulatory networks and identify cell states. SCENIC's approach, which combines coexpression networks with TF motif discovery enables researchers to get a deeper understanding of the mechanisms driving cellular heterogeneity in different biological processes such as development (Aibar et al., 2017). However, this method has also some limitations. For example, it is not able to identify the exact cis-regulatory elements targeted by a TF. To solve this, SCENIC+ has been built upon the foundation of SCENIC by integrating chromatin accessibility data with gene expression profiles from individual cells. This method improves the accuracy of TF binding site (TFBS) predictions and allows for the inference of enhancer-driven GRNs (Bravo González-Blas et al., 2023). SCENIC+ identifies genomic enhancers and their candidate upstream TFs, and then linking these enhancers to their potential target genes. This approach facilitates a deeper exploration of gene regulation dynamics, for example along differentiation trajectories and the impact of TF perturbations on cell state conversion (Bravo González-Blas et al., 2023) Figure 1.3.1b.

Moreover, CellOracle (Kamimoto et al., 2023) further extends the capabilities of GRN analysis. Using GRNs inferred from single-cell multi-omics data, it performs in silico TF perturbations,

simulating changes in cell identity using only unperturbed wild-type data. This machine-learning-based approach, by predicting perturbations in silico, can improve and simplify our understanding of development and differentiation processes (Kamimoto et al., 2023) Figure 1.3.1b.

1.3.1 Benchmarking gene regulatory networks

GRNs have become an important component in many biological research studies. With the recent development of numerous tools focused on GRNs, there arises a need for their systematic assessment, a process known as benchmarking, to validate their accuracy and efficiency. The benchmarking of GRNs, particularly those derived from single-cell and multi-omics data, presents significant challenges. One primary obstacle is the absence of universally accepted ground-truth datasets for assessing algorithm accuracy (Pratapa et al., 2020; Aibar et al., 2017; Bravo González-Blas et al., 2023; Sanguinetti and Huynh-Thu, 2018; Kartha et al., 2022; Kamal et al., 2023; Xu et al., 2021; Badia-I-Mompel et al., 2023). This is further complicated by the inherent characteristics of single-cell data, such as cellular heterogeneity, variations in sequencing depth, and high sparsity. These challenges highlight the difficulty in evaluating the accuracy of GRN inference methods and their comparison through benchmarking, particularly when adapted to single-cell data (Pratapa et al., 2020).

In order to address the above-mentioned challenges in GRN comparisons, one approach is to use synthetic networks to create an in silico network, serving as benchmarks for GRN reconstruction algorithms (Pratapa et al., 2020). While this strategy provides a controlled environment for evaluation, it may not fully capture the complexity of natural gene regulatory mechanisms found in biological systems (Badia-I-Mompel et al., 2023). In addition, when using in silico network, the choice of evaluation metrics for GRN inference algorithms is also crucial. Naive approaches like threshold-based accuracy metrics can be misleading due to the sparsity of GRNs. However, more sophisticated methods, such as precision-recall curves, provide a better-detailed view of an algorithm's performance by considering the trade-off between true positive identifications and false positive rates (Sanguinetti and Huynh-Thu, 2018).

Perturbation experiments offer another way for GRN evaluation, where the connections can be determined based on the experimental design. However, these experiments are also not without limitations. In addition to high costs, sometimes they produce unexpected results and may be influenced by compensatory mechanisms in a real biological system. Moreover, the temporal aspect of gene regulation adds another layer of complexity, as experiments conducted at different time frames can generate inconsistent results due to the dynamic nature of gene expression and regulatory interactions (Badia-I-Mompel et al., 2023).

Notably, the inclusion of chromatin accessibility data (ATAC-seq) in GRN inference attempts to fine-tune TF-gene connections by considering TF binding to open regions of DNA and incorporating cis-regulatory elements. However, this approach also has its limitations. Just because a TF binds to a gene doesn't automatically mean that the TF regulates that gene. TFs often bind

stochastically to open regions of DNA and require cooperation with other molecules to effectively regulate transcription. (Badia-I-Mompel et al., 2023).

In summary, in the absence of a definitive gold standard for GRN evaluation and validation, the field remains relying upon a collection of 'silver standards'. These standards measures contain various methodologies, each contributing to the overall understanding and evaluation of GRN inference methods (Badia-I-Mompel et al., 2023).

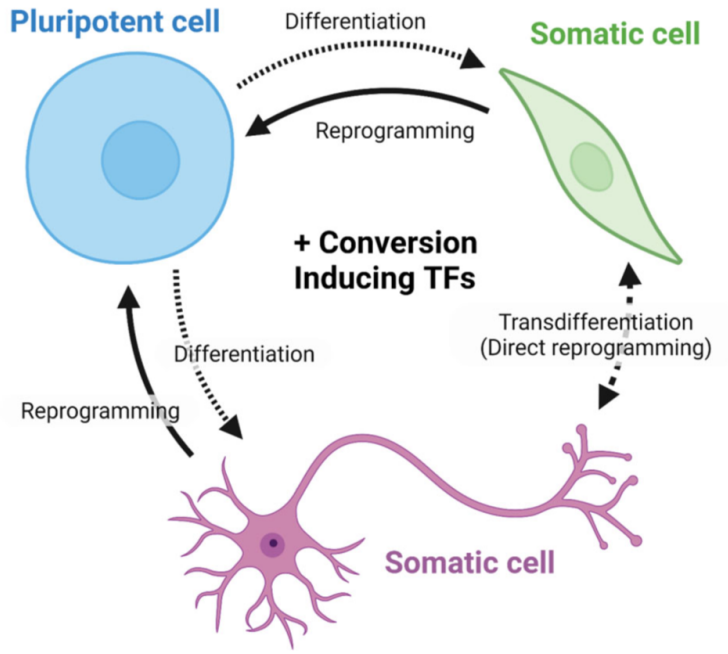

Figure 1.4.1. Cell reprogramming
Visual Representation of Cell Fate Dynamics: The illustration demonstrates the transformation of pluripotent stem and progenitor cells into specialized cell types, directed by TFs. Additionally, it highlights the reprogramming of somatic cells to pluripotent states and their direct conversion to different somatic cell types via specific TFs.
This Figure was adopted from (Larcombe et al., 2022), under the Creative Commons Attribution License.

1.4 Cell identity regulation by transcription factors

TFs (explained in 1.1.1) are one of the main players in the transcriptional regulation of cell identity (1.2.2), while their expression along with their binding to DNA regulatory regions is closely tied to cell types and external conditions. TFs are indeed the key to determining whether a cell maintains its current state or undergoes a transformation at each certain time point. On the other hand, by revealing the complex interplay of TFs and target genes in regulating different cell states and conditions, GRNs can help us to identify and understand the biological processes, and describe their changes under diverse circumstances (Larcombe et al., 2022; Kamimoto et al., 2023).

Indirect reprogramming is one of the very noteworthy examples demonstrating how cell identity can be altered by modifying TF expression. A prominent example is the use of Yamanaka factors (OCT4, SOX2, KLF4, and C-MYC), which initiate changes in previously inaccessible DNA regions (Takahashi and Yamanaka, 2006), and facilitate a huge gene expression profile alteration towards less differentiated cell state. This process effectively lowers the barriers to cell fate changes, illustrating the power of TFs in reprogramming cellular identity (Larcombe et al., 2022) Figure 1.4.1.

Direct reprogramming or transdifferentiation applies to converting one differentiated cell type directly into another specialized form of cell. For instance, fibroblasts can be transformed into neurons using TFs like ASCL1, BRN2, and MYT1L, and into cardiomyocytes with GATA4, MEF2C, and TBX5 (Davis et al., 1987; Wapinski et al., 2013). These examples highlight the role of

TFs in initiating significant shifts in a cell's gene expression profile and chromatin state, thereby rewriting its identity (Larcombe et al., 2022) Figure 1.4.1.

Safeguard mechanisms are important in maintaining an established cell identity. Myt1l, for example, a pan neuron-specific TF, actively represses various non-neuronal programs, thereby stabilizing the neuronal identity. Its role in silencing multiple lineage programs except the neuronal fate illustrates how certain TFs function as key stabilizers of cellular identity, actively maintaining the cell's state against potential reprogramming triggers, or external stimuli (Mall et al., 2017). In addition, safeguard repressor factors, by preserving cell identity play an essential role in preventing cells from reverting to a more versatile, stem-like state, which is important to avoid tumor formation in various tissues (Mall et al., 2017).

In conclusion, the role of TFs in cell identity, especially when studied in the context of GRNs, is central to understanding how different cell types maintain or change their state. This insight is vital in developmental biology and regenerative medicine, where altering cell identity could not only enable us to understand the disease mechanism and progression but also offer new treatment strategies. Besides, investigating TF dynamics within GRNs provides a comprehensive view of cellular behavior and potential therapeutic applications.

1.5 Cellular plasticity and cancer

Initially, the understanding of cancer biology was framed around six fundamental hallmarks (Hanahan and Weinberg, 2011). These included sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. Recent research, however, has broadened this perspective, introducing additional hallmarks and underlying factors that paint a more comprehensive picture of cancer's complexity (Hanahan and Weinberg, 2011).

Expanding on the recognition of additional hallmarks in cancer research, the field now includes important insights like the reprogramming of cellular energy metabolism. This reprogramming is not just a byproduct but a strategic alteration by cancer cells to support their rapid growth and survival, indicating a shift in cellular priorities from normal function to malignancy. Alongside, the evasion of immune destruction emerges as a critical hallmark, highlighting cancer cells' ability to cleverly avoid detection and destruction by the host's immune system. These added hallmarks, together with the underlying factors of genome instability and inflammation, underscore the adaptive and complex nature of cancer, illustrating how it manipulates basic cellular and molecular processes for its progression (Hanahan, 2022).

Cell fate plasticity in cancer is an intricate and essential process that enables cancer cells to adopt various cellular identities, adapting to different environments and therapeutic pressures. This plasticity is not just limited to de-differentiation, where cancer cells revert to a more stem-like undifferentiated state, but also includes the ability of cells to maintain themselves in a partially differentiated state or to undergo transdifferentiation, shifting entirely into different cell lineages. For instance, certain cancer types form when cells originally destined for a specific lineage abnormally change their developmental trajectory, adopting characteristics of a completely different tissue type, and therefore inducing the malignancy in that tissue. This dynamic shift in cellular identity is central to cancer's progression, metastasis, and its often formidable resistance to conventional treatments (Hanahan, 2022).

The complex mechanisms driving cell fate plasticity in cancer are rooted in a complex interplay of genetic and epigenetic factors. Transcription factors play an important role in this process, acting as key upstream regulators that can either trigger or suppress different cell differentiation pathways. In colorectal cancer, for example, the loss of transcription factors like HOXA5 and SMAD4, which are crucial for maintaining a differentiated state, results in cells gaining stem-like properties (Ordóñez-Morán et al., 2015). This increase in 'stemness' not only enhances the metastatic potential of cancer cells but also contributes significantly to the heterogeneity within tumors. This heterogeneity, in turn, poses a substantial challenge to effective treatment, as it enables the cancer to adapt to and resist various drugs and therapeutic strategies. Therefore, understanding different regulatory mechanisms exposed by distinct families of TFs opens the door not only to understanding and possibly preventing the cancer formation and progression

but also to developing targeted therapies that can specifically inhibit or reverse these plasticity processes, potentially leading to more effective and personalized approaches in cancer treatment (Hanahan, 2022).

In summary, the evolving landscape of cancer hallmarks, especially with the addition of cell fate plasticity, underscores the complexity of tumor development and progression, while the role of transcription factors in these processes is important, providing insights into new therapeutic options that target these pathways for more effective cancer treatment.

1.6 Aims of study

The main goal of this study is to understand and identify essential TFs in a context- and cell type-specific manner. This involves determining high-confidence regulons within a reconstructed GRN. Therefore, first aim of this thesis is to propose an unbiased method for evaluating GRNs, providing insights into the importance of TFs in specific contexts or cell types.

Second objective of this thesis is to develop and apply a framework for inferring enhancer-based and cell-type specific GRNs, particularly to study cellular responses in macrophages. This aim extends beyond only evaluating the GRNs, rather includes: one. benchmarking the networks, and two. identifying important TFs to explain the variations in specific perturbations.

In line with the previous goals, this research also aimed to investigate a specific group of TFs known as Safeguard Repressors. This objective involves: One. Evaluating potential TFs acting as safeguard repressors in hepatocytes differentiation, (identified previously through computational screening), and two. Investigating their roles in Hepatocellular carcinoma (HCC), as well as in hepatic injury, and reprogramming.

Chapter 2

GRaNPA - Gene Regulatory Network Performance Analysis

In this chapter, I present GRaNPA, a method I developed for unbiased evaluation of gene regulatory networks, focusing on identifying important transcription factors that explain gene expression variations in specific cell types or contexts. This method was conceptualized and refined under supervision from Dr. Judith Zaugg. The content in this chapter is based on my original work, proofread by large language model-based tools, and has been adapted from:

Aryan Kamal*, Christian Arnold**, Annique Claringbould, Rim Moussa, Nila H Servaas, Maksim Kholmatov, Neha Daga, Daria Nogina, Sophia Mueller-Dott, Armando Reyes-Palomares, Giovanni Palla, Olga Sigalova, Daria Bunina, Caroline Pabst, Judith B Zaugg. GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks.* Mol Syst Biol. (2023) 19: e11627. https://doi.org/10.15252/msb.202311627

2.1 Introduction

In the Section 1.3.1, I explored the considerable challenge of benchmarking gene regulatory networks (GRNs) due to the absence of a "true" gold standard. This gap is particularly notable when relying on in silico networks based on synthetic models, which may not accurately reflect true biological processes. Additionally, utilizing small, biologically-derived networks, identified through experimental methods, tends to bias the evaluation towards well-established TFs (Weidemüller et al., 2021). Such approaches fail to effectively validate the roles of lesser-known TFs.

To address this, I introduce GRaNPA, a novel and unbiased methodology designed to evaluate the biological relevance of a GRN Figure 2.1.1. The fundamental principle behind GRaNPA is based on the hypothesis that a network that accurately captures actual biological factors—specifically, the true connections between TFs and their target genes—should be capable of predicting variations in gene expression to a certain extent. In simpler terms, the topology of the network itself

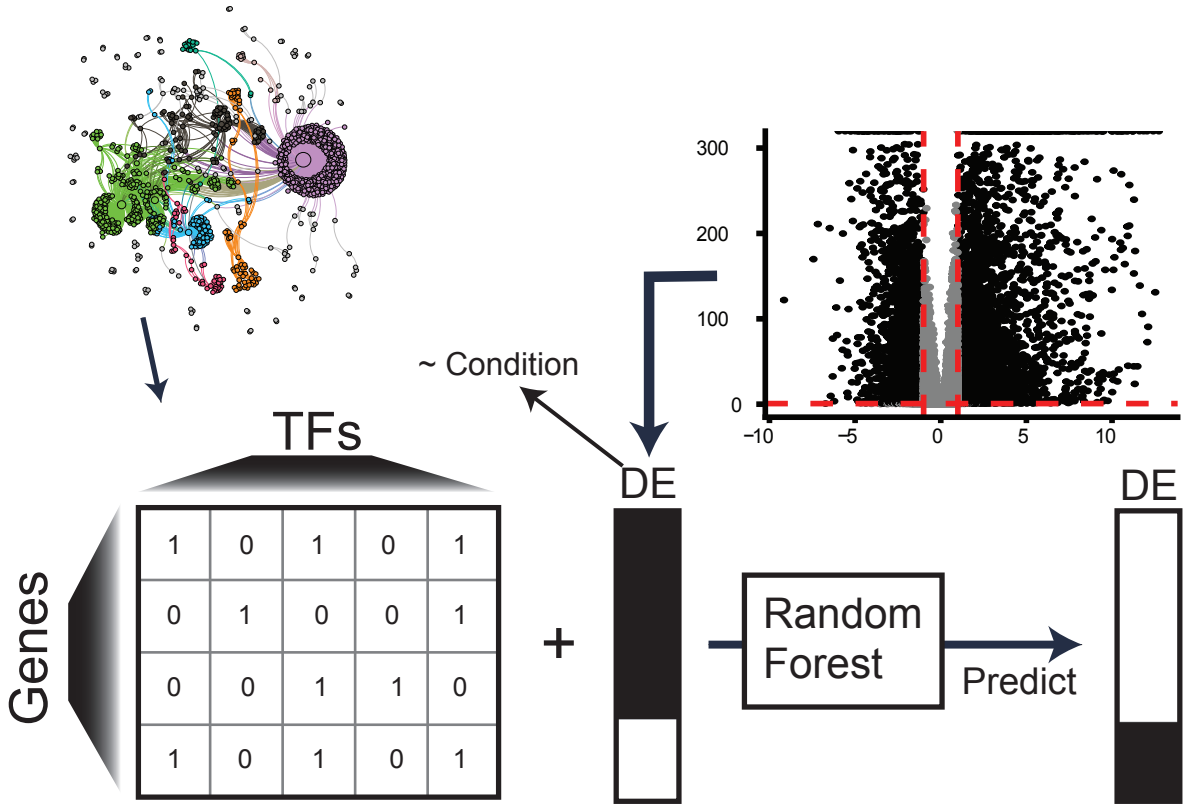

Figure 2.1.1. GRaNPA overview
Overview of the GRaNPA algorithm: utilizing a GRN to construct a model that predicts differential expression values.
This figure is produced by myself and is part of (Kamal et al., 2023).

should offer predictive insights into the effects of gene expression perturbations. This approach unbiasedly allows us to assess whether the network truly encapsulates regulatory connections or if it predominantly includes false positives.

2.2 Method

2.2.1 Requirements

In the context of GRaNPA, essential requirements include data on gene expression perturbations and a TF to gene GRN. Specifically, GRaNPA requires differential expression values between two conditions within a particular cell type or context. These values can be computed using various tools, such as DESeq2 (Love et al., 2014), though GRaNPA remains indifferent to the method of calculation.

The GRN input should be a TF to gene network, as GRaNPA utilizes TFs as features to predict gene expression. Gene-to-gene networks are unsuitable because they lack this TF information. For networks incorporating TF-peak-gene interactions, GRaNPA uses only the TF-to-gene component.

Furthermore, many GRNs include weighted information for connections, which GRaNPA can directly integrate into its model.

To summarize, GRaNPA requires log2 fold changes and adjusted p-values for each gene based on differential expression analysis. For the GRN, it necessitates a list of connections between TFs and genes, while additional details on edge weights and directionality are optional but could be beneficial for improving the prediction accuracy.

Mathematical notation of differential expression

The differential expression of a gene is represented mathematically as $DE(j)$, signifying the expression difference of the j^{th} gene. It is given by:

$$DE(j) = \log_2\text{fold change of the } j^{th} \text{ gene} \tag{2.1}$$

Mathematical notation of the GRN

The GRN is the main component of our prediction model. The GRN's mathematical representation is expressed as a function $grn(gene, tf)$, defined as follows:

$$grn(gene, tf) = \begin{cases} \text{weight} & \text{if tf and gene are connected in the GRN} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

In this function, the 'weight' represents the strength or confidence of the connection between a gene and a TF as provided by the GRN. This relationship is used to construct a matrix X, which outlines the connections between genes and TFs in the GRN. In matrix X, each row corresponds to a gene and each column to a TF. The cell value is the 'weight' from the GRN; if a weight is not provided, the value defaults to '1' to indicate a connection between the gene and TF.

2.2.2 Prediction model in GRaNPA

The key part of GRaNPA is its prediction model. As previously indicated, the model utilizes network topology for predicting gene expression perturbations, requiring pre-calculated differential expression values. I have chosen to conduct the prediction within GRaNPA through supervised regression, selecting random forest as the preferred method after testing a range of models, from simpler ones like Lasso and linear regression to more complex ones such as SVM.

Additionally, I explored more advanced decision tree methodologies, specifically XGBoost (Chen and Guestrin, 2016), which showed comparable performance but were slightly slower. This led me to choose Random Forest as the primary predictive method in GRaNPA. In addition, another key reason for this choice was the ability of decision tree-based models to offer importance scores for all features, which I will discuss in subsequent sections. The mathematical representation of this prediction model is as follows:

$$DE \sim \hat{F}(X_{TFs}) \tag{2.3}$$

I implemented the Random Forest model using the "ranger" package (Wright and Ziegler, 2017) within the R programming environment. To reduce the risk of overfitting, I employed 10-fold cross-validation during the model training. Importantly, I did not apply any hyper-parameter tuning to the Random Forest model, maintaining the integrity of my predictive approach. I evaluated the model's performance by analyzing the cross-validation R^2 score.

In addition to regression, I also used GRaNPA for classification tasks, specifically for predicting the directionality of DE. I assessed the effectiveness of this classification using metrics like the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic (AUROC). The results from these metrics were consistent with the R^2 performance indicators.

2.2.3 Creating a permuted network to evaluate connection specificity

To evaluate the specificity of connections within the network during the random forest regression, I created a permuted control network mirroring the structure of the actual GRN. This control network retains the same number of edges and the same degree distribution for TFs, but the gene labels are shuffled. This permutation randomizes the connections between TFs and genes. The differential expression values remain unchanged. Additionally, if any weighting method is applied to the edges of the actual GRN, the same method is also used to recalibrate the weights for the permuted network's edges if necessary.

If the prediction accuracy for the randomized network is high, it implies that the original network may include many incorrect connections, indicating a lack of specificity. Although the original network might reflect true biological relationships, a strong predictive performance by the randomized version suggests that the network contains significant noise.

2.2.4 Assessing overfitting in GRaNPA through random signal generation

To ensure the network is not overfitted despite cross-validation, I created another randomized version of the actual GRN. I kept the structure of the original GRN but replaced the actual gene expression differences with random numbers. These numbers were uniformly spread between -5 and 5. Choosing a uniform distribution for these numbers ensures that the model isn't unintentionally trained on patterns resembling real gene expression data.

If the model predicts well under these conditions, it suggests that it might be overfitting, especially when the dataset has a limited number of genes or the network is too small, consisting of a few hundred connections or less. In such situations, it's better to be cautious with interpreting

GRaNPA's outputs. Ideally, the prediction accuracy for this test should be around zero, confirming that the model isn't overfitting.

2.2.5 Calculating feature importance

In random forest models, variable importance measurement is an essential tool for identifying the most influential predictors within a dataset. When applied in GRaNPA, this technique evaluates the impact of each TF on the prediction of differential gene expression. This approach enables the identification of key TFs, providing more understanding of the biological processes.

In random forests, a popular method for measuring variable importance is the Gini index method. It calculates the importance based on the improvement in the Gini gain splitting criterion each variable brings in individual trees. This method takes into account how each variable contributes to the model's accuracy (Strobl et al., 2007).

However, a more advanced and robust method is the permutation accuracy importance measure. This technique involves randomly shuffling a predictor variable's values, disrupting its original correlation with the response variable. If shuffling this variable significantly lowers the model's accuracy, it indicates the variable's strong association with the response. The measure of importance is the difference in prediction accuracy before and after permuting the variable. This method is particularly effective because it evaluates the impact of each predictor both individually and in interaction with other variables (Strobl et al., 2007).

For my work with GRaNPA, I chose the permutation method for its detailed approach to assessing variable importance, particularly due to its ability to handle complex interactions between variables. Although it's based on the "ranger" and "caret" packages, users have the option to use different alternatives.

2.3 Application

For the application of GRaNPA, detailed exploration within the scope of two distinct projects will be presented in the subsequent chapters of this thesis.

2.4 Discussion

The introduction highlighted a fundamental challenge in GRN analysis: the absence of ground truth and a lack of unbiased evaluation methods. To address this, I developed GRaNPA, based on the hypothesis that a network accurately representing biological insights can predict perturbations in gene expression. The strength of GRaNPA lies in its ability to identify key transcription factors that are essential in explaining gene expression variations. This capability is not only beneficial for analyzing the GRN itself but also enriches downstream analysis of differential expression datasets.

In discussing GRaNPA's limitations, it's important to note a few key points. First, GRaNPA assumes that changes in gene expression are mostly due to transcription factors. This is an oversimplification, as other factors like RNA stability also affect gene expression levels. Second, while GRaNPA generally performs better than randomized networks, its performance is not always high. This could mean that the gene regulatory networks it uses might not capture all the underlying biological signals. Improving its performance might require adding more gene-specific features. Third, GRaNPA doesn't account for the joint action of multiple transcription factors binding together. In addition, it struggles with datasets where only a few genes are differentially expressed compared. Lastly, transcription factors with fewer connections in the network tend to be less prominent in GRaNPA's feature importance results, as they influence fewer genes. These areas highlight where GRaNPA could be improved in the future.

To summarize, GRaNPA serves as an evaluation tool for TF-to-gene GRNs, helping in the identification of key features, which improve our understanding of the underlying biological mechanisms of the network.

Chapter 3

GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks

In the following chapter, I will explain an application of GRaNPA based on different macrophages enhancer-based gene regulatory networks (eGRNs). This study introduces GRaNIE, which has been implemented by Dr. Christian Arnold, and I used it to construct eGRNs based on induced pluripotent stem cells (iPSC) derived macrophage dataset. Furthermore, I evaluate these networks using GRaNPA and explore macrophages' biological insights using important transcription factors (TFs) reported by GRaNPA. The text in this chapter was originally written by me, has been proofread by large language model-based tools, and is adapted from the following reference:

Aryan Kamal, Christian Arnold*, Annique Claringbould, Rim Moussa, Nila H Servaas, Maksim Kholmatov, Neha Daga, Daria Nogina, Sophia Mueller-Dott, Armando Reyes-Palomares, Giovanni Palla, Olga Sigalova, Daria Bunina, Caroline Pabst, Judith B Zaugg. GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks.* Mol Syst Biol. (2023) 19: e11627. https://doi.org/10.15252/msb.202311627

3.1 Introduction

Enhancers, important genomic regions, play a significant role in cell-type specific gene regulation. Their malfunction is increasingly associated with a range of diseases (Claringbould and Zaugg, 2021). Genome-wide association studies (GWAS) have linked numerous genetic variants with various traits and diseases. Interestingly, most of these disease-associated genetic variants are located in non-coding regions (Claringbould and Zaugg, 2021), away from gene promoters, suggesting they likely impact enhancers and serve a regulatory function.

One of the major post-GWAS challenges is understanding the role of these disease-associated genetic variants in non-coding regions of the genome. It's often unclear which genes these

genetic variants target, complicated by the cell- and condition-specific activity of gene regulatory elements. TFs are believed to play a key role in regulating these elements. Recent studies highlight the significance of studying TFs, especially in the context of genetic variants linked to autoimmune diseases (Freimer et al., 2022). Research suggests that trans-expression Quantitative Trait Loci (eQTL), likely mediated by TFs, are more closely associated with disease variants than cis-eQTLs (Võsa et al., 2021). Additionally, the integration of enhancer analysis is essential for understanding TF functions in cell fate determination (Xu et al., 2021; Janssens et al., 2022). Studies using enhancer-based GRNs, such as those derived from paired RNA and ATAC-seq data, have improved our understanding of regulatory diversity in various cell types, including neurons and disease mechanisms like pulmonary arterial hypertension (Reyes-Palomares et al., 2020). Therefore, interpreting disease-associated genetic variants requires a combined analysis of TF activity, enhancers, and gene expression in a cell-type specific manner.

Various methods have been proposed to infer TF-gene networks, utilizing different data sources like co-expression in bulk (Huynh-Thu et al., 2010b; Haynes et al., 2013) or single-cell data (Kamimoto et al., 2023; Aibar et al., 2017; Bravo González-Blas et al., 2023), time-course data (Huynh-Thu and Geurts, 2018), or data curation (Garcia-Alonso et al., 2019; Han et al., 2018; Liu et al., 2015; Keenan et al., 2019). Similarly, techniques to infer enhancer-gene links include analyzing co-variation of peaks or targeted enhancer perturbations (Fulco et al., 2019; Schraivogel et al., 2020). However, there's a gap in tools that can jointly infer TF-enhancer and enhancer-gene links, essential for understanding TF interactions with regulatory elements in specific contexts. A critical step in regulatory network reconstruction is assessing their biological significance. Common approaches include benchmarking against simulated or known biological networks, but these methods have limitations, such as being based on assumptions or suffering from literature bias and low complexity (Chen and Mar, 2018; Pratapa et al., 2020). Considering these biases and limitations, it's important to develop an unbiased method for confirming the biological importance of identified regulatory interactions.

3.2 Results

3.2.1 Basic explanation of the GRaNIE method

GRaNIE was developed to analyze genetic and epigenetic variations in specific regions referred to as "peaks," which are identified through ATAC-seq data. The software is an R/Bioconductor package designed to infer interactions between TFs and enhancers/promoters, as well as enhancers/promoters and genes, using the same dataset in a context-specific way. This approach is based on a method used in a study that explored how enhancers contribute to the development of pulmonary arterial hypertension (Reyes-Palomares et al., 2020) and were implemented by Dr.

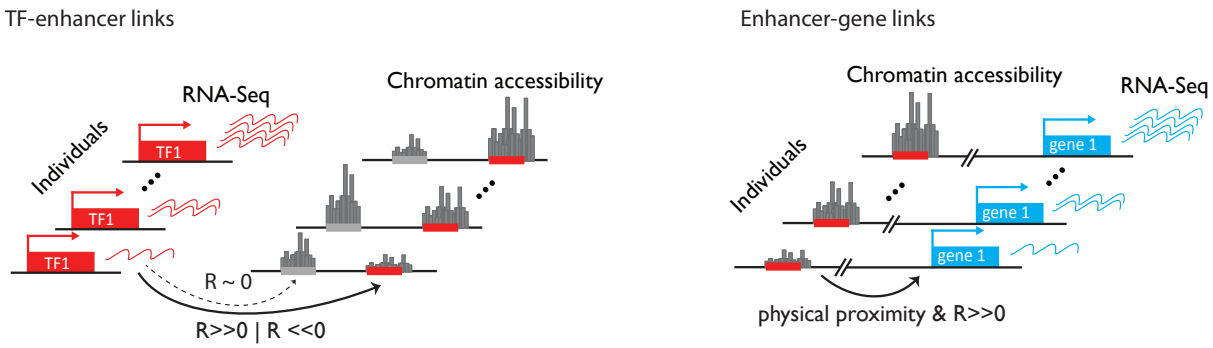

Figure 3.2.1. GRaNIE methodology overview
Schematic representation of eGRN construction via GRaNIE: TF to Peak (Left) and Peak to Gene (Right) connections. This figure was provided by Dr. Christian Arnold and is published in the (Kamal et al., 2023).

Christian Arnold. In brief, connections between TFs and peaks, as well as between peaks and genes, are first identified by GRaNIE. Then, these connections are combined to create an eGRN.

Identifying TF-Peak connections

The TF-peak links are established based on the statistically significant co-variation of TF expression and peak accessibility across samples (e.g., individuals, with a recommended minimum number of approximately 10-15), considering predicted TF binding sites. One key criterion for linking a TF to a peak is the presence of a TF binding site within the peak, which is mainly determined using datasets of predicted binding sites. For this, the primary choice is HOCOMOCO v11 (Kulakovskiy et al., 2018), but one can also use JASPAR (Castro-Mondragon et al., 2022) or the provided binding sites to establish these connections. To obtain these links, pairwise correlations between TF expression levels (measured by RNA-seq) and peak signals (measured by ATACseq or ChIP-seq) are calculated by GRaNIE.

For each TF, the distribution of all peaks that do not contain its predicted binding site is used as a background reference to calculate an empirical false discovery rate (FDR) for assessing the significance of TF-peak links Figure 3.2.1. In a previous study, it was demonstrated that negative TF-peak correlations suggest that the TF functions as a transcriptional repressor, while positive correlations indicate an activator role. This allows for the classification of TFs into activators and repressors (Berest et al., 2019). As a quality control measure, the recommendation is made to compare the number of actual TF-peak links to those obtained from a background set of links inferred from randomized data.

Identifying Peak-Gene connections

For the process of linking peaks with genes, a key criterion is their physical proximity to each other. This proximity is determined by measuring the distance between the peak and the gene's transcription start site (TSS). To ensure that both enhancers and promoters are considered, a

default distance of 250kb is used. In addition to distance, GRaNIE calculates the co-variation between gene expression and peak signals, and establishes connections for those with significant positive correlations Figure 3.2.1.

The quality of peak-to-gene links is assessed by evaluating the ratio of negative correlations between peaks and genes. In this assessment, focus on the directionality of correlations is important as it is widely recognized in genetics that the accessibility of chromatin at regulatory elements is associated with active gene regulation and transcription. Therefore, there is an expectation that effective connections between peaks and genes will typically show positive correlations. This principle holds true even when the regulatory elements are bound by repressors, as it is well-established that repressors generally reduce both the accessibility of these elements and the transcription of the genes they regulate (Berest et al., 2019). Consequently, negative correlations often have no clear biological meaning, potentially indicating noise or residual batch effects. Therefore, in evaluating the reliability of eGRN peak to gene connections, the proportion of negative to positive correlations is compared against a background dataset. A higher ratio of negative correlations is indicative of a noisier network.

Ultimately, GRaNIE integrates two types of connections: those between TFs and peaks, and those between peaks and genes. This integration is based on certain criteria such as the FDR threshold for TF-peak connections, the distance of peaks to gene TSS, and the statistical significance of the links between peaks and genes. By combining these elements, GRaNIE creates a three-part network consisting of TFs, genes, and enhancers.

This entire pipeline, which includes methods for identifying communities within the network and analyzing their gene ontology (GO), is accessible as part of R/Bioconductor packages.

3.2.2 Using GRaNIE to generate cell-type-specific eGRN in macrophages

Macrophages are large white blood cells belonging to the innate immune system. They are present in nearly all tissues and are involved in inflammatory conditions. Several autoimmune and other diseases have genetic variants that are enriched in enhancers active in macrophages (Alasoo et al., 2018; Novikova et al., 2021). Since inflammatory conditions are linked to many common diseases, macrophages are a critical cell type where disease-associated variations have a noticeable impact. Therefore, macrophages are an ideal choice for applying and testing the eGRN framework.

I gathered paired RNA-seq and ATAC-seq data from macrophages derived from iPSCs from (Alasoo et al., 2018). This dataset includes information from 31 to 45 individuals and includes macrophages in four different states: naive, primed with interferon gamma (IFN-γ), infected with Salmonella, and both primed with IFN-γ and infected (Alasoo et al., 2018). After initial data preparation and prepossessing, I constructed eGRNs for each of these conditions using the GRaNIE workflow. For identifying binding motifs within these networks, I used HOCOMOCO v11 (Kulakovskiy et al., 2018).

Evaluation of eGRNs by molecular evidence

Besides the usual quality control for these networks, I also aimed to assess them using separate molecular evidence. This additional evaluation helped in setting default and meaningful values for the statistical significance of both TF to peak and peak to gene connections. A major obstacle in this process was the absence of a molecular ground truth for TF-peak-gene connections, requiring separate evaluations for each link type. For the TF-peak connections, I utilized cell type-specific ChIP-seq data. For the evaluation of peak-gene connections, cell type-specific eQTL data were used.

ChIP-seq evaluation for macrophage eGRNs

I specifically gathered ChIP-seq data for macrophages from the Remap 2022 database (Hammal et al., 2021). For all the TFs available in Remap, I calculated the enrichment of GRaNIE's TF-peak connections within the ChIP-seq peaks. In this process, I used all ATAC-seq peaks that contained TF motifs as a background reference (Method Section 3.4.5). From this evaluation, I discovered the strongest signal for naive, primed, and infected macrophage eGRNs at an FDR of 0.2. I observed that the enrichment decreased as the FDR increased, suggesting that an FDR of 0.2 could be an appropriate default threshold for TF to peak connections Figure 3.2.2A. Similar to the direct quality control from GRaNIE Figure 3.2.2C, the network for primed-infected macrophages did not show significant enrichment. Based on this finding, I decided to exclude this particular network from further analysis.

cis-eQTL evaluation for macrophage eGRNs

For the next phase, focusing on the connection between peaks and genes, macrophage-specific cis-eQTL data was utilized to assess the enrichment of eQTL links within GRaNIE's connections. To perform this enrichment analysis, a set of distance-matched links was used across various FDR levels (Method Section 3.4.5). The findings from this analysis indicated that as the FDR increased, the odds ratio decreased Figure 3.2.2B. Consequently, an FDR of 0.1 was identified as a suitable threshold for the eGRNs. This peak-to-gene evaluation was carried out by Dr. Annique Claringbould.

Considering the results from these two molecular evaluations, I decided to establish default parameters for the GRaNIE algorithm. For TF-to-peak connections, the default FDR is set at 0.2, and for peak-to-gene connections, it is set at 0.1.

GRaNIE quality control for macrophage eGRNs

Applying the default settings established through molecular evidence, I proceeded to analyze the eGRNs I had constructed. These networks comprised between 92 to 126 TFs, 1411 to 6742 enhancers, and 1454 to 3869 genes Figure 3.2.3. When GRaNIE was applied to randomized data

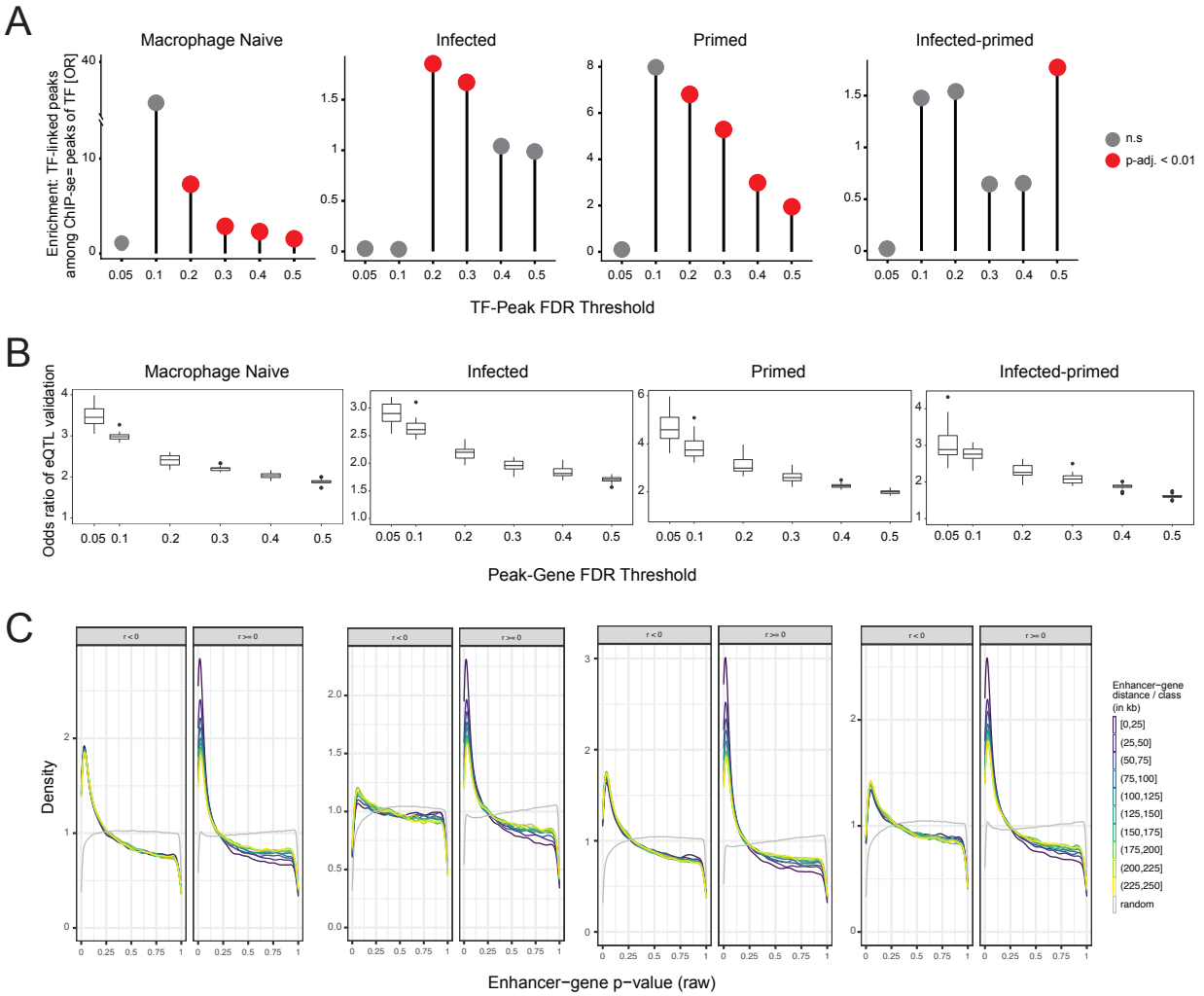

Figure 3.2.2. Validation and QC of macrophage eGRNs
(A) Validating eGRN TF-peak links with ChIP-seq data. Displaying ChIP-seq peak enrichment overlapping GRaNIE-inferred TF-bound peaks (same TF) at different TF-peak FDRs in naive (left), infected (second from left), primed (second from right), and primed-infected (right) macrophage eGRNs. Background: peaks containing the respective TF motif but lacking significant links.
(B) Validating eGRN peak-gene links with macrophages eQTLs. Illustrating the enrichment of eGRN links overlapping an eQTL compared to randomly sampled distance-matched peak-gene links overlapping an eQTL for various peak-gene FDRs in naive (left), primed (second from left), infected (second from right), and primed-infected (right) macrophages eGRNs.
(C) Observing QC and Summary of Peak-Gene Connections in Macrophage eGRNs. The distributions of p-values are presented for positive (right) and negative (left) peak-gene correlations. Peak-gene pairs are stratified by their distance (heat colors), while permuted gene-peak pairs from the randomized network are depicted in gray.
This figure was contributed by Dr. Annique Claringbould (Part B) and Dr. Christian Arnold (Part C) and myself (Part A) and is published in the (Kamal et al., 2023).

(where sample labels, peak labels, and motif labels were permuted), significantly fewer meaningful connections were observed in all eGRNs. This indicates that the TF-peak links meet quality control standards.

Likewise, I evaluated the ratio of positively correlated peak-to-gene connections against negatively correlated ones, which I regarded as noise. For Naive, Primed, and Infected networks, I

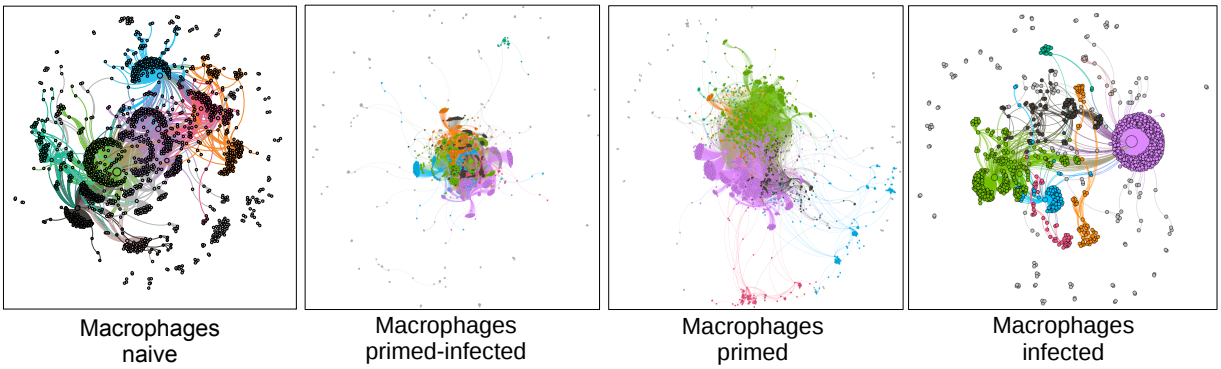

Figure 3.2.3. Visualization of macrophage eGRNs
Network visualizations and community identification in other macrophage eGRNs within this manuscript using forced-directed visualization. Colors indicate identified network communities.
This figure was produced by myself and is published in the (Kamal et al., 2023).

observed a satisfactory ratio, indicating a good signal quality Figure 3.2.2C. However, as previously mentioned, the network for the primed-infected condition did not demonstrate a favorable quality in this ratio Figure 3.2.2C. This finding aligned with my decision to exclude the primed-infected network from further detailed analysis.

Evaluating enhancers using CAGE database

Additionally, a significant enrichment of TF-peak-gene links among active enhancers specific to the cell type, based on CAGE data (Andersson et al., 2014), was discovered. This further confirms that biologically meaningful eGRNs were inferred by GRaNIE Figure 3.2.4A, a task completed by Makism Kholmatov.

Noticeably, within all eGRNs, it is evident that approximately 20-30% of the peaks exhibit connections to their nearest gene TSS, as depicted in Figure 3.2.4B. This observation aligns with earlier findings in pulmonary arterial endothelial cells (Reyes-Palomares et al., 2020) and iPSC-derived cardiomyocytes (Bunina et al., 2021).

Well-known macrophage TFs among the most connected in eGRNs

Furthermore, I also calculated the average number of enhancer peaks associated with each gene under different conditions. In the naive condition, a gene typically linked to about 4.4 peaks, whereas in the infected condition, the average was around 2.9 peaks. For the primed condition, this number increased to approximately 5.9 peaks. Within these connections, an average of 1.8 peaks in the naive condition, 1.5 in the infected condition, and a notable 5.7 in the primed condition were connected to TFs and formed part of the eGRNs identified by the GRaNIE algorithm Figure 3.2.5A.

These observations highlighted a limitation of the GRaNIE algorithm, as it seemed to miss some potential connections, indicating that there are likely more interactions in the actual biological systems than what was captured in the networks. Additionally, it was evident that

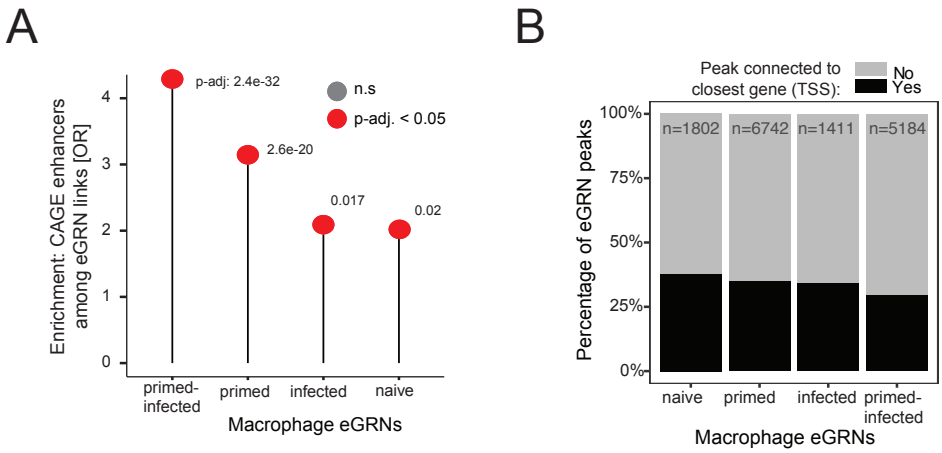

Figure 3.2.4. Enhancer validation of macrophage eGRNs
(A) Analyzing enrichment of macrophage-specific FANTOM5 CAGE (Andersson et al., 2014) enhancers in macrophage eGRN Peaks. Statistical significance was assessed using Fisher's exact test, with the test set comprising all peaks considered for peak-gene connections within each eGRN (n = 210,083, 227,035, 227,120, and 219,823 peaks for the naive, infected, primed, and primed-infected eGRN, respectively). Categories included overlap with CAGE enhancer and being part of the GRaNIE network.
(B) illustration of the fraction of eGRN peaks connected to the closest gene (depicted in black) versus other genes (in grey) for the macrophage eGRNs.
This figure was Provided by Maksim Kholmatov(part A) and Dr. Christian Arnold(part B) and is published in the (Kamal et al., 2023).

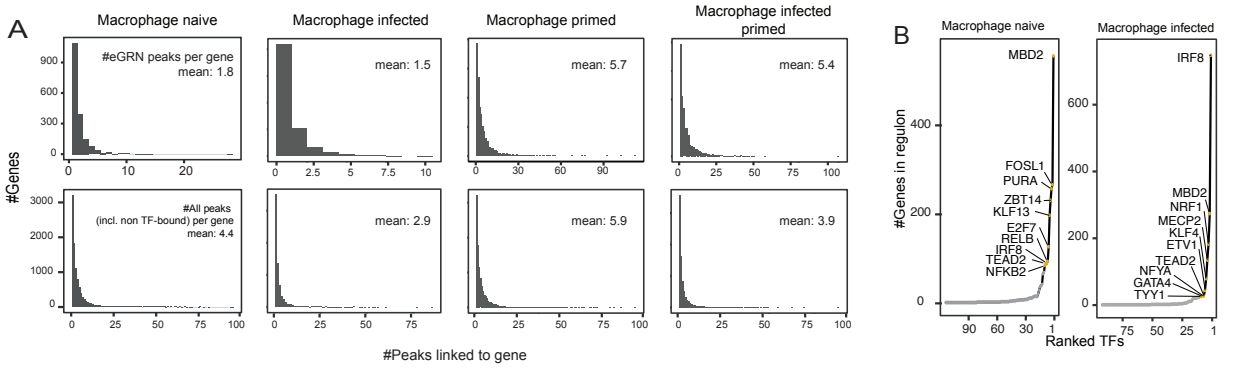

Figure 3.2.5. Connection analysis of macrophage eGRNs
(A) Histograms depicting the count of peaks linked to a gene across various macrophage eGRNs, along with their respective mean values. The upper row tallies peaks exclusively if they are TF-bound, as determined by GRaNIE as proper TF-peak-gene connections. In contrast, the lower row includes all peaks, encompassing those not associated with a TF, thus emphasizing all significant TF-gene connections while disregarding the TF-peak FDR.
(B) Illustration of the number of genes connected to each TF within the naive macrophage eGRN (with labeling for the top 10 TFs).
This figure was provided by Dr. Christian Arnold and is published in the (Kamal et al., 2023).

most TFs were connected to only a few genes Figure 3.2.5B. This pattern is consistent with the typical scale-free structure of GRNs, where a small number of nodes (in this case, TFs) have a large number of connections, while the majority have very few (Ouma et al., 2018).

The most highly connected TFs in the infected and naive eGRNs include several well-known macrophage TFs such as IRF8, NFKB2, and RELB (Langlais et al., 2016; Grigoriadis et al., 1996), as well as non-conventional macrophage TFs like MBD2, FOSL1, and NRF1. These non-canonical TFs were only recently linked to macrophage biology in mouse studies (Jones et al., 2020; Morishita et al., 2009; An et al., 2020).

Naive macrophage eGRNs uncover M1 and M2 polarization potentials in detected communities

To explore the biological insights provided by eGRNs, GRaNIE offers features for detecting subnetworks or communities (utilizing Louvain clustering by default, following the implementation in the R igraph package (Blondel et al., 2008)) and conducting GO term enrichment analysis on these communities. In line with the scale-free structure of these networks, I typically observe a few substantial communities and a long tail of very small and isolated nodes in each eGRN Figure 3.2.3.

Within the communities of the naive macrophage eGRN Figure 3.2.6A, one community is enriched in GO terms related to pro-inflammatory processes (such as the response to IL-1, chemotaxis, and response to IFN-γ). Another community is enriched in GO terms related to anti-inflammatory processes (including angiogenesis, cytoskeleton reorganization, and positive regulation of osteoblast differentiation; Figure 3.2.6E). This reflects the potential of naive macrophages to polarize into either M1 (pro-inflammatory) or M2 (anti-inflammatory) cell states (Murray, 2017). The M1-phenotype cluster is regulated by NFKB1/2 and REL, while the M2-phenotype cluster is regulated by TEAD1/2 and GATA4.

Among the communities of the infected macrophage eGRN Figure 3.2.6C, one community is enriched in pro-inflammatory processes, another is enriched in phagocytosis-related processes, and the third is enriched in chemotaxis-related processes Figure 3.2.6F. This recapitulates the most important facets of the macrophage function (Parameswaran and Patial, 2010; Nathan et al., 1983; Meng et al., 2014). Importantly, each of these functional communities is regulated by a specific set of TFs: IRF8 for the pro-inflammatory community, MBD2, NFR1, and ETV1 for the phagocytosis community, and MECP2 for the chemotaxis community.

For evaluating the utility of the GRaNIE eGRNs, a comparison was made between real and permuted eGRNs in terms of the number and biological relevance of GO terms enriched in the TF regulons. It's important to note that the regulons of the permuted networks had the same degree distribution and size distribution as the regulons of the real eGRN. However, the regulons derived from the permuted networks were enriched in less specific GO terms that were not related to macrophage biology, unlike the regulons of the real eGRN. This analysis has been done by Rim Moussa.

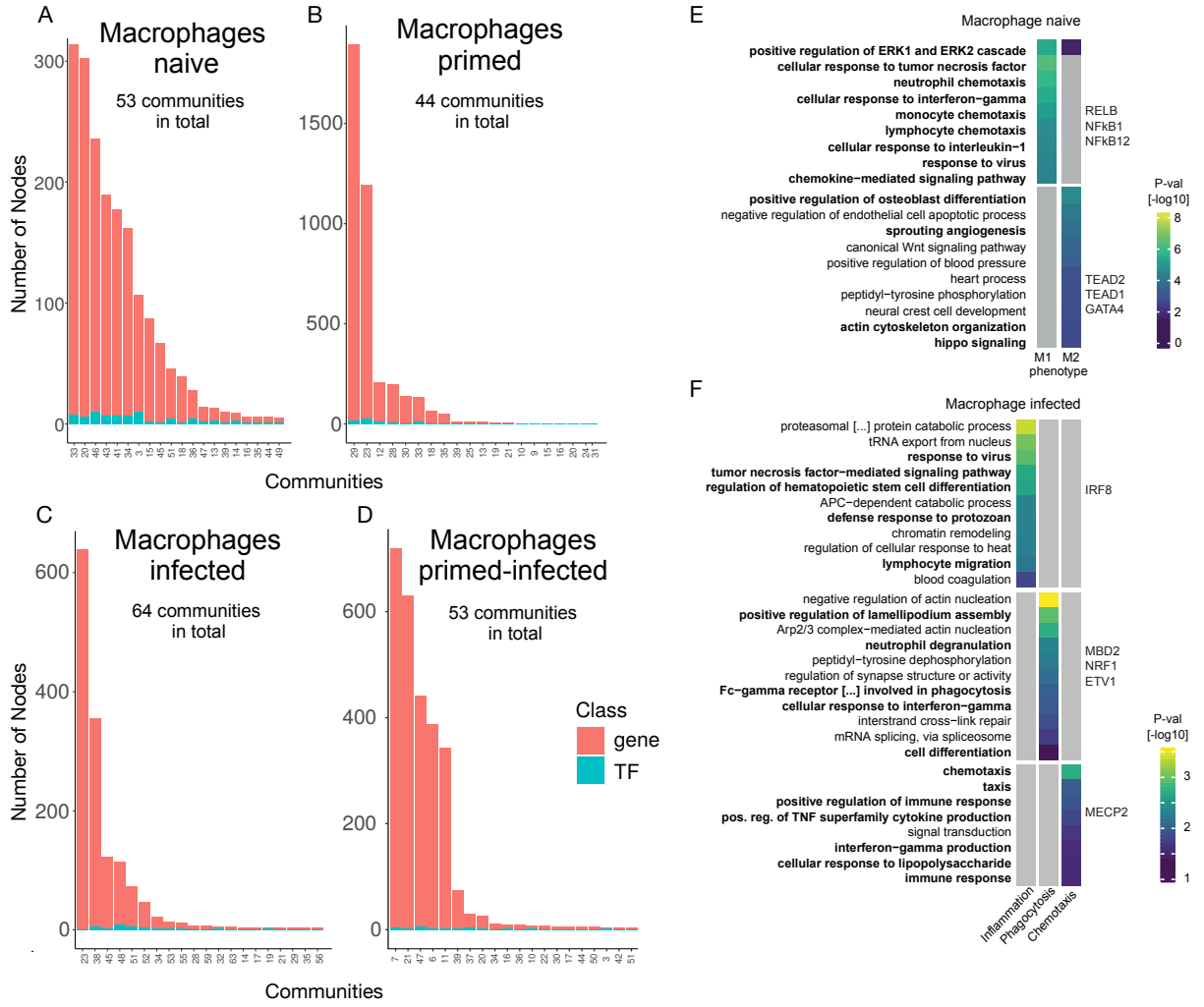

Figure 3.2.6. Community detection analysis and GO enrichment in macrophage eGRNs
(A-D) Community sizes in various macrophage eGRNs, including the number of genes and TFs within each community.
(E, F) Presentation of GO enrichment results and their corresponding P-values for chosen communities derived from the naive (E) and infected (F) macrophage eGRNs.
This figure was provided by Dr. Christian Arnold and Rim Moussa and is published in the (Kamal et al., 2023).

In conclusion, these findings indicate that the GRaNIE-inferred eGRNs effectively capture molecular evidence from eQTLs, ChIP-seq, and CAGE data. They prove to be valuable for investigating TF-driven biological processes in a cell type or state-specific manner.

3.2.3 Evaluation of macrophage eGRNs using GRaNPA

I used the GRaNPA tool (see Chapter 2) to assess the predictive power of the macrophage eGRNs that I had previously constructed. The aim was not only to evaluate the networks but also to identify key transcription factors explaining differential expression variations, therefore gaining deeper biological insights. To prepare for this, I conducted a differential expression analysis using DESeq2 (Love et al., 2014) on RNA-seq data from naive and Salmonella-infected macrophages

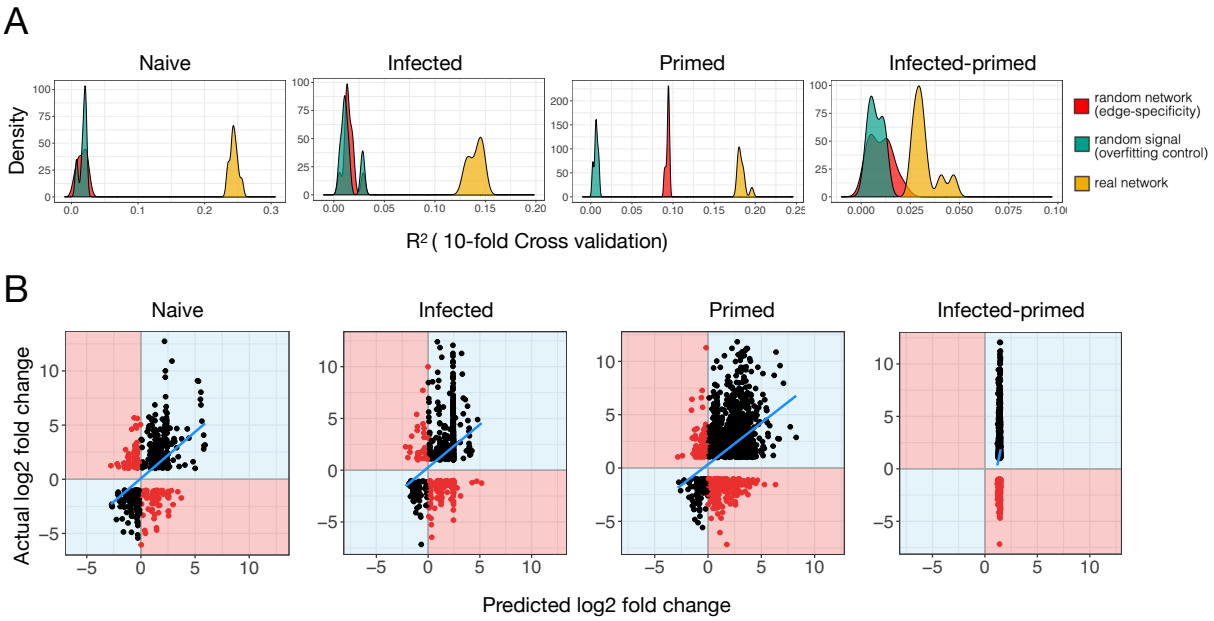

Figure 3.2.7. GRaNPA evaluation of macrophage eGRNs
(A) The results from GRaNPA are presented as a density distribution of R^2 values obtained from 10 random forest runs for all macrophage eGRNs, predicting differential expression in response to Salmonella infection. This analysis includes two permuted control datasets.
(B) The GRaNPA output is illustrated as a comparison between the actual and predicted log2 fold-changes in gene expression, specifically in the context of the macrophage response to Salmonella infection.
This figure was produced by myself and is published in the (Kamal et al., 2023).

from (Alasoo et al., 2018). I excluded the samples that were initially used in the construction of the eGRNs with GRaNIE to avoid circular arguments.

The evaluation with GRaNPA revealed that the three macrophage eGRNs were effective in predicting differential expression, with R^2 values ranging from 0.15 to 0.25 Figure 3.2.7A. Beyond just examining R^2 values, I also calculated the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) scores to predict the directionality of differential expression using the eGRN networks. The results, with AUROC scores between 0.65 and 0.75 and AUPRC scores from 0.71 to 0.88 Figure 3.2.8, indicated that network structure alone could predict the direction of differential expression to a substantial degree.

To ensure the robustness of these findings and prevent overfitting, I compared them against permuted networks. The significant differences in performance, as evidenced by t-test p-values less than 1e-6 for all comparisons, confirmed the networks' validity. Notably, the primed-infected macrophage network, which was previously excluded due to unsuccessful ChIP-seq validation Figure 3.2.2A, also failed to predict differential expression in the GRaNPA evaluation. This further validated the correspondence between GRaNPA results and molecular evidence. In summary, the significant difference between the actual networks and their permuted counterparts emphasized

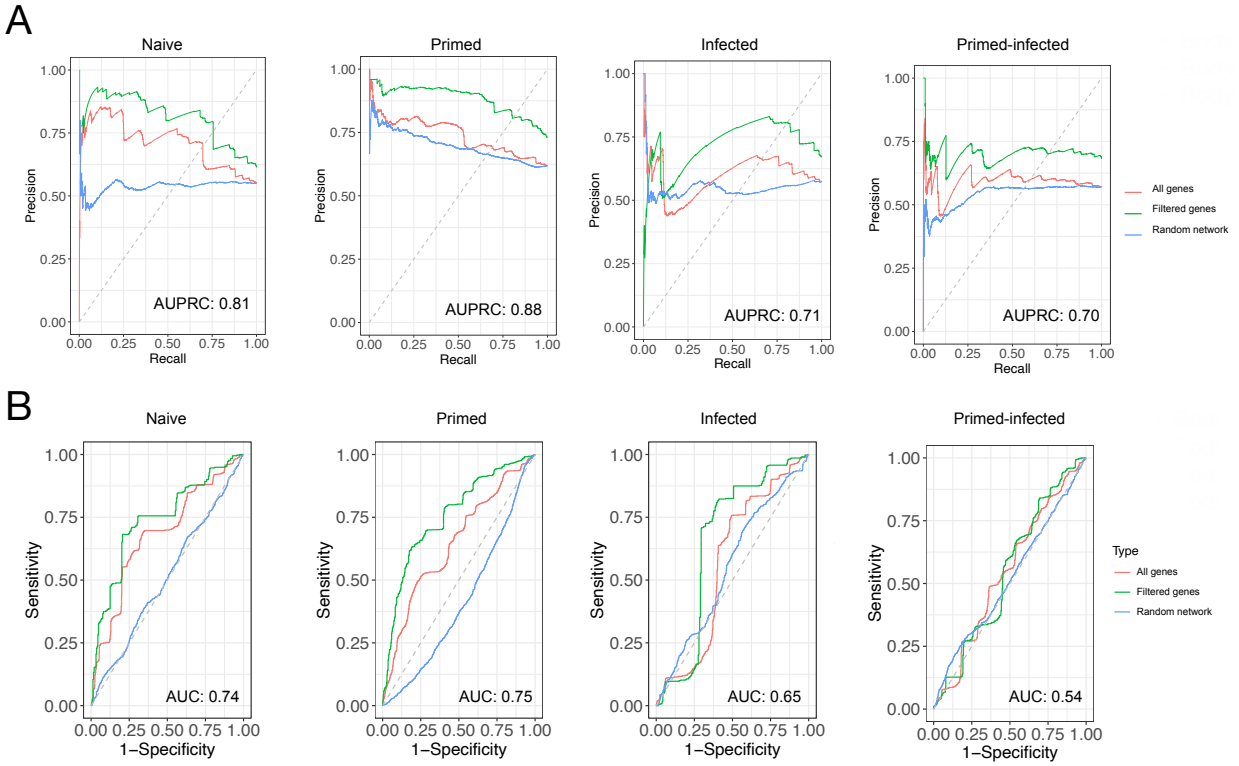

Figure 3.2.8. GRaNPA analysis of differential expression classification in macrophage eGRNs
(A) The output from GRaNPA is presented as a precision-recall curve for the classification of differential expression directions within various macrophage eGRNs. In this analysis, random forest classification models were trained using the real network for predicting all genes (depicted in red), the real network for predicting genes with an absolute log-fold change > 1 (shown in green), and the random network for predicting all genes (represented in blue). (B) The GRaNPA output for the classification of differential expression directions in different macrophage eGRNs is showcased here, similar to part A. In this instance, the receiver operating characteristic (ROC) curve is displayed. For more comprehensive details, please refer to the caption for part A.
This figure was produced by myself and is published in the (Kamal et al., 2023).

that the eGRNs constructed by GRaNIE effectively capture biologically relevant interactions between transcription factors and genes.

3.2.4 eGRNs constructed from single cell types demonstrate cell-type-specific predictions

To examine the specificity of the GRaNIE-inferred eGRNs for different cell types, I used datasets from various cell types with matched RNA and chromatin accessibility data. These datasets included primary human CD4+ T-cells sourced from (Freimer et al., 2022), and data for Acute Myeloid Leukemia (AML) obtained from (Garg et al., 2019; He et al., 2022). Using the same parameters as previously described, I ran GRaNIE, resulting in additional eGRNs for primary CD4+ T-cells and AML.

By this approach, I generated two unique eGRNs, each for a different cell type, using their specific RNA and ATAC data. This ensures that the connections in each eGRN reflect variations

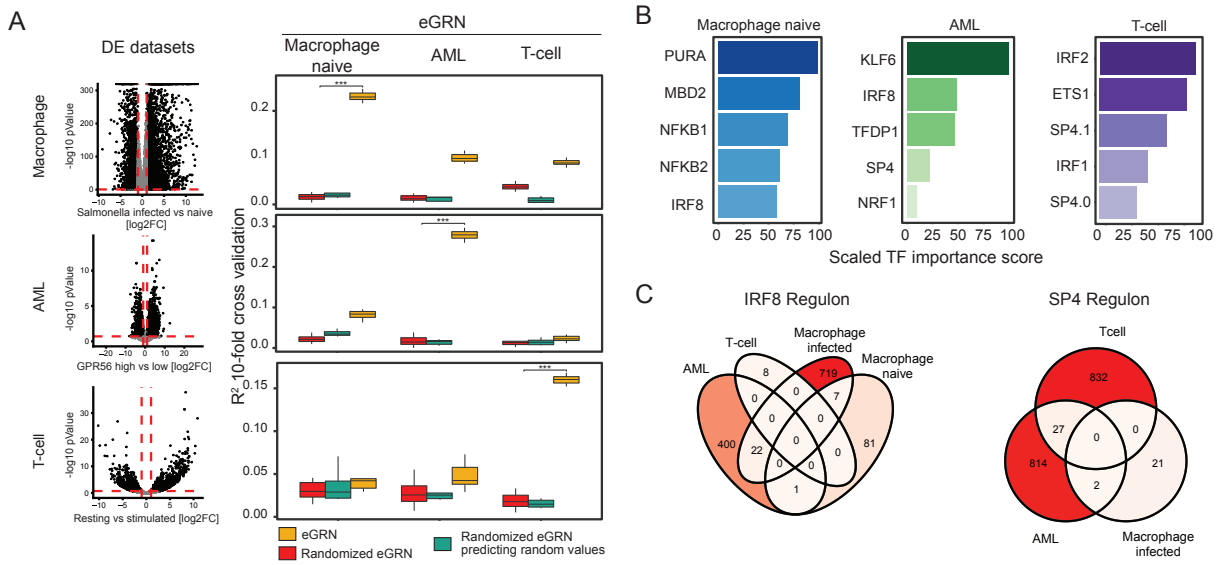

Figure 3.2.9. Cell type specificity assessment of eGRNs using GRaNPA
(A) The evaluation of eGRNs using GRaNPA is presented for naive macrophages (left), AML (middle), and T-Cells (right), focusing on differential expression analysis involving macrophages infected with Salmonella versus naive cells (top), two subtypes of AML (middle), and resting versus stimulated T-cells (bottom). The red lines denote the log2 fold-change (vertical line) and P-value (horizontal line) thresholds for genes included in the GRaNPA analysis. The boxplots display the distributions of R^2 values obtained from distinct random forest runs (n = 10). T-tests were conducted to compare GRaNPA performance between permuted and real networks (***P < 0.001).
(B) Identification of the top 5 most important TFs (0.0 and 0.1 indicate distinct TF motifs as defined by the HOCOMOCO database) for each of the eGRNs in (A) based on predictions within the same cell type.
(C) The overlap of SP4 (right) and IRF8 (left) regulons is depicted between eGRNs from different cell types, focusing on eGRNs with at least one connection to the respective TF.
This figure was produced by myself and is published in the (Kamal et al., 2023).

unique to that cell type. Therefore, the prediction of differential expression is specific to each eGRN and doesn't apply to eGRNs of other cell types.

To assess this, I used GRaNPA analysis on the eGRNs from naive macrophages, T-cells, and AML. I then compared their performance in predicting differential gene expression within each of these three cell types.

Specifically, I conducted differential expression analysis within the context of each cell type. For the T-cell eGRN, I calculated the differences in gene expression between resting and lipopolysaccharide (LPS) stimulated follicular CD4+ T-cells, using data from (Calderon et al., 2019). In the case of AML, I examined the differential expression between two AML subtypes (GPR56-high vs. GPR56-low), based on data from (Garg et al., 2019). Lastly, for the Macrophage cell type, I assessed the differential expression between naive and salmonella-infected macrophages, using data from (Alasoo et al., 2018).

In my analysis, I found that the eGRN matching the respective cell type provided the most accurate predictions Figure 3.2.9A. This indicates that indeed the eGRNs are highly specific to their respective cell types and can only effectively predict gene expression within those specific cell types.

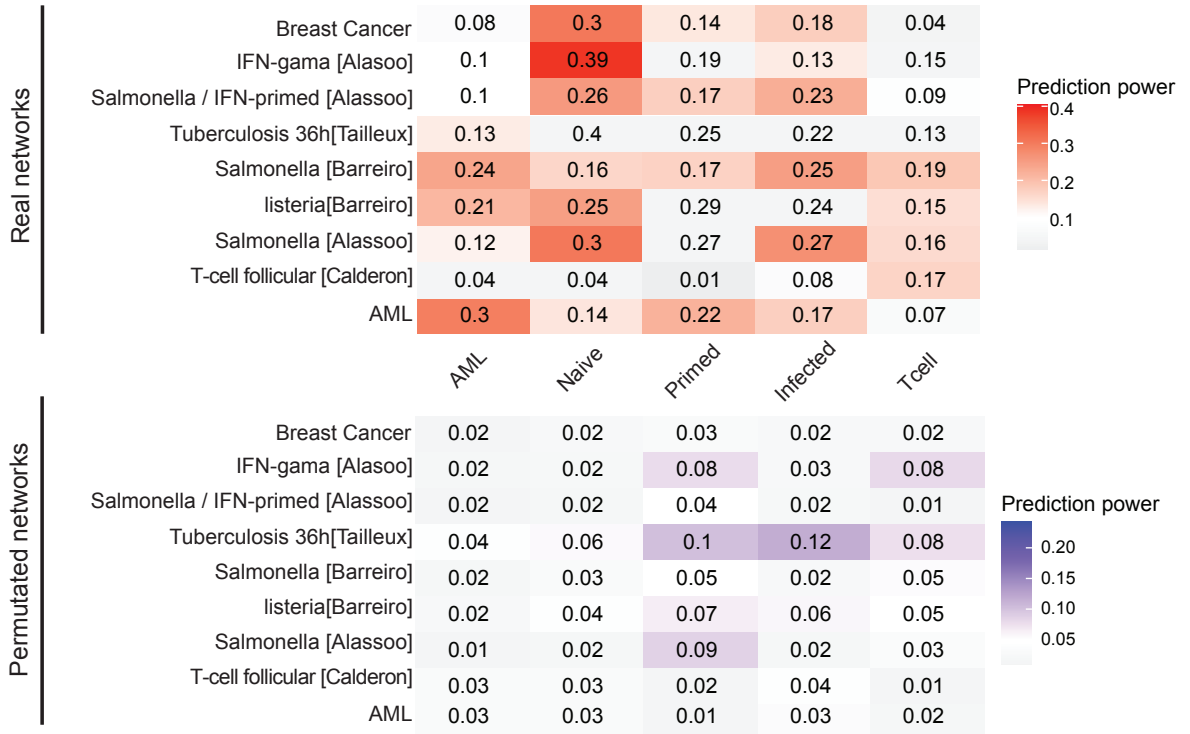

Figure 3.2.10. Incorporating gene-specific expression variation in GRaNPA analysis of eGRNs
GRaNPA Output with Gene-Specific Expression Variation Incorporation. The mean R^2 values, averaged across 10 random forest runs, are presented using a modified model that includes expression variation as a gene-specific feature, calculated from GTEx data. R^2 values are displayed, indicating the predictive performance for differential expression responses across nine distinct perturbations (rows) with unique GRaNIE-inferred eGRNs (columns; top), in comparison with the corresponding random networks (columns; bottom). eGRNs displaying a performance > 0.05 when compared to the random eGRN (bottom) are shaded in gray for the real network (top).
This figure was produced by myself and is published in the (Kamal et al., 2023).

While T-cells and macrophages exhibited predictive accuracy exclusively within their respective cell types, the AML eGRN, to a lesser extent, demonstrated the capability to predict macrophage responses. This observation suggests a possible shared regulatory framework between AML cells and macrophages, both belonging to the myeloid lineage.

Improving predictive power with gene-specific features

Furthermore, I observed that by incorporating gene-specific features, such as the variation in gene expression among individuals (as previously explored in (Sigalova et al., 2020)) Figure 3.2.10, R^2 values can be improved. However, it's important to clarify that these additional features provide insights into the sources of variation in differential expression. They do not, however, contribute to the evaluation of the eGRN itself or the identification of important TFs responsible for capturing a portion of this variation.

My primary focus remains on evaluating TF-gene links and assessing the cell-type specificity of eGRNs. Therefore, GRaNPA does not incorporate these gene-specific features by default.

However, it's worth noting that GRaNPA offers flexibility, allowing users to include additional features if desired.

Regulon specificity across cell types in GRaNPA analysis

Utilizing the TF-importance estimation within GRaNPA, I observed that the top 5 most important TFs, explaining the variation in differential expression for specific cell types Figure 3.2.9B, are mostly unique to one cell type. An exception to this pattern is IRF8, which was found to be important in both AML and macrophages, as well as SP4, which played a significant role in both AML and T-cells.

It's worth considering whether the regulons (groups of genes regulated by these TFs) are different between different cell types or if they regulate almost the same set of genes. Notably, the IRF8 regulons in AML and macrophages had only 22 genes in common, with no shared enhancers, while each included hundreds of distinct genes Figure 3.2.9C. Similarly, the SP4 regulons in T-cells and AML were nearly mutually exclusive. This suggests that even though SP4 and IRF8 play important roles in different cell types, their regulons have minimal overlap, highlighting their high cell-type specificity.

Validating TF regulons through knockout analysis

As an additional validation of the cell-type specificity of the TF regulons obtained from GRaNIE eGRNs, Dr. Nila Servaas conducted a comparison between the regulons and differential expression data resulting from TF knockout (K/O) within the same cell type. The data included information for one or two of the top five important TFs in each cell type: NFKB1 in macrophages (Somma et al., 2021), IRF8 in AML (Liss et al., 2021), and IRF1 and IRF2 in T-cells (Freimer et al., 2022). This analysis revealed a significant enrichment of genes that were downregulated upon TF K/O within the TF regulons of their respective cell types, providing further support for the cell-type specificity of these regulons Figure 3.2.11A.

Remarkably, the genes that exhibited downregulation following IRF8 knockout in AML were distinctly enriched in the IRF8 regulon specific to AML. This observation is particularly compelling, given that IRF8 is also a critical TF in macrophages, as depicted in Figure 3.2.9C. This finding underscores that the cell-type specificity of the GRaNPA predictions is not solely determined by distinct sets of TFs governing the cellular response. It also depends on the genes that these TFs regulate within that specific cell type. This emphasizes the crucial role of cell type-specific eGRNs in understanding these complexities.

As part of the validation process to confirm the cell-type specificity of enhancers in GRaNIE, Maksim Kholmatov obtained cell-type specific enhancer maps from FANTOM5 (Andersson et al., 2014). This data was derived from CAGE data for both T-cells and macrophages.

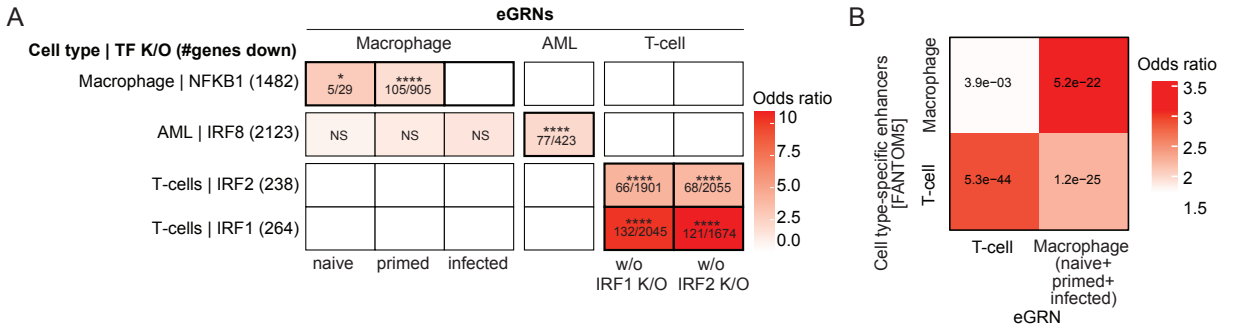

Figure 3.2.11. Odds ratio of target gene enrichment in cell-specific TF knockouts and eGRN enhancer overlaps
A) The odds ratios indicating the presence of target genes for NFKB1, IRF8, IRF1, and IRF2 in cell-type-specific knockouts (rows) within the corresponding eGRN regulons for macrophages, AML, and T-cells (columns). The figures within the chart represent the ratio of the number of genes within the regulon and downregulated post-TF knockout to the total number of genes in the regulon. The significance levels, determined by Fisher's exact test, are marked by asterisks, with the number of asterisks indicating the adjusted P-value thresholds. White squares represent regulons with no associated genes.
B) Enrichment of cell-specific enhancers from the FANTOM5 CAGE database within the T-cell and macrophage eGRN peaks. The numbers displayed in the tiles are Benjamini-Hochberg adjusted P-values from Fisher's exact test, comparing the number of peaks within the eGRN that overlap with identified CAGE enhancers against the total peaks in the respective cell types. The macrophage eGRN includes data from three states: infected, naive, and primed.
This figure was provided by Dr. Nila Servaas and Maksim Kholmatov and is published in the (Kamal et al., 2023).

Quantifying the degree of overlap between these enhancer maps and the enhancers from the T-cell and macrophage eGRNs, it became evident that there was a significantly stronger enrichment among the enhancers originating from the same cell type, in comparison to enhancers from different cell types. This finding is illustrated in Figure 3.2.11B and underscores the cell-type specificity of the identified enhancers in the GRaNIE framework.

Assessing promoter vs enhancer links in eGRNs using GRaNPA

The eGRNs establish connections between TFs and genes through active regulatory regions, encompassing both promoters and enhancers. In my evaluation, I set up a framework to assess the relative significance of promoter links (located within 10kb of the TSS) and enhancer links (situated more than 10kb away from TSS) within various eGRNs.

To perform this analysis, I categorized the gene-peak pairs into ten groups based on their distance from the TSS and ran GRaNPA separately for each group. The results revealed that the eGRNs that solely involved promoters, particularly in the case of infected and primed macrophages, exhibited limited or negligible predictive capacity Figure 3.2.12. This underscores the critical role of enhancers in this context and aligns with a recent study emphasizing the importance of considering enhancers when predicting the potential cell fate regulated by TFs (Xu et al., 2021).

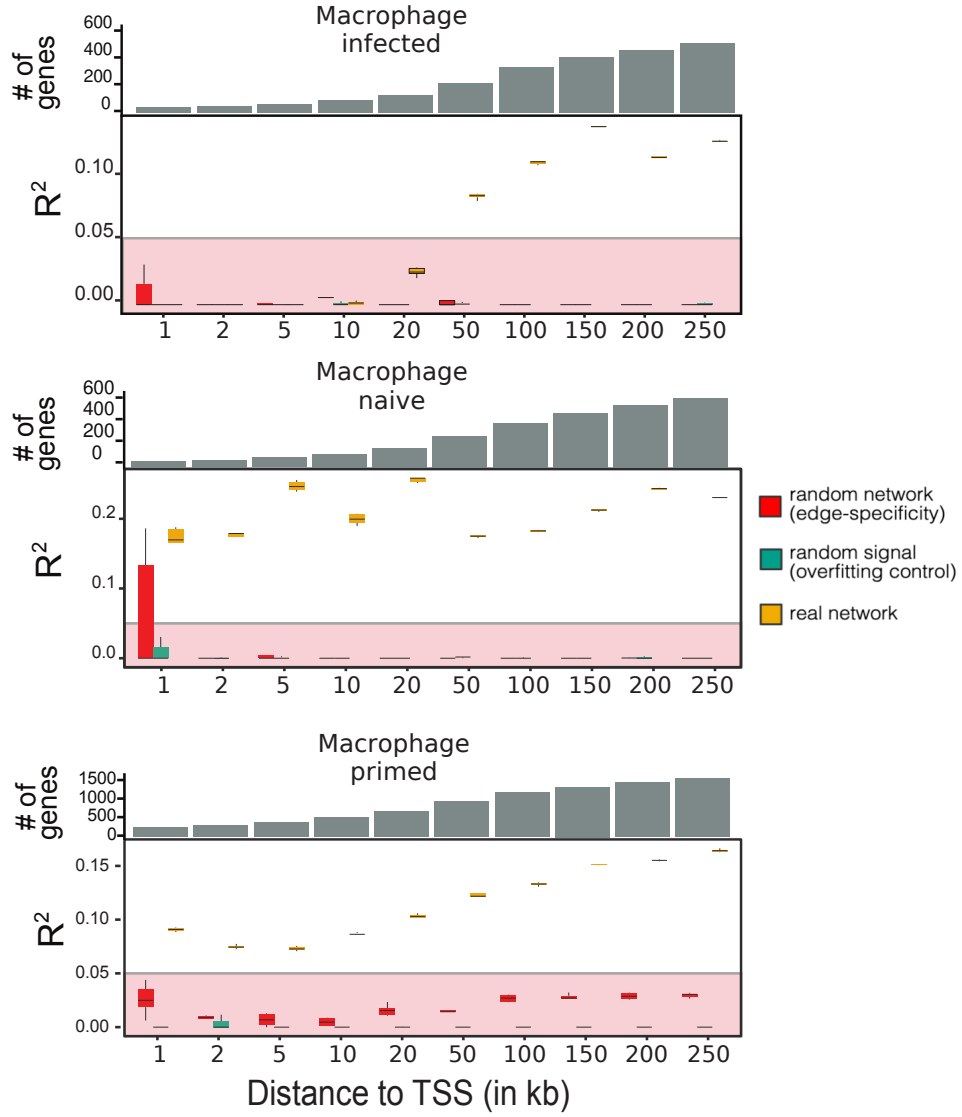

Figure 3.2.12. Analysis of subnetworks based on proximity between genes and enhancers in macrophage eGRNs
Analysis conducted by GRaNPA on the subnetworks, which are categorized according to the proximity between genes and their regulatory enhancers within eGRNs from three types of macrophages: infected, naive, and primed (arranged from top to bottom). The differential expression used for this evaluation is based on macrophage response to Salmonella infection at 5 hours compared to the naive state. Subnetworks are defined for each distance threshold 'k,' including all gene-enhancer connections that fall within the range from 0 up to that specific threshold distance 'k'.
This figure was Produced by myself and is published in the (Kamal et al., 2023).

3.2.5 Using GRaNPA to compare GRaNIE eGRNs with other GRN methods

It's worth noting that GRaNPA is a versatile tool that can be applied to evaluate any type of bipartite TF-gene network. Therefore, I utilized GRaNPA to assess the effectiveness of various previously published TF-gene GRNs. These networks establish connections between TFs and

genes using different methodologies. Here's a summary of the networks I evaluated, with more detailed information about each network available in the Section 3.4.6.

- **DoRothEA:** This network combines manual curation with data-driven approaches, including co-expression analysis, to establish TF-gene links (Garcia-Alonso et al., 2019; Holland et al., 2020).
- **ChEA3:** ChEA3 draws TF-gene links based on ChIP-seq experiments from sources like ENCODE, ReMap, or published literature (Keenan et al., 2019).
- **RegNet:** This curated network integrates TFs and miRNAs (Liu et al., 2015).
- **TRRUST:** TRRUST compiles TF-gene links from articles indexed in PubMed (Han et al., 2018).

Additionally, I included an enhancer-based GRN inferred using ANANSE (Xu et al., 2021) from macrophage data in the evaluation.

Next, I found that the cell-type-specific GRaNIE eGRNs and DoRothEA ABC (DoRothEA filtered by ABC confidence level) provided accurate predictions for all the datasets I tested. However, the TRRUST, RegNet, and ChEA3 networks had limited predictive power for macrophages. On the other hand, the ANANSE network, which is enhancer-based, performed poorly across all cell types Figure 3.2.13A. It's important to note that DoRothEA ABC outperformed DoRothEA all, suggesting that computationally derived connections in DoRothEA may not accurately represent the quality of validated connections. In summary, the GRaNIE networks generally outperformed most other networks and were comparable in performance to the well-curated DoRothEA.

Next, I compared cell-type specificity by examining the overlap between the TF-regulons identified in the networks with strong predictive capabilities and the genes that were downregulated when the same TF was knocked out (as introduced in Figure 3.2.11A).

The results showed that, overall, the cell-type-matched GRaNIE eGRNs outperformed all other networks in terms of recall Figure 3.2.13B. However, it's important to note that the absolute recall values were relatively small. This is likely because TF knockout can have many indirect downstream effects that aren't necessarily captured by the direct mechanistic links in eGRNs.

GRaNIE also demonstrated better performance than all other networks in terms of precision when evaluating AML. While DoRoThEA achieved the highest precision for IRF1 and IRF2 knockout in T-cells, the recall was lower. In summary, this cell-type-specific TF knockout evaluation emphasizes the significance of unbiased and cell type-specific eGRNs.

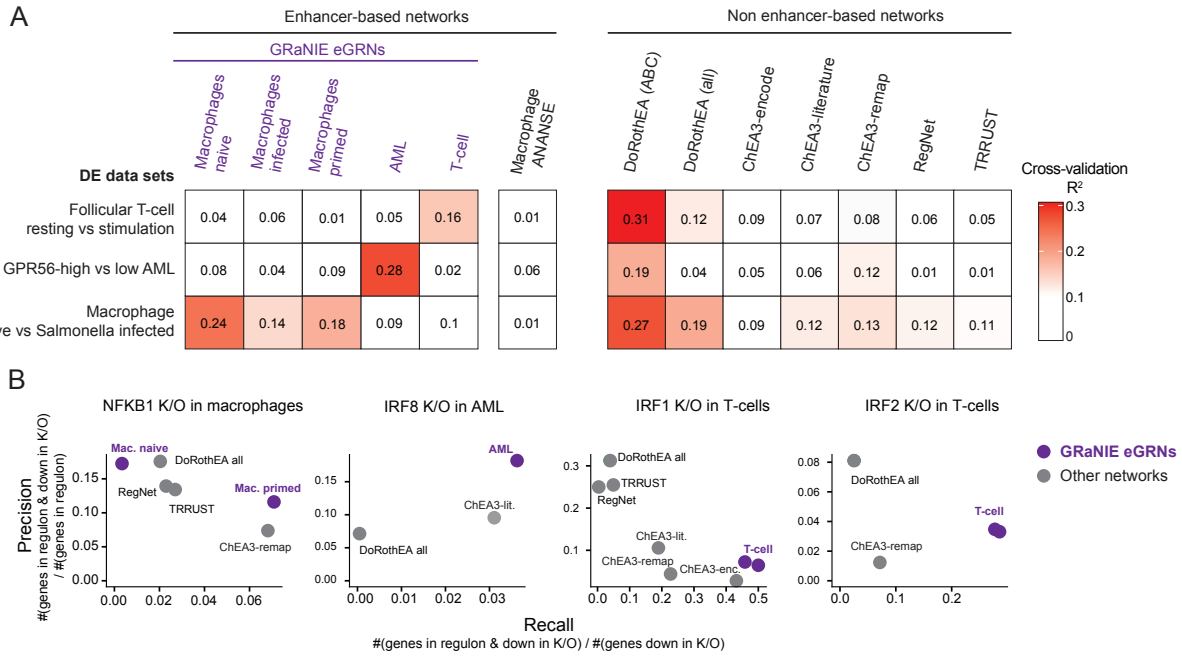

Figure 3.2.13. Comparative analysis of GRNs in predicting differential expression and TF knockout impact
A) Evaluation by GRaNPA of various GRNs, including five derived from GRaNIE (covering naive, primed, and infected macrophages, AML, and T-cells), an enhancer-based eGRN inferred using ANANSE, and several publicly available TF-gene networks. These include networks compiled from data curation (DoRothEA ABC and all), ChIP-seq datasets (ChEA3 encode, literature, and ReMap), and those created through manual curation (TRRUST and REGNET). These GRNs are assessed for their effectiveness in predicting differential gene expression in three scenarios: resting vs. stimulated follicular T-cells, GPR56 high vs. low AML, and naive vs. Salmonella-infected macrophages. The R^2 quantifies the predictive performance.
B) This part of the figure evaluates the precision-recall performance of the NFKB1, IRF8, IRF1, and IRF2 regulons from the networks mentioned in part A. The evaluation focuses on their ability to identify genes that are downregulated following the knockout of the respective TFs. For GRaNIE-derived eGRNs (shown in purple), the analysis highlights the performance of networks that match the cell type of interest, while other networks maintain a consistent performance across all analyses.
This figure was contributed by Dr. Nila Servaas (Part B) and myself and is published in the (Kamal et al., 2023).

3.2.6 Macrophage eGRNs uncover unique TFs governing diverse infection responses

GRaNIE and GRaNPA can also provide biological insights. Here, I utilized these tools to study various pro-inflammatory M1-like responses of macrophages to bacterial infections, as well as the anti-inflammatory M2-like response observed in breast cancer-associated macrophages.

Several datasets for Macrophages were collected and analyzed to investigate the impact of key TFs in M1 and M2:

- **Macrophages infected with Mycobacterium Tuberculosis (MTB)** - Data preprocessed by Daria Nogina (Giraud-Gatineau et al., 2020).
- **Macrophages infected with Listeria monocytogenes** - Data preprocessed by Daria Nogina (Pai et al., 2016).
- **Macrophages infected with Salmonella Typhimurium** - Data preprocessed by Daria Nogina.
- **Macrophages stimulated with IFN-y** - Data preprocessed by Sophia Mueller-Dott (Alasoo et al., 2018).
- **Comparative study of tumor-associated macrophages with tissue resident macrophages in breast cancer tissue** - Data preprocessed by Dr. Nila Servaas (Cassetta et al., 2019).

To gain a deeper understanding of macrophage responses under various conditions, I initially applied GRaNPA separately to the eGRNs of naive, infected, and primed macrophages Figure 3.2.14. However, the primed macrophage network did not meet the specificity criteria in several instances. Based on this, I decided to create a union network comprising only the naive and infected eGRNs. By forming a union network, I refer to the aggregation of connections from both the naive and infected eGRNs, rather than analyzing them individually.

This union approach proved to be more effective. The network, encompassing both naive and infected macrophage states, showed robust predictive capabilities across all the DE datasets that I used in this study, as demonstrated in Figure 3.2.15A. The ability of a network containing combined information from both naive and infected states to predict variations across the selected DE datasets was a significant finding, pointing to the value of this integrated approach in identifying key TFs relevant to each response.

Figure 3.2.15B illustrates the significance of TFs in both macrophages and the specific conditions of differential expression. One well-understood macrophage response is the activation of the NFKB family of TFs when exposed to IFN-γ (Medzhitov and Horng, 2009). NFKB, a key transcription factor of M1 macrophages, is essential for the induction of numerous inflammatory genes. These genes include those responsible for encoding TNF-α, IL-1β, IL-6, IL-12p40 and cyclooxygenase 2 (Liu et al., 2017).

In line with this, I identified NFKB2 as a highly important TF in response to IFN-γ stimulation Figure 3.2.15B. The NFKB2 regulon exhibited enrichment in GO terms related to chemokine signaling and taxis, and it displayed a substantial upregulation in response to IFN-γ Figure 3.2.15C. This underscores the ability of GRaNPA to pinpoint biologically relevant TFs.

To assess the reliability of GRaNPA, I compared the TF importance predictions across two independent datasets from Salmonella-infected macrophages. These profiles were strikingly similar,

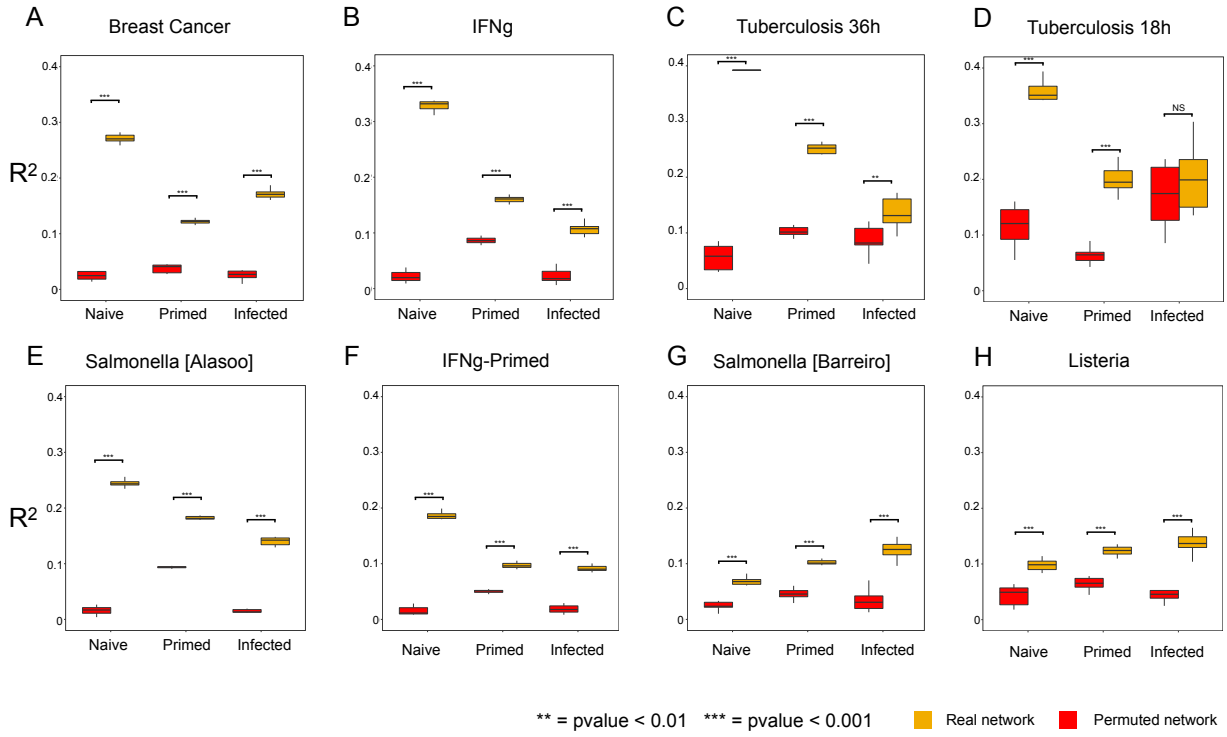

Figure 3.2.14. Evaluation of macrophage eGRNs in various differential expression conditions
GRaNPA evaluation of eGRNs for naive (left), primed (middle), and infected (right) macrophages across seven macrophage differential expression conditions. The results indicate that both the naive and the combined naive-infected eGRNs were effective in predicting the differential expression responses in most scenarios, with R^2 values greater than 0.1, contrasting with R^2 values below 0.05 for their respective random networks. The primed network, however, did not meet the specificity criteria in some cases. Owing to the high predictive accuracy of both naive and infected eGRNs, the evaluation of pathogen response was conducted using their union network, as illustrated in Figure 3.2.15A.
This figure was Produced by myself and is published in the (Kamal et al., 2023).

despite differences in the experimental setup (iPSC-derived vs. monocyte-derived macrophages) and time points (5 hours and 2 hours post-infection, respectively). This emphasizes the robustness of GRaNPA and underscores the biological consistency between the experiments.

In contrast, the TF importance profiles varied significantly across different conditions Figure 3.2.15B. This variation likely reflects the diverse roles of macrophages, such as M1 and M2 phenotypes, and the distinct defense mechanisms activated in response to various pathogens (Leseigneur et al., 2020). Notably, breast cancer-associated macrophages exhibited a distinct profile with TEAD1 and TEAD2 emerging as significant TFs. The GO analysis of the TEAD1/2 regulon revealed strong enrichment in angiogenesis, osteoblast differentiation, ERK signaling, and more Figure 3.2.15D, aligning with a more M2-like phenotype (Chen et al., 2020; Corliss et al., 2016).

The primary TF for predicting the response to Salmonella infection was IRF8, with MBD2 and ZBT14 following as the next significant contributors Figure 3.2.15B. IRF8 is a well-known pro-inflammatory interferon response factor associated with the pro-inflammatory (M1) polarization

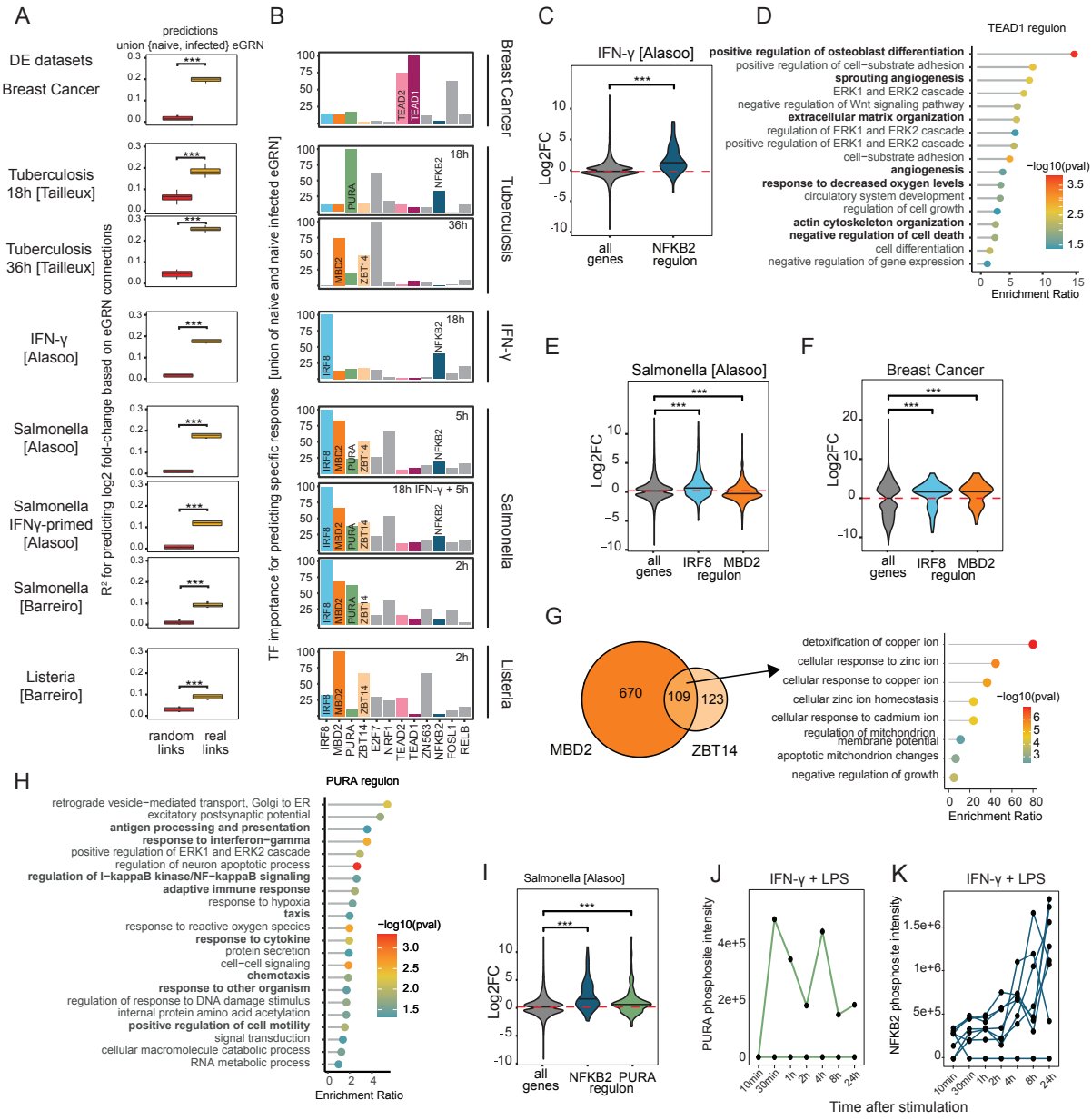

Figure 3.2.15. For the following figure's caption, please refer to the next page.

of macrophages (Chistiakov et al., 2018). This association was validated in our data through gene set enrichment analysis (GSEA) of the IRF8 regulon Figure 3.2.16, conducted by Dr. Nila Servaas.

While less is known about MBD2 and ZBT14 in the context of macrophages, MBD2 has been linked to intestinal inflammation in mice (Jones et al., 2020) and with an M2 macrophage program in pulmonary fibrosis (Wang et al., 2021). Consistent with this, the MBD2 regulon was down-regulated in response to infection Figure 3.2.15E but upregulated in breast cancer-associated macrophages Figure 3.2.15F, showing the opposite pattern to the IRF8 regulon Figure 3.2.15E-F. Furthermore, I found an enrichment of the M2 gene set among the MBD2 regulon in breast cancer-associated macrophages Figure 3.2.14. The MBD2 and ZBT14 regulons exhibited significant overlap (Figure 3.2.15F, p=3.3e-13, hypergeometric test), and genes jointly regulated by them

Figure 3.2.15. Exploring macrophage eGRNs through GRaNPA evaluations and predictive important TFs
(A) GRaNPA evaluation compares the combined naive and infected macrophage eGRNs (naive+infected eGRN; real links) with their permuted control counterparts (random links) across eight macrophage perturbation experiments. R^2 values from ten random forest iterations are represented as boxplots, with significant differences between real and permuted networks established through two-sided t-tests (***P < 0.001). Boxplots depict the interquartile range (25–75%) and median, while whiskers extend to 1.5 times the interquartile range from the box.
(B) The top five predictive transcription factors (TFs) for each experimental setting are detailed, highlighting specific TFs discussed in the text with individual labeling and color-coding.
(C) Log2 fold-changes for genes in the NFKB2 regulon, within the context of IFN-γ stimulation versus naive macrophages, are compared to the overall gene response.
(D) Included is the GO enrichment analysis for the TEAD1 regulon.
(E-F) Log2 fold-changes for genes in the IRF8 and MBD2 regulons are examined in response to Salmonella infection versus naive macrophages and in breast cancer-associated macrophages, alongside the total gene response.
(G) A Venn diagram illustrates the overlap between the MBD2 and ZBT14 regulons, complemented by a lollipop plot showing enriched GO terms for genes in the overlap.
(H) The GO enrichment analysis for the PURA regulon.
(I) The log2 fold-changes for genes in the NFKB2 and PURA regulons are compared to the general gene response in Salmonella infection versus naive macrophages.
(J-K) Normalized mass spectrometry intensity values for phosphosites on PURA and NFKB2 are plotted. These values are recorded in macrophages under M1 polarizing stimuli (IFN-γ and LPS) across different time points, with each phosphosite on the respective TFs individually charted.
This figure was contributed by Dr. Nila Sarvaas (J, K) and Rim Moussa (D, H) and the rest is by myself and is published in the (Kamal et al., 2023).

were enriched for terms related to the response to metal ions Figure 3.2.15G. The use of zinc and copper ions in macrophage defense strategies is well-documented (Stafford et al., 2013; Festa and Thiele, 2012). Since ZBT14 and MBD2 are crucial for predicting responses to pathogens but not to IFN-γ stimulation Figure 3.2.15B, I speculate that ZBT14 and MBD2 may collectively trigger a macrophage-intrinsic mechanism to counteract toxic metal ions. This mechanism may be aimed at overcoming the toxic effects of their own weapons.

3.2.7 GRaNPA pinpoints PURA as a potential proinflammatory TF in macrophages

Among the TFs less recognized for their role in macrophages, I identified PURA as significant in various infection settings. Indicative of a pro-inflammatory function, PURA's regulon was enriched with genes related to chemotaxis and IFN-$gamma$ response Figure 3.2.15H. This observation was reinforced GSEA (Provided by Rim Moussa), which showed a significant enrichment of M1-related genes in PURA-regulated differentially expressed genes during Salmonella infection Figure 3.2.15H. Moreover, genes in the PURA regulon were upregulated to a similar extent as genes in the NFKB2 regulon, a well-known pro-inflammatory TF Figure 3.2.15I.

To delve deeper into PURA's potential role in M1 polarization, the phosphoproteomics data, collected from macrophages stimulated with LPS and IFN-$gamma$ to induce the M1 phenotype (He et al., 2021), were obtained, and this analysis was performed by Dr. Nila Servaas. This data revealed a specific increase in phosphorylation at Thr187 upon LPS/IFN-$gamma$ stimulation

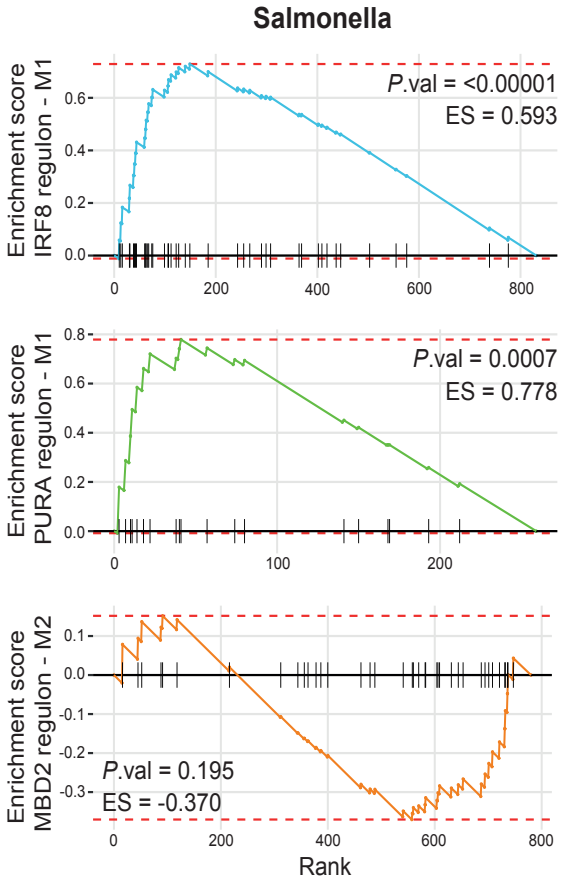

Figure 3.2.16. Differential gene expression and enrichment analysis in macrophage regulons post-salmonella infection
Gene Set Enrichment Analysis (GSEA) for genes within the IRF8 (top, blue), PURA (middle, green), and MBD2 (bottom, orange) regulons from the combined naive and infected macrophage eGRNs. These genes show differential expression (Padj < 0.05) in Salmonella-infected macrophages compared to controls. The genes are ranked based on their log2 fold-change (displayed on the x-axis). The enrichment score for each regulon is illustrated by colored lines (on the y-axis), with vertical black bars marking the positions of M1-associated genes in the IRF8 and PURA regulons, and M2-associated genes in the MBD2 regulon. The normalized enrichment score (ES) and the p-value (P.val) are also indicated.
This figure was provided by Dr. Nila Servaas and is published in the (Kamal et al., 2023).

Figure 3.2.15J. This phosphorylation pattern mirrored the increasing phosphorylation over time of phosphosites on NFKB2 Figure 3.2.15K. Significantly, the phosphorylation site in PURA is situated in the Purα repeats region, which is crucial for DNA binding and PURA function (Weber et al., 2016). Phosphorylation of DNA-binding regions is known to activate other TFs (Hirata et al., 1993), suggesting that PURA's activation plays a role in M1 polarization, further substantiating its significance in the pro-inflammatory response of macrophages.

In summary, these findings underscore the utility of GRaNPA in combination with cell-type-specific eGRNs for investigating a TF's specific functions in a particular cell type.

3.3 Discussion

Individual differences in physical traits arise mainly from two sources: genetic diversity and external factors. These external factors can have lasting effects (like epigenetics) or temporary ones (like signals). Both these sources can create variations in molecular traits, affecting more complex traits. Understanding these variations, especially in diseases, requires studying how genetics, epigenetics, and external cell signals interact. I introduce GRaNPA, a tool that, alongside GRaNIE, helps analyze these aspects together and explore their significance in biology.

GRaNIE is an R/Bioconductor software package designed to establish connections between TFs and peaks, and then link these peaks to genes. These connections are identified through statistically significant co-variation across individual samples or groups. GRaNIE ultimately forms a tripartite network comprising TFs, genes, and enhancers. The tool also includes downstream analyses of eGRNs, such as community detection and GO analysis based on communities or regulons.

As outlined in the previous chapter, another R tool, GRaNPA, is employed to impartially assess the biological significance of a network based on its structure. It also offers a platform to identify key TFs responsible for explaining the variations in specific datasets. In the absence of a known "true" network, GRaNPA serves as a valuable tool for comparing different eGRNs.

GRaNIE distinguishes itself by including enhancers in its networks, which proves beneficial in various applications. For instance, it has been instrumental in predicting how macrophages respond to infections by incorporating enhancer connections. The inclusion of enhancers in GRNs is crucial for studying genetic traits linked to specific TFs, thereby enhancing our understanding of the genetic variations associated with these traits. Due to the cell type-specific nature of enhancers, eGRNs derived from GRaNIE are likely more tailored to particular cell types compared to other TF-gene networks. Furthermore, GRaNIE addresses some of the limitations found in other methods for predicting TF binding sites. It does this by requiring a correlation between the expression of a TF and the accessibility of an enhancer, which helps to avoid establishing false connections. Although similar to a previous method (Marbach et al., 2016), GRaNIE is unique because the earlier data and software from that method are no longer available or maintained.

In scientific research, understanding the specific role of TFs in different cellular contexts is crucial. Studies have shown that TFs can bind to distinct sites, leading to varied roles across cell types or conditions (Zhang et al., 2023). Therefore, when analyzing GRNs to investigate a particular cell type or context, it is imperative to use networks constructed specifically for that context. GRaNIE, a tool within the R software environment, enables the creation of GRNs tailored to a specific cell type or context by analyzing variations across samples from that particular group.

More importantly, GRaNPA, another tool in R, validates the predictive capability of these context-specific GRNs. It has been observed that these networks are only predictive within

their specific context and fail to predict differential gene expression in other scenarios. This highlights the significance of understanding the network's context before employing it for biological interpretation. Both GRaNIE and GRaNPA are invaluable in guiding users toward more accurate and biologically relevant analyses, emphasizing the importance of context in understanding gene regulation.

GRaNIE and GRaNPA are powerful tools that not only assist in constructing and evaluating eGRNs but also enable detailed data analysis after eGRN creation. GRaNIE, in particular, offers insights into communities within networks. For instance, in the study of macrophages, it was found that each community centers around a few key TFs that are closely linked to specific biological processes in macrophages. Additionally, GO analysis based on these communities or regulons can be instrumental in understanding the roles of specific TFs or communities in a context or cell-type specific manner.

Moreover, GRaNPA sheds light on potentially significant TFs for specific research interests. In this study, examples in macrophages revealed that TFs like MBD2 and PURA, not widely recognized in macrophage studies, play crucial roles in fighting infections. For instance, certain TFs essential for infection response, which are not IFN-γ dependent, are known to bind to methylated DNA (Hainer et al., 2016). This finding correlates with recent discoveries about DNA methylation changes induced by pathogens, showcasing the depth of new biological insights that can be gained using GRaNIE and GRaNPA (Qin et al., 2021).

These tools collectively facilitate a more detailed understanding of gene regulation in specific contexts, highlighting less known but biologically significant TFs and their roles in cellular responses. This not only advances our knowledge in specific fields like macrophage biology but also opens new avenues for exploring gene regulation in various cell types and conditions.

3.4 Method

3.4.1 GRaNIE data sets

All data sets were subjected to PCA, accompanied by an inspection of metadata in the PCA space, to determine if any samples should be excluded as outliers. Details on any such exclusions are provided in the respective paragraphs.

RNAseq and ATACseq data for iPSC-derived macrophages

I utilized a publicly accessible data set (ERP020977) for my study. This data set includes RNA-seq profiles of naive and primed macrophages, both uninfected and infected with Salmonella for 5 hours, as detailed in (Alasoo et al., 2018). From this data set, I gathered 304 RNA-seq profiles from 86 individuals. Notably, 145 of these profiles were paired with ATACseq data, which is available here.

The samples were categorized into four groups: primed, primed-infected, naive, and naive-infected. The availability of samples was as follows: 41 paired RNA/ATAC and 43 only RNA-seq for primed, 31 paired and 55 only RNA-seq for primed-infected, 42 each for naive, and 31 paired and 55 only RNA-seq for naive-infected.

Along with these samples, the data set included metadata and peak coordinates. I used the paired samples to reconstruct the eGRNs with GRaNIE, and the unpaired RNA-seq data to evaluate the eGRNs with GRaNPA.

Expression and chromatin accessibility data for CD4+ T-cells

The paired RNA-seq and ATAC-seq data were collected from the public dataset GSE171737 (Freimer et al., 2022). The preprocessing has been done using a snakemake pipeline provided by Dr. Christian Arnold.

Expression and chromatin accessibility for AML

Raw RNA-seq data for 23 AML patients was obtained from a study by (Garg et al., 2019). Processed and quality-controlled ATAC-seq data and peaks for the same patients were obtained from (He et al, 2022) (He et al., 2022).

3.4.2 GRaNIE construction for other cell types

For using GRaNIE in this study, I needed the following inputs:

1. Either raw or prenormalized data on chromatin accessibility, such as ATAC-seq, DNase-Seq, or ChIP-seq data for histone modifications like H3K27ac.

2. Raw or prenormalized RNA-seq count data.

3. Pre-assembled lists of transcription factor binding site (TFBS) predictions for each TF, which I obtained as described in (Berest et al., 2019), for human and mouse TFBS.

In all the data sets for this study, I applied the default parameters as outlined in the subsequent section for constructing the eGRNs. The conceptual basis of GRaNIE involves several main steps.

3.4.3 Chromatin accessibility and RNA-Seq data

Both chromatin accessibility and RNA-seq data were processed, which could be either in raw counts or prenormalized form. For RNA-seq data with raw counts, quantile normalization was applied to minimize the effects of outliers on the correlations. Chromatin accessibility data was processed using size factor normalization, following the approach outlined in DESeq2. The choice of normalization method for each data type was at the researcher's discretion. Additionally, filters were employed to exclude specific chromosomes, like sex chromosomes, or genes/peaks with low counts. This entailed removing genes or peaks if their average counts across all samples fell below a specified threshold, which was set at 5 by default.

Matching TFBS with ATAC-Seq peaks

The process involved overlaying the TFBS for each transcription factor, as per the provided list (refer to (Berest et al., 2019) for methodology), with the open chromatin peaks identified in the ATAC-Seq data. For each peak and transcription factor, it was recorded whether there was at least one potential TFBS within the peak. This created a binary matrix of TF-peak bindings, which was then utilized in the following stages of the analysis.

Determining statistically significant TF-Peak connections

To determine statistically significant connections between TFs and chromatin peaks, a cell-type-specific, data-driven approach was employed. This process began with the calculation of Pearson's correlation coefficients, quantifying the association between the expression levels of each TF and the open chromatin signals observed in each peak across the collected samples.

An empirical FDR method was then applied to determine statistically significant connections between TFs and chromatin peaks. In this process, for each TF, the peaks were divided into two groups: a foreground group containing peaks with predicted TFBS and a background group comprising peaks without such predictions, as defined by the earlier established TF-peak binding matrix. The TF-peak correlation was then categorized into 40 intervals, progressing in steps of 0.05, ranging from -1 through 0 to 1. For each of these intervals, a specific FDR value was calculated, considering both positive (moving from -1 to 1) and negative (moving from 1 to -1) correlation directions.

For each bin, representing a correlation threshold k, the empirical FDR was calculated using the formula $efdr_k = \frac{nfp_k}{nfp_k+ntp_k}$. In this formula, nfp and ntp represent the total number of TF-peak connections in the background and foreground, respectively, where the correlation r is greater than or equal to k for the positive direction, and less than k for the negative direction. To align the data from the foreground and background, the number of TF-peaks in the background nfp was normalized relative to the foreground ntp using the formula $nfp_k = nfp_k \times \frac{ntp}{nfp}$, ensuring comparability between these two sets. This method was critical for accurately assessing the statistical significance of the TF-peak connections across different correlation thresholds.

Identifying statistically significant Peak-Gene connections

In the next stage, peak-gene connections were added to the network. Highly correlated peak-gene pairs were identified based on Pearson's correlation and the associated P-value, calculated using `cor.test` in R, between the normalized RNA-seq data for gene expression and the corresponding open chromatin peak.

GRaNIE employs a local neighborhood-based method for choosing peak-gene pairs to examine their correlation. This approach uses a specified neighborhood size, with a default setting of 250 kb both upstream and downstream of the peak, to select these pairs.

GRaNIE further records various properties for each peak-gene pair, such as their distance and the gene type & status as annotated by Gencode. By default, the eGRN includes only protein-coding and lincRNA genes, but this can be customized to incorporate other gene types.

Filtering GRN connections and calculating Peak-Gene FDR

In the final step of the process, various options are provided to combine and filter TF-peak and peak-gene connections to construct the complete GRN for further analysis. These connections can be filtered based on their FDRs or correlation values. Specifically, peak-gene links can be additionally filtered by criteria such as their distance and gene type. By default, only positively correlated peak-gene pairs are retained.

After applying all the filters to the peak-gene links, a multiple testing adjustment is conducted using the Benjamini–Hochberg method. The predefined thresholds for filtering TF-peak and peak-gene links are set at FDR < 0.2 and FDR < 0.1, respectively.

GRaNIE quality controls

The GRaNIE package provides optional PCA plots for both RNA-seq and open chromatin data. When additional metadata is available, it can be incorporated into these PCA plots, allowing for the data to be color-coded accordingly. This feature is particularly useful in identifying batch effects and outlier samples that might introduce unwanted variations into the data.

Furthermore, a series of quality control measures are implemented for different stages of the analysis. These measures aim to assess both the quantity and the signal-to-noise ratio for TF-enhancer and enhancer-gene links.

To evaluate TF-enhancer links, the number of links from actual data is compared to a background set obtained using randomized data. This involves a twofold randomization approach: permuting the TF-peak overlap matrix and sample labels for the RNA counts while applying the same methods as previously described.

For enhancer-gene link assessment, a background link set is similarly constructed. This process begins by shuffling the peaks in the real table of peak-gene pairs that meet the specified criteria for correlation testing. Importantly, this shuffling maintains the degree distribution for both peaks and genes. Additionally, the sample labels for the RNA data are shuffled.

Quality controls are based on the premise that peak accessibility and gene expression are positively correlated. Therefore, analyzing the proportion of significant positive versus negative correlations in the actual versus background enhancer-gene links acts as an indicator of the signal-to-noise ratio. Specifically, a higher number of positive correlations compared to negative ones is expected in real enhancer-gene links, which is unlikely to be observed in the background set. Additionally, several other quality control plots, including those representing enhancer-gene distance, are used, with the expectation of observing a discernible signal difference in real links but not in the background links.

3.4.4 Differential expression analysis for other cell types

Differential expression analysis was conducted using the DESeq2 package (Love et al., 2014) for all datasets, typically comparing treatment and no treatment or disease and control conditions. The general design formula was " condition," unless otherwise specified. Details specific to each dataset are provided below. The input for GRaNPA typically involved shrunken log2 fold-changes, generated with the lfcShrink function in DESeq2 using the apeglm method (Zhu et al., 2019), though it's not mandatory to use a particular transformation with GRaNPA.

1. iPSC-derived macrophages infected with Salmonella (Alasoo et al., 2018):
    - Contrasts: naive vs. infected, naive vs. IFN-$gamma$ primed, IFN-$gamma$ primed vs. IFN-$gamma$ primed-infected.
    - DESeq2 formula: "$\sim$condition."
    - Only RNA-seq data not used for eGRN reconstruction was considered for differential expression analysis.
    - This analysis has been done by myself.
2. Macrophages infected with Listeria and Salmonella (Pai et al., 2016):

- Differential expression compared to control, listeria-infected, and salmonella-infected samples separately.
- Donor information was used as a covariate in the design formula: "∼patient + condition."
- No samples were removed.
- This analysis has been done by Daria Nogina.

3. Macrophages infected with Tuberculosis (Giraud-Gatineau et al., 2020):

   - Differential expression was calculated between monocyte-derived macrophages from healthy donors infected with tuberculosis and control samples.
   - Datasets GSE133145 and GSE143731 were analyzed separately but shared a common design formula.
   - The design formula: "∼patient + treatment + condition."
   - Outliers in the PCA plot led to the removal of one control and one infected sample from the GSE143731 series.
   - This analysis has been done by Daria Nogina.

4. CD4+ follicular T-cells resting vs. LPS-stimulated (Calderon et al., 2019):

   - Differential expression between CD4-positive follicular T-cells in resting vs. stimulated conditions.
   - DESeq2 design formula: "∼condition."
   - This analysis has been done by myself.

5. AML subtypes (Garg et al., 2019):

   - Differential expression compared samples with high leukemia stem cell burden (GPR56-high) vs. low leukemia stem cell burden (GPR56-low samples) based on immunophenotyping criteria.
   - The design formula: "∼GPR56status."
   - Shrunken log fold-changes were not used as input for GRaNPA, although results remained qualitatively unchanged when applied.
   - This analysis has been done by Maksim Kholmatov.

6. Tumor-associated and tissue-resident macrophages from human breast tissue (Cassetta et al., 2019):

- Differential expression analysis was performed to compare tissue-resident and tumor-associated macrophages.
- DESeq2 design formula: "∼condition."
- This analysis has been done by Dr. Nila Servaas.

In summary, differential expression analysis was customized for each dataset based on the specific biological context, and results were used as input for further analyses, including GRaNPA.

3.4.5 Molecular analysis of TF-Peak connections in GRaNIE using ChIP-seq

I collected macrophage-specific ChIP-seq data from ReMap 2022 (Hammal et al., 2021) for specific TFs (CEBPA, CEBPB, FOS, GABPA, GFI1, IRF8, IRF9, LYL1, MYB, NR1H3, PPARG, RUNX1, STAT1, STAT2, VDR) that appeared in any of the GRaNIE inferred eGRNs. Then, I assessed the overlap between the GRaNIE inferred TF-associated peaks (regardless of whether they were also linked to a gene) and the corresponding ChIP-seq peaks (within a 50bp range). To determine the significance of this overlap, I used Fisher's exact test and compared it to a background set of ATAC-seq peaks that only included the TF motif. I decided to exclude two transcription factors, SPI1 and CTCF, due to an observed discrepancy in their connection counts. Specifically, at FDR thresholds of 0.4 or 0.5, these TFs displayed over 10,000 connections in the GRaNIE analysis. However, when I applied lower FDR thresholds, these connections disappeared, indicating a potential noise issue.

Molecular analysis of Peak-Gene connections inferred by GRaNIE using eQTL data

For peak-gene evaluation, cis-eQTL data were collected from the eQTL catalog, specifically targeting monocytes and macrophages. Six distinct datasets were selected for this analysis, and the eQTLs from these datasets were combined. A stringent selection criterion was applied, only considering associations with a permutation-based FDR below 0.3.

For each eGRN peak, the presence of eQTL single nucleotide polymorphisms (SNPs) was examined. A peak was deemed valid for inclusion in the study if it housed an eQTL SNP that impacted the same gene as identified in the GRN. To ascertain the meaningfulness of these peak-gene associations, they were compared against a randomly constructed background. This background consisted of links between GRN peaks and genes, randomly sampled and distance-matched based on 50 kb bins. This random sampling and comparison process was repeated 20 times to ensure reliability, with the odds ratio calculated between validated GRN links and those in the background.

The validity of peak-gene links across various FDR thresholds was assessed by analyzing their enrichment for four different macrophage GRNs. While additional eQTLs closely linked to the primary eQTLs were not included to avoid redundancy, the diverse range of datasets ensured

multiple variant inclusions per peak. This approach helped in identifying the most significant genetic variants influencing gene expression in these specific cell types, contributing to a robust analysis.

3.4.6 GRN benchmarking against other networks/tools

Dorothea (Garcia-Alonso et al., 2019; Holland et al., 2020) is a resource that lists interactions between transcription factors (TFs) and their targets. These connections are ranked based on how reliable they are, with rankings from A (very reliable) to E (based only on computational methods).

TRRUST (Han et al., 2018) is a database of TF regulatory networks for humans and mice, created through text mining and careful manual review. The human network in TRRUST v2 includes 795 TFs, 2,067 genes, and 8,427 links. In my study, I didn't apply any specific weight to these connections as TRRUST v2 does not provide this information.

ChEA3 (Keenan et al., 2019) compiles libraries of TF-target genes. These libraries include targets identified by ChIP-seq experiments from various sources like ENCODE and ReMap, as well as RNA-seq data-based co-expression connections from resources like GREx and ARCHs4.

ANANSE (Xu et al., 2021) offers an enhancer-based, cell-type-specific network useful for identifying key TFs in cell fate decisions. We utilized the ANANSE network specifically tailored for macrophages, selecting links with a probability threshold of 0.8.

3.4.7 Visualisation (Shiny App)

I've created a web application for the eGRNs I constructed in this study, using a Shiny app framework. The main purpose of this app is to visualize specific parts of the network that users are interested in. It allows for the selection and examination of the regulon for a particular TF of interest, although any part of the network can be visualized. Users can filter the network based on their chosen gene, enhancer, TF, or SNP, and adjust various thresholds for filtering connections from TF to peak or peak to gene. This makes it a versatile tool for exploring the intricate details of gene regulatory networks. The app is designed to be user-friendly and is available for use at https://apps.embl.de/grn/.

Chapter 4

Active repression of alternative cell fates safeguards hepatocyte identity and prevents liver tumorigenesis

In the next chapter, I'll investigate how certain transcription factors are crucial for preserving cell identity. This work is a joint project with Dr. Mortiz Mall from DKFZ, with the experimental part carried out by Dr. Bryce Lim and Dr. Juan Segarra. The text in this chapter was originally written by me, has been proofread by large language model-based tools, and is adapted from:

Aryan Kamal*, Bryce Lim**, Juan M. Adrian Segarra, Ignacio L. Ibarra, Borja Gomez Ramos, Kai Volz, Mohammad Rahbari, Nuh Rahbari, Eric Poisel, Kanela Kafetzopoulou, Lio Böse, Suchira Gallage, Jose Efren Barragan Avila, Hendrik Wiethoff, Ivan Berest, Sarah Schnabellehner, Taija Mäkinen, Darjus F. Tscharhaganeh, Mathias Heikenwalder, Judith B. Zaugg*, Moritz Mall** *Active repression of cell fate plasticity by PROX1 safeguards hepatocyte identity and prevents liver tumorigenesis* manuscript is under revision.

4.1 Introduction

The development and maintenance of cell identity in complex organisms depend on precise gene regulation, primarily controlled by transcription factors (TFs) (Mannervik et al., 1999). These TFs identify and attach to specific DNA sequences in gene promoters or enhancers, regulating gene expression. This process involves both the activation of genes specific to a particular cell type and the suppression of genes linked to other cell fates (Vos, 2021). Human cells contain around 1,600 types of TFs, with each cell expressing several hundred to manage cell fate induction and preservation (Vaquerizas et al., 2009).

While numerous TFs contribute to gene regulation in any given cell, only a select few, known as "selectors" or "master regulators" are capable of initiating specific cell lineages (Hobert, 2008). These TFs not only trigger cell fate but also continue to influence mature cells, ensuring

lasting cell identity. Some TFs, known as "pioneer" factors, can activate cell-specific genes in diverse cellular contexts (Balsalobre and Drouin, 2022). However, these factors often act broadly, sometimes activating genes of multiple cell types, which can lead to unintended cell identities. This observation underscores the importance of not just activating cell-specific genes, but also repressing genes related to other lineages (Treutlein et al., 2016).

Recognizing the significance of transcriptional repression in cell identity, my collaborator Mortiz Mall identified a new type of TF termed 'safeguard repressor.' Unlike other repressors, safeguard repressors like MYT1l are highly specific to certain cell types and act to repress genes related to alternative cell fates, thus ensuring the stability of the desired cell identity (Mall et al., 2017). Based on this study, my collaborator Ignacio Ibarra developed a computational method to identify safeguard repressors across various cell types. Utilizing this method, my collaborators and I have shown that PROX1 acts as a safeguard repressor in hepatocytes. It enhances their identity by suppressing genes associated with other cell fates. This discovery underscores the role of repression in maintaining cell fate and suggests the potential of safeguard repressors in disease prevention through the preservation of cell identity.

4.2 Results

4.2.1 Uncovering cell-specific safeguard repressors in eighteen cell types

In this study, my collaborator and I identified a new category of TFs known as safeguard repressors. These TFs are predominantly and specifically expressed in a particular cell type. Their targets, identified based on motifs in the gene promoters, are typically markers and specific genes of other cell types. This characteristic enables these repressor TFs to block alternative cell fates and safeguard the desired one. An example of such a TF is MYT1L, previously recognized for promoting neuronal fate by suppressing other cell fates (Mall et al., 2017). The concept of a safeguard repressor depends on two important criteria: the TFs' specific and predominant expression in the desired cell type, and their target binding motifs being present in other cell types, thereby preventing the development of alternate cell fates.

To computationally identify candidates with the potential to act as safeguards for specific cell types, a screening methodology was developed by Dr. Ignacio Ibarra. This screening approach involves analyzing the expression of TFs across various cell types, searching for those with a high z-score in their median expression. Alongside, it examines the presence of binding motifs in the promoter regions of genes, where these target genes should be specifically and highly expressed in cell fates other than the desired one. In simpler terms, the expression of these TFs should be enriched in the desired cell type, while their binding motifs should be absent or depleted, as shown in the Figure 4.2.1.

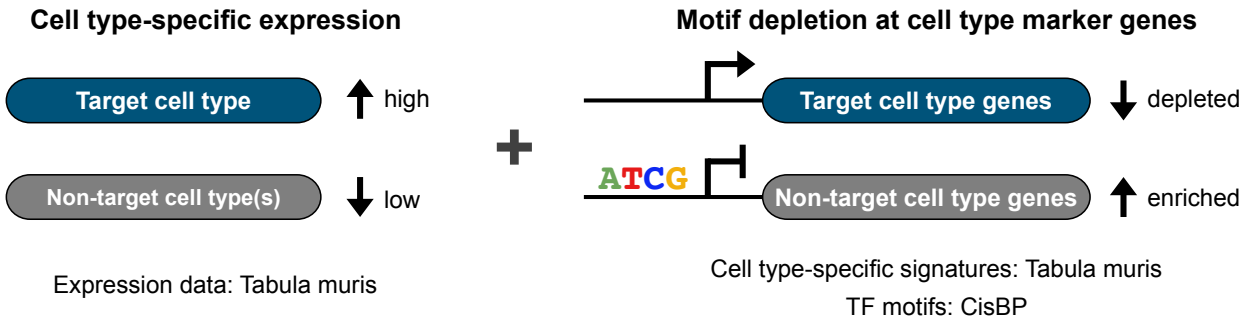

Figure 4.2.1. Screening for cell type-specific TFs with safeguard repressor potential
Transcription factors that are highly expressed in a particular cell type and show reduced motif presence in other cell types' specific promoter regions achieve increased safeguard repressor scores.
This figure was provided by Dr. Bryce Lim and and is part of a manuscript in (Kamal, Lim et al.).

This screening included 18 well-known cell types from the three primary layers of cells in an embryo: ectoderm, mesoderm, and endoderm. From the ectoderm, neurons, astrocytes, skin cells, epithelial cells, keratinocytes, and oligodendrocytes were included. The mesoderm contributed B cells, heart muscle cells, blood vessel cells, fibroblasts, granulocytes, microglia, monocytes, muscle stem cells, and T cells. Finally, from the endoderm, bladder cells, pancreatic beta cells, and liver cells were part of the study.

The Tabula Muris project (Schaum et al., 2018) provided the basis for this computational screening, offering single-cell gene expression data across 18 different cell types from 12-week-old mice. The screening commenced with an analysis of the expression of 1,296 TFs across these cell types, with a z-score calculation based on their median expression. A potential safeguard candidate for a specific cell type should exhibit a high z-score in comparison to the other 17 cell types. Subsequently, a set of 1,000 highly expressed genes was selected for each of the 18 cell types. A TF qualifying as a potential safeguard candidate for a specific cell type should demonstrate enriched binding motifs in the other 17 cell types and depleted presence in the targeted cell type. Ultimately, the final normalized score was calculated, scaling the z-scores for expression and motif (ranging from -1 to 1) for all TFs in each cell type. Candidates with a safeguard score above 0 were identified as potential safeguard repressors.

In addition to the analysis, I have created a website that serves as a search tool for this screening database. This website, based on a Shiny app, allows users to explore the data in various ways. Users can select a specific cell type and view scores for all the TFs associated with it. The website highlights the top candidate TFs for each cell type, making it easier to identify potential safeguard repressors.

Moreover, the website is designed to accommodate searches based on a user's specific TF of interest. This feature enables users to determine if their chosen TF ranks as a top potential safeguard for any of the cell types included in the study.

Candidate safeguard repressors in 18 cell types

In the subsequent phases of the study, 59 potential safeguard repressor candidates were identified, each selected based on their respective safeguard scores. These candidates represent the top six from each of the 18 cell types, chosen for their minimum qualifying scores, and are showcased in the accompanying Figure 4.2.2. Through my analysis of the TRRUST v2 dataset, complemented by a thorough literature review, it was determined that 50 of these TFs function either as repressors or possess dual roles as both activators and repressors. Furthermore, 33 of these TFs exhibit lifelong expression, a characteristic identified using the Tabula Muris senis (Schaum et al., 2018) dataset, which includes data from 27 mice approximately two years of age. Notably, 27 of these TFs meet both the criteria of being repressors and having lifelong expression.

Research indicates that 14 out of these 27 candidates actively promote their respective safeguarded cell fates during development or reprogramming. Notably, some of these factors function as repressors in specific contexts, such as TBX5 in cardiomyocytes and OLIG2 in oligodendrocytes (Waldron et al., 2016; Zhou et al., 2001). The screening process also confirmed MYT1L as a safeguard repressor, in line with prior experimental findings. These findings suggest that MYT1L helps maintain neuronal identity by suppressing non-neuronal genes (Mall et al., 2017) Figure 4.2.2.

Additionally, through a literature review conducted by my collaborator Dr. Bryce Lim, it was discovered that 16 of the 27 candidates exhibit tumor suppressor effects in their respective cell types. This trait was particularly prominent among endodermal candidates, especially those associated with hepatocyte cell fate, as demonstrated in Figure 4.2.2.

4.2.2 Hepatocyte safeguard repressor candidates

Next step, I focused on the critical impact of cell fate loss in liver diseases, particularly how it contributes to dedifferentiation and transdifferentiation in the emergence of primary liver cancer (Seehawer et al., 2018). Recognizing this, I analyzed the expression patterns of hepatocyte-specific candidates in hepatocellular carcinoma (HCC) patients, using data from The Cancer Genome Atlas (TCGA).

I then concentrated on the top six hepatocyte candidates, as highlighted in the Figure 4.2.3ab. Analysis of TCGA data indicated that high expression of PROX1, KLF15, ONECUT2, and ZNF771 in HCC patients is associated with longer survival (Method Section 4.4.1) Figure 4.2.3c, suggesting a more favorable outcome. This finding led me to hypothesize that these factors might function as tumor suppressors in liver cancer development. Furthermore, a lifelong expression analysis based on data from (Cardoso-Moreira et al., 2019), carried out by Dr. Bryce Lim, revealed that only Prox1, Onecut2, and KLF15 exhibit lifelong expression Figure 4.2.3d. Combining these insights effectively narrowed down our focus to these three candidates, which not only demonstrate

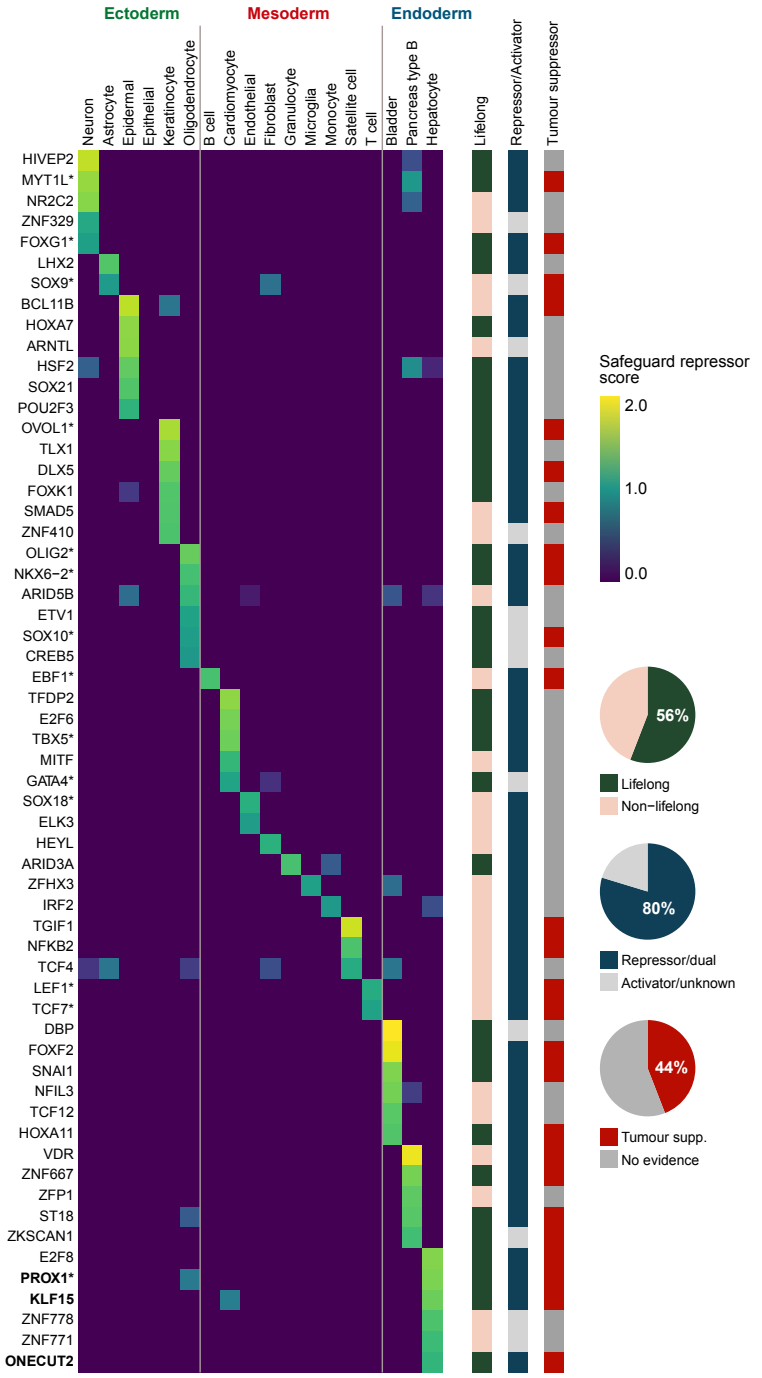

Figure 4.2.2. Top safeguard repressors in 18 cell types with lifelong expression and functional roles
Six leading safeguard repressor candidates were identified in eighteen cell types, each with a safeguard repressor score above 0. This includes data on persistent expression from Tabula Muris Senis (Schaum et al., 2018), roles as repressors/activators, and tumor-suppressing functions from existing studies.
This figure was provided by Dr. Bryce Lim and is part of a manuscript in (Kamal, Lim et al.).

lifelong expression but also show potential tumor-suppressing capabilities based on my survival analysis.

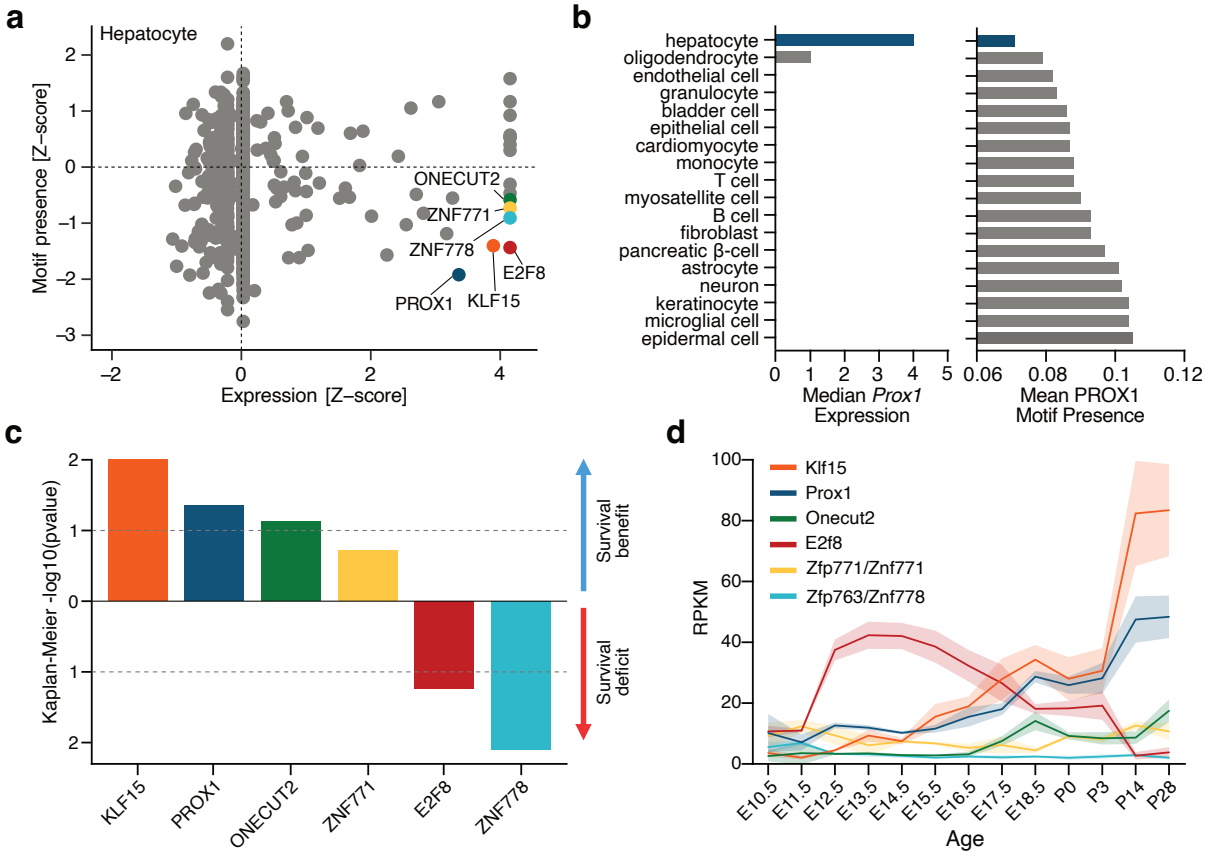

Figure 4.2.3. Narrowing down hepatocyte safeguard candidates with tumor suppressor role and lifelong expression
(a) Analysis of expression and motif presence in 1,296 transcription factors identifies six key candidates as hepatocyte safeguard repressors.
(b) On the left, the chart shows Prox1 expression across 18 cell types from Tabula Muris (Schaum et al., 2018) On the right, the graph displays the count of PROX1 motifs in the promoter regions of genes specific to each cell type.
(c) Evaluation of survival impact for liver safeguard repressor candidates, using a log rank test on Kaplan-Meier survival curves of hepatocellular carcinoma patients from TCGA. This analysis distinguishes between high and low expression levels of each candidate.
(d) Chart depicting the expression of the six foremost hepatocyte repressor candidates during liver development in mice (Cardoso-Moreira et al., 2019)
Part d of this figure was contributed by Dr. Bryce Lim, while the remaining sections were provided by myself. The complete figure is included in (Kamal, Lim et al.).

Validation of hepatocyte safeguard repressor candidates

The effects of the three primary safeguard repressor candidates on hepatocyte identity were examined in the context of cell fate reprogramming (Dr. Bryce Lim did the experiment). Prior research, demonstrated that mouse embryonic fibroblasts (MEFs) could be reprogrammed into induced hepatocytes by overexpressing four liver transcription factors Foxa3, Gata4, Hnf1a, and Hnf4a (4in1) (Song et al., 2016). Consequently, the impact of overexpressing Prox1, Onecut2, or Klf15 during hepatocyte reprogramming was evaluated, as shown in Figure 4.2.4a.

All the factors under consideration exhibited similar expression levels, yet Prox1 stood out with a significantly more effective hepatocyte reprogramming pattern Figure 4.2.4b. This was

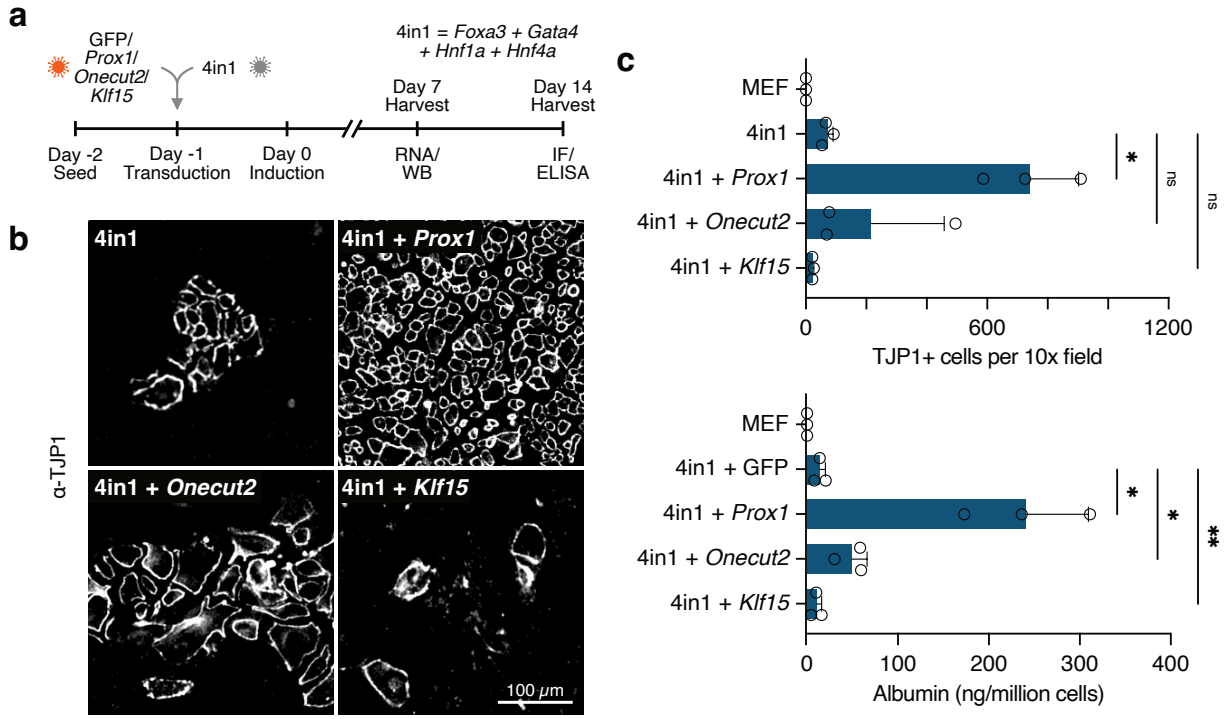

Figure 4.2.4. Validation and effects of key hepatocyte repressor candidates in reprogramming
(a) Verification of safeguard repressor candidates through lentiviral overexpression in MEFs during hepatocyte reprogramming using the 4in1 method.
(b) Display of TJP1 immunofluorescence in hepatocytes induced after overexpressing the top three hepatocyte repressor candidates, compared to a GFP control.
(c) Analysis of TJP1-positive cells through immunofluorescence (upper panel) and albumin levels assessed by ELISA (lower panel) at day 14 of hepatocyte reprogramming with selected candidates.
This figure was provided by Dr. Bryce Lim and and is part of a manuscript in (Kamal, Lim et al.).

determined by assessing several hepatocyte-specific protein markers, such as albumin, as shown in the Figure 4.2.4c. Moreover, overexpression of Prox1 led to a notably higher increase in the number of hepatocyte-like cells, evaluated based on their morphology and TJP1 expression Figure 4.2.4c. The outcomes of these experiments suggest that Prox1 significantly improved the efficiency of hepatocyte reprogramming. These results, coupled with additional evidence from my TCGA survival analysis and lifelong expression analysis, strongly indicate that Prox1 is a promising candidate as a hepatocyte safeguard. Consequently, this prompted me to delve deeper into its potential role as a liver tumor suppressor.

4.2.3 Association of PROX1 expression with increased survival in HCC and reduction in cancer cell proliferation

Various studies present conflicting views on PROX1, with some suggesting it acts as a tumor suppressor, while others indicate it may promote tumor growth (Liu et al., 2013; Shimoda et al., 2006). Therefore, I and my collaborators decided to investigate the role of Prox1 in HCC as a potential hepatocyte safeguard. In the initial phase of the study, PROX1 expression levels were compared

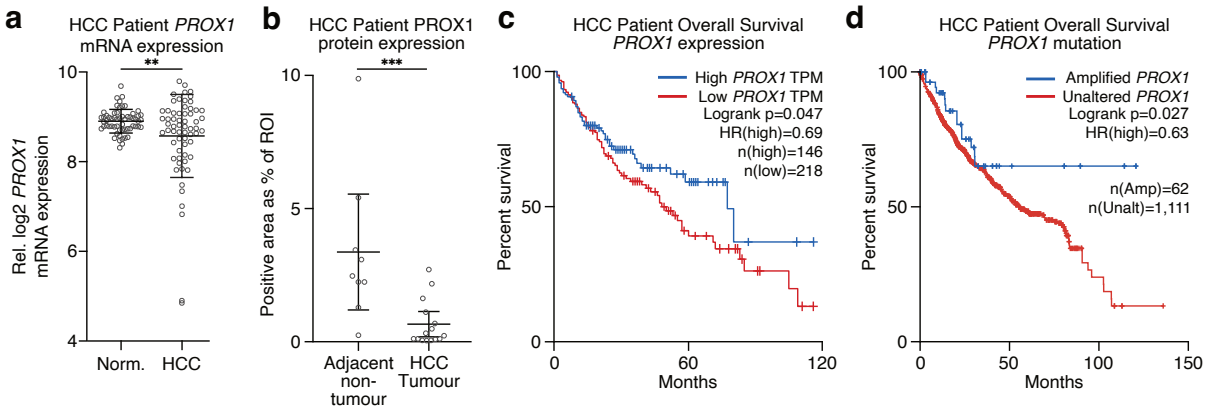

Figure 4.2.5. Effects of PROX1 in HCC
(a) Analysis of PROX1 gene expression in tumor and paired normal liver tissues from 62 HCC patients, based on the TIGER dataset (Chaisaingmongkol et al., 2017).
(b) Comparison of PROX1 protein levels in tumor versus adjacent non-tumor liver sections from HCC patients. Statistical significance was assessed with the Mann-Whitney test, *** $p < 0.001$.
(c) Survival analysis of 364 HCC patients classified by PROX1 expression levels, with a 40% cutoff for the high-expression group (Menyhárt et al., 2018). Evaluated using the log-rank test.
(d) Survival comparison of 1,173 HCC patients based on PROX1 mutation status, distinguishing between chromosomal amplification including PROX1 and no alteration (Ahn et al., 2014; Cerami et al., 2012; Gao et al., 2013). Analysis was performed using the log-rank test.
This figure was provided by Dr. Bryce Lim and is part of a manuscript in (Kamal, Lim et al.).

between patient samples and corresponding healthy liver tissue controls. This comparison revealed that PROX1 expression is significantly higher in normal liver tissues Figure 4.2.5a, aligning with its potential role as a tumor suppressor (Chaisaingmongkol et al., 2017). Subsequently, a comparison of PROX1 protein levels between HCC tumors and adjacent non-tumor liver tissues was conducted. Consistent with Prox1's tumor-suppressing activity, the protein levels were found to be higher in adjacent non-tumor tissues Figure 4.2.5b.

To further reinforce PROX1's role as a tumor suppressor, another survival analysis involving a cohort of 364 patients with available transcriptome and survival data was performed by Dr. Bryce Lim (Menyhárt et al., 2018). This analysis indicated that higher PROX1 expression correlates with a median survival of 81.9 months, compared to a significantly reduced median survival of 47.7 months in cases with low PROX1 expression, as shown in the Figure 4.2.5c.

In the final stage of the analysis, a survival study was conducted by Dr. Bryce Lim, specifically examining the impact of chromosomal amplifications that include PROX1. This study revealed that such amplifications in PROX1 are associated with increased survival rates in HCC patients (Ahn et al., 2014) (Cerami et al., 2012) (Gao et al., 2013). This significant correlation underscores the potential tumor-suppressing role of PROX1 in liver cancer and is illustrated in Figure Figure 4.2.5d.

4.2.4 Inhibition of HCC development and progression by Prox1 in Mice

In order to explore how PROX1 levels impact the behavior of human liver cancer cells, an examination was conducted using an HCC cell line (Hep3B) *in vitro* (The experiment was done by Dr. Bryce Lim). This involved creating Hep3B cell lines with the capability to either increase (overexpress) or decrease (knockdown) PROX1 levels through inducible constructs.

It was found that overexpressing PROX1 levels in these cells led to a notable decrease in their proliferation rate, with a reduction of up to 60%. Conversely, when PROX1 levels were reduced using shRNA (short hairpin RNA), a significant increase in cell proliferation was observed, as shown in Figure 4.2.6a.

The research was directed at elucidating the molecular mechanisms responsible for the antiproliferative effect of PROX1, particularly its influence on chromatin organization. To explore whether PROX1 contributes to the condensation of chromatin in HCC cells, ATAC-seq experiment was conducted by Dr. Bryce Lim. I analyzed the ATAC-seq Section 4.4.3 data using the DiffBind tool and found that on the second day, there were 4,353 peaks with differential accessibility (padj < 0.05). Notably, 4,074 of these peaks, representing 93.6%, showed reduced accessibility or 'closure' when PROX1 was overexpressed, compared to a GFP control Figure 4.2.6b.

To investigate into the downstream effects of PROX1 in cancer, I utilized GRaNPA (See Chapter 2). Specifically, I worked with RNAseq data from the overexpression of PROX1 at days 2 and 7, as well as from the knockout of PROX1 at day 7. Following quality control and preprocessing, I calculated differentially expressed genes between overexpression (OE) or knockout (KD) versus GFP for each timepoint. Next, I merged ATACseq data with cut&run (which was also produced by Dr. Bryce Lim) and PROX1 motif analysis to identify a set of potential target genes for PROX1. These targets were required to have a cut&run peak within 1500 bases of their transcription start site (TSS), and this peak had to be differentially accessible on day 2 according to ATACseq data. This process helped me define the cancer PROX1 regulon.

For predicting differential expression using GRaNPA, a GRN is needed. Since I knew these variations stemmed from OE or KD of PROX1, I constructed a GRN using the PROX1 regulon identified via ATAC and cut&run, supplemented with regulons for TFs connected to PROX1 identified using DoRothEA Garcia-Alonso et al., 2019. With this GRN centered around PROX1, and differential expression data for days 2 and 7 of overexpression and day 7 of knockout, I ran GRaNPA. This analysis revealed the top 5 important TFs in each condition Figure 4.2.6c. Among these, only PROX1 and MYC consistently emerged as the most important TFs explaining the variation between GFP and OE/KD conditions. The regulon change of MYC, based on the DoRothEA regulon, aligned perfectly with the expected direction for both OE and KD, as shown in Figure 4.2.6d. This finding confirms that PROX1 reduces proliferation in cancer by directly targeting MYC.

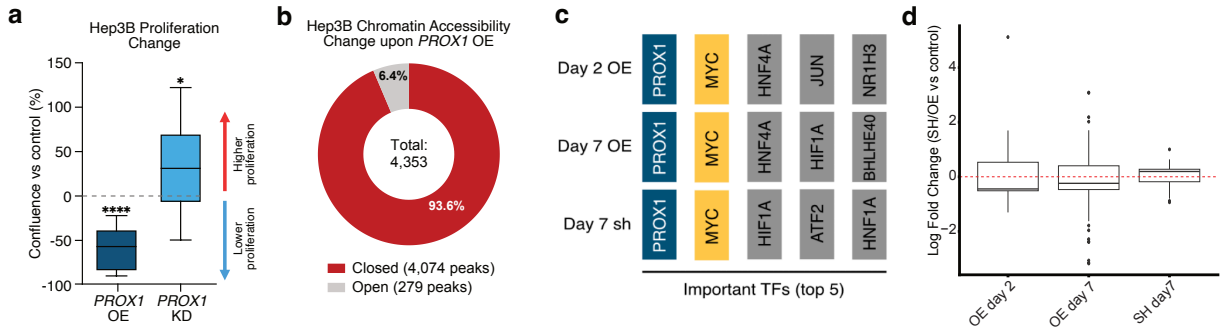

Figure 4.2.6. PROX1 impact on Hep3B cells with proliferation, chromatin accessibility, and transcriptional regulation analysis
(a) Proportion of confluence in Hep3B cells after 7 days, following the induction of PROX1 knockdown (KD) or overexpression (OE), normalized to uninduced controls. Statistical significance assessed with a one-sided t-test, using 0 as the theoretical mean.
(b) Comparison of chromatin regions that became more or less accessible two days after PROX1 overexpression in Hep3B cells, as identified by ATAC-seq, relative to control conditions.
(c) The top 5 most influential TFs identified by GRaNPA in each condition, not ranked. Notably, Myc and Prox1 are the only TFs common to all three conditions.
(d) The log2 fold change in Myc regulon across different time points and conditions.
This Figure was contributed by Dr. Bryce Lim(parts a,b) and myself and is part of a manuscript in (Kamal, Lim et al.).

4.2.5 Prox1 suppresses multiple non-hepatocyte cell types

In addition to its roles in development and cancer, cell fate plasticity is crucial in other transitions, like regeneration after injury or direct cell reprogramming. I was interested in exploring if PROX1 could universally regulate cell fate plasticity in these contexts as well. In cases of liver injury, mature hepatocytes often undergo dedifferentiation, marked by the reactivation of progenitor-like programs, followed by proliferation and differentiation to restore functional hepatocytes (Li et al., 2023).

I re-examined single-cell data across pseudotime from a mouse model of chemical-induced liver injury (Li et al., 2023), which included stages before, during, and after injury. After preprocessing this data and conducting PCA and uniform manifold approximation and projection (UMAP), I employed Monocle3 (Trapnell et al., 2014) to determine cell trajectory across these timepoints, utilizing their cell annotations for more detailed insights.

My analysis revealed that Prox1 expression decreased following injury and then gradually increased during the regeneration of hepatocytes, as shown in Figure 4.2.7. Simultaneously, there was a notable reduction in hepatocyte identity, which I determined by calculating the z-score expression of a set of hepatocyte marker genes, during the dedifferentiation process triggered by an injury, followed by a rapid resurgence in their identity during the recovery phase. This observation indicates a link between the reduction of Prox1 levels and the dedifferentiation of hepatocytes. Yet, the question of whether overexpressing Prox1 expression could beneficially influence hepatocyte cell identity still needs further exploration.

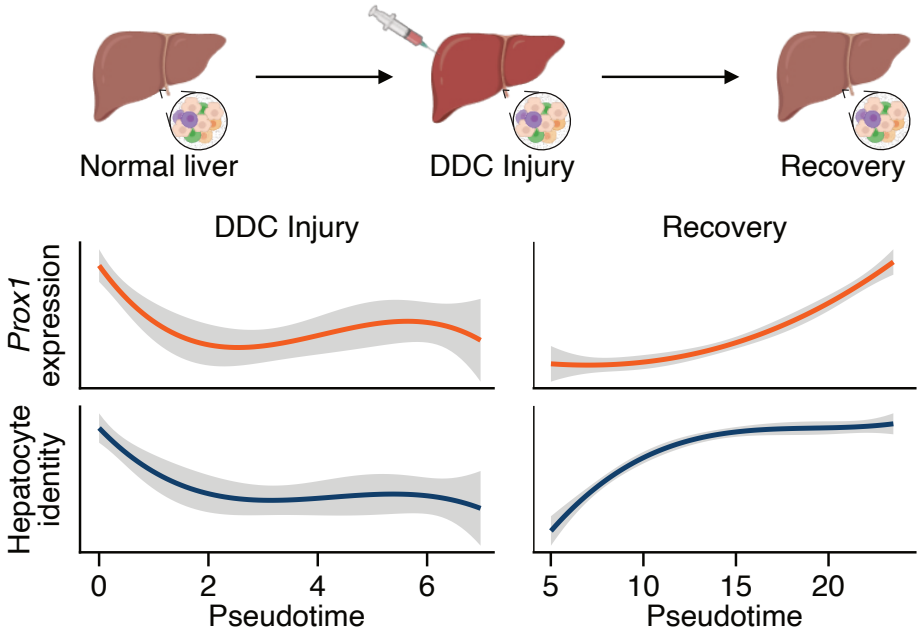

Figure 4.2.7. Prox1 expression trends in liver injury and regeneration
Analysis of Prox1 and hepatocyte marker gene expression throughout the stages of liver injury and regeneration in a mouse model using DDC induction, as per single-cell gene expression pseudotime analysis (Li et al., 2023)
This figure has been produced by myself and is part of a submitted manuscript (Kamal, Lim et al.)

4.2.6 Prox1 promotes hepatocyte reprogramming

To experimentally explore the potential of PROX1 in enhancing liver cell fate acquisition, the effects of Prox1 overexpression during hepatocyte reprogramming were evaluated. This involved conducting single-cell RNA sequencing (scRNA-seq) on MEFs both untreated and at various stages - days 2, 7, and 14 - of hepatocyte reprogramming induced by the 4in1 method. This process was performed alongside the overexpression of Prox1, with GFP serving as the control, as illustrated in Figure 4.2.8. The experiment was conducted by Dr. Juan Segarra and Dr. Bryce Lim.

Next, I began analyzing the data by demultiplexing the pooled samples, as detailed in the methods Section 4.4.2. After demultiplexing, it was necessary to exclude cells incorrectly labeled as Prox1 but not exposed to the virus, as well as those that had not received the 4in1 virus. To do this, I assessed the activity of Prox1 and 4in1 in each cell, ensuring their presence and activity. For Prox1, ATACseq and cut&run data were produced on day 2 for both Prox1 overexpression and GFP control (The experiment has been performed by Dr. Bryce Lim). Following the preprocessing of these datasets, I combined the Prox1 motif with cut&run peaks, selecting those within 1000 bases of the TSS of a gene and filtering them based on differential accessibility in the ATACseq data. This approach allowed me to establish a stringent set of target genes for Prox1. I then calculated Prox1 activity based on these targets, identifying cells labeled as Prox1 but not showing activity Section 4.4.2 and shown in the Figure 4.2.9a.

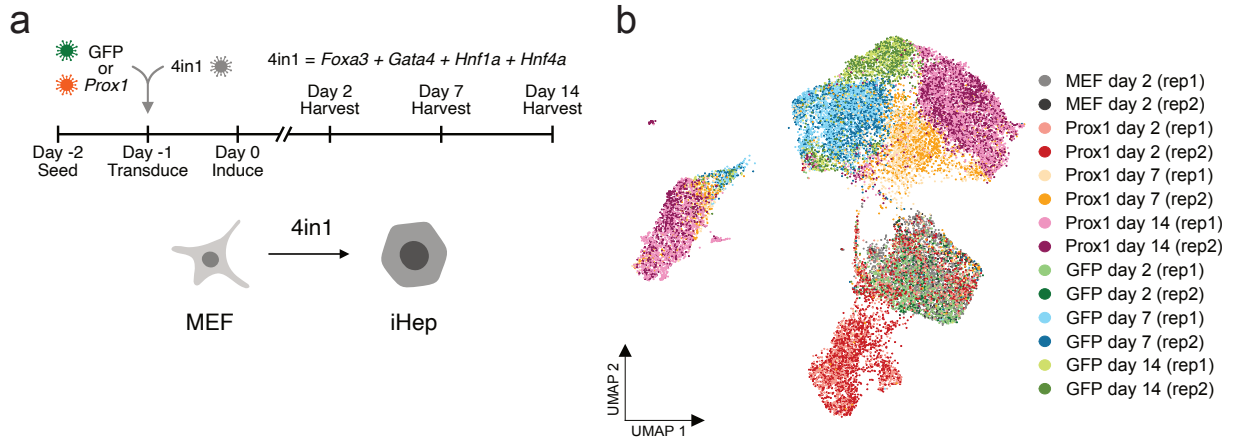

Figure 4.2.8. Prox1's role in hepatocyte reprogramming through single-cell RNA-seq analysis
Analysis of hepatocyte reprogramming over time with and without Prox1 intervention, as examined through single-cell RNA sequencing. This study includes data from 22,761 cells across two biological replicates, post-clustering and UMAP projection.
This figure has been contributed by Dr. Bryce Lim (part a) and myself and is part of a submitted manuscript (Kamal, Lim et al.).

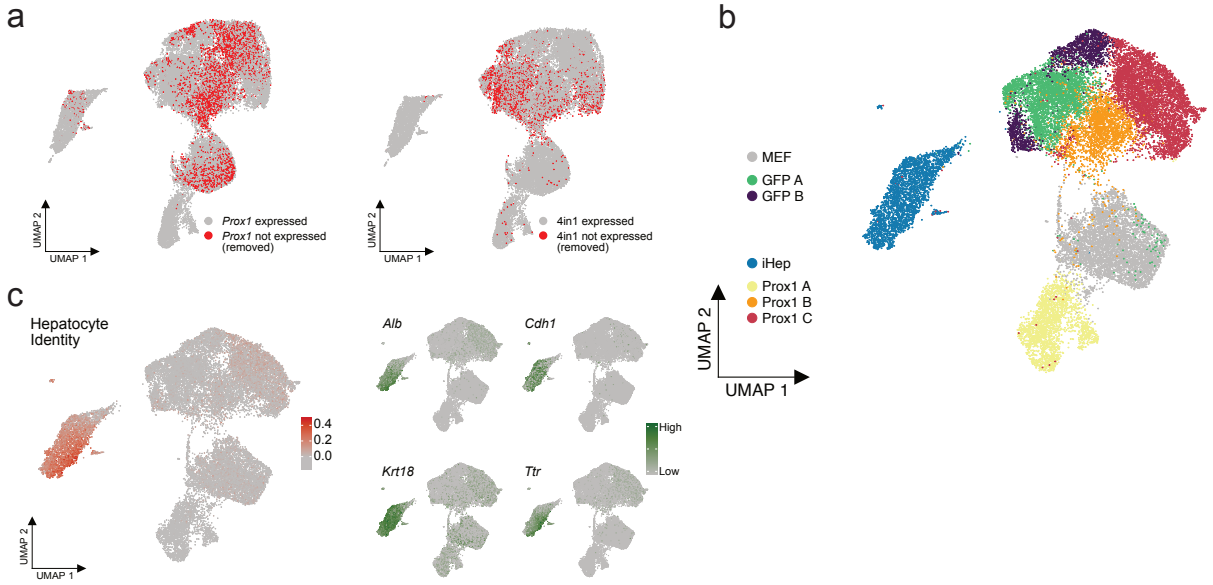

Figure 4.2.9. Cellular distribution and hepatocyte identity in reprogramming analyzed with UMAP for Prox1 and 4in1 activity
a) UMAP visualization of cells categorized by Prox1 expression and 4in1, deduced from the activity levels of PROX1 and 4in1.
b) UMAP representation of all cells, distinguished by their respective clusters.
c) Distribution of hepatocyte identity scores across cells on the UMAP, accompanied by the expression profiles of select hepatocyte marker genes in all cells.
This figure has been produced by myself and is part of a submitted manuscript (Kamal, Lim et al.).

For 4in1, lacking ATAC or cut&run data, I employed DoRothEA (Garcia-Alonso et al., 2019) for each of the four TFs, merging their targets to ascertain 4in1 activity. Subsequently, I removed cells lacking activity for either Prox1 or 4in1, except for MEF cells, which I retained in the dataset.

To minimize unwanted effects, I regressed out cycling genes, then normalized the data, and performed PCA and UMAP, followed by clustering (see Method Section 4.4.2) Figure 4.2.9b. To identify hepatocyte clusters, I gathered cell type marker genes from the Panglao database (Franzén et al., 2019) and created an exclusive set of markers for each of the 18 cell types, including MEFs and hepatocytes. By calculating a z-score identity score for each cell type for all cells, I was able to confirm the MEF cells, identify hepatocyte clusters, and gain insights into cells from other cell types Figure 4.2.9c.

Prox1 enhances hepatocyte identity and suppresses MEF identity

Next, I found that overexpressing Prox1 significantly boosted the number of successfully reprogrammed hepatocytes by more than sevenfold (2,527 Prox1 cells versus 347 control, $p < 2.2e\text{-}16$). Notably, this overexpression significantly enhanced hepatocyte cell identity at each time point, which I measured by assessing the expression of hepatocyte marker genes, as shown in Figure 4.2.10a.

Interestingly, cells with Prox1 overexpression not only downregulated the original MEF cell fate more effectively but also showed a more pronounced reduction in this regard Figure 4.2.10b. They also exhibited lower levels of markers associated with alternative cell identities, including fibroblasts, adipocytes, myocytes, and neurons, as detailed in Figure 4.2.10c-e. These findings strongly suggest that PROX1 plays a significant role in reinforcing hepatocyte cell fate while simultaneously suppressing gene expression programs related to fibroblasts and other cell types. This analysis led me to conclude the important role of PROX1 in guiding and stabilizing the identity of hepatocyte cells during reprogramming Figure 4.2.10.

Enhancement of liver cell fate through active suppression of alternative cell identities

I then examined how the Prox1 activity, based on the regulon I had previously identified, related to other cell fates. I found that Prox1 activity was significantly inversely correlated with the fibroblast identity score. This finding implies that Prox1 could suppress the original fibroblast cell fate Figure 4.2.11, When I expanded this analysis to include additional cell fates, it became evident that Prox1 activity was inversely related to all tested cell identities, with the exception of hepatocytes and oligodendrocytes, as shown in Figure 4.2.11

Interestingly, the activity of liver inducers 4in1 did not solely enhance hepatocyte identity. Instead, their activity also showed a positive correlation with several other cell identities, including astrocytes, epithelial cells, and cholangiocytes, as detailed in Figure 4.2.11. Importantly, these cell signatures were negatively associated with PROX1 activity. This suggests that PROX1 might suppress many non-hepatic gene signatures that could potentially be activated by the 4in1 factors, thereby focusing the cellular reprogramming process on achieving the desired hepatocyte fate.

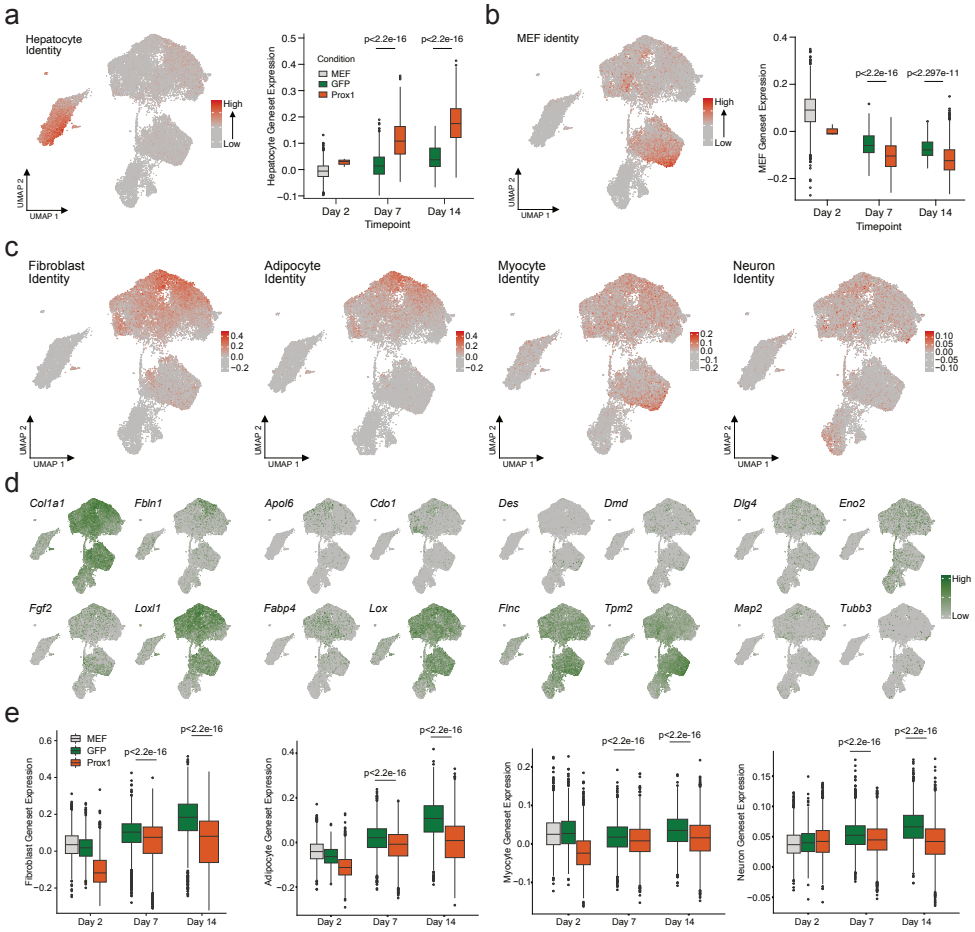

Figure 4.2.10. Quantifying cell identity scores and projecting them through UMAP in hepatocyte reprogramming
(a, b) Boxplot representation of hepatocyte (a) and fibroblast (b) identity scores within the hepatocyte cluster, including p-values from a two-tailed t-test for each timepoint and treatment. UMAP visualization of their identity scores is also presented.
c, d, f) UMAP projection of fibroblast, adipocyte, myocyte, and neuron identity scores. Display of specific marker gene expression for fibroblasts, adipocytes, myocytes, and neurons in all cells. Boxplot quantification of identity scores for fibroblasts, adipocytes, myocytes, and neurons, accompanied by p-values from two-tailed t-tests for each timepoint and treatment.
This figure has been produced by myself and is part of a submitted manuscript (Kamal, Lim et al.).

Prox1 inhibits alternative neuronal and muscle cell reprogramming processes

Subsequently, Prox1's influence on the development and reprogramming of other cell types was explored. For this, reprogramming experiments targeting neuronal and myocyte fates were designed. Specifically, the neuronal fate reprogramming experiment, conducted by Dr. Bryce Lim, involved transforming MEFs into neurons by overexpressing Ascl1 (Chanda et al., 2014). Results from this experiment indicated that co-overexpressing Prox1 with Ascl1 substantially lowered the reprogramming efficiency. This was evident from the observed decrease in TUBB3 protein level, as illustrated in the Figure 4.2.12ab.

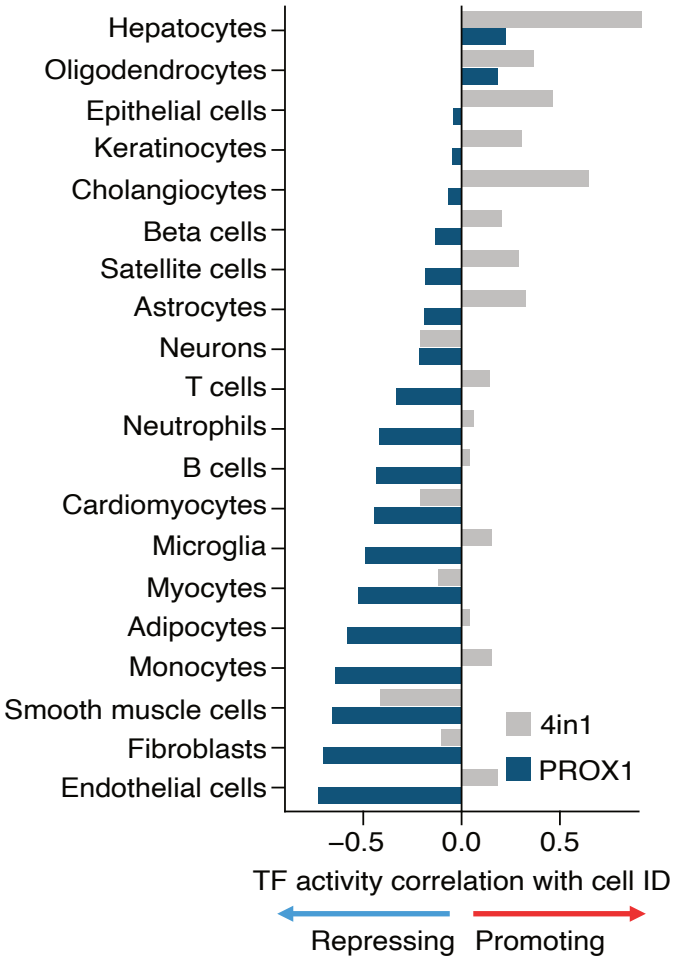

Figure 4.2.11. Assessing the relationship between cell identities and 4in1/Prox1 activities
Analysis of the correlation between diverse cell identity scores and the activities of 4in1 or Prox1
This figure has been produced by myself and is part of a submitted manuscript (Kamal, Lim et al.).

Following observations from hepatocyte reprogramming, where a decrease in the muscle marker protein Desmin was noted, it was hypothesized that Prox1 might influence myocyte reprogramming. To test this, an experiment, conducted by Dr. Bryce Lim, was devised to reprogram MEFs into myocytes via the overexpression of Myod1 (Davis et al., 1987). As anticipated, introducing Prox1 alongside Myod1 significantly reduced, and almost entirely eliminated, Desmin expression Figure 4.2.12a-b. This suggests that Prox1 overexpression can profoundly modify the reprogramming process towards myocytes. These effects of Prox1 align with previous studies that have highlighted MYOD1's versatile activity (Lee et al., 2020), suggesting that safeguard repressors like Prox1 or Myt1l can potentially guide such activity towards a preferred cell fate. In conclusion, Prox1's role extends beyond merely suppressing alternative cell fates in hepatocyte reprogramming; its overexpression in neuronal or myocyte reprogramming contexts can also significantly alter the process and counteract the effects of master regulators such as Ascl1 or Myod1.

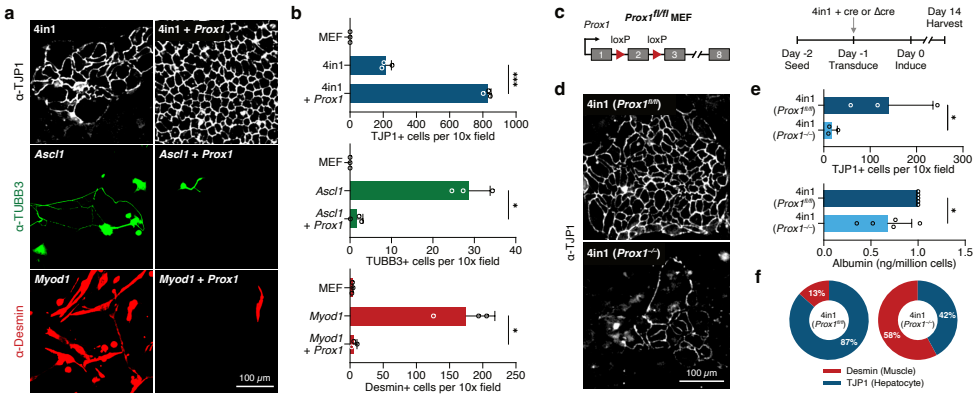

Figure 4.2.12. Impact of Prox1 on reprogramming MEFs to hepatocytes, neurons, and myocytes
(a) Conversion of MEFs to induced hepatocytes (top), neurons (middle), or myocytes (bottom) using overexpression of 4in1, Ascl1, or Myod1, respectively. Displayed are representative immunofluorescence images showing TJP1 (hepatocyte), TUBB3 (neuronal), or Desmin (myocyte) marker proteins on day 14 of reprogramming, both with and without Prox1 overexpression.
(b) Quantification of immunofluorescence for cells in (a) across three biological replicates.
(c) Prox1 genetic knockout during hepatocyte reprogramming achieved through cre-mediated deletion of exon 2 in Prox1fl/fl MEFs.
(d) TJP1 immunofluorescence observed on day 14 of hepatocyte reprogramming in Prox1−/− or Prox1fl/fl cells.
(e) Count of TJP1-positive cells from the analysis in (d) (n=3), and Albumin secretion measurement in cells from (d) following Prox1 deletion (n=5).
(f) Percentage of reprogrammed cells in (d) expressing either Desmin or TJP1.
This figure is provided by Dr. Bryce Lim and is part of a submitted manuscript (Kamal, Lim et al.).

The essential role of Prox1 in inhibiting alternative cell fates during hepatocyte reprogramming

In previous stages of the study, it was established that overexpressing Prox1 can significantly enhance hepatocyte reprogramming by suppressing alternative cell fates. Following this, the necessity of endogenous Prox1 expression for successful hepatocyte reprogramming was investigated. This phase involved a knockout experiment designed to study hepatocyte reprogramming, conducted by Dr. Bryce Lim Figure 4.2.12c. The results of this experiment demonstrated that the absence of Prox1 significantly impacts hepatocyte reprogramming Figure 4.2.12d. Specifically, the removal of Prox1 led to a considerable reduction in the number of TJP1-positive hepatocyte-like cells, as evidenced in the Figure 4.2.12e. At the same time, there was a notable decrease in the level of albumin protein when Prox1 was knocked out Figure 4.2.12e-f.

These findings collectively highlight that Prox1's overexpression is not just beneficial for more effective reprogramming and the production of better-quality hepatocyte cells. It also emerges as a crucial factor for efficient liver cell fate induction, with its knockout resulting in a less effective reprogramming process.

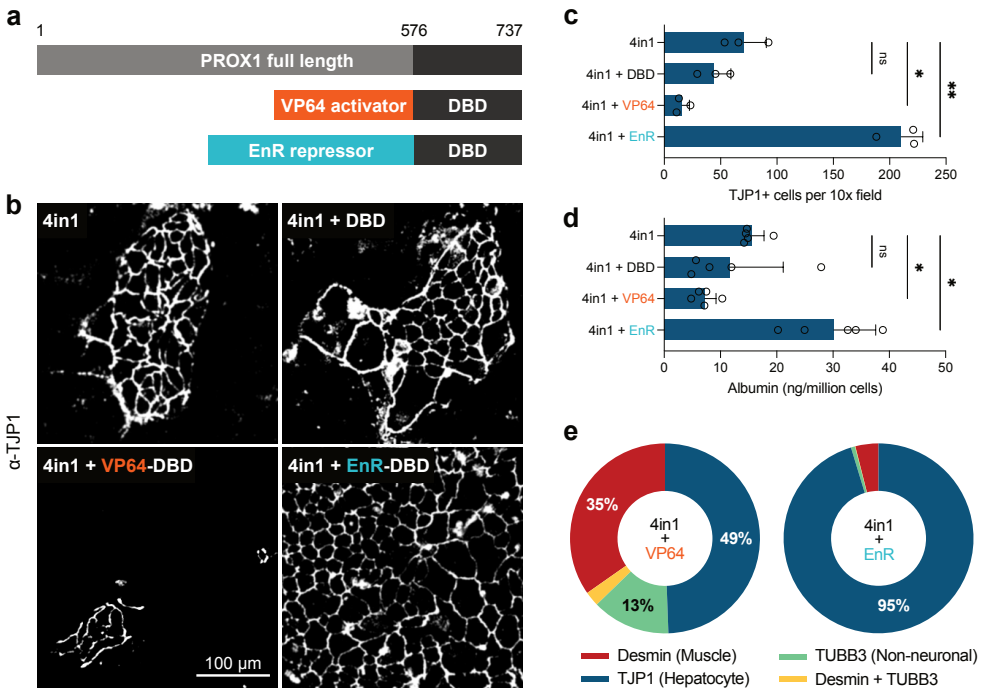

Figure 4.2.13. Evaluating Prox1 DBD fusions in hepatocyte reprogramming a) Diagram of fusion proteins comprising the PROX1 DNA-binding domain (DBD) linked with either the VP64 activator or EnR repressor domains (illustration not to scale).
b) Display of TJP1 immunofluorescence in hepatocytes reprogrammed using 4in1 and various PROX1 fusion constructs, observed on day 14.
c) Statistical analysis of the number of TJP1-positive induced hepatocytes produced in (b) across three biological replicates.
d) Measurement of Albumin secretion normalized to cell count in the experiment shown in (b), with data collected from five replicates.
e) Percentage of reprogrammed cells from (b) that express specific markers and exhibit morphology characteristic of the indicated cell types.
This figure has been provided by Dr. Bryce Lim and is part of a submitted manuscript (Kamal, Lim et al.).

4.2.7 Targeted inhibition of Prox1 genes promotes liver identity, while activation enables cell fate flexibility

Prox1, in mice, is notably present in specific neural stem cells within the hippocampus and cerebellum, where it contributes to neurogenesis (Karalay et al., 2011). It also plays a vital role in the development and sustenance of lymphatic endothelial cells (Petrova et al., 2002). In these contexts, Prox1 typically activates gene expression, often in collaboration with coactivators such as NR2F2, also known as COUP-TFII (Aranguren et al., 2013). In contrast, within liver environments, Prox1 has an association with the corepressor HDAC3 (Armour et al., 2017). The aim of this study was to differentiate between the effects dependent on cofactors and to ascertain the specific gene regulatory networks that Prox1 directly influences in the regulation of cell identity.

To manipulate the target genes of PROX1 directly, the DNA-binding domain (DBD) of PROX1 was merged either with a transcriptional activator (VP64) or with the Engrailed repressor (EnR).

Additionally, a control involving only the DBD without any effector domain was implemented Figure 4.2.13a. The investigation revealed that the sole DBD did not markedly affect hepatocyte reprogramming. However, the integration with the repressor mirrored the effects of the complete PROX1 protein, demonstrated by an uptick in both the number of TJP1-positive hepatocyte-like cells and Albumin secretion per cell Figure 4.2.13b-d. Inversely, the activator combination had a notably negative impact, considerably hindering hepatocyte reprogramming. Moreover, the ratio of cells reprogrammed into alternate neuronal or myocyte identities decreased with the repressor and increased with the activator Figure 4.2.13e. The experimental and analysis work for this part was carried out by Dr. Bryce Lim.

PROX1 represses alternative fate inducers

To examine how PROX1 affects chromatin organization during the reprogramming of hepatocytes, ATAC-seq was conducted (The experiment has been done by Dr. Bryce Lim). During the analysis of ATAC-seq data from day 2 of 4in1-induced hepatocyte reprogramming, I discovered 111,411 peaks showing differential accessibility (padj < 0.05) with Prox1 overexpression compared to the GFP control. Among these, 85,140 peaks (76.4%) exhibited reduced accessibility Figure 4.2.14a. I then analyzed chromatin remodeling at PROX1-bound target genes using CUT&RUN DNA-binding data for PROX1. This revealed a repressive signature and decreased accessibility at PROX1-bound sites with Prox1 overexpression, confirming its role in repression. Moreover, 74% of genes having a PROX1-bound and differentially accessible region in their promoters were downregulated upon Prox1 overexpression. These findings indicate that PROX1 primarily closes chromatin at its direct target genes, thereby reducing their expression in both fibroblasts and the initial stages of hepatocyte reprogramming.

To investigate the role of PROX1 in enhancing hepatocyte identity in the later stages of reprogramming and to pinpoint suppressed PROX1 target genes that could contribute to this process, a time-course transcriptomic study was conducted using bulk RNA sequencing. Dr. Bryce Lim has produced this data. I specifically examined differential gene expression during hepatocyte reprogramming with either GFP (as a control) or Prox1 overexpression at days 7, 14, and 28 Figure 4.2.14a. Across these time points, I identified 8,036 protein-coding genes showing differential expression between Prox1 and the control Section 4.4.4. Among these, 3,629 genes were consistently upregulated, and 3,264 were consistently downregulated across all three time points with Prox1 overexpression compared to the control.

Based on the promoter accessibility at day two and differential gene expression over time, I categorized these genes into two clusters Figure 4.2.14a. Cluster 1, which was upregulated throughout the four weeks and showed increased promoter accessibility at day two upon Prox1 overexpression (t-test, p < 1.487e-05) Figure 4.2.15a, included genes for hepatocyte identity but also some non-hepatic genes like cholangiocyte and oligodendrocyte markers Figure 4.2.14b. In contrast, cluster 2, downregulated throughout the 28 days of reprogramming and showing

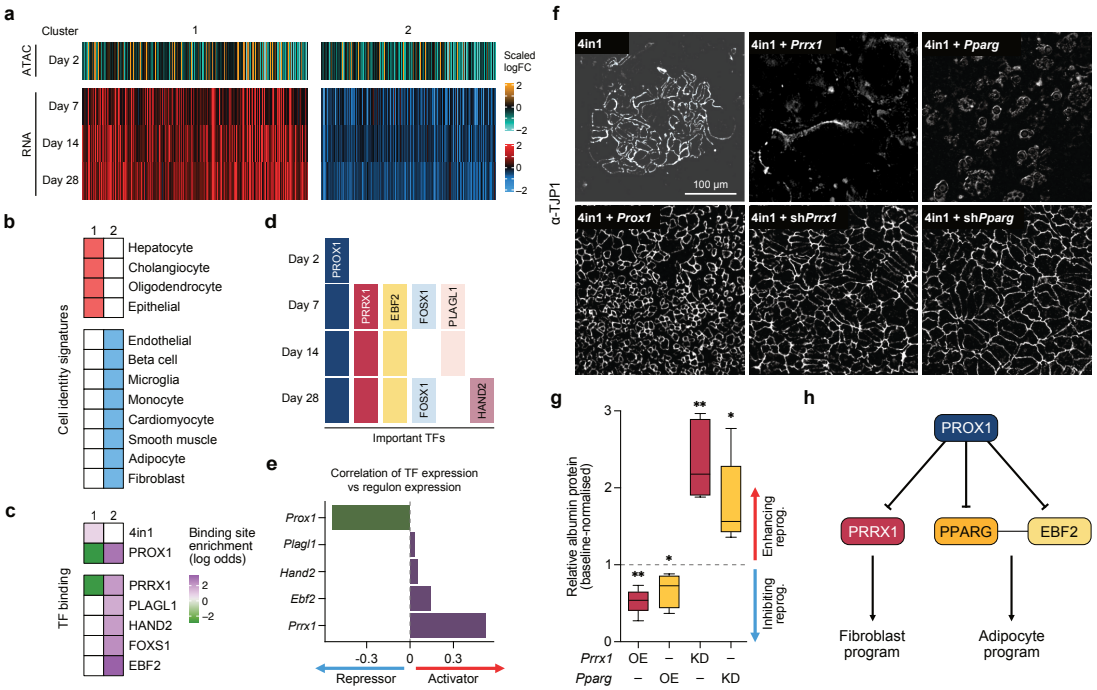

Figure 4.2.14. PROX1 influence on hepatocyte reprogramming including chromatin, gene expression, and TF networks

(a) Comparative analysis of chromatin accessibility and gene expression in hepatocyte reprogramming with and without Prox1 over various time points. Genes were clustered based on expression and chromatin accessibility, displayed as scaled logFC relative to control.

(b) Identification of common cell identity marker genes within each gene cluster, highlighting significant overlaps (p-adj < 0.01).

(c) Assessment of transcription factor binding enrichment or depletion, based on CUT&RUN (PROX1) or motif presence (other factors), in gene promoters within each cluster. Displayed are log2 odds ratios with significant values (p-adj < 0.05).

(d) Computational prediction of important transcription factors driving PROX1-enhanced hepatocyte reprogramming over time using GRaNPA.

(e) Correlation analysis between selected transcription factors and their target genes, aiding in predicting their activator or repressor roles.

(f) TJP1 immunofluorescence in hepatocytes reprogrammed using 4in1 with day 14 overexpression or shRNA-mediated knockdown of Prrx1 or Pparg.

(g) Quantitative analysis of Albumin protein levels in cells from (f) using Western blot. Data normalized to controls and presented in boxplots (n=5), Fisher's LSD test was applied for significance (* p-adj < 0.05, ** p-adj < 0.01).

(h) Proposed regulatory network by PROX1, illustrating how repression of downstream transcription factors can suppress alternative cell identities.

This figure was contributed by Dr. Bryce Lim (f,g) and is part of a submitted manuscript (Kamal, Lim et al.).

reduced promoter accessibility at day two (t-test, p < 6.637e-09) (Figure 4.2.14a; Figure 4.2.15a), was enriched in non-hepatocyte identity markers, such as neuron, muscle, fibroblast, and adipocyte genes Figure 4.2.14b.

During the reprogramming period, the expression levels of genes in cluster 2 increased in control cells but decreased upon Prox1 overexpression Figure 4.2.15a. Interestingly, genes from cluster 2 were mostly expressed in the PROX1-repressed alternate fate clusters in our single-cell dataset Figure 4.2.15b, while genes from cluster 1 were highly expressed in reprogrammed hepatocytes.

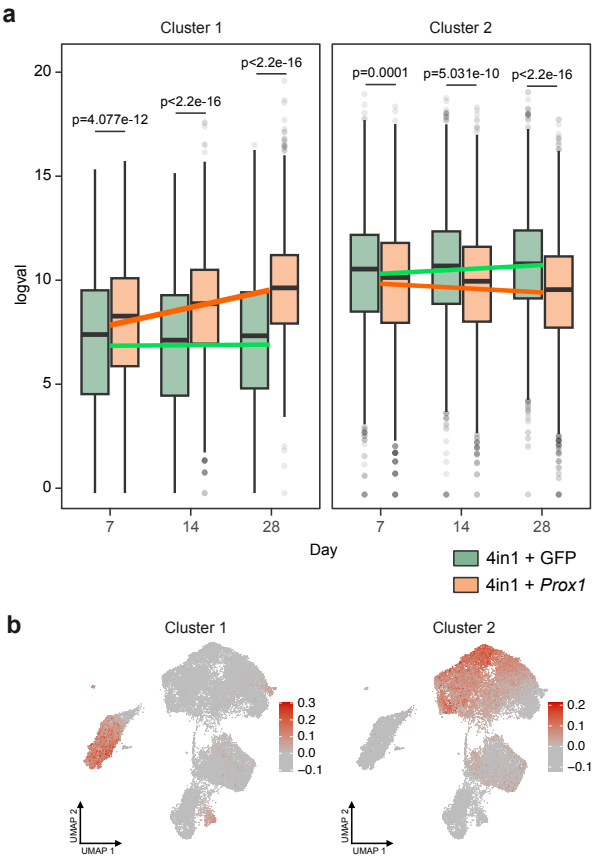

Figure 4.2.15. Gene expression dynamics and cluster association in hepatocyte reprogramming
(a) Analysis of gene expression changes in 4in1-induced hepatocyte reprogramming at specified time points, categorized by differential expression and chromatin accessibility. Boxplots depict the median gene expression of differentially expressed genes between Prox1 and GFP control across both clusters, with significance assessed by a two-tailed t-test (p-values indicated).
(b) Combined expression profiles of genes from clusters 1 and 2, derived from bulk gene expression data, superimposed on UMAP from 4in1-induced reprogramming single-cell transcriptomics. This illustration emphasizes the alignment of cluster 1 with the hepatocyte cluster and cluster 2 with clusters indicative of alternative cell fates.
This figure has been produced by myself and is part of a submitted manuscript (Kamal, Lim et al.).

I then sought to establish if any of these clusters were directly regulated by the overexpressed Prox1. For this, I retrieved the target genes of the liver reprogramming factors 4in1 from the DoRothEA database Garcia-Alonso et al., 2019 and combined them into a 4in1 regulon containing 145 target genes. Additionally, I defined a PROX1 regulon comprising 1,411 genes, bound by PROX1 as per CUT&RUN data and showing direct regulation based on activator or repressor fusion transcriptome analysis Figure 4.2.13a. Cluster 1, rich in hepatocyte genes, was significantly enriched for 4in1 target genes and strongly induced during reprogramming Figure 4.2.14a-c. Conversely, PROX1 target genes were significantly underrepresented in cluster 1, suggesting that while 4in1 directly enhances hepatocyte maturation, PROX1 indirectly supports this effect Figure 4.2.14b-c. Indeed, cluster 2, containing eight alternate fate signatures, was significantly enriched for direct PROX1 targets Figure 4.2.14b-c. Since this cluster was downregulated with

Prox1 overexpression Figure 4.2.14a, it further implies that PROX1 actively represses unwanted fates to promote liver cell induction and maturation.

Prrx1 and Pparg are two alternate fate inducers repressed by PROX1

To interpret how PROX1 effectively silences various non-hepatocyte cell fates, I once again utilized GRaNPA (see Chapter 2). This time, with differential expression data available for different time points Figure 4.2.14a, and for GRN, I created a GRN around PROX1. This involved using the Prox1 regulon and the Dorothea database for transcription factors targeted by Prox1.

As anticipated, the differential expression on day two was almost entirely attributable to PROX1 Figure 4.2.14d. By day 7, my predictions included additional transcription factors, specifically the direct PROX1 targets PRRX1 and EBF2, as significant regulators of differential gene expression Figure 4.2.14d. Significantly, network analysis indicated that PROX1, along with its direct targets PRRX1 and EBF2, remained crucial at days 14 and 28. This persistent relevance of PROX1 throughout the four-week period underlines its potential role in maintaining hepatocyte fate Figure 4.2.14d.

From day 28 onwards, I identified further transcription factors potentially regulated by PROX1, such as the cardiac regulator HAND2 (Fernandez-Perez et al., 2019) Figure 4.2.14d. Notably, the regulons of all transcription factors targeted by PROX1 were predominantly found in the repressed non-hepatocyte cluster 2 Figure 4.2.14c-d. In contrast to PROX1, all downstream transcription factors were predicted to function as activators, as inferred from the expression correlation with their target genes Figure 4.2.14e.

Therefore, my gene regulatory network analysis suggests that PROX1 upholds hepatocyte identity by directly silencing genes associated with alternate cell fates and by repressing transcription factors that activate non-hepatocyte gene programs.

Next, I discovered that the identities of the donor fibroblasts and alternate adipocytes were among the gene signatures most effectively repressed by PROX1. This was evident from both bulk and single-cell transcriptomic inferences during hepatocyte reprogramming Figure 4.2.10 and Figure 4.2.14. Consistent with these findings, GRaNPA analysis identified key regulators of both cell types as direct targets of PROX1. EBF2, known as a co-activator of PPARG, collaborates to drive adipogenesis (Jimenez et al., 2007). Additionally, PRRX1, a master transcription factor for stromal fibroblasts, regulates the differentiation of mesodermal cell types and is implicated in fibrosis during wound healing (Leavitt et al., 2020). We confirmed that Prrx1 and Pparg are directly targeted by PROX1 at their promoters, showing reduced chromatin accessibility upon Prox1 overexpression.

When Prrx1 or Pparg were overexpressed during 4in1-mediated hepatocyte reprogramming, similar effects to Prox1 deletion were observed, including approximately 50% reduction in Albumin expression per cell and impaired liver fate induction as indicated by TJP1 immunofluorescence Figure 4.2.14f-g. Moreover, overexpression of Prrx1 or Pparg along with Prox1 negated the

beneficial effects of PROX1 on liver reprogramming. Conversely, shRNA-mediated depletion of Prrx1 or Pparg led to an approximate 1-2 fold increase in Albumin expression per cell and in the number of TJP1-positive hepatocyte-like cells upon 4in1 expression Figure 4.2.14f-g, mirroring the results of Prox1 overexpression. The depletion of Prrx1 or Pparg alongside Prox1 overexpression did not further enhance hepatocyte reprogramming, as judged by Albumin protein levels and TJP1 immunofluorescence, suggesting that they function downstream of PROX1 Figure 4.2.14h. The overexpression and sh knockout experiment has been conducted by Dr. Bryce Lim.

In conclusion, My collaborators and I have found that PROX1 limits increased cellular plasticity by actively repressing non-hepatic cell identities. This is achieved through the direct transcriptional silencing of master regulators of alternate lineages, including the fibroblast-specific Prrx1 and the adipocyte regulator Pparg.

4.2.8 PROX1 blocks hepatocellular carcinoma transdifferentiation in mice

Cellular plasticity and transdifferentiation play important roles in liver cancer development, as evidenced by previous studies (Seehawer et al., 2018). Cholangiocarcinoma (CCA) and HCC are identified as the two principal forms of liver cancer, each with distinct cellular compositions and morphologies (Farazi and DePinho, 2006). However, the process of transformation from hepatocytes to HCC or CCA remains largely unexplored. Given Prox1's significant role as a safeguard in hepatocyte identity, its potential influence in HCC and CCA was examined.

An initial comparison between 153 liver cancer patients (Chaisaingmongkol et al., 2017) revealed that the median expression of PROX1 is notably higher in HCC compared to CCA, hinting at PROX1's possible involvement in HCC Figure 4.2.16a.

My observations also showed a clear positive correlation between PROX1 expression and the HCC marker HNF4A Figure 4.2.16b. Importantly, in addition to the CCA marker KRT19, PROX1 targets PRRX1 and PPARG displayed a negative correlation with PROX1 expression in cancer tissues, indicating that PROX1 might contribute to reducing cellular plasticity by suppressing genes linked to alternative cell fates, even in primary liver cancers.

To delve deeper into the role of PROX1 in HCC and CCA, an in vivo experiment was devised, focusing on the effects of Prox1 knockdown during HCC formation Figure 4.2.16c. Dr. Bryce Lim conducted this experiment. The results aligned with previous observations, showing that Prox1 knockdown led to an increase in tumor sizes, as depicted in another Figure 4.2.16d. A critical finding was that reducing Prox1 levels triggered a morphological shift from HCC to CCA. This shift was characterized by the development of cholangiocyte-like tubular structures, a decrease in HNF4A (a primary hepatocyte marker), and an increase in KRT19 expression (a key cholangiocyte marker) Figure 4.2.16e-f. Transcriptomic analysis of these tumor nodules revealed that Prox1 knockdown results in the loss of hepatocyte-associated markers such as Alb,

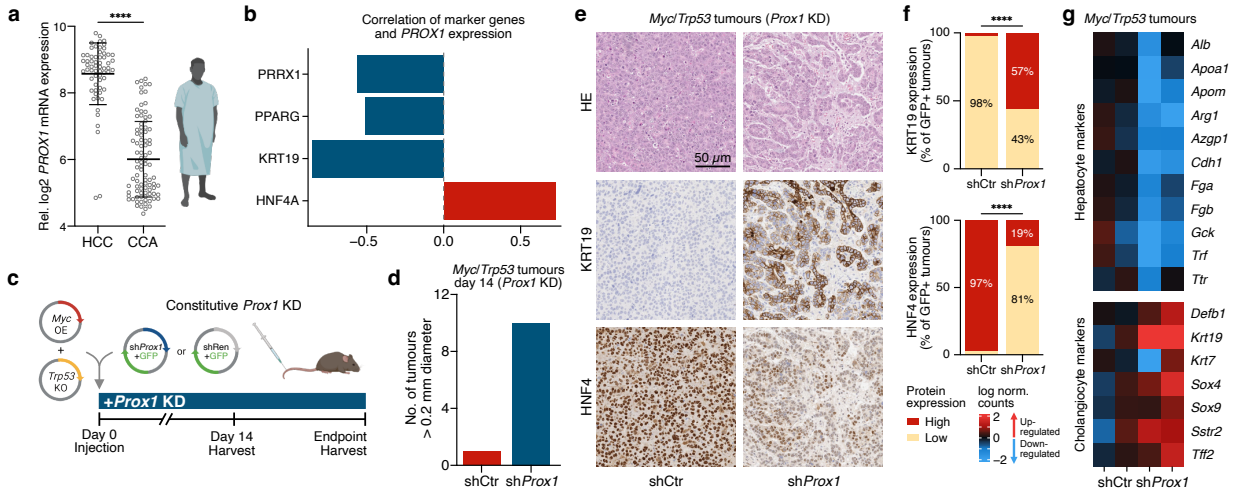

Figure 4.2.16. PROX1 expression in liver cancer and its implications in HCC and CCA subtypes along with impact on tumor development

(a) Comparison of PROX1 gene expression in tumor samples from patients with HCC (n=62) and CCA (n=91) using the TIGER-LC dataset (Chaisaingmongkol et al., 2017).

(b) Analysis of the correlation between PROX1 expression and specific markers and fate regulators in 153 HCC and CCA cancer patients from the TIGER-LC dataset (Chaisaingmongkol et al., 2017).

(c) Description of the hydrodynamic tail vein (HDTV) injection method used to induce HCC-like tumors with Myc overexpression and Trp53 knockout, combined with Prox1 knockdown.

(d) Total number of tumor nodules larger than 0.2 mm in diameter at day 14 post-induction, as detailed in (b), following shRNA-mediated control or Prox1 knockdown (n=5).

(e) Immunohistological analysis of tumor models at endpoint, as established in (c), using hematoxylin and eosin staining and KRT19 and HNF4 antibodies, comparing control and Prox1-knockdown samples.

(f) Quantitative assessment of CCA (KRT19) and HCC (HNF4) marker protein levels in GFP-positive tumors from (e), presented as a percentage. Statistical significance tested using Fisher's exact test, **** $p < 0.0001$.

(g) Transcriptomic analysis of tumor nodules from mice in (e) using RNA-seq post-Prox1 knockdown (n=2). A heatmap of selected differentially expressed genes related to cholangiocytes and hepatocytes is provided.

This figure was contributed by Dr. Bryce Lim (part a,c-g) and myself (part b) and is part of a submitted manuscript (Kamal, Lim et al.).

Apoa1, Cdh1, Trf, and Ttr, and an increase in cholangiocyte markers like Krt19 and Sox9, further detailed in an additional Figure 4.2.16g. These findings suggest that Prox1 not only functions as a safeguard repressor for hepatocyte identity and aids in preventing liver diseases but also plays a role in lowering cellular plasticity, thereby reducing the potential for cellular transformation and transdifferentiation.

4.3 Discussion

In various cells, while numerous transcription factors may be expressed, only a few key selectors or master regulators are crucial for inducing specific lineage identities by activating gene expression during development. Loss of these factors can increase cellular plasticity and potentially lead to cancer. My collaborator, Mortiz Mall, uncovered a new category of TFs, called 'safeguard repressors'. These are specifically expressed in one cell type and function to both induce and maintain its fate by targeting marker TFs of alternative cell fates.

In this study, Dr. Ignacio Ibbara developed a computational screening method using 18 cell types to identify potential safeguard repressors. Focusing on hepatocytes, PROX1 emerged as a significant candidate, noted for its role in cancer suppression and consistent lifelong expression. Additionally, Dr. Bryce Lim and I have demonstrated the tumor-suppressing effects of PROX1 in an HCC cell line, notably through its ability to reduce cell proliferation by targeting the MYC gene. I also demonstrated PROX1's function as a safeguard repressor in liver injury and hepatocyte reprogramming, where it enhances reprogramming efficiency by repressing key regulators of other cell fates, such as PPARG and PRRX1. Conclusively, our findings illustrated that PROX1 prevents transdifferentiation in cancer, as shown by HCC cells transitioning to CCA following the knockout of PROX1.

In summary, our research demonstrates that cell type-specific safeguard repressors are crucial in both inducing and maintaining cell fate by actively repressing alternative cell fates. Employing computational methods to identify these key factors offers the potential for generating cells for biomedical applications and preventing diseases associated with cell fate plasticity.

4.4 Method

4.4.1 Survival analysis

In all survival analyses I conducted, which included evaluating the survival of HCC patients and the survival in a mouse HCC model, I utilized Kaplan-Meier curves to plot the survival durations. To analyze the statistical significance of the differences in survival outcomes, I employed a log-rank test. Specifically, for the analysis of the impact on HCC patient survival illustrated in Figure 4.2.3c, I presented the results as -log10(p-values) derived from the log-rank test. In this context, positive values indicate an improvement in survival correlated with higher expression of the candidate gene, while negative values suggest poorer survival outcomes associated with higher candidate expression.

4.4.2 single cell RNA-seq processing

I analyzed reprogramming single-cell RNA-seq data using 10x Genomics Cell Ranger (version 4.0.0) and Seurat (version 4.3) (Hao et al., 2021). During the data preprocessing phase, I removed cells that had fewer than 1,000 features or fewer than 2,000 reads. Additionally, cells where mitochondrial genes constituted more than 20% of all genes were also discarded to ensure data quality. To identify and exclude cell doublets, I applied Scrublet (Wolock et al., 2019) with a threshold set at 0.35.

For the demultiplexing process based on their hashtag oligos, I first realigned the cells, concentrating on reads that had been discarded but contained hashtag oligos. The purpose of this realignment was to accurately determine the count of each hashtag in every cell. Following this, I used the HTODemux function in Seurat for the demultiplexing process, implementing a recursive approach to ensure precise and efficient sorting of the cells.

In explaining the HTODemux process: HTODemux performs k-medoid clustering on normalized HTO values, initially segregating cells into K (number of samples) + 1 clusters. For each HTO, a 'negative' distribution is calculated using the cluster with the lowest average value as the negative group. A negative binomial distribution is then fitted to this negative cluster for each HTO, and the 0.99 quantile of this distribution is used as a threshold. Cells are classified as positive or negative for each HTO based on these thresholds, and cells positive for more than one HTO are labeled as doublets.

Despite these measures, a considerable number of cells remained undefined or were categorized as doublets. To tackle this issue, I removed the cells that had already been annotated and reran the algorithm exclusively on those labeled as unknown or doublets. I repeated this recursive process three additional times, which significantly improved the accuracy in identifying and categorizing a substantial proportion of the cells.

Next, In order to minimize the impact of cell cycle variation on the analysis, I regressed out the expression values of cell cycle genes using the vars.to.regress parameter in ScaleData().

The cells were then visualized in a 2-dimensional space using the UMAP algorithm. For achieving optimal clustering of cells based on their gene expression profiles, I utilized 40 dimensions and set the resolution parameter at 0.35. To determine the best possible number of clusters, I employed the clustree package (Zappia and Oshlack, 2018), which provided guidance on the most suitable clustering resolution. This approach allowed for a detailed and nuanced analysis of the single-cell RNA-seq data, highlighting the complexity of cellular reprogramming processes.

Activity score quantification for 4in1 TFs and Prox1

I determined the activity of the 4in1 transcription factors by aggregating the expression of all genes within the 4in1 regulon. This regulon was sourced from Dorothea, and the aggregate expression was calculated using Seurat's addModuleScore() function. For PROX1 activity, I took a different approach: I calculated the inverse of the aggregate expression of a select subset of the PROX1 regulon. This subset, detailed in Supplementary Table 9, comprised 79 high-confidence PROX1 repressed target genes.

These target genes were chosen because they each have a PROX1 CUT&RUN peak and a motif within 1 kb of their TSS. Additionally, the promoters of these genes showed differential closure at day two following PROX1 overexpression, as indicated by an ATAC-seq log2 fold change (log2FC) threshold (comparing GFP vs Prox1) of greater than 1. By focusing on this subset of genes, I aimed to gain a more nuanced understanding of PROX1's role and its repressive impact on specific target genes during the process of transcription regulation.

Filtering cells without transduction

I utilized PROX1 activity as a criterion to exclude cells from the Prox1 condition that were not effectively transduced with the Prox1 overexpression lentivirus. For each cell labeled as Prox1, I calculated two proportions: the proportion of GFP-labeled cells with lower PROX1 activity than the given Prox1-labeled cell (denoted as $GFP_{proportion}$), and the proportion of Prox1-labeled cells with higher PROX1 activity than the same cell (referred to as $Prox1_{proportion}$). Cells were then excluded if the $GFP_{proportion}$ exceeded the $Prox1_{proportion}$.

A parallel method was applied for cells labeled with 4in1 and MEF. This approach helped in filtering out cells from the dataset that had not been successfully transduced with the 4in1 overexpression lentivirus.

Regulon analysis

I defined the PROX1 regulon as the set of genes that showed downregulation when fused with the PROX1 repressor, or upregulation with the activator fusion. Additionally, these genes needed to

have a PROX1 CUT&RUN peak and motif within 1 kb of their TSS. To identify target gene regulons for all other transcription factors, I used the Dorothea database (version 1.7.2, encompassing all confidence levels) (Garcia-Alonso et al., 2019).

Further, I constructed a sub-gene-regulatory network that included the PROX1 regulon and the regulons of transcription factors directly regulated by PROX1. This network allowed for a comprehensive analysis of the gene regulatory relationships and dependencies orchestrated by PROX1 and its associated transcription factors, providing deeper insights into their collective roles in gene regulation.

4.4.3 ATACseq preprocessing

The ATAC-seq data were processed using a custom Snakemake pipeline. The initial step involved checking the quality of raw reads with FastQC (v0.11.8). The reads were then trimmed using Trimmomatic (v0.38) and aligned to the UCSC mm10 or hg38 genomes with Bowtie2 (v2.3.4.3) (Langmead and Salzberg, 2012). Following alignment, the reads underwent cleaning and base recalibration with Samtools (v1.10) (Danecek et al., 2021) and Picard (v2.18.16) (Broad Institute 2019) to account for Tn5 insertion biases. The filtering process involved Bedtools (v2.27.1) (Quinlan and Hall, 2010), Samtools, and Picard. Peak calling was conducted using MACS2 (2.1.2), and coverage calculations were performed with deepTools (v3.1.3) (Ramirez et al., 2016). Final quality assessments were completed with MultiQC (v1.6) (Ewels et al., 2016). For the differential peak analysis, I employed DiffBind (v3.4.11) (Ross-Innes et al., 2012), following the methodology outlined in the author's vignette (same version). This approach allowed for a detailed and robust analysis of the ATAC-seq data.

4.4.4 RNAseq preprocessing

The raw reads were aligned to the reference genomes mm10 or hg38 using the STAR alignment tool (Dobin et al., 2013). To identify differential gene expression, I utilized DESeq2 (R package version 1.28.1) (Love et al., 2014), applying size factor normalization and Wald significance tests in the analysis. For the bulk MEF reprogramming data, the primary MEF line was incorporated as a covariate in the analysis. For the visualization of data, particularly in heatmaps, I used the ComplexHeatmap package (version 2.12.1) (Gu et al., 2016). In the analysis presented in Figure 4.2.14a genes were selected for inclusion if they exhibited an absolute log2 fold change (log2FC) greater than 0.75 at any of the examined time points, ensuring a focus on genes with significant expression changes.

Chapter 5

Conclusions and future perspectives

Cells vary in their roles and structures, yet they all contain identical genetic information. This highlights the significance of each cell's unique gene expression profile, which differs across cell types and states. A key group of proteins, known as transcription factors, plays a key role in regulating gene expression in cells. Research indicates that altering the expression of just a few transcription factors can transform the identity of cells (Takahashi and Yamanaka, 2006). This underscores the importance of understanding the gene regulatory networks that define each cell type and state. Moreover, studies have demonstrated that disruptions in gene regulation are often linked to various diseases (Cha, 2007).

Building gene regulatory networks initially relied on curated and experimental methods. However, with the advent of high-throughput data, computational methods have gained prominence and become increasingly important. These computational approaches are crucial for developing GRNs that are more specifically aligned with particular biological questions, moving beyond the generalized networks derived from databases. As a result, a wide range of computational methods have emerged. Initially, these methods focused on transcriptomic data and used gene-to-gene correlations. However, this approach, lacking information about transcription factors, often led to many false positives (Badia-I-Mompel et al., 2023; Langfelder and Horvath, 2008). Significant improvements were made with the introduction of methods like GENIE3 (Huynh-Thu et al., 2010a). The integration of chromatin accessibility data into GRN construction marked a significant advancement, greatly enhancing the accuracy and relevance of the networks.

Enhancers are key cis-regulatory elements where transcription factors bind to regulate gene expression. In our work, my colleague and I have developed a method that merges information about TFs with chromatin accessibility, particularly focusing on enhancers, and links these elements to genes. This approach helped us construct a three-part network, GRaNIE, which is useful for exploring biological questions related to specific contexts or cell types. However, since there is no established 'gold standard' gene regulatory network for evaluating various methods, I introduced a new evaluation technique called GRaNPA. GRaNPA is based on the hypothesis that a network which accurately reflects real biological processes — specifically, the true connections

between TFs and their target genes — should be able to predict changes in gene expression to a certain degree. This method allows us to determine if the network genuinely represents regulatory connections or if it mainly consists of false positives, in an unbiased manner.

One limitation of our GRaNIE method was its reliance on bulk data. However, with the advancement of single-cell technologies, It is now understood that much more detailed information about cell types and states can be obtained. One challenge in adapting single-cell data for GRaNIE is the lack of individual samples. In bulk data, GRaNIE determines correlations between individuals to link TFs to peaks or peaks to genes. However, in single-cell data, we often lack a sufficient number of individual samples (typically 15-20 are needed). To overcome this, scGRaNIE has been implemented by Dr. Christian Arnold, which leverages the sparsity of single-cell data. This method involves pseudo-bulking cells based on specific categories, which can be user-defined based on prior knowledge, such as cell type annotations or clusters. By using these pseudo-bulk categories, scGRaNIE can provide insights into specific variations. For instance, if scGRaNIE is built based on different cell types, it reveals variations between these cell types. Alternatively, we can construct cell type-specific scGRaNIE based on variations across different individuals.

I have improved the GRaNPA algorithm by adopting the scenario of predicting differential expression using multiple networks at once, and specifically identifying important transcription factors for each network. It has been shown that scGRaNIE can produce networks that capture variation across different biological concepts. This enables the evaluation of which networks are more predictive for specific questions or biological expression variations. Additionally, this approach allows for a comparison of important TFs reported by GRaNPA across different networks. For example, if two different networks contain distinct regulons for the same TF within the same concept, it can be determined which one is more important, indicating the regulon's relevance to the concept.

Analyzing gene regulatory networks along developmental trajectories offers valuable insights into biological processes. Questions such as how regulons of specific transcription factors change over time, which TFs are crucial at different development stages, and understanding the patterns of expression changes for a TF and its regulon are essential for deeper knowledge of development and differentiation. The transition from static to dynamic GRNs has been facilitated by a method known as Dictys (Wang et al., 2023). Dictys is a dynamic gene regulatory network inference and analysis tool that combines multiomic single-cell data, context-specific transcription factor footprinting, and probabilistic modeling, enhancing the accuracy of GRN reconstruction and enabling the analysis of cell-type-specific and dynamic networks in developmental contexts (Wang et al., 2023). However, a limitation of Dictys is its restriction to specific GRNs, indicating a gap in methodologies for analyzing dynamic GRN changes throughout developmental trajectories. Here, GRaNPA, in combination with scGRaNIE, presents a promising approach. By reconstructing multiple GRNs at various cell resolutions across trajectories using scGRaNIE, and employing GRaNPA to identify important TFs between developmental stages, this method provides critical

insights into the significant TFs in individual states. This approach, along with scGRaNIE, forms a part of the ongoing project.

Cell fate plasticity is another crucial factor in gene regulation, particularly linked to diseases like cancer. Despite its significance, there is a notable absence of computational tools capable of assessing cell fate plasticity at the single-cell level, especially in conjunction with gene regulatory network knowledge. As discussed in Chapter 4, Terminal Selector and safeguard repressors are vital in maintaining and defining cell identity, suggesting the feasibility of computationally determining cellular plasticity based on these TFs' activity. Additionally, the current computational screening for safeguard repressors is constrained, as it identifies TF targets solely based on motifs in promoters. This screening, when integrated with cell state and cell type-specific regulons derived from GRNs like scGRaNIE, and utilizing GRaNPA to identify important TFs in each stage, could yield a more refined set of potential Terminal Selectors and safeguard repressors. By identifying these TFs, and based on the hypothesis that cell fate plasticity can be defined as a function of Terminal Selector and safeguard repressor activity, it becomes possible to computationally calculate cell fate plasticity.

In summary, understanding gene regulatory networks is important for unraveling the gene expression profiles of individual cells. Despite the primary challenge of evaluating these networks, there is a lack of extensive downstream analysis options post-GRN construction. GRaNPA, an easy-to-use tool, facilitates deeper insights into biological questions through the application of gene regulatory networks. It holds the potential to be integrated into more advanced tools, offering an improved understanding of complex biological questions.

References

Mazzarello, P (May 1999). “A unifying concept: the history of cell theory”. en. In: *Nat. Cell Biol.* 1.1, E13–5.

Zeng, Hongkui (July 2022). “What is a cell type and how to define it?” en. In: *Cell* 185.15, pp. 2739–2755.

Arendt, Detlev (Nov. 2008). “The evolution of cell types in animals: emerging principles from molecular studies”. en. In: *Nat. Rev. Genet.* 9.11, pp. 868–882.

Bernstein, Bradley E et al. (Feb. 2007). “The mammalian epigenome”. en. In: *Cell* 128.4, pp. 669–681.

Ladewig, Julia et al. (Apr. 2013). “Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies”. en. In: *Nat. Rev. Mol. Cell Biol.* 14.4, pp. 225–236.

Takahashi, Kazutoshi and Shinya Yamanaka (Aug. 2006). “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors”. en. In: *Cell* 126.4, pp. 663–676.

Davis, R L et al. (Dec. 1987). “Expression of a single transfected cDNA converts fibroblasts to myoblasts”. en. In: *Cell* 51.6, pp. 987–1000.

Vierbuchen, Thomas et al. (Feb. 2010). “Direct conversion of fibroblasts to functional neurons by defined factors”. en. In: *Nature* 463.7284, pp. 1035–1041.

Kamimoto, Kenji et al. (Feb. 2023). “Dissecting cell identity via network inference and in silico gene perturbation”. en. In: *Nature* 614.7949, pp. 742–751.

Basu, Amitava and Vijay K Tiwari (July 2021). “Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine”. en. In: *Clin. Epigenetics* 13.1, p. 144.

Steens, Jennifer and Diana Klein (Sept. 2022). “HOX genes in stem cells: Maintaining cellular identity and regulation of differentiation”. en. In: *Front. Cell Dev. Biol.* 10, p. 1002909.

Lee, Tong Ihn and Richard A Young (Mar. 2013). “Transcriptional regulation and its misregulation in disease”. en. In: *Cell* 152.6, pp. 1237–1251.

Graf, Thomas (Dec. 2011). “Historical origins of transdifferentiation and reprogramming”. en. In: *Cell Stem Cell* 9.6, pp. 504–516.

Ng, Huck-Hui and M Azim Surani (May 2011). “The transcriptional and signalling networks of pluripotency”. en. In: *Nat. Cell Biol.* 13.5, pp. 490–496.

Orkin, Stuart H and Konrad Hochedlinger (June 2011). “Chromatin connections to pluripotency and cellular reprogramming”. en. In: *Cell* 145.6, pp. 835–850.

Casamassimi, Amelia et al. (July 2017). "Transcriptome profiling in human diseases: New advances and perspectives". In: *Int. J. Mol. Sci.* 18.8, p. 1652.

Casamassimi, Amelia and Alfredo Ciccodicola (Mar. 2019). "Transcriptional regulation: Molecules, involved mechanisms, and misregulation". en. In: *Int. J. Mol. Sci.* 20.6, p. 1281.

Cha, Jang-Ho J (Nov. 2007). "Transcriptional signatures in Huntington's disease". en. In: *Prog. Neurobiol.* 83.4, pp. 228–248.

Shu, Xing-Sheng et al. (Oct. 2012). "Chromatin regulators with tumor suppressor properties and their alterations in human cancers". en. In: *Epigenomics* 4.5, pp. 537–549.

Chen, Taiping and Sharon Y R Dent (Feb. 2014a). "Chromatin modifiers and remodellers: regulators of cellular differentiation". en. In: *Nat. Rev. Genet.* 15.2, pp. 93–106.

Feng, Suhua et al. (Oct. 2010). "Epigenetic reprogramming in plant and animal development". en. In: *Science* 330.6004, pp. 622–627.

Lejeune, Erwan and Robin C Allshire (June 2011). "Common ground: small RNA programming and chromatin modifications". en. In: *Curr. Opin. Cell Biol.* 23.3, pp. 258–265.

Chen, Taiping and Sharon Y R Dent (Feb. 2014b). "Chromatin modifiers and remodellers: regulators of cellular differentiation". en. In: *Nat. Rev. Genet.* 15.2, pp. 93–106.

Oksuz, Ozgur et al. (July 2023). "Transcription factors interact with RNA to regulate genes". en. In: *Mol. Cell* 83.14, 2449–2463.e13.

Weidemüller, Paula et al. (Dec. 2021). "Transcription factors: Bridge between cell signaling and gene regulation". en. In: *Proteomics* 21.23-24, e2000034.

Badia-I-Mompel, Pau et al. (Nov. 2023). "Gene regulatory network inference in the era of single-cell multi-omics". en. In: *Nat. Rev. Genet.* 24.11, pp. 739–754.

Wang, Guohua et al. (Sept. 2015). "Understanding transcription factor regulation by integrating gene expression and DNase I hypersensitive sites". en. In: *Biomed Res. Int.* 2015, p. 757530.

Panigrahi, Anil and Bert W O'Malley (Apr. 2021). "Mechanisms of enhancer action: the known and the unknown". en. In: *Genome Biol.* 22.1, p. 108.

Zaugg, Judith Barbara et al. (Dec. 2022). "Current challenges in understanding the role of enhancers in disease". en. In: *Nat. Struct. Mol. Biol.* 29.12, pp. 1148–1158.

Claringbould, Annique and Judith B Zaugg (Nov. 2021). "Enhancers in disease: molecular basis and emerging treatment strategies". en. In: *Trends Mol. Med.* 27.11, pp. 1060–1073.

Su, Emily Y et al. (Feb. 2022). "Reconstruction of dynamic regulatory networks reveals signaling-induced topology changes associated with germ layer specification". en. In: *Stem Cell Reports* 17.2, pp. 427–442.

Langfelder, Peter and Steve Horvath (Dec. 2008). "WGCNA: an R package for weighted correlation network analysis". en. In: *BMC Bioinformatics* 9.1, p. 559.

Huynh-Thu, Vân Anh et al. (Sept. 2010a). "Inferring regulatory networks from expression data using tree-based methods". en. In: *PLoS One* 5.9, e12776.

Xu, Quan et al. (July 2021). "ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination". In: *Nucleic Acids Research* 49.14, pp. 7966–7985. ISSN: 0305-1048. DOI: 10.1093/nar/gkab598. eprint: https://academic.oup.com/nar/article-pdf/49/14/7966/39802888/gkab598.pdf.

Pratapa, Aditya et al. (Feb. 2020). "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data". en. In: *Nat. Methods* 17.2, pp. 147–154.

Aibar, Sara et al. (Nov. 2017). "SCENIC: single-cell regulatory network inference and clustering". en. In: *Nat. Methods* 14.11, pp. 1083–1086.

Bravo González-Blas, Carmen et al. (Sept. 2023). "SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks". en. In: *Nat. Methods* 20.9, pp. 1355–1367.

Sanguinetti, Guido and Van Anh Huynh-Thu, eds. (Dec. 2018). *Gene regulatory networks*. en. 1st ed. Methods in molecular biology (Clifton, N.J.) New York, NY: Humana Press.

Kartha, Vinay K et al. (Sept. 2022). "Functional inference of gene regulation using single-cell multi-omics". en. In: *Cell Genom.* 2.9, p. 100166.

Kamal, Aryan et al. (2023). "GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks". In: *Molecular Systems Biology* 19.6, e11627. DOI: https://doi.org/10.15252/msb.202311627. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.202311627.

Larcombe, Michael R et al. (Dec. 2022). "Indirect mechanisms of transcription factor-mediated gene regulation during cell fate changes". en. In: *Adv. Genet. (Hoboken)* 3.4, p. 2200015.

Wapinski, Orly L et al. (Oct. 2013). "Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons". en. In: *Cell* 155.3, pp. 621–635.

Mall, Moritz et al. (Apr. 2017). "Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates". In: *Nature* 544.7649, pp. 245–249. ISSN: 1476-4687. DOI: 10.1038/nature21722.

Hanahan, Douglas and Robert A Weinberg (Mar. 2011). "Hallmarks of cancer: the next generation". en. In: *Cell* 144.5, pp. 646–674.

Hanahan, Douglas (Jan. 2022). "Hallmarks of cancer: New dimensions". en. In: *Cancer Discov.* 12.1, pp. 31–46.

Ordóñez-Morán, Paloma et al. (Dec. 2015). "HOXA5 counteracts stem cell traits by inhibiting Wnt signaling in colorectal cancer". en. In: *Cancer Cell* 28.6, pp. 815–829.

Love, Michael I. et al. (Dec. 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

Chen, Tianqi and Carlos Guestrin (Aug. 2016). "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM.

Wright, Marvin N. and Andreas Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1, pp. 1–17. DOI: 10.18637/jss.v077.i01.

Strobl, Carolin et al. (Jan. 2007). "Bias in random forest variable importance measures: illustrations, sources and a solution". en. In: *BMC Bioinformatics* 8.1, p. 25.

Freimer, Jacob W et al. (Aug. 2022). "Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks". en. In: *Nat. Genet.* 54.8, pp. 1133–1144.

Võsa, Urmo et al. (Sept. 2021). "Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression". en. In: *Nat. Genet.* 53.9, pp. 1300–1310.

Janssens, Jasper et al. (Jan. 2022). "Decoding gene regulation in the fly brain". en. In: *Nature* 601.7894, pp. 630–636.

Reyes-Palomares, Armando et al. (Apr. 2020). "Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension". In: *Nature Communications* 11.1, p. 1673. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15463-x.

Huynh-Thu, Vân Anh et al. (Sept. 2010b). "Inferring regulatory networks from expression data using tree-based methods". en. In: *PLoS One* 5.9, e12776.

Haynes, Brian C et al. (Aug. 2013). "Mapping functional transcription factor networks from gene expression data". en. In: *Genome Res.* 23.8, pp. 1319–1328.

Huynh-Thu, Vân Anh and Pierre Geurts (Feb. 2018). "dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data". en. In: *Sci. Rep.* 8.1.

Garcia-Alonso, Luz et al. (Aug. 2019). "Benchmark and integration of resources for the estimation of human transcription factor activities". en. In: *Genome Res.* 29.8, pp. 1363–1375.

Han, Heonjong et al. (Jan. 2018). "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions". en. In: *Nucleic Acids Res.* 46.D1, pp. D380–D386.

Liu, Zhi-Ping et al. (Sept. 2015). "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse". en. In: *Database (Oxford)* 2015, bav095.

Keenan, Alexandra B et al. (July 2019). "ChEA3: transcription factor enrichment analysis by orthogonal omics integration". en. In: *Nucleic Acids Res.* 47.W1, W212–W224.

Fulco, Charles P et al. (Dec. 2019). "Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations". en. In: *Nat. Genet.* 51.12, pp. 1664–1669.

Schraivogel, Daniel et al. (June 2020). "Targeted Perturb-seq enables genome-scale genetic screens in single cells". en. In: *Nat. Methods* 17.6, pp. 629–635.

Chen, Shuonan and Jessica C Mar (Dec. 2018). "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data". en. In: *BMC Bioinformatics* 19.1.

Kulakovskiy, Ivan V et al. (Jan. 2018). "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis". In: *Nucleic Acids Res.* 46.D1, pp. D252–D259.

Castro-Mondragon, Jaime A et al. (Jan. 2022). "JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles". en. In: *Nucleic Acids Res.* 50.D1, pp. D165–D173.

Berest, Ivan et al. (Dec. 2019). "Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF". en. In: *Cell Rep* 29.10, 3147–3159.e12.

Alasoo, Kaur et al. (Mar. 2018). "Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response". In: *Nature Genetics* 50.3, pp. 424–431. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0046-7.

Novikova, Gloriia et al. (Mar. 2021). "Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes". In: *Nature Communications* 12.1, p. 1610. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21823-y.

Hammal, Fayrouz et al. (Nov. 2021). "ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments". In: *Nucleic Acids Research* 50.D1, pp. D316–D325. ISSN: 0305-1048. DOI: 10.1093/nar/gkab996. eprint: https://academic.oup.com/nar/article-pdf/50/D1/D316/42058627/gkab996.pdf.

Andersson, Robin et al. (Mar. 2014). "An atlas of active enhancers across human cell types and tissues". en. In: *Nature* 507.7493, pp. 455–461.

Bunina, Daria et al. (2021). "Pathological LSD1 mutations cause HDAC-mediated aberrant gene repression during early cell differentiation". In: *bioRxiv*. DOI: 10.1101/2021.08.11.455900. eprint: https://www.biorxiv.org/content/early/2021/08/11/2021.08.11.455900.full.pdf.

Ouma, Wilberforce Zachary et al. (Apr. 2018). "Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties". en. In: *PLoS Comput. Biol.* 14.4, e1006098.

Langlais, David et al. (Apr. 2016). "The macrophage IRF8/IRF1 regulome is required for protection against infections and is associated with chronic inflammation". en. In: *J. Exp. Med.* 213.4, pp. 585–603.

Grigoriadis, G et al. (Dec. 1996). "The Rel subunit of NF-kappaB-like transcription factors is a positive and negative regulator of macrophage gene expression: distinct roles for Rel in different macrophage populations". en. In: *EMBO J.* 15.24, pp. 7099–7107.

Jones, Gareth-Rhys et al. (Feb. 2020). "The methyl-CpG-binding protein Mbd2 regulates susceptibility to experimental colitis via control of CD11c+ cells and colonic epithelium". en. In: *Front. Immunol.* 11, p. 183.

Morishita, Hideaki et al. (Apr. 2009). "Fra-1 negatively regulates lipopolysaccharide-mediated inflammatory responses". en. In: *Int. Immunol.* 21.4, pp. 457–465.

An, Yanying et al. (Mar. 2020). "TRIM59 expression is regulated by Sp1 and Nrf1 in LPS-activated macrophages through JNK signaling pathway". en. In: *Cell. Signal.* 67.109522, p. 109522.

Blondel, Vincent D et al. (Oct. 2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

Murray, Peter J (Feb. 2017). "Macrophage polarization". en. In: *Annu. Rev. Physiol.* 79.1, pp. 541–566.

Parameswaran, Narayanan and Sonika Patial (2010). "Tumor necrosis factor-α signaling in macrophages". en. In: *Crit. Rev. Eukaryot. Gene Expr.* 20.2, pp. 87–103.

Nathan, C F et al. (Sept. 1983). "Identification of interferon-gamma as the lymphokine that activates human macrophage oxidative metabolism and antimicrobial activity". en. In: *J. Exp. Med.* 158.3, pp. 670–689.

Meng, Aihong et al. (Dec. 2014). "Role of p38 MAPK and STAT3 in lipopolysaccharide-stimulated mouse alveolar macrophages". en. In: *Exp. Ther. Med.* 8.6, pp. 1772–1776.

Garg, Swati et al. (July 2019). "Hepatic leukemia factor is a novel leukemic stem cell regulator in DNMT3A, NPM1, and FLT3-ITD triple-mutated AML". en. In: *Blood* 134.3, pp. 263–276.

He, Lixiazi et al. (Apr. 2022). "CDK7/12/13 inhibition targets an oscillating leukemia stem cell network and synergizes with venetoclax in acute myeloid leukemia". en. In: *EMBO Mol. Med.* 14.4, e14990.

Calderon, Diego et al. (Oct. 2019). "Landscape of stimulation-responsive chromatin across diverse human immune cells". en. In: *Nat. Genet.* 51.10, pp. 1494–1505.

Sigalova, Olga M et al. (Aug. 2020). "Predictive features of gene expression variation reveal mechanistic link with differential expression". en. In: *Mol. Syst. Biol.* 16.8, e9539.

Somma, Domenico et al. (Oct. 2021). "Defining the role of nuclear factor (NF)-κB p105 subunit in human macrophage by transcriptomic analysis of NFKB1 knockout THP1 cells". en. In: *Front. Immunol.* 12, p. 669906.

Liss, Franziska et al. (Feb. 2021). "IRF8 is an AML-specific susceptibility factor that regulates signaling pathways and proliferation of AML cells". en. In: *Cancers (Basel)* 13.4, p. 764.

Holland, Christian H et al. (June 2020). "Transfer of regulatory knowledge from human to mouse for functional genomics analysis". en. In: *Biochim. Biophys. Acta Gene Regul. Mech.* 1863.6, p. 194431.

Giraud-Gatineau, Alexandre et al. (May 2020). "The antibiotic bedaquiline activates host macrophage innate immune resistance to bacterial infection". en. In: *Elife* 9.

Pai, Athma A et al. (Sept. 2016). "Widespread shortening of 3' untranslated regions and increased Exon inclusion are evolutionarily conserved features of innate immune responses to infection". en. In: *PLoS Genet.* 12.9, e1006338.

Cassetta, Luca et al. (Apr. 2019). "Human tumor-associated macrophage and monocyte transcriptional landscapes reveal cancer-specific reprogramming, biomarkers, and therapeutic targets". en. In: *Cancer Cell* 35.4, 588–602.e10.

Medzhitov, Ruslan and Tiffany Horng (Oct. 2009). "Transcriptional control of the inflammatory response". en. In: *Nat. Rev. Immunol.* 9.10, pp. 692–703.

Liu, Ting et al. (July 2017). "NF-κB signaling in inflammation". en. In: *Signal Transduct. Target. Ther.* 2.

Leseigneur, Clarisse et al. (Sept. 2020). "Emerging evasion mechanisms of macrophage defenses by pathogenic bacteria". en. In: *Front. Cell. Infect. Microbiol.* 10, p. 577559.

Chen, Kaixuan et al. (Dec. 2020). "Communications between bone marrow macrophages and bone cells in bone remodeling". en. In: *Front. Cell Dev. Biol.* 8, p. 598263.

Corliss, Bruce A et al. (Feb. 2016). "Macrophages: An inflammatory link between angiogenesis and lymphangiogenesis". en. In: *Microcirculation* 23.2, pp. 95–121.

Chistiakov, Dimitry A et al. (Jan. 2018). "The impact of interferon-regulatory factors to macrophage differentiation and polarization into M1 and M2". en. In: *Immunobiology* 223.1, pp. 101–111.

Wang, Yi et al. (Jan. 2021). "MBD2 serves as a viable target against pulmonary fibrosis by inhibiting macrophage M2 program". en. In: *Sci. Adv.* 7.1, eabb6075.

Stafford, Sian L et al. (July 2013). "Metal ions in macrophage antimicrobial pathways: emerging roles for zinc and copper". en. In: *Biosci. Rep.* 33.4, pp. 541–554.

Festa, Richard A and Dennis J Thiele (Sept. 2012). "Copper at the front line of the host-pathogen battle". en. In: *PLoS Pathog.* 8.9, e1002887.

He, Lizhi et al. (Nov. 2021). "Global characterization of macrophage polarization mechanisms and identification of M2-type polarization inhibitors". en. In: *Cell Rep.* 37.5, p. 109955.

Weber, Janine et al. (Jan. 2016). "Structural basis of nucleic-acid recognition and double-strand unwinding by the essential neuronal protein Pur-alpha". en. In: *Elife* 5.

Hirata, Y et al. (Nov. 1993). "The phosphorylation and DNA binding of the DNA-binding domain of the orphan nuclear receptor NGFI-B". en. In: *J. Biol. Chem.* 268.33, pp. 24808–24812.

Marbach, Daniel et al. (Apr. 2016). "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases". en. In: *Nat. Methods* 13.4, pp. 366–370.

Zhang, Qinhu et al. (Jan. 2023). "Computational prediction and characterization of cell-type-specific and shared binding sites". en. In: *Bioinformatics* 39.1.

Hainer, Sarah J et al. (Nov. 2016). "DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells". en. In: *Elife* 5.

Qin, Wanhai et al. (July 2021). "The role of host cell DNA methylation in the immune response to bacterial infection". en. In: *Front. Immunol.* 12, p. 696280.

Zhu, Anqi et al. (June 2019). "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences". en. In: *Bioinformatics* 35.12, pp. 2084–2092.

Mannervik, M et al. (Apr. 1999). "Transcriptional coregulators in development". en. In: *Science* 284.5414, pp. 606–609.

Vos, Seychelle M (Apr. 2021). "Understanding transcription across scales: From base pairs to chromosomes". en. In: *Mol. Cell* 81.8, pp. 1601–1616.

Vaquerizas, Juan M et al. (Apr. 2009). "A census of human transcription factors: function, expression and evolution". en. In: *Nat. Rev. Genet.* 10.4, pp. 252–263.

Hobert, Oliver (Dec. 2008). "Regulatory logic of neuronal diversity: terminal selector genes and selector motifs". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 105.51, pp. 20067–20071.

Balsalobre, Aurelio and Jacques Drouin (July 2022). "Pioneer factors as master regulators of the epigenome and cell fate". en. In: *Nat. Rev. Mol. Cell Biol.* 23.7, pp. 449–464.

Treutlein, Barbara et al. (June 2016). "Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq". en. In: *Nature* 534.7607, pp. 391–395.

Schaum, Nicholas et al. (Oct. 2018). "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris". In: *Nature* 562.7727, pp. 367–372. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0590-4.

Waldron, Lauren et al. (Feb. 2016). "The cardiac TBX5 interactome reveals a chromatin remodeling network essential for cardiac septation". en. In: *Dev. Cell* 36.3, pp. 262–275.

Zhou, Q et al. (Sept. 2001). "The bHLH transcription factor Olig2 promotes oligodendrocyte differentiation in collaboration with Nkx2.2". en. In: *Neuron* 31.5, pp. 791–807.

Seehawer, Marco et al. (Oct. 2018). "Necroptosis microenvironment directs lineage commitment in liver cancer". en. In: *Nature* 562.7725, pp. 69–75.

Cardoso-Moreira, Margarida et al. (July 2019). "Gene expression across mammalian organ development". en. In: *Nature* 571.7766, pp. 505–509.

Song, Guangqi et al. (June 2016). "Direct reprogramming of hepatic myofibroblasts into hepatocytes in vivo attenuates liver fibrosis". en. In: *Cell Stem Cell* 18.6, pp. 797–808.

Liu, Yanfeng et al. (Aug. 2013). "PROX1 promotes hepatocellular carcinoma metastasis by way of up-regulating hypoxia-inducible factor 1α expression and protein stability". en. In: *Hepatology* 58.2, pp. 692–705.

Shimoda, Masayuki et al. (Oct. 2006). "A homeobox protein, prox1, is involved in the differentiation, proliferation, and prognosis in hepatocellular carcinoma". en. In: *Clin. Cancer Res.* 12.20 Pt 1, pp. 6005–6011.

Chaisaingmongkol, Jittiporn et al. (July 2017). "Common molecular subtypes among Asian hepatocellular carcinoma and cholangiocarcinoma". en. In: *Cancer Cell* 32.1, 57–70.e3.

Menyhárt, Otılia et al. (Dec. 2018). "Determining consistent prognostic biomarkers of overall survival and vascular invasion in hepatocellular carcinoma". en. In: *R. Soc. Open Sci.* 5.12, p. 181006.

Ahn, Sung-Min et al. (Dec. 2014). “Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification”. en. In: *Hepatology* 60.6, pp. 1972–1982.

Cerami, Ethan et al. (May 2012). “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data”. en. In: *Cancer Discov.* 2.5, pp. 401–404.

Gao, Jianjiong et al. (Apr. 2013). “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”. en. In: *Sci. Signal.* 6.269, p. l1.

Li, Lu et al. (Mar. 2023). “Kupffer-cell-derived IL-6 is repurposed for hepatocyte dedifferentiation via activating progenitor genes from injury-specific enhancers”. en. In: *Cell Stem Cell* 30.3, 283–299.e9.

Trapnell, Cole et al. (Apr. 2014). “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. en. In: *Nat. Biotechnol.* 32.4, pp. 381–386.

Franzén, Oscar et al. (Jan. 2019). “PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data”. en. In: *Database (Oxford)* 2019.

Chanda, Soham et al. (Aug. 2014). “Generation of induced neuronal cells by the single reprogramming factor ASCL1”. en. In: *Stem Cell Reports* 3.2, pp. 282–296.

Lee, Qian Yi et al. (Apr. 2020). “Pro-neuronal activity of Myod1 due to promiscuous binding to neuronal genes”. en. In: *Nat. Cell Biol.* 22.4, pp. 401–411.

Karalay, Ozlem et al. (Apr. 2011). “Prospero-related homeobox 1 gene (Prox1) is regulated by canonical Wnt signaling and has a stage-specific role in adult hippocampal neurogenesis”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.14, pp. 5807–5812.

Petrova, Tatiana V et al. (Sept. 2002). “Lymphatic endothelial reprogramming of vascular endothelial cells by the Prox-1 homeobox transcription factor”. en. In: *EMBO J.* 21.17, pp. 4593–4599.

Aranguren, Xabier L et al. (Mar. 2013). “COUP-TFII orchestrates venous and lymphatic endothelial identity by homo- or hetero-dimerisation with PROX1”. en. In: *J. Cell Sci.* 126.Pt 5, pp. 1164–1175.

Armour, Sean M et al. (Sept. 2017). “An HDAC3-PROX1 corepressor module acts on HNF4α to control hepatic triglycerides”. en. In: *Nat. Commun.* 8.1.

Fernandez-Perez, Antonio et al. (May 2019). “Hand2 selectively reorganizes chromatin accessibility to induce pacemaker-like transcriptional reprogramming”. en. In: *Cell Rep.* 27.8, 2354–2369.e7.

Jimenez, Maria A et al. (Jan. 2007). “Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade”. en. In: *Mol. Cell. Biol.* 27.2, pp. 743–757.

Leavitt, Tripp et al. (Nov. 2020). “Prrx1 fibroblasts represent a pro-fibrotic lineage in the mouse ventral dermis”. en. In: *Cell Rep.* 33.6, p. 108356.

Farazi, Paraskevi A and Ronald A DePinho (Sept. 2006). “Hepatocellular carcinoma pathogenesis: from genes to environment”. en. In: *Nat. Rev. Cancer* 6.9, pp. 674–687.

Hao, Yuhan et al. (June 2021). “Integrated analysis of multimodal single-cell data”. en. In: *Cell* 184.13, 3573–3587.e29.

Wolock, Samuel L et al. (Apr. 2019). “Scrublet: Computational identification of cell Doublets in Single-cell transcriptomic data”. en. In: *Cell Syst.* 8.4, 281–291.e9.

Zappia, Luke and Alicia Oshlack (July 2018). “Clustering trees: a visualization for evaluating clusterings at multiple resolutions”. en. In: *Gigascience* 7.7.

Langmead, Ben and Steven L Salzberg (Mar. 2012). “Fast gapped-read alignment with Bowtie 2”. en. In: *Nat. Methods* 9.4, pp. 357–359.

Danecek, Petr et al. (Feb. 2021). “Twelve years of SAMtools and BCFtools”. en. In: *Gigascience* 10.2.

Quinlan, Aaron R and Ira M Hall (Mar. 2010). “BEDTools: a flexible suite of utilities for comparing genomic features”. en. In: *Bioinformatics* 26.6, pp. 841–842.

Ramirez, Fidel et al. (July 2016). “deepTools2: a next generation web server for deep-sequencing data analysis”. en. In: *Nucleic Acids Res.* 44.W1, W160–W165.

Ewels, Philip et al. (Oct. 2016). “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19, pp. 3047–3048.

Ross-Innes, Caryn S et al. (Jan. 2012). “Differential oestrogen receptor binding is associated with clinical outcome in breast cancer”. en. In: *Nature* 481.7381, pp. 389–393.

Dobin, Alexander et al. (Jan. 2013). “STAR: ultrafast universal RNA-seq aligner”. en. In: *Bioinformatics* 29.1, pp. 15–21.

Gu, Zuguang et al. (Sept. 2016). “Complex heatmaps reveal patterns and correlations in multidimensional genomic data”. en. In: *Bioinformatics* 32.18, pp. 2847–2849.

Wang, Lingfei et al. (Sept. 2023). “Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics”. en. In: *Nat. Methods* 20.9, pp. 1368–1378.