4D Nucleome-Guided Cellular Reprogramming

by

Gabrielle A. Dotson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Indika Rajapakse, Chair
Professor Anthony M. Bloch
Professor Daniel M. Burns
Research Assistant Professor Lindsey A. Muir
Professor Pavan R. Reddy
Professor Max S. Wicha

Gabrielle A. Dotson
dotsonga@umich.edu
ORCID iD: 0000-0001-6624-5332

DEDICATION

It is with deep gratitude that I dedicate this dissertation to

> My parents, Dr. Glenn Averil Dotson and Mrs. Karen Olivia Dotson, who instilled in me the importance of faith and perseverance, without which I could not have completed this journey;
>
> My brother, Gabriel Alexander Dotson, who has always been there for me with uplifting words and a listening ear;
>
> My grandma, Barbara Ann Small, who has been my most ardent cheerleader and prayer warrior.

You have witnessed each life chapter that has culminated in this hour. Your love and support is what made this journey possible.

"Write the vision and engrave it plainly on [clay] tablets so that the one who reads it will run. For the vision is yet for the appointed [future] time. It hurries toward the goal [of fulfillment]; it will not fail." (Habakkuk 2:2-3)

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Indika Rajapakse for his mentorship throughout my doctoral training. Your devotion to teaching and endless enthusiasm for science made each day in lab a day to look forward to. Thank you for encouraging me to ”always read more” and challenging me to use my imagination - I am a better researcher and person because of it. I also wish to thank the members of my thesis committee: Anthony Bloch, Daniel Burns, Lindsey Muir, Pavan Reddy, and Max Wicha for their guidance and feedback as I worked towards my dissertation.

I would like to extend a sincere thank you to the former and current members of the Rajapakse lab for your valuable insight on analyses, manuscripts, and presentations, vibrant discussions on all things science and beyond, and for your general camaraderie: Walter Meixner, Cooper Stansbury, Maria Lawas, Joshua Pickard, Christian Kelley, Anthony Cicalo, Stephen Lindsly, Can Chen, Scott Ronquist, Sam Dilworth, Christopher York, Emily Crossette, Drew McKearney, Charles Ryan, Sivakumar Jeyarajan, and Haiming Chen. I would also like to thank my collaborators and research associates: Thomas Ried, Markus Brown, Peter Cook, Amit Surana, and Yaping Sun.

I am also grateful to the Bioinformatics graduate program coordinators and staff for their support and for fostering a welcoming academic community. Thank you also to Rackham for funding me through the Rackham Graduate Fellowship and to the National Institutes of Health (NIGMS, NHGRI) for funding me through the Bioinformatics Training Grant and Genome Science Training Program.

Finally, I want to express my gratitude to family and friends, near and far, who made my graduate school experience full and enjoyable. Thank you for the prayers, calls, food, favors, laughs, and for believing in me even when I did not believe in myself - I cannot thank you all enough.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ACRONYMS

1D One-dimensional

3D Three-dimensional

4D Four-dimensionl

4DN 4D nucleome

ATAC-seq Assay for transposase-accessible chromatin using sequencing

ChIP-seq Chromatin immunoprecipitation sequencing

CRC Colorectal cancer

FISH Fluorescence *in situ* hybridization

GVHD Graft-versus-host disease

Hi-C Genome-wide chromosome conformation capture

kb kilobase

Mb Megabase

MR Master regulator

NGS Next-generation sequencing

RNA-seq RNA sequencing

TAD Topologically associating domain

TF Transcription factor

scRNA-seq Single-cell RNA sequencing

ST Spatial transcriptomics

ABSTRACT

The 4D Nucleome (4DN) refers to the dynamical relationship between chromatin architecture and gene expression in the nucleus that gives rise to cellular phenotype. Our ability to steer cellular phenotype hinges on our ability to understand and precisely characterize a dynamic cellular state. The 4DN framework can be instrumental in helping to realize such reprogramming potential [288, 198]. In this dissertation, I demonstrate how 4DN data can be used to evaluate the impact of perturbations to a cell, examining how TF silencing sensitizes colorectal cancer cells to treatment and assessing how altering genome structure could be a viable therapeutic strategy harnessed for directly modulating *in vivo* disease processes. I then showcase how recent experimental innovations have allowed us to refine the 4DN framework. Profiling spatially-resolved gene expression in tissue, I characterize the disruption of adult adipose tissue morphogenesis during obesity progression. Finally, I offer a computational framework to analyze multi-way chromatin interactions and reveal organization principles of the genome that govern how transcription factors regulate the formation of transcription clusters. The culmination of this work has allowed us to move beyond the traditional scope of 4DN and consider insights from spatial networks and complex higher-order genome organization. Together, these advances can enable us to develop reprogramming regimes that reach target states with higher fidelity through exploring endogenous nucleome modifications, considering the anatomical context of cells, and ensuring recapitulation of a target cell's conformational state.

CHAPTER 1

Introduction

1.1 Research Overview

Cellular reprogramming refers to the transformation of one cell type into another cell type, whether directly or passing through a pluripotent state. This is most commonly achieved through the overexpression of externally introduced transcription factors (TFs) though can also be achieved by introducing chemical compounds to cells, RNAi-mediated TF silencing, or CRISPR/Cas9-mediated activation or silencing. Regardless of the approach, understanding the state of the source and target cell is integral to the successful realization of cellular transformations. While gene regulatory networks and gene expression profiles have proven sufficient in guiding the discovery of TF sets that can enable direct or pluripotent reprogramming, reprogramming efforts could further benefit from additional genomic measurements that take into consideration genome conformation and nuclear organization. Towards this end, the 4D Nucleome (4DN), which captures the phenotype-driving dynamical relationship between chromatin architecture and gene expression in the nucleus, offers a robust means of elucidating defining features of a cell [284, 61].

With the right set of TFs, it is largely guaranteed that a target cell type's transcriptional program will be successfully recovered. Maintaining that established state, however, remains a formidable challenge in the field. Low cell fate conversion efficiencies owing to the presence of TFs that maintain the starting cell's gene regulatory network, lineage-specific repressors, inaccessible chromatin, and more, contribute to cell types reverting back to their initial state or stalling in an intermediate state. Recent attempts to overcome this barrier have considered the role of epigenetic factors like DNA methylation and histone modifications, as certain chromatin features can interfere with reprogramming [21]. Structural information and tissue dynamics are often neglected in reprogramming studies but could have implications towards improving reprogramming efficiency as well.

In this dissertation, I outline my contributions towards furthering the field of cellular reprogramming through insights from 4DN analyses. This includes the analysis of novel 4DN datasets and the development of novel methods to perform this analysis. The remainder of this research

overview will summarize the work presented in each chapter and following sections will provide a background of relevant themes introduced in these chapters.

Chapter II, "Cellular Reprogramming: Mathematics Meets Medicine" provides an overview of computational approaches that have guided characterization of cell identity and prediction of transitions between cell identities to facilitate TF-mediated cellular reprogramming. Cellular reprogramming has far-reaching therapeutic potential with the promise of generating new cells to repair tissues throughout the body. Data-guided mathematics present a valuable intersection through which to assess and improve reprogramming outcomes. Computational approaches incorporating perspectives ranging from synthetic biology to information and control theory have made considerable advancements towards fine-tuning the prediction of optimal sets of factors and conditions that will ensure efficient cell conversion.

Chapter III, "Partial Reprogramming of Colorectal Cancer Cells Towards Treatment Sensitivity", examines how TF silencing partially reprograms colorectal cancer (CRC) cells to a treatment-sensitive state. In CRC, aberrant Wnt signaling leads to constitutive activity of the Wnt genetic program including several genes belonging to the T cell factor (TCF) family. The overexpression of TCF4, one of the main downstream effectors of the Wnt signaling network, in CRC cells has previously been implicated in mediating resistance to chemoradiotherapy [162]. In this work, we used RNAi-mediated silencing to knockdown *TCF7L2*, the gene encoding TCF4, in CRC cells and collected paired bulk RNA-seq and Hi-C data over a 72 hour time series to understand the changes in genome structure and function that facilitate the development of this treatment-sensitive state. We used properties of linear algebra and spectral graph theory to find that the regulation of the epithelial-to-mesenchymal transition (EMT) and E2F networks were most impacted by TCF4 silencing. Evaluating the dynamics of these two networks revealed additional TFs with the potential to further interfere with the transcriptional program of CRC cells. This work emphasizes how TF silencing can transform the integrity of a cancer cell and drive it towards a destabilized state for better intervention and that further disturbance towards a senescent or apoptotic state could be achieved by combinatorially targeting TFs critical to relevant pathways.

Chapter IV, "Rearrangement of T Cell Genome Architecture Regulates GVHD", explores the genome architecture of mature T cells and evaluates the effect of disrupting genome architecture on the regulation of mature T cell function. Mature T cells are critical to immunopathologies like graft-versus-host disease (GVHD) and we set out to understand how such a disease state would respond in the absence of a key regulator of genome architecture. We again employ principles from spectral graph theory to extract changes in genome structure and function from population-level Hi-C and RNA-seq data of mature T cells in the presence and absence of the cohesin release factor, WAPL, *in vivo* at baseline and following non-antigen and antigen stimulation. We observed dysregulation of cell cycling regions of the genome that were accompanied by a reduction in GVHD

in the absence of WAPL. Void of a temporal dimension, this study does not fully embody the 4DN, but nonetheless identifies coupling between genome structure and function that characterizes a clinically-relevant disease state. The clinical implications of WAPL regulating *in vivo* immune responses mediated by T cells in this work inspire us to consider what could be possible if we couple TF-mediated cellular reprogramming with direct alterations to genome architecture.

Chapter V, "Obesity Disrupts Innate-Adaptive Immune Network Patterning in Adipose Tissue", describes the shifting adipose tissue landscape from a lean state to obesity. Immune cells and inflammation contribute to adipose tissue dysfunction, but these changes are difficult to profile outside of their native tissue context. In this study, we combined spatial transcriptomics (ST) with single cell RNA-sequencing (scRNA-seq) across a time course of diet-induced obesity in mice to capture adipose tissue immune cell niches. Through integration of these data, we identified dominant cell type signatures preserved in their anatomical context, quantified spatial gene expression patterns, performed cell type network analysis, and identified changes in ligand-receptor signaling. We further adapted Alan Turing's mathematical theory of morphogenesis [354, 270] to summarize tissue dynamics, revealing disrupted multicellular tissue function between lean and obese adipose tissue. This work introduces an extension to the 4DN framework, wherein features of genome function are preserved in their spatial context, a critical layer of information when studying disease. This work also highlights how communication between cells is crucial to the tissue microenvironment and impresses the need to view cellular reprogramming through a multicellular lens that could ultimately allow us to reprogram entire tissues.

Chapter VI,"Deciphering Multi-way Interactions in the Human Genome", proposes a computational framework for analyzing higher-order chromatin organization. We highlight the recently-developed Pore-C approach [355], a derivative of Hi-C, that leverages long-read sequencing to map genome-wide multi-way contacts as opposed to only pairwise contacts. The long-range chromatin interactions directly captured through this technique allow us to define clusters of genomic loci that colocalize for more efficient transcription, which we term "transcription clusters". Combining multi-way contact data collected from B lymphocytes, adult fibroblasts, and neonatal fibroblasts with chromatin accessibility, gene expression, and transcription factor binding data, we provide a comprehensive summary of genome-wide transcription clusters, cell-type specific architectural signatures, and the transcription factors that likely regulate these clusters. Our hypergraph representation of these multidimensional relationships allow us to compare cell-type specific chromatin architectures using hypergraph entropy and hypergraph-based similarity measures. This work constitutes a significant advancement for the 4DN framework, as multi-way chromatin interactions can uncover more complex and nuanced features of genome structure. Moreover, we gain insight into how TFs support cell type-specific conformational states, demonstrating a role that chromatin organization plays in transcriptional regulation not often appreciated in cellular reprogramming

studies.

Together, the work presented in this dissertation provide a refined 4DN framework for better understanding the dynamics of genome organization within and between cells, with applications for cellular reprogramming. Insights from this work will guide future research with implications in cancer treatment, tissue regeneration, wound healing, and personalized medicine.

1.2 The 4D Nucleome

The genome is more than a linear sequence of base pairs. Meticulous arrangement of the genome in three-dimensional (3D) nuclear space dictates regulatory relationships that drive gene function, as the spatial proximity produced by chromatin folding enables interactions between regulatory elements, like enhancers, and target genes that are non-adjacent along the linear genome sequence. Consequently, genome structure (chromatin folding) and genome function (gene activity) are inextricably linked. Moreover, this relationship is dynamically changing over time, introducing the fourth dimension at play in the 4D nucleome (4DN).

There are several classes of data that can capture 4DN properties, most notably emerging from imaging, modeling, and next-generation sequencing (NGS) [89, 214, 278]. Early efforts to characterize the 4DN leveraged fluorescence *in situ* hybridization (FISH), revealing placement of chromosomes relative to the center of the nucleus, periphery of the nucleus, or other nuclear landmarks and enabling measurement of physical distance between genomic loci [394, 35, 343]. Though limited in throughput and coverage compared to genomics-based approaches, the development of oligopaints [25], multiplexed FISH [241], and high-throughput FISH [108] have helped overcome some of those limitations. Theoretical and data-driven modeling [146] have further filled knowledge gaps, leveraging insights from the laws of polymer physics to describe constraints on possible chromatin configurations [219], adapting mathematical methods like multidimensional scaling (MDS) to predict chromatin folding and reconstruct 3D genome structures [347, 185], and testing hypotheses to propose molecular mechanisms like loop extrusion [238, 111].

Still, some of the most prominent 4DN studies have turned to genomics to characterize genome structure and function at the molecular level. Chromosome conformation capture (3C)-based approaches have been at the forefront of these studies. 3C involves cross-linking, digesting, then ligating genomic loci in close spatial proximity followed by sequencing to delineate pairwise chromatin contact frequencies [91]. Hi-C, a high-throughput, genome-wide derivative of 3C that captures pairwise chromatin interactions between all genomic loci [190] has led to remarkable insights pertaining to the hierarchical organization of DNA into chromosome territories [269, 77, 107], compartmentalization into active and inactive chromatin regions [190], and preferential contact within local domains [274, 94]. Other -omics approaches capturing one-dimensional (1D) features

like gene expression (RNA-seq), chromatin accessibility (ATAC-seq), and epigenetic marks (ChIP-seq) supplement these findings and are typically integrated with 3C-based data to provide a more comprehensive view of synergistic behavior in the cell. Sequencing-based approaches for determining higher-order chromatin interactions like GAM [22], SPRITE [263], and ChIA-Drop [405] have also garnered attention in recent years for more reliably characterizing long-range interactions involving multiple loci and uncovering associated functional roles.

The field of 4DN analysis is continually expanding and evolving. The last decade has ushered in single-cell methods to replace population-level approaches [234, 324, 271] and seen inclusion of haplotype and karyotype information to characterize allele-specific genome organization [194] and enable elucidation of cancer-specific genome organization [306], respectively, among other advances. The following work presents additional extensions, introducing analytical frameworks incorporating two new data types compatible with the notion of 4DN - spatial transcriptomics, capturing position-specific gene expression [323], and Pore-C, capturing multi-way chromatin interactions [355]. The power of the 4DN framework lies in our ability to integrate all these data modalities to paint a fuller picture of the inner-workings of the cell, offering an ideal platform for further studying cellular control.

1.3 Cancer Reprogramming

As with TF-mediated somatic cell reprogramming, cancer cells can be directly reprogrammed to another differentiated state or reprogrammed to a pluripotent state. For cancer cells, that pluripotent state is referred to as a cancer stem cell (CSC), which like iPSCs possess the ability to self-renew and specialize [149]. Yamanaka factor-driven reprogramming of cancer cells has led to considerable success in stripping cancer cells of their tumorigenicity, malignancy, and methylation signatures, yielding a cellular state with a benign phenotype capable of differentiating into normal lineages [357, 222]. Lineage-specific TFs have demonstrated comparable success as well [402, 216, 141]. In spite of such promising outcomes, as with normal cells, reprogramming efficiency in cancer cells remains low. Even more difficult to reconcile is the genomic variability from cancer cell to cancer cell in comparison to the largely fixed genetic profile among normal cells of the same lineage, rendering a 'one size fits all' reprogramming regime infeasible.

While the transcriptional landscape of a cancer cell is variable, mutational profiles reveal that there exist core modules of master regulators (MRs) consistently found across diverse cancer contexts that regulate cancer hallmarks (i.e. apoptosis, G2M checkpoint, inflammatory response) [47, 255]. It is also believed that tumor-specific MRs operate within these tumor-independent modules to maintain a dysregulated homeostatic state, making both classes of MRs attractive therapeutic targets. Pharmacological intervention has proven promising in this area as well, as the

advent of differentiation therapy demonstrated that pharmacological agents could be used to drive cancer cells to more mature cell states, disrupting the cancer phenotype by compromising their ability to self-renew [85].

It is understood that the structural landscape of cancer cells is an important factor in defining cancer cell identity [286, 404]. Marked by chromosomal translocations that enable oncogene activation, finding reprogramming factors that alter cancer-promoting spatial organization is imperative as well. Ultimately, cancer cell reprogramming requires precision in characterizing cell state and once obtained can give rise to an array of reprogramming possibilities including a state of senescence (indefinite growth arrest), apoptosis (cell death), or simply a destabilized state that can more feasibly be attenuated.

1.4 Genome Structure and Immunity

Immune cells help maintain tissue homeostasis by clearing pathogens and malignant cells. When their transcriptional program is compromised, immune cells can contribute to a number of pathologies including cancer, immunodeficiency disorder, and autoimmune disease. Immune cells undergo a myriad of processes to reach terminal differentiation and activation, many of which rely on timely genomic rearrangements to mediate regulatory interactions. This relationship between genome organization and immune response has been established most extensively within adaptive immune cells, comprised of B cells and T cells. Studies reveal a highly dynamic topological landscape that is actively responding to cues that help to gatekeep developmental stage-specific phenotypes and prevent premature maturation that can adversely alter immune response.

In particular, chromatin remodeling drives antigen receptor loci recombination, better known as V(D)J rearrangement, a process that occurs in B cells and T cells [14, 158, 364]. FISH and 3C-based analyses have attributed this genomic shuffling to compartment and domain-level spatial organization [84, 224]. Gene segments of a given immunoglobulin (Ig) or T cell receptor (Tcr) locus are held in distinct sub-domains of a common TAD until at a specific stage of differentiation, compartment switching releases key lineage-commitment TF loci (like PAX5 in B cells or BCL11B in T cells) from inactive regions, activating their expression and allowing their protein products to be recruited to the locus. These TFs in conjunction with more universal structural proteins (like CTCF and YY1) then initiate a process called ”locus contraction” in which the elements of sub-domains are brought into close spatial proximity to recombine [155]. Similar to cancer cells, there is substantial diversity among immune cells, owing to the several stages of development and vast array of gene segment combinations during recombination, estimated at more than 10^{11} [233], that encode antigen-specific receptors and thus sub-classes of antibody-producing immune cells.

Genome structure dynamics continue to play a critical role after maturation of immune cells

to facilitate proper activation. CD4+ T cells, for example, can further differentiate into one of several T helper cell subtypes upon activation of particular transcriptional signatures that are dependent in part on the repositioning of chromatin that brings key TF loci and cytokine-encoding loci together [317, 123, 137]. Additionally, beyond antigen-receptor rearrangement and mature lymphocyte activation, genome structure has also been implicated in guiding clonal expansion, somatic hypermutation, and allelic exclusion [293, 57, 136].

Overall, TFs play a central role in regulating immune response through genome structure. Moreover, key lineage-specific TFs have been demonstrated to drive immune cell commitment, in a manner independent of transcription [301, 159]. Such observations go to show that genome structure is essential to shaping cell identity and provide a proof of principle that lineage-specific TFs associated with these conformational dynamics are viable reprogramming targets. Some models suggest that such TFs inevitably carry properties of global genomic organization, namely the ability to maintain the epigenome and transcriptional program specific to a cell [237].

1.5 Tissue Reprogramming

Tissues are composed of diverse and differentially localized cells that form distinct patterns throughout the tissue. This heterogeneous cellular composition and crosstalk at the molecular level across the tissue give rise to the emergence of tissue function. Understanding the spatial context of a cell, including its precise position in the tissue, relationship with surrounding cells, and its tissue-specific properties is critical to the development of effective and timely therapeutic interventions.

Such spatial tissue profiling has already seen strides in mapping the trajectory of developing tissue [11], elucidating the tissue architecture of cancers [223, 156, 186], and exploring the spatial dynamics of other pathologies [210]. This is not to suggest that a core transcriptional program conserved across tissue types and even disease contexts is not integral to therapeutic discovery, but that recent technologies like spatial transcriptomics [323, 273] have uncovered new layers of information that enable us to detect more nuanced dynamics. These technologies have the potential to guide cellular reprogramming towards engineering tissue or restoring cellular function through *in situ* treatments with higher fidelity.

Computational frameworks for studying and integrating these data with other data modalities further invites the application of rigorous mathematical principles like tensor and network theory to deconvolute the nested network that is tissue comprised of cells driven by their respective nuclear networks, as well as Alan Turing's mathematical theory of tissue morphogenesis, modeling tissue function in terms of a linear system of equations that capture the production and diffusion of morphogens within and between cells, respectively [353, 270]. Ultimately, the translational suc-

cess of tissue reprogramming requires precise biomanufacturing, employed to mimic aspects of the native tissue microenvironment from inter-cellular behavior down to the construction of the extra-cellular matrix (ECM) and 3D tissue architecture [43]. This has been achieved with recent success through 3D bioprinting [183, 1, 407, 199, 310] which involves positioning cells and biomolecules on a tissue scaffold and incorporating biological structural components to give rise to functioning multicellular tissue. Beyond regenerative tissue therapy, using cellular reprograming to artificially construct tissue also has implications in *in vitro* modeling and drug screening [43, 6].

CHAPTER 2

Cellular Reprogramming: Mathematics Meets Medicine

This chapter is based on a paper by Gabrielle A. Dotson, Charles W. Ryan, Can Chen, Lindsey A. Muir, and Indika Rajapakse [101].

2.1 Abstract

Generating needed cell types using cellular reprogramming is a promising strategy for restoring tissue function in injury or disease. A common method for reprogramming is addition of one or more transcription factors that confer a new function or identity. Advancements in transcription factor selection and delivery have culminated in successful grafting of autologous reprogrammed cells, an early demonstration of their clinical utility. Though cellular reprogramming has been successful in a number of settings, identification of appropriate transcription factors for a particular transformation has been challenging. Computational methods enable more sophisticated prediction of relevant transcription factors for reprogramming by leveraging gene expression data of initial and target cell types, and are built on mathematical frameworks ranging from information theory to control theory. This review highlights the utility and impact of these mathematical frameworks in the field of cellular reprogramming.

2.2 Introduction

In 1989, pioneering work by Hal Weintraub demonstrated conversion of human skin cells into muscle cells through overexpression of a single transcription factor, *MYOD1*. In 2007, Shinya Yamanaka reprogrammed human skin cells into induced pluripotent stem cells (iPSCs) using four transcription factors, a discovery that would earn him a Nobel Prize in 2012. These remarkable findings demonstrated that the genome is a system that can be controlled via an external input of

transcription factors. Following these experimental discoveries, engineering, statistical and mathematical methods predicting candidate transcription factors for cellular reprogramming began to emerge.

In this review, we show how experimental methodologies for cellular reprogramming and mathematics have come together to enhance our understanding and control of cell fate conversions. We start with a comprehensive overview of key milestones in cellular reprogramming before touching on the translational impact in the field thus far. We then offer a mathematical perspective of cellular reprogramming, covering the use of mathematical principles to describe cellular states and their dynamics. Additionally, we explore how we can harness mathematics to improve the specificity and efficacy of cellular reprogramming by summarizing existing computational frameworks designed to facilitate and validate reprogramming methodologies. Finally, we discuss future applications and implications of engaging mathematics in the domain of cellular control.

2.3 Background

The premise of cellular reprogramming arose following the discovery of DNA as the "genetic code", with all cells in an organism possessing the same sequences of DNA. From this point forward, biologists sought to understand how cells with the same underlying code take on different phenotypes. From neurons and cardiomyocytes to dermal epithelia, cells with identical DNA somehow differentiate into subsets of cells with complementary functions in a beautifully coordinated process. These specialized cells coalesce to form tissues with emergent functions, which in turn form organs and ultimately organisms. While this process is tightly controlled with awe-inspiring fidelity, errors do occur leading to a wide spectrum of human pathology. Errors along the developmental process, either due to germline mutations or environmental exposures, may lead to developmental malformations with high morbidity and mortality. In addition, tumorigenesis is often considered the de-differentiation of a cell, wherein a cell loses its identity and fails to conform to its intended function [390]. Finally, tissues begin to fail with age as the epigenetic factors that define cell identity wane, rendering a need for restorative cellular reprogramming techniques [167, 160]. For these reasons and more, it is necessary to develop our understanding of the underlying mechanisms of cell fate decision making, and to use that knowledge to improve methods for control of cellular differentiation.

2.3.1 Cellular Differentiation

Typically, cells are envisioned as having a fixed identity capable of performing specialized functions. Indeed, fully differentiated cells show remarkable functional stability over a range of

physiological conditions. However, cells do not start off this way but rather develop from precursor cells called stem cells. The earliest stem cells, embryonic stem cells (ESCs), form following conception of a zygote and are totipotent, meaning they can give rise to all cell types. These totipotent cells proliferate, and some of their progeny take on specialized identities through a process called differentiation. As a cell differentiates, the spectrum of possible progeny cells that it can give rise to narrows and it begins to express only those genes needed for its subsequent role. As a zygote develops, gradients of cell signals across the principle axes of the organism lead to differential expression of transcription factors, which modify the expression levels of genes in such a way as to give rise to distinct phenotypes, thus differentiating cells from one another (Figure 2.1A). Over time, this difference in gene expression caused by transcription factor availability is thought to be reinforced by epigenetic markers, which package the DNA depending on the needs of the cell. Actively expressed genes are left relatively unpacked (euchromatin) while silenced genes are tightly stowed away (heterochromatin) [7].

The mechanism by which cells suppress unused genes as they differentiate and specialize their function remains an open question, though biologists have proposed theories since the mid-twentieth century. A popular model of unidirectional differentiation was postulated by Conrad Waddington in 1957 suggesting that stem cells can be thought of as a high-energy state in which all genes are active [367]. As cells differentiate, genes unnecessary for the cell's subsequent role were thought to be silenced as the cell enters a lower-energy stable state, analogous to a ball rolling down a hill. While an insightful model for its time, this analogy of a ball rolling down a hill implies a unidirectional fatalistic process. Cellular reprogramming challenges this notion of cell state permanence, and allows a cell's fate to be redirected. This requires us to level the hierarchical epigenetic landscape [176] envisioned by Waddington and instead adopt a flattened landscape in which cells are free to shift from one cell potency and state to another, under the right conditions (Figure 2.1B-C). That is, one or more core genetic elements may be manipulated in a way that forces a cell to take on a new form and function.

Typically, this cellular transformation is facilitated by an exogenous stimulus like the addition or removal of one or more transcription factors. Transcription factors (TFs) are DNA-binding proteins that modulate how often a gene is transcribed into messenger RNA (mRNA) for protein production. Of the roughly 20,000 known proteins in the genome, around 1,500 are known to function as TFs [144]. Therefore, focusing on TFs as target inputs significantly reduces the dimensionality of the problem. Notably, the exogenous expression of only one to three TFs is often sufficient to achieve the acquisition of specialized lineages from other cell types [339].

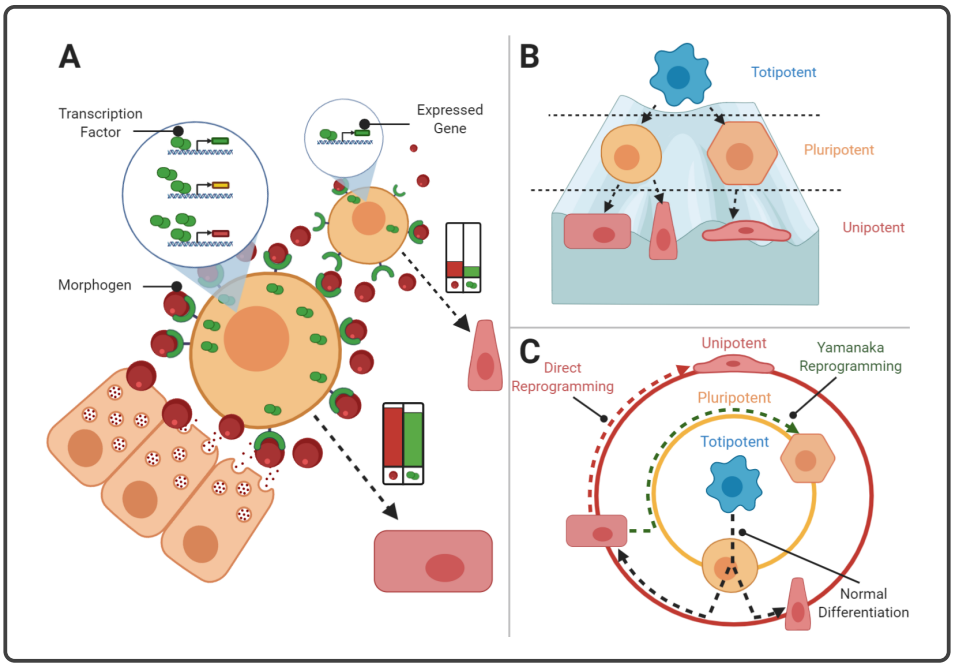

Figure 2.1: Cellular Differentiation and Reprogramming. (A) Cellular differentiation in developmental biology. During normal development, concentration gradients of morphogens (red spheres) lead to differing levels of transcription factor (dimeric green spheres) activation in a distance-dependent manner. This leads to differing transcription profiles of cells as a function of spatial location, allowing for body patterning. (B,C) Re-imagining of Waddington's epigenetic landscape. Waddington's original model (B) best describes the process of cellular differentiation and presents an intuitive illustration of non-specialized cell types traversing down peaks and settling in valleys once specialized. In a re-imagined version of Waddington's landscape (C), the essence of cellular reprogramming is captured, where a cell is unimpeded by the gravity of a hierarchical model and can flow between any potency, germ layer, and cell state. The common center of the concentric circles represents the totipotent state while the subsequent inverted circles represent decreasing levels of cellular potency moving outwards. Here, direct reprogramming is analogous to 'direct conversion' or 'transdifferentiation', and refers to a change in cell fate that does not incorporate a pluripotent or progenitor state.

2.3.2 Cellular Reprogramming

This notion of cellular reprogramming is far from a theoretical whim. Rather, it has been established over the last four decades in several contexts. Reprogramming has been achieved by introducing critical TFs to susceptible cell types and by transferring the nucleus of a somatic cell into an oocyte to "rejuvenate" the nucleus into a totipotent zygote.

Epigenetics	The study of heritable changes in gene activity caused by mechanisms other than changes in the underlying DNA sequence. These can include biochemical modifications of genomic DNA or histones, the proteins that help package genomic DNA
Differentiation	The unidirectional progression of a cell towards a specialized cell type
Reprogramming	The directed transformation of one cell type into a less or differently specialized cell type, typically mediated by exogenous TF expression
Cell Potency	The potential for a cell to develop into a more specialized cell type. The more lineages the cell can give rise to, the higher the potency and the closer it is to a stem cell-like state
Reprogramming Factor	One or more transcription factors capable of driving a cell from one state to another when exogenously overexpressed
Gene Regulatory Network	Cell-type specific set of interactions between genes, governed by the regulatory influence they exert on each other

Table 2.1: Glossary of Terms

2.3.2.1 Transcription Factor Mediated Reprogramming

In 1989, Harold Weintraub *et al.* demonstrated that non-muscle cells could be driven to express muscle-specific genes by *MYOD1* activation [378]. This was the first success in directly reprogramming one differentiated cell to another and demonstrated that TFs such as *MYOD1* possess the capability to override expression programs. This inspired others to search for additional TFs capable of driving differentiation. In 2006, Shinya Yamanaka's group redefined what was possible when they demonstrated that murine adult fibroblasts could return to a pluripotent state following lentiviral transduction of just four TFs: *Oct3/4*, *Sox2*, *Klf4*, and *c-Myc* [341]. These factors would be coined the "Yamanaka factors" and Yamanaka would be awarded the Nobel Prize in Physiology or Medicine in 2012 for this groundbreaking discovery. Not only were these cells transcriptionally similar to embryonic stem cells, but they successfully differentiated into all three germ layers - endoderm, ectoderm, and mesoderm - when introduced into murine blastocysts, indicating true pluripotency. The following year, the Yamanaka factors were again used to convert human fibroblasts to iPSCs whose differentiation to neurons and cardiomyocytes could be subsequently induced, and that could form teratomas with all germ layers present when injected into nude mice [340]. These results indicate that cell identity remains malleable through adulthood, and when the correct TFs are applied, cells may be driven towards identities that meet experimental or therapeutic needs.

2.3.2.2 Somatic Cell Nuclear Transfer

Another route to reprogram cells is by transferring the nucleus of one cell type into the cytoplasm of another. The cell-signaling molecules in the recipient cytoplasm, including TFs, act to reprogram the donor nucleus to exhibit an expression profile more reminiscent of the recipient cell. The oocyte is an ideal recipient cell - its large cytoplasmic volume buffers it against the donor signaling molecules, as it contains sufficient quantities of TFs to overpower those contained in the donor nucleus. In addition, the oocyte's role in generating the embryo and resetting the sperm nucleus suggests that it contains factors necessary for nuclear reprogramming to a totipotent state. These ideas became reality when in 1952, it was demonstrated that a nucleus from a blastula cell may be introduced into an enucleated frog egg to give rise to a normal embryo [38]. Following this discovery, John Gurdon demonstrated that even nuclei of intestinal epithelial cells were capable of giving rise to an entire embryo when implanted into an oocyte [126]. In 1996, the first mammalian clone was produced by transferring sheep mammary nuclei into an enucleated oocyte [366]. Finally, in 2013, the first hESC cell line was produced using somatic cell nuclear transfer techniques [338].

These nuclear transfer experiments highlight a potential application to regenerative medicine, where a nucleus derived from a patient skin cell could be introduced into an oocyte for reprogramming into embryonic stem cells, which may in turn be differentiated into cell types of clinical interest. Since the generated cells would contain DNA of the graft recipient, this would provide an avenue for autologous tissue grafting without concern for rejection.

2.3.2.3 Translational Success

Recently, cellular reprogramming has made the journey from bench to bedside as efficiency and safety have improved. The therapeutic potential of autologous grafts derived from reprogrammed cells is broad. Easily accessible yet malleable cells such as dermal fibroblasts, for example, may be reprogrammed to various cell types for a range of downstream medical applications (Figure 2.2). One such example is the application of reprogrammed cells to treat age-related macular degeneration. As described in a 2017 report, retinal pigment epithelium (RPE) cells were derived from fibroblasts induced with non-integrating episomal vectors carrying TFs to treat a 77-year old patient with wet macular degeneration [209]. A skin sample was taken and reprogrammed to an iPSC state, and subsequently differentiated into RPE cells. These reprogrammed cells exhibited DNA methylation and gene expression patterns consistent with RPE cells. Additionally, genome sequencing indicated that no mutagenesis had occurred during the process, mitigating the possibility for oncogenesis. Since these cells are autologous and therefore carry the DNA of the patient, there is no risk of graft rejection upon re-introduction. The reprogrammed cells were surgically grafted

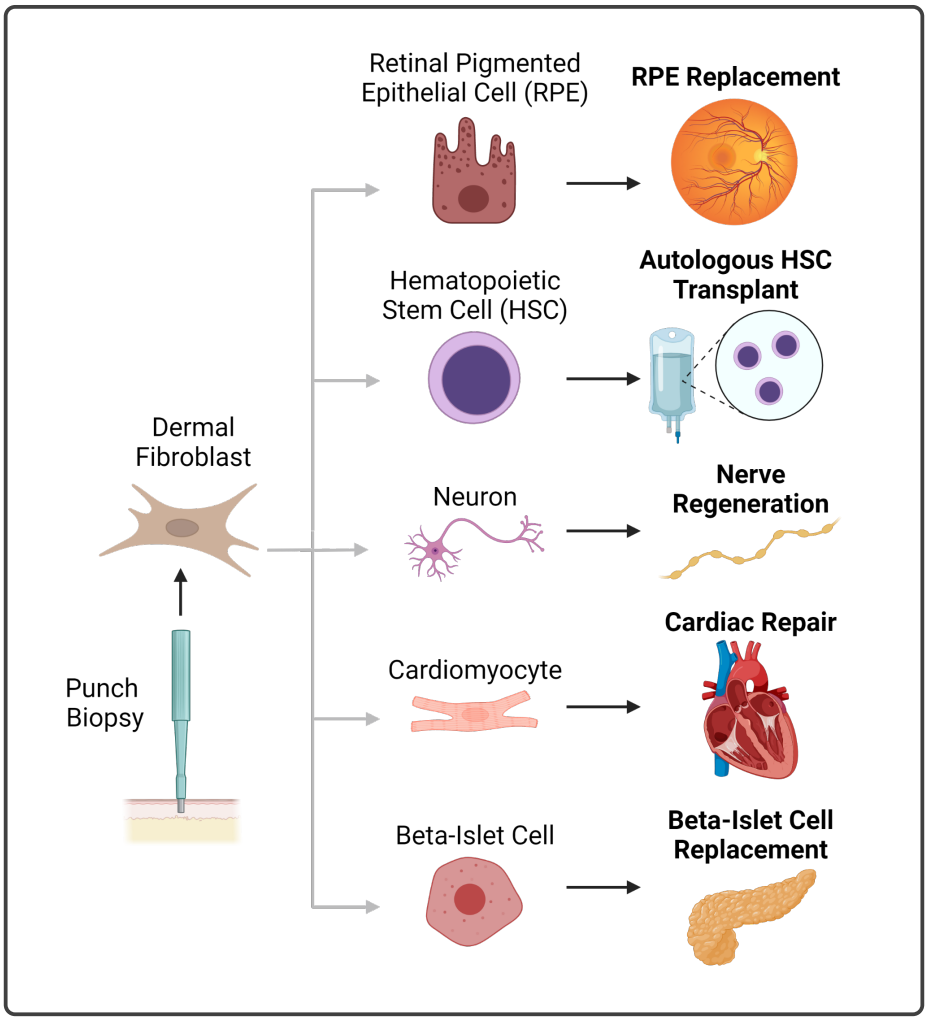

Figure 2.2: Medical Applications of Cellular Reprogramming. Dermal fibroblasts can be acquired via a minimally invasive punch biopsy, and subsequently differentiated into a desired cell type via the addition of select TFs. Finally, autologous cultured cells of the desired type can be reintroduced into the patient without concern for graft rejection.

into the patient's retina and found to have engrafted successfully after one year without evidence of vision loss. This early success demonstrates the clinical potential of cellular reprogramming using non-integrating TF delivery methods, as an autologous graft was generated from reprogrammed cells without introducing potentially harmful mutations.

Clinically-relevant reprogramming successes in laboratory settings have presented exciting potential as well. For example, critical steps have been made towards autologous hematopoetic stem cell therapies for post-chemotherapy leukemia patients, as murine committed lymphoid and myeloid progenitors [283] as well as human fibroblasts [314] were converted into hemogenic cells. This success extends to the field of nerve regeneration therapy, as fibroblasts were reprogrammed

into glutaminergic neurons [393]. Yet another clinically-relevant success is the differentiation of human fibroblasts to cardiomyocytes [50, 110], as the ability to cultivate autologous cardiomyocytes may allow replacement therapy following myocardial injury in the future. Finally, with type-I and late-stage type-II diabetes mellitus characterized by the loss of insulin-producing beta-islet cells, success in generating autologous beta cells has curative potential. The feasibility of this solution was recently demonstrated with the conversion of pancreatic acinar cells to beta-islet cells [55] as well as the conversion of fibroblasts to beta-islet-like cells via directed differentiation through endodermal intermediates [409]. Taken together, these results indicate that an easily accessible and malleable cell type such as the fibroblast may be converted into a range of clinically applicable cell types for autologous grafting, potentially challenging the permanence of diseases ranging from macular degeneration to diabetes mellitus.

2.4 Mathematics of Cellular Reprogramming

Cellular reprogramming is a very calculated process involving inputs, transition states, and outputs - a control theory problem at its core. Long before control theory concepts would be implemented in cellular reprogramming studies, however, mathematicians were already conceptualizing how biological systems could be systematically controlled [200]. From observing state-dependent network entropy during differentiation [267] to modeling cells as complex networks subject to perturbations that can modulate the system's equilibrium and placement in an n-dimensional state space[75], it became clear that mathematics could be exploited to optimize cellular reprogramming.

2.4.1 Cellular States

During cellular reprogramming, cells encounter a start state and end state, occupying one or more transition states in between. These states are most commonly described in terms of their gene expression profiles or gene regulatory networks [46, 104, 264, 92, 288]. Derived from RNA-sequencing data, gene expression can be represented as a vector of n non-negative values, where n is the roughly 20,000 genes in the human genome. A single expression vector represents a state-dependent snapshot of a genome's transcriptional landscape in time. Collecting data as the cell evolves can then give rise to a sequence of expression profiles reflecting discrete points along a reprogramming trajectory. While gene expression has provided sufficient cellular state representation thus far, additional measurements such as those from ChIP-seq, DNase-seq, or chromosome conformation data can contribute to a more comprehensive definition of cellular states and may be considered in future studies [348].

Entropy plays an important role in cellular dynamics and offers a quantitative measure of a cell's differentiation status (i.e., a cell's placement within the Waddington landscape). One frame of thought is that the specialization of cells minimizes entropy and thus controllability [267, 268]. The closer a cell is to a pluripotent state, the higher its entropy - as there exists a wider availability of signaling pathways that can be activated and conformations the cell can adopt (more uncertainty) in the lesser committed state [206, 76]. We can glean local and global entropy from data on gene expression and protein networks [16]. Local entropy indicates the susceptibility of specific signaling pathways in the cell to perturbation and is derived simply from the Shannon entropy for a particular gene or protein. Global entropy indicates how specialized the cell is and can be captured by finding the average of those local entropies.

2.4.2 Cellular State Perturbations

Perturbations push cells to overcome the epigenetic barriers that consign them to a single differentiated state [207]. In the context of cellular reprogramming, these perturbations are often the external introduction of one or more TFs that evoke changes in transcriptional activation and inhibition throughout the cell's gene regulatory network. The internal response to these external perturbations, however, is not fully deterministic. That is, perturbations may drive the cell towards the vicinity of a target state (i.e., a basin of attraction) with the expectation that the cell will converge to the intended target state (attractor) in a stochastic manner.

In their work on perturbation patterns and network topology, Marc Santolini and Albert-László Barabási demonstrate how topological models can offer substantial insight into accurately predicting the influence of perturbations on biological networks [298]. Santolini and Barabási explain that the effects of external stimuli do not diffuse throughout the genome uniformly. Rather, perturbation spread is confounded by interaction properties like directedness, modularity, or whether an interaction is activating or repressive. As such, every node in a biological network will correspond to a differential equation with parameters encompassing these properties. Network topologies can then be captured from a set of differential equations by constructing the Jacobian matrix, a matrix of partial derivatives that describes the impact of node i on the activity of node j and that quantifies direct interactions between all pairwise nodes in the network. An adjacency matrix, describing the influence of the network, can subsequently be derived as $\mathbf{A} = \text{sign}(\mathbf{J}^\top(\mathbf{x}^*))$, where the *sign* function represents the directionality of the edges between nodes and is applied element-wise, $\mathbf{J}^\top$ is the transpose of the Jacobian, and $\mathbf{x}^*$ is a vector of magnitudes (i.e. gene expression) representing the steady state of the system. To account for indirect interactions in a network, Santolini and Barabási also describe a sensitivity matrix, $\mathbf{S}$, where each element is the full derivative between two nodes.

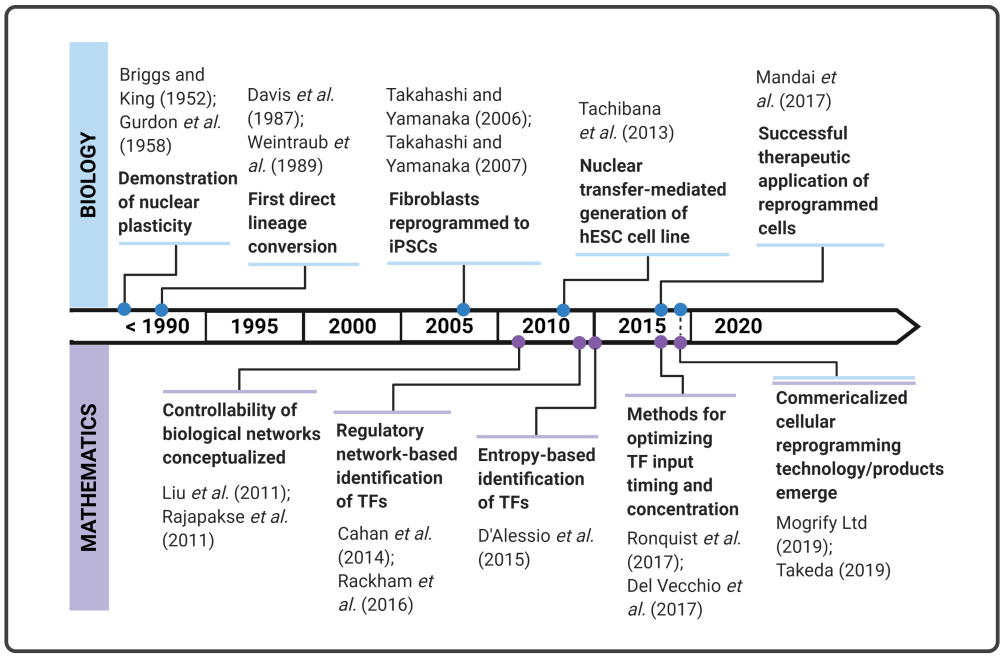

Figure 2.3: Timeline of Key Experimental and Computational Cellular Reprogramming Advancements.

The relationship between the sensitivity and Jacobian matrices becomes:

$$\mathbf{S} = (\mathbf{I} - \mathbf{J})^{-1}\mathbf{D}(1/(\mathbf{I} - \mathbf{J})^{-1}),$$

where, $\mathbf{I}$ is the identity matrix, $\mathbf{D}$ is the diagonal operator, and the $/$ operator denotes element-wise division. By modifying (perturbing) the differential equation of nodes of interest (TFs) and analytically deriving the sensitivity matrix, we get a sense of how that node's influence propagates throughout the network. Taken together, this predicative framework facilitates evaluation of how cells might react to a perturbation, providing an avenue for which to test the influence of candidate reprogramming factors.

2.5 Computational Tools

The earliest cellular reprogramming studies incorporated a systematic "guess-and-check" method for identifying TFs that would be necessary and sufficient for reprogramming in their experimental system. The four Yamanaka factors were narrowed down from an initial list of 24 candidate reprogramming factors that were selected based on their properties and suspected involvement in promoting and maintaining an embryonic stem cell-like state. It took several trials of transducing different permutations of these factors to finally converge upon the four factors capable

of inducing pluripotency in differentiated cells. Several computational frameworks for streamlined data-guided prediction of reprogramming factors have since been developed (Table 2.2), demonstrating that we can bypass the "guess-and-check" method of validating TFs, which tends to be time consuming and cost prohibitive.

2.5.1 CellNet

The successful realization of a target cell identity following reprogramming can be assessed by establishing transcriptional similarity using gene expression profiling and hierarchical clustering analyses. This validation approach, however, fails to capture functional dissimilarities that may exist between the reprogrammed cell and its naturally occurring counterpart. A more quantitative and rigorous system is therefore required to verify the integrity of reprogrammed cells. To this end, in 2014, Patrick Cahan and co-author Samantha Morris *et al.* introduced CellNet, a computational tool employing network biology to assess and improve the fidelity of reprogrammed cells [46]. In a companion paper, Morris and Cahan put CellNet to the test with B cell to macrophage and fibroblast to hepatocyte-like cell conversions, experimentally verifying that original reprogramming schemes could not achieve full conversion but that CellNet-refined schemes could [227].

As described in the original publication, CellNet first generates a training dataset containing cell- and tissue- (C/T) specific GRNs derived from a large array of publicly sourced microarray datasets. The CellNet algorithm has since been extended for compatibility with RNA-sequencing data [265]. Subsequently, based on input gene expression data from a reprogramming experiment, CellNet classifies the query cell by its most likely C/T-specific GRN and evaluates how close it is to the target phenotype's GRN. If the two GRNs deviate significantly, CellNet proposes TFs that could further drive the reprogrammed cell to better mimic its target identity. Therefore, CellNet's measure of reprogramming success is the extent to which a reprogrammed cell has established its target state's GRN. To evaluate the reprogrammed cell's GRN in comparison to the target cell type's GRN, the CellNet algorithm computes a GRN status score as follows:

$$\text{RGS} = \sum_{i=1}^{n} \text{Zscore}(\text{gene}_i)_{\text{C/T}} \times (\text{gene weight}_i)_{\text{GRN}},$$

where, n is the number of genes in the reprogrammed cell's GRN, the z-score of gene i is derived from the distribution of expression values for gene i in the C/T-specific training data, and the gene weight is the expression of gene i as determined from the reprogrammed cell's expression profile. If the reprogrammed and target GRNs are not equivalent, CellNet then investigates how transcriptional regulators (TRs) of the target GRN can push the reprogrammed cell to fully convert.

This is determined by computing a network influence score (NIS) for TRs (i.e., TFs) in the GRN:

$$\text{NIS(TR)} = \sum_{i=1}^{n} (\text{Zscore}(\text{target}_i)_{\text{C/T}} \times \text{weight}_{\text{target}_i}) + n \times \text{Zscore(TR)}_{\text{C/T}} \times \text{weight}_{\text{TR}},$$

where, n is the number of genes in the target C/T-specific GRN, the z-score of the target is derived from the distribution of expression values for gene i in the training data, $\text{weight}_{\text{target}}$ is the mean expression value of gene i in the training data, the z-score of the given TR is based on the distribution of expression values for the TR in the training data, and $\text{weight}_{\text{TR}}$ is the mean expression value of the TR in the training data.

With CellNet, Cahan *et al.* uncover variations in reprogrammed cells' regulatory networks compared to the target state thereby highlighting inherent limitations in specificity that accompany current reprogramming methods. This technique is unique in that it seeks to address both how well a reprogrammed cell recapitulates its target and how the administered combination of TFs can be revised to better recapitulate that target - what Cahan later designates the 'assessment' and 'improvement' problems, respectively [45].

2.5.2 D'Alessio *et al.* Method

In contrast to CellNet, a later method for identifying reprogramming factors described by D'Alessio *et al.* in 2015 incorporates only the expression of TFs instead of genome-wide expression data in its prediction algorithm. [104]. Specifically, TFs are ranked by their specificity and uniqueness, where a given TF is highly ranked and ideal for reprogramming if it is highly expressed in the query (i.e., target cell type) dataset and not expressed in cell types comprising the background dataset. To identify such TFs, D'Alessio *et al.* define the distribution of a TF's expression level across cell types, such that the expression of a TF in a query cell type may be compared to the expression of that TF in all other cell types. The cell type specificity of each TF is evaluated using an entropy-based score adapted from previously described applications of Jensen-Shannon (JS) divergence for describing tissue-specific gene expression [349] and lincRNA expression [42]. Specifically, a distance metric describing the difference between a query expression pattern and the ideal expression pattern is obtained by taking the square root of the JS divergence as follows:

$$\text{JS}_{\text{dist}}(\mathbf{e}, \mathbf{e^I}) = \left[\mathbf{H}(\frac{\mathbf{e} + \mathbf{e^I}}{2}) - \frac{\mathbf{H}(\mathbf{e}) + \mathbf{H}(\mathbf{e^I})}{2} \right]^{1/2},$$

where, $\mathbf{e}$ is an n-dimensional vector, $\mathbf{e} = [e_1, e_2, \ldots, e_n]$, consisting of the expression-derived abundance density of a specific TF across n cell types. Vector $\mathbf{e^I}$ represents the ideal distribution in the form of a binary n-dimensional vector where $e_i^{\mathbf{I}} = 1$ if i = query cell type and 0 for all

other cell types. Lastly, $\mathbf{H}(\cdot)$ represents the Shannon entropy. TFs with cross-cell type expression distributions closely aligned to the ideal distribution, (i.e., expressed in the query cell type but not in other cell types) will yield JS distances close to zero. D'Alessio *et al.* identified the top ten TFs with expression profiles most similar to the ideal expression profile for nearly 200 cell type and tissue pairs using this approach. Experimental validation of TFs predicted to define RPE cells demonstrated successful reprogramming from fibroblasts into RPE cells. This success lends credence to the identification of reprogramming factors based on the influence that TFs have on the identity of a desired cell type, as determined by the level of unique and relative expression.

2.5.3 Mogrify

Similar to CellNet, Mogrify leverages regulatory networks to predict TFs for direct conversion of cells, selecting TFs that exert regulatory influence on genes that identify most with the target cell type [264]. Using publicly available gene expression data, the Mogrify algorithm starts by accumulating TFs that are differentially expressed between the target cell type and a background set of non-target cell types. A tree-based approach is employed to ensure that the background set is not saturated with cell types that are too highly or distantly related to the target cell type. Selected TFs are subsequently ranked based on the combination of their differential expression (scored as the product between the log-transformed fold change and adjusted p-value) and local network influence based on protein-DNA and protein-protein interaction data. This network influence is derived as follows:

$$N_x = \sum_{r \in V_x} G_r \times \frac{1}{L_r} \times \frac{1}{O_r},$$

where, $r \in V_x$ are genes (r) that are nodes (V_x) in the local subnetwork of TF x, G_r is the aforementioned differential expression-based score for gene r, L_r is the number of edges between gene r and TF x in the local subnetwork, and O_r is the out-degree of gene r's parent node in the network. This score is determined for the network of known gene protein product interactions and separately for the network of known TF-DNA interactions. With the described distance-based and connectivity-based weightings, Mogrify carefully selects for TFs with high influence and regulatory specificity. The subsequent combination of gene and network scores yields a set of TFs that are highly ranked if they are non-ubiquitous with a high fold change and low p-value. To further prune the list, TFs that regulate over 98% of the same genes are deemed redundant and the higher ranked TF preserved. Additionally, TFs in the ranked list that are already highly expressed in the starting cell type are removed from consideration.

While Mogrify's systematic identification of reprogramming factors is only as good as the fidelity of publicly available gene expression data and breadth of known TF binding interactions, it

offers robust predictive power in mediating direct cell conversion, having recovered 84% of previously discovered conversion factor cocktails and validated two novel conversions at its time of publication. Mogrify-enabled predictions for TF-mediated conversions between cell types represented in the FANTOM5 gene expression atlas have been compiled into an online directory (http://mogrify.net/), providing a convenient one-stop shop for a wide range of direct reprogramming objectives [250]. Additionally, since its inception, Mogrify has undergone its own transformation from technology to corporation, now leading the way in commercialized direct human cell conversion and revolutionizing cell therapy.

2.5.4 Del Vecchio *et al.* Method

The natural dynamics of a cell's GRN are not always amenable to preset overexpression of TFs [228, 302, 120, 388]. Del Vecchio *et al.* address this uncertainty in reprogramming success by employing mathematical modeling to facilitate the design of a genetic feedback controller, independent of GRN dynamics, that iteratively adjusts TF input concentrations based on the difference between current and target state TF concentrations [92]. In this method, a cellular GRN ($\mathbf{x}$) with n TFs, $\mathbf{x} = [x_1, \ldots, x_n]$ is modeled as a set of ordinary differential equations represented as:

$$\frac{dx_i}{dt} = H_i(\mathbf{x}) - \gamma_i x_i + u_i \text{ for } i \in \{1, ..., n\},$$

where $H_i(\mathbf{x})$ is the Hill function capturing the GRN regulation of TF x_i, γ_i is the constant decay rate of TF x_i, and u_i is a non-negative scalar indicating the input perturbation corresponding to TF x_i. In contrast to open loop control which takes a constant or time-varying TF concentration as input, the authors employ closed loop feedback control which updates the input concentration for a system in real time. Since the feedback control is dependent on the error between the most current concentration, x_i, and the target concentration, x_i^*, the u_i term in the former equation can be replaced with the expression $G_i(x_i^* - x_i)$, where G_i is a large positive constant representing gain that renders H_i and $\gamma_i x_i$ negligible. The resulting form describes the simultaneous production and degradation of TF x_i mRNA that govern the behavior of Del Vecchio *et al.*'s novel synthetic genetic controller circuit. Operation of this circuit is facilitated by two inducers, one acting on a synthetic copy of gene x_i and the other on siRNA complementary to both the endogenous and synthetic mRNA, where the inducer concentrations are informed by x_i^*. Implemented for each TF in the GRN, or a predetermined subset, the circuit steers the concentration of a given TF x_i and its mRNA, m_i, towards x_i^* and m_i^* by inducing the concurrent production of the synthetic TF and degradation of the endogenous TF. Once the desired concentration is achieved, the inducers are set to zero to stop the synthetic controller and the endogenous system takes over TF production to maintain the stability of the reprogrammed state.

At its time of publication, the proposed approach had been simulated on a two-node model of the pluripotency network but not yet demonstrated on a real cellular GRN. The authors suggested, however, that the application of their controller may ensure more success in iPSC reprogramming as pluripotent states can be difficult to attain with preset TF overexpression due to the potential for the TF to dominate the network's behavior. Though highly theoretical at this time, this work offers a promising and well thought out synthetic biology approach for fine-tuning cellular reprogramming.

2.5.5 Data-Guided Control (DGC)

Coupling gene expression and TF binding data with control theory, Ronquist *et al.*'s universal algorithm for cellular reprogramming offers a slightly more mathematically rigorous framework than its predecessors for data-guided prediction of reprogramming factors, or data-guided control (DGC) [288]. In addition to identifying candidate reprogramming factors, Ronquist *et al.* also present an optimization method for identifying the ideal time during the cell cycle for addition of TFs. The foundation of this algorithm is the following linear, control theory difference equation:

$$\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \mathbf{B}\mathbf{u}_k$$

where $\mathbf{x}_k \in \mathbb{R}^N$ is a gene expression vector representing the cell state at time k. To ease the computational burden, the authors reduce the dimensions of this vector from $\sim 20,000$ to $\sim 2,000$ ($\mathbb{R}^{\tilde{N}}$) by summing the expression levels of genes that are in close proximity to one another, or more specifically, that occupy the same cell-type invariant topological domains. $\mathbf{A}_k \in \mathbb{R}^{\tilde{N}\times\tilde{N}}$ is the transition state matrix representing time-varying changes to the cellular state at time k. $\mathbf{B} \in \mathbb{R}^{\tilde{N}\times M}$ is the input matrix indicating interactions between genomic domains ($\tilde{N}$) and TFs (M). Elements of the matrix are weighted by the magnitude of regulatory influence, determined from publicly available TF binding site data, and signed according to whether a given TF activates or represses the genes in each domain. Finally, $\mathbf{u}_k \in \mathbb{R}^M$ is a binarized input vector, with nonzero elements indicating which TF(s) to introduce at time k. An overview of this framework is reproduced in Figure 2.4.

Once the solution to the difference equation is obtained, TFs are subsequently scored by exe-

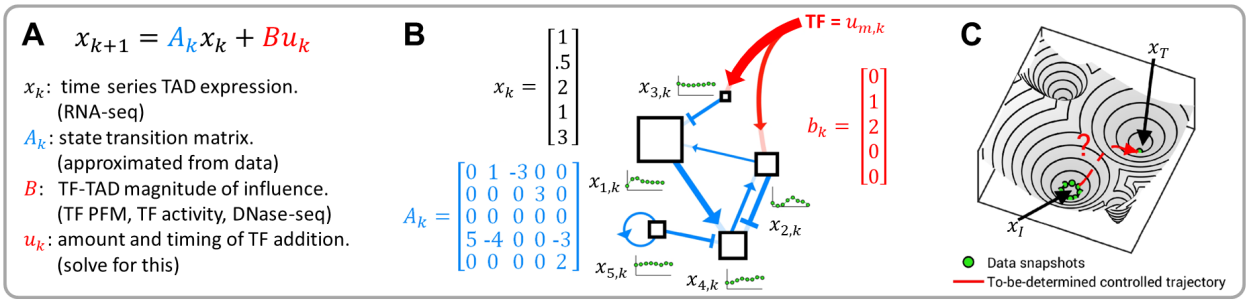

Figure 2.4: Data-Guided Control Overview. (A) Summary of control equation variables. (B) Each box represents a topological domain containing several genes. The blue connections represent the edges of the network and are determined from time series RNA-seq data. The small green plots at each node represent the expression of each domain changing over time. The red arrows indicate additional regulation imposed by exogenous TFs. (C) Conceptual illustration of determining TFs to push a cell state from one basin to another. Figure reproduced with permission from [288].

cuting the following optimization problem for all possible input signals:

$$\text{minimize}_{\mathbf{u}} \; \|\mathbf{x}_T - \mathbf{z}_F(\mathbf{u})\|$$
$$\text{subject to} \begin{cases} \mathbf{u}_{m,k}, \; k = 1, 2, \ldots, F \\ \mathbf{u}_{m,k}, \; \text{if } m \notin \hat{p} \\ \mathbf{u}_{m,k+1} \geq \mathbf{u}_{m,k} \end{cases},$$

where, the objective is to minimize the distance between the final state, $\mathbf{z}_F$, and target state, $\mathbf{x}_T$, with $\mathbf{z}_F(\mathbf{u})$ denoting that the final state is dependent on the input signal, $\mathbf{u}$. We can evaluate this distance as the Euclidean norm ($||\cdot||$) of the difference between the gene expression levels at each state. The first constraint addresses the restriction that TFs only be added to the cell and not removed. The second constraint requires all elements of input signal $\mathbf{u}_k$ to be zero if they do not correspond to the subset of TFs ($\hat{p}$) selected to drive the system. Lastly, the third constraint indicates that select TFs can be added consecutively or added sequentially, but that once added, the TFs must continue to be exogenously expressed until the final time point. Notably, Ronquist *et al.* demonstrate that TF scores are dependent on the time of input, with some TFs showing preference for addition towards the beginning of the cell cycle and others towards the end. Overall, Ronquist *et al.* present a convincing case for incorporating time-varying data into TF prediction frameworks and considering the time of TF input to a system. This approach mimics the natural course of differentiation and may increase the efficiency of reprogramming to achieve yields sufficient for tissue grafting, though experimental validation is needed.

Method	Input[1]	Output	Approach	Validation
CellNet [46]	Gene expression data and GRN information	Assessment of original reprogramming scheme and proposal for refined reprogramming scheme	Scores TFs based on GRN status and network influence	Experimentally validated
D'Alessio et al. Method [104]	Gene expression data for known TFs	Set of core TFs specific to target cell type	Scores TFs by expression specificity to target cell type based on entropy-based measure	Experimentally validated
Mogrify [264]	Gene expression data	Set of candidate TFs for conversion to target cell type	Ranks TFs based on influence over differentially expressed genes in the target cell type	Comparison to previously confirmed conversion cocktails and novel conversions were experimentally validated
Del Vecchio et al. Method [92]	TFs in target GRN	Desired endogenous TF concentrations for target cell type	Synthetic genetic feedback controller that steers TF concentrations based on discrepancy between actual and desired TF concentrations	Simulation on model of a cellular network
Data-Guided Control (DGC) [288]	Time-series gene expression and TF binding data	Set of candidate TFs for conversion to target cell type and suggested time of input	Genome dynamics for initial and target cell states modelled using control theory difference equation and TFs scored based on Euclidean distance between states under TF control	Comparison to previously confirmed conversion cocktails

[1] Data necessary to facilitate prediction of reprogramming factors. Input of initial and target cell types is assumed.

Table 2.2: Summary of Computational Tools for Reprogramming

2.6 Experimental Realization

With the guidance of predictive models such as those mentioned above, we can narrow down the list of candidate TFs to quantities achievable with modern high-throughput methods. Rather than considering any combination of the approximate 1,500 TFs in the human genome, the problem is reduced to only those with favorable predictions. Now that the problem is of a more manageable size, experimental validation is critical.

2.6.1 Transcription Factor Delivery

In order to experimentally achieve reprogramming, it is necessary to determine a method of adding TFs to cells. TFs are proteins coded by genes, and thus could be introduced to the cell directly or via DNA or mRNA coding for that protein (Figure 2.5).

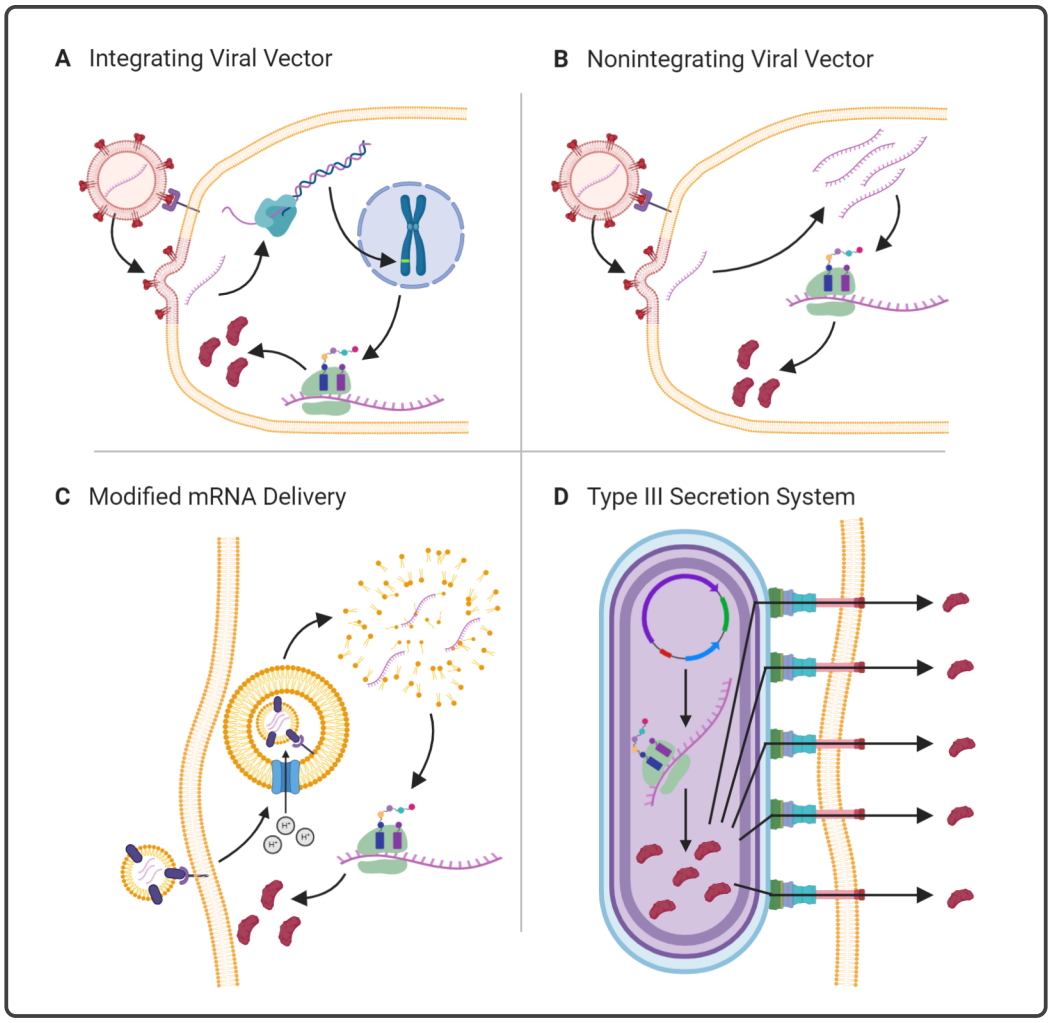

Figure 2.5: Transcription Factor Delivery. (A) Transduction of a TF-encoding gene by a DNA integrating virus such as a lentivirus. The virus first binds to the host membrane and fuses its viral envelope with the eukaryotic phospholipid membrane to enter the cell. Nucleic acids are released in the cytosol, where RNA-dependent DNA polymerase creates double stranded DNA. Viral DNA is integrated into the host genome. Host cells then express virally-delivered genes, which are translated into functional proteins. (B) Introduction of mRNA coding for a TF by a non-integrating virus such as a Sendai virus. As above, the virus docks, fuses with the cell membrane, and releases nucleic acids into the cytosol. RNA-dependent RNA polymerase then generates positive sense RNA, which is translated to functional protein. (C) Delivery of modified mRNA via a lipid nanoparticle. First, the lipid particle-embeded peptides bind target cell receptors and trigger receptor-mediated endocytosis. Next, the endosome is acidified by $\mathbf{H}^+$ pumps. The acidified endosome and lipid nanoparticle are destabilized, releasing mRNA into the cytosol. Free mRNA is translated to functional protein. (D) Bacterial type-III secretion system as a TF delivery mechanism. First, a bacterial DNA plasmid is expressed and protein is produced. Next, the bacterial T3SS delivers the protein to the eukaryotic cell through a molecular needle.

2.6.1.1 Viral Vectors

Perhaps the most common modality by which biologists introduce genes into cells is lentiviral transduction. Lentiviruses are a class of retroviruses commonly used in experimental procedures. Retroviruses, such as the lentivirus, are made up of an RNA-based genome, RNA-dependent DNA polymerase (reverse transcriptase), and DNA integrase surrounded by a protein capsid and a phospholipid envelope. These viruses function by fusing with the membrane of cells and releasing their RNA into the cytoplasm, which is reverse transcribed into DNA and integrated into the host genome [295]. These viruses serve as a useful tool, as one can package a gene of interest into their RNA genome and use these viruses to integrate that gene into the genome of a cell in culture. In addition, antibiotic selection genes and fluorescent proteins can be added to aid in selecting cells that have taken up and readily express the viral genome. Despite these advantages, insertional mutageneis may confound experimental results and increase the risk of tumorigenesis in therapeutic applications [356]. To circumvent this risk, non-integrating viruses such as the Sendai virus can be used to temporarily express exogenous TFs and initialize reprogramming processes within the cell. While this solves one problem, both retroviruses and non-integrating viruses trigger host cell innate immune responses as viral RNA is sensed by pattern recognition receptors (PRRs). This leads to a wide range of cell signaling events and transcriptional changes that alter cell behavior, potentially interfering with experimental and therapeutic applications [296].

2.6.1.2 Modified mRNA

An alternative to viral vectors is direct application of TF-encoding mRNA packaged in lipid nanoparticles (LNPs). These LNPs are taken up via receptor-mediated endocytosis, and nucleic acids are released into the cytoplasm as acidification of the endosome leads to dissolution of the lipid nano-particle and endosome [277]. An advantage of liposomal delivery of synthetic mRNA molecules over viral transduction is the avoidance of the host cell immune response, as synthetic mRNAs can be designed to mimic host mRNA and avoid PRR binding [191]. In addition, similar to non-integrating viruses, modified mRNA does not cause changes to the host genome and therefore is not associated with risk of mutagenesis of oncogenes or tumor suppressor genes. Much progress over the last decade has led to efficient protocols to generate iPSCs from differentiated cells by TF introduction via modified mRNA [374]. Nanoparticle-delivery of mRNA has also been explored as a COVID-19 vaccine candidate in human trials, highlighting the safety and stability of LNPs *in vivo* [153].

2.6.1.3 Bacterial Type-III Secretion System

A disadvantage of mRNA-based gene introduction is the lag time between initiation of translation and accumulation of functional protein due to the time it takes for a cell to translate delivered mRNA molecules into proteins of interest. In addition, the dependence of the rate of translation on cellular properties may further confound experiments, as the number of available ribosomes and tRNA substrate is not universal [290]. As suggested in the Del Vecchio *et al.* and Ronquist *et al.* methods above, cellular reprogramming experiments may benefit from introducing high concentrations of TFs at a precise point in time which presents a challenge for mRNA-delivery methods considering the time needed for translation. An emerging technique to introduce proteins directly into cells takes advantage of the bacterial type-III secretion system (T3SS). The bacterial T3SS functions as a molecular needle, puncturing the membrane of eukaryotic cells and injecting proteins that carry an N-terminal secretion signal. Originally a mechanism of toxin delivery by pathogens such as *Salmonella*, *Shigella*, and *Yersinia* species, this system has been commandeered for use in various peptide delivery experiments. This system has several advantages over nucleic acid-based approaches. First, the work of protein production is outsourced to bacterial cells, which generate and deliver the protein of interest to the eukaryotic cell at a rate that is independent of recipient cell state. In addition, the host response to foreign nucleic acids is avoided, as proteins are delivered directly without a nucleic acid intermediate. Finally, no changes to the genome are made, mitigating the risk of mutagenesis seen with lentiviral delivery methods. This system therefore has the potential to deliver TFs to cells without permanent genetic aberrations while minimizing confounding cell behavior caused by foreign nucleic acids.

Bacterial T3SS delivery methods have recently been employed in several contexts with promising results. A T3SS expressing strain of *Pseudomonas aeruginosa* successfully delivered *MYOD* to mouse embryonic fibroblasts (MEFs) to reprogram MEFs to myotubes [28], and facilitated differentiation of human ESCs and iPSCs into cardiomyocytes by sequential addition of five TFs [157]. In addition to the aforementioned advantages, the versatility of bacteria allows for engineering of the system. Recently, optogentic interaction control switches were added to the *Yersinia enterocolitica* T3SS such that protein injection occurred in a light-dependent manner, further increasing the ability to control the delivery of peptides with respect to space and time [192]. The T3SS therefore represents a novel path forward in TF delivery for cellular reprogramming, both experimentally and therapeutically.

2.7 Future Directions

2.7.1 Biology

Incredible progress has been made in the development of cellular reprogramming regimes that give rise to a broad spectrum of cell types across all germ layers. However, overcoming low cell fate conversion efficiencies is a primary hurdle facing the future of cellular reprogramming [13, 122, 140]. While direct reprogramming reduces some of the common side-effects of iPSC reprogramming like mutagenesis and tumorigenesis, full conversion to an intended cell identity is still not guaranteed. It is common for reprogramming perturbations to drive cells to a stable yet hybrid state where some of the starting cell's transcriptional program is retained and the target identity is only partially acquired. Such failures to fully convert have been attributed to the presence of TFs that maintain the starting cell's gene regulatory network, lineage-specific repressors, and inaccessible chromatin among other factors. More rigorous reprogramming regimens involving the silencing of endogenous genes via CRISPR/Cas9 [370], endogenous gene activation using CRISPR/Cas9-based methods [189, 143], and cocktails of chemical compounds with or without TFs [51, 114, 135] have already shown considerable success in combating these roadblocks. Recent achievements in direct reprogramming spanning these and other conversion methods are summarized in Table 2.3. Further improvements in determining effective combinations of source cell types, reprogramming mechanisms, delivery methods, and cultivation conditions will be necessary to achieve consistent large-scale production of reprogrammed cells.

Stable cultivation and expansion of reprogrammed cells remains a challenge [64], limiting sufficient quantities of cells to reconstitute tissues and organs. Once methods of reprogrammed cell proliferation are improved, it is theoretically possible to regenerate organs derived from a patient's somatic cells for autologous transplantation [399]. Since these organs would be recipient-derived, there would be little concern for transplant rejection. Moreover, lifelong immunosuppression, which is currently a significant source of mortality among transplant recipients, would no longer be necessary.

2.7.2 Mathematics

Looking forward, a vast arena of mathematical principles with the potential to further sufficient understanding and facilitation of cellular reprogramming still remains. The high dimensional nature of cellular states, for example, invites an opportunity to examine cellular data in a tensor state space [59]. Tensors are multidimensional arrays generalized from vectors and matrices, and have wide applications in many domains such as social sciences, biology, applied mechanics, machine learning and signal processing [70, 202, 373, 380]. Classical linear control systems, as used in Ron-

Starting Cell Fate	Target Cell Fate	Reprogramming Factors	Delivery Method	Reference
Dermal fibroblasts	Adipocyte-like cells	PPARγ2	Lentivirus	[62]
Dermal fibroblasts	Endothelial progenitor cells	ETV2	mRNA transfection	[363]
Embryonic fibroblasts	Antigen-presenting dendritic cells	PU.1, IRF8, and BATF3	Lentivirus	[289]
Glioblastoma cells	Neuronal cells	Forskolin, ISX9, CHIR99021, I-BET 151, and DAPT	Small molecule cocktail treatment	[182]
Embryonic fibroblasts	Hepatocytes	GATA4, FOXA2, HHEX, HNF4A, HNF6A, MYC, and P53-siRNA	Lentivirus	[387]
Bone marrow-derived cells, Fibroblasts, and Keratinocytes	Neural precursor cells	MSI1, NGN2, and MBD2	Plasmid transfection	[3]
Dermal fibroblasts	Cardiomyocyte-like cells	GATA4, MEF2C, and TBX5	Nanoparticles	[164]
Foreskin fibroblasts	Cardiac progenitor cells	GATA4, HAND2, MEF2C, TBX5, and MEIS1	CRRISPR-cas9 based endogenous gene activation	[371]

Table 2.3: Recent Successes in Direct Reprogramming

quist *et al.*'s work, often fail to fully capture the dynamics of cellular reprogramming because state vectors only represent gene expression, neglecting structural information. Chen *et al.* generalized the classical systems notion of controllability into multilinear systems in which the states, inputs, and outputs are preserved as tensors [59]. Multilinear control systems can significantly relieve the difficulty of describing genome-wide structure and gene expression simultaneously, and will be beneficial in analyzing the dynamics of cellular reprogramming more comprehensively. Further, Chen *et al.* exploited the notion of hypergraphs in modeling the network dynamics of cellular reprogramming [58]. A hypergraph is a generalization of a graph in which its edges can join any number of nodes. The notion of transcription factories supports the existence of multiway interactions involving multiple genomic loci [72], which implies that the human genome configuration can be more accurately captured by hypergraphs. Chen *et al.* developed the notions of entropy and controllability for hypergraphs, which will be potentially advantageous in investigating network dynamics of cellular reprogramming (e.g. detecting cell identities transition points and identifying minimum TF inputs) [58, 60]. Ultimately, cellular reprogramming would be required to account for nonlinearity or nonlinear control in the multiway dynamical system representation and analysis framework, and is an important direction for future research.

2.7.3 Medicine

Reprogrammed cells have many translational applications, from their direct use in replacement therapy to use as experimental models of disease and pharmacologic screens [320]. Many human diseases are characterized by the loss of a functioning tissue, such as the loss of hematopoietic stem cells in aplastic anemia or insulin-secreting beta-islet cells in diabetes mellitus. Reprogramming methods to generate cells lost in such disease states have been described and are rapidly maturing in the laboratory setting, such as those to produce hematopoetic stem cells, neurons, cardiomyocytes and pancreatic beta-islet cells [312]. Once refined, their translation to clinical trials will pave the way for curative approaches to chronic diseases. Early clinical successes, such as the successful autologous graft of RPE cells, described above, indicate that autologous grafts using reprogrammed cells are safe and teeming with potential.

Moreover, reprogrammed human cells have the potential to make superior models of disease compared to animal models. A patient-specific disease model may be created by generating iPSCs from skin fibroblasts and differentiating them into the cell type relevant to the patient's condition. This approach allows the experimentalist to shed light on the mechanisms of pathogenesis, as cells of patients with susceptibility to disease may be closely observed from their developmental infancy to their diseased state. Such iPSC-derived models have already been developed for several conditions, from microcephaly [178] to autism spectrum disorder [213]. While a useful tool in several settings, iPSCs may be limited in their ability to model age-related disease such as neurodegenerative conditions, as epigenetic rejuvenation caused by transit through pluripotent states may reverse aspects of pathogenesis. Direct reprogramming techniques have been employed to circumvent this, as this approach avoids widespread reversal of age-related epigenetic changes [350, 184]. Using these techniques, reprogrammed cell lines may be generated from patients to study the mechanisms and pharmacologic susceptibilities of their disease, enabling better understanding of the connections between patient genotype and pathologic phenotype [67].

2.8 Conclusion

This review sought to examine advancements in cellular reprogramming and computational methods that contribute to them. Many promising approaches in cellular reprogramming are under development, though discovery of new TF recipes to efficiently convert source cells to target cells remains limiting. Mathematically-based approaches such as those outlined here may facilitate their systematic discovery.

Moreover, convergent ideas in the areas of biology, mathematics and medicine point to hybrid (concurrent TF addition and removal), time-dependent, and concentration-dependent reprogram-

ming regimes as viable next steps to improve reprogramming methods. These refinements could offer increased control over cell identity of normal and abnormal cells, and their *in vivo* regenerative potential.

CHAPTER 3

Partial Reprogramming of Colorectal Cancer Cells Towards Treatment Sensitivity

This chapter is motivated by a paper by Markus A. Brown, Gabrielle A. Dotson, Scott Ronquist, Georg Emons, Indika Rajapakse, and Thomas Ried [39].

3.1 Abstract

Canonical Wnt signaling is crucial for intestinal homeostasis as TCF4, the major Wnt signaling effector in the intestines, is required for stem cell maintenance. The capability of TCF4 to maintain the stem cell phenotype is contingent upon β-catenin, a potent transcriptional activator, which interacts with histone acetyltransferases and chromatin remodeling complexes. We used RNAi to explore the influence of TCF4 on chromatin structure (Hi-C) and gene expression (RNA sequencing) across a 72-hour time series in colon cancer. We found that TCF4 reduction results in a disproportionate up-regulation of gene expression, including a powerful induction of *SOX2*. Integration of RNA sequencing and Hi-C data revealed a TAD boundary loss, which occurred concomitantly with the over-expression of a cluster of *CEACAM* genes on chromosome 19. We identified EMT and E2F as the two most deregulated pathways upon TCF4 depletion and *LUM*, *TMPO*, and *AURKA* as highly influential genes in these networks using measures of centrality. Results from gene expression, chromatin structure, and centrality analyses were integrated to generate a list of candidate transcription factors crucial for colon cancer cell homeostasis. The top ranked factor was c-JUN, an oncoprotein known to interact with TCF4 and β-catenin, confirming the usefulness of this approach.

3.2 Introduction

The Wnt signaling transcription factor, TCF4, is crucial for homeostasis of the mammalian intestine [170]. Loss of TCF4, in either embryonic or adult mice, results in ablation of the proliferative compartment of the intestine [362]. The capability of TCF4 to drive a crypt progenitor phenotype and maintain intestinal homeostasis is contingent upon its binding partner. When Wnt signaling is active, cytoplasmic β-catenin migrates to the nucleus where it binds TCF4, resulting in target gene expression. β-catenin is a potent transcriptional activator which influences the surrounding chromatin by recruiting histone acetyltransferases, chromatin remodeling factors, and RNA polymerase associated factors [2, 133, 18, 229]. Therefore, the binding of a βcat/TCF4 complex significantly enhances the recruitment of the cellular machinery necessary for transcription, thereby driving a crypt progenitor phenotype [360].

Conversely, when TCF4 is bound by the TLE family of transcriptional repressors, target gene expression is repressed. Despite the presence of TCF4 throughout the intestine, active Wnt signaling is sequestered to the base of the colonic crypts, thereby limiting βcat/TCF4 complex formation, and the resulting crypt progenitor phenotype, to cells of the crypt base [19, 300, 360].

In colorectal cancer, mutations in the Wnt signaling pathway, primarily in *APC*, result in constitutive Wnt signaling activity [168, 291]. High levels of nuclear β-catenin result in the constitutive expression of Wnt target genes, such as *MYC* and *CCND1* [225, 171, 132, 345]. Given that epigenetic modifications and chromatin remodeling have been reported to occur prior to the expression of Wnt target genes, the high levels of nuclear β-catenin in colorectal cancer likely influence the surrounding chromatin structure [18, 253].

It has become evident that understanding chromatin structure lends insight into the regulation of gene expression [190, 94]. Aspects of chromatin structure include the accessibility of genomic loci to the transcriptional machinery as well as the 3-dimensional configuration of the chromatin, which may facilitate cooperativity between genomic regions sharing a particular 3-dimensional space. Division of the genomeinto euchromatin and heterochromatin domains, referred to as A and B compartments, respectively, reflects the interaction between chromatin structure and function at the chromosomal level [190]. Chromosome folding brings distant sites along the linear genome in close spatial proximity, forming topologically associating domains (TADs), which are insulated regions of the genome sharing epigenetic modifications, gene expression, and replication timing [94, 274, 251]. Given the influence of chromatin structure on gene expression, the dissection of a cellular response to a stimulus requires an understanding of both structural and functional dynamics [266, 284].

Herein, we explored the dynamical influence of silencing *TCF7L2*, the gene encoding TCF4, on chromatin structure and gene expression across a 72-hour time series in the colon cancer cell line,

SW480. Silencing of *TCF7L2* not only allows us to investigate the changes which occur as a result of silencing a major transcription factor, it also represents a clinically relevant model as *TCF7L2* expression correlates with resistance to chemoradiotherapy (CRT), a treatment modality in rectal cancer [119, 162, 80]. For the time series, an inverse RNAi transfection protocol was designed and optimized, which significantly reduced confounding factors typically found in time series data. A 4DN approach to data analysis was used, which integrated structural (Hi-C) and functional (RNA sequencing) data and allowed us to identify influential genes in the most dynamically changing networks 3.1. We then identified candidate reprogramming factors, to be perturbed alongside, or in lieu of, *TCF7L2* to debilitate the colorectal cancer cell.

3.3 Results

3.3.1 Gene Expression is Disproportionately Up-Regulated Across the Time Series

The time series data was generated by sequential addition of a small, interfering RNA (siRNA) targeting the 3' UTR of *TCF7L2* at equally spaced time points (0, 24, 48, and 72 hours) according to a modified siRNA transfection protocol (Figure S1A in Brown *et al.* [39], see Methods). The modified protocol was designed to mitigate varying cell cycle distributions, which occur naturally due to varying growth times and are a major confounding factor in extended (>48 hours) time series data (Figure S1C and D in Brown *et al.* [39]). The modified protocol resulted in 70% of the cells in the G1 phase of the cell cycle at each time point (Figure S1B in Brown *et al.* [39]).

To assess the efficacy of siRNA-mediated *TCF7L2* silencing, the abundance of *TCF7L2* transcripts was determined using quantitative PCR (qPCR). Transcript levels decreased progressively across the time series resulting in an 2-fold decrease by 72 hours (Figure 3.1B). The amount of TCF4, the protein product of *TCF7L2*, was determined using a Western Blot. TCF4 protein levels decreased marginally after 24 hours, 35% after 48 hours and 70% after 72 hours (Figure 3.1C). Principal component analysis (PCA) of the RNA sequencing data demonstrated a directed change in the global gene expression program, with biological replicates clustering together (Figure 3.1D). The number of differentially regulated genes which exhibited both a statistically significant ($p < 0.05$) and a 4-fold change in expression increased dramatically across the time series (Figure 3.1E). However, the balance was positively skewed with nearly twice the number of genes up-regulated as down-regulated. Considering that SW480 cells have constitutive Wnt signaling activity, and therefore persistent transcription through β-cat/TCF4 complexes, the general increase in gene expression following the reduction in TCF4 is unexpected.

To determine the biological relevance of the disproportionate up-regulation of gene expres-

sion, the expression profiles of individual genes with known biological roles were assessed. The expression of prominent Wnt signaling target genes, such as *MYC* and *CCND1*, decreased consistently across the time series, similar to that of *TCF7L2* [132, 376, 342]. It has been previously reported that loss of β-cat/TCF4- mediated transcription results in the up-regulation of genes associated with intestinal differentiation [360]. We confirm this observation as genes associated with differentiation, such as *MUC2* and *LGALS4*, were significantly up-regulated following *TCF7L2* silencing ($p < 0.05$) (Figure S2A in Brown *et al.* [39]).

3.3.2 *SOX2* Up-Regulation is a Consequence of *TCF7L2* Silencing

We identified a powerful ($\log_2$FC $\sim$3.55) up-regulation of *SOX2*, which began after 24 hours and continued to increase until 72 hours post-transfection, which we found was dose-dependent in relation to the levels of *TCF7L2* (Figure 3.1F, Figure S2B in Brown *et al.* [39]). This may represent direct repression of *SOX2* by TCF4, as TCF4 has been previously reported to bind the *SOX2* promoter in colorectal cancer [131]. To determine the influence of nuclear β-catenin on this interaction, we performed a Western Blot (Figure 3.1G). Levels of active β-catenin did not fluctuate over the time series, indicating that the interaction between TCF4 and *SOX2* is independent of β-catenin, in this system, supporting our direct repression hypothesis. Given the potency of the *SOX2* up-regulation, we also investigated whether a concomitant shift in nuclear structure occurred at the *SOX2* locus. *SOX2* resides at a TAD boundary in SW480 cells, however the boundary remained unchanged across the time series, indicating that nuclear structure was not a contributing factor in the up-regulation of *SOX2* (Figure S3A in Brown *et al.* [39]). The observation that *TCF7L2* silencing results in the over-expression of *SOX2* suggests that *TCF7L2* may enforce a transcriptional module which potently represses *SOX2* under physiological conditions.

3.3.3 Genome-wide A/B Compartments are Maintained Over Time

Hi-C was performed to capture genome-wide changes in chromosome structure across the time series to determine the influence of TCF4 on chromatin organization. *TCF7L2* silencing resulted in a significant, genome-wide increase in both the number and strength of pairwise chromatin interactions, as evidenced by matrix subtraction between 0 and 72 hour Hi-C contact maps (Figure 3.2A). The contact maps demonstrate a higher interaction frequency at 72 hours, while the algebraic difference between the two contact maps reveals that the overwhelming direction of change is an increase in chromatin interaction from 0 to 72 hours. Out of all possible 1Mb-scale chromatin interactions, SW480 cells at 0 hours harbored 74% of those interactions, whereas by 72 hours the proportion of interactions had risen to 87%, demonstrating that the chromatin became more compact over time.

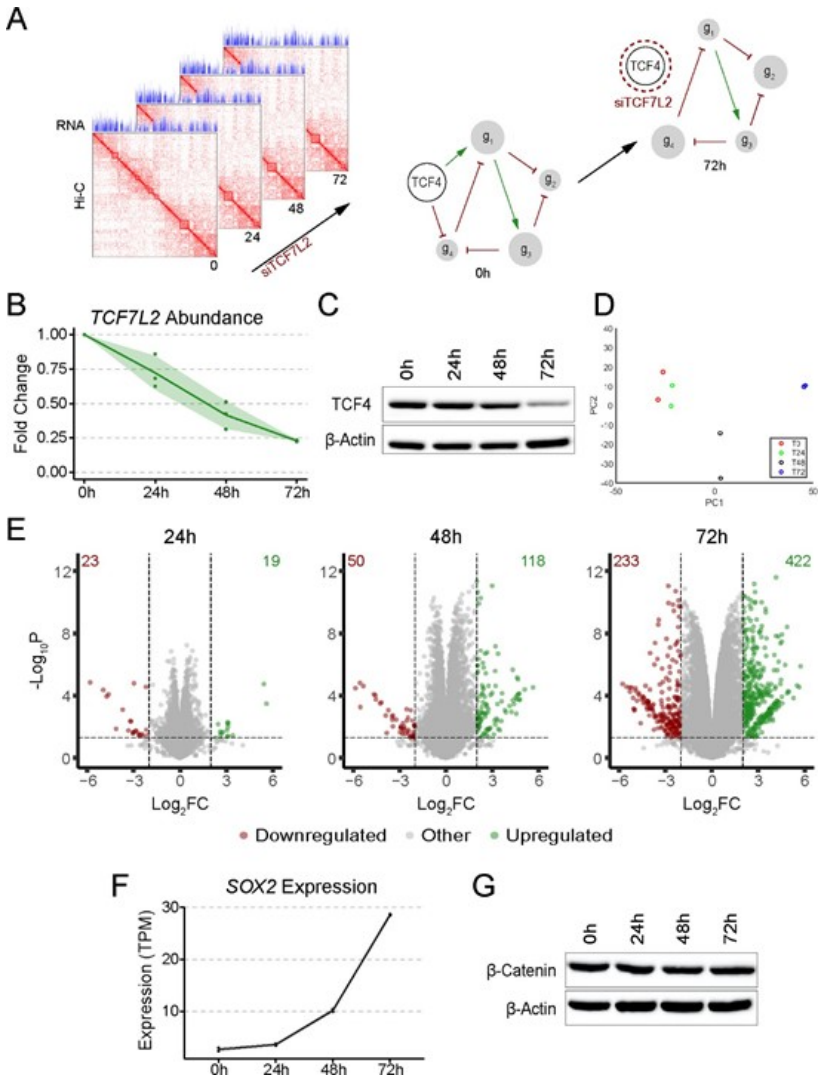

Figure 3.1: Conceptual Approach and Gene Expression Dynamics Following *TCF7L2* Silencing. (A) Schematic summary of the experimental approach. We sought to explore how silencing *TCF7L2* impacted the colorectal cancer gene network in terms of chromatin structure and gene expression. (B) qPCR demonstrated progressive silencing of *TCF7L2* over time with a 75% reduction in *TCF7L2* transcript levels after 72 hours. Each dot represents a biological replicate, with three replicates plotted per time point (n = 3). The green line represents the mean, while the green-shaded ribbon represents the standard deviation.(C) Western blot analysis demonstrated a 70% decrease in TCF4 protein abundance by the last time point with β-actin as loading control. (D) PCA of gene expression (TPM) over the time course, with time points differentiated by color (n = 2). (E) Volcano plots show differential gene expression genome-wide with thresholds set at $\log_2$FC = 2, and p = 0.5 (equivalent to 1.3 on the $-\log_{10}$P scale). Genes in red are significantly down-regulated and undergo a $>$ 4-fold decrease in expression, while those in green are significantly up-regulated and experience a $>$ 4-fold increase in expression. The number of genes which fall within these regions are shown. (F) Expression profile of *SOX2*, which increases dramatically across the time series. The dots correspond to each biological replicate (n = 2), the shaded ribbon represents the standard deviation. (G) Western Blot for active β-catenin over time with β-actin as loading control. No noticeable change in active β-catenin protein levels occurs during the time series.

To further assess changes in genome-wide chromatin organization, the genome was assigned to A/B compartments using a spectral graph theory approach, which calculates the Fiedler number for each 100kb bin (Figure S3B in Brown *et al.* [39])[63]. The sign of the Fiedler number, positive or negative, indicates within which compartment the bin resides. An A/B compartment switch occurred in ∼4% of the genome at any given point during the time series, which encompassed 1,305 genes. K-means clustering identified eight distinct switching patterns in which a specific genomic locus switched compartments unidirectionally (A-to-B or B-to-A) or bidirectionally (a permutation of A-to-B-to-A-to-B) over time (Figure 3.2B). Approximately 52% of the regions demonstrating an A/B compartment switch underwent a single switch, while the remaining 48% underwent multiple switching events. Despite the switching events, the overall proportion of A/B compartments genome-wide remained relatively unchanged with most loci residing in the same chromatin state across the time series (52% A and 48% B)(Table S2 in Brown *et al.* [39]). Furthermore, ∼1.3% of the genes within the A/B switching regions were significantly differentially expressed following the compartment switch.

3.3.4 TAD Boundary Loss in the *CEACAM* Gene Cluster

Local partitioning of chromatin into TADs provided insight into the coupling of structure and function. Genes occupying the same TADs have been shown to be co-expressed, likely by accessing the same cluster of transcriptional machinery as a result of their spatial proximity [94, 74]. The number of TADs genome-wide fluctuated by 5% over the time series, while the number of common TADs shifted by 7% (Table S3 and S4 in Brown *et al.* [39]). To gain insight regarding which factors may be involved in the TAD domain switching, we used publicly available CTCF and TCF4 ChIP-seq data sets from the HCT116 colorectal cancer cell line (see Methods, Table S5 and S6 in Brown *et al.* [39]). We observed that the binding locations of CTCF in HCT116 matched 75% of the TAD boundaries of SW480 while the binding locations of TCF4 matched 41% of the TAD boundaries. However, while TAD boundary binding by CTCF was relatively consistent across chromosomes, the binding of TCF4 fluctuated, with chromosome 19 showing the highest levels of TCF4 binding.

On chromosome 19, a ∼10Mb span of the genome houses a group of *CEACAM* family genes, which are involved in colorectal cancer progression and metastasis [23]. At the initial time point, *CEACAM5*, *CEACAM6*, and *CEACAM7* were located within the same TAD, while *CEACAM1* and *CEACAM8*, were located in an adjacent, smaller TAD (Figure 3.2C). Changes in expression of the two *CEACAM* gene clusters after 24 hours were negligible. However, at the 48 hour time point, the boundary separating the two TADs was lost, joining *CEACAM5*, *CEACAM6*, and *CEACAM7* into a larger TAD with *CEACAM1* and *CEACAM8*. Combination of the TAD domains as well as

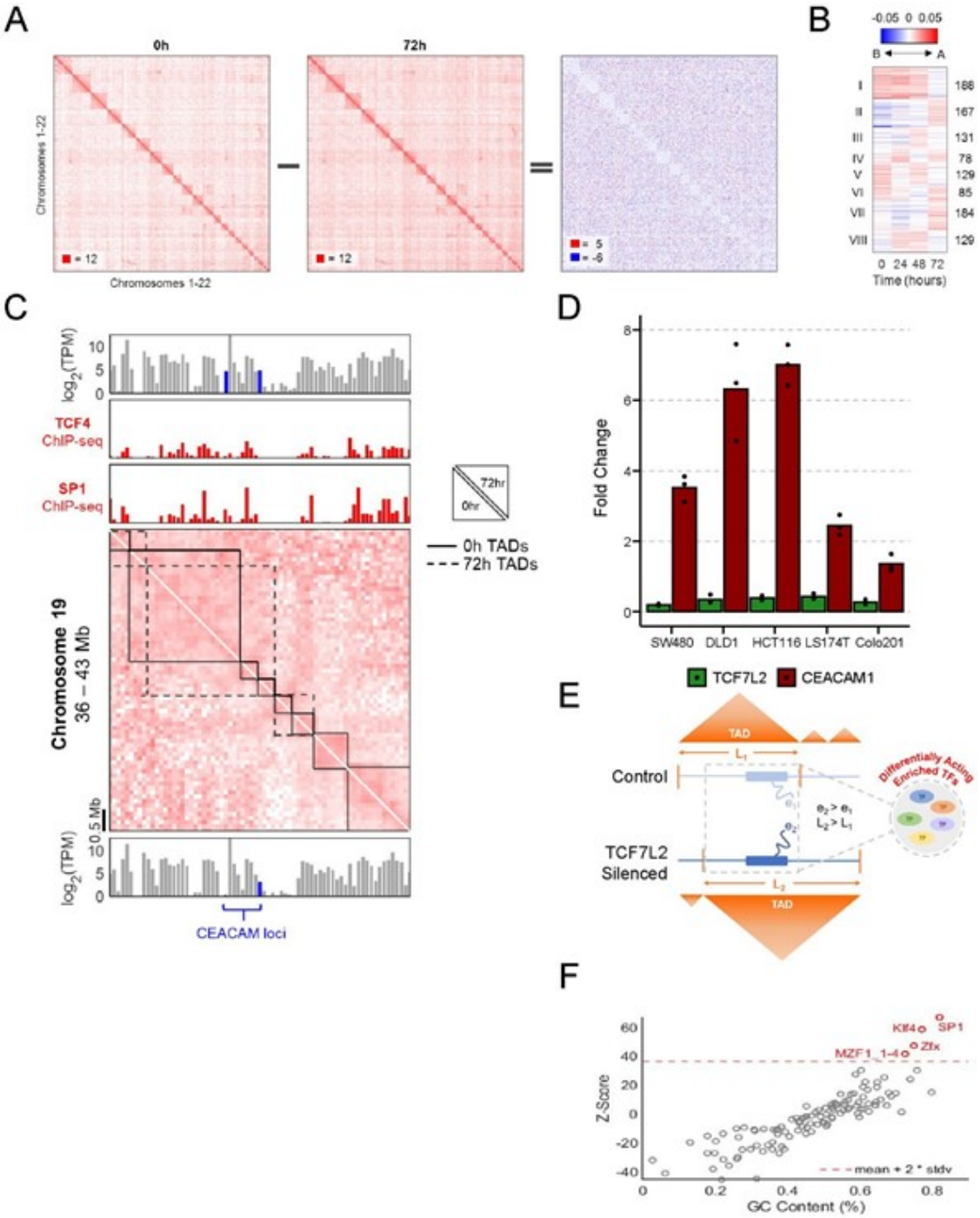

Figure 3.2: Changes in Global and Local Hi-C Partitioning Over Time. (A) Algebraic difference between genome-wide Hi-C contact maps at 0 and 72 hours demonstrates an overall increase in contact frequency over time. (B) Global partitioning (A/B compartmentalization) dynamics revealed that 1,091 100-kb genomic bins switched compartments at one or more time points genome-wide. The Fiedler number (see Methods) of each loci are k-means clustered into eight groups (denoted to the left) exhibiting different switching dynamics. The number of genomic bins in each cluster is specified on the right. A 95% change in Fiedler number from one time point to another was set as a threshold to remove inherent noise.

Figure 3.2: (C) Local partitioning (TAD organization) of the region on Chromosome 19 (34.6 - 45.2 Mb) containing two *CEACAM* gene groups. Hi-C contact maps are shown at 100kb resolution with TAD domains at 0h denoted by solid, black lines and at 72h denoted by dashed, black lines. The CEACAM gene groups are denoted by blue lines in the expression array. ChIP-seq data for TCF4 and SP1 binding was over-layed, demonstrating that TCF4 and SP1 may bind within the TAD boundary region. (D) qPCR was performed on *TCF7L2*-silenced cells at the 72 hour time point to determine the expression of *CEACAM1* in various colon cancer cell lines. Expression of *TCF7L2* and *CEACAM1* was normalized to the negative control (Time 0 - not plotted), which was uniformly set to 1 for each cell line. All cell lines tested demonstrated an up-regulation of *CEACAM1*, however the weakest response was observed in *COLO201*.(E) Diagram illustrating coupled chromosome structure and gene expression between two experimental states as well as a possible explanation for this coupling, i.e., cell state-specific transcription factor activity. (F) Transcription factor enrichment analysis for the 64 genes found in the differentially conformed region of chromosome 19 (see 2C) showed enrichment for SP1, KLF4, ZFX, and MZF1 when compared against a background of 24,752 genes.

a significant increase in expression for both *CEACAM* groups occurred seemingly simultaneously at 48 hours. Gene expression for both groups continued to increase following the TAD boundary loss (Figure 3.2C). The increase in gene expression likely represents increased accessibility to the transcriptional machinery for both gene clusters. Silencing of *TCF7L2* in multiple colon cancer cell lines demonstrates that the up-regulation of *CEACAM1* is present in both APC mutated (SW480, DLD1) as well as APC wild-type (HCT116, LS174T) colon cancer cell lines (Figure 3.2D). The up-regulation of *CEACAM1* is however weaker in COLO201, which is derived from a metastatic site.

3.3.5 TAD Boundaries of the *CEACAM* Gene Cluster Show Enrichment for the SP1, KLF4, ZFX, and MZF1 Transcriptions Factors

To identify which transcription factors may be mediating this TAD partitioning loss, a transcription factor enrichment analysis was performed using oPOSSUM-3 (Figure 3.2E) [175]. Over-representation of transcription factor binding sites (TFBS) in the DNA sequences of the 64 genes occupying the differentially conformed region of chromosome 19 were investigated. The top four identified transcription factors were SP1, KLF4, ZFX, and MZF1 (Figure 3.2F). The top hit, SP1, is found ubiquitously and is involved in chromatin remodeling. KLF4 is involved in embryonic development and binds a GC-rich motif highly similar to SP1. ZFX plays a role in self-renewal of hematopoietic stem cells and MZF1 has been associated in inflammatory bowel disease. To further explore the role of SP1, and TCF4, in mediating this TAD boundary change, we used publicly available SP1 and TCF4 ChIP-seq data sets (see Methods) from the HCT116 colorectal cancer cell line. We found that both SP1 and TCF4 bind the site of the TAD boundary observed in SW480 and

it is therefore feasible that they are responsible for the repartitioning of this domain (Figure 3.2C).

3.3.6 Coordinated Pathway Responses Suggest a Loss in Cell Cycle Progression and DNA Synthesis Capabilities

To assess the dynamics of TCF4 loss on pathway behavior, fast Gene Set Enrichment Analysis (fgsea) was performed using the 50 Hallmark gene sets and the $\log_2$FC expression values from each time point to generate a normalized enrichment score (NES) [328, 308]. The NESs from fgsea were compared and the 20 gene sets (10 up-regulated and 10 down-regulated) with the greatest change by the last time point were plotted (Figure 3.3A). Strong decreases in expression for both MYC Target gene sets as well as the E2F Targets and G2M Checkpoint gene sets were observed. All four of these gene sets are involved in growth and the regulation of cell cycle progression. The E2F pathway is also involved in DNA synthesis, thereby linking DNA synthesis to cell cycle progression [148, 281].

Interestingly, the Notch, Cholesterol homeostasis, Estrogen Response, and MTORC1 signaling pathways all demonstrated a similar NES profile over time, suggesting coordinated function. Several articles identify interplay between these pathways, supporting this hypothesis [56, 34]. The EMT, Apoptosis, and Interferon Gamma pathways demonstrated a coordinated network of up-regulated pathways. The most dynamically up-regulated and down-regulated gene sets upon *TCF7L2* silencing were EMT and E2F, respectively.

3.3.7 Pathway Level Structure-Function Relationships

We then sought to compare structure-function relationships in these pathways, utilizing the Frobenius norm, which describes the space of a vector (gene expression) or matrix (chromatin interaction frequencies) [326] (Figure 3.3B). We find that, generally, the pathways which demonstrate more drastic changes in expression also occupy a more open conformation. For instance, strongly down-regulated E2F Targets, MYC Targets V1, V2, and Unfolded Protein Response pathway genes reside in a more open conformation than do the moderately down-regulated genes in the MTORC1, Notch, G2M Checkpoint, and Estrogen Response Early pathways. This remains largely true for the up-regulated pathways with the Coagulation, Apical Junction, and KRAS Signaling pathway genes residing in a more open conformation than the Allograft Rejection, Interferon Alpha Response, Apoptosis, and Interferon Gamma Response pathway genes.

However, the exceptions are the EMT, Cholesterol Homeostasis, and DNA Repair pathways. The EMT pathway undergoes the most dramatic shift in expression, yet is in the most constricted conformation, while the Cholesterol Homeostasis and DNA Repair pathways have the least change in expression yet still reside in one of the most open conformations.

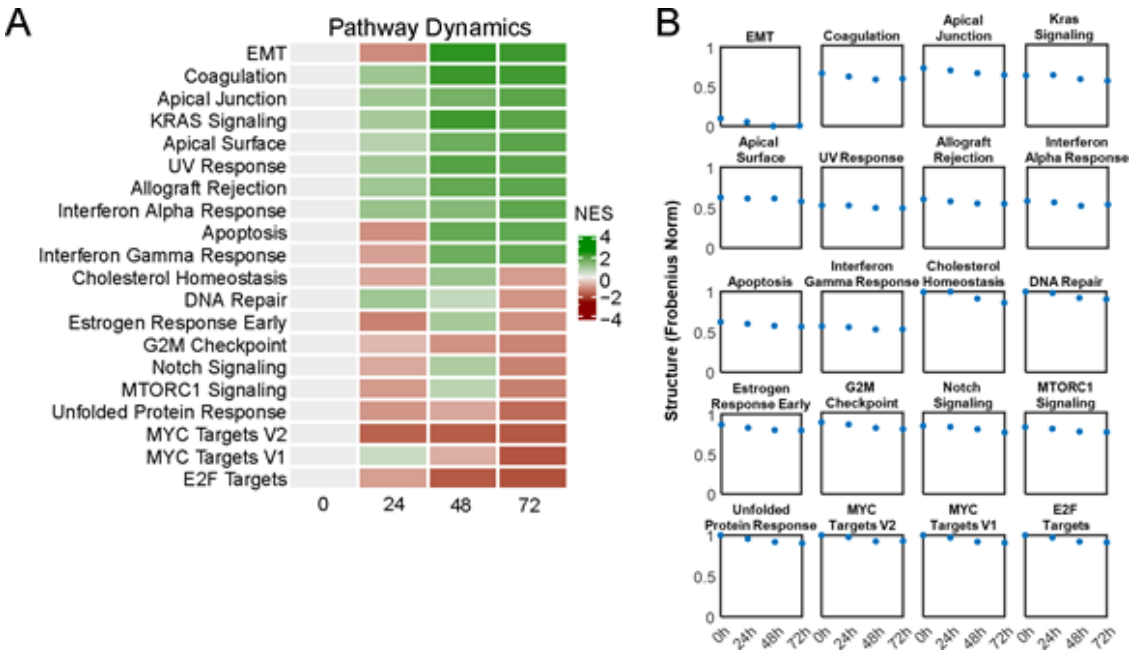

Figure 3.3: Pathway Level Gene Expression and Structural Dynamics. (A) Fast Gene Set Enrichment Analysis (fgsea) was performed using the Hallmark Gene Sets and $\log_2$FC for each time point to generate a normalized enrichment score (NES). The NESs of the ten most up- and down-regulated pathways were then plotted as a heatmap to show pathway level gene expression over time. The NES at the initial time point is set to 0. (B) The montage of plots reflect the change in Frobenius norm of pathway-specific contact frequency data for each time point and pathway of interest. Norm values are min-max normalized across all pathways represented. Each pathway is represented as four points, with each point corresponding to a time point, with higher values indicating a more open conformation. The E2F Targets, MYC Targets V1 and V2, Coagulation, and KRAS Signaling pathways all demonstrate high degrees of openness.

3.3.8 EMT and E2F Signaling Genes are Highly Connected in SW480 Cells

The two most dynamic gene sets, EMT and E2F, were then probed to gauge the connection between these networks in three-dimensional space. A Hi-C-derived, synthetic 5C contact map was generated to highlight the inter-loci interaction frequencies among EMT and E2F signaling genes. The contact map is comprised of 1Mb bins containing genes in either of the pathways (see Methods). Compared to a synthetic 5C contact map composed of randomly sampled, gene-containing genomic bins, we observe that the EMT and E2F signaling genes contact map is denser, suggesting that these genes are more closely interacting than other genes in the genome (Figure 3.4A). Permutation testing reveals that these genes are indeed more inter-connected than genes outside either network ($p < 0.001$), with a Fiedler number that surpasses those of 1,000 sets of randomly sampled genes (Figure S4A-B in Brown *et al.* [39]).

To understand the evolving behavior of individual genes in the EMT, E2F, and Wnt signaling

networks, a gene-level, structure-function analysis was performed. Each gene was examined at 5kb resolution and was assigned a 35kb-length window – 5kb overlapping the transcription start site and 15kb flanking either side of that 5kb region – creating a seven by seven sub-matrix for each gene. Several network- based approaches have previously been used to characterize behaviors in dynamically changing genomes [305, 198]. We apply one such approach - a derivative of Von Neumann Entropy (VNE) - to measure local chromatin organization of individual gene regions [65]. Higher VNE values indicate that the number of conformations available to the gene and its immediate neighborhood are higher, indicating that chromatin is more accessible. We computed VNE on the EMT, E2F, and Wnt signaling gene sub-matrices and plotted them as a function of gene expression, creating a phase portrait (Figure 3.4B).

The phase portraits show the gene-level trajectory of chromatin accessibility and expression for the ten most dynamic genes, as determined by the area of their ellipse. Overlap between the gene ellipses demonstrates a homogeneous concerted pathway response in the EMT and E2F gene sets, indicating that the genes, in their respective pathways, are in a similar chromatin environment and demonstrate coordinated changes in expression. The WNT gene network undergoes a greater diversity in its response, indicative of varying chromatin environments and diverse transcriptional regulation at the individual gene loci. Genes with different expression patterns, yet whose phase portraits still overlap, are indicative of regulation at the transcriptional level (such as transcription factor binding), rather than at the chromatin level. For example, *CCNY* and *LEF1* are present in similar chromatin environments, as indicated by their overlapping phase portraits, yet show opposite functional patterns, indicating that their regulation occurs at the transcriptional level. Overall, we observe that the fitted ellipses change dynamically in terms of structure and function, confirming that *TCF7L2* silencing influences chromatin structure.

3.3.9 Network Interactions are Preserved as Genes Move Spatially Closer Over Time

To further explore chromatin topology, we performed network centrality analysis on inter-loci interaction data from our synthetic 5C contact maps. Network centrality describes the influence or importance of nodes in a network, which yields insight into how the network relays information [33, 240, 198]. We calculated four measures of centrality: betweenness, closeness, degree, and eigenvector. Betweenness centrality identifies nodes which lie along the shortest path between other nodes, closeness centrality is determined by the average farness to all other nodes, degree centrality captures the number of edges connected to a node, and eigenvector centrality determines the influence of a node based on the influence of the neighboring nodes.

Roughly 57% of the EMT and E2F network genes underwent a significant change in between-

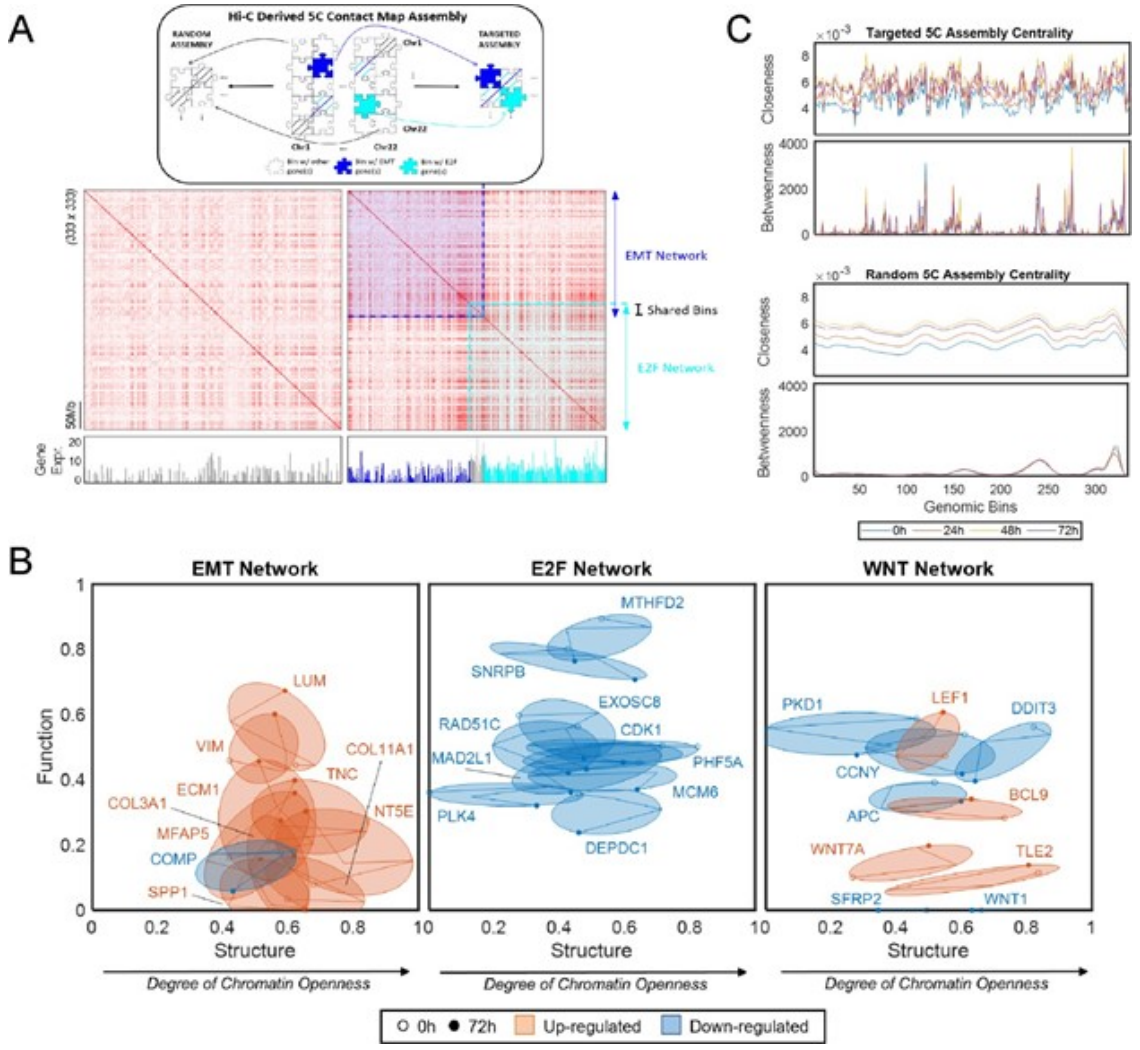

Figure 3.4: Centrality Analyses and Gene Level Structure-Function Relationships. (A) Gene network-level synthetic 5C contact maps were generated by extracting bins corresponding to genes in the EMT and E2F signaling networks from a 1Mb adjacency matrix and stitched together (top) to create a synthetic contact map (right). Synthetic 5C contact map for randomly sampled gene bins (left) shown for comparison. (B) Structure versus function phase portrait for the EMT, E2F, and WNT signaling pathway genes. Min-Max normalized Von Neumann Entropy (see Methods) and gene expression represent structure and function, respectively. The ten most dynamic genes (those whose fitted ellipse have the largest areas) in each network are shown. Genes that decrease in expression over time are labelled in blue and genes that increase in expression are labelled in orange. Phase portraits reveal the extent to which chromatin structure and gene expression are coupled. (C) Closeness centrality increases over time, indicating that the network becomes increasingly compact. An increase in betweenness centrality is also observed over time which reaches a maximum at 48 hours. Peaks in betweenness centrality plot identify bins (genes) which significantly regulate the network. The Random 5C Assembly demonstrates a progressive increase in closeness centrality, reflective of the general increased contact frequencies over time.

ness centrality, the majority of which experienced a decrease in their betweenness centrality score (Figure 3.4C). *LUM*, *TMPO*, and *AURKA* had the highest betweenness centrality scores at any given time point, suggesting that they are highly influential in their respective networks. Since silencing of *TCF7L2* resulted in decreased betweenness centrality scores for these genes, they represent the most likely nodes by which TCF4 may influence the EMT and E2F networks.

Closeness centrality scores provided insight into the evolving proximity of the network genes. We observe that closeness centrality increased with time for all loci with either of the latter two time points (48 and 72 hours) having the highest scores (Figure 3.4C). This accompanies the previous observation that pairwise chromatin interactions increase over time, i.e., chromatin organization is becoming denser. Since a higher closeness centrality score indicates that a node is relatively closer to all other nodes in the network, this trend suggests that the genes become more organized with time. We observe that the eigenvector and degree centralities of our network genes fluctuate very little with time - meaning the same gene interactions with a similar magnitude of connectivity occur at each time point (Figure S4C in Brown *et al.* [39]).

3.3.10 4DN Analysis Provides Insight Into TF-Driven Controllability of CRC Cells

Our findings demonstrate that silencing *TCF7L2* is sufficient to mediate structural and functional changes that influence the behavior of colorectal cancer cells. We suspect that there are a number of reprogramming factors that, when silenced alongside *TCF7L2*, could further destabilize the colorectal cancer cell. Genes that have significant nodal influence represent ideal candidate reprogramming factors, as a result of their ability to destabilize and promote wide-spread changes in the network. Betweenness centrality is one measure of nodal influence and is particularly characteristic of cross-talk control. We find that in the colorectal cancer network, 187 1Mb bins containing 215 genes from the EMT and E2F signaling gene networks harbored a significant change in betweenness centrality from 0 to 72 hours. These genes were ranked according to their magnitude of change in betweenness centrality, with higher magnitudes corresponding to a higher ranking. The genes were then ranked separately by their percent change in gene expression and Frobenius norm between 0 and 72 hours to incorporate aspects of gene function and structure. The average of the three rankings for each gene was computed and the genes ordered accordingly, with lower rankings indicating a gene with the most controllability potential. Considering that transcription factors regulate a network of genes, we summarize the rankings and dynamics for the most likely transcription factor-encoding genes from our candidate gene list in Table 3.1. The highest ranked candidate was c-JUN, an oncogenic subunit of the AP-1 transcription factor, whose activity is augmented in many cancer types, and has been shown to interact with TCF4 and β-catenin to form a

ternary, transcriptionally-competent complex [236].

Gene	Network	Gene Expression (Fold Change)	Structure (Frobenius Norm)	Betweenness Centrality (Fold Change)	Biological Function
JUN	EMT	0.56	87.11	4.29	Stabilizes βcat/TCF4 complexes, decreases MYC degradation
SNAI2	EMT	-1.35	81.57	2.81	Promotes invasion and metastasis in EMT
TP53	E2F	-0.74	66.56	2.45	Crucial tumor suppressor involved in regulating DNA repair/apoptosis
MXD3	E2F	-0.10	78.36	-4.21	Suppresses MYC dependent cell transformation
MSX1	EMT	0.12	85.91	-3.35	Transcriptional repressor in limb-pattern formation
MYBL2	E2F	-0.68	149.56	-2.27	Regulates cell survival, proliferation, and differentiation
PRRX1	EMT	-1.36	59.32	3.00	Regulator of muscle creatine kinase
E2F8	E2F	-0.21	151.82	-1.29	Regulates progression from G_1 to S phase
DNMT1	E2F	-0.12	79.52	1.02	May silence tumor suppressor genes by methylation

Table 3.1: Ranked Candidate Reprogramming Factors. A total of 215 genes from the EMT and E2F signaling networks were ranked according to the magnitude of their change in gene expression, frobenius norm (structure), and betweenness centrality, between the 0 and 72 hour time points, irrespective of direction. The list of candidate factors was then filtered by removing genes which are not well characterized as transcription factors as well as those whose motif binding information was unavailable.

3.4 Discussion

3.4.1 TCF4 as a Transcriptional Repressor

Colorectal cancers exhibit constitutively active Wnt signaling, most commonly due to mutations in APC [168]. The over-expression of Wnt target genes is driven by β-cat/TCF4 complexes, as TCF4 is the major Wnt signaling transcription factor in colorectal cancer [171]. Due to the central role of Wnt signaling in colorectal tumorigenesis as well as the capability of β-catenin to modulate the chromatin environment, we explored the dynamical influence of TCF4 on chromatin structure and gene expression in the SW480 colorectal cancer cell line.

TCF7L2 silencing resulted in a progressive deregulation of the transcriptome with the most dynamic changes occurring at the last time point. Surprisingly, nearly twice the number of genes were up-regulated upon *TCF7L2* silencing. While difficult to distinguish between direct and indirect effects, these results hint at a repressive role for TCF4. In support of this, naturally occurring dominant-negative variants of TCF4, have been described [358, 44]. The gene *SOX2* demonstrated a powerful increase in expression, which occurred progressively over the time series. We reason that this induction is a result of decreased repression by dominant negative TCF4 isoforms, as TCF4 has been shown to bind the *SOX2* promoter [131]. Additionally, this interaction occurs independently of fluctuating β-catenin levels, in line with repression by a dominant-negative isoform. Furthermore, interactions between canonical Wnt and *SOX2* have been described in various cell types. In lung epithelia, an antagonizing role between canonical Wnt signaling and *SOX2* has been described in which activation of Wnt signaling led to the loss of *SOX2* and interfered with the devel-

opment of the bronchiolar epithelium [130]. In gastric tumorigenesis, *SOX2* functioned as a tumor suppressor which antagonized Wnt-driven adenomas [299]. These two examples demonstrate a mutual, antagonizing relationship between *SOX2* and canonical Wnt signaling, which influences the development of normal tissue (lung) as well as adenomas (stomach). Here, we observe an antagonizing role between TCF4 and *SOX2* in colorectal cancer. Taken together, these results suggest that TCF4 maintains a transcriptional module for repressing *SOX2*, which may be necessary for generating and/or maintaining the intestinal lineage.

3.4.2 TCF4 Influences Local Genome Organization

Time series Hi-C experiments demonstrated that TCF4 influences chromatin organization both globally and locally. Silencing of *TCF7L2* led to an increase in pairwise chromatin interactions, demonstrating an overall compaction of the chromatin across the genome. This may be due to a decrease in β-cat/TCF4 binding, which would lead to decreased β-catenin-mediated chromatin acetylation. The high levels of nuclear β-catenin present in colorectal cancer may be sufficient for mediating a genome-scale chromatin effect. However, the compaction was not sufficient to cause a decrease in gene expression.

Approximately 4% of the genome underwent an A/B compartment switch at any given time point. The silencing of a single transcription factor leading to compartment switching in 4% of the genome is comparable to over-expressing four transcription factors which influence $\sim$20% of the genome [322]. Despite changes in A/B compartment switching, the total number of bins which resided in the A/B compartments remained stable over time (52% A, 48% B). This supports the existence of a genome-wide 'balance' of open and closed chromatin domains. However, the A/B compartment balance in normal cells of the mouse is different, at 40% A and 60% B [322]. This may be a reflection of the difference in species or could be an indication that, since cancer cells have enlarged nuclei, they may be able to support a larger percentage of the genome in the A compartment, and therefore support a more diverse transcriptome. However, an increased nuclear volume could merely reflect the increased DNA content of cancer cells. Another explanation for the relatively stable A/B compartment balance is a result of noise within the data. Since a single time series was performed for Hi-C, we are unable to exclude this possibility.

TAD analysis demonstrated that the number of TAD domains fluctuated by $\sim$5% over time. Incorporating CTCF ChIP-seq data from HCT116 demonstrated that the majority (80%) of the TAD boundaries are likely bound by CTCF, whereas 41% are likely to be bound by TCF4. On chromosome 19, a TAD domain change was coordinated with an increase in gene expression in the *CEACAM* family genes. The most dramatic change in the expression of these genes occurred concomitantly as the TAD boundary loss, hinting at a causal link between the two. Following the

TAD domain coalescence, the expression of both *CEACAM* gene groups continued to increase, likely by accessing the same transcriptional machinery, as they are part of the same gene family. Incorporation of ChIP-seq data for TCF4 and SP1, the transcription factor with the highest degree of enrichment in the switching region, suggested that TCF4 and SP1 could feasibly be responsible for the TAD domain change as peaks for both transcription factors were present near the initial TAD boundary. Given the current understanding of structure-function relationships, we interpret the dynamics of the event thus: *TCF7L2* silencing results in the decreased activity of a factor, such as TCF4 itself, which then influences the activity of SP1, which results in loss of the TAD boundary. Boundary loss increases the probability that the two *CEACAM* groups will come in close contact, thereby allowing them to coordinate recruitment of the transcriptional machinery.

3.4.3 Structure-Function Relationships in SW480

From fgsea, we observed down-regulation of the MYC, E2F, and the G2/M gene sets, all which mediate growth and cell cycle progression. In addition to regulating growth, genes in the E2F pathway are also responsible for DNA synthesis [148, 281]. Additionally, despite showing the greatest decrease in pathway gene expression, E2F genes are still expressed to a higher degree than genes in EMT, the most up-regulated gene set (Figure 3.4A - Gene Expression). Previous evidence has demonstrated that TCF4 expression corresponds with resistance to CRT and that loss of TCF4 increases sensitivity [119, 162, 37]. Furthermore, it has been suggested that the mutation rate in intestinal stem cells is relatively equal to the mutation rate in liver stem cells, despite the higher proliferative rate of the former [31, 118]. We hypothesize that high levels of DNA synthesis/DNA proofreading activity, mediated by E2F, could maintain DNA sequence integrity in the face of continued stem cell proliferation and, in colorectal cancers, this same DNA program would defend against CRT. Loss of TCF4 results in cell cycle arrest, ensuring that cells with hampered DNA repair capabilities do not propagate in the stem cell crypt.

Investigation of structure and function relationships in the top 20 most dynamic pathways demonstrated that pathways whose genes resided in a more open conformation also demonstrated more dramatic changes in expression. Pathways which demonstrated less dramatic changes in expression also resided in chromatin environments which were more constricted. This trend was observed in 85% of the most dynamically acting pathways. According to these results, we find that the conformation of the chromatin is not a determinant of gene expression, but rather, functions as a filter. Genes which reside in open conformations are easily accessed by the transcriptional machinery and thus experience a full response, while genes in a more constricted environment have less exposure and therefore undergo less dynamic changes in expression. The general decrease in chromatin accessibility across the time points is a result of the genome-wide chromatin com-

paction which occurred as a result of *TCF7L2* silencing. In the remaining 15% of the pathways, specifically EMT, Cholesterol Homeostasis, and DNA Repair, the connection between expression and structure was not pronounced. Despite showing the most dynamic change in expression, the EMT pathway genes resided in the most constricted conformation out of the top 20 pathways. Conversely, the Cholesterol and DNA Repair pathway genes resided in the most accessible conformation yet underwent less dynamic changes in expression than other pathways. Therefore, in some instances, chromatin conformation is not a determinant of gene expression.

3.4.4 Candidate Reprogramming Factors for Colorectal Cancer

Centrality analysis allowed us to determine the most influential genes in their respective networks. The genes with the highest betweenness centrality were LUM, TMPO, and AURKA, identifying these genes as the most useful targets for propagating a stimulus through their respective pathways (EMT or E2F). We subsequently identified a list of factors which can be used in concert with loss of TCF4 to profoundly perturb the colorectal cancer cell. The top ranked candidate reprogramming factor was c-JUN, an oncoprotein which forms part of the AP-1 transcription factor. Previous work has shown that c-JUN forms a complex with TCF4 and β-catenin, which stabilizes β-catenin, and drives expression of the JUN promoter, initiating a self-stimulating feedback loop [212, 113]. Abrogation of the c-JUN, TCF4, β-catenin complex led to reduced tumor number and size, as well as prolonged lifespan [236]. High levels of c-JUN and Wnt signaling activity resulted in increased resistance to cisplatin whereas abrogation of c-JUN or Wnt signaling activity increased the sensitivity of ovarian cancer cells to treatment.

3.5 Conclusion

In conclusion, we find that *TCF7L2* influences both structure and function in the SW480 colon cancer cell line. Loss of TCF4 results in a potent up-regulation of *SOX2*, which in a saturated WNT environment, is not dependent upon β-catenin, hinting at a repressive role for TCF4. When TCF4 is present, chromatin resides in a more open conformation genome-wide, which is likely mediated by βcat/TCF4 complexes. At the local chromatin level, the expression of a group of *CEACAM* genes situated at a TAD boundary was greatly increased upon TAD boundary loss, a change which is likely mediated by TCF4 or SP1. The expression of pathway specific genes corresponded to the accessibility of those genes, suggesting that chromatin conformation acts as a filter to attenuate the potency of gene expression regulation. The two most dynamic pathways following *TCF7L2* silencing – EMT and E2F – interact in 3-dimensional space and the most dynamic pathway genes inhabit overlapping structure-function space. Incorporation of chromatin structure, gene expres-

sion, and centrality data allowed us to generate a list of transcription factors which would be most effective at perturbing the colorectal cancer cell alongside TCF4. Our top ranked transcription factor, c-JUN, is known to bind TCF4 and β-catenin, thereby confirming the validity of network connectivity analysis in identifying relevant biological interactions.

3.6 Materials and Methods

3.6.1 Cell Culture

The SW480, DLD1, COLO201, HCT116, and LS174T colon cancer cell lines were obtained from the American Type Culture Collection. SW480, DLD1, and COLO201 cells were grown in RPMI-1640 growth medium (Thermo Fisher; 11875093), HCT116 cells were grown in McCoy's 5A (Modified) medium (Thermo Fisher; 16600082), and LS174T cells were grown in Minimum Essential medium (Thermo Fisher; 11095080). All growth media were supplemented to 10% FBS (Thermo Fisher; 10082147) and 2mM L-glutamine (Thermo Fisher; 25030149). Cells were incubated at 37°C in 5% CO2. No antibiotics were used. Proper cell line identity was confirmed using STR profiling and the absence of mycoplasma contamination was confirmed routinely using PCR (Genlantis; MY01050).

3.6.2 siRNA Inverted Transfections

Cells were seeded in a 6-well tissue culture plate (Corning; 353046) at a density of 150,000 cells per well. The wells contained 2mL of RPMI-1640 growth medium. Silencer Select siRNAs (Thermo Fisher), Negative Control No.1 (4390843) and si*TCF7L2*-s13880 (4392420), were delivered to the cells using the lipofectamine RNAiMAX reagent (Thermo Fisher; 13778150) according to the manufacturer's instructions. This resulted in the addition of 7.5µL of RNAiMAX and 25pmol of siRNA per well of the 6-well plate. The siRNA:lipofectamine complexes were not removed during the time course. The time series was constructed thus: 0 and 72 hour time points were transfected with siNEG and si*TCF7L2*, respectively, 24 hours after seeding. The 48 hour time point was transfected, with si*TCF7L2*, 24 hours after the 0 and 72 hour time points. The 24 hour time point was transfected, with si*TCF7L2*, 24 hours after the 48 hour time point. The cells were harvested 24 hours later, resulting in exposure to si*TCF7L2* for 72, 48, and 24 hours. The 0 hour time point spent 0 hours in si*TCF7L2* and serves as the transfection control. This procedure is illustrated in Figure S1 in Brown *et al.*[39].

3.6.3 Quantitative PCR

RNA was extracted from at least three independent inverted transfections using the RNeasy Mini Kit and on-column DNase treatment (Qiagen; 74104, 79254), according to the manufacturer's instructions. RNA concentration was determined using a NanoDrop 1000 spectrophotometer (Thermo Fisher). Synthesis of cDNA was performed using the Verso cDNA Synthesis kit (Thermo Fisher; AB1453B) based on the manufacturer's instructions. The optional RT Enhancer was used and equal amounts of both the Anchored oligo dT and Random Hexamers were added. Quantitative PCR was performed for *TCF7L2*, *SOX2*, *CEACAM1*, and *YWHAZ* on an ABI PRISM 7000 sequence detection system (Applied Biosystems) using *Power* SYBR Green PCR Master Mix (Thermo Fisher; 4367659), according to the manufacturer's instructions. Primer sequences can be found in Table S7 [39]. The thermocycling protocol consisted of a 10 minute hold at 95°C followed by 40 two-step cycles consisting of 15 seconds at 95°C and 1 minute at 60°C. Fold change was determined using the *YWHAZ* reference gene and the $2\hat{(}$-$\Delta\Delta$CT) method.

3.6.4 Western Blot

Protein was extracted using a 1% NP-40 Lysis Buffer solution containing a protease inhibitor cocktail (Millipore; 11836153001). To extract protein, 350µL of the 1% NP-40 solution was added to each well of a 6-well plate and cells were scraped into the buffer using a cell scraper and transferred to a 1.5mL eppendorf tube. The mixture was incubated on ice for 30 minutes and spun at 14,000 rpm for 30 minutes at 4°C. The supernatant was transferred to a new tube. Protein was prepared for quantification using the Pierce BCA Protein Assay Kit (Thermo Fisher; 23225), according to the manufacturer's instructions, and quantified on a SpectraMAX M2e microplate reader (Molecular Devices). Nuclear protein samples (20µg) were electrophoresed using a 4-12% Bis-Tris gel (Thermo Fisher; NP0322BOX) in a XCell SureLock Mini-Cell (Thermo Fisher; EI0002) with MOPS Running Buffer (Thermo Fisher; NP0001). Proteins were transferred into a PVDF membrane (Millipore; IPVH00010) using Transfer Buffer (Thermo Fisher; NP00061) and an XCell II Blot Module. Membranes were incubated with antibodies targeting TCF4 (Origene; TA333645), active β-catenin (Millipore; 05-665), or β-actin (CST; 4970S). An HRP-secondary antibody (CST; 7074S, 7076S) was applied and detected using Pierce ECL Western Blotting Substrate (Thermo Scientific; 32209). Blots were imaged in an Azure c600 Gel Imaging Station (Azure Biosystems).

3.6.5 RNA Sequencing and Data Processing

RNA was extracted from three independent inverted transfections using the RNeasy Mini Kit and on-column DNAse treatment (Qiagen; 74104, 79254), according to the manufacturer's instruc-

tions. RNA concentration was determined using a NanoDrop 1000 spectrophotometer (Thermo Fisher). RNA integrity was determined using the 4200 TapeStation (Agilent). RNA samples with an RNA integrity number (RIN) of 9 or higher were used for library preparation. Libraries were prepared using the TruSeq Stranded Total RNA Library Prep Gold kit (Illumina; 20020598). Ribosomal RNA (rRNA) was removed using biotinylated, target-specific oligos combined with Ribo-Zero rRNA removal beads. The RNA was then fragmented and the cleaved RNA fragments were copied into first strand cDNA using reverse transcriptase and random primers, followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. The resulting double-strand cDNA was used as input for Illumina library preparation with end-repair, adapter ligation, and high-fidelity PCR. The final purified product was quantified by qPCR. Samples were sequenced on a HiSeq2500 using Illumina TruSeq v4 chemistry generating 125 base pair, paired-end reads. RNA sequencing data was processed using the CCBR Pipeliner Version 3.0. Briefly, FASTQC was used to assess sequencing quality and Cutadapt was used to remove adapter sequences and perform quality trimming, respectively [9, 215]. Kraken, KronaTools, and FastQ Screen were used to assess microbial contamination [384, 245]. Our samples were confirmed to be free of microbial contamination. STAR (two-pass) was used to align reads to the hg19 reference genome [95]. Picard, Preseq, SAMtools, and RSeQC were used to assess alignment quality [82, 188, 372]. Gene expression was quantified using RSEM [187]. Transcripts per million (TPM) were computed to normalize gene expression values. Differential gene expression was per- formed using limma [180, 316]. Volcano plots were generated using the EnhancedVolcano package [30].

3.6.6 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis was performed using R and the fgsea package, available through Bioconductor [328, 308]. Differentially expressed genes were pre-ranked according to their $\log_2$FC values. Gene set enrichment was calculated using the fgsea command and the Hallmark gene sets available from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp). The minSize and maxSize parameters were set to 15 and 500, respectively. The heatmap was generated using the ComplexHeatmap package [124].

3.6.7 Hi-C Sequencing

The *in situ* Hi-C protocols from Rao et al. [19] were adapted with slight modifications. For each Hi-C library, approximately 3 x 10^6 cells were incubated in 250µl of ice-cold Hi-C lysis buffer (10mM Tris–HCl pH 8.0, 10mM NaCl, 0.2% Igepal CA630) with 50µl of a protease inhibitor cocktail on ice for 30 minutes and washed with 250µl lysis buffer. The nuclei were pelleted by centrifugation at 2,500g for 5 minutes at 4°C, re-suspended in 50µl of 0.5% sodium dodecyl sulfate

(SDS) and incubated at 62°C for 10 minutes. Afterwards, 145µl of water and 25µl of 10% Triton X-100 were added and incubated at 37°C for 15 minutes.

Chromatin was digested with 200 units of MboI (NEB) overnight at 37°C with rotation. Chromatin end overhangs were filled in and marked with biotin-14-dATP (Thermo Fisher Scientific) by adding the following components to the reaction: 37.5µl of 0.4mM biotin-14-dATP (Life Technologies), 1.5µl of 10mM dCTP, 1.5µl of 10mM dGTP, 1.5µl of 10mM dTTP, and 8µl of 5U/µl DNA Polymerase I, Large (Klenow) Fragment (NEB). The marked chromatin ends were ligated by adding 900µl of ligation master mix consisting of 663µl of water, 120µl of 10X NEB T4 DNA ligase buffer (NEB), 100µl of 10% Triton X-100, 12µl of 10mg/ml BSA, 5µl of 400U/µl T4 DNA Ligase (NEB), and incubated at room temperature for 4 hours.

Chromatin reverse crosslinking was performed by adding 50µl of 20mg/ml proteinase K (NEB) and 120µl of 10% SDS and incubated at 55°C for 30 minutes, adding 130µl of 5M sodium chloride and incubated at 68°C overnight. DNA was precipitated with ethanol, washed with 70% ethanol, and dissolved in 105µl of 10 mM Tris–HCl, pH 8. DNA was sheared on a Covaris S2 sonicator. Biotinylated DNA fragments were pulled with the MyOne Streptavidin C1 beads (Life Technologies). To repair the ends of sheared DNA and remove biotin from unligated ends, DNA-bound beads were resuspended in 100µl of mix containing 82µl of 1X NEB T4 DNA ligase buffer with 10mM ATP (NEB), 10µl of 10 (2.5mM each) 25mM dNTP mix, 5µl of 10U/µl NEB T4 PNK (NEB), 4µl of 3U/µl NEB T4 DNA polymerase (NEB), and 1µl of 5U/µl NEB DNA polymerase I, Large (Klenow) Fragment (NEB).

After end-repair, dATP attachment was carried out in 100µl of mix consisting of 90µl of 1X NEBuffer 2, 5µl of 10mM dATP, 5µl of 5U/µl NEB Klenow exo minus (NEB), and incubated at 37°C for 30 minutes. The beads were then cleaned for Illumina sequencing adaptor ligation which was done in a mix containing 50µl of 1X T4 ligase buffer, 3µl T4 DNA ligases (NEB), and 2µl of a 15µM Illumina indexed adapter at room temperature for 1 hour. DNA was dissociated from the bead by heating at 98°C for 10 minutes, separated on a magnet, and transferred to a clean tube.

Final amplification of the library was carried out in multiple PCR reactions using Illumina PCR primers. The reac-tions were performed on a 25µl scale consisting of 25ng of DNA, 2µl of 2.5mM dNTPs, 0.35µl of 10µM each primer, 2.5µl of 10X PfuUltra buffer, and 0.5µl of PfuUltra II Fusion DNA polymerase (Agilent). The PCR cycle conditions were set to 98°C for 30 seconds as the denaturing step, followed by 16 cycles of 98°C 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds, then with an extension step at 72°C for 7 minutes.

After PCR amplification, the products from the same library were pooled and fragments ranging in sized of 300–500 bp were selected with AMPure XP beads (Beckman Coulter). The sized selected libraries were sequenced to produce paired-end reads on the Illumina HiSeq 2500 platform with the V4 for 125 cycles. Quality control metrics of the Hi-C data is given in Table S1 of Brown

et al.[39].

3.6.8 Hi-C Matrix Generation

Paired end reads were processed using the juicer pipeline with default parameters [103]. Reads were mapped to reference genome hg19, with "-s" (site parameter) MboI. Reads with MAPQ >30 were kept for further analysis. Data was extracted and input to MATLAB using the Juicebox tools command "dump". Knight-Ruiz (KR) normalization was applied to all matrices, observed over expected (O/E) matrices were used for A/B compartmentalization and identification of topologically associating domains (TADs). Rows and columns for which more than 10% of entries had zeros were removed from the matrix.

3.6.9 A/B Compartmentalization

Hi-C data was partitioned into A/B compartments according to the sign of the chromatin accessibility metric termed the Fiedler vector (the eigenvector corresponding to the second smallest eigenvalue of the normalized Laplacian matrix). Positive vector values were assigned to the 'A' compartment representing euchromatic loci and negative values were assigned to the 'B' compartment representing heterochromatic loci. The magnitude of the Fiedler values indicate the level of openness or closedness of the chromatin at a given loci. Hi-C data was at 100kb resolution.

3.6.10 TAD Calling

Topologically associating domains (TADs) were designated using spectral identification as described in [63]. Briefly, the Fiedler vector is calculated for a normalized Hi-C adjacency matrix and initial TADs are organized according to neighboring regions whose Fiedler values have the same sign. The initial TAD structure is repartitioned if, for a given domain, the Fiedler number falls below a user-specified threshold (λ_{thr}). This ensures that TADs are not overly large and repartitions the domains until the Fiedler number is larger than the threshold or until the TAD reaches the smallest allowable TAD size (default is 3). This process is performed iteratively for a set of chromosome-specific Hi-C contact maps. For our analysis, λ_{thr} was chosen to ensure a median TAD size of 900kb, as the expected median TAD size in mammalian genomes is 880kb (rounded to 900kb since bins are in intervals of 100kb)[32]. λ_{thr} was chosen individually for each chromosome to ensure each TAD clustering set would have the same approximate median TAD size. Hi-C data was at 100kb resolution.

3.6.11 Transcription Factor Enrichment Analysis

Transcription factor enrichment analysis was performed using oPOSSUM-3 software [175]. oPOSSUM-3 tests for over-representation of transcription factor binding sites (TFBS) in the DNA sequences of a set of genes, using Z-score and Fisher exact probability to determine significance. For our analysis, 64 genes from a differentially conformed region (*CEACAM* region) of chromosome 19 were submitted as query and compared against a background set of genes in the oPOSSUM-3 database. Both sets of genes were compared to the JASPAR CORE database of preferential TFBS, representing 719 vertebrate TFBS, and the rate of TFBS occurrence for each set of genes was measured [163]. The following analysis parameters were used: 8 bit minimum profile specificity, 0.40 conservation cutoff, 85% matrix score threshold, and a 5kb flanking length.

3.6.12 ChIP Sequencing Analysis

To explore key TF binding patterns throughout the genome and specifically at TAD boundaries, we obtained publicly available SP1, CTCF, and TCF4 ChIP-seq data from the HCT116 colorectal cancer cell line. We accessed the ChIP-seq data from GEO using the accession numbers GSM1010902 (SP1), GSM1010903 (CTCF), and GSM782123 (TCF4). Two replicates were available for SP1 and CTCF, hence replicate profiles were averaged together. Profiles were then binned at 100kb resolution to be compatible with our RNA-seq and Hi-C data. The percentage of nonzero CTCF and TCF4 peak intensities that coincided with TAD boundaries called on the Hi-C data are reported in Tables S5 and S6 of Brown *et al.* [39]. Additionally, because TFEA analysis revealed that SP1 is the most highly enriched TF near the *CEACAM* loci on chromosome 19, we evaluated SP1 peak intensities at these loci. To evaluate the influence of TCF4 binding on the *CEACAM* loci, we aligned the TCF4 ChIP-seq profile to the region as well. Because ChIP-seq data for SP1 and TCF4 were acquired from different experiments, direct comparison of their binding behavior at the *CEACAM* loci was facilitated by performing min-max normalization on each profile.

3.6.13 Synthetic 5C Map

We constructed a synthetic 5C contact map derived from a genome-wide 1Mb Hi-C contact map for genomic regions containing genes in the EMT and E2F gene networks. This was done by locating the genomic bins corresponding to the network genes, extracting the inter-chromosomal and network-specific intra-chromosomal interaction frequencies for those bins, and stitching them together in genomic order. Our working set of network genes was sourced from the KEGG database and included 189 EMT genes and 189 E2F genes distributed across all chromosomes in the genome. Some of the 378 total genes occupy the same 1Mb genomic bins resulting in a 333

by 333 1Mb-resolution synthetic 5C adjacency matrix.

3.6.14 Permutation Test

To determine whether the EMT and E2F gene networks are more closely interacting than other regions of the genome, we performed a permutation test with 1000 randomly sampled sets of 333 1Mb genomic bins across the genome. The Fiedler number for each set was computed and the distribution of numbers was plotted with the observed Fiedler number corresponding to the combined EMT and E2F gene networks. Our null hypothesis is that the EMT/E2F gene networks and random gene sets do not differ and our significance level is 0.1.

3.6.15 Von Neumann Entropy

Entropy is a measure of "disorder" in a given system – the higher the entropy, the greater the disorder. In the context of genome structure, the higher the entropy, the more conformations available to the system [258]. If the distant ends of a genomic region, e.g., a gene, interact to form a loop, there are fewer conformations available to the gene and thus the entropy of that genomic region is reduced. In this way, entropy can be considered a signature for chromatin conformation. We compute the Von Neumann Entropy (VNE) - multivariate entropy - using the following equation:

$$\mathbf{H} = -\sum_{i=1}^{n} \lambda_i \ln \lambda_i. \tag{3.1}$$

Here, **H** takes on positive values with larger values indicating higher chromatin accessibility.

3.6.16 Data Availability

The RNA sequencing and Hi-C datasets generated in this study have been deposited to the Gene Expression Omnibus (GEO) and can be accessed via their accession numbers, GSE151934 and GSE151970, respectively.

CHAPTER 4

Rearrangement of T Cell Genome Architecture Regulates GVHD

This chapter is based on a paper by Yaping Sun, Gabrielle A. Dotson, Lindsey A. Muir, Scott Ronquist, Katherine Oravecz-Wilson, Daniel Peltier, Keisuke Seike, Lu Li, Walter Meixner, Indika Rajapakse, and Pavan Reddy [330].

4.1 Abstract

The cohesin complex modulates gene expression and cellular functions by shaping three-dimensional (3D) organization of chromatin. WAPL, cohesin's DNA release factor, regulates 3D chromatin architecture. The 3D genome structure and its relevance to mature T cell functions *in vivo* is not well understood. We show that *in vivo* lymphopenic expansion, and allo-antigen driven proliferation, alters the 3D structure and function of the genome in mature T cells. Conditional deletion of *Wapl* in T cells reduced long-range genomic interactions, altered chromatin A/B compartments and interactions within topologically associating domains (TADs) of the chromatin in T cells at baseline. Comparison of chromatin structure in normal and WAPL-deficient T cells after lymphopenic and allo-antigen driven stimulation revealed reduced loop extensions with changes in cell cycling genes. WAPL-mediated changes in 3D architecture of chromatin regulated activation, cycling and proliferation of T cells *in vitro* and *in vivo*. Finally, WAPL-deficient T cells demonstrated reduced severity of graft-versus-host disease (GVHD) following experimental allogeneic hematopoietic stem cell transplantation. These data collectively characterize 3D genomic architecture of T cells *in vivo* and demonstrate biological and clinical implications for its disruption by cohesin release factor WAPL.

4.2 Introduction

The three-dimensional (3D) architecture of the genome includes coiling of genomic DNA around histone proteins to form the chromatin fiber, which folds into higher-order structures such as loops, domains, compartments, and chromosomes [32, 107, 220, 267, 266]. High resolution chromatin conformation capture experiments reveal that the 3D spatial architecture of the chromatin at various scales is conserved, reproducible at the cellular level, and regulates gene expression [77, 78, 90, 91]. It is increasingly appreciated that higher spatial organization can have specific alterations during mammalian development and in some pathologies including cancers and infections [107, 368]. However, whether *a priori* disruption of this 3D organization alters *in vivo* cellular functions and disease processes remains poorly understood.

The multi-unit cohesin ring complex plays a critical role in 3D genomic organization and in cell division. It consists of SMC1/3, SCC1 (RAD21), and STAG subunits that are loaded by the SCC2/SCC4 complex onto genomic DNA and establish the cohesin ring structure [79, 235, 259, 280]. Cohesin dependency has been demonstrated by depletion of various cohesin units [259]. Cohesin release from chromatin is driven by WAPL, which opens an exit site at the interface of the SMC3/SCC1 subunits of the cohesin ring [127, 129, 313]. Prior studies have elegantly demonstrated that the absence of WAPL reduces cohesin turnover, alters chromatin loop extensions, and leads to defects in interphase chromosome organization [41, 127, 129, 139, 313, 344].

WAPL is essential during mammalian embryonic development [344]. However, the role of WAPL-dependent chromatin alterations on *in vivo* functions after embryonal development or in mature T cells is not known. The 3D chromatin landscape has been recently described in T cell development, in T cell lines and following *in vitro* stimulation [142, 150, 220, 275, 285, 391, 40, 24]. It remains unclear however whether the specific changes in the 3D chromatin landscape influence or emerge from T cell development. Furthermore, the functional chromatin landscape in mature T cells and their alterations following *in vivo* activation remain unknown. The role of cohesin ring formation in mature T cell development, its function *in vivo*, and its disruption by the absence of WAPL in mature T cell immunity are also not fully understood.

Mature T cells can cause graft-versus-host disease (GVHD) following allogeneic hematopoietic cell transplantation (HCT), a potentially curative therapy against many hematological malignant and non-malignant diseases [29, 385, 398]. GVHD has precluded widespread utilization of this effective therapy. The chromatin landscape of allogeneic T cells, and whether the disruption of this landscape can regulate the severity of T cell-mediated GVHD remains unexplored.

Herein, we utilized genome-wide chromosome conformation capture (Hi-C) and RNA-sequencing to describe the 3D chromatin architecture of mature T cells *in vivo*, specifically at baseline (unstimulated, naive, pre-transplant) and following non-antigen stimulated lymphopenia

induced proliferation (syngeneic) and allo-antigen driven (allogeneic) stimulation [208] (Figure S1 in Sun *et al.* [330]). We evaluated 3D chromatin architecture at the chromosome, TAD, and sub-TAD levels to capture both global and local trends of T cell genome structure (Figure 4.1A). We generated T cell-specific WAPL-deficient mice and demonstrated that WAPL regulates the 3D chromatin structure, function, and *in vivo* biological response of T cells such as GVHD. These data collectively demonstrate that *a priori* disruption of chromatin structure regulates mature T cell function.

4.3 Results

4.3.1 Characterization of mature naïve T cell genome architecture following *in vivo* stimulation

We first determined the genome architecture of mature naïve T cells at baseline and following *in vivo* lymphopenic and antigen-driven stimulation. To mimic clinically-relevant *in vivo* stimulation, we compared naïve T cells before and after experimental syngeneic and allogeneic transplantation. To this end, we utilized MHC-disparate B6 into a BALB/c model of transplantation [276]. $CD62L^{+}$ naive donor T cells were harvested from the splenocytes of B6 donors and transplanted into congenic syngeneic B6 or allogeneic BALB/c recipients (see Methods). Recipient animals were sacrificed seven days after transplantation and their splenic T cells were isolated using congenic markers and analyzed for genomic architecture. While we generated Hi-C contact maps to profile genome-wide chromatin interactions [274] for harvested T cells, variability in the number of mappable Hi-C reads among unstimulated and stimulated T cell settings (Table S1 in Sun *et al.* [330]) precluded us from reliably making direct comparisons between the genome architecture of naive, syngeneic, and allogeneic T cells. From the RNA-seq data, however, we did observe significant differences in gene expression ($|\log_2 FC| \geq 1$, $p \leq 0.001$) before and after *in vivo* stimulation (Figure S2 in Sun *et al.* [330]).

4.3.2 Generation of T cell conditional WAPL-deficient mice

Because T cell function changed following stimulation, we next sought to evaluate which conformational features in the genome are critical for T cell function. Cohesin promotes chromatin looping while WAPL is important for the release of cohesin from chromatin. WAPL is essential for embryonal development and its deficiency has been shown to cause defects in chromatin structure [344]. Specifically, WAPL has been shown to restrict chromosome loop extension [41, 129, 128, 139]. In addition, we previously demonstrated that mir142 regulates mature T cell

proliferation, targets WAPL expression in T cells and that it may modulate T cell activation [333]. Therefore, we determined whether WAPL regulates 3D chromatin architecture and the function of mature T cells.

Because WAPL is critical for embryonal development, we generated a T cell conditional *Wapl* knock-out (KO) mouse using CRISPR-Cas9 and CD4-CRE systems [272, 337]. The *Wapl* locus on chromosome 14q has 18 exons and one non-coding exon (Figure S3A in Sun *et al.* [330]). We designed two sgRNAs specific to exon 2 of the *Wapl* gene to generate a double strand break in exon 2 (Figure 1A and Figure S3B in Sun *et al.* [330]). Our first line of mice carried an insert of two sgRNAs targeting exon 2 in *Wapl* (Figure 1A and Figure S3C in Sun *et al.* [330]). The second and third lines carried Rosa26-floxed STOP-Cas9 knock-in on B6J (The Jackson Laboratory, Stock No:026175) or were CD4-CRE transgenic mice (The Jackson Laboratory, Stock No: 017336) (Figure S3C in Sun *et al.* [330]). Triple crosses were screened for sgRNA insert showing the positive 457 bp band (sgRNA-*Wapl*) (Figure 1B in Sun *et al.* [330]), CRE positive as a 100 bp band, loxP-SV40pA x3-loxP-CAS9-GFP (LSL) cleavage activity demonstrating 1,123 bp for WT LSL, and a 285 bp band for cleaved LSL after CRE recombination processing. The conditional KO mice developed normally. The *Wapl* KO T cells were verified for GFP expression (Figure 1B in Sun *et al.* [330]). The higher levels of LSL cleavage and GFP expression assured successful depletion of WAPL (Figure 1B, lane 4 in Sun *et al.* [330]). We further confirmed WAPL protein depletion through Western blotting (Figure 1C) in T cells isolated from multiple knockout pups. Finally, to further confirm efficient deletion, we performed RNA-seq on the T cells sorted from these mice and compared them with littermate WT T cells, which demonstrated efficient loss of exon 2 in the *Wapl* gene (Figure 1D in Sun *et al.* [330]).

4.3.3 WAPL regulates T cell genome architecture

Because WAPL is known to regulate genome architecture and genome architecture entrains transcription [174, 344, 386], we next evaluated the impact of WAPL deficiency on genome architecture and gene expression of unstimulated naïve T cells. To analyze whole genome architecture (structure) and gene expression (function), we integrated Hi-C and RNA-seq data (see Methods). We studied changes in genome structure and function at the chromosome, TAD, and sub-TAD levels (Figure 4.1A). RNA-seq was performed on T cells harvested from naïve B6 and those harvested on day+7 from transplanted syngeneic and allogeneic B6 recipients (see Methods).

We evaluated the statistical dissimilarity between wildtype (WT) and knockout (KO) Hi-C contact matrices at the chromosome level in each setting by employing the Larntz-Perlman (LP) procedure (see Methods) which revealed regions critically impacted by *Wapl* knockout. WT and KO matrices were significantly dissimilar across all chromosomes ($p < .001$). Though statistically

different, WT and KO chromosomes in unstimulated naïve T cells, except for chromosomes 16 and 19, had a lesser number of differential regions when compared to WT and KO chromosomes in syngeneic or allogeneic T cells (Figure 4.1B and Table S2).This conservation of structure is consistent with the notion that unstimulated naïve T cells are in a quiescent-like state and therefore less impacted by the loss of WAPL. Further, differential regions in the naive setting appear more concentrated between ends of chromosomes – suggesting that long-range interactions may be critical to maintaining structural integrity in unstimulated T cells.

We highlight regions in the 99th percentile of significant changes across chromosomes 1, 2, 7, and 11 in Figure 4.1B, due to their abundance of differential expression and compartmentalization (Figures S4 and S5 in Sun *et al.* [330]). Across all naïve chromosomes, 15-25% of genes within these LP-dissimilar regions are significantly differentially expressed ($p < 0.05$) compared to 8.5-13% across LP-dissimilar regions of syngeneic chromosomes and 3-7% across LP-dissimilar regions of allogeneic chromosomes. Interestingly, there are differential regions in common between the syngeneic and allogeneic settings that are not detected in the naive setting (red arrows in Figure 4.1B), indicating stimulation-specific genome architecture.

We then bi-partitioned chromosomes into individual stretches of accessible (active) and inaccessible (inactive) chromatin, termed A (euchromatin) and B (heterochromatin) compartments, respectively [190]. We demarcated these regions using the signed values of the Fiedler vector which measures underlying chromatin accessibility [61]. The positive values of the Fiedler vector reflect compartment A and negative values reflect compartment B. We observed A/B compartment switch events mediated by the loss of WAPL in all settings (Figure S5 in Sun *et al.* [330]). Across all chromosomes, 184 genomic bins (100kb-length) containing 217 genes in the unstimulated naïve KO T cells occupied a different compartment when compared to WT T cells (Figure S5 in Sun *et al.* [330]). These switch events demonstrated a bias from compartment B to compartment A (70.7% of switch events) (Figure S5 in Sun *et al.* [330]).

In the context of lymphopenic stimulation (syngeneic), KO T cells exhibited 303 switch events involving 375 genes when compared to WT T cells. In the allogeneic context, KO T cells demonstrated 413 switch events involving 485 genes when compared to WT T cells (Figure S5 in Sun *et al.* [330]). The switch bias was once again in the direction of compartment B to compartment A in syngeneic T cells (60.4% of switch events) and in allogeneic T cells (53.8% of switch events). Additionally, allogeneic T cells had more switch events per chromosome than unstimulated naïve and syngeneic T cells (Figure S5 in Sun *et al.* [330]). Interestingly, switched compartment loci tended to congregate towards the ends of chromosomes rather than in the middle or spread evenly throughout. No switch events were observed among contiguous regions of the genome in any of these settings. Overall, 14%, 68%, and 70% of switch events in naive, syngeneic, and allogeneic T cells, respectively, occurred within the previously identified LP-dissimilar regions. While there are

coordinated changes in expression and chromatin accessibility genome-wide and most notably at LP-dissimilar regions, chromatin compartmentalization largely remained stable between WT and KO T cells (Table S2 in Sun *et al.* [330]).

4.3.4 WAPL impacts internal structure of TADs and local gene transcription

Stability in global compartment organization throughout the genome does not preclude changes to local genome organization. Chromatin preferentially interacts within locally-distributed and insulated regions called topologically associating domains (TADs) that regulate transcription [94]. Thus, we next analyzed the impact of WAPL depletion in TADs. One highly supported mechanism of TAD formation is loop extrusion [111] mediated by the ring-shaped cohesin complex and WAPL, which enables TAD dynamics by promoting cohesin turnover [129]. Specifically, genes residing in the same TAD experience coordinated regulation and expression [94, 242, 311, 181] while changes to TAD boundaries due to altered CTCF binding influence anomalous gene-enhancer interactions [109, 205]. We therefore utilized spectral graph theory to identify the positional boundaries that define TADs [63]. TADs have been widely suggested to be highly conserved across cell types and conditions in mammalian genomes [94, 242, 274], so we pooled our Hi-C data across settings to enable higher resolution binning of the data and reliable detection of TAD boundaries within chromosomes (see Methods).

TAD analysis revealed the emergence of a unique feature in the context of syngeneic and allogeneic T cells in the absence of WAPL - the appearance of 'corner peaks'. Corner peaks are the enrichment of interaction frequency at domain boundaries seen at the bottom left and top right corners of TADs [274]. To evaluate differences in corner peak signal between WT and KO T cells, we established a local neighborhood around the corners of each TAD and determined the average number of observed contacts in each neighborhood (Figure 4.2A). On aggregate, we observed that corner peaks became more pronounced upon WAPL depletion in syngeneic and allogeneic T cells (Figure 4.2B and Figure S6 in Sun *et al.* [330]), consistent with the known activity of WAPL to limit loop extension [128, 129]. Further, several studies have demonstrated that corner peaks are associated with a longer residence time of the cohesin complex at TAD borders [129, 304, 336]. We did not, however, observe this rise in corner peak abundance in the KO unstimulated naïve T cells, but rather a decrease (Figure 4.2B and Figure S6 in Sun *et al.* [330]), suggesting that the development of corner peaks in stimulated T cells might be a consequence of their proliferation following activation.

When compared against a background set of interactions – surrounding interactions occurring at a comparable distance to opposing TAD boundaries - we found that the rise in corner peak

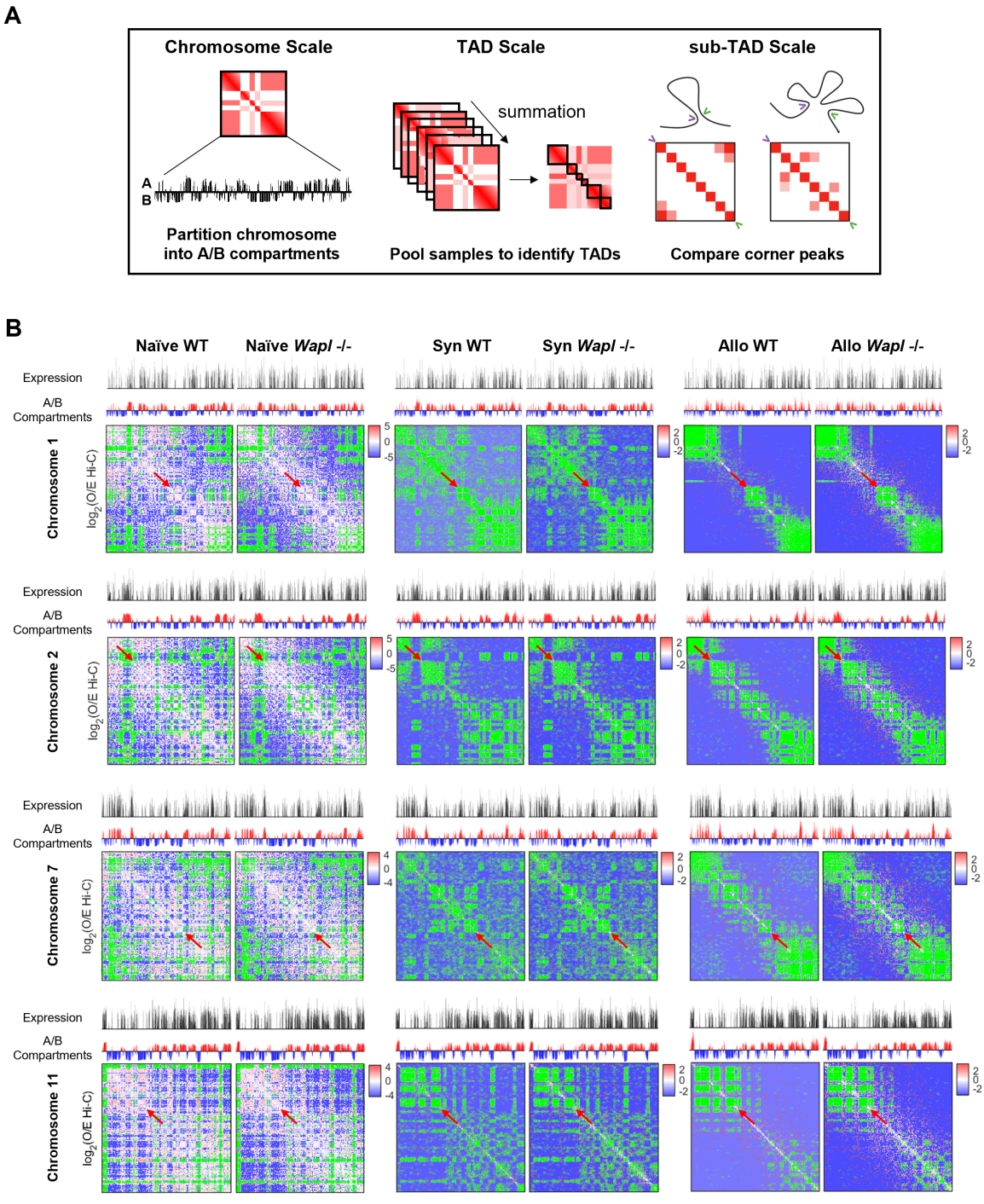

Figure 4.1: Genome-wide Effects of WAPL Knockout in Unstimulated Naïve T Cells and After Syngeneic/Allogeneic Transplantation. (A) Illustration of hierarchical Hi-C analysis workflow.

Figure 4.1: (B) Chromosome-level features and contact maps for chromosomes 1 and 2 from wild-type and *Wapl* knockout T cells. Gene expression (top track) is shown as a vector of $\log_2$(TPM) values binned at 100kb resolution and chromatin accessibility (bottom track) is shown as signed values from the Fiedler vector of the Hi-C contact map where positive values (red) denote A compartments and negative values (blue) denote B compartments. ICE and O/E normalized Hi-C contact maps are shown at 100-kb resolution and log-scale. Matrix dissimilarity between WT and KO conditions was detected by the Larntz-Perlman procedure (see Methods) and genomic regions with dissimilarity in the 99th percentile are shaded in green. These areas indicate the largest perturbations to the chromatin architecture upon *Wapl* knockout.

signal from WT to KO in syngeneic and allogeneic T cells was statistically significant in a large proportion of TADs ($p < 0.01$, Table S3 in Sun *et al.* [330]). We next determined whether there was any functional significance to the observed corner peak dynamics. Syngeneic T cells harbored the most significantly differentially expressed genes (DEGs) ($|\log_2 FC| \geq 1$, $p \leq 0.01$) between WT and KO out of all settings (297 in naïve, 566 in syngeneic, and 46 in allogeneic) and had a particularly high concentration of DEGs on chromosome 11 (58 DEGs, Figure S4 in Sun *et al.* [330]). Given this functional dissimilarity, we further investigated intra-TAD interactions on Chromosome 11 at 50kb resolution (Figure 4.2C) and present an earlier iteration of this analysis on Chromosome 7 at a lower 100kb resolution in Figure S6 (in Sun *et al.* [330]).

We were particularly interested in identifying regions where differential expression overlapped with increases in corner peak signal. We noted statistically significant corner peak increases in 3 naïve, 47 syngeneic, and 50 allogeneic TADs upon WAPL deletion across Chromosome 11, respectively. Of the TADs exhibiting these corner peak increases, two contained DEGs in naïve T cells, 10 in syngeneic T cells, and three in allogeneic T cells. We highlighted two such regions on Chromosome 11 (Figure 4.2C, bottom left and right). The first region we observed ranged from position 105.5Mb to 107.2 Mb on Chromosome 11 and encompassed a TAD exhibiting a significant corner peak signal increase most notably in the syngeneic and allogeneic contexts. Furthermore, this TAD contains the gene, *Wipi1*, which exhibited a 2.8-fold increase in expression from WT to KO in syngeneic cells yet a less significant 0.8-fold increase in expression from WT to KO in both naïve and allogeneic T cells. Another stand-out TAD within region 109.4Mb to 111.1Mb on Chromosome 11 demonstrated differential expression of *Cdc42ep4* in syngeneic T cells with a 1.5-fold increase in expression upon WAPL deletion. Overall, Chromosome 11 contained 25 DEGs in naïve T cells, 58 DEGs in syngeneic T cells, and 4 DEGs in allogeneic T cells dispersed across several TADs.

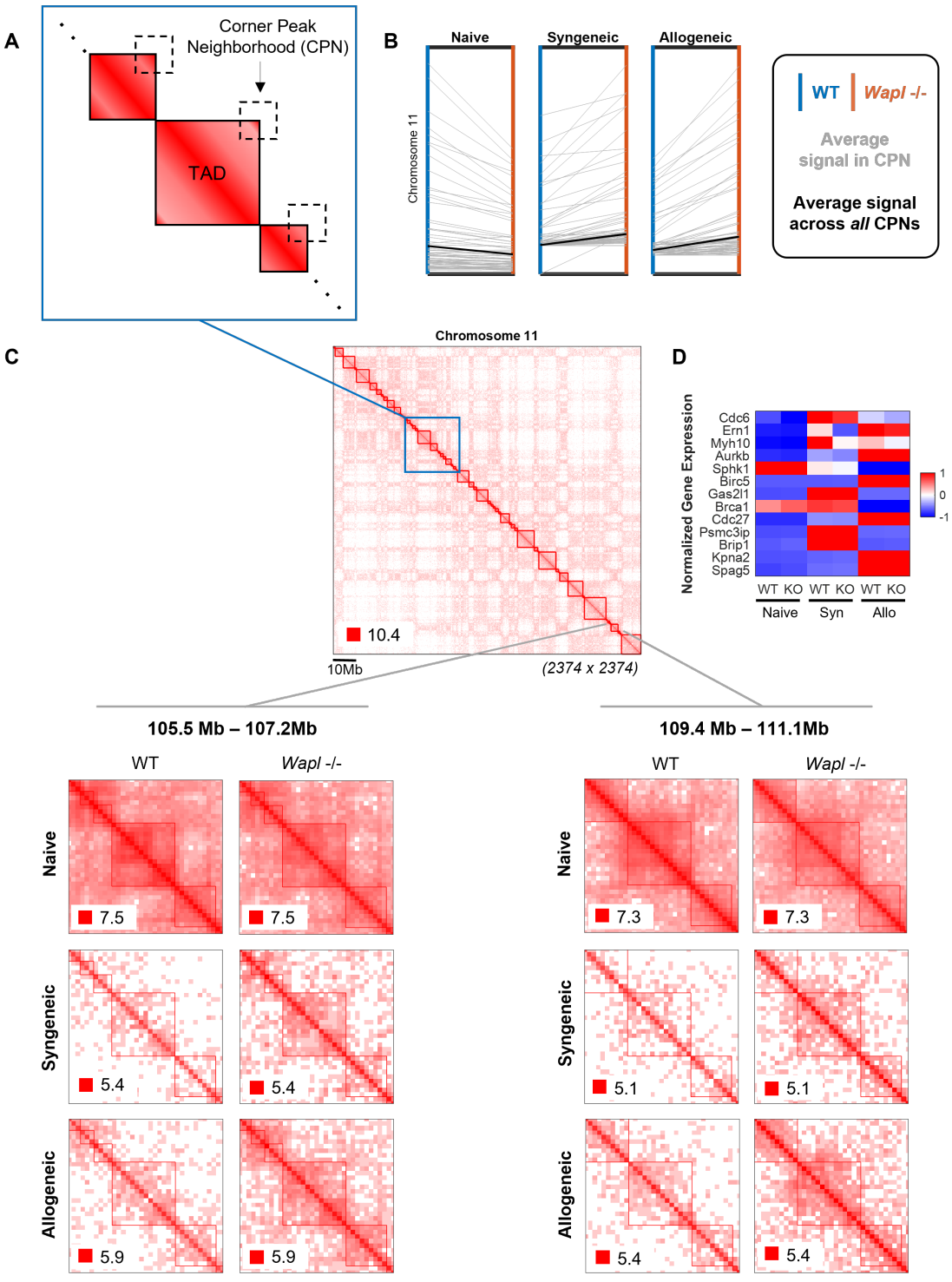

Figure 4.2: Comparison of Internal TAD Organization Between WT and KO T Cells. (A) Schematic of TADs illustrating how 'corner peaks' – interactions between opposite TAD boundaries – were characterized in terms of their local neighborhoods.

Figure 4.2: The nonzero mean of absolute read counts in each local corner peak neighborhood (CPN) was used to define the strength of TAD boundary interactions. (B) Change in corner peak signal from WT (blue axis) to KO (orange axis) for each TAD (individual gray lines) on Chromosome 11. The average change in corner peak signal between the two conditions is plotted as a solid black line in each panel to capture the overall trend across settings. TADs and their corner peak signals are represented at 50kb resolution. (C) (Top center) Chromosome 11 observed contact map pooled across samples at 50kb resolution. (Bottom left) TAD region of Chromosome 11 extending from 105.5 Mb to 107.2 Mb capturing corner peak increase from WT to KO most notably in the syngeneic and allogeneic settings; differential expression of gene *Wipi1* occurs in the centermost TAD in this region. (Bottom right) TAD region of Chromosome 11 extending from 109.4 Mb to 111.1 Mb capturing corner peak increase from WT to KO; differential expression of gene *Cdc42ep4* occurs in the centermost TAD in this region. (D) Heatmap of differentially expressed cell cycle genes residing on Chromosome 11.

Impact of WAPL on the Cell Cycle Gene Network

WAPL is critical for sister chromatid cohesion and loop extrusion dynamics. These processes correlate with cellular proliferation and cell cycling, and because we observed notable changes in corner peaks of Chromosomes 7 and 11 only in the context of syngeneic and allogeneic settings in the absence of WAPL, we next explored changes in cell cycle genes and the TADs that these genes reside in. One gene-rich region in Chromosome 7, extending between positions 60Mb and 70Mb on the chromosome, exhibited considerable intra-TAD reorganization in the KO T cells when compared to WT unstimulated (Figure S7D in Sun *et al.* [330]), syngeneic (Figure S7E in Sun *et al.* [330]), and allogeneic (Figure S7F in Sun *et al.* [330]) T cells. This region contains cell cycle genes [96, 147]: *Fanci*, *Prc1*, and *Blm*. These cell cycle genes (*Fanci*, *Prc1*, and *Blm*) were not significantly differentially expressed, as they did not meet the cutoff of $|\log_2 \text{FC}| \geq 1$. However, there were seven non-cell cycling genes (*Mesp2*, *Arpin*, *Fes*, *Homer2*, *Saxo2*, *Cemip*, and *Arnt2*) directly upstream and downstream of the cell cycle genes in this region that were significantly differentially expressed. These data suggest that in the unstimulated KO T cells, these three cell cycle genes on Chromosome 7 were not differentially expressed despite the changes in intra-TAD interactions when compared to WT T cells.

Similar to the unstimulated context, in the syngeneic T cells, there was a differential expression of non-cell cycle genes upstream and downstream of cell cycle genes in this region (*Agbl1*, *Isg20*, *Can*, *Hapln3*, *Ribp1*, *Mesp2*, *Anpep*, *Fes*, *Slc28a1*, *Homer2*, *Adamtsl3*, and *Tmc3*). In allogeneic T cells, non-cell cycle genes (*Agbl1*, *Can*, *Rhcg*, and *Adamtsl3*) once again were differentially expressed. Thus, in the absence of WAPL, the highlighted region of Chromosome 7 containing the three cell cycle genes did not dynamically change but demonstrated significant changes in the expression ($|\log_2 \text{FC}| \geq 1$, $\text{p} \leq 0.05$) and internal structural rearrangement of genes in their vicinity. No other cell cycle genes on Chromosome 7 demonstrated differential expression.

Chromosome 11 contained 13 differentially expressed cell cycle genes (Figure 4.2D). Of these 13, only one gene (*Gas2l1*) was significantly differentially expressed between naïve WT and KO T cells, two genes (*Myh10* and *Sphk1*) between syngeneic WT and KO T cells, and none in the allogeneic setting. We did not however observe an overlap between the differential expression of these genes and TADs with altered internal TAD structure. While WAPL depletion did not disrupt the cell cycling transcriptional program within unstimulated and stimulated contexts much on Chromosome 11, the expression of these genes was quite different between contexts (Figure 4.2D).

While we did not see a change in the three cell cycle genes noted above in the analysis of Chromosome 7 nor significant coupling of cell cycle differential expression and corner peak dynamics on Chromosome 11, WAPL is known to regulate cell cycling so we next explored the entire breadth of cell cycle genes throughout the genome. To this end, we extracted and stitched together 141 Hi-C genomic bins that corresponded to a curated set of 170 cell cycle genes genome-wide, generating a Hi-C-derived 5C contact matrix (see Methods). 5C, like Hi-C, is a derivative of the original chromosome conformation capture technique [91], and is useful in identifying interactions among select genomic regions that bear relationship to one another [98]. In unstimulated naïve T cells, we observed that the connectivity of the cell cycle network decreased in the absence of WAPL, as determined by the element-wise Pearson correlations moving toward zero (Figure 4.3A). In syngeneic and allogeneic T cells, however, connectivity appeared to strengthen in the absence of WAPL, as demonstrated by the correlation tending towards ± 1 (Figure 4.3A).

The connectivity (structure) of cell cycle genes changed between the WT and KO T cells in all settings, but the maximal change was noted in the syngeneic setting (Figure 4.3B). However, changes in the structure of the cell cycle network did not trend with the function (expression). The expression (function) of the cell cycle gene network changed the most between WT and KO T cells in the context of allogeneic setting (Figure 4.3B). The genome-wide structural analysis of the cell cycle gene network highlighted two subgroups of highly connected genes (shown in purple and green boxes in Figure 4.3A). One of these subgroups negatively correlated with most of the network (Figure 4.3A, purple box) while the other demonstrated positive correlation with most of the network (Figure 4.3A, green box) in all three settings. The measure of inter-connectivity within these two subgroups, in terms of degree centrality, is shown in Figure 4.3C. Ultimately, while gene expression of cell cycle genes *Fanci*, *Prc1*, and *Blm*, did not change in the absence of WAPL, other genes in the cell cycle network did change significantly (Table S4 in Sun *et al.* [330]). Five cell cycle genes were up-regulated in the absence of WAPL in unstimulated naive T cells. By contrast, 12 cell cycle genes were down-regulated and 4 up-regulated between WT and KO T cells in the syngeneic context. In the context of allo-antigen stimulation, only one gene was down-regulated.

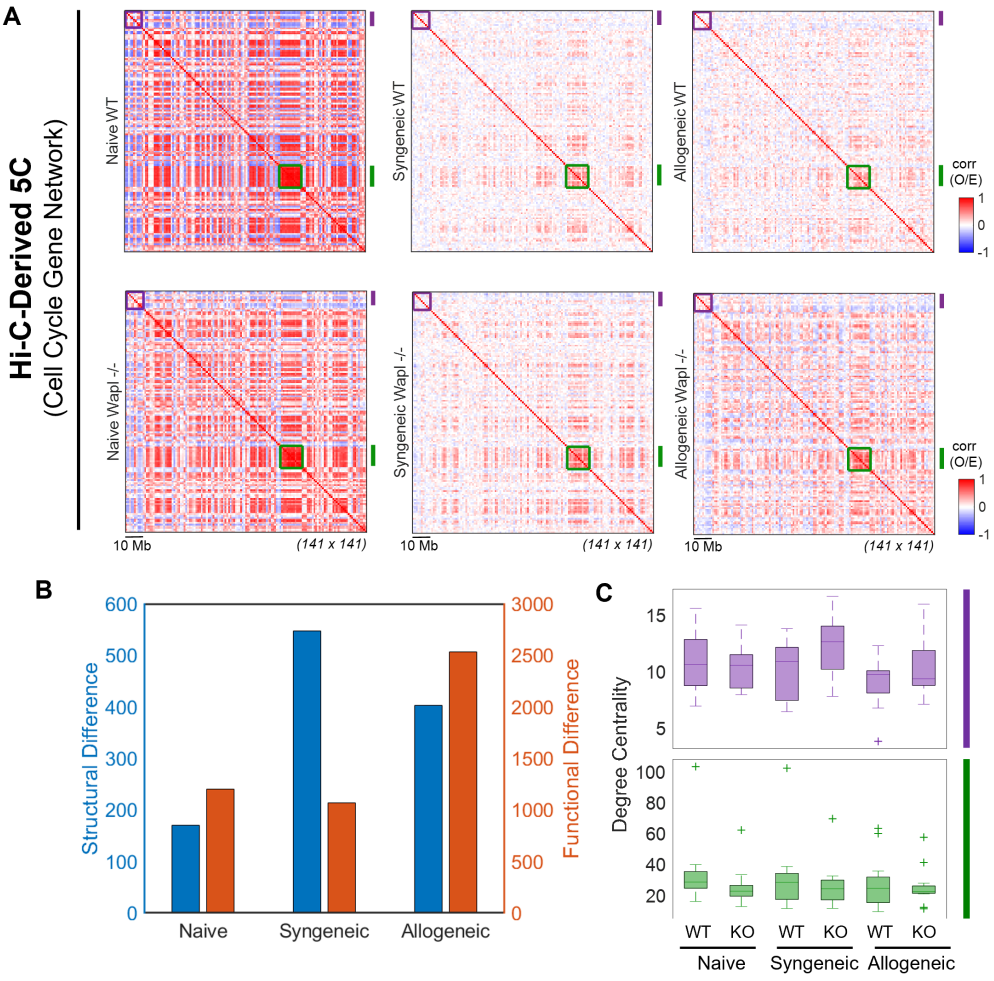

Figure 4.3: Cell Cycle Gene Network Across T Cells. (A) Hi-C-derived 5C contact maps representing the cell cycle gene network. Rows and columns correspond to genomic bins across all chromosomes that contain cell cycle genes. Maps are shown at 1-Mb resolution as a Pearson correlation of normalized (observed/expected) contacts. The purple and green boxes highlight subgroups of interest that are assessed in C. The purple subgroup is comprised of 10 loci (1Mb length) ranging non-contiguously from 18 Mb to 197 Mb and containing 10 cell cycle genes. The cell cycle gene loci in WT and KO T cells in each subgroup is comprised of 12 loci (1Mb length) ranging non-contiguously from 1,489 Mb to 1,606 Mb and containing 16 cell cycle genes. (B) Structural and functional differences between all cell cycle gene loci in WT and KO T cells in each setting. Difference is measured as the Frobenius norm of the difference between WT and KO data (see Methods). The left y-axis (blue) reflects the Frobenius norm of the difference between WT and KO Hi-C matrices (measure of structural change) in each setting and the right y-axis (orange) reflects the Frobenius norm of the difference between WT and KO gene expression vectors (measure of functional change) in each setting. (C) Degree centrality (i.e. the row sum of a matrix) of subgroups of interest.

4.3.5 WAPL-induced changes in genome structure alter T cell gene expression

We next determined whether the changes in the chromatin architecture related to WAPL deficiency in T cells affected genome function. Three-dimensional genome structural changes have been suggested to affect T cell development [307]. Whether *a priori* disruption of the cohesin ring affected T cell function, however, is not clear. The transgenic mice with conditional *Wapl* KO in T cells displayed normal birth and growth rates and generated enough mature naïve T cells. To better analyze the developmental impact of WAPL deficiency on T cell development, we next analyzed the thymii from 8–10-week-old WT and *Wapl* KO littermates. The thymocytes from the *Wapl* KO animals showed significant changes when compared to the WT littermate controls. Specifically, they showed reduction in the total number of thymocytes, double positive (DP), and CD8 SP thymocytes (Figure 5A and Figure S8A in Sun *et al.* [330]).

We next analyzed the secondary lymphoid organ (spleens) for peripheral T cell subsets. *Wapl* KO littermates demonstrated lower numbers of total T cells in the spleen (Figure 5B in Sun *et al.* [330]) and lower numbers of $CD3^+CD4^+$ and $CD3^+CD8^+$ T cells (Figure S8B and C in Sun *et al.* [330]) when compared to WT littermates. During the β-selection checkpoint in the thymus, the β chain of the T cell receptor rearranged by the thymocytes must retain the structural properties allowing it to be presented on the surface of the thymocytes. We therefore screened for TCRβ^+ $CD4^+$ and $CD8^+$T cells in the spleen and found that the *Wapl* KO mice demonstrated lower numbers of TCRβ^+ $CD4^+$ and $CD8^+$ T cells in the spleen when compared to WT littermates (Figure S8D and E in Sun *et al.* [330]). The splenocytes from the KO animals also demonstrated lower numbers of $CD3^+CD69^+$, $CD4^+CD69^+$, $CD8^+CD69^+$ and $CD4^+CD25^+$ T cells (Figure S8A-D in Sun *et al.* [330]).

Because WAPL deficiency altered chromatin architecture, we next analyzed whether these structural changes were associated with changes in gene expression and proliferation, at baseline and following *in vivo* stimulation. To this end, we first determined whether *Wapl* expression itself changed in T cells following stimulation. Consistent with our previous observation, we observed significant upregulation of WAPL protein in T cells upon co-culture with allogeneic DCs for 60 hours (Figure 5C in Sun *et al.* [330]) [331, 332]. To determine the impact of this structure-function relationship, we performed gene set enrichment analysis (GSEA) for the cell cycle gene network highlighted in our earlier Hi-C-derived 5C analyses (Figure 4.3). We found that cell cycle genes were differentially regulated between WT and WAPL-deficient T cells (Figure 5D in Sun *et al.* [330]). We verified results with real-time quantitative PCR. *Gas2l1*, a gene induced upon growth arrest [121], and *Mcm3* that increases as cells progress from G0 into the G1/S phase and regulates the cell cycle [352], were significantly upregulated in WAPL-deficient T cells compared

to WT T cells (Figure 5E and F in Sun *et al.* [330]). In contrast, *Sphk1*, a gene that plays a key role in TNF-a signaling and the canonical NF-κB activation pathway important in inflammatory, apoptotic, and immune processes [8], and *Myh10*, a gene required for completion of cell division during cytokinesis [165, 325] were both downregulated in WAPL-deficient T cells (Figure 5G and H in Sun *et al.* [330]). These observations demonstrated that changes in genome structure related to WAPL deficiency in T cells promote differential gene expression that suppresses apoptosis, cell proliferation and cell cycle progress.

To further investigate the impact of WAPL deficiency on T cell cellular responses, we stimulated WAPL-deficient and WT T cells *in vitro*, either with allo-antigen stimulation in mixed lymphocytes reaction (MLR) by co-culturing T cells with allogeneic DCs for four days or with CD3/CD28Ab for two or three days and pulsed with H3-TdR. The WT T cells demonstrated significantly greater H3-TdR incorporation when compared with *Wapl* KO T cells suggesting that WAPL deficiency caused reduced proliferation (Figure 5I and Figure S9E in Sun *et al.* [330]). T cell proliferation was directly assessed *in vitro* using a dye dilution assay. Specifically, we utilized CellTrace™FarRed dilution to avoid the interference with GFP fluorescence in *Wapl* KO T cells. The *Wapl* KO T cells demonstrated reduced dye dilutions when compared to WT T cells upon stimulation by either allogeneic DCs or CD3/CD28ab (Figure 5J and Figure S8F in Sun *et al.* [330]). It is possible that cell death from apoptosis could contribute to the reduction in T cell expansion related to WAPL deficiency. We therefore also determined apoptosis after stimulation of both WT and *Wapl* KO T cells. The WAPL-deficient T cells also showed significantly decreased apoptosis when compared to littermate WT T cells (Figure 5K in Sun *et al.* [330]). These data demonstrate that the absence of WAPL altered T cell proliferation and apoptosis following *in vitro* stimulation.

The interaction between cohesin and WAPL plays an essential role in maintaining chromosome structure and separation of sister chromatids during mitosis and cell proliferation [256]. With the coordinated changes in expression of cell cycling genes and genomic architecture in the absence of WAPL, and the reduction in proliferation, we next examined cellular mechanisms underlying the reduced proliferative capacity of WAPL-deficient T cells following *in vitro* and *in vivo* stimulation. Specifically, we explored the hypothesis that altered gene expression of cell cycling genes from the change in genomic architecture related to WAPL deficiency leads to a reduction in cell cycling. To assess kinetic cell cycling we utilized flow cytometry analyses of DNA content with FxCycle™Far-red Stain to avoid interference from the GFP fluorescence in *Wapl* KO T cells. Purified WT or *Wapl* KO T cells were stimulated with CD3/CD28T cell activator dynabeads and analyzed for 2C and >2C populations. *Wapl* KO T cells demonstrated significantly lower 2C percentages but higher >2C percentages when compared with WT T cells at several time points after CD3/CD28 stimulation (Figure 5L and Figure S10A in Sun *et al.* [330]). We next analyzed whether the effect on cell cycling was observed *in vivo*. To this end, we once again utilized the allogeneic

bone marrow transplantation (BMT) model system, where the host splenocytes were harvested and analyzed for WT or *Wapl* KO donor T cells as above. As shown in Figure 5M and Figure S9B (in Sun *et al.* [330]), the *in vivo* stimulation of WT and *Wapl* KO T cells demonstrated similar differences as from *in vitro* stimulation. These data indicated that *Wapl* KO T cells had cell cycle deficiency, which impairs T cell proliferation and alters T cell function *in vitro* and *in vivo*, consistent with cell cycle gene analyses by Hi-C derived 5C (Figure 5 in Sun *et al.* [330]), GSEA, and qPCR analyses (Figure 5D-H in Sun *et al.* [330]).

4.3.6 WAPL deficiency in T cells improves survival after allogeneic BMT

Mature T cells in the allogeneic donor T cells are the principal mediators of GVHD, a major cause of mortality after allogeneic BMT. Because WAPL regulated mature T cell responses following *in vitro* and *in vivo* allo-antigen stimulation, we next determined whether this has a clinical impact on GVHD severity following experimental allo-BMT. To this end, we once again utilized the MHC mismatched B6 (H2) $\rightarrow$ BALB/c (H2d) mouse model where the congenic B6 animals served as the syngeneic controls. BALB/c recipient mice were lethally irradiated (800 cGy total body irradiation, split dose) and transplanted with T cell-depleted WT BM from B6 donors along with purified mature T cells from the spleen of either WT or *Wapl* KO B6 animals[332]. The recipient animals were monitored for survival and GVHD severity as described in Methods.

We first determined whether WAPL expression changed after allo-BMT. Consistent with the *in vitro* allo-antigen stimulation shown in Figure 5C (in Sun *et al.* [330]), WT cells demonstrated higher expression of WAPL protein in donor T cells harvested from recipient spleens 7 days after allogeneic BMT when compared to syngeneic controls (Figure 6A in Sun *et al.* [330]). Survival analysis demonstrated that all the syngeneic animals survived, but the allogeneic animals that received WT T cells died with signs of severe clinical GVHD (Figure 6B in Sun *et al.* [330]). In contrast, allogeneic animals that received WAPL-deficient T cells showed significantly improved survival (53% versus 17%; $p < 0.01$) (Figure 6B in Sun *et al.* [330]) and reduced clinical severity of GVHD ($p < 0.01$) (Figure 6C in Sun *et al.* [330]). We confirmed the reduction in GVHD with detailed histopathological analyses of GVHD target organs including the liver, gastrointestinal (GI) tract, and skin. As shown in Figure 6D (in Sun *et al.* [330]), allogeneic animals that received T cells from WAPL-deficient donors had significantly reduced histopathological GVHD in the liver ($p < 0.01$), GI tract (SI and LI) ($p < 0.05$) and skin ($p < 0.05$) on day +21 after BMT. Consistent with reduced mortality, the recipients of allogeneic *Wapl* KO T cells demonstrated reduced serum levels of proinflammatory cytokines such as IFN-γ and TNF-α when compared with WT T cell recipients (Figure 6E and F in Sun *et al.* [330]).

The allogeneic animals that received WT or KO T cells demonstrated >98% donor engraftment

on day 21, ruling out mixed chimerism or engraftment as a cause for reduction in GVHD. Furthermore, consistent with above results, *Wapl* KO T cells showed significantly reduced expansion when compared with WT donor T cells (Figure 6G in Sun *et al.* [330]) in the recipient spleens harvested 7 days after BMT, suggesting that the reduction in GVHD was a consequence of reduced expansion of allo-antigen stimulated T cells. Consistent with this, the allogeneic *Wapl* KO T cells demonstrated fewer absolute numbers of $CD3^{+}CD69^{+}$ cells (Figure 6H in Sun *et al.* [330]) and regulatory T cells (despite a higher percent) when compared to WT T cells (Figure 6I and J in Sun *et al.* [330]).

4.4 Discussion

The cohesin complex and its regulators, such as WAPL, are critical determinants of 3D genome architecture, which regulates replication, repair, and transcriptional processes [79, 220, 235, 280]. Given the paucity of data on the genomic organization of mature T cells and its impact on T cell function *in vivo*, we describe the genome architecture of mature T cells following *in vivo* lymphopenic and allogeneic stimulation.

Prior to the availability of Hi-C methods, structural features of the T cell genome following *in vitro* stimulation were explored in a seminal paper by Kim *et al.* [166]. Since then, other studies have built on it with the advent of Hi-C techniques following *in vitro* stimulation of T cells with anti-CD3/CD28 [391, 40, 24]. We now expand on those studies by exploring the changes following *in vivo* homeostatic and antigen-driven stimulation. Previously, chromosome 6 in differentiating T cells was explored in its entirety, though global spatial characterization was limited to chromosome territories [166]. Here, we profile the entire genome architecture of mature T cells following *in vivo* activation, including local structures like TADs. We further connect those changes to genome-wide functional changes in gene expression and T cell responses. Hu *et al.* demonstrated a key role for transcription factor BCL11B in the development of T cells and associated genome structure changes [142], but did not characterize the relevance of structural integrity to mature T cell functions. Our data focus on the genomic landscape of mature T cells. We define a mechanistic role for genomic architecture in the regulation of gene expression, cellular function, and biological responses *in vivo* that impact clinically-relevant disease states such as GVHD. We find that differences in 3D genome architecture are a consequence of the cellular state of activation, and that its disruption, by altering the function of the cohesin complex, contributes to the regulation of mature T cell functions in response to allogeneic-stimulation.

We explored the mechanistic relevance of genome structure to function (gene expression) and to the cellular functions and biological impact of mature T cells by deleting WAPL, a key regulator of genome structure. WAPL deletion led to reduction of long-range interactions in the baseline

unstimulated state of naïve T cells. However, following lymphopenic (syngeneic) and antigen (allogeneic) activation, there was a greater amplification of longer-range interactions following WAPL deletion. This is consistent with previous reports on WAPL-deficient cell lines [129, 197] and with the notion of extruded DNA loops beyond CTCF barriers [5]. These data suggest that WAPL alters genomic architecture, at baseline and after activation of T cells.

The *in vivo* role for WAPL-mediated regulation of chromatin structure is crucial for embryonal development [344]. We now extend these studies and demonstrate that WAPL also regulates *in vivo* immune responses mediated by T cells. T cell-specific WAPL-deficient animals developed normally despite the T cells showing genomic structural changes at homeostasis. Upon stimulation, WAPL-deficient T cells demonstrated mitosis defects and more axial structures during interphase suggesting that they exit mitosis with less intact cohesin. However, it is important to note that the absence of WAPL neither caused a complete loss of development of T cells nor a total shutdown of mature T cell proliferation. The T cells developed and proliferated in the absence of WAPL, albeit at a much lower level, the reasons for which remain unclear. One possible explanation is that separation of chromatids during mitosis in T cells may be independent, or only partially dependent, on WAPL, based on their strength and duration of stimulation/activation [244, 319, 395]. This notion is consistent with the observation that several cohesinopathies in humans are caused by mutations in various components of cohesin complex and yet do not appear to cause T cell defects [259, 280, 315]. Future studies may determine the role of WAPL and disruption of cohesin in thymopoiesis and in regulation of T cells in secondary lymphoid organs.

The structural changes in our study reflect a polyclonal response from a combination of allore-active/lymphopenia induced proliferating cells, and the non-proliferating mature T cells. The T cells therefore might be in various stages of early/mid G1, S, G2, M phases. Thus, the genome contact frequency and associated structural changes are reflective of the T cells in these various stages after stimulation. T cells from naïve, syngeneic, and allogeneic settings demonstrated significant changes in internal TAD structure, which were consistent despite the polyclonal nature of the T cell subsets. Nonetheless, while the genome architectures were significantly different, whether these are the direct cause or a consequence of proliferation differences cannot be definitively ascertained. Furthermore, whether antigen-specific T cell responses vary based on the type of antigen cannot be determined from our study. However, our assessment of polyclonal responses is akin to biologically and clinically relevant situations such as allogeneic transplantation.

Our data collectively demonstrate for the first time, to our knowledge, that altering genomic structure *a priori*, at baseline, regulates T cell gene and cellular functions. Though possible that WAPL could regulate gene expression independent of genome structure, it is widely held that form precedes function and that genome structure and function are coordinated [52, 266]. Mechanistically, the data show that WAPL alters chromatin architecture at cell cycling genes and therefor

links genome structure and function. Nonetheless, validation studies to confirm the exact mechanistic role of WAPL in this context will need to be explored. Additionally, while we only validate select cell cycling genes concomitant with changes in genome structure in this study, it is possible that other cell cycling genes might be involved in the proliferative defects we observe and future studies will need to validate these in a systematic manner as well. Future studies may also benefit from integrating single-cell Hi-C and single-cell RNA-seq to refine observations, as features at the TAD and sub-TAD levels observed in this study via bulk Hi-C are a population average and could be variable and distinct between various T cell subsets that develop and differentiate after stimulation. Regardless, our data indicate that disruption of 3D chromatin architecture by WAPL may directly regulate gene expression and cellular function of T cells in a physiologically and clinically relevant disease context.

Importantly, we introduce a simple yet robust approach for evaluating TADs and their internal structure. Due to the experimental limitations of performing Hi-C in an *in vivo* setting we obtained a lower number of cells and thus mappable reads than what Hi-C performed on cell lines or in an *in vitro* setting would typically yield. Consequently, to reliably investigate TAD and sub-TAD level features, we pooled data across all our samples allowing us to bin our data at a higher resolution. From this pooled matrix, we were able to determine TAD boundaries and subsequently superimpose them onto each sample's raw Hi-C contact map. Informed by this common TAD backbone, we were able to perform targeted analysis of the internal TAD structure in each sample and evaluate how sensitive the backbone and its internal organization are to perturbations across the different settings.

There are no reported cases of isolated germline WAPL mutations in humans [116, 259, 280, 315]. This is likely because of its critical role during embryonic development. However, somatic mutations in WAPL have been linked to epithelial carcinogenesis [368]. Our study demonstrates that WAPL deficiency in T cells did not cause a profound defect in development of T cells, nor cause T cell malignancy. Thus, WAPL may play a nuanced role in different cell subtypes, depending on their developmental stage, context, and stimulation. Future studies will need to carefully assess the biological implications of WAPL deficiency on T cell subsets and other immune cells. Because WAPL-deficient T cells demonstrated a reduction in GVHD, it is tempting to speculate whether targeting it uniquely in T cells might be a novel strategy to mitigate immunopathologies such as GVHD, allograft rejection, or autoimmunity. Such a strategy may be feasible with emerging gene editing strategies for adoptive transfer of T cells, however, the viability of the strategy to delete WAPL clinically will need significant further investigation. At a broader level, our data provide a proof of concept for the notion that targeting 3D genomic architecture may be a therapeutic strategy that can be potentially harnessed for directly modulating *in vivo* disease processes.

4.5 Materials and Methods

4.5.1 Mice

B6 (H2b, CD4.1 and CD45.2), BALB/c (H2d, CD45.1 and CD45.2), Rosa26-floxed STOP-Cas9 knockin (B6J) and CD4CRE (B6) mice were purchased from The Jackson Laboratory and the National Cancer Institute. The ages of the mice used for experiments ranged between 8 and 12 weeks. Mice were housed in sterilized microisolator cages and received filtered water and normal chow.

4.5.2 Generating conditional *Wapl* KO T cells

We designed 2 sgRNAs targeting exon 2 of the *Wapl* locus which are localized in mouse Chromosome 14 qB (Figure S3 in Sun *et al.* [330]) [272]. Three lines of mice with B6 background were used to generate conditional *Wapl* KO mice in T cells. The first line was generated to carry CRISPR-sgRNAs targeting insert to identify exon 2 in *Wapl* (Figure 1A in Sun *et al.* [330]). The second and third lines are Rosa26-floxed STOP-Cas9-GFP knockin on B6J (The Jackson Laboratory, Stock No:026175) and CD4-CRE transgenic mice (The Jackson Laboratory, Stock No: 017336). Tail DNA was screened for a positive sgRNAs-*Wapl* insert, positive CRE, and a stop signal in loxP-SV40pA x3-loxP cleavage or deletion. The potential *Wapl* KO T cells were sorted for positive GFP (Figure 1B in Sun *et al.* [330]). To confirm the success of conditional in KO in T cells, T cells were processed for SDS-PAGE and detected with anti-*Wapl* antibody. Additionally, total RNAs were isolated from *Wapl* KO and WT T cells and processed for RNA-seq.

4.5.3 DC isolation and purification

Dendritic cells (DC) were isolated from splenocytes either from WT B6 or BALB/c mice. Single-cell suspensions were prepared according to the manufacturer's instruction then subjected to CD11c microbead (MACS) staining and positive selection using LS column (Miltenyi Biotec). The purity of enriched $CD11c^+$ DC preparation was 85.6-90%.

4.5.4 T cell isolation and purification, and mixed lymphocyte reaction (MLR)

WT and *Wapl* KO B6 T cells were isolated by negative selection (>95% purity) (Pan T Cell Isolation Kit II; Miltenyi Biotec). T cells were co-cultured with B6 WT or BALB/c DCs at a ratio of 40:1 (T cells versus DCs 3 × 105:7.5 × 103) for 96 hours using 96-well flat-bottomed plates

(Falcon Labware), or stimulated with Dynabeads T cell activator CD3/CD28 (25 ul/106/ml) for 2 or 3 days respectively. Proliferation was determined by incubating the cells with H3-thymidine (1 Ci/well [0.037 MBq]) for the last 20 or 6 hours, respectively. H3-thymidine incorporation in T cells was counted on a 1205 Betaplate reader (Wallac, Turku, Finland).

4.5.5 BMT and systemic analyses of GVHD

Bone marrow transplantations (BMTs) were performed as described previously [332]. The donor T cells (WT B6 or WAPL deficiency) were isolated from spleens and purified by negative selection (using the Pan T cell Isolation Kit II; Miltenyi Biotec). Bone marrow cells from tibia and fibula were harvested and TCD (T cell deletion) BM cells were isolated with positive deletion using anti-CD90.2 microbeads and LS column (Miltenyi Biotec). The recipient BALB/c mice received an 800-cGy total body irradiation on day -1 (split dose) and T cells (1 × 106, either B6 WT or WAPL deficiency T cells, and TCD BM cells (5 × 106, from WT B6 mice) were injected intravenously into the recipients on day 0. The syngeneic B6 control mice received a 1000-cGy total body irradiation on day -1. T cells (2 × 106, isolated from B6 WT mice) and TCD BM cells (5 × 106 from B6 WT mice) were injected intravenously into the recipients on day 0. Mice were housed in sterilized microisolator cages and received normal chow and autoclaved hyperchlorinated drinking water for the first 3 weeks after BMT.

4.5.6 Immunoblotting

T cell lysates were extracted as previously described [332], and 50 to 100 μg of protein extract was separated in SDS-PAGE and transferred onto a PVDF membrane (GE Healthcare). The membrane was blocked with 5% nonfat milk for 30 minutes and then incubated overnight at 4°C with the following Abs in 5% nonfat milk: anti-*Wapl* rabbit polyclonal Ab (1:500 in nonfat milk, Proteintech Cat 16370-1-AP), anti–β-actin mouse mAb (1:1,000 in 5% nonfat milk; Abcam, catalog ab8226). After washing 3 times with TBST for 5 minutes, the blot was incubated with specific HRP-labeled secondary Ab, washed again with TBST, and signals generated and visualized using the Enhanced Chemiluminescence Kit (Thermo Scientific, Cat 32106).

4.5.7 Study approval

Study approval. All animal studies were reviewed and approved by the University Committee on Use and Care of Animals of the University of Michigan, based on University Laboratory Animal Medicine guidelines.

4.5.8 GVHD and pathology scoring

Survival was monitored daily. The degree of systemic GVHD was assessed by a standard scoring system with four criteria scores: percentage of weight change, posture, activity, fur texture, and skin integrity, and subsequently graded from 0 to 2 for each criterion (maximum index = 10) [332]. Acute GVHD was also assessed by histopathologic analysis of the ileum and the ascending colon, liver, and ear skin. Specimens were harvested from animals on day 21 after BMT, then processed and stained with hematoxylin and eosin. Coded slides were examined systematically in a blind manner by using a semi-quantitative scoring system to assess the following abnormalities known to be associated with GVHD, small intestine: villous blunting, crypt regeneration, loss of enterocyte brush border, luminal sloughing of cellular debris, crypt cell apoptosis, outright crypt destruction, and lamina propria lymphocytic infiltrate; colon: crypt regeneration, surface colonocytes, colonocyte vacuolization, surface colonocyte attenuation, crypt cell apoptosis, outright crypt destruction, and lamina propria lymphocytic infiltrate; liver: portal tract expansion, neutrophil infiltrate, mononuclear cell infiltrate, nuclear pleomorphism, intraluminal epithelial cells, endothelialitis, hepatocellular damage, acidophilic bodies, mitotic figures, neutrophil accumulations, macrophage aggregates, macrocytosis; skin: apoptosis in epidermal basal layer or lower malpighian layer or outer root sheath of hair follicle or acrosyringium, lichenoid inflammation, vacuolar change, lymphocytic satellitosis. The scoring system denoted 0 as normal, 0.5 as focal and rare, 1.0 as focal and mild, 2.0 as diffuse and mild, 3.0 as diffuse and moderate, and 4.0 as diffuse and severe. Scores were summed together to provide a total score for each specimen [332].

4.5.9 RNA-seq library generation and data processing

Naïve WT and *Wapl* KO T cells, or WT and *Wapl* KO T cells isolated from syngeneic or allogeneic BMT on day 7 were first purified as described previously [332]. All dead cells were excluded by sorting for far-red fluorescent reactive dye (Invitrogen, Cat. L10120). Then, WT T cells were sorted for PE-CD3 and APC-CD45.2 while *Wapl* KO T cells were sorted for APC-CD3, PE-CD45.2 and positive GFP. Each sample contained pooled T cells from 3-4 mice, and samples in each group were biologically triplicated. RNA was isolated using DNA/RNA mini Kit (Qiangen Cat 80204) by following the RNA isolation procedures. Sequencing was performed by the University of Michigan (UM) DNA Sequencing Core, using the Illumina Hi-Seq 4000 platform, paired-end, 50 cycles and Ribosomal Reduction library prep.

We obtained read files from the UM Sequencing Core and concatenated those into a single FASTQ file for each sample. We checked the quality of the raw read data for each sample using FastQC (version 0.11.3) to identify features of the data that may indicate quality problems (e.g., low quality scores, overrepresented sequences, inappropriate GC content). We used the Tuxedo

Suite software package for alignment, differential expression analysis, and post-analysis diagnostics. Briefly, we aligned reads to the reference genome (GRCm38) using TopHat (version 2.0.13) and Bowtie2 (version 2.2.1). We used default parameter settings for alignment, with the exception of: "–b2-very-sensitive" telling the software to spend extra time searching for valid alignments. We used FastQC for a second round of quality control (post-alignment), to ensure that only high-quality data would be input to expression quantitation and differential expression analysis. We used Cufflinks/CuffDiff (version 2.2.1) for expression quantitation, normalization, and differential expression analysis, using GRCm38.fa as the reference genome sequence. For this analysis, we used parameter settings: "–multi-read-correct" to adjust expression calculations for reads that map in more than one locus, as well as "–compatible-hits-norm" and "–upper-quartile–norm" for normalization of expression values. We generated diagnostic plots using the CummeRbund R package. We used locally developed scripts to format and annotate the differential expression data output from CuffDiff. Briefly, we identified genes and transcripts as being differentially expressed based on FDR $\leq$ 0.05, and fold change $\geq$ ± 1.5. We annotated genes with NCBI Entrez GeneIDs and text descriptions. RNA-seq reads in bam format were mapped to the most recent mouse genome (mm10) using the Integrative Genomics Viewer (IGV). The RNA-seq data reported here can be found in the Gene Expression Omnibus (GEO) database with the series accession ID GSE134975.

4.5.10 Generation of Hi-C libraries for sequencing

The *in situ* Hi-C protocols from Rao *et al.* [274] were adapted with slight modifications. For each Hi-C library, approximately 1 x 10^6 cells were incubated in 250 μl of ice-cold Hi-C lysis buffer (10mM Tris-HCl pH8.0, 10mM NaCl, 0.2% Igepal CA630) with 50μl of protease inhibitors (Sigma) on ice for 30 minutes and washed with 250 μl lysis buffer. The nuclei were pelleted by centrifugation at 2500xg for 5 minutes at 4°C, re-suspended in 50 μl of 0.5% sodium dodecyl sulfate (SDS) and incubated at 62°C for 10 minutes. Afterwards 145 μl of water and 25 μl of 10% Triton X-100 (Sigma) were added and incubated at 37°C for 15 minutes.

Chromatin was digested with 200 units of restriction enzyme MboI (NEB) overnight at 37°C with rotation. Chromatin end overhangs were filled in and marked with biotin-14-dATP (Thermo Fisher Scientific) by adding the following components to the reaction: 37.5 μl of 0.4mM biotin-14-dATP (Life Technologies), 1.5 μl of 10mM dCTP, 1.5 μl of 10mM dGTP, 1.5 μl of 10mM dTTP, and 8 μl of 5U/μl DNA Polymerase I, Large (Klenow) Fragment (NEB). The marked chromatin ends were ligated by adding 900 μl of ligation master mix consisting of 663 μl of water, 120 μl of 10X NEB T4 DNA ligase buffer (NEB), 100 μl of 10% Triton X-100, 12 μl of 10mg/ml BSA, 5 μl of 400 U/μl T4 DNA Ligase (NEB), and incubated at room temperature for 4 hours.

Chromatin de-crosslinking was performed by adding 50 μl of 20mg/ml proteinase K (NEB)

and 120 μl of 10% SDS and incubated at 55°C for 30 minutes, adding 130 μl of 5M sodium chloride and incubate at 68°C overnight. DNA was precipitated with ethanol, washed with 70% ethanol, and dissolved in 105 μl of 10 mM Tris-HCl, pH 8. DNA was sheared on a Covaris S2 sonicator. Biotinylated DNA fragments were pulled with the MyOne Streptavidin C1 beads (Life Technologies). To repair the ends of sheared DNA and remove biotin from unligated ends, DNA-bound beads were re-suspended in 100 μl of mix containing 82 μl of 1X NEB T4 DNA ligase buffer with 10mM ATP (NEB), 10 μl of 10 (2.5mM each) 25mM dNTP mix, 5 μl of 10U/μl NEB T4 PNK (NEB), 4 μl of 3U/μl NEB T4 DNA polymerase (NEB), and 1 μl of 5U/μl NEB DNA polymerase I, Large (Klenow) Fragment (NEB).

After end-repair, dATP attachment was carried out in 100 μl of reaction solution, consisting of 90 μl of 1X NEBuffer 2, 5 μl of 10mM dATP, and 5 μl of 5U/μl NEB Klenow exo minus (NEB). The reaction was incubated at 37°C for 30 minutes. The beads were then cleaned for Illumina sequencing adaptor ligation which was done in a mix containing 50 μl of 1X T4 ligase buffer, 3 μl T4 DNA ligases (NEB), and 2 μl of a 15 μM Illumina indexed adapter at room temperature for 1 hour. DNA was dissociated from the bead by heating at 98°C for 10 min, separated on a magnet, and transferred to a clean tube.

Final amplification of the library was carried out in multiple PCRs using Illumina PCR primers. The reactions were performed in 25 μl scale consisting of 25 ng of DNA, 2 μl of 2.5mM dNTPs, 0.35 μl of 10 μM each primer, 2.5 μl of 10X PfuUltra buffer, PfuUltra II Fusion DNA polymerase (Agilent). The PCR cycle conditions were set to 98°C for 30 seconds as the denaturing step, followed by 14 cycles of 98°C 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds, then with an extension step at 72°C for 7 minutes.

After PCR amplification, the products from the same library were pooled and fragments ranging in size from 300-500 bp were selected with AMPure XP beads (Beckman Coulter). The size-selected libraries were sequenced to produce paired-end Hi-C reads on the Illumina HiSeq 2500 platform with the V4 of 125 cycles.

4.5.11 Hi-C data processing

We generated the Hi-C matrices using Juicer [103]. Juicer uses BWA-mem to align each paired-end read separately. It then determines which reads can be mapped uniquely and keeps unambiguously mapped read pairs. Each read is assigned to a "fragment", determined by the restriction enzyme cut sites and paired reads that map to the same fragment are removed. The Juicer pipeline creates a binary data file (namely, the ".hic" file), which contains Hi-C contacts. The .hic file is imported into MATLAB-compatible variables via our in-house MATLAB toolbox, 4DNvestigator [193]. Centromeric and telomeric regions were removed in the process. Then, Hi-C data were

subsequently normalized and binned at 50kb and 100kb resolution. ICE [145] and "observed over expected" (O/E) normalized contact maps at 100kb resolution were used for chromosome-level analysis of A/B compartments. Observed contact maps at 50kb resolution were used for detection and analysis of TADs. Hi-C data for this study are available through the NCBI BioProject database (accession number: PRJNA608895).

4.5.12 Integration of Hi-C and RNA-seq data

To integrate our Hi-C and RNA-seq datasets, we first binned measurements at the same resolution. Accordingly, each element in the binned RNA-seq expression vector corresponds to a row (or column) of the Hi-C contact matrix that captures the same set of genomic loci. Genomic bins spanning a given region contain measurements reflecting the sum of its parts, where expression values for genes sharing the same bin are summed and contact frequencies for loci occupying the same bin are summed. This allows us to directly compare the transcriptional and conformational behavior across regions of the genome.

4.5.13 Frobenius Norm

The Frobenius norm describes the size or volume of a matrix [326]. As such, the Frobenius norm can be used as a measure of variance or variability in data [102]. For matrix A of dimension m, that is $\mathrm{A} \in \mathrm{R}^{\mathrm{m}\times\mathrm{n}}$, the mathematical definition of the Frobenius norm is as follows:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a^2{}_{ij}}, \tag{4.1}$$

where a_{ij} is an element of the data matrix A. Equivalently this expression can be written as,

$$\|A\|^2{}_F = trace(A^T A), \tag{4.2}$$

where $trace$ is the sum of the diagonal elements of a matrix [105]. When comparing two sets of data, the notation can is modified as follows:

$$\|A - B\|^2{}_F = trace((A - B)^T (A - B)), \tag{4.3}$$

where A and B are matrices of the same dimensions. The Frobenius norm of a vector is equivalent to the Euclidean norm.

4.5.14 Larntz-Perlman Procedure

The Larntz-Perlman (LP) procedure is a method designed to test the equivalence of correlation matrices [179]. We applied the LP procedure as described in the 4DNvestigator Toolbox [193] to compare Hi-C contact matrices between WT and *Wapl* KO conditions. Briefly, correlation matrices are derived from data by taking the pairwise linear correlation coefficient between each pair of columns in the Hi-C contact matrices. Then, a null hypothesis is defined from corresponding population correlation matrices. We then compute a Fisher z-transformation on the correlation matrices and calculate a test statistic for the chi-squared distribution to determine whether or not the null hypothesis (that Hi-C matrices are equivalent) should be rejected with a corresponding p-value.

4.5.15 A/B Compartmentalization

Chromatin can either take on a condensed heterochromatic form or a looser euchromatic form. We use the Fiedler vector – the eigenvector corresponding to the second smallest eigenvalue (Fiedler number) of the normalized Laplacian matrix – to describe this feature mathematically which allows us to bi-partition the data into signed compartments – "A" (positive values) corresponding to euchromatin and "B" (negative values) corresponding to heterochromatin. The Fiedler vector and first principal component (PC1) vector are mathematically equivalent. However, unlike the PC1 vector, the Fiedler vector is accompanied by its corresponding Fiedler number which allows us to assign a magnitude of connectivity to a network. We characterize compartment switch events by identifying genomics bins with differing sign and magnitude above the 90th percentile of vector values between settings.

4.5.16 Identification of TADs

Topologically associating domains (TADs) were identified using spectral clustering as described in chen *et al.* [63]. This technique calculates the Fiedler vector for a normalized Hi-C adjacency matrix and initially organizes neighboring regions whose Fiedler vector values have the same sign into shared domains. This initial TAD structure is repartitioned if for a given domain, the Fiedler number is less than a user-defined threshold (λ_{thr}) to ensure that the TADs represent well-connected regions and are not too large. Resulting TADs are iteratively repartitioned until their respective Fiedler numbers are larger than λ_{thr} or until the smallest allowable TAD size (default is 300kb) is reached. We perform this procedure for chromosome-level Hi-C contact maps pooled across samples at 50kb resolution. For our analysis, thr was chosen to ensure a median TAD size of 900kb, as the expected median TAD size in mammalian genomes is 880kb (rounded to 900kb

for our data since bins are in intervals of 50kb) [94, 32]. A specific λ_{thr} was chosen for each chromosome and sample to ensure each TAD clustering set would have the same approximate median TAD size.

4.5.17 Quantifying internal TAD organization

The internal structure of TADs is comprised of transient interactions over varying lengths that form chromatin 'loops'. In the absence of WAPL, loop length increases, enriching contact frequency between either boundary of a given TAD [129]. Visually, we identify these interactions by the notable density of contacts present in the top right or bottom left corners of TADs. We quantify the strength of these interactions by designating a local neighborhood with a window size of 150kb by 150kb centered around the corner of each TAD. We then find the nonzero average of observed contacts in each window. TADs that are less than or equal to the window size are ignored (corner peak signal set to 0). This is performed for every TAD across each setting and chromosome. To evaluate the statistical significance of a given change in corner peak signal between WT and KO cells, we compared each corner peak signal to a background set of interactions. This background set is made up of contacts in the region between the diagonals encompassing the corner peak neighborhoods. We evaluate the p-value of each corner peak signal based on the proportion of background interactions with a signal less than the corner peak signal. As such, we determined that if a corner peak has a higher signal in the KO cells compared to WT cells and is significantly higher than interactions at comparable distances ($p < 0.01$), then the corner peak enrichment is statistically significant.

4.5.18 Hi-C derived 5C contact map generation

We constructed a synthetic 5C contact map derived from a genome-wide 1Mb Hi-C contact map for genomic regions containing genes in the cell cycle gene networks. This was done by locating the genomic bins corresponding to cell cycle genes, extracting the inter-chromosomal and cell cycle loci-specific intra-chromosomal interaction frequencies for those bins, and stitching them together in genomic order. Our working set of cell cycle genes was sourced from the KEGG database and included 170 genes distributed across all chromosomes in the genome. Some of the genes occupy the same 1Mb genomic bins resulting in a 141 by 141 dimension adjacency matrix at 1Mb-resolution.

4.5.19 Cell proliferation assay

CellTrace™Far-Red was utilized for our cell proliferation assay. Purified T cells were labeled with far-red at a final concentration of 1 μmol/L according to the manufacturer's instruction (Molecular Probes, C34564). Far-red-labeled T cells were cultured similar to MLR for allogeneic reaction or Dynabeads T cell activator CD3/CD28 stimulation for up to 7 days. Then, collected cells were examined with flow cytometry, gated for CD3 (APC) positive, and Far-red dilution was determined.

4.5.20 FxCycle™Far-Red Stain for DNA content measurement

For *in vitro* experiments, WT or *Wapl* KO T cells were treated with Dynabeads T cell activator CD3/CD28 stimulation for up to 4 days. For *in vivo* experiments, transferred WT and *Wapl* KO T cells were isolated and purified from spleens of BALB/c recipient mice on day 7 after allogeneic BMT. The T cells were fixed with 10% formaldehyde for 30 mins, washed 3 times with PBS and the sample cell concentration was adjusted at 1 × 106 cells/mL. FxCycle™Far-red stain (200 nM) (Molecular Probes, F10348) and 5 μL of RNase A (20 mg/mL) (Roche Cat. 70294823) were added to flow cytometry samples and continued for incubation at room temperature for 30 minutes and protected from light. Samples were analyzed with flow cytometer (AttuneNxT) without washing using 633nm excitation and emission collected in a 660 bandpass. The DNA contents were determined as 2C and >2C by fluorescence intensities.

4.5.21 ELISA

Concentrations of TNF-α and IFN-γ in sera on day 21 after allogeneic BMT were measured with specific anti-mouse ELISA kits from BD Biosciences. Assays were performed per the manufacturer's protocol and read at 450nm in a microplate reader (Bio-Rad). The concentrations were calculated from triplicate samples as mean ± SEM.

4.5.22 FACS

Single-cell suspensions of spleens and thymii were prepared as previously described [332]. Briefly, to analyze surface phenotype, purified T cells and thymocytes from B6 WT, *Wapl* KO deficiency mice, or transplanted animals, were washed with FACS wash buffer (2% bovine serum albumin [BSA] in phosphate-buffered saline [PBS]), pre-incubated with the rat anti-mouse FcR mAb 2.4G2 for 15 minutes at 4°C to block nonspecific FcR binding of labeled antibodies, then resuspended in FACS wash buffer and stained with conjugated monoclonal antibodies purchased from BD Biosciences (San Jose, CA): allophycocyanin (APC)-conjugated monoclonal antibodies

(MoAbs) to CD4, CD8, CD3, CD45.2, CD45.1, CD25 and CD69; phycoerythrin (PE)-conjugated MoAbs to CD3, CD4, CD8, CD25, and TCRβ; allophycocyanin (APC)-conjugated MoAbs to CD3, CD4, and PerCP/Cy5.5-conjugated MoAbs to CD3, CD4 and CD8 were purchased from eBioscience (SanDiego, CA). Next, cells were analyzed using an AttuneNxT flow cytometer. For intra-cellular staining, cells were stained for CD4 and CD25 antibodies as above, then fixed with IC Fixation Buffer (Biolegend, Cat. No.420801), incubated 20-60 minutes at room temperature, followed by continued addition of 2 mL of 1X Permeabilization Buffer (Biolegend Cat. No. 421002) and centrifugation at 400-600 x g for 5 minutes at room temperature. Cell pellets were resuspended in 100 μL of 1X permeabilization buffer and PE-conjugated FoxP3 antibody (eBioscience, SanDiego, CA, Cat. 126403) was added at 0.5 μg/million cells/100 μl and incubated for 30 minutes at room temperature. Stained cells were resuspended in an appropriate volume of Flow Cytometry Staining Buffer for flow cytometry analyses. Apoptotic cells were detected by PE-Annexin V staining.

CHAPTER 5

Obesity Disrupts Innate-Adaptive Immune Network Patterning in Adipose Tissue

This chapter is based on a paper by Gabrielle A. Dotson, Indika Rajapakse, and Lindsey A. Muir [100] (under review).

5.1 Abstract

Obesity drives significant changes in adipose tissue that precede development of tissue and systemic insulin resistance. Immune cell infiltration and inflammation are known contributors to these changes but there is limited understanding of their spatial context tissue-wide. We sought to identify spatial patterning in epididymal adipose tissue immune cells in a time course of diet-induced obesity in mice. Using spatial transcriptomics and single-cell RNA-sequencing, we identified dominant cell type signatures preserved in their anatomical context, quantified gene expression patterns at spots throughout adipose tissue, performed cell type network analysis, and investigated ligand-receptor colocalization. Our data support increased innate immune cells, including macrophages, monocytes, and innate lymphoid cells with tissue-wide interspersion, and dampened adaptive immune cell signatures with obesity. Network analysis identified increased heterogeneity in all major immune cell types, consistent with increased subtypes. To capture tissue dynamics at obesity onset, we draw on mathematical principles from linear algebra and spectral graph theory and provide a framework for better understanding cell cooperation toward emergence of multicellular tissue function. The culmination of these analyses revealed a widespread paradigm shift in ligand-receptor activity with near-exclusive macropahge-macropahge or monocyte-macropahge interactions at crown-like structures across adipose tissue during obesity.

5.2 Introduction

The global increase in obesity raises significant concern about the development of cardiovascular and metabolic disease, increased risk of adverse events in viral infection, and economic and health care costs [221, 81]. Changes in white adipose tissue (WAT) immune cell types and polarization occur in obesity and are associated with metabolic dysfunction in mice and humans [204, 203, 231, 230].

Population and single-cell transcriptomics enable low-bias characterization of these obesity-related changes in WAT immune cells [154, 365, 230]. However, the molecular events triggering tissue dysfunction remain poorly understood in their anatomical context, which is lost with tissue digestion. Tissue immunostaining captures anatomical context but is limited to a small number of concurrent markers that may not fully distinguish cells with similar surface proteins. In adipose tissue, macrophages (ATMs) and dendritic cells (ATDCs) have been challenging to distinguish, yet have unique participation in tissue homeostasis and development of insulin resistance [294, 69, 261]. Thus low-bias capture of immune cells and subtypes within a tissue landscape will be critical for identifying functional distinctions relevant to disease.

Spatial transcriptomics (ST) captures signatures of gene expression across tissue sections, which in studies of amyotrophic lateral sclerosis (ALS) led to discovery of new anatomical subregions and spatiotemporal patterns underlying tissue dysfunction [211, 11, 66, 369, 86]. ST was recently used to profile different adipocyte subtypes with distinct responses to insulin stimulation in human adipose tissue [15]. Here, we mapped murine adipose tissue and immune cell populations in obesity using a combination of spatial transcriptomics and single cell RNA-sequencing. Over a time course of diet-induced obesity and development of the pre-diabetic state, one of our primary goals was to capture dynamics of early immune cell infiltration as tissue dysfunction appears. We present the tissue landscape as a network of cells and interrogate spatial patterning and cell-cell communication in early obesity. Our analyses capture early monocyte infiltration, progressive dominance of monocyte and macrophage crosstalk, and dampening of adaptive immune cell signatures. Furthermore, ligand-receptor analyses show putative monocyte-macrophage interactions in pre-crown-like anatomical neighborhoods by 8 weeks of high-fat diet feeding, implicating monocyte signaling in development of crown-like structures in a role distinct from their differentiation into ATMs.

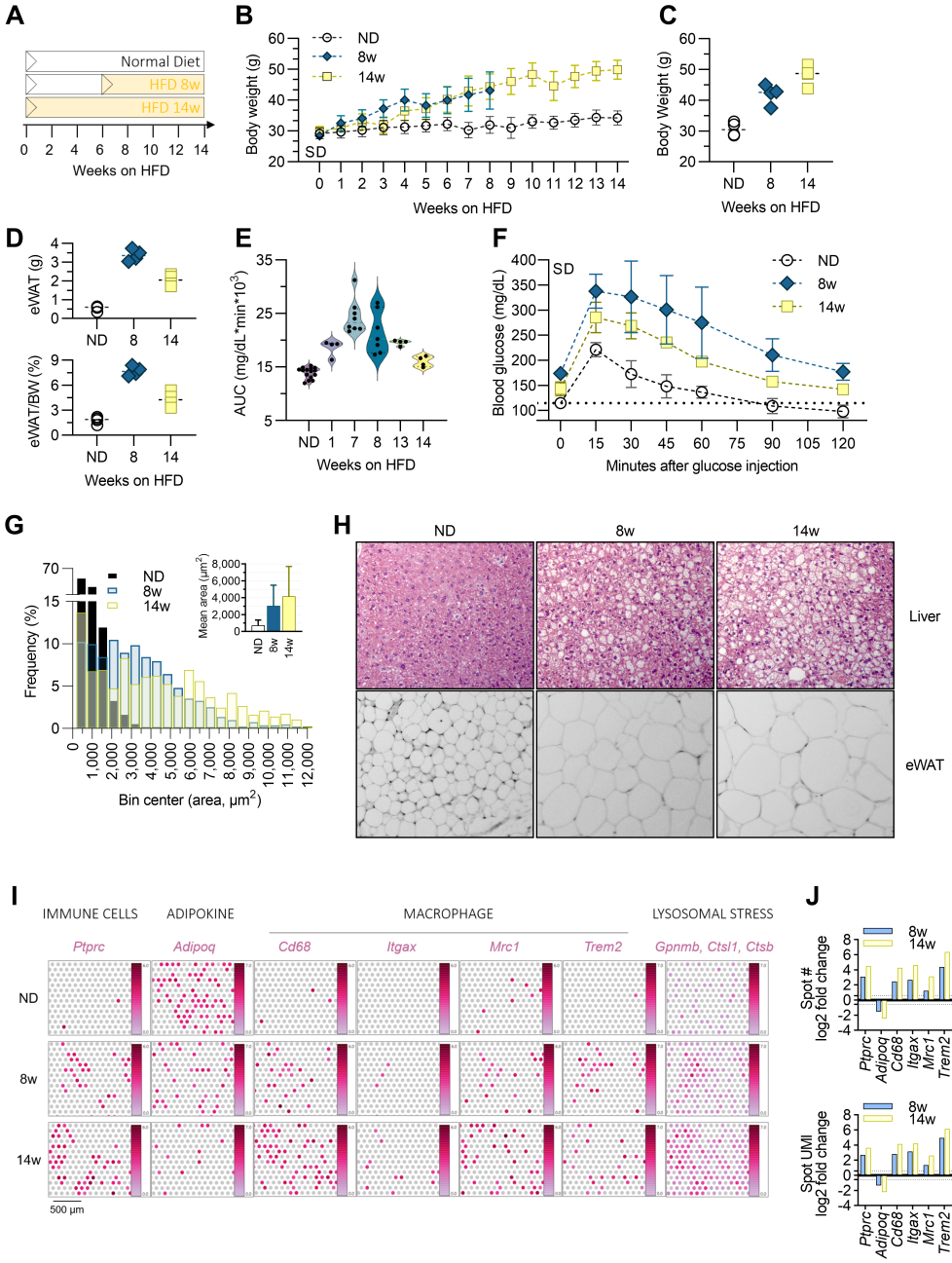

Figure 5.1: Diet-Induced Obesity and Adipose Tissue Remodeling. (A) Time course for mice fed a 60% high-fat diet (HFD) for 8 weeks (8w) or 14 weeks (14w), versus normal diet (ND) controls. Measurements included weight gain (B), final body weight (C), epididymal adipose tissue (eWAT) weight and eWAT as a percentage of body weight (D), and glucose tolerance test data showing area under the curve (AUC) across time points (E) and glucose measurements for cohorts one week prior to endpoint tissue collection (F). (G) Frequency distribution and average adipocyte size in eWAT of ND, 8w, and 14w cohorts. (H) Representative images from H&E staining of eWAT and liver. (I) Representative maps showing spatially preserved gene expression in eWAT. Each spot is 55 μm diameter location of transcript capture from a tissue section, colored corresponding to the expression level ($\log_2$-transformed) of select obesity-related genes. (J) Gene expression changes

Figure 5.1: in HFD-fed mice compared to ND mice corresponding to the genes shown in part (I). Upper graph, $\log_2$(fold change) in the quantity of spots in HFD cohorts; Lower graph, $\log_2$(fold change) in average tissue-wide expression of a gene in HFD cohorts.

5.3 Results

5.3.1 Spatial Analysis of Adipose Tissue Across Early Obesity

We evaluated tissue and metabolic function in our diet-induced obesity model using mice fed a normal diet (ND), a high-fat diet for 8 weeks (8w), and a high-fat diet for 14 weeks (14w) (Figure 5.1A). As expected, mice fed a high-fat diet (HFD) gained body and epididymal white adipose tissue (eWAT) weight (Figure 5.1B-D). Glucose tolerance tests (GTT) showed increased area under the curve (AUC) starting at one week of HFD feeding, with the highest AUC and variability at intermediate time points (Figure 5.1E,F). Adipocyte sizing showed increased frequency of large adipocytes and greater mean adipocyte size in the 8w and 14w cohorts (Figure 5.1G,H). These data are consistent with pre-diabetes of this HFD-fed model, showing early insulin resistance during HFD feeding and a period of variable insulin resistance that aligns with dynamic restructuring of adipose tissue.

Changes in SVCs in murine diet-induced obesity include immune cell infiltration and activation that progressively disrupts homeostatic mechanisms [226, 231, 261, 260]. To capture this changing tissue landscape, we performed spatiotemporal molecular profiling, which can uncover key patterns in breakdown of tissue function [11, 66]. Here, we spatially profiled genome-wide transcripts across eWAT sections from ND and HFD-fed mice using the Visium (10X Genomics) spatial transcriptomics (ST) platform [323, 15]. We initially evaluated expected immune cell and obesity-related gene expression, and found increased pan-immune cell *Ptprc* (*CD45*), myeloid-associated *CD68*, recruited ATM marker *Itgax* (*CD11c*), resident ATM marker *Mrc1*, and *Trem2*, which is expressed in lipid-associated ATMs and has proposed functions in protective lipid homeostasis [154, 195, 309] (Figure 5.1I,J). *Adipoq* progressively decreased as expected in the HFD-fed conditions.

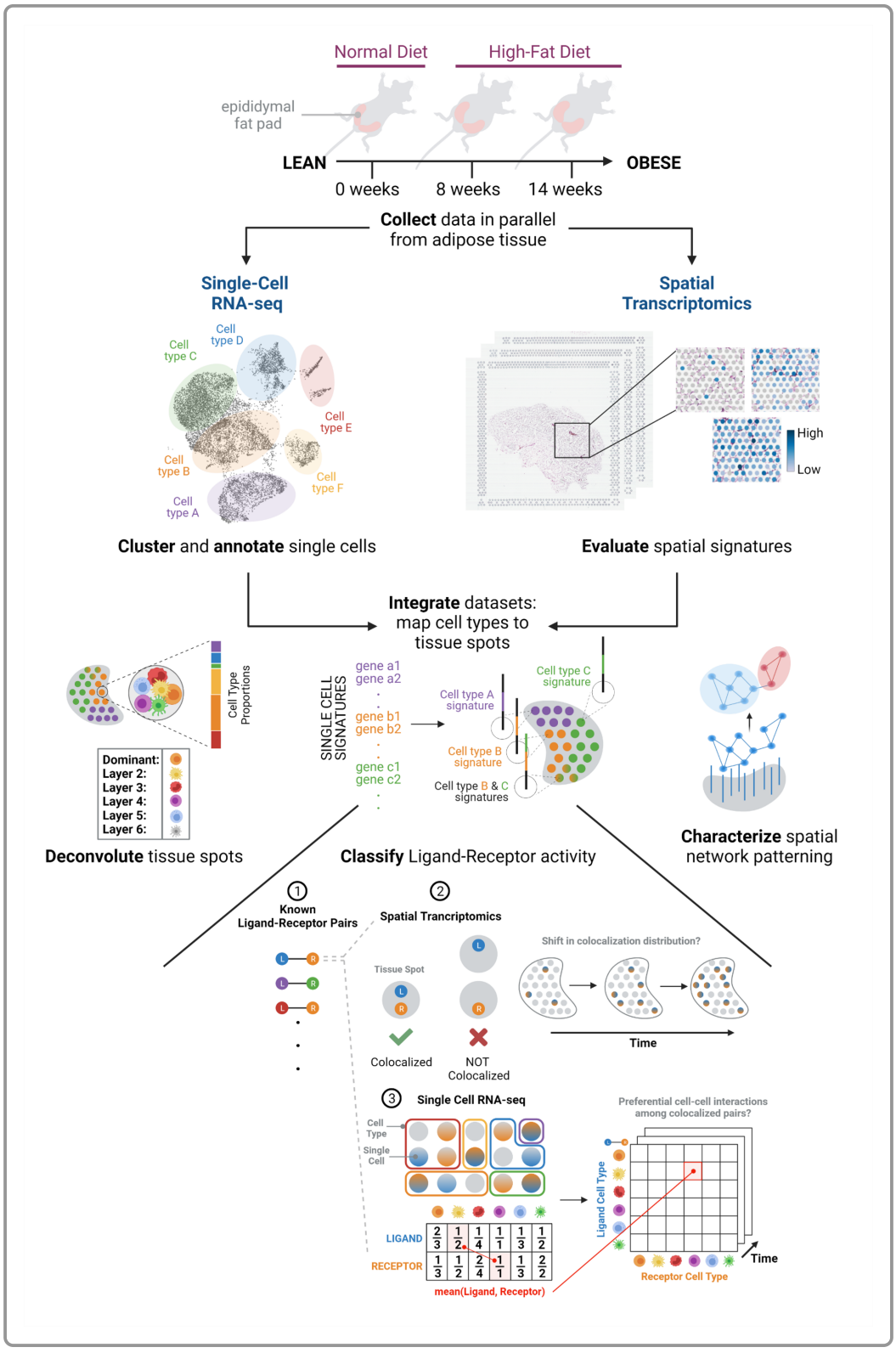

Figure 5.2: Workflow of Spatial Transcriptomics (ST) and Single-Cell RNA-Sequencing (scRNA-seq) Data Analyses. Epididymal white adipose tissue was collected from mice in a time course of high fat diet feeding. scRNA-seq data were obtained from within-cohort pooled samples, and ST data were collected from a section of fresh frozen adipose tissue from each diet condition.

Figure 5.2: scRNA-seq data were then pooled across diet conditions, clustered (Algorithm 2), and cell populations were annotated using a data-driven approach (see Methods). Existing literature and databases were used to query highly-scored markers and call cell types for each cluster. ST data were evaluated for distribution of obesity marker genes. scRNA-seq and ST datasets were integrated by mapping cell types to the adipose tissue sections, performed by finding the overlap between identified cell type signatures (scRNA-seq) and tissue spot expression profiles (ST). Given the size of each tissue spot, more than one cell type signature was likely to be found at each spot. Consequently, every cell type was assigned at each tissue spot based on the proportion of its signature expressed at the spot, generating layers of tissue spot assignments with decreasing likelihood, a process we refer to as spot deconvolution (Algorithm 3). Once tissue spots were annotated, patterns in the intra-cell type spatial tissue network were characterized. Finally, ligand-receptor signaling was classified and colocalization and cell-cell interaction trends characterized (Algorithm 4).

5.3.2 Breakdown of Adipose Tissue Immune Cell Networks in Early Obesity

Multiple nuclei can contribute transcripts to each spatial data spot, therefore in parallel we performed single-cell RNA-sequencing (scRNA-seq) in SVCs to clarify changes in immune cell types and facilitate signature identification in spatial data. SVCs were isolated from fat pads contralateral to the fat pads used for spatial data and enriched for $CD45^+$ immune cells. We identified cell types in scRNA-seq data using spectral clustering and analysis of signature genes for each cluster, considering expression and uniqueness (Algorithm 2, Table S1 in Dotson *et al.* [100], Methods and Materials). Cell type signatures were further verified by alignment with ImmGen profiles (Figure S1 in Dotson *et al.* [100]). To identify cell types present at spots, we integrated scRNA-seq and ST datasets, determining overlap between a cell type signature and tissue spot gene expression (Figure 5.2, Algorithm 3, Methods and Materials).

Spectral clustering and data-guided cell type identification yielded six broad immune cell types: monocytes, T cells, ATMs, ATDCs, natural killer cells (NKCs), and B cells (Figure 5.3A, Figure S1 in Dotson *et al.* [100]). Changes in the cellular composition of adipose tissue in lean versus obese mice were largely driven by T cells, B cells, and ATMs, where the proportion of single cells classified as T cell and B cell decreased with obesity (17% to 5% in T cells and 22% to 3% in B cells), while the proportion classified as ATMs sharply increased (43% to 76%). These findings are consistent with prior scRNA-seq studies of immune cell populations in diet-controlled mouse models [154, 377], where the same panel of immune cells has been observed and where monocytes and ATMs have been the predominant populations present.

We next established cell type localization patterns in lean and obese adipose tissue. Cell types were assigned to tissue spots hierarchically based on the proportion of their signature expressed at the spot. In this way, the dominant cell type at a spot is the one with the highest number of its

representative genes expressed. Dominant cell types across all spots make up the dominant layer, or Layer 1. Layer 2 is assigned as the cell type with the second highest number of signature genes expressed, and the remaining cell types are assigned subsequent layers in the same manner. The outcome of this assignment is a series of layers representing a decreasing likelihood of a given cell type's presence at a particular spot (Figure S2 in Dotson *et al.* [100]). We refer to this process as spot deconvolution and the resulting tissue assignment layers as 'deconvoluted layers' (Algorithm 3, Figure 5.2, Methods and Materials). Cell type localization in the dominant layer over time was consistent with trends in cellular composition from the scRNA-seq data, suggesting decreasing T cells and B cells and increasing ATMs in the HFD settings (Figure 5.3B).

5.3.3 Spatial Network Patterning Highlights Activation of Innate and Dampening of Adaptive Immune Cell Signatures

Inflammation is a key factor in development of adipose tissue dysfunction and insulin resistance in obesity. Cells of the adaptive and innate immune system both contribute to inflammatory processes in adipose tissue, including T cells, B cells, ATMs, ATDCs, and innate lymphoid cells (ILCs) [217]. This network of cells, and their subtypes, modulate inflammation through cytokine and chemokine secretion and cell-cell interactions, reducing anti-inflammatory signaling and promoting proinflammatory cell types and processes over time. To broadly compare patterns of adaptive and innate immune cell networks in obesity, each tissue spot in the spatial data was classed as adaptive (T cell, B cell), innate (monocyte, NKCs, ATM, or ATDC), or mixed type based on the first two tissue assignment layers from Algorithm 3. We found that spots containing only adaptive immune cells consistently decreased with obesity, from 21% to 6% (3.5 fold change), while spots containing only innate immune cells increased with obesity, from 1% to 9% (9 fold change) by 14w (Figure 5.3C), consistent with known shifts in cellularity with HFD feeding [231]. Among total spots, 75%-85% had mixed type signatures, with the higher frequencies found in obesity. Overall, these data suggest a stronger influence of adipose tissue innate immune cells than adaptive immune cells with obesity, through increased quantity and frequency, strong spatial interspersion, and increased potential for cross-talk with the adaptive immune system (Figure 5.3C).

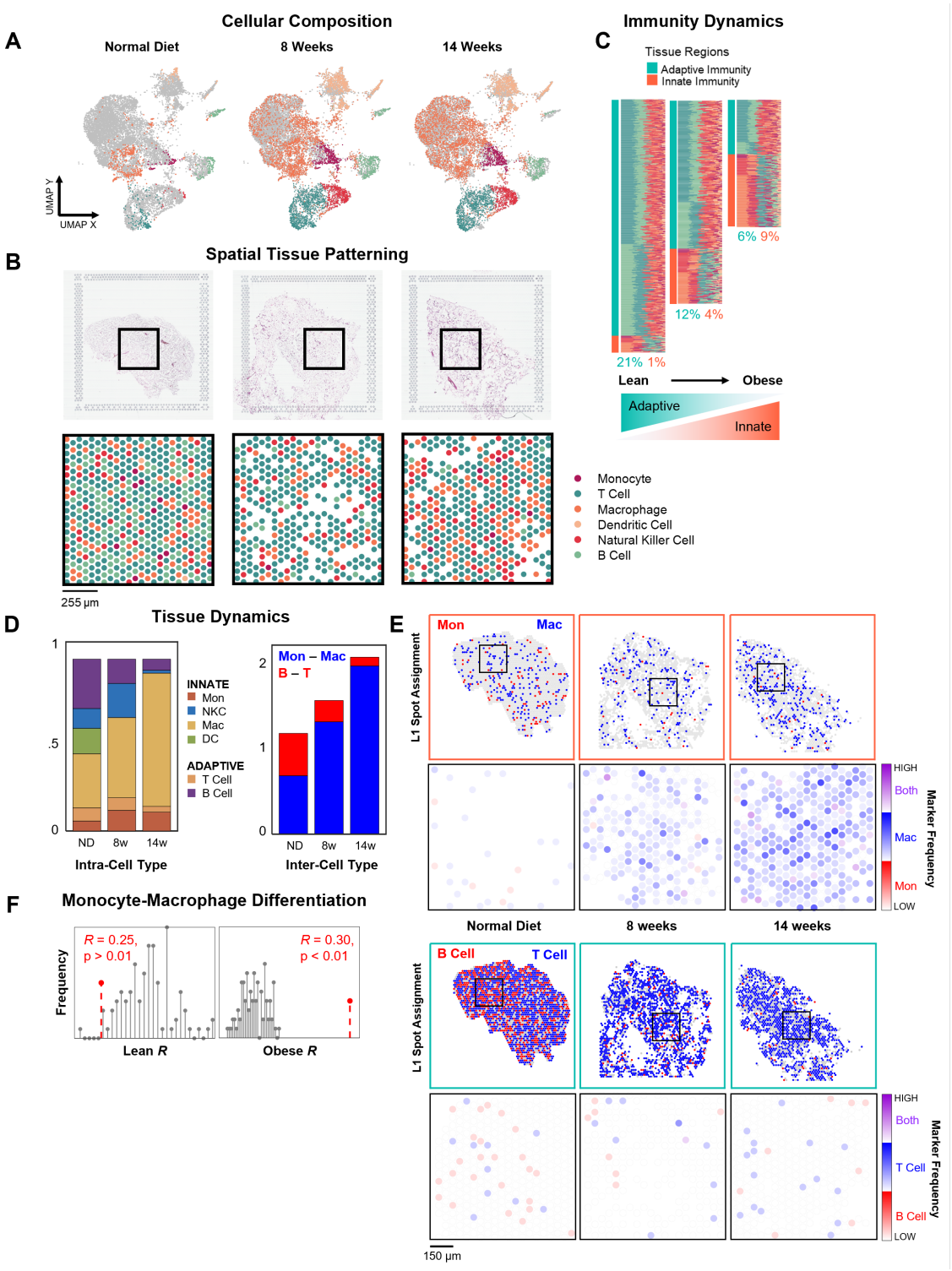

Figure 5.3: Immune Cell Type Identification and Localization in Adipose Tissue. (A) Cellular composition. UMAP projections of scRNA-seq data over time, stratified by cell type. Gray points represent the UMAP projection from pooled clustering of all three samples and colored points represent sample-specific single cells. (B) Spatial tissue patterning. (Top) H&E staining of adipose tissue sections. (Bottom) Zoomed view of annotated ST-derived tissue landscapes. Cell type assignments represent the dominant spot assignments after performing spot deconvolution (Algorithm 3).

Figure 5.3: White spaces with no spots indicate non-viable areas of the tissue where there were visible holes and therefore no detectable mRNA levels. (C) Immunity dynamics. Stacked bar plots representing cell type proportions at tissue spots assigned to adaptive or innate immune cell types at each timepoint (denoted by orange and teal vertical bars to the left of the plots). Tissue spots were classified as contributing to adaptive immunity if adaptive cell types colocalized at the spot, meaning deconvoluted tissue assignment layers 1 and 2 (L1 and L2) were assigned as T cell or B cell for those spots. Tissue spots were classified as contributing to innate immunity if innate cell types colocalized at the spot, meaning L1 and L2 were assigned as ATM, monocyte, NKC, or ATDC. Remaining tissue spots were classified as ambiguous tissue regions where L1 was assigned an adaptive cell type and L2 an innate cell type, or vice versa, and are not shown in the figure. Each stacked bar represents a single tissue spot. The height of each component in a stacked bar reflects cell type proportions (the number of genes expressed out of each 50-gene cell type-specific signature). Below plots is a model of the observed adaptive and innate immunity trade-off during obesity progression. We propose that obesity weakens the adaptive immune system and innate immunity concurrently compensates for that weakened immune response over time. (D) Tissue dynamics. (Left) Individual cell type contributions to tissue function (normalized to sum to 1). Intra-cell type interactions are quantified by computing the tradeoff between cell and genome connectivity for each cell type. (Right) Prominent inter-cell type relationships. Inter-cell type interactions are quantified by dividing the tradeoff between cell and genome connectivity for two cell types by the the tradeoff between cell and genome connectivity for all cell types (Supplementary Figure 3). (E) Inter-cell type spatial distribution. (Boxed in orange) Monocyte (red) and macrophage (blue) spot assignments in the dominant tissue assignment layer. Directly below each plot is a sample region of the tissue showing where monocyte- and macrophage-specific markers localize. A high color intensity indicates a high proportion of monocyte or macrophage-specific markers at the spot. Purple colored regions indicate spots where both monocyte and macrophage markers were found and represent inter-cell type interaction. (Boxed in teal) B cell (red) and T cell (blue) spot assignments in the dominant tissue assignment layer. Directly below each plot is a sample region of the tissue showing where B cell and T cell-specific markers localize. A high color intensity indicates a high proportion of B cell or T cell-specific markers at the spot. Purple colored regions indicate spots where both B cell and T cell markers were found and represent inter-cell type interaction. (F) Monocyte-Macrophage differentiation. Stem plot of Spearman correlation (R) between monocyte and macrophage single-cell gene expression profiles in lean versus obese mice. Red stems are the real observed Spearman correlation coefficients in normal diet (left) and high-fat diet (right) mice. Gray stems are the background distribution of Spearman correlation coefficients between randomly-selected single-cell gene expression profiles.

5.3.4 Turing-Inspired Analysis Reveals Increased Interconnectivity Among Monocytes and Macrophages

To further assess how immune cells contribute to shifting multicellular tissue function, we described intra- and inter-cell type relationships at the cellular level and identified transcriptional diversity within cell type populations at the genome level. For modeling this tissue morphogenesis,

we introduce the notion of the Turing system. Briefly, the Turing system considers the coordinated effect that between-cell and within-cell dynamics have on emerging tissue function (Supplementary Notes)[270]. Here, we characterized between-cell dynamics in terms of correlation between position-specific gene expression profiles weighted by Euclidean distance from ST data and within-cell dynamics in terms of correlation between single-cell gene expression profiles from scRNA-seq data (Methods and Materials). We then captured emerging tissue function in terms of individual cell type contributions, by computing the trade-off between cell and genome connectivity for each immune cell type, derived from between-cell and within-cell dynamics, respectively (Figure S3 in Dotson *et al.* [100], Methods and Materials). Our results were consistent with increased contributions from macrophage and monocytes to the changing adipose tissue landscape in obesity (Figure 5.3D). The individual contributions of T cells and B cells to emerging tissue function decreased with obesity, consistent with localization trends described in Figure 5.3B.

To determine the contribution of pairs of cell types to tissue function in a potentially coordinated manner – emergence of *multicellular* tissue function – we normalized the emergence of local tissue function (cell and genome connectivity trade-off derived from merged data across *two* cell types) by the emergence of global tissue function (cell and genome connectivity trade-off derived from merged data across *all* cell types). While the majority of cell type inter-connectivities demonstrated decreased or stable coupling with obesity (Figure S4 in Dotson *et al.* [100]), the inter-cell type contributions between monocyte-macrophage and monocyte-dendritic cell increased in obesity (Figure 5.3D). In addition, macrophage inter-connectivities were high overall, particularly with other innate immune cell types (Figure S4 in Dotson *et al.* [100]), while the B-T cell axis had one of the lowest inter-cell type connectivities, decreasing over time (Figure 5.3D).

The increased monocyte-macrophage and monocyte-dendritic cell inter-cell type relationships in HFD feeding may reflect signatures related to monocyte differentiation [392] or crosstalk between these cells. We further explored the distribution of key myeloid cell markers, and found growing colocalization of common monocyte (*Traf1*, *Ccr2*, *Il1b*, *Napsa*, *Plac8*, *S100a9*, *Msrb1*, *Ifitm6*, *Ms4a4c*, *Cd209a*) and macrophage (*Adgre1*, *Ctsk*, *Lyve1*, *Cd209f*, *Mertk*, *Ccl8*, *Trem2*, *Cd63*, *Mmp12*, *Gpnmb*, *Rhoc*, *Ctsd*, *Mfge8*) markers across the tissue (Figure 5.3E), indicating a potential role for undifferentiated monocytes in cooperation with macrophages in obesity. In contrast, little to no overlap between B cell and T cell marker colocalization was found 5.3E), consistent with the low and decreasing B cell-T cell inter-connectivity with obesity.

We next evaluated transcriptional correlation between monocytes and macrophages grew over time as an indicator of a lineage relationship. The correlation between monocyte and macrophage expression profiles increased non-randomly from R = 0.25 to R = 0.30 ($p < 0.01$) (Figure 5.3F). While this supports monocyte to macrophage differentiation, the magnitude of the correlation was low, potentially due to increased diversity of subtypes in these populations. Overall, these data sup-

port the presence of monocyte-macrophage crosstalk as well as ongoing differentiation of monocytes.

5.3.5 Increased Heterogeneity of Signatures Within Immune Cells Captures Phenotype Shifts Across Adipose Tissue

Obesity promotes shifts in immune cell polarization and the appearance of subtypes that affect tissue function. In mice, proportional increases in ATMs relative to tissue weight and patterns of proliferation and apoptosis during weight gain and loss suggest active mechanisms of homeostatic signaling [231] that may eventually be overridden by chronic obesity [396]. We hypothesized that evaluation of immune cell networks in our time course data would provide insight into phenotype shifts that disrupt adipose tissue function.

To better quantify these shifts, we constructed spatial cell networks for each major cell type, where nodes were the tissue spots assigned to the cell type based on the dominant layer and edges between nodes represented a correlation based on transcriptional signature. Surprisingly, cell type network edges decreased universally in obesity, indicating loss of transcriptional correlation among cells of the same type (Figure 5.4A). This was concurrent with increased heterogeneity in expression of cell type-specific signature genes. Loss of spatial patterning and increased heterogeneity support the emergence of immune cell subtypes in obesity that are not present in the lean state. Macrophages in particular show many phenotypic states [329, 389, 294]. Upon clustering macrophage-classified single cells into two sub-clusters, only one of the two clusters was represented in the normal diet setting while both were represented in the two high-fat diet settings (Figure S5 in Dotson *et al.* [100]). The emergent cluster in the HFD settings was significantly enriched in *Cd9* ($\log_2$FC of 11.8), which has previously been observed in lipid-laden macrophages in obesity [138, 154, 287]. In contrast to the observed macrophage subpopulations, T cell, B cell, and NKC sub-clusters were less distinct with fewer differentially expressed genes.

5.3.6 Monocyte and Macrophage Ligand-Receptor Pairs Dominantly Colocalize in Obese Adipose Tissue

To better understand specific interactions that might mediate shifted spatial patterning in obesity, we evaluated colocalized expression of paired ligands and receptors over time. We first captured ligand and receptor expression in adaptive and innate classes, characterizing four interaction states: (1) adaptive ligand and adaptive receptor, (2) innate ligand and innate receptor, (3) adaptive ligand and innate receptor, and (4) innate ligand and adaptive receptor (Materials and Methods). For this study, we determined transcript colocalization of known ligand-receptor pairs as a proxy

for cell-cell communication given the restricted area of each tissue spot.

Interrogating known ligand-receptor (LR) pairs from the literature, we observed an increase in colocalization across the tissue over time, with both the number of colocalized spots and the number of LR pairs at each colocalized spot increasing (Figure 5.4B). From the scRNA-seq data, we found the proportions of adaptive single cells expressing colocalized pairs' ligand and receptor, respectively, and the proportions of innate single cells expressing colocalized pairs' ligand and receptor, respectively (Materials and Methods). This allowed us to infer whether signaling interactions sourced from adaptive, innate, or an interplay of both contributed to the changes we saw between lean and obese adipose tissue. We defined a score for each state by taking the mean proportion between all pairwise combinations of adaptive- or innate-sourced ligands and receptors (Materials and Methods). This score represents the likelihood that a LR pair exhibits signaling activity within (or between) the particular class(es) of cell types. Then, for each colocalized LR pair, we identified which of the four interaction states exhibited the highest score and found that the percentage of ligand-receptor pairs dominantly involved in adaptive ligand-adaptive receptor, adaptive ligand-innate receptor, and innate ligand-adaptive receptor signalling decreased during obesity progression while the percentage of pairs dominantly involved in innate ligand-innate receptor signaling increased (Figure 5.4C).

We performed the same steps to explore the contribution of different cell types to specific ligand-receptor interactions in lean versus obese adipose tissue (Algorithm 4, Figure 5.2). Interestingly, we found that there was a preference for ligands and receptors interacting from different cell types rather than the same cell type, though decreasingly so with obesity (Figure 5.4D). Interactions where the ligand cell type and/or the receptor cell type was a macrophage scored highly among most LR pairs compared to other cell types (Figure 5.4E). Global interaction states – derived by taking the average of interaction state scores across all colocalized LR pairs – revealed a diverse panel of cell-cell interactions involving LR signaling in ND that shifted towards a growing preference for LR signaling between macrophage-macrophage and monocyte-macrophage by 8w (Figure 5.4F).

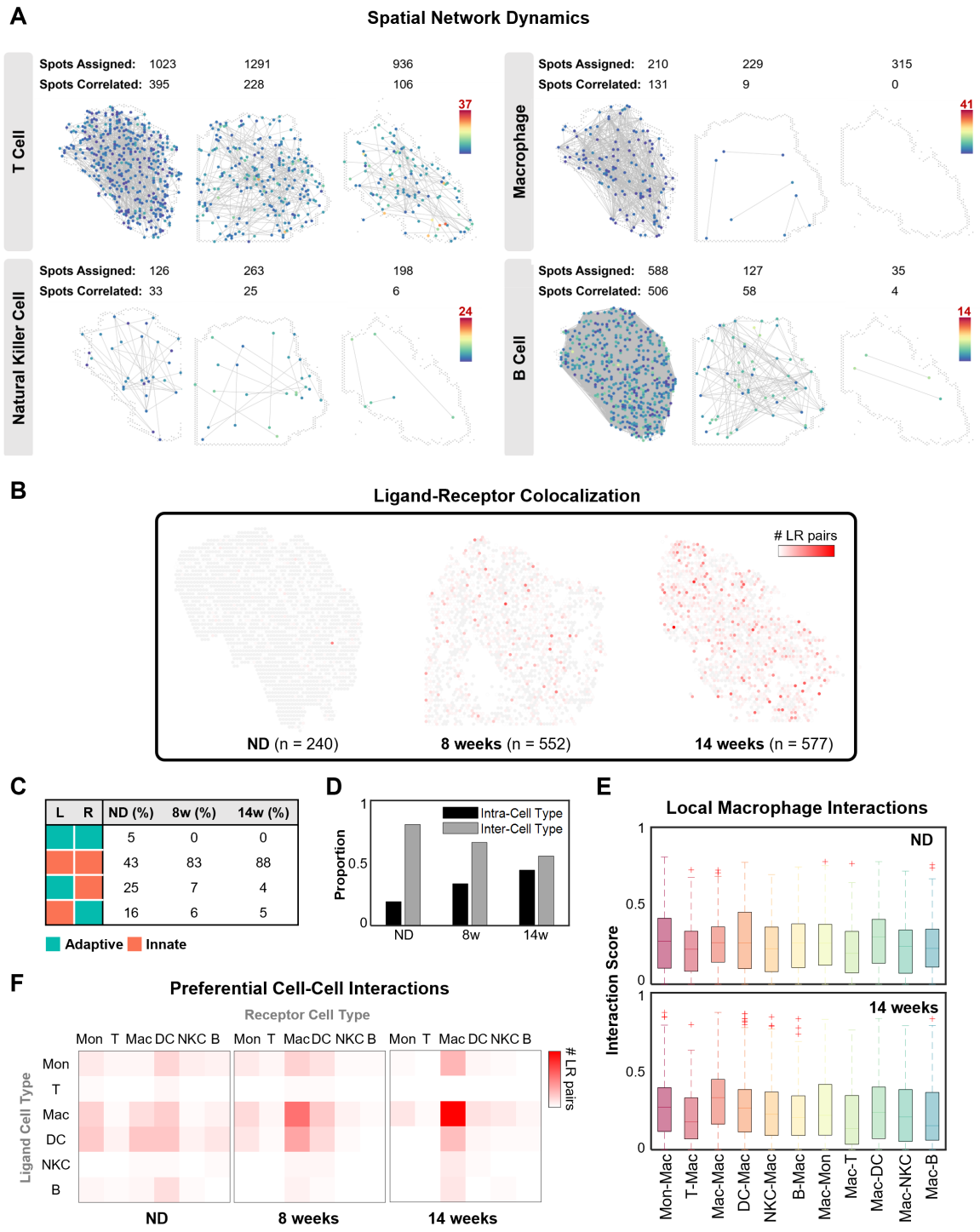

Figure 5.4: Intra-Cell Type Spatial Patterning. (A) Network nodes represent transcriptionally correlated tissue spots assigned to the cell type in the dominant tissue assignment layer. Color intensity of network nodes represents the number of cell type-specific signature genes expressed at the spot. The color bar reflects the maximum number of signature genes (out of 50) expressed at any given spot across all three tissue settings. Network edges represent high correlation (correlation coefficient ≥ 0.85) between cell type-specific gene expression profiles at two spots.

Figure 5.4: (B) Distribution of colocalized ligand-receptor (LR) pairs. Color intensity denotes the number of LR pairs colocalized at the tissue spot (normalized by the total number of tissue spots). (C) Percent of LR pairs that preferentially interact between adaptive and/or innate cell types. (D) Proportion of ligand-receptor interactions between single cells of the same cell type (intra-cell type) versus single cells of different cell types (inter-cell type) (E) Range of interaction scores across all colocalized LR pairs in all macrophage-involved cell-cell pairs. Each LR pair's interaction score between cell types is evaluated by computing the mean between the proportion of single cells of one cell type expressing the ligand and the proportion of single cells of another cell type expressing the receptor. (F) Number of LR pairs that preferentially interact between two cell types. Frequencies are normalized by the total number of colocalized LR pairs in the sample. Preferential interaction was determined by evaluating individual LR interaction scores based on the mean proportion of single cells in a cell type expressing the ligand and the mean proportion of single cells in a cell type expressing the receptor (described in Algorithm 4).

5.3.7 Pre-Crown-Like Structure Neighborhoods Appear in Early Obesity

Crown-like structures (CLS) appear in adipose tissue in chronic obesity, where macrophages accumulate around dead or dying adipocytes (Figure 5.5A). We used *Itgax* and *Trem2*, which are involved in CLS formation [195, 151], and *Cd9*, a marker of lipid-laden macrophages, to identify CLSs across the 8w and 14w tissue sections, verifying the known trend for their emergence in response to HFD feeding (Figure 5.5B). Since CLSs do not form in lean adipose tissue, we screened for *Cd9* on its own in ND tissue to identify tissue spots that could serve as a baseline comparison. We then demarcated neighborhoods around each CLS signature-expressing tissue spot to capture surrounding cells involved in CLS dynamics and explore features that could reveal key attributes of CLS dynamics during obesity (Figure 5.5C, Materials and Methods).

We first determined whether CLS neighborhoods exhibited differential expression. We compared expression pooled across all ND baseline neighborhoods with expression pooled within each individual CLS neighborhood at 14w, finding that 939 genes were uniquely (compared to control neighborhoods, see Materials and Methods) and significantly differentially expressed ($|\log_2 \text{FC}| \geq 1.5$, $p < 0.01$). Of those, 24 genes were differentially expressed in at least 20% of CLS neighborhoods (Figure 5.5D). Moreover, all differentially expressed genes were up-regulated, with *Trem2*, *Cd9*, and *Itgax* among the most frequently occurring DEGs across CLS neighborhoods, as expected. In addition, we found several genes related to tumor progression (*Soat1*, *Hk3*, *Syngr1*) [282, 262], which has been associated with CLS formation.

We then evaluated how LR colocalization trends with CLS emergence. We found that the increase in CLS numbers with obesity was concordant with an increased proportion of LR colocalization within CLS neighborhoods (Figure 5.5E). Specifically, LR colocalization at CLS neighborhoods increased from 18% in ND (17% in control neighborhoods) to 58% at 14w (30% in control

neighborhoods). These colocalized LRs exhibited a variety of dynamical patterns, becoming more enriched or depleted across CLSs over time and interacting between different cell types (Materials and Methods). An overwhelming majority of LR pairs demonstrated an overall increase in the proportion of CLS neighborhoods they colocalized from the lean to obese states (Figure 5.5F). Most of these were LR pairs that increased from ND to 8w and again from 8w to 14w (54%). Interestingly, there was a considerable proportion (20%) of LR pairs whose involvement at CLS neighborhoods increased from ND to 8w then decreased from 8w to 14w. We previously observed this similar "amplification-attenuation" trend among the intra-cell type tissue contributions where both natural killer cells and monocytes demonstrated increased connectivity within their respective cell type networks followed by weakened connectivity (Figure 5.3). This trend led us to speculate about a pre-CLS state in our intermediate HFD feeding time point.

To elucidate cell types responsible for driving dynamic signaling at CLS neighborhoods, we revisited cell types found to dominantly express each LR pair from our earlier analysis in Figure 5.4. Monocytes and dendritic cells dominantly express a large proportion of LR pairs at baseline neighborhoods in the increase-increase group in the ND setting. By 14w, macrophage accounted for the majority of ligands (60.7%) secreted and picked up by receptors (77%) in the CLS niche. We further examined specific LR pairs from the increase-increase group, screening for LR pairs that appeared in CLS neighborhoods only after HFD feeding.

We identified 12 LR pairs that colocalized at no baseline neighborhoods in lean adipose tissue then emerged prominently at these sites in obese adipose tissue (Table S2 in Dotson *et al.* [100]). Of the remaining LR pairs, some involved a ligand and/or receptor with a previously characterized role in obesity, like *C3*, linked to lipid metabolism and contribution to inflammation in adipose tissue during obesity [17]. *C3* was a recurrent, dominantly monocyte-expressed ligand in this group, with an average percent increase in colocalization with receptors at CLS neighborhoods of 54.8% from ND to 14w. *C3* was associated with receptors *Itgax*, *Itgb2*, *C3ar1*, and *Lrp1*, which were expressed by a mix of innate cell types in the data at 8w. At 14w, *C3* only colocalized with receptors associated with macrophages. Several LR pairs exhibited similar behavior.

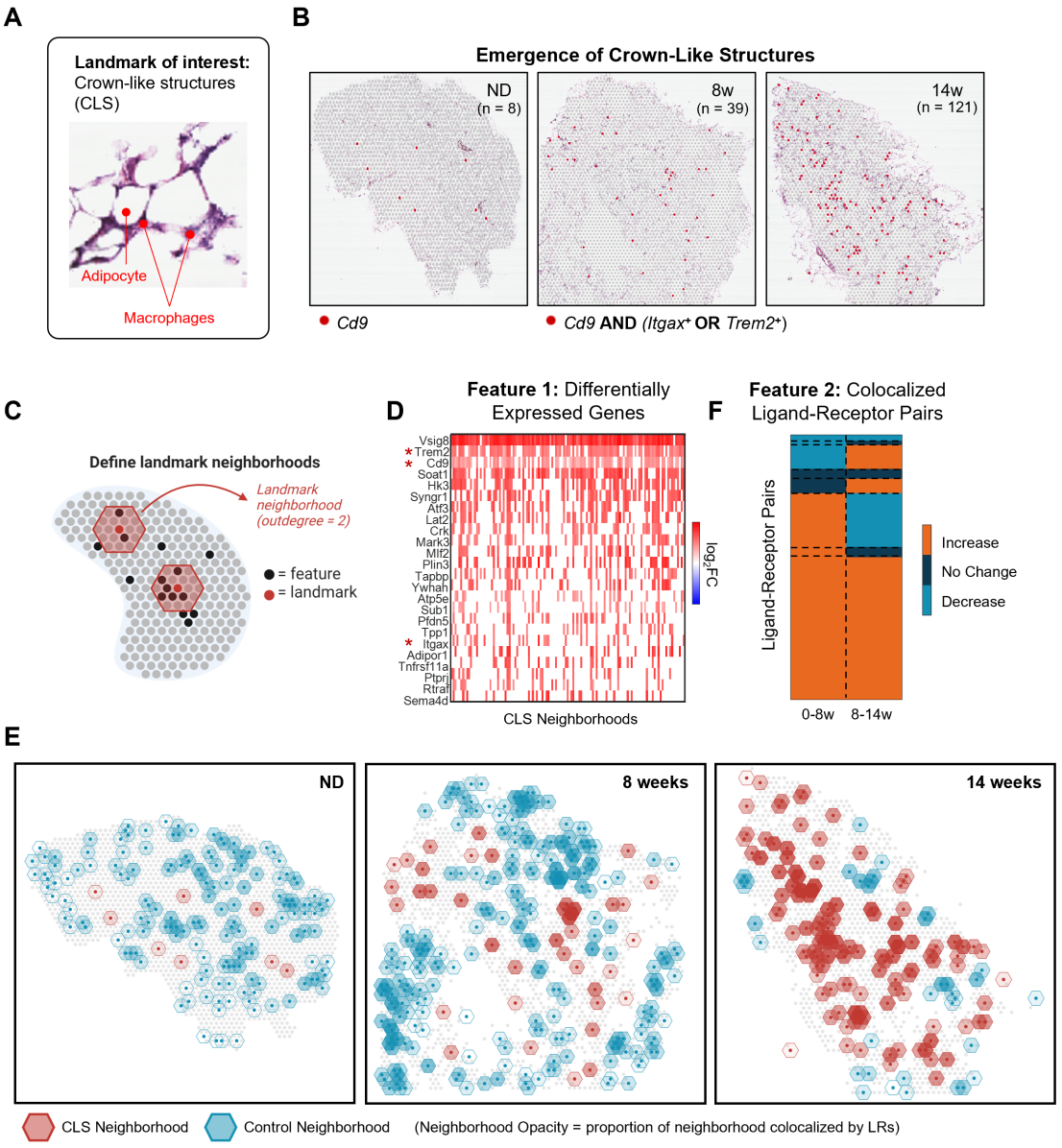

Figure 5.5: Emergence of Crown-Like Structures. (A) H&E staining of the 14w adipose tissue section showing a characteristic crown-like structure (CLS). (B) Identification of CLSs in the ST data. In HFD settings, tissue spots were classified as CLSs if they expressed *Cd9* with either *Itgax* or *Trem2*. As CLSs are not found in lean adipose tissue, only the more general lipid-laden macrophage marker, *Cd9*, was queried in the ND setting. (C) Illustration of CLS neighborhood demarcation. Spots within two outdegrees of the CLS-classified spot (landmark) were defined as the CLS neighborhood. Following analyses evaluated different features relative to these neighborhoods. (D) Differentially expressed genes ($|\log_2 FC| \geq 1.5$, $p < 0.01$) found in at least 20% of CLS neighborhoods. Color intensity indicates fold change between ND and 14w. Red values indicate an increase in expression (positive FC) and blue indicates a decrease in expression (negative FC).

Figure 5.5: (E) LR pairs in CLS neighborhoods. Red neighborhoods correspond to CLSs while blue neighborhoods correspond to a control population containing housekeeping gene *Ppia*. The transparency of the filled-in neighborhoods indicates the number of colocalized LR pairs found in the neighborhood, with more opaque shading corresponding to a higher number of colocalized pairs. For visual clarity, neighborhoods are shown to an outdegree of 1, but analyses were performed on neighborhoods with an outdegree of 2. (F) Patterns of LR colocalization dynamics at CLS neighborhoods. The majority of LRs increased in colocalization at CLSs from ND to 8w and 8w to 14w (orange-orange block). The second most frequent pattern observed was LRs that increased in colocalization at CLSs from ND to 8w, then decreased from 8w to 14w (orange-light blue block).

5.4 Discussion

Inflammation is a prominent feature of obese adipose tissue that contributes to development of insulin resistance. Here, investigation of immune cell signatures in their spatial context highlights the tissue-wide scale of innate and myeloid cell population expansion. While scRNA-seq data improved inference of cell identity, exploring sequencing-based spatial transcriptomic platforms with higher resolution will be valuable [68].

A positive cooperative relationship is consistent with monocyte infiltration and differentiation into macrophages in obesity, and is further supported by the increase in monocyte contribution to tissue function shown in Figure 5.3D. Our spatial patterning analysis showed increased heterogeneity for all immune cell types. This finding is consistent with adoption of additional polarization states or phenotypes in immune cells in obesity, for example in *Cd9*-enriched macrophages in obesity [138, 154, 287]. Emerging subtypes in other immune cells were less identifiable, possibly due to lower potential for different functional states in those cell types. Known shifts in subtypes include increased $Cd8^+$ T effector and $Cd4^+$ T_H1 cells and decreased regulatory T cells in obesity [279, 261, 260].

Additionally, while increased IgG^+ B cells have been observed in obese murine visceral adipose tissue [382], the only differentially expressed gene in our B cell sub-clustering was the immunoglobulin gene, *Ighg3*, which was enriched in the high-fat diet settings ($\log_2$FC of 1).

Ligands and receptors expressed in innate cells were preferentially colocalized across the obese tissue compared to ligands and receptors expressed in adaptive cells or a combination of innate and adaptive cells. These interactions were predominantly from monocytes and ATMs, consistent with their increased quantity in obesity. Our adaptation of Turing's system for within-cell and between-cell dynamics similarly showed an increasing contribution of monocytes and ATMs to tissue function in obesity, alongside a decreasing contribution of adaptive immune cells, NK cells,

and ATDCs.

Finally, differential and ligand-receptor expression at mapped crown-like structures revealed a shift toward monocyte-driven signaling to lipid-associated ATMs in short-term high-fat diet feeding, prior to architectural formation of crown-like structures. These data support a model in which monocytes, recruited from circulation, both differentiate into lipid-associated macrophages and participate in cell-cell signaling via specific obesity-activated ligands. Notably, this pattern was present after 8w of HFD feeding, which precedes observable CLS architecture.

Future work will stain crown-like structure biomarkers to confirm their presence, though colocalization of CLS-forming markers in addition to properties such as increased gene signatures associated with tumor-infiltrating cell types provide initial confidence towards their emergence in our samples. Additionally, while we cannot say with certainty that the cell type interactions reported account for the observed ligand-receptor activity at these tissue landmarks, or rule out the involvement of adipocytes and other non-immune cell types, our data capture the dominant immune cell types participating in colocalized ligand-receptor expression during HFD feeding. High numbers of CLSs have been linked to more severe clinical outcomes of obesity and associated disorders like cancers [232, 152, 53, 172]. Further assessment and validation of the trends we observed in this study could help uncover mechanisms of CLS initiation and therapeutic targets in early obesity.

This study provides proof-of-concept for predicting the trajectory of tissue states in disease. Our model of disrupted tissue morphogenesis in disease could be generalized to other tissues to define a disease index that distinguishes healthy from disrupted tissue states predictive of metabolic dysfunction or other diseases. Relevant to obesity are liver and intestinal tissue, which have defined tissue changes in gene expression in obesity [54, 88, 87]. Integrated analysis of spatial and single-cell transcriptomics in these tissues could allow us to establish a consistent inter-tissue index in obesity.

Some obesity-related changes in adipose tissue immune cells persist even in weight loss [396], highlighting the need to better understand mechanisms that promote adipose tissue dysfunction. These mechanisms could also tie obesity to the risk for other conditions, including cardiovascular disease and cancer [48, 112], with one study finding that metabolically activated macrophages in mammary adipose tissue promotes triple-negative breast cancer [346]. Future studies using time course spatial transcriptomics could be particularly powerful in understanding tissue dynamics and the evolution of cell-cell communication in adipose tissue.

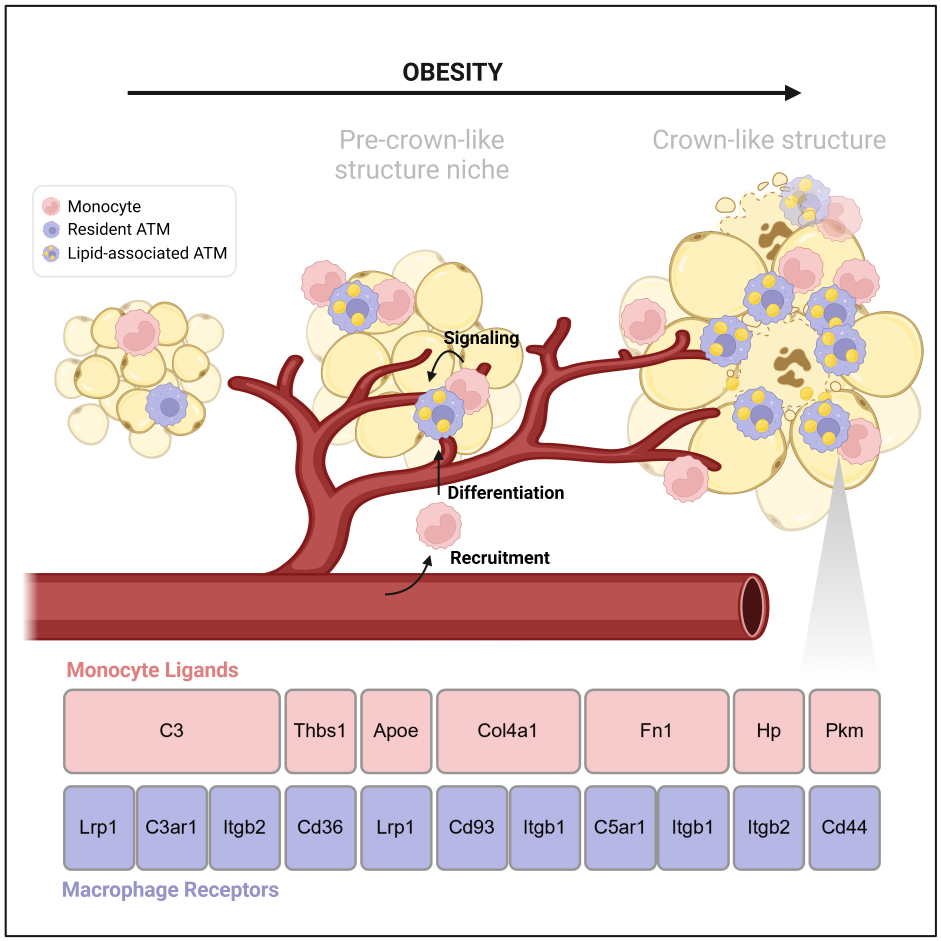

Figure 5.6: Mechanism of Crown-Like Structure Emergence. As obesity progresses, we observe increasing co-localization and signaling between monocytes and macrophage, supporting a mechanism for CLS formation in which monocytes, recruited from circulation, both differentiate into lipid-associated macrophage and cooperate with existing lipid associated macrophage through ligand-receptor signaling. Remarkably, this pattern is present in 8w tissue, preceding the formation of these structures. As macrophage accumulate at formed CLSs in 14w tissue, monocytes persist in their communication with macrophage.

5.5 Materials and Methods

5.5.1 Animals

C57BL/6J mice were used for these experiments (Jackson Laboratories 000664). Male mice were fed ad libitum a control normal chow diet (ND; 13.4% fat, 5L0D LabDiet) or high-fat diet (HFD; 60% calories from fat, Research Diets D12492) for the indicated amount of time starting at 9 weeks old. Animals were housed in a specific pathogen-free facility with a 12 h light/12 h dark cycle and given free access to food and water except for withdrawal of food for temporary fasting associated with glucose tolerance tests. All mouse procedures were approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Michigan (Animal Welfare Assurance Number D16-00072 (A3114-01), #PRO00008583), and care was taken to minimize suffering adhering to the Institute of Laboratory Animal Research Guide for the Care and Use of Laboratory Animals.

5.5.2 Glucose Tolerance Tests

For glucose tolerance tests (GTT), starting four hours into the light cycle, mice were fasted with ad libitum access to water for six hours in clean cages. A 100 mg/mL D-glucose (Sigma G7021) solution was prepared in sterile -/- DPBS and injected at 0.7 g/kg of body weight. Area under the curve (AUC) calculations were performed using the log trapezoidal method.

5.5.3 Stromal Cell Isolation

Stromal vascular cells (SVCs) were collected from adipose tissues as in [231]. After cardiac perfusion, adipose tissues were collected, minced finely to 3-5 mm pieces, and added to ice cold HBSS+Ca/Mg. Up to 1.5 g of tissue per sample was digested in 10 ml of 1 mg/mL collagenase II (Sigma C68850) in HBSS+Ca/Mg at 37°C for 45 minutes with vigorous shaking. Digests were filtered through buffer-soaked 100 micron cell strainers and centrifuged at 300 x g at 4C to pellet SVCs.

5.5.4 Immune Cell Enrichment and Single-Cell RNA-Sequencing

SVCs were enriched for $CD45^+$ immune cells using Biolegend MojoSort Mouse CD45 Nanobeads (Biolegend 480027), following the manufacturer's protocol. Briefly, SVC pellets were resuspended in 1 mL MojoSort Buffer, pooling the four samples from each cohort into a single respective cohort tube (ND, 8w, 14w), then filtered through a 70 micron cell strainer and placed in 5 mL polypropylene tubes. After addition of nanobeads, samples were sequentially processed for

magnetic separation. To increase purity, three magnetic separations in total were performed on the labeled fractions. Final cell suspensions were filtered through 40 micron pipette tip filters. Cell viability was >80% with <15% aggregation.

5.5.5 Spatial Transcriptomics Tissue Preparation

Within 30 minutes of cardiac perfusion, adipose tissues were pre-soaked in ice cold O.C.T. Compound (VWR 25608-930) and placed in biopsy cryomolds (VWR 25608-922) with fresh O.C.T., rapidly frozen by immersion in liquid nitrogen-cooled isopentane, and kept on dry ice or at -80°C until sectioning. Fresh tissue sections were cut at 10 μm after 20 minute equilibration in a cryochamber set to -26°C or below with specimen arm at -40°C. Samples were placed on a 10X Genomics/Visium Spatial Gene Expression slide and processed by the University of Michigan Advanced Genomics Core according to the manufacturer's protocol.

5.5.6 Single-Cell RNA-Sequencing Data Processing

Raw single-cell RNA-sequencing data were processed using the 10X Genomics CellRanger (version 4.0.0) pipeline and resulting feature-barcode matrices were loaded into MATLAB for further processing using an in-house pipeline. Single cells (barcodes) with fewer than 500 expressed genes were filtered from the data matrix leaving 1,231 cells in the ND setting, 6,011 cells in the HFD8 setting, and 6,258 cells in the HFD14 setting. Genes (features) were filtered out according to the same criteria used to filter genes in the spatial transcriptomics data. The median number of genes expressed per cell for each setting was: 1,707, 1,644, and 1583, respectively. Additionally, filtered matrices were normalized in the same way as the spatial transcriptomics data matrices.

5.5.7 Spatial Transcriptomics Data Processing

Raw sequencing data were processed using the 10X Genomics SpaceRanger (version 1.0.0) pipeline with mouse reference GRCm38, and resulting feature-barcode matrices were loaded into MATLAB for further processing using an in-house pipeline. Matrices contain absolute transcript counts, reported as unique molecular identifier (UMI) counts, pertaining to each feature (rows) and tissue-associated barcode (columns). Informed by the 10X Genomics Loupe Browser visualization of barcoded tissue spots overlaying tissue histology images, we filtered out barcoded spots that overlapped visible holes in the tissue and therefore had no tissue material for gene expression to be detected at. This resulted in no barcoded spots being discarded from from the ND setting, 1,458 being discarded from the HFD8 setting, and 490 from the HFD14 setting. After filtering, there were 2,035 , 1,936, and 1,504 viable spots remaining in each setting, respectively. Genes that

were detected at no barcoded tissue spots (indicated by a row sum of 0) were filtered from the feature-barcode matrices. Additionally, we filtered out mitochondrial and ribosomal genes as well as genes known to be linked to mapping errors (*Malat*, *Lars2*, *Kcnq10t1*, and *Gm42418*) [361]. Our working features list contained 19,941 total genes. The median number of expressed genes per tissue spot for each setting was: 91, 105, and 173, respectively. Finally, we normalized each feature-barcode matrix by dividing by a scaling factor - computed as the median UMI counts per barcode divided by the total UMI counts per barcode - then log-transforming the matrix.

5.5.8 Integration of Spatial and Single-Cell Transcriptomics Data

The cellular composition of each tissue section was characterized by mapping expression profiles from scRNA-seq data to their paired spatial landscape. To achieve this, we first aggregated our scRNA-seq data matrices, mean-centered and dimensionally reduced the resulting matrix to the first 20 singular vectors, then performed spectral clustering using Euclidean distance as the metric for generating the similarity graph. Informed by known cell types found in adipose tissue from literature, we selected for 15 clusters (k = 15). We then employed a data-guided cell type identification approach where we scored markers for each cluster based on ubiquity (percent of cells in the cluster expressing each gene), cluster-averaged expression (average expression of each gene among all cells in the cluster), and uniqueness (see 'Marker Uniqueness' subsection). For each cluster, the top 50 genes with the highest cumulative score that were also expressed across all three dietary settings and expressed in the spatial transcriptomics data, were selected as cluster-specific signatures. We used prior literature and the CellMarker database [403] to query these signatures and call cell types for each cluster. We further narrowed our set of clusters by removing those indicative of fibroblast contamination, removing those of unknown lineage whose signatures contained non-essential genes, and merging clusters with redundant cell type classifications, leaving us with six total clusters. We validated our cell type annotations using Immgen and found that they aligned with Immgen signatures [134]. To annotate our spatial transcriptomics tissue sections, we mapped cell types identified from the scRNA-seq data to tissue spots, based on the proportion of signature genes expressed at each spot (see 'Spot Deconvolution' subsection).

5.5.8.1 Marker Uniqueness

To evaluate the uniqueness of potential cell type markers to a cluster, we found the ratio of the total number of clusters overexpressing a given gene (fold change ≥ 2) to the number of times that gene is overexpressed by the cluster of interest. The resulting value increases a marker's score proportional to how frequently we see the gene expressed highly in one cluster compared to all other clusters, but offsets the score if the gene is highly expressed in multiple clusters. Ultimately,

non-unique genes rank low with this approach. We found that this approach is similar to the Term Frequency-Inverse Document Frequency (TF-IDF) technique commonly applied in machine learning contexts to determine the importance of a word within a document or set of documents [297]. For proof-of-concept, we adapted the TF-IDF technique for marker ranking by treating documents as clusters and words as genes, where term frequency corresponded to the number of times a given gene is overexpressed by the cluster of interest and inverse document frequency to the number of clusters overexpressing the gene. In doing so, we yielded similar results (data not shown).

5.5.8.2 Spot Deconvolution

Given the thickness of tissue sections and diameter of tissue spots, it is likely that two or more cell identities will colocalize at the same tissue spot. To deconvolute each spot into their possible cellular components, we start by defining a dominant tissue layer wherein a cell type is assigned to a spot if at least five of the cell type-specific signature genes are expressed at the spot and a higher proportion of its signature is expressed at the spot than any other cell type. We define subsequent tissue layers by assigning remaining cell types to each tissue spot following the same criteria as long as the cell type has not already been assigned to the spot in a previous layer. Additionally, we impose a tiebreak condition for when two or more cell types express the same number of genes at a spot. When this happens, the tied cell types are assigned according to highest to lowest cumulative expression of their expressed signature genes. With this approach, each cell type has an equal likelihood of being assigned to a layer, giving us a total number of layers equivalent to the number of cell types identified in the tissue.

5.5.9 Modeling Tissue Function

While our spatial transcriptomics data is not single-cell resolution, we are able to assign a dominant cell type to each tissue spot based on the expression of cell type-specific signatures. We can then treat each tissue spot as an individual cell in our model of emerging tissue function. While we also described non-dominant layers of cell type assignments at each spot, we focus our model on the dominant cell type assignment. Our model draws on two subnetworks of the overall tissue network - the cell-cell network and within-cell genome network. Taken together, we can represent each cell type's individual contribution to tissue function and define an index describing the tissue's state in terms of cellular and genome connectivity.

5.5.9.1 Cellular Connectivity

Spatial transcriptomics captures the expression profiles of cells tissue-wide in Euclidean space, allowing us to observe patterns of between-cell connectivity across the tissue. We start by characterizing *intra*-cell type relationships within individual cell populations across the tissue (emergence of tissue function) by finding the transcriptome-wide correlation between tissue spots assigned to a given cell type. We then weight each pairwise correlation coefficient by the Euclidean distance between their corresponding tissue spots, generating an adjacency matrix. We next perform eigenvector decomposition on the normalized Laplacian of the adjacency matrix to obtain the Fiedler number, representing the connectivity of the network of transcriptionally similar cells across the tissue – cell connectivity. The cell type-specific Fiedler number is then divided by the number of tissue spots assigned to the cell type to normalize for the variability in network sizes across the tissue.

To capture *inter*-cell type contributions to tissue function (emergence of *multicellular* tissue function) we consider a network of networks where each cell type is a subnetwork within the larger tissue network. We repeat our previous steps, this time generating an adjacency matrix based on the correlation in expression between tissue spots from two cell types instead of one. We do this for every pair of cell types represented in our tissue samples to demonstrate how cell type coupling contributes to tissue function.

5.5.9.2 Genome Connectivity

From single-cell RNA-seq, we compare the diversity in gene expression profiles within a given cell type to capture within-cell genome connectivity across the tissue. Similar to our computation for cellular connectivity, we find the transcriptome-wide correlation between single cells classified as a given cell type. We then perform eigenvector decomposition on the normalized Laplacian of this correlation matrix to obtain the Fiedler number which we use to define genome connectivity. Again, we normalize the Fiedler number, this time dividing by the number of single cells assigned to the cell type.

5.5.10 Ligand-Receptor Analysis

Known ligand-receptor pairs in mice were taken from a compendium of pairs reported throughout the literature compiled by the Lewis Lab at UCSD. [10]. For each ligand-receptor pair and timepoint, we scanned the tissue space to identify tissue spots where both the ligand and receptor of the pair colocalize. We define colocalization as the spatial overlap or co-expression of the ligand and receptor (UMI count > 0) belonging to a known pair at the same x-,y- coordinate (tissue spot) in the spatial transcriptomics data. Paired ligands and receptors that colocalized at at least one

tissue spot were further queried for cell-type specific expression at the single-cell level, where the independent expression of the ligand and receptor in adaptive and innate immunity cells was evaluated. The proportions of single cells annotated as: (1) adaptive immunity cell types that expressed the ligand, (2) adaptive immunity cell types expressing the receptor, (3) innate immunity cell types expressing the ligand, and (4) innate immunity cell types expressing the receptor were captured in a 2-by-2 contingency table. The UMI count cutoff for ligand or receptor expression was 3 and pairs where the ligand and/or receptor were not expressed in more than 1% of cells in at least one cell type were discarded. Based on these proportions, we developed an interaction score between the immunity classes for each LR pair where we took the mean of the ligand immunity class proportion and a receptor immunity class proportion, allowing scores to represent the likelihood that the LR pair interacts between those two immune cell classes. We generated local state matrix to reflect these scores, where for a given ligand-receptor pair the intra-interaction between the ligand and receptor in adaptive cells (mean of proportions 1 and 2), the intra-interaction between the ligand and receptor in innate cells (mean of proportions 2 and 3), the inter-interaction between the ligand in adaptive and receptor in innate cells (mean of proportions 1 and 4), and the inter-interaction between the ligand in innate cells and receptor in adaptive cells (mean of proportions 3 and 2) were captured. For each LR pair, we could then determine which class had the highest score, which we interpret as the preferential immunity class pairing exhibiting that LR pair's signaling. Finally, a global state matrix summarizing these intra- and inter-interactions among all colocalized ligand-receptor pairs was derived by taking the element-wise mean across the local state tables. Local and global state matrices were derived for individual cell types as well, as described in Algorithm 4.

5.5.11 Tissue Landmark Analysis

The tissue landmark of focus for this study was crown-like structures (CLSs). To identify tissue positions associated with this tissue niche, we queried each barcoded spot for expression of the macrophage marker *Cd9* in addition to either *Itgax* or *Trem2*, two markers known to be highly expressed at CLS sites. This selection criteria was applied to our two HFD settings. CLSs are a hallmark of obesity and are not present in lean adipose tissue, so we screened tissue positions in the ND setting using only *Cd9* for a baseline comparison. CLSs involve multiple cells (a ring of macrophages surrounding an adipocyte) and because adipocytes are expanded in size in obese tissue, activity related to these structures span multiple tissue spots. To capture all relevant signals at and around the CLS positions, we defined landmark neighborhoods, centered around each identified CLS that encompass tissue spots within two outdegrees from the center. We also defined control neighborhoods, centered around spots expressing hallmark gene *Ppia*. Control neighborhoods that overlapped with CLS neighborhoods were removed from analysis. We then assessed

features at CLS neighborhoods and used control neighborhoods as a background distribution of tissue spots to gauge the significance of observed trends.

For neighborhood differential expression analysis, expression profiles at spots across all neighborhoods in the ND setting were first pooled together by taking the average expression across spots. Differential expression was performed between pooled expression in ND and each individual CLS neighborhood in the 14w setting, where average expression across spots within a neighborhood were used. The same analysis was performed for ND and 14w control neighborhoods. Genes were assessed for how many CLS neighborhoods they were significantly differentially expressed in ($|\log_2 FC| \geq 1.5$, $p < 0.01$). Overlaps with DEGs in control neighborhoods were filtered.

We evaluated ligand-receptor (LR) signaling at CLS neighborhoods by evaluating the proportion of neighborhood spots colocalized by an LR pair. Only LR pairs found to be colocalized in all three settings were used in this analysis (132 pairs). We evaluated the same proportion for control neighborhoods and compared the average proportion across both sets of neighborhoods for each setting. We also determined how the colocalization of each LR pair at CLS neighborhoods changed over time, relative to the control neighborhoods, grouping LRs based on their pattern of change from ND to 8w and 8w to 14w. For example, LR pairs whose colocalization at CLS neighborhoods (percent of neighborhoods where the LR pair was co-expressed in at least one spot) consistently increased over time were grouped together while those whose colocalization increased from ND to 8w then decreased from 8w to 14w were grouped together. Within each group, LR pairs were then ranked in descending order according to their cumulative change over time and those whose rate of colocalization at CLS neighborhoods were not at least 10% different than at control neighborhoods were further disregarded. Individual LR pairs across groups were further evaluated based on observed trends for biological relevance.

CHAPTER 6

Deciphering Multi-way Interactions in the Human Genome

This chapter is based on a paper by Gabrielle A. Dotson, Stephen Lindsly, Can Chen, Anthony Cicalo, Sam Dilworth, Charles Ryan, Sivakumar Jeyarajan, Walter Meixner, Nicholas Beckloff, Amit Surana, Max Wicha, Lindsey A. Muir, and Indika Rajapakse [99] (under review).

6.1 Abstract

Chromatin architecture, a key regulator of gene expression, can be inferred using chromatin contact data from genome-wide chromosome conformation capture, or Hi-C. However, classical Hi-C does not preserve multi-way contacts. Here, we use long sequencing reads to map genome-wide multi-way contacts and investigate higher order chromatin organization in the human genome. We use hypergraph theory for data representation and analysis, and quantify higher order structures in neonatal fibroblasts, biopsied adult fibroblasts, and B lymphocytes. By integrating multi-way contacts with chromatin accessibility, gene expression, and transcription factor binding, we introduce a data-driven method to identify cell type-specific transcription clusters. We provide transcription factor-mediated functional building blocks for cell identity that serve as a global signature for cell types.

6.2 Introduction

Structural features of the genome are integral to the regulation of gene expression and corresponding generation of cellular phenotypes [190, 220, 61]. Aspects of genome structure have been inferred by studying genomic regions that are in close physical proximity. Chromosome conformation capture (3C)-based methods capture these interactions (contacts) through chemical fixation,

digestion of DNA, and proximity ligation, followed by sequencing of ligated DNA to identify genomic regions that are in contact. A variety of cell types have now been characterized using Hi-C, a genome-wide 3C-based method, adding substantially to our understanding of genome architecture. However, limitations on read length during sequencing lead to over-representation of simple interactions, predominantly pairwise. Identification of more complex, higher-order interactions can help us build a more complete set of principles of genome architecture.

Multi-way contacts have been identified using targeted 3C-based methods [12, 243, 249], through inference from pairwise contacts [196], and on occasion using classical Hi-C [83]. Ligation-free approaches, such as GAM and SPRITE, have recently enabled large scale capture of multi-way interactions [22, 263, 405], though comparisons of different methods find under- and over-representation of higher order contacts in the absence of proximity ligation [93, 161].

A recent extension of Hi-C preserves multi-way interactions and uses sequencing of long reads (e.g. Pore-C) [93] to unambiguously identify sets of contacts among multiple loci. Multi-contact 4C sequencing (MC-4C) also uses long-read sequencing to capture contact complexity [5], however, it was designed to capture local topology for individual genes and regulatory regions and does not generate multi-way contacts genomewide. While direct capture of multi-way contacts can clarify higher order structures in the genome, new frameworks are needed to address unique analysis and representation challenges posed by the multi-way data.

To address this gap, we generated Pore-C data from neonatal and biopsied adult fibroblasts and collected publicly available Pore-C data for B lymphocytes [93] and constructed hypergraphs to represent the multidimensional relationships of multi-way contacts among loci. Hypergraphs are similar to graphs, but hypergraphs contain hyperedges instead of edges. Hyperedges can connect any number of nodes at once, while edges can only connect two nodes [26, 60, 58]. Prior work on neural networks highlights the utility of hypergraph representation learning to denoise and analyze existing multi-way contact data and to predict *de novo* multi-way contacts [401]. Here, we use incidence matrix-based representation and analysis of multi-way chromatin structure directly captured by Pore-C data, which is mathematically simple and computationally efficient, and yet can provide new insights into genome architecture.

In our hypergraph framework, nodes are genomic loci and hyperedges are multi-way contacts among loci. In our incidence matrices, rows are genomic loci and columns are individual hyperedges. This representation enabled quantitative measurements of chromatin architecture through hypergraph entropy and the comparison of different cell types through hypergraph similarity measures. In addition, we integrated Pore-C with other data modalities to discover biologically relevant multi-way interactions, which we term transcription clusters. The cell-type specific transcription clusters we identified support a role in maintaining cell identity, consistent with prior work on transcriptional hubs or factories [247, 248, 303, 408, 321]. Furthermore, the formation of transcription

clusters in the nucleus is consistent with small world phenomena in networked systems [375, 36].

We use the following definitions. **Entropy**: a measure of structural order in the genome. **Hyperedge**: an extension of edges where each hyperedge can contain any number of nodes (multi-way contact). **Hypergraph**: an extension of graphs containing multiple hyperedges. **Hypergraph motifs**: an extension of network motifs that describe connectivity patterns of 3-way, 4-way, ..., n-way hyperedges. **Incidence matrix**: a representation for hypergraphs where rows are nodes and columns are hyperedges. **Transcription cluster**: a group of genomic loci that colocalize for efficient gene transcription.

6.3 Results

6.3.1 Capturing Multi-way Contacts

We conducted Pore-C experiments using adult human dermal fibroblasts obtained from a skin biopsy and neonatal human dermal fibroblasts, and obtained additional publicly available Pore-C data from B lymphocytes [93]. The experimental protocol for Pore-C is similar to Hi-C, including cross-linking, restriction digestion, and ligation of adjacent ends followed by sequencing (Figure 6.1A). Alignment of Pore-C long reads to the genome enables fragment identification and classification of multi-way contacts (Figure 6.1B).

Hypergraphs represent multi-way contacts, where individual hyperedges contain at least two loci (Figure 6.1C, left). Hypergraphs provide a simple and concise way to depict multi-way contacts and allow for abstract representations of genome structure. Computationally, we represent multi-way contacts as incidence matrices (Figure 6.1C, right). For Hi-C data, adjacency matrices are useful for assembly of pairwise genomic contacts. However, since rows and columns represent individual loci, adjacency matrices cannot be used for multi-way contacts in Pore-C data. In contrast, incidence matrices permit more than two loci per contact and provide a clear visualization of multi-way contacts. Multi-way contacts can also be decomposed into pairwise contacts, similar to those in Hi-C, by extracting all pairwise combinations of loci (Figure 6.1D).

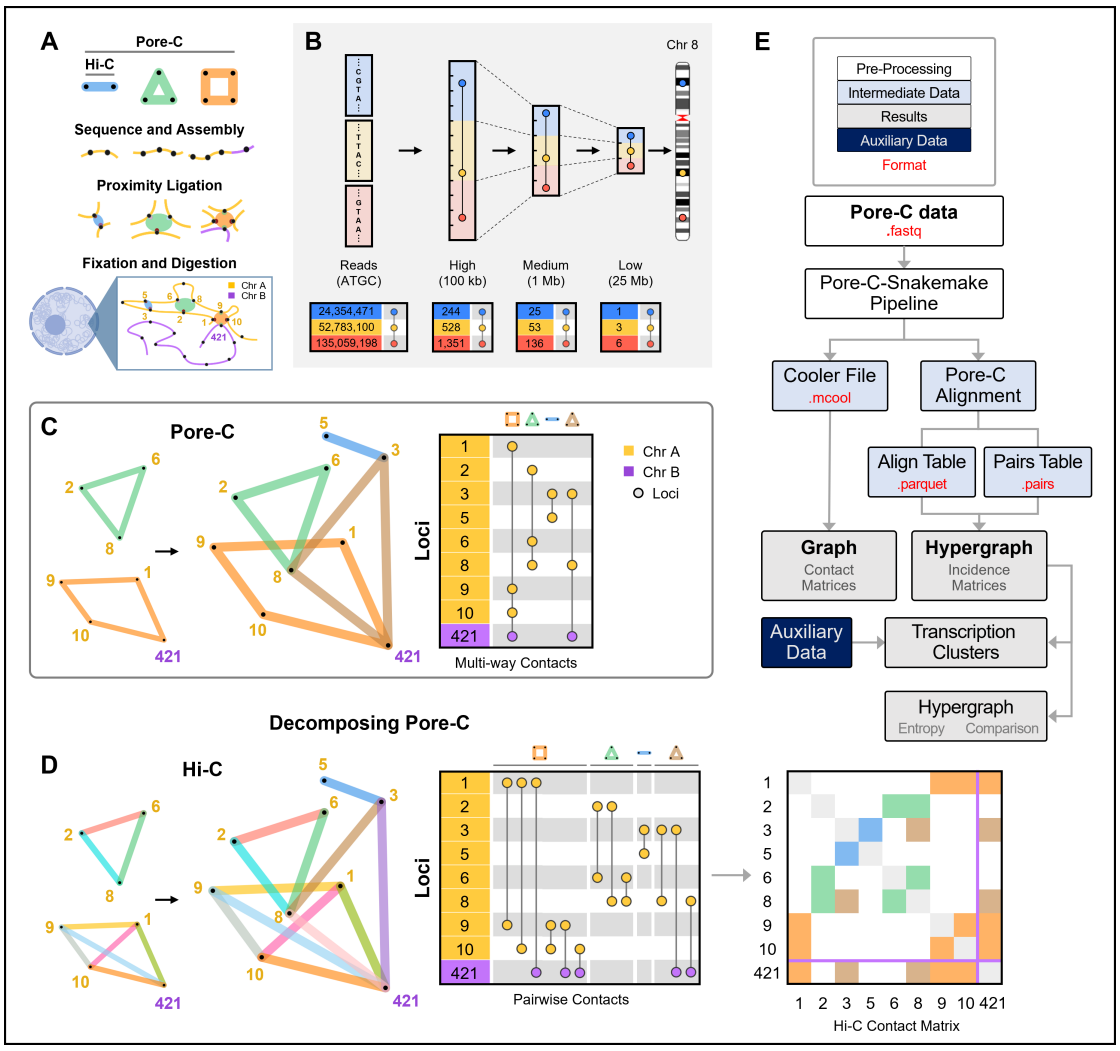

Figure 6.1: Pore-C Experimental and Data Workflow. (A) The Pore-C experimental protocol, which captures pairwise and multi-way contacts (*Materials and Methods*). (B) Representation of multi-way contacts at different resolutions (top). Incidence matrix visualizations of a representative example from Chromosome 8 in adult human fibroblasts at each resolution (bottom). The numbers in the left columns represent the location of each genomic locus present in a multi-way contact, where values are either the chromosome base-pair position (read-level) or the bin into which the locus was placed (binning at 100 kb, 1 Mb, or 25 Mb). (C) Hypergraph representation of Pore-C contacts (left) and an incidence matrix (right) of four multi-way contacts within (yellow-to-yellow) and between (yellow-to-purple) chromosomes. Contacts correspond to examples from A. The numbers in the left column represent genomic bins in which a locus resides. Each vertical line represents a multi-way contact, with nodes at participating genomic loci. (D) Multi-way contacts can be decomposed into pairwise contacts. Decomposed multi-way contacts can be represented using graphs (left) or incidence matrices (middle), which when decomposed are interchangeable with traditional Hi-C contact matrices (right). Contacts correspond to examples from A and C. (E) Flowchart overview of the computational framework. Descriptions of file type formats (red text) are in Table S1 in Dotson *et al.* [99].

6.3.2 Decomposing Multi-way Contacts

From our Pore-C experiments using adult human dermal fibroblasts, neonatal human dermal fibroblasts, and additional publicly available Pore-C data from B lymphocytes, we constructed hypergraphs at multiple resolutions (read level, 100 kb, 1 Mb, and 25 Mb) [93]. We first analyzed individual chromosomes at 100 kb resolution by decomposing multi-way contacts into their pairwise contacts. Decomposing Pore-C data into pairwise contacts provides more information than Hi-C, as each Pore-C read can contain many pairwise contacts [93]. It also allows us to identify topologically associated domains (TADs) using established methods (*Materials and Methods*) [94, 274, 63]. We demonstrate identification of TAD boundaries from decomposed multi-way contacts and show intra- and inter-TAD relationships using multi-way contacts (Figures 6.2, S1 in Dotson *et al.* [99]). The loci that frequently participate in these multi-way contacts give rise to the block-like pattern of chromatin interactions often seen in Hi-C data.

6.3.3 Chromosomes as Hypergraphs

To gain a better understanding of genome structure with multi-way contacts, we constructed hypergraphs for entire chromosomes at 1 Mb resolution. We show an incidence matrix of Chromosome 22 as an example in Figure 6.3A, and in Figure 6.3B, we visualize the distribution of 1 Mb contacts at multiple orders (2-way contacts, 3-way contacts, etc.) on Chromosomes 22. Figure 6.3C highlights the most common intra-chromosomal multi-way contacts on Chromosome 22 using multi-way contact "motifs", which we use as a simplified way to show hyperedges. Figure 6.3D shows how multi-way contacts at lower resolutions (25 Mb, 1 Mb) are composed of many multi-way contacts at higher resolutions (100 kb, read level), and Figure 6.3E visualizes the multi-way contacts contained in Figure 6.3D as a hypergraph.

We also identified multi-way contacts that contain loci from multiple chromosomes. These inter-chromosomal multi-way contacts can be seen at 1 Mb resolution in Figure 6.3F and in 25 Mb resolution for both adult fibroblasts and B lymphocytes in Figure 6.4A and 6.4B, respectively. Figure 6.4 gives a summary of the entire genome's multi-way contacts, by showing the most common intra- and inter-chromosomal multi-way contacts across all chromosomes. We highlight examples of multi-way contacts with loci that are contained within a single chromosome ("intra only"), spread across unique chromosomes ("inter only"), and a mix of both within and between chromosomes ("intra and inter"). We note that many of the "inter only" contacts observed in Figure 6.4 may also have intra-chromosomal contacts when viewed at a higher resolution, similar to Figure 6.3D. Finally, we found the most common inter-chromosomal multi-way contacts across all chromosomes, which we summarize with five example chromosomes in Figure 6.5 using multi-way contact motifs. These multi-way contacts between distant genomic loci may offer insights

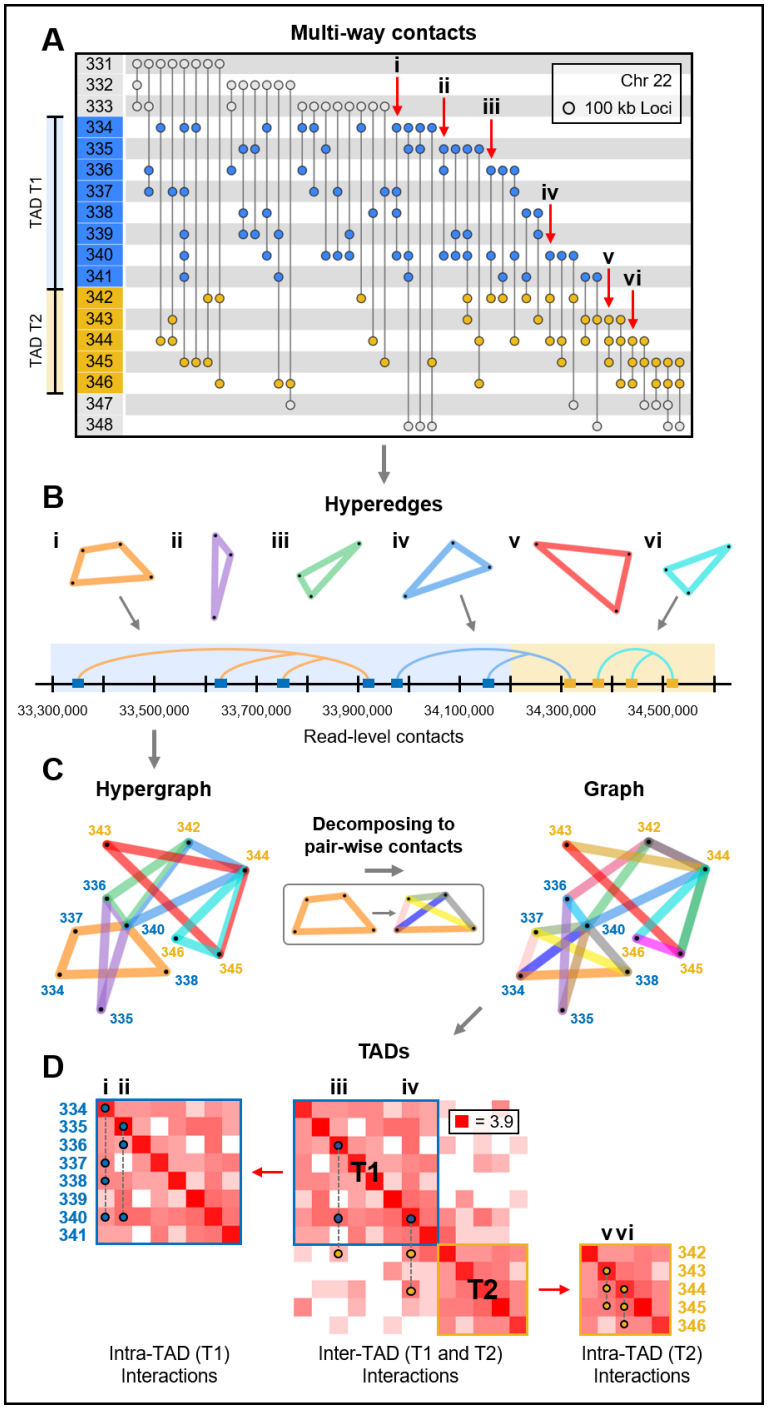

Figure 6.2: Local Organization of the Genome. (A) Incidence matrix visualization of a region in Chromosome 22 from adult fibroblasts (V1-V4). The numbers in the left column represent genomic loci at 100 kb resolution, vertical lines represent multi-way contacts, where nodes indicate the corresponding locus' participation in this contact. The blue and yellow regions represent two TADs, T1 and T2. The six contacts, denoted by the labels i-vi, are used as examples to show intra- and inter-TAD contacts in B, C, and D. (B) Hyperedge and read-level visualizations of the multi-way contacts i-vi from the incidence matrix in A. Blue and yellow shaded areas (bottom) indicate which TAD each locus corresponds to.

Figure 6.2: (C) A hypergraph is constructed using the hyperedges from B (multi-way contacts i-vi from A). The hypergraph is decomposed into its pairwise contacts in order to be represented as a graph. (D) Contact frequency matrices were constructed by separating all multi-way contacts within this region of Chromosome 22 into their pairwise combinations. TADs were computed from the pairwise contacts using the methods from [63]. Example multi-way contacts i-vi are superimposed onto the contact frequency matrices. Multi-way contacts in this figure were determined at 100 kb resolution after noise reduction, originally derived from read-level multi-way contacts (*Materials and Methods*).

into the higher-order structural patterning of the genome and its relationship with transcriptional regulation.

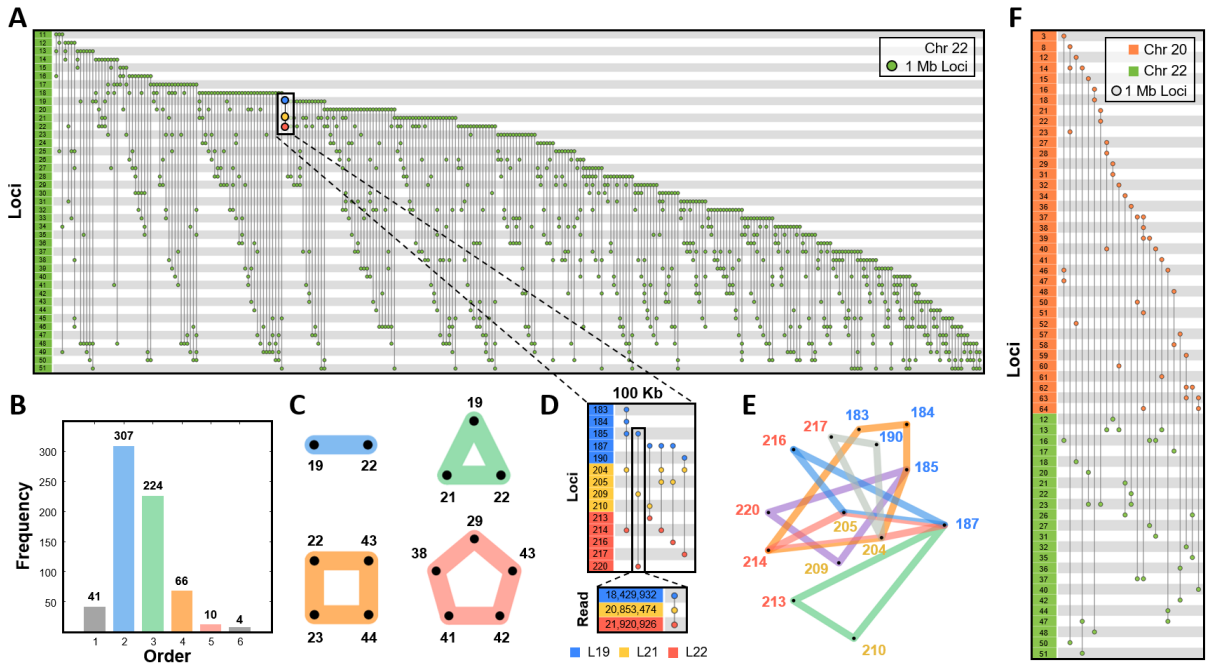

Figure 6.3: Patterning of Intra- and Inter-chromosomal Contacts. (A) Incidence matrix visualization of Chromosome 22 in adult fibroblasts. The numbers in the left column represent genomic loci at 1 Mb resolution. Each vertical line represents a multi-way contact, in which the nodes indicate the corresponding locus' participation in this contact. (B) Frequencies of Pore-C contacts in Chromosome 22. Bars are colored according to the order of contact. Blue, green, orange, and red correspond to 2-way, 3-way, 4-way, and 5-way contacts. (C) The most common 2-way, 3-way, 4-way, and 5-way intra-chromosome contacts within Chromosome 22 are represented as motifs, color-coded similarly to B. (D) Zoomed in incidence matrix visualization in 100 kb resolution shows the multi-way contacts between three 1 Mb loci: L19 (blue), L21 (yellow), and L22 (red). An example 100 kb resolution multi-way contact is zoomed to read-level resolution. (E) Hypergraph representation of the 100 kb multi-way contacts from D. Blue, yellow, and red labels correspond to loci L19, L21, and L22, respectively.

Figure 6.3: (F) Incidence matrix visualization of the inter-chromosomal multi-way contacts between Chromosome 20 (orange) and Chromosome 22 (green) in 1 Mb resolution. Within this figure, all data are from one adult fibroblast sequencing run (V2) and multi-way contacts were determined after noise reduction at 1 Mb or 100 kb resolution accordingly (*Materials and Methods*).

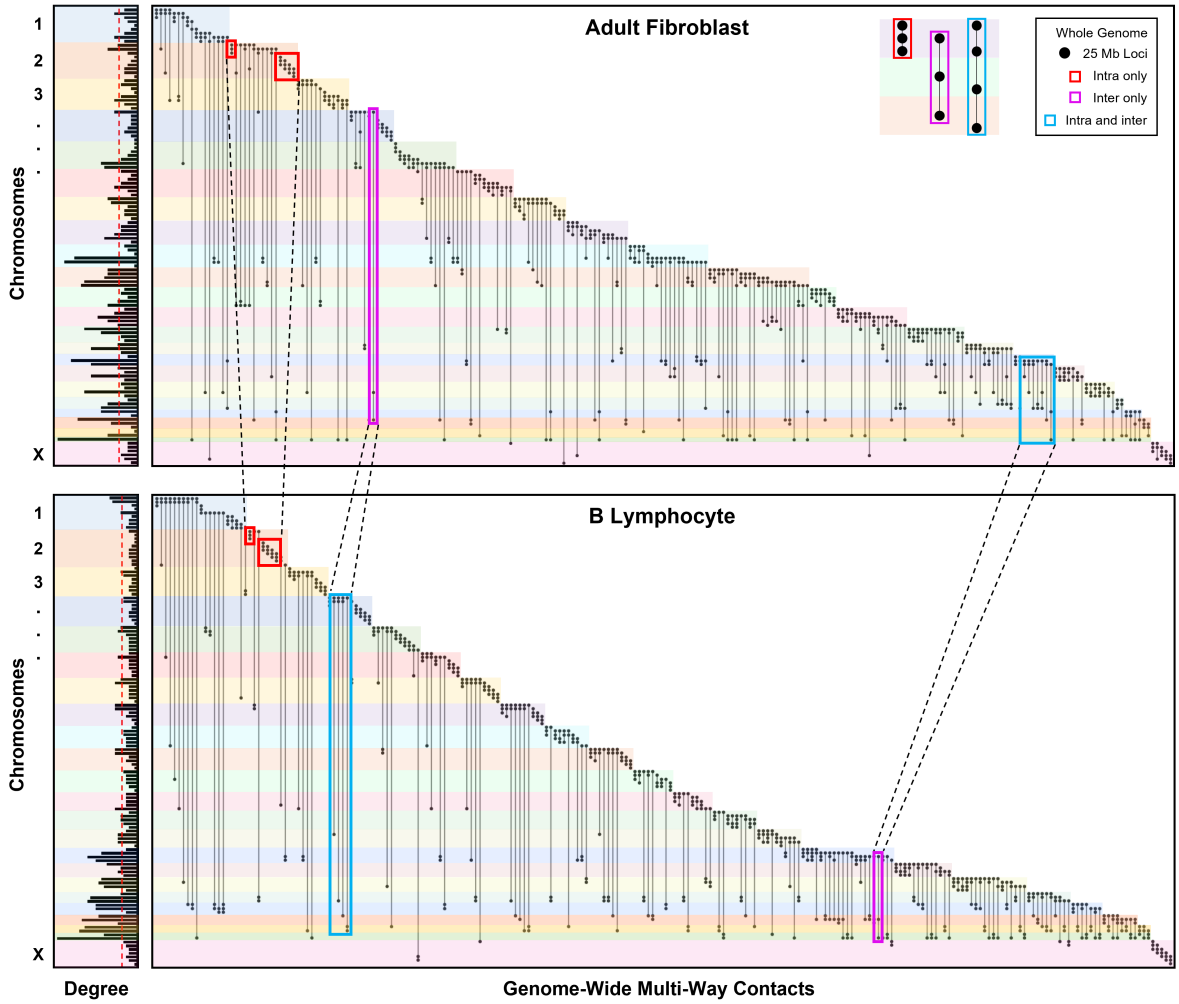

Figure 6.4: Genome-wide Patterning of Multi-way Contacts. Incidence matrix visualization of the top 10 most common multi-way contacts per chromosome. Matrices are constructed at 25 Mb resolution for both adult fibroblasts (top, V1-V4) and B lymphocytes (bottom). Specifically, 5 intra-chromosomal and 5 inter-chromosomal multi-way contacts were identified for each chromosome with no repeated contacts. If 5 unique intra-chromosomal multi-way contacts are not possible in a chromosome, they are supplemented with additional inter-chromosomal contacts. Vertical lines represent multi-way contacts, nodes indicate the corresponding locus' participation in a multi-way contact, and color-coded rows delineate chromosomes. Highlighted boxes indicate example intra-chromosomal contacts (red), inter-chromosomal contacts (magenta), and combinations of intra- and inter-chromosomal contacts (blue). Examples for each type of contact are shown in the top right corner.

Figure 6.4: Multi-way contacts of specific regions are compared between cell types by connecting highlighted boxes with black dashed lines, emphasizing similarities and differences between adult fibroblasts and B lymphocytes. Normalized degree of loci participating in the top 10 most common multi-way contacts for each chromosome in adult fibroblast and B lymphocytes are shown on the left. Red dashed lines indicate the mean degree for adult fibroblasts and B lymphocytes (top and bottom, respectively). Genomic loci that do not participate in the top 10 most common multi-way contacts for adult fibroblasts or B lymphocytes were removed from their respective incidence matrices and degree plots. Multi-way contacts were determined at 25 Mb resolution after noise reduction (*Materials and Methods*).

6.3.4 Transcription Clusters

We use the following definitions: **Transcription cluster**: a group of genomic loci that colocalize for efficient gene transcription. **Master regulator**: a self-regulating transcription factor that binds its gene analog. **Specialized transcription cluster**: a transcription cluster where at least one master regulator binds. **Self-sustaining transcription cluster:** a transcription cluster where a TF binds and its gene analog is expressed.

Genes are transcribed in short sporadic bursts in areas with high concentrations of transcriptional machinery [335, 73, 74], including transcriptionally engaged polymerase and accumulated transcription factors (TFs). Colocalization of multiple genomic loci in these areas may facilitate more efficient transcription, which is supported by studies using fluorescence *in situ* hybridization (FISH) that show colocalization during active transcription [247]. Simulations also provide evidence that genomic loci that are bound by common transcription factors can self-assemble into clusters, forming structural patterns commonly observed in Hi-C data [74]. We refer to these instances of highly concentrated areas of transcriptional machinery and genomic loci as *transcription clusters*. The colocalization of multiple genomic loci naturally leads to multi-way contacts, but these interactions cannot be fully captured from the pairwise contacts of Hi-C. Multi-way contacts derived from Pore-C reads can detect interactions between many genomic loci, and are well suited for identifying potential transcription clusters (Figure 6.6).

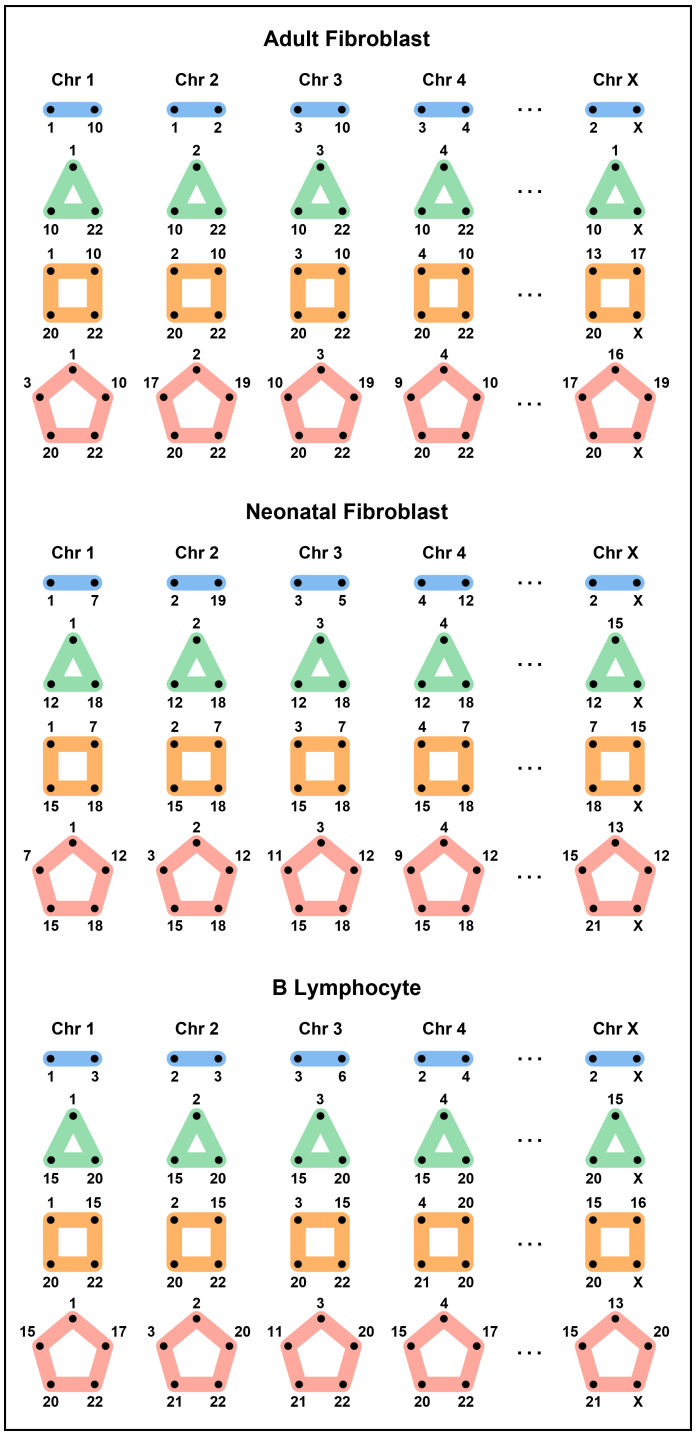

Figure 6.5: Inter-chromosomal Interactions. The most common 2-way, 3-way, 4-way, and 5-way inter-chromosome combinations for each chromosome are represented using motifs from adult fibroblasts (top), neonatal fibroblasts (center), and B lymphocytes (bottom). Rows represent the combinations of 2-way, 3-way, 4-way, and 5-way inter-chromosomal interactions, and columns are the chromosomes. Inter-chromosomal combinations are determined using 25 Mb resolution multi-way contacts after noise reduction (*Materials and Methods*) and are normalized by chromosome length. Here we only consider unique chromosome instances (i.e. multiple loci in a single chromosome are ignored).

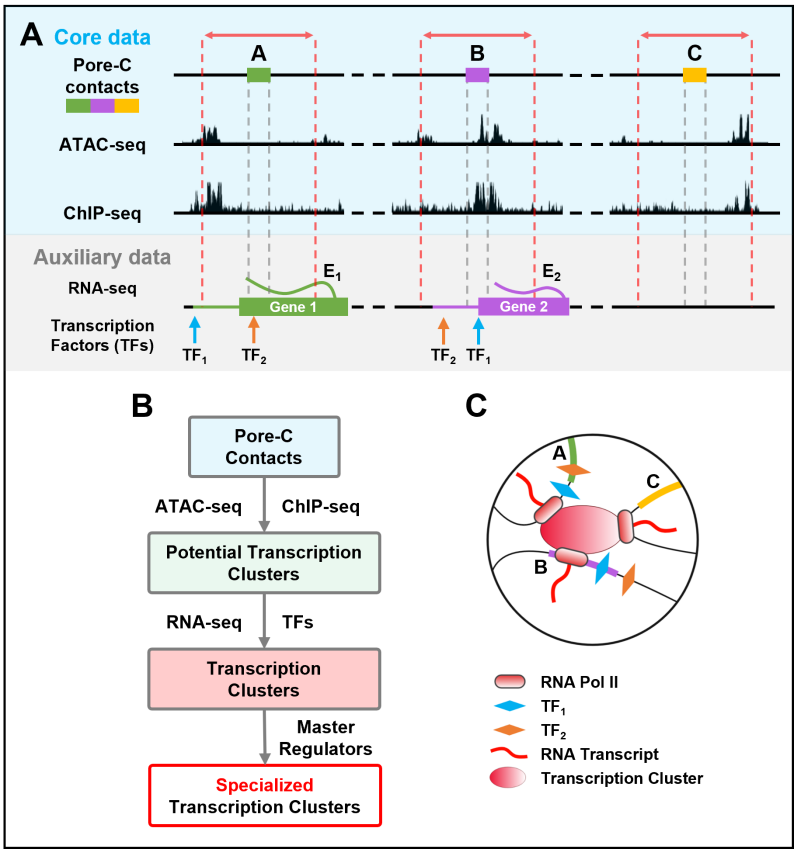

Figure 6.6: Data-driven Identification of Transcription Clusters. (A) Blue shaded area: A 5 kb region before and after each locus in a Pore-C read (region between red dashed lines) is queried for chromatin accessibility and RNA Pol II binding (ATAC-seq and ChIP-seq, respectively). Multi-way contacts between accessible loci that have ≥ 1 instance of RNA Pol II binding are indicative of potential transcription clusters. Gray shaded area: Gene expression (RNA-seq, E_1 for gene 1 and E_2 for gene 2, respectively) and transcription factor binding sites (TF_1 and TF_2) are integrated to determine potential coexpression and coregulation within multi-way contacts with multiple genes. Transcription factor binding sites are queried $\pm$ 5 kb from the gene's transcription start site (*Materials and Methods*). Genes are colored based on the overlapping Pore-C locus, and the extended horizontal line from each gene represents the 5 kb flanking region used to query transcription factor binding sites. (B) Pipeline for extracting transcription clusters (*Supplementary Materials*). (C) Schematic representation of a transcription cluster.

As actively transcribed regions are accessible and transcription is carried out by locally-bound RNA Pol II [169], we require all loci in a candidate transcription cluster to be accessible, evidenced by ATAC-seq, and at least one of those loci to contain an RNA Pol II ChIP-seq signature. Using these criteria, we identified 12,364, 16,080, and 16,527 potential transcription clusters from neonatal fibroblasts, adult fibroblasts, and B lymphocytes, respectively (Table 6.1, *Materials and Methods*). The majority of these clusters involved at least one expressed gene (94.2% in neonatal fibroblasts, 95.0% in adult fibroblasts, 90.5% in B lymphocytes) as well as at least two expressed

genes (69.6% in neonatal fibroblasts, 71.9% in adult fibroblasts, 58.7% in B lymphocytes). While investigating the colocalization of expressed genes in transcription clusters, we found that over half of clusters containing multiple expressed genes had common transcription factors based on binding motifs in fibroblasts (61.9% in neonatal fibroblasts, 65.2% in adult fibroblasts) and that over half of these common transcription factors were master regulators (55.9% in neonatal fibroblasts, 63.4% in adult fibroblasts). These proportions were slightly lower in B lymphocytes where we observed that 50.0% of clusters containing multiple expressed genes had common transcription factors while 46.8% of these common transcription factors were master regulators. Example transcription clusters derived from 3-way, 4-way, and 5-way contacts in fibroblasts and B lymphocytes are shown in Figure 6.7. Transcription clusters contained at least two expressed genes with at least one common transcription factor binding motif.

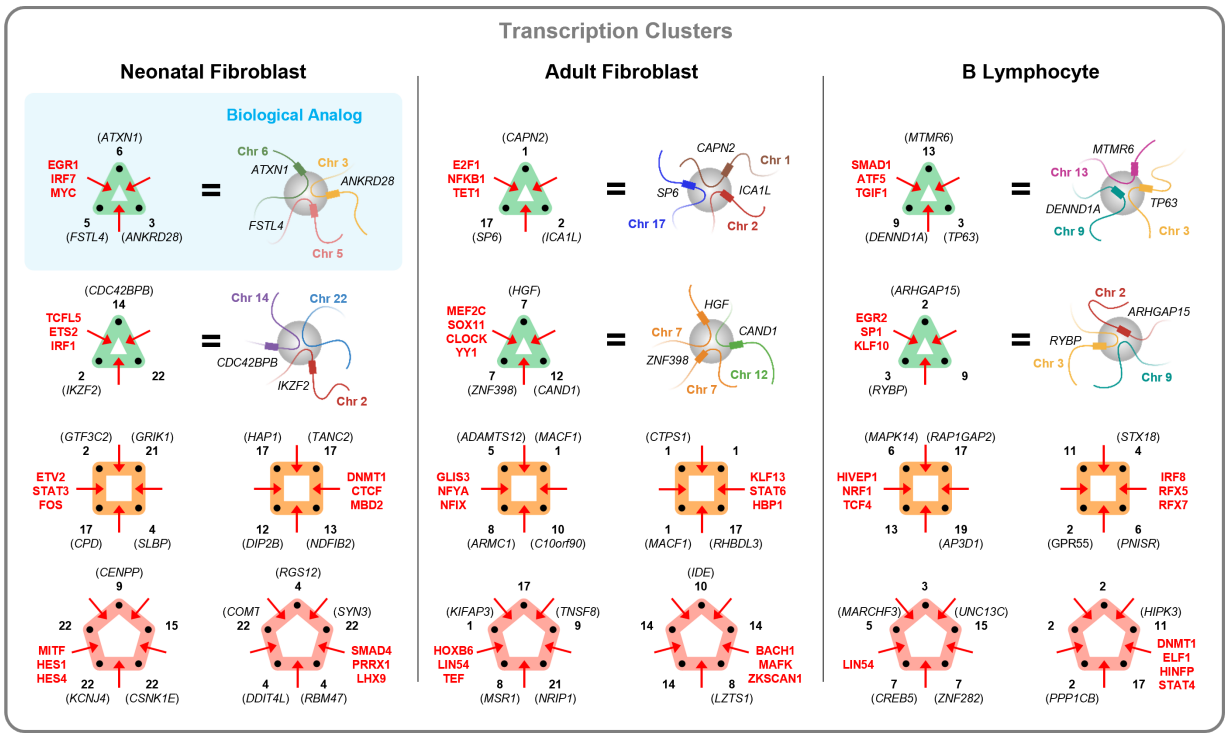

Figure 6.7: Example Transcription Clusters. Six examples of transcription clusters are shown for neonatal fibroblasts (left), adult fibroblasts (center), and B lymphocytes (right) as multi-way contacts (hypergraph motifs). Black labels indicate genes and chromosomes (bold). Red labels correspond to transcription factors shared between the majority of genes within the transcription cluster. For three-way contacts (green motifs), we highlight the transcription clusters' biological analog (blue-shaded box), showing how fragments of chromatin fold and congregate at a common transcription cluster (grey sphere). Each node (black dot) of the hyperedge and its denoted chromosome and gene in the hypergraph motif corresponds to a single chromatin fragment, colored according to chromosome, in the biological analog. Thus, a three-way hyperedge is depicted by three chromatin fragments in close spatial proximity. Multi-way contacts used for adult and neonatal fibroblasts include all experiments (V1-V4). Examples were selected from the subset of multi-way contacts summarized in the "Clusters with Common TFs" column of Table 6.1.

We tested the criteria for potential transcription clusters for statistical significance (*Materials and Methods)*. That is, we tested whether the identified transcription clusters are more likely to include genes, and if these genes were more likely to share common transcription factors, than random multi-way contacts. We found that the identified transcription clusters were significantly more likely to include ≥ 1 gene and ≥ 2 genes than random multi-way contacts ($p < 0.01$). In addition, transcription clusters containing ≥ 2 genes were significantly more likely to have transcription factors and master regulators in common ($p < 0.01$). After testing all orders of multi-way transcription clusters together, we also tested the 3-way, 4-way, 5-way, and 6-way (or more) cases individually. We found that all cases were statistically significant ($p < 0.01$) except for clusters

with common transcription factors or master regulators in the 6-way (or more) case for both fibroblasts and B lymphocytes. We hypothesize that these cases were not statistically significant due to the large number of loci, naturally leading to an increased overlap with genes. This increases the likelihood that at least two genes will have common transcription factors or master regulators. Over half of the transcription clusters where the majority of genes contained common transcription factors also contained at least one enhancer in adult fibroblasts (98.2%) and B lymphocytes (87.9%), further suggesting regulatory function within these multi-way contacts [49, 318] (Table S2 in Dotson *et al.* [99]). In contrast, only 11.6% of transcription clusters in neonatal fibroblasts exhibited the same properties, which may be a factor of the significantly sparser enhancer annotation available for this cell type compared to the others from the EnhancerAtlas 2.0 database.

To understand how our transcription clusters aligned with factors that have known involvement in chromatin architecture and transcriptional regulation, we evaluated CCCTC-binding factor (CTCF) for binding using ChIP-seq data. CTCF specifically mediates chromatin looping and TAD boundary insulation [257, 246] and binds generously throughout the genome [274]. We found significantly higher CTCF binding in our identified transcription clusters compared to multi-way contacts that were not classified as transcription clusters (Figure S3 in Dotson *et al.* [99]). In adult and neonatal fibroblasts, CTCF binding was nearly two-fold greater in transcription clusters compared to randomly selected multi-way contacts (80% vs. 45% and 81% vs. 47%). In B lymphocytes, CTCF binding was present at 82% of transcription clusters compared to 59% of random multi-way contacts. We additionally investigated cohesin for its colocalization with CTCF and involvement in regulation of chromatin architecture, such as chromatin loop extrusion [379, 252, 4, 129]. In particular, the cohesin subunits RAD21 and SMC3 have been previously linked to CTCF-mediated transcriptional regulation [274]. ChIP-seq data showed preferential binding of RAD21 and SMC3 at transcription clusters compared to random multi-way contacts in adult fibroblasts (79% vs. 42% for RAD21, 71% vs. 37% for SMC3, $p < 0.01$) and B lymphocytes (76% vs. 55% for RAD21, 79% vs. 53% for SMC3, $p < 0.01$) (Figure S3 in Dotson *et al.* [99]). Together these data suggest that the identified transcription clusters are important sites of transcriptional regulation, and support a model in which CTCF and cohesin actively mediate multi-way interactions.

We next sought to determine which TFs might be involved in cell type-specific regulation in transcription clusters. For each cell type, we ranked expressed TFs by frequency of binding sites across transcription clusters. Among TFs with the most frequent binding sites, 39% were shared across all three cell types, compared to 72% between adult and neonatal fibroblasts (Table S6 in Dotson *et al.* [99]). Fibroblast and B lymphocyte TF binding sites had less overlap, at 52% (adult) and 45% (neonatal), than binding sites between fibroblasts, supporting cell type-specific regulation of transcription cluster subsets. Of 18 TFs whose binding sites were unique to the transcription clusters of neonatal fibroblasts, the most frequently occurring was RARB, found at

10.2% of clusters, while in adult fibroblasts, binding sites for ZNF667 were the most frequent among 14 TFs at 6.7% of transcription clusters. In B lymphocytes, binding sites for TFEC were the most frequent among 161 TFs at 7.6% of transcription clusters. Prior studies support the cell type-specific roles of these uniquely-binding TFs in fibroblasts and B lymphocytes [351, 292].

Given the role of TFs in coordinating transcription among clusters of genes [303, 400], we hypothesized that TF loci might feature in a subset of transcription clusters. To investigate this question, we looked for the binding motif and encoding gene locus for a given TF within the same transcription cluster, defining this class as a self-sustaining transcription cluster (Figure 6.8A-B). We identified nine, eight, and thirteen self-sustaining transcription clusters in adult fibroblasts, neonatal fibroblasts, and B lymphocytes, respectively (Table S7 in Dotson *et al.* [99]). In adult fibroblasts, we observed that the binding motif for FOXO3, a master regulator, exists at a 4-way transcription cluster expressing the *FOXO3* gene. The neonatal fibroblast and B lymphocyte datasets had a self-sustaining transcription cluster in common where STAT3 had a binding motif and the *STAT3* gene was expressed. While self-sustaining transcription clusters demonstrate the capacity for a TF to regulate itself, not every TF co-occupying a transcription cluster with its gene analog is classified as a master regulator (Table S7 in Dotson *et al.* [99]). Therefore, we further stratify these clusters into self-sustaining transcription clusters where the TF is a master regulator and thus binds its gene analog (stronger coupling) and self-sustaining transcription clusters where the TF binds in the cluster but not at is gene analog (weaker coupling). We propose that these strongly-coupled self-sustaining transcription clusters are 'core' transcription clusters that serve as transcriptional signatures for a cell type. It also follows that strongly-coupled self-sustaining transcription clusters are specialized transcription clusters (Figure S4 in Dotson *et al.* [99]). We then considered two classes of analog-independent transcription clusters - where either a TF and its gene analog occupy different clusters (Figure 6.8C) or a TF occupies a cluster, but its gene analog occupies no cluster (Figure 6.8D). Since both the TF and gene analog belong to transcription clusters in Figure 6.8C, they are coupled, though lesser so than either class of self-sustaining transcription clusters. In contrast, Figure 6.8D represents an architecturally uncoupled state - 23.3%, 25.7%, and 40.1% of TF gene analogs in adult fibroblasts, neonatal fibroblasts, and B lymphocytes, respectively, were not expressed in any transcription cluster. Lastly, we binned all multi-way contact loci involved in self-sustaining transcription clusters at 100 kb resolution and plotted the interaction frequencies of their decomposed pairwise components (Figure 6.8E).

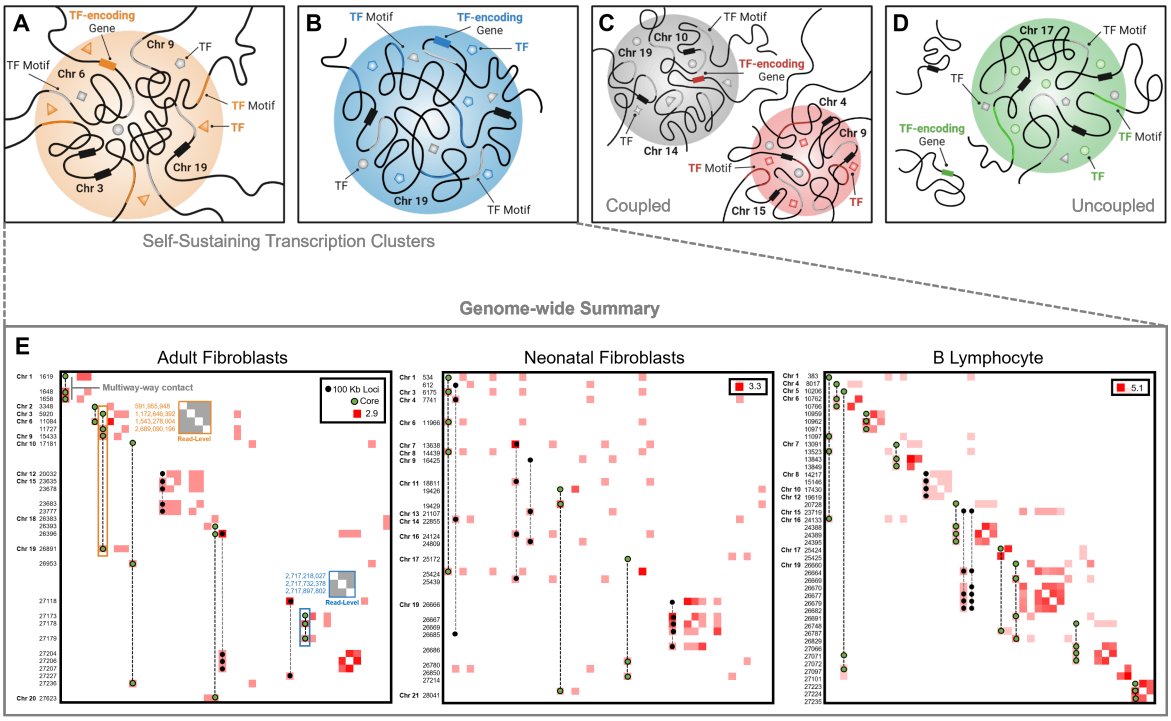

Figure 6.8: Classes of Transcription Clusters. In a self-sustaining transcription cluster, a TF and the gene encoding that TF are both present. The inter- and intra-chromosomal examples in (A) and (B), respectively, illustrate this phenomenon where in (A) we see the TF of interest (orange triangle) circulating at the cluster, its binding motif present on the chromatin (orange portion), and its corresponding gene expressed (orange rectangle on Chromosome 6). The gray shapes represent additional TFs with binding motifs (gray portion of chromatin) at the cluster. Black rectangles on Chromosomes 3, 9, and 19 represent additional genes present in the cluster. (C) An analog-independent class of transcription clusters where we observe a TF (red square) bind at a transcription cluster (red cluster) and its corresponding gene expressed in a separate transcription cluster (grey cluster), yet not in the same cluster. (D) An analog-independent class of transcription clusters where we observe a TF (green circle) bind at a transcription cluster (green cluster) and its corresponding gene expressed but not within a transcription cluster. (E) Genome-wide cell type-specific self-sustaining transcription clusters extracted from multi-way contact data and decomposed into Hi-C contact matrices at 100 kb resolution. Contact frequencies are log-transformed for better visualization. Frequencies along the diagonal indicate interaction between two or more unique multi-way loci that fall within the same 100 kb bin. Axis labels are non-contiguous 100 kb bin coordinates in chromosomal order. Multi-way contacts that make up the self-sustaining transcription clusters are superimposed. Multi-way contacts with green-colored loci represent 'core' transcription clusters - transcription clusters containing a master regulator and its gene analog. An example read-level contact map for the inter-chromosomal FOXO3 self-sustaining transcription cluster is denoted by the orange highlighted box in the adult fibroblast contact matrix and a read-level contact map for the intra-chromosomal ZNF320 self-sustaining transcription cluster is denoted by the blue highlighted box. Values along the left axis of these read-level contact matrices are base-pair positions of the contacting loci in the genome.

Order	Multi-way Contacts	Transcription Clusters	Clusters with $\geq$ 1 Gene	Clusters with $\geq$ 2 Genes	Clusters with Common TFs	Clusters with Common MRs
3	240,477	8,384	7,384	4,157	3,788	3,645
	301,366	8,182	7,615	5,208	4,839	4,439
	379,165	11,261	10,581	7,518	6,890	6,778
4	227,352	4,345	3,972	2,686	2,435	2,341
	156,742	2,593	2,467	2,008	1,868	1,729
	181,554	3,254	3,159	2,658	2,515	2,468
5	196,423	1,996	1,881	1,434	1,103	957
	98,172	999	976	834	572	443
	98,272	1,021	999	877	688	614
6+	1,000,231	1,802	1,727	1,419	932	783
	178,705	590	583	549	368	303
	142,575	544	542	514	395	340

Table 6.1: Summary of Multi-way Contacts and Transcription Clusters. Multi-way contacts from B lymphocytes (white rows), neonatal fibroblasts (light gray rows), and adult fibroblasts (dark gray rows, V1-V4) are listed after different filtering criteria. Multi-way contacts are considered to be potential transcription clusters if all loci within the multi-way contact are accessible and at least one locus binds RNA Pol II. These multi-way contacts are then queried for nearby expressed genes. If a transcription cluster candidate has at least two expressed genes, we determine whether the majority of these genes have common transcription factors (TFs) through binding motifs. If only two expressed genes are contained within a transcription cluster candidate, we require both genes to have common TFs. From the set of transcription clusters with common TFs, we calculate how many clusters have at least one common master regulator (MR) (Algorithm 3 in Dotson *et al.* [99]).

Algorithm 1: Multi-way Contact Analysis

Input: Aligned Pore-C data (**A**), RNA-seq (**R**: gene expression), RNA Pol II (**P**: ChIP-seq), ATAC-seq (**C**: chromatin accessibility), transcription factor binding motifs (**B**)

for each set of Pore-C data $\mathbf{A}_l \in \mathbf{A}$ **do**

    Construct incidence matrix $\mathbf{H}_l$ using Algorithm 2 in Dotson *et al.* [99]

    Identify transcription clusters $\mathbf{T}_{lp}$, $\mathbf{T}_{lc}$, and $\mathbf{T}_{ls}$ using Algorithm 3 in Dotson *et al.* [99]

    Calculate entropy S_l using Algorithm 4 in Dotson *et al.* [99]

end for

Compute hypergraph distance d_{ij} between pairs $\mathbf{H}_i$ and $\mathbf{H}_j$ with $p \geq 1$ using Algorithm 5 in Dotson *et al.* [99]

Calculate the statistical significance α_{ij} for hypergraph distance d_{ij} using the permutation test in Algorithm 6 in Dotson *et al.* [99].

Return: Hypergraph incidence matrices $\mathbf{H}_l \in \mathbb{R}^{n \times m}$, hypergraph entropy S_l, potential transcription clusters $\mathbf{T}_{lp}$, transcription clusters $\mathbf{T}_{lc}$, specialized transcription clusters $\mathbf{T}_{ls}$, and hypergraph distance matrix $[d_{ij}]$ with statistical significance $[\alpha_{ij}]$.

6.4 Discussion

In this work, we introduce a hypergraph framework to study higher-order genome organization from Pore-C long-read sequence data. We demonstrate that higher-order genome architecture can be precisely represented and analyzed using hypergraph theory. Using direct capture of multi-way contacts, we identified transcription clusters with physical proximity and coordinated gene expression. Our framework thus enables study of explicit structure-function relationships that are observed directly from data, without needing to infer multi-way contacts. In engineering and social systems, hypergraph representation of data has revealed higher-order organization principles efficiently [26, 60, 58, 334]. Our work here extends the application of hypergraphs, demonstrating a natural way to represent and analyze genome organization across scales.

Exploring long-range, inter-chromosomal interactions genome-wide offers the opportunity to establish fundamental principles of genome organization. Unbiased capture and study of multi-way contacts can help identify biologically important assemblies that affect transcription, such as transcription clusters [74, 267]. This approach can also connect genome organization principles to the study of transcription factors and how they govern cell type-specific network architecture, which resembles small world phenomena [36, 327]. Our results support the idea of cell type-specific formation of transcription clusters that serve as a basis for efficient navigation of information within the nucleus, and thereby reflect a signature of small world architecture. Analogous to the behavior

of short-path information propagation in social networks, we posit that transcription clusters act as decentralized nodes, or critical architectures relevant to cell identity. Future work to explore these phenomena systematically will undoubtedly help us understand cell type-specific organization principles. Another exciting direction will be to investigate time series multi-way interactions during cellular transitions such as differentiation and cell reprogramming, with single cell observations. Furthermore, we imagine that multi-way chromatin structure together with spatial transcriptomics will guide us to uncover formation principles in tissue patterning and organogenesis [273, 125].

6.5 Materials and Methods

6.5.1 Cell Culture

Primary human adult dermal fibroblasts were obtained from a donor and were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1X Glutamax (Thermo Fisher Scientific Cat no. 35050061) and 1X nonessential amino acid (Thermo Fisher Scientific Cat no. 11140050).

6.5.2 Cross-linking

Protocols for cross-linking were based on Ulahannan et al. [93]. 2.5 million cells were washed three times in chilled 1X phosphate buffered saline (PBS) in a 50 centrifuge tube, pelleted by centrifugation at 500 x g for 5 min at 4 between each wash. Cells were resuspended in 10 room temperature 1X PBS 1% formaldehyde (Fisher Scientific Cat no. BP531-500) by gently pipetting with a wide bore tip, then incubated at room temperature for 10 min. To quench the cross-linking reaction 527 of 2.5 M glycine was added to achieve a final concentration of 1% w/v or 125 mM in 10.5. Cells were incubated for 5 min at room temperature followed by 10 min on ice. The cross-linked cells were pelleted by centrifugation at 500 x g for 5 min at 4.

6.5.3 Restriction Enzyme Digest

The cell pellet was resuspended in 500 of cold permeabilization buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630, 100 of protease inhibitor cock-tail Roche Cat no. 11836170001) and placed on ice for 15 min. One tablet of protease inhibitor cocktail was dissolved in 1 ml nuclease free water and 100 from that was added to a 500 permeabilization buffer. Cells were centrifuged at 500 x g for 10 min at 4 after which the supernatant was aspirated and

replaced with 200 of chilled 1.5X New England Biolabs (NEB) cutsmart buffer. Cells were centrifuged again at 500 x g for 10 min at 4, then aspirated and re-suspended in 300 of chilled 1.5X NEB cutsmart buffer. To denature the chromatin, 33.5 of 1% w/v sodium dodecyl sulfate (SDS, Invitrogen Cat no. 15553-035) was added to the cell suspension and incubated for exactly 10 min at 65 with gentle agitation then placed on ice immediately afterwards. To quench the SDS, 37.5 of 10% v/v Triton X-100 (Sigma Aldrich Cat no. T8787-250) was added for a final concentration of 1%, followed by incubation for 10 min on ice. Permeabilized cells were then digested with a final concentration of 1 U/ of NlaIII (NEB-R0125L) and brought to volume with nuclease-free water to achieve a final 1X digestion reaction buffer in 450. Cells were then mixed by gentle inversion. Cell suspensions were incubated in a thermomixer at 37 for 18 hours with periodic rotation.

6.5.4 Proximity Ligation and Reverse Cross-linking

NlaIII restriction digestion was heat inactivated at 65 for 20 min. Proximity ligation was set up at room temperature with the addition of the following reagents: 100 of 10X T4 DNA ligase buffer (NEB), 10 of 10 mg/mL BSA and 50 of T4 Ligase (NEB M0202L) in a total volume of 1000 with nuclease-free water. The ligation was cooled to 16 and incubated for 6 hours with gentle rotation.

6.5.5 Protein Degradation and DNA Purification

To reverse cross-link, proximity ligated sample was treated with 100 Proteinase K (NEB P8107S-800U/ml), 100 10% SDS (Invitrogen Cat no. 15553-035) and 500 20% v/v Tween-20 (Sigma Aldrich Cat no. P1379) in a total volume of 2000 with nuclease-free water. The mixture was incubated in a thermal block at 56 for 18 hours. In order to purify DNA, the sample was transferred to a 15 centrifuge tube, rinsing the original tube with a further 200 of nuclease-free water to collect any residual sample, bringing the total sample volume to 2.2. DNA was then purified from the sample using a standard phenol chloroform extraction and ethanol precipitation.

6.5.6 Nanopore Sequencing

Purified DNA was Solid Phase Reversible Immobilization (SPRI) size selected before library preparation with a bead ratio of 0.48X for fragments $>$ 1.5 kb. The $>$ 1.5 kb products were prepared for sequencing using the protocol provided by Oxford Nanopore Technologies. In brief, 1 of genomic DNA input was used to generate a sequencing library according to the protocol provided for the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford Science Park, UK, version GDE_9063_v109_revU_14Aug2019). After the DNA repair, end prep, and adapter ligation steps, SPRI select bead suspension (Cat No. B23318, Beckman Coulter Life Sciences, Indianapolis, IN,

USA) was used to remove short fragments and free adapters. A bead ratio of 1X was used for DNA repair and end prep while a bead ratio of 0.4X was used for the adapter ligation step. Qubit dsDNA assay (ThermoFisher Scientific, Waltham, MA, USA) was used to quantify DNA and ∼300-400 ng of DNA library was loaded onto a GridION flow cell (version R9, Flo-MIN 106D). For adult fibroblasts, 4 sequencing runs were conducted generating a total of 6.25 million reads (referred to as V1-V4). For neonatal fibroblasts, 4 sequencing runs were conducted generating a total of 11.85 million reads.

6.5.7 Sequence Processing.

Reads which passed Q-score filtering (`--min_qscore 7`, 4.56 million reads) after base calling on the Oxford Nanopore GridION (Guppy, version 4.0.11) were used as input for the Pore-C-Snakemake pipeline (https://github.com/nanoporetech/Pore-C-Snakemake, commit 6b2f762). The pipeline maps multi-way contacts to a reference genome and stores the hyperedge data in a variety of formats. The reference genome used for mapping was GRCh38.p13 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/). Each of the four sequencing runs were assigned a sequencing run label and then concatenated. The combined pipeline outputs were used as standard inputs for all downstream analysis.

6.5.8 Hypergraphs

A hypergraph is a generalization of a graph. Hypergraphs are composed of hyperedges, which can join any number of nodes [27]. Mathematically, a hypergraph is a pair such that $\mathsf{G} = \{\mathbf{V}, \mathbf{E}\}$ where $\mathbf{V}$ is the node set and $\mathbf{E}$ is the hyperedge set. Each hyperedge in $\mathbf{E}$ is a subset of $\mathbf{V}$. Examples of hypergraphs include email communication networks, co-authorship networks, film actor/actress networks, and protein-protein interaction networks. For genomic networks, traditional graph-based methods fail to capture contacts that contain more than two genomic loci once, which results in a loss of higher order structural information. Hypergraphs can capture higher order connectivity patterns and represent multidimensional relationships unambiguously in genomic networks [26, 383]. In hypergraphs obtained from Pore-C data, we defined nodes as genomic loci at a particular resolution (e.g. read level, 100 kb, 1 Mb, or 25 Mb bins), and hyperedges as contacts among genomic loci. We switch between these different resolutions both for computational efficiency and visual clarity. Most higher order contacts are unique in Pore-C data at high resolution (read level or 100 kb), so for these data we considered unweighted hypergraphs (i.e. ignore the frequency of contacts). For lower resolutions (1 Mb or 25 Mb), we considered edge weights (frequency of contacts) to find the most common intra- and inter-chromosomal contacts.

6.5.9 Hypergraph Filtering

We performed an additional filtering step while constructing genomic hypergraphs. We first decomposed each multi-way contact into its pairwise combinations at a particular resolution. From these pairwise contacts, we counted the number of times a contact was detected for each pair of loci, and identified the highest frequency locus pairs. Pairwise contacts were kept if they occurred above a certain threshold number, which was set empirically at the 85^{th} percentile of the most frequently occurring locus pairs. For example, in fibroblast data binned at 1 Mb resolution, a locus pair with six detected contacts corresponded to the 85^{th} percentile. Thus all pairs of loci with fewer than six detected contacts were not considered, which increases confidence in the validity of identified multi-way contacts.

6.5.10 Incidence Matrices

An incidence matrix of the genomic hypergraph was an n-by-m matrix containing values zero and one. The row size n was the total number of genomic loci, and the column size m was the total number of unique Pore-C contacts (including self-contacts, pairwise contacts, and higher order contacts). Nonzero elements in a column of the incidence matrix indicate genomic loci contained in the corresponding Pore-C contact. Thus, the number of nonzero elements (or column sum) gives the order of the Pore-C contact. The incidence matrix of the genomic hypergraph can be visualized via PAOHvis [359]. In PAOHvis, genomic loci are parallel horizontal bars, while Pore-C contacts are vertical lines that connect multiple loci (see Figures 6.1, 6.2, 6.3, and 6.4). Beyond visualization, incidence matrices play a significant role in the mathematical analysis of hypergraphs.

6.5.11 Data-driven Identification of Transcription Clusters

We used Pore-C data in conjunction with multiple other data sources to identify potential transcription clusters (Figure 6.6). Each locus in a Pore-C read, or multi-way contact, was queried for chromatin accessibility and RNA Pol II binding (ATAC-seq and ChIP-seq peaks, respectively). Multi-way contacts were considered to be potential transcription clusters if all loci within the multi-way contact were accessible and at least one locus had binding of RNA Pol II. The loci in potential transcription clusters were then queried for nearby expressed genes. A 5 kb flanking region was added upstream and downstream of each locus when querying for chromatin accessibility, RNA Pol II binding, and nearby genes [239]. Gene expression (RNA-seq) and transcription factor binding site data were then integrated to determine coexpression and coregulation of genes in potential transcription clusters. If a potential transcription cluster candidate had at least two genes

present, and these genes had common transcription factors based on binding motifs in the cluster, the potential transcription cluster was determined to be a real transcription cluster. From the set of transcription clusters with common transcription factors, we calculated how many clusters were regulated by at least one master regulator, a transcription factor that also regulates its own gene, and classified these as specialized transcription clusters (Figure 6.6).

6.5.12 Transcription Factor Binding Motifs

Transcription factor binding site motifs were obtained from "The Human Transcription Factors" database [177]. FIMO (https://meme-suite.org/meme/tools/fimo) was used to scan for motifs within $\pm$ 5 kb of genes' transcription start sites. The results were converted to a 22,083 $\times$ 1,007 MATLAB table, where rows were genes, columns were transcription factors, and entries were the number of binding sites for a particular transcription factor and gene. The table was then filtered to only include entries with three or more binding sites in downstream computations. This threshold was determined empirically and is adjustable in the MATLAB code.

6.5.13 Identifying Self-Sustaining Transcription Clusters

From identified transcription clusters (Table 6.1), we obtained a subset containing TF-encoding genes specific to each cell type, yielding 79, 54, and 144 transcription clusters from the adult fibroblast, neonatal fibroblast, and B lymphocyte data, respectively. We then classified these clusters as self-sustaining if the TF binding motif corresponding to the expressed TF-encoding gene was also at the cluster. We further determined whether the self-sustaining TFs were master regulators based on protein-DNA interaction data. Results are summarized in Figure 6.8 and Table S7.

6.5.14 Public Data Sources

Pore-C data for B lymphocytes were downloaded from Ulahannan *et al.* [93]. ATAC-seq and ChIP-seq data were obtained from the Encyclopedia of DNA Elements (ENCODE) to assess chromatin accessibility and RNA Pol II binding, respectively. These data were compared to read-level Pore-C contacts to determine whether colocalizing loci belong to accessible regions of chromatin and had RNA Pol II binding for both fibroblasts and B lymphocytes. RNA-seq data were also obtained from ENCODE to ensure that genes within potential transcription clusters were expressed in their respective cell types. Additionally, ChIP-seq data for CTCF, RAD21, and SMC3 binding were obtained from ENCODE to evaluate binding preference at transcription clusters. A summary of these data sources can be found in Table 6.2.

Data Type	Cell Type	Data Description and Source
Pore-C	IR	Human adult primary dermal fibroblasts were derived from a donor skin biopsy
Pore-C	BJ	Human foreskin fibroblasts from BJ cell line (Cat no. CRL-2522, ATCC, Manassas, VA)
Pore-C	GM12878	B lymphocyte Pore-C data obtained from Ulahannan *et al*. [93]
ATAC-seq	IMR-90	Adult fibroblast chromatin accessibility (ENCFF310UDS)
DNase-seq	BJ	Foreskin fibroblast chromatin accessibility (ENCFF310UDS)
ATAC-seq	GM12878	B lymphocyte chromatin accessibility data (ENCFF410XEP)
ChIP-seq	IMR-90	Adult fibroblast RNA Polymerase II binding data (ENCFF676DGR)
ChIP-seq	GM12878	B lymphocyte RNA Polymerase II binding data (ENCFF912DZY)
ChIP-seq	IMR-90	Adult fibroblast CTCF binding data (ENCFF203SRF)
ChIP-seq	BJ	Human foreskin fibroblasts CTCF binding data (ENCFF518RUC)
ChIP-seq	GM12878	B lymphocyte CTCF binding data (ENCFF951PEM)
ChIP-seq	IMR-90	Adult fibroblast RAD21 binding data (ENCSR000EFJ)
ChIP-seq	GM12878	B lymphocyte RAD21 binding data (ENCSR000EAC)
ChIP-seq	IMR-90	Adult fibroblast SMC3 binding data (ENCSR000HPG)
ChIP-seq	GM12878	B lymphocyte SMC3 binding data (ENCSR000DZP)
RNA-seq	IMR-90	Adult fibroblast gene expression data averaged over two replicates (ENCFF353SBP, ENCFF496RIW)
RNA-seq	IR	Adult fibroblast primary gene expression data
RNA-seq	BJ	Neonatal fibroblast gene expression data averaged over two replicates (ENCFF477JDG, ENCFF005WBQ)
RNA-seq	BJ	Foreskin fibroblast primary gene expression data on BJ cell line
RNA-seq	GM12878	B lymphocyte gene expression data averaged over two replicates (ENCFF306TLL, ENCFF418FIT)
Enhancers	IMR-90	Adult fibroblast enhancer location data from EnhancerAtlas 2.0 [115]
Enhancers	BJ	Foreksin fibroblast enhancer location data from EnhancerAtlas 2.0 [115]
Enhancers	GM12878	B lymphocyte enhancer location data from EnhancerAtlas 2.0 [115]

Table 6.2: Data Sources. Data obtained from ENCODE unless otherwise specified [71].

6.5.15 Hypergraph Entropy

Network entropy is often used to measure the connectivity and regularity of a network [58, 254, 218]. We defined a new notion of hypergraph entropy to quantify the organization of chromatin structure from Pore-C data. Denote the incidence matrix of the genomic hypergraph as $\mathbf{H}$. The Laplacian matrix is then an n-by-n matrix (n is the total number of genomic loci in the hypergraph), which can be computed by

$$\mathbf{L} = \mathbf{D} - \mathbf{H}\mathbf{E}^{-1}\mathbf{H}^{\top} \in \mathbb{R}^{n\times n}, \tag{6.1}$$

where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the degrees of nodes along its diagonal, and $\mathbf{E} \in \mathbb{R}^{m\times m}$ is a diagonal matrix containing the orders of hyperedges along its diagonal. Inspired by von Neumann graph entropy (which utilizes the distribution of the eigenvalues from the graph

Laplacian matrix), we define the hypergraph entropy as

$$\textbf{Hypergraph Entropy} = -\sum_{i=1}^{n} \bar{\lambda}_i \ln \bar{\lambda}_i, \tag{6.2}$$

where $\bar{\lambda}_i$ are the normalized eigenvalues of $\mathbf{L}$ such that $\sum_{i=1}^{n} \bar{\lambda}_i = 1$, and the convention $0 \ln 0 = 0$ is used. In mathematics, eigenvalues can quantitatively represent different features of a matrix [326]. Biologically, genomic regions with high entropy are likely associated with high proportions of euchromatin (i.e. less organized folding patterns), as euchromatin is more structurally permissive than heterochromatin [206, 268, 193].

We computed the entropy of intra-chromosomal genomic hypergraphs for both fibroblasts and B lymphocytes as shown in Table S8 in Dotson *et al.* [99]. It is expected that larger chromosomes have larger hypergraph entropy because more potential genomic interactions occur in the large chromosomes. However, there are still subtle differences between the fibroblast and B lymphocyte chromosomes, indicating differences in their genome structure. In order to better quantify the structural properties of chromosomes and compare between cell types, it may be useful to introduce normalizations to hypergraph entropy in the future.

6.5.16 Hypergraph Distance

Comparing graphs is a ubiquitous task in data analysis and machine learning [97, 97, 106, 381]. In order to quantify difference between two genomic hypergraphs G_1 and G_2 at different scales, we propose using several hypergraph distance or similarity measures. These measures are based on the conversion of a hypergraph into a graph representation, see [334] for details. Denote the incidence matrices of two genomic hypergraphs by $\mathbf{H}_1 \in \mathbb{R}^{n \times m_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{n \times m_2}$, respectively. For $i = 1, 2$, construct the adjacency matrices $\mathbf{A}_i$ and normalized Laplacian matrices $\tilde{\mathbf{L}}_i$:

$$\mathbf{A}_i = \mathbf{H}_i \mathbf{E}_i^{-1} \mathbf{H}_i^{\top}, \qquad \tilde{\mathbf{L}}_i = \mathbf{I} - \mathbf{D}_i^{-\frac{1}{2}} \mathbf{H}_i \mathbf{E}_i^{-1} \mathbf{H}_i^{\top} \mathbf{D}_i^{-\frac{1}{2}} \in \mathbb{R}^{n \times n}, \tag{6.3}$$

respectively, where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix, $\mathbf{E}_i \in \mathbb{R}^{m_i \times m_i}$ is a diagonal matrix containing the orders of hyperedges along its diagonal, and $\mathbf{D}_i \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the degrees of nodes along its diagonal [406]. The degree of a node is equal to the number of hyperedges that contain that node. Given these adjacency and normalized Laplacian matrices, we use following three distance measures in our application to determine differences in the two genomic hypergraphs at both local and global scales:

- **Hamming Distance:** measures local similarity and is based on absolute values of difference

between the two adjacency matrices, i.e.,

$$D_H(\mathsf{G}_1, \mathsf{G}_2) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} |\mathbf{A}_{1,jk} - \mathbf{A}_{2,jk}|,$$

where, the notation $\mathbf{A}_{i,jk}$ implies jk-th entry of the matrix $\mathbf{A}_i$.

- **Spectral Distance:** measures global similarity and is based on the p-norm for difference between ordered set of eigenvalues of the two Laplacians, i.e.,

$$D_\lambda(\mathsf{G}_1, \mathsf{G}_2) = \frac{1}{n}(\sum_{j=1}^{n} |\lambda_{1,j} - \lambda_{2,j}|^p)^{1/p}, \tag{6.4}$$

where $\lambda_{i,j}$ is the jth eigenvalue of $\tilde{\mathbf{L}}_i$ for $i = 1, 2$, and $p \geq 1$. In our analysis, we choose $p = 2$.

- **DeltaCon Distance:** measures both local and global similarity, and is based on the fast belief propagation method of measuring node affinities using the matrix [173],

$$\mathbf{S}_i = \left(\mathbf{I} + \epsilon^2 \mathbf{D}_i^a - \epsilon \mathbf{A}_i\right)^{-1},$$

where, $0 < \epsilon \ll 1$ is small constant capturing the influence between neighboring nodes, and $\mathbf{D}_i^a$ is the $n \times n$ diagonal matrix with the diagonal entries $\mathbf{D}_{i,jj}^a = \sum_{k=1}^{n} \mathbf{A}_{i,jk}$. DeltaCon then compares the two matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ via the Matusita difference as the measure,

$$D_\Delta(\mathsf{G}_1, \mathsf{G}_2) = \frac{1}{n^2} \left(\sum_{j=1}^{n} \sum_{k=1}^{n} (\mathbf{S}_{1,jk}{}^{1/2} - \mathbf{S}_{2,jk}{}^{1/2})^2 \right)^{1/2}. \tag{6.5}$$

where we have added a normalization factor $\frac{1}{n^2}$. In our analysis, we found results too insensitive to the choice of ϵ, and report the results for $\epsilon = 10^{-3}$.

Further details on the properties of these different distances can be found in Supplemental Notes.

We computed hypergraph distance between genome-wide hypergraphs derived from adult fibroblasts, neonatal fibroblasts, and B lymphocytes using the Hamming, spectral, and DeltaCon distances described above and examined distances statistically through a permutation test. Figure S2A1-3 in Dotson *et al.* [99] demonstrates that the adult fibroblast and B lymphocyte hypergraphs are significantly different at the chromosome level, especially along Chromosome 21, in stark contrast to the distance between adult and neonatal fibroblasts. Additionally, we computed the same distance measures at the genome level, incorporating inter-chromosomal data, and found that the

genomic hypergraphs between fibroblasts and B lymphocytes were significantly different, with a p-value of 0 compared to an observed insignificant difference between adult and neonatal fibroblasts (p-value of 1) (Figure S2B in Dotson *et al.* [99]).

6.5.17 Statistical Significance via Permutation Test

In order to assess the statistical significance of the transcription cluster candidates we determined using our criteria (Figure 6.6), we used a permutation test which builds the shape of the null hypothesis (i.e. the random background distribution) by resampling the observed data over N trials. We randomly selected n 3rd, 4th, 5th, and 6th or more order multi-way contacts from our Pore-C data, where n was based on the number of transcription cluster candidates we determined for each order. For example, we randomly selected $n = 11,261$ multi-way contacts from the set of 3rd order multi-way contacts in fibroblasts (Table 6.1). For each trial, we determined how many of these randomly sampled "transcription clusters" match our remaining criteria: transcription clusters with ≥ 1 gene, ≥ 2 genes, common TFs, and common MRs. The background distribution for each of the criteria was then constructed from these values. The proportion of values in the background distributions that was greater than their counterparts from the data-derived transcription cluster candidates yielded the p-value. This analysis was based on the assumption that transcription clusters will be more likely to contain genes and that those genes are more likely to have common transcription factors than random multi-way contacts. For this analysis, we chose $N = 1,000$ trials.

Similarly, we used a permutation test to determine the significance of the measured distances between two hypergraphs. Suppose that we are comparing two hypergraphs G_1 and G_2. We first randomly generate N hypergraphs $\{\mathsf{R}_i\}_{i=1}^{N}$ that are similar to G_1 ("similar" means similar number of node degree and hyperedge size distribution). The background distribution therefore can be constructed by measuring the hypergraph distances between G_1 and R_i for $i = 1, 2, \ldots, N$. The proportion of distances that was greater than the distance between G_1 and G_2 in this background distribution yielded the p-value. For this analysis, we again chose $N = 1,000$ trials. See Supplemental Notes in Dotson *et al.* [99] for details.

CHAPTER 7

Concluding Remarks

Cellular reprograming has seen exciting clinical success in the 21st century. Our growing understanding of cell identity has translated into age-related macular degeneration [397], Parkinson's disease [20], and myocardial injury [201] treatment potential. The commercial success and scalability of cellular reprogramming solutions in recent years has even made direct-to-consumer cell therapies a tangible option. Persisting low conversion efficiencies and discrepancies between reprogrammed cells and their native counterparts, however, challenge us to consider what we still do not know about cell identity and lineage commitment decision-making. In Chapter 2, we touch on some of those challenges and cover computational approaches designed to predict how cells will respond to perturbation and better facilitate reprogramming success. We highlight how integration of different data modalities has enabled more creativity in predicting reprogramming regimes, from factoring in time of perturbation and implementing feedback control to adjust transcription factor input concentrations real-time, to designing hybrid schemes that both introduce and inhibit transcription factors in a system and more. Fortunately, the advent of next-generation sequencing nearly 20 years ago and the rapidly-growing development of new techniques like spatial transcriptomics and Pore-C that are more sensitive to the nuances of cellular and genome dynamics, allow us to better elucidate cell phenotype from the dynamical relationship between genome structure and function. With such informative data at our fingertips, our next step is to let this refined 4DN framework guide cellular reprogramming solutions into an era of higher efficiency rates and ultimately tissues biomanufactured from reprogrammed cells.

A uniform regime for transforming cancer cells into normal cells remains elusive in the field of cellular reprogramming due to robustness and clonal diversity. In Chapter 3, we highlight that cellular reprogramming need not be limited to inducing a new lineage's transcriptional program, but that merely attenuating the transcriptional program of a cancer cell can open the door to several more feasible and effective outcomes like acquisition of a treatment-sensitive, senescent, or apoptotic state. We used 4DN analysis to demonstrate that silencing the transcription factor, TCF4, in colorectal cancer cells progressively disrupted pathways promoting colon cancer cell homeostasis, a perturbation sufficient for restoring the cells' sensitivity to chemoradiotherapy. In this work,

we characterized specific changes in chromatin partitioning at key loci that were coordinated with functional changes and appeared to contribute to the observed treatment sensitization phenotype. Our application of network centrality analysis of the local genome architecture further recovered TF-encoding genes previously implicated in co-regulating tumor progression, demonstrating the utility of the 4DN framework in extracting clinically-relevant reprogramming factors.

Coordinated changes in genome structure and function have been described in several studies and architectural proteins have been shown to drive key lineage-commitment events in cells, yet genome structure continues to be an unappreciated target for cell therapy. In Chapter 4, we perform 4DN analysis on Hi-C and RNA-seq data collected *in vivo* from T cells at baseline and in two stimulated, proliferative states before and following knockout of the cohesin-related WAPL protein, necessary for proper chromatin looping. We found that the absence of WAPL correlated with universal defects to internal TAD structure that were coordinated with differential expression in relevant cell cycling genes. Importantly, the loss of WAPL was also accompanied by a reduction in GVHD severity, suggesting that disruption of key regulators of genome structure can invoke desirable clinical outcomes. Further investigation will need to be performed to ensure that such a perturbation would also drive changes in the transcriptional program that are compatible with the target state.

Current gold-standard functional analyses capture gene expression in cells from dissociated or homogenized tissue, removing molecular measurements from their spatially-aware context. In Chapter 5, we enter the emerging field of spatial transcriptomics to interrogate the morphing cellular and genomic landscape of adipose tissue during obesity progression. We combined spatial transcriptomics with single-cell RNA-seq in lean and diet-induced obese mice over time to reveal extensive disruption of innate-adaptive immune response. Dynamically changing ligand-receptor signaling at key tissue landmarks showcased a decisive shift from adaptive immune cell-driven interactions to innate immune-cell driven interactions, highlighting patterns of associated position-specific activity. These data allowed us to profile cross-tissue heterogeneity and represent a valuable extension to the 4DN framework. Moreover, knowledge of gene expression spatial constraints could better aid in selection of factors for reprogramming and inform placement of reprogrammed cells for tissue biomanufacturing. Advances in refining tissue spot resolution to the single-cell level will be needed to make more precise observations about the intricate cell-cell network in the tissue moving forward. Additionally, if chromosome conformation technologies rise to the challenge of preserving spatial context, the promise of spatially-resolved 4DN analysis could soon be on the horizon.

Since its introduction nearly two decades ago, chromosome conformation capture and its many derivatives like Hi-C have taken the genomics world by storm, revealing important regulatory relationships, characterizing hierarchical genome organization, and uncovering properties of key

architectural proteins. While informative, it has become clear in recent years that pairwise chromatin interactions only tell part of the genome organization story. Several new technologies have since emerged facilitating the capture or recovery of simultaneously-occurring chromatin interactions involving more than two loci in the vicinity of one another. In Chapter 6, we introduce a framework for processing and analyzing these interactions captured using Pore-C sequencing. The foundation of our proposed framework is hypergraph representation of multi-way contacts, allowing us to use sophisticated quantitative measurements to describe entropy and similarity among these complex structures. We further integrated Pore-C contacts with additional data modalities to define transcription clusters where transcription factors and functionally-related loci converge in three-dimensional space to undergo efficient transcription. This approach allows us to elucidate key regulatory relationships occurring at an inter-chromosomal scale that were previously undetected through pairwise conformation capture assays. Further, the cell type-specific multi-way interactions that we are able to classify have the potential to guide prediction of reprogramming factors that regulate the conformational properties of specific cell lineages. Adapting this framework to existing single-cell Pore-C data and further validating the existence of multi-way chromatin interactions using imaging remain critical next steps before clinical application of this framework.

APPENDIX A

Supplemental Materials

Supplementary materials, including figures, tables, and additional notes can be found at [39] for Chapter 3, [330] for Chapter 4, [100] and [99] for Chapter 6.

A.1 Chapter 5 Algorithms

Algorithm 2: Clustering and Visualization

Input: Data matrix $\mathbf{X}_{m\times n} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^{m\times n}$ where m rows are genes and n columns are cells.

Output: Cell clusters and a low dimensional projection

1 Compute the sample mean μ_n and the centered matrix $\mathbf{X}_c = \mathbf{X} - \mu_n \mathbf{1}^\top$ where $\mathbf{1}$ is a vector of ones

2 Compute the SVD of $\mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$

3 Construct $\mathbf{P}_{n\times r} = \begin{bmatrix} v_1 & v_2 & \dots & v_r \end{bmatrix}$ where each column in $\mathbf{P}$ is a right singular vector of $\mathbf{X}_c$. Here r can be chosen using the optimal hard threshold [117] on $\mathbf{X}_c$

4 Construct a similarity matrix $\mathbf{A}_{n\times n}$ from $\mathbf{P}$ by determining the distance between each row. The choice of distance measure depends on the data type and user preference. Examples include Gaussian similarity, Euclidean distance, Manhattan distance (city block distance), Kullbeck-Liebler divergence, and correlation

5 Perform clustering: spectral or modularity clustering on $\mathbf{A}$ with k clusters. k can be chosen using domain knowledge or by testing multiple values of k and evaluating the best performance. Note: k may be $\leq r$

6 Visualization: t-SNE or UMAP to reduce the dimensions of $\mathbf{P}$ and visualize data colored according to clusters

Algorithm 3: Cell Type Localization and Spot Deconvolution

Input: scRNA-seq feature-barcode matrices $\mathbf{F}_{(s)} \in \mathbb{R}^{m \times n}$ where m rows are genes and n columns are cells; ST feature-barcode matrices $\mathbf{T}_{(s)} \in \mathbb{R}^{m \times n}$ where m rows are genes and n columns are tissue spots

Output: Tissue spot cell type assignments

1 Perform spectral clustering on $\mathbf{F}_{(s)}$ aggregated across all timepoints s

2 Rank m genes for each cell type (cluster) k. Genes are ranked by ubiquity, cluster-averaged expression, and uniqueness (see Methods and Materials)

3 Construct cell type-specific signatures $\mathbf{S}_{(k)} \in \mathbb{R}^{50 \times 1}$, containing highly ranked genes common across timepoints

4 Query expression of cell type-specific signature genes at tissue spots by finding $\mathbf{T}_{(s)} \subset \mathbf{S}_{(k)}$

5 Assign to each tissue spot the cell type expressing the most and at minimum 5 of its signature genes

6 Deconvolute tissue spots: repeat step 5 until an attempt to assign each cell type to every tissue spot has been made. Each successive assignment represents a deconvoluted tissue layer.

Algorithm 4: Ligand-Receptor Colocalization Dynamics

Input: Table of ligand-receptor pairs $\mathbf{L}$, from literature

Output: Global state matrix, $\mathbf{S} \in \mathbb{R}^{n \times n}$

1 **ST Data:** For each ligand-receptor pair p in $\mathbf{L}$, assign colocalization score:

$$c = \begin{cases} 1 & \text{if } l > 0 \text{ and } r > 0 \\ 0 & \text{otherwise} \end{cases},$$

where l and r are UMI counts for the ligand and receptor in p, respectively, at a given tissue spot

2 **scRNA-seq Data:** If p has a colocalization score of 1 at least once across the tissue, construct contingency table $\mathbf{C}_{2 \times n}$, where $\mathbf{C_{11}}$ and $\mathbf{C_{21}}$ are the proportions of single cells classified as 'cell type 1' with l and $r > 3$, respectively, $\mathbf{C_{12}}$ and $\mathbf{C_{22}}$ are the proportions of single cells classified as 'cell type 2' with l and $r > 3$, respectively, and so on up to 'cell type n' ($\mathbf{C_{1n}}$ and $\mathbf{C_{2n}}$)

3 Construct local state matrix for $\mathbf{S}_{(p)} \in \mathbb{R}^{n \times n}$, by finding the mean between all pairwise combinations of ligand-specific ($\mathbf{C_{11}}$,$\mathbf{C_{12}}$,...,$\mathbf{C_{1n}}$) and receptor-specific ($\mathbf{C_{21}}$,$\mathbf{C_{22}}$,...,$\mathbf{C_{2n}}$) proportions in $\mathbf{C}$

4 After performing steps 1-3 for all p in $\mathbf{L}$, define global state matrix, $\mathbf{S} \in \mathbb{R}^{n \times n}$, by computing the element-wise mean across all $\mathbf{S}_{(p)} \in \mathbb{R}^{n \times n} : \sum_{p=1}^{m} \mathbf{S}_{(p)}/m$, where m is the number of ligand-receptor pairs, p

BIBLIOGRAPHY

[1] Emily Abelseth, Laila Abelseth, Laura De la Vega, Simon T Beyer, Samuel J Wadsworth, and Stephanie M Willerth. 3d printing of neural tissues derived from human induced pluripotent stem cells using a fibrin-based bioink. *ACS Biomaterials Science & Engineering*, 5(1):234–243, 2018.

[2] Theodora Agalioti, Stavros Lomvardas, Bhavin Parekh, Junming Yie, Tom Maniatis, and Dimitris Thanos. Ordered recruitment of chromatin modifying and general transcription factors to the ifn-β promoter. *Cell*, 103(4):667–678, 2000.

[3] Jan-Eric Ahlfors, Ashkan Azimi, Rouwayda El-Ayoubi, Alexander Velumian, Ilan Vonderwalde, Cecile Boscher, Oana Mihai, Sarathi Mani, Marina Samoilova, Mohamad Khazaei, et al. Examining the fundamental biology of a novel population of directly reprogrammed human neural precursor cells. *Stem Cell Research & Therapy*, 10(1):1–17, 2019.

[4] Elnaz Alipour and John F Marko. Self-organization of domain structures by dna-loop-extruding enzymes. *Nucleic acids research*, 40(22):11202–11212, 2012.

[5] Amin Allahyar, Carlo Vermeulen, Britta AM Bouwman, Peter HL Krijger, Marjon JAM Verstegen, Geert Geeven, Melissa van Kranenburg, Mark Pieterse, Roy Straver, Judith HI Haarhuis, et al. Enhancer hubs and loop collisions identified from single-allele topologies. *Nature genetics*, 50(8):1151–1160, 2018.

[6] Quentin Alle, Enora Le Borgne, Ollivier Milhavet, and Jean-Marc Lemaitre. Reprogramming: Emerging strategies to rejuvenate aging cells and tissues. *International Journal of Molecular Sciences*, 22(8):3990, 2021.

[7] Robin C Allshire and Hiten D Madhani. Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology*, 19(4):229, 2018.

[8] Sergio E Alvarez, Kuzhuvelil B Harikumar, Nitai C Hait, Jeremy Allegood, Graham M Strub, Eugene Y Kim, Michael Maceyka, Hualiang Jiang, Cheng Luo, Tomasz Kordula, et al. Sphingosine-1-phosphate is a missing cofactor for the e3 ubiquitin ligase traf2. *Nature*, 465(7301):1084–1088, 2010.

[9] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[10] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[11] Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegård, Fredrik Salmén, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660, 2019.

[12] Ferhat Ay, Thanh H Vu, Michael J Zeitz, Nelle Varoquaux, Jan E Carette, Jean-Philippe Vert, Andrew R Hoffman, and William S Noble. Identifying multi-locus chromatin contacts in human cells using tethered multiple 3c. *BMC genomics*, 16(1):1–17, 2015.

[13] Begüm Aydin and Esteban O Mazzoni. Cell reprogramming: the many roads to success. *Annual Review of Cell and Developmental Biology*, 35:433–452, 2019.

[14] Alba Azagra, Ester Marina-Zárate, Almudena R Ramiro, Biola M Javierre, and Maribel Parra. From loops to looks: transcription factors and chromatin organization shaping terminal b cell differentiation. *Trends in Immunology*, 41(1):46–60, 2020.

[15] Jesper Bäckdahl, Lovisa Franzén, Lucas Massier, Qian Li, Jutta Jalkanen, Hui Gao, Alma Andersson, Nayanika Bhalla, Anders Thorell, Mikael Rydén, et al. Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell metabolism*, 33(9):1869–1882, 2021.

[16] Christopher RS Banerji, Diego Miranda-Saavedra, Simone Severini, Martin Widschwendter, Tariq Enver, Joseph X Zhou, and Andrew E Teschendorff. Cellular network entropy as the energy potential in waddington's differentiation landscape. *Scientific Reports*, 3(1):1–7, 2013.

[17] Andreea Barbu, Osama A Hamad, Lars Lind, Kristina N Ekdahl, and Bo Nilsson. The role of complement factor c3 in lipid metabolism. *Molecular immunology*, 67(1):101–107, 2015.

[18] Nick Barker, Adam Hurlstone, Hannah Musisi, Antony Miles, Mariann Bienz, and Hans Clevers. The chromatin remodelling factor brg-1 interacts with β-catenin to promote target gene activation. *The EMBO journal*, 20(17):4935–4943, 2001.

[19] Nick Barker, Johan H Van Es, Jeroen Kuipers, Pekka Kujala, Maaike Van Den Born, Miranda Cozijnsen, Andrea Haegebarth, Jeroen Korving, Harry Begthel, Peter J Peters, et al. Identification of stem cells in small intestine and colon by marker gene lgr5. *Nature*, 449(7165):1003–1007, 2007.

[20] Roger A Barker, Malin Parmar, Lorenz Studer, and Jun Takahashi. Human trials of stem cell-derived dopamine neurons for parkinson's disease: dawn of a new era. *Cell stem cell*, 21(5):569–573, 2017.

[21] Amitava Basu and Vijay K Tiwari. Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine. *Clinical Epigenetics*, 13(1):1–11, 2021.

[22] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee CA Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, 2017.

[23] Nicole Beauchemin and Azadeh Arabzadeh. Carcinoembryonic antigen-related cell adhesion molecules (ceacams) in cancer progression and metastasis. *Cancer and Metastasis Reviews*, 32(3-4):643–671, 2013.

[24] Naiara G Bediaga, Hannah D Coughlan, Timothy M Johanson, Alexandra L Garnham, Gaetano Naselli, Jan Schröder, Liam G Fearnley, Esther Bandala-Sanchez, Rhys S Allan, Gordon K Smyth, et al. Multi-level remodelling of chromatin underlying activation of human t cells. *Scientific reports*, 11(1):1–16, 2021.

[25] Brian J Beliveau, Eric F Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y Fonseka, Ruth B McCole, Yiming Chang, Jin Billy Li, Tharanga Niroshini Senaratne, Benjamin R Williams, et al. Versatile design and synthesis platform for visualizing genomes with oligopaint fish probes. *Proceedings of the National Academy of Sciences*, 109(52):21301–21306, 2012.

[26] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[27] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[28] Candace Bichsel, Dennis Neeld, Takashi Hamazaki, Lung-Ji Chang, Li-Jun Yang, Naohiro Terada, and Shouguang Jin. Direct reprogramming of fibroblasts to myocytes via bacterial injection of myod protein. *Cellular Reprogramming*, 15(2):117–125, 2013.

[29] Bruce R Blazar, Geoffrey R Hill, and William J Murphy. Dissecting the biology of allogeneic hsct to enhance the gvt effect whilst minimizing gvhd. *Nature Reviews Clinical Oncology*, 17(8):475–492, 2020.

[30] Kevin Blighe, Sharmila Rana, and Myles Lewis. Enhancedvolcano: Publication-ready volcano plots with enhanced colouring and labeling. *R package version*, 1(0), 2019.

[31] Francis Blokzijl, Joep De Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Nobuo Sasaki, Meritxell Huch, Sander Boymans, Ewart Kuijk, Pjotr Prins, et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264, 2016.

[32] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661, 2016.

[33] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[34] Anne Boulay, Joelle Rudloff, Jingjing Ye, Sabine Zumstein-Mecker, Terence O'Reilly, Dean B Evans, Shiuan Chen, and Heidi A Lane. Dual inhibition of mtor and estrogen receptor signaling in vitro induces cell death in models of breast cancer. *Clinical Cancer Research*, 11(14):5319–5328, 2005.

[35] Shelagh Boyle, Susan Gilchrist, Joanna M Bridger, Nicola L Mahy, Juliet A Ellis, and Wendy A Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, 10(3):211–220, 2001.

[36] CA Brackley, Nick Gilbert, Davide Michieletto, Argyris Papantonis, MCF Pereira, PR Cook, and Davide Marenduzzo. Complex small-world regulatory networks emerge from the 3d organisation of the human genome. *Nature communications*, 12(1):1–14, 2021.

[37] Rüdiger Braun, Lena Anthuber, Daniela Hirsch, Darawalee Wangsa, Justin Lack, Nicole E McNeil, Kerstin Heselmeyer-Haddad, Irianna Torres, Danny Wangsa, Markus A Brown, et al. Single cell-derived primary rectal carcinoma cell lines reflect intratumor heterogeneity associated with treatment response. *Clinical Cancer Research*, 2020.

[38] Robert Briggs and Thomas J. King. Transplantation of living nuclei from blastula cells into enucleated frogs' eggs. *Proceedings of the National Academy of Sciences*, 38(5):455–463, 1952.

[39] Markus A Brown, Gabrielle A Dotson, Scott Ronquist, Georg Emons, Indika Rajapakse, and Thomas Ried. Tcf7l2 silencing results in altered gene expression patterns accompanied by local genomic reorganization. *Neoplasia*, 23(2):257–269, 2021.

[40] Oliver S Burren, Arcadio Rubio García, Biola-Maria Javierre, Daniel B Rainbow, Jonathan Cairns, Nicholas J Cooper, John J Lambourne, Ellen Schofield, Xaquin Castro Dopico, Ricardo C Ferreira, et al. Chromosome contacts in activated t cells identify autoimmune disease candidate genes. *Genome biology*, 18(1):1–19, 2017.

[41] Georg A Busslinger, Roman R Stocsits, Petra Van Der Lelij, Elin Axelsson, Antonio Tedeschi, Niels Galjart, and Jan-Michael Peters. Cohesin is positioned in mammalian genomes by transcription, ctcf and wapl. *Nature*, 544(7651):503–507, 2017.

[42] Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & Development*, 25(18):1915–1927, 2011.

[43] Melissa Cadena, Liqun Ning, Alexia King, Boeun Hwang, Linqi Jin, Vahid Serpooshan, and Steven A Sloan. 3d bioprinting of neural tissues. *Advanced healthcare materials*, 10(15):2001600, 2021.

[44] Ken M Cadigan and Marian L Waterman. Tcf/lefs and wnt signaling in the nucleus. *Cold Spring Harbor perspectives in biology*, 4(11):a007906, 2012.

[45] Patrick Cahan. Enabling direct fate conversion with network biology. *Nature Genetics*, 48(3):226–227, 2016.

[46] Patrick Cahan, Hu Li, Samantha A Morris, Edroaldo Lummertz Da Rocha, George Q Daley, and James J Collins. Cellnet: network biology applied to stem cell engineering. *Cell*, 158(4):903–915, 2014.

[47] Andrea Califano and Mariano J Alvarez. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nature Reviews Cancer*, 17(2):116–130, 2017.

[48] Eugenia E Calle, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of us adults. *New England Journal of Medicine*, 348(17):1625–1638, 2003.

[49] Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how, and why? *Molecular cell*, 49(5):825–837, 2013.

[50] Nan Cao, Yu Huang, Jiashun Zheng, C Ian Spencer, Yu Zhang, Ji-Dong Fu, Baoming Nie, Min Xie, Mingliang Zhang, Haixia Wang, et al. Conversion of human fibroblasts into functional cardiomyocytes by small molecules. *Science*, 352(6290):1216–1220, 2016.

[51] Shangtao Cao, Shengyong Yu, Yan Chen, Xiaoshan Wang, Chunhua Zhou, Yuting Liu, Junqi Kuang, He Liu, Dongwei Li, Jing Ye, et al. Chemical reprogramming of mouse embryonic and adult fibroblast into endoderm lineage. *Journal of Biological Chemistry*, 292(46):19122–19132, 2017.

[52] Yi Cao, Zizhen Yao, Deepayan Sarkar, Michael Lawrence, Gilson J Sanchez, Maura H Parker, Kyle L MacQuarrie, Jerry Davison, Martin T Morgan, Walter L Ruzzo, et al. Genome-wide myod binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Developmental cell*, 18(4):662–674, 2010.

[53] Jodi M Carter, Tanya L Hoskin, M Alvaro Pena, Rushin Brahmbhatt, Stacey J Winham, Marlene H Frost, Melody Stallings-Mann, Derek C Radisky, Keith L Knutson, Daniel W Visscher, et al. Macrophagic "crown-like structures" are associated with an increased risk of breast cancer in benign breast disease. *Cancer Prevention Research*, 11(2):113–119, 2018.

[54] Joseph F Cavallari, Emmanuel Denou, Kevin P Foley, Waliul I Khan, and Jonathan D Schertzer. Different th17 immunity in gut, liver, and adipose tissues during obesity: the role of diet, genetics, and microbes. *Gut Microbes*, 7(1):82–89, 2016.

[55] Claudia Cavelti-Weder, Weida Li, Adrian Zumsteg, Marianne Stemann, Takatsugu Yamada, Susan Bonner-Weir, Gordon Weir, and Qiao Zhou. Direct reprogramming for pancreatic beta-cells using key developmental genes. *Current Pathobiology Reports*, 3(1):57–65, 2015.

[56] Steven M Chan, Andrew P Weng, Robert Tibshirani, Jon C Aster, and Paul J Utz. Notch signals positively regulate activity of the mtor pathway in t-cell acute lymphoblastic leukemia. *Blood, The Journal of the American Society of Hematology*, 110(1):278–286, 2007.

[57] Julie Chaumeil, Mariann Micsinai, Panagiotis Ntziachristos, Ludovic Deriano, Joy M-H Wang, Yanhong Ji, Elphege P Nora, Matthew J Rodesch, Jeffrey A Jeddeloh, Iannis Aifantis, et al. Higher-order looping and nuclear organization of tcra facilitate targeted rag cleavage and regulated rearrangement in recombination centers. *Cell reports*, 3(2):359–370, 2013.

[58] Can Chen and Indika Rajapakse. Tensor entropy for uniform hypergraphs. *IEEE Transactions on Network Science and Engineering*, 2020.

[59] Can Chen, Amit Surana, Anthony Bloch, and Indika Rajapakse. Multilinear time invariant system theory. In *2019 Proceedings of the Conference on Control and its Applications*, pages 118–125. SIAM, 2019.

[60] Can Chen, Amit Surana, Anthony Bloch, and Indika Rajapakse. Controllability of hypergraphs. *arXiv preprint arXiv:2005.12244*, 2020.

[61] Haiming Chen, Jie Chen, Lindsey A Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4d nucleome. *Proceedings of the National Academy of Sciences*, 112(26):8002–8007, 2015.

[62] Jian-Hua Chen, Kim Jee Goh, Nuno Rocha, Matthijs P Groeneveld, Marina Minic, Timothy G Barrett, David Savage, and Robert K Semple. Evaluation of human dermal fibroblasts directly reprogrammed to adipocyte-like cells as a metabolic disease model. *Disease Models & Mechanisms*, 10(12):1411–1420, 2017.

[63] Jie Chen, Alfred O Hero III, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, 2016.

[64] Kevin G Chen, Barbara S Mallon, Ronald DG McKay, and Pamela G Robey. Human pluripotent stem cell culture: considerations for maintenance, expansion, and therapeutics. *Cell Stem Cell*, 14(1):13–26, 2014.

[65] Pin-Yu Chen, Lingfei Wu, Sijia Liu, and Indika Rajapakse. Fast incremental von neumann graph entropy computation: Theory, algorithm, and applications. *arXiv preprint arXiv:1805.11769*, 2018.

[66] Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, et al. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991, 2020.

[67] Anne BC Cherry and George Q Daley. Reprogrammed cells for disease modeling and regenerative medicine. *Annual Review of Medicine*, 64:277–290, 2013.

[68] Chun-Seok Cho, Jingyue Xi, Yichen Si, Sung-Rye Park, Jer-En Hsu, Myungjin Kim, Goo Jun, Hyun Min Kang, and Jun Hee Lee. Microscopic examination of spatial transcriptome using seq-scope. *Cell*, 2021.

[69] Kae Won Cho, Brian F Zamarron, Lindsey A Muir, Kanakadurga Singer, Cara E Porsche, Jennifer B DelProposto, Lynn Geletka, Kevin A Meyer, Robert W O'Rourke, and Carey N Lumeng. Adipose tissue dendritic cells are independent contributors to obesity-induced inflammation and insulin resistance. *The Journal of Immunology*, 197(9):3650–3661, 2016.

[70] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, and Danilo P Mandic. Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning*, 9(4-5):249–429, 2016.

[71] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[72] Peter Cook and Davide Marenduzzo. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic acids research*, 46, 09 2018.

[73] Peter R Cook. The organization of replication and transcription. *Science*, 284(5421):1790–1795, 1999.

[74] Peter R Cook and Davide Marenduzzo. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic Acids Research*, 46(19):9895–9906, 2018.

[75] Sean P Cornelius, William L Kath, and Adilson E Motter. Realistic control of network dynamics. *Nature Communications*, 4(1):1–9, 2013.

[76] Thomas M Cover. *Elements of Information Theory*. John Wiley & Sons, 1999.

[77] Marion Cremer and Thomas Cremer. Nuclear compartmentalization, dynamics, and function of regulatory dna sequences. *Genes, Chromosomes and Cancer*, 58(7):427–436, 2019.

[78] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292–301, 2001.

[79] Ana Cuadrado and Ana Losada. Specialized functions of cohesins stag1 and stag2 in 3d genome architecture. *Current opinion in genetics & development*, 61:9–16, 2020.

[80] D Cunningham. atkin w, lenz h, lynch ht, minsky b, nordlinger b, et al. colorectal cancer. *The Lancet*, 375(9719):1030–1047, 2010.

[81] Haijiang Dai, Tariq A. Alsalhe, Nasr Chalghaf, Matteo Riccò, Nicola Luigi Bragazzi, and Jianhong Wu. The global burden of disease attributable to high body mass index in 195 countries and territories, 1990–2017: An analysis of the global burden of disease study. *PLOS Medicine*, 17(7):1–19, 07 2020.

[82] Timothy Daley and Andrew D Smith. Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325–327, 2013.

[83] Emily M Darrow, Miriam H Huntley, Olga Dudchenko, Elena K Stamenova, Neva C Durand, Zhuo Sun, Su-Chen Huang, Adrian L Sanborn, Ido Machol, Muhammad Shamim, et al. Deletion of dxz4 on the human inactive x chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31):E4504–E4512, 2016.

[84] Claudia Ribeiro de Almeida, Rudi W Hendriks, and Ralph Stadhouders. Dynamic control of long-range genomic interactions at the immunoglobulin κ light-chain locus. *Advances in Immunology*, 128:183–271, 2015.

[85] Hugues de Thé. Differentiation therapy revisited. *Nature Reviews Cancer*, 18(2):117–127, 2018.

[86] Natasja L De Vries, Ahmed Mahfouz, Frits Koning, and Noel FCC De Miranda. Unraveling the complexity of the cancer microenvironment with multidimensional genomic and cytometric technologies. *Frontiers in oncology*, page 1254, 2020.

[87] Nicole JW De Wit, Mark V Boekschoten, Eva-Maria Bachmair, Guido JEJ Hooiveld, Philip J de Groot, Isabel Rubio-Aliaga, Hannelore Daniel, and Michael Müller. Dose-dependent effects of dietary fat on development of obesity in relation to intestinal differential gene expression in c57bl/6j mice. *PLoS One*, 6(4):e19145, 2011.

[88] Nicole Jw De Wit, Hanneke Bosch-Vermeulen, Philip J de Groot, Guido JEJ Hooiveld, Mechteld M Grootte Bromhaar, Jenny Jansen, Michael Müller, and Roelof van der Meer. The role of the small intestine in the development of dietary fat-induced obesity and insulin resistance in c57bl/6j mice. *BMC medical genomics*, 1(1):1–16, 2008.

[89] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. The 4d nucleome project. *Nature*, 549(7671):219–226, 2017.

[90] Job Dekker and Leonid Mirny. The 3d genome as moderator of chromosomal communication. *Cell*, 164(6):1110–1121, 2016.

[91] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[92] Domitilla Del Vecchio, Hussein Abdallah, Yili Qian, and James J Collins. A blueprint for a synthetic genetic feedback controller to reprogram cell fate. *Cell Systems*, 4(1):109–120, 2017.

[93] Aditya S Deshpande, Netha Ulahannan, Matthew Pendleton, Xiaoguang Dai, Lynn Ly, Julie M Behr, Stefan Schwenk, Will Liao, Michael A Augello, Carly Tyer, et al. Identifying synergistic high-order 3d chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature Biotechnology*, pages 1–12, 2022.

[94] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[95] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[96] Soheila Dolatabadi, Julián Candia, Nina Akrap, Christoffer Vannas, Tajana Tesan Tomic, Wolfgang Losert, Göran Landberg, Pierre Åman, and Anders Ståhlberg. Cell cycle and cell size dependent gene expression reveals distinct subpopulations at single-cell level. *Frontiers in genetics*, 8:1, 2017.

[97] Claire Donnat and Susan Holmes. Tracking network dynamics: A survey using graph distances. *The Annals of Applied Statistics*, 12(2):971 – 1012, 2018.

[98] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, 2006.

[99] Gabrielle A. Dotson, Stephen Lindsly, Anthony Cicalo, Can Chen, Sam Dilworth, Charles Ryan, Sivakumar Jeyarajan, Walter Meixner, Nicholas Beckloff, Amit Surana, Max Wicha, Lindsey A. Muir, and Indika Rajapakse. Deciphering multi-way interactions in the human genome. *bioRxiv*, 2022.

[100] Gabrielle A Dotson, Indika Rajapakse, and Lindsey A Muir. Monocyte-macrophage spatial dynamics promote pre-crown-like niches in early obesity. *bioRxiv*, 2022.

[101] Gabrielle A Dotson, Charles W Ryan, Can Chen, Lindsey Muir, and Indika Rajapakse. Cellular reprogramming: Mathematics meets medicine. *WIREs Mechanisms of Disease*, 13(4):e1515, 2021.

[102] Petros Drineas and Michael W Mahoney. Lectures on randomized numerical linear algebra. *The Mathematics of Data*, 25(1), 2018.

[103] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas SP Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell systems*, 3(1):95–98, 2016.

[104] Ana C D'Alessio, Zi Peng Fan, Katherine J Wert, Petr Baranov, Malkiel A Cohen, Janmeet S Saini, Evan Cohick, Carol Charniga, Daniel Dadon, Nancy M Hannett, et al. A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports*, 5(5):763–775, 2015.

[105] Lars Eldén. *Matrix methods in data mining and pattern recognition*. SIAM, 2007.

[106] Katherine Faust and John Skvoretz. Comparing networks across space and time, size and species. *Sociological Methodology*, 32(1):267–299, 2002.

[107] Elizabeth H Finn and Tom Misteli. Molecular basis and biological function of variability in spatial genome organization. *Science*, 365(6457):eaaw9498, 2019.

[108] Elizabeth H Finn, Gianluca Pegoraro, Hugo B Brandão, Anne-Laure Valton, Marlies E Oomen, Job Dekker, Leonid Mirny, and Tom Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, 176(6):1502–1515, 2019.

[109] William A Flavahan, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suvà, and Bradley E Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, 2016.

[110] Ji-Dong Fu, Nicole R Stone, Lei Liu, C Ian Spencer, Li Qian, Yohei Hayashi, Paul Delgado-Olguin, Sheng Ding, Benoit G Bruneau, and Deepak Srivastava. Direct reprogramming of human fibroblasts toward a cardiomyocyte-like state. *Stem cell reports*, 1(3):235–247, 2013.

[111] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15(9):2038–2049, 2016.

[112] Emily Jane Gallagher and Derek LeRoith. Obesity and diabetes: the increased risk of cancer and cancer-related mortality. *Physiological reviews*, 95(3):727–748, 2015.

[113] Xiao-qing Gan, Ji-yong Wang, Ying Xi, Zhi-li Wu, Yi-ping Li, and Lin Li. Nuclear dvl, c-jun, β-catenin, and tcf form a complex leading to stabilization of β-catenin–tcf interaction. *The Journal of cell biology*, 180(6):1087–1100, 2008.

[114] Longfei Gao, Wuqiang Guan, Min Wang, Huihan Wang, Jiali Yu, Qing Liu, Binlong Qiu, Yongchun Yu, Yifang Ping, Xiuwu Bian, et al. Direct generation of human neuronal cells from adult astrocytes by small molecules. *Stem Cell Reports*, 8(3):538–547, 2017.

[115] Tianshun Gao and Jiang Qian. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, 48(D1):D58–D64, 2020.

[116] Scarlett Gard, William Light, Bo Xiong, Tania Bose, Adrian J McNairn, Bethany Harris, Brian Fleharty, Chris Seidel, Jason H Brickner, and Jennifer L Gerton. Cohesinopathy mutations disrupt the subnuclear organization of chromatin. *Journal of Cell Biology*, 187(4):455–462, 2009.

[117] Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040 – 5053, 2014.

[118] Helmuth Gehart and Hans Clevers. Tales from the crypt: new insights into intestinal stem cells. *Nature Reviews Gastroenterology & Hepatology*, 16(1):19–34, 2019.

[119] B Michael Ghadimi, Marian Grade, Michael J Difilippantonio, Sudhir Varma, Richard Simon, Cristina Montagna, Laszlo Füzesi, Claus Langer, Heinz Becker, Torsten Liersch, et al. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 23(9):1826, 2005.

[120] Pollyanna A Goh, Sara Caxaria, Catharina Casper, Cecilia Rosales, Thomas T Warner, Pete J Coffey, and Amit C Nathwani. A systematic evaluation of integration free reprogramming methods for deriving clinically relevant patient specific induced pluripotent stem (ips) cells. *PLoS One*, 8(11):e81622, 2013.

[121] Dmitri Goriounov, Conrad L Leung, and Ronald KH Liem. Protein products of human gas2-related genes on chromosomes 17 and 22 (hgar17 and hgar22) associate with both microfilaments and microtubules. *Journal of cell science*, 116(6):1045–1058, 2003.

[122] Alexander Grath and Guohao Dai. Direct cell reprogramming for tissue engineering and regenerative medicine. *Journal of Biological Engineering*, 13(1):14, 2019.

[123] Jane L Grogan, Markus Mohrs, Brian Harmon, Dee A Lacy, John W Sedat, and Richard M Locksley. Early transcription and silencing of cytokine genes underlie polarization of t helper cell subsets. *Immunity*, 14(3):205–215, 2001.

[124] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

[125] Charlène Guillot and Thomas Lecuit. Mechanics of epithelial tissue homeostasis and morphogenesis. *Science*, 340(6137):1185–1189, 2013.

[126] John B Gurdon. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Development*, 10(4):622–640, 1962.

[127] Judith HI Haarhuis, Ahmed MO Elbatsh, Bram van den Broek, Daniel Camps, Hasan Erkan, Kees Jalink, René H Medema, and Benjamin D Rowland. Wapl-mediated removal of cohesin protects against segregation errors and aneuploidy. *Current biology*, 23(20):2071–2077, 2013.

[128] Judith HI Haarhuis and Benjamin D Rowland. Cohesin: building loops, but not compartments. *The EMBO Journal*, 36(24):3549–3551, 2017.

[129] Judith HI Haarhuis, Robin H van der Weide, Vincent A Blomen, J Omar Yáñez-Cuna, Mario Amendola, Marjon S van Ruiten, Peter HL Krijger, Hans Teunissen, René H Medema, Bas van Steensel, et al. The cohesin release factor wapl restricts chromatin loop extension. *Cell*, 169(4):693–707, 2017.

[130] Shuichi Hashimoto, Huaiyong Chen, Jianwen Que, Brian L Brockway, Jeffrey A Drake, Joshua C Snyder, Scott H Randell, and Barry R Stripp. β-catenin–sox2 signaling regulates the fate of developing airway epithelium. *Journal of cell science*, 125(4):932–942, 2012.

[131] Pantelis Hatzis, Laurens G van der Flier, Marc A van Driel, Victor Guryev, Fiona Nielsen, Sergei Denissov, Isaäc J Nijman, Jan Koster, Evan E Santo, Willem Welboren, et al. Genome-wide pattern of tcf7l2/tcf4 chromatin occupancy in colorectal cancer cells. *Molecular and cellular biology*, 28(8):2732–2744, 2008.

[132] Tong-Chuan He, Andrew B Sparks, Carlo Rago, Heiko Hermeking, Leigh Zawel, Luis T Da Costa, Patrice J Morin, Bert Vogelstein, and Kenneth W Kinzler. Identification of c-myc as a target of the apc pathway. *Science*, 281(5382):1509–1512, 1998.

[133] Andreas Hecht, Kris Vleminckx, Marc P Stemmler, Frans Van Roy, and Rolf Kemler. The p300/cbp acetyltransferases function as transcriptional coactivators of β-catenin in vertebrates. *The EMBO journal*, 19(8):1839–1850, 2000.

[134] Tracy SP Heng, Michio W Painter, Kutlu Elpek, Veronika Lukacs-Kornek, Nora Mauermann, Shannon J Turley, Daphne Koller, Francis S Kim, Amy J Wagers, Natasha Asinovski, et al. The immunological genome project: networks of gene expression in immune cells. *Nature immunology*, 9(10):1091–1094, 2008.

[135] Joseph Herdy, Simon Schafer, Yongsung Kim, Zoya Ansari, Dina Zangwill, Manching Ku, Apua Paquola, Hyungjun Lee, Jerome Mertens, and Fred H Gage. Chemical modulation of transcriptionally enriched signaling pathways to optimize the conversion of fibroblasts into neurons. *Elife*, 8:e41356, 2019.

[136] Susannah L Hewitt, Deborah Farmer, Katarzyna Marszalek, Emily Cadera, Hong-Erh Liang, Yang Xu, Mark S Schlissel, and Jane A Skok. Association between the igk and igh immunoglobulin loci mediated by the 3 igk enhancer induces' decontraction'of the igh locus in pre–b cells. *Nature immunology*, 9(4):396–404, 2008.

[137] Susannah L Hewitt, Frances A High, Steven L Reiner, Amanda G Fisher, and Matthias Merkenschlager. Nuclear repositioning marks the selective exclusion of lineage-inappropriate transcription factor loci during t helper cell differentiation. *European journal of immunology*, 34(12):3604–3613, 2004.

[138] David A Hill, Hee-Woong Lim, Yong Hoon Kim, Wesley Y Ho, Yee Hoon Foong, Victoria L Nelson, Hoang CB Nguyen, Kavya Chegireddy, Jihoon Kim, Andreas Habertheuer, et al. Distinct macrophage populations direct inflammatory versus physiological changes in adipose tissue. *Proceedings of the National Academy of Sciences*, 115(22):E5096–E5105, 2018.

[139] Louisa Hill, Anja Ebert, Markus Jaritz, Gordana Wutz, Kota Nagasaka, Hiromi Tagoh, Daniela Kostanova-Poliakova, Karina Schindler, Qiong Sun, Peter Bönelt, et al. Wapl repression by pax5 promotes v gene recombination by igh loop extrusion. *Nature*, 584(7819):142–147, 2020.

[140] Kenichi Horisawa and Atsushi Suzuki. Direct cell-fate conversion of somatic cells: Toward regenerative medicine and industries. *Proceedings of the Japan Academy, Series B*, 96(4):131–158, 2020.

[141] Hiromitsu Hoshino, Hiroaki Nagano, Naotsugu Haraguchi, Shimpei Nishikawa, Akira Tomokuni, Yoshihiro Kano, Takahito Fukusumi, Toshiyuki Saito, Miyuki Ozaki, Daisuke Sakai, et al. Hypoxia and tp53 deficiency for induced pluripotent stem cell-like properties in gastrointestinal cancer. *International journal of oncology*, 40(5):1423–1430, 2012.

[142] Gangqing Hu, Kairong Cui, Difeng Fang, Satoshi Hirose, Xun Wang, Darawalee Wangsa, Wenfei Jin, Thomas Ried, Pentao Liu, Jinfang Zhu, et al. Transformation of accessible chromatin and 3d nucleome underlies lineage commitment of early t cells. *Immunity*, 48(2):227–242, 2018.

[143] Hua Huang, Liang Zhong, Jin Zhou, Yanping Hou, Zhiyuan Zhang, Xiaoyu Xing, and Jie Sun. Leydig-like cells derived from reprogrammed human foreskin fibroblasts by crispr/dcas9 increase the level of serum testosterone in castrated male rats. *Journal of Cellular and Molecular Medicine*, 24(7):3971–3981, 2020.

[144] Elena V Ignatieva, Victor G Levitsky, and Nikolay A Kolchanov. Human genes encoding transcription factors and chromatin-modifying proteins have low levels of promoter polymorphism: a study of 1000 genomes project data. *International Journal of Genomics*, 2015, 2015.

[145] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[146] Maxim V Imakaev, Geoffrey Fudenberg, and Leonid A Mirny. Modeling chromosomes: Beyond pretty pictures. *FEBS letters*, 589(20):3031–3036, 2015.

[147] Toshiya Inaba, Hiroaki Honda, and Hirotaka Matsui. The enigma of monosomy 7. *Blood, The Journal of the American Society of Hematology*, 131(26):2891–2898, 2018.

[148] Seiichi Ishida, Erich Huang, Harry Zuzan, Rainer Spang, Gustavo Leone, Mike West, and Joseph R Nevins. Role for e2f in control of both dna replication and mitotic functions as revealed from dna microarray analysis. *Molecular and cellular biology*, 21(14):4684–4699, 2001.

[149] Banu Iskender, Kenan Izgi, and Halit Canatan. Reprogramming bladder cancer cells for studying cancer initiation and progression. *Tumor Biology*, 37(10):13237–13245, 2016.

[150] Takeshi Isoda, Amanda J Moore, Zhaoren He, Vivek Chandra, Masatoshi Aida, Matthew Denholtz, Jan Piet van Hamburg, Kathleen M Fisch, Aaron N Chang, Shawn P Fahl, et al. Non-coding transcription instructs chromatin folding and compartmentalization to dictate enhancer-promoter communication and t cell fate. *Cell*, 171(1):103–119, 2017.

[151] Michiko Itoh, Hideaki Kato, Takayoshi Suganami, Kuniha Konuma, Yoshio Marumoto, Shuji Terai, Hiroshi Sakugawa, Sayaka Kanai, Miho Hamaguchi, Takahiro Fukaishi, et al. Hepatic crown-like structure: a unique histological feature in non-alcoholic steatohepatitis in mice and humans. *PloS one*, 8(12):e82163, 2013.

[152] Neil M Iyengar, Xi Kathy Zhou, Ayca Gucalp, Patrick G Morris, Louise R Howe, Dilip D Giri, Monica Morrow, Hanhan Wang, Michael Pollak, Lee W Jones, et al. Systemic correlates of white adipose tissue inflammation in early-stage breast cancer. *Clinical Cancer Research*, 22(9):2283–2289, 2016.

[153] Lisa A Jackson, Evan J Anderson, Nadine G Rouphael, Paul C Roberts, Mamodikoe Makhene, Rhea N Coler, Michele P McCullough, James D Chappell, Mark R Denison, Laura J Stevens, et al. An mrna vaccine against sars-cov-2—preliminary report. *New England Journal of Medicine*, 2020.

[154] Diego Adhemar Jaitin, Lorenz Adlung, Christoph A Thaiss, Assaf Weiner, Baoguo Li, Hélène Descamps, Patrick Lundgren, Camille Bleriot, Zhaoyuan Liu, Aleksandra Deczkowska, et al. Lipid-associated macrophages control metabolic homeostasis in a trem2-dependent manner. *Cell*, 178(3):686–698, 2019.

[155] Suchit Jhunjhunwala, Menno C van Zelm, Mandy M Peak, and Cornelis Murre. Chromatin architecture and the generation of antigen receptor diversity. *Cell*, 138(3):435–448, 2009.

[156] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020.

[157] Yongxin Jin, Ying Liu, Zhenpeng Li, Katherine Santostefano, Jing Shi, Xinwen Zhang, Donghai Wu, Zhihui Cheng, Weihui Wu, Naohiro Terada, et al. Enhanced differentiation of human pluripotent stem cells into cardiomyocytes by bacteria-mediated transcription factors delivery. *PloS one*, 13(3):e0194895, 2018.

[158] Timothy M Johanson, Wing Fuk Chan, Christine R Keenan, and Rhys S Allan. Genome organization in immune cells: unique challenges. *Nature Reviews Immunology*, 19(7):448–456, 2019.

[159] Timothy M Johanson, Aaron TL Lun, Hannah D Coughlan, Tania Tan, Gordon K Smyth, Stephen L Nutt, and Rhys S Allan. Transcription-factor-mediated supervision of global genome architecture maintains b cell identity. *Nature immunology*, 19(11):1257–1264, 2018.

[160] Riya R Kanherkar, Naina Bhatia-Dey, and Antonei B Csoka. Epigenetics across the human lifespan. *Frontiers in Cell and Developmental biology*, 2:49, 2014.

[161] Rieke Kempfer and Ana Pombo. Methods for mapping 3d chromosome architecture. *Nature Reviews Genetics*, 21(4):207–226, 2020.

[162] Emil Kendziorra, Kerstin Ahlborn, Melanie Spitzner, Margret Rave-Fränk, Georg Emons, Jochen Gaedcke, Frank Kramer, Hendrik A Wolff, Heinz Becker, Tim Beissbarth, et al. Silencing of the wnt transcription factor tcf4 sensitizes colorectal cancer cells to (chemo-) radiotherapy. *Carcinogenesis*, 32(12):1824–1831, 2011.

[163] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Cheneby, Shubhada R Kulkarni, Ge Tan, et al. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1):D260–D266, 2018.

[164] Hye Jin Kim, Hyun Jyung Oh, Ji Sun Park, Jung Sun Lee, Jae-Hwan Kim, and Keun-Hong Park. Direct conversion of human dermal fibroblasts into cardiomyocyte-like cells using cicmc nanogels coupled with cardiac transcription factors and a nucleoside drug. *Advanced Science*, 7(7):1901818, 2020.

[165] Kye-Young Kim, Mihály Kovács, Sachiyo Kawamoto, James R Sellers, and Robert S Adelstein. Disease-associated mutations and alternative splicing alter the enzymatic and motile activity of nonmuscle myosins ii-b and ii-c. *Journal of Biological Chemistry*, 280(24):22769–22775, 2005.

[166] SH Kim, PG McQueen, MK Lichtman, EM Shevach, LA Parada, and T Misteli. Spatial genome organization during t-cell differentiation. *Cytogenetic and genome research*, 105(2-4):292–301, 2004.

[167] Jacob C Kimmel, Lolita Penland, Nimrod D Rubinstein, David G Hendrickson, David R Kelley, and Adam Z Rosenthal. Murine single-cell rna-seq reveals cell-identity-and tissue-specific trajectories of aging. *Genome Research*, 29(12):2088–2103, 2019.

[168] Kenneth W Kinzler and Bert Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87(2):159–170, 1996.

[169] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[170] Vladimir Korinek, Nick Barker, Petra Moerer, Elly van Donselaar, Gerwin Huls, Peter J Peters, and Hans Clevers. Depletion of epithelial stem-cell compartments in the small intestine of mice lacking tcf-4. *Nature genetics*, 19(4):379–383, 1998.

[171] Vladimir Korinek, Nick Barker, Patrice J Morin, Dick Van Wichen, Roel De Weger, Kenneth W Kinzler, Bert Vogelstein, and Hans Clevers. Constitutive transcriptional activation by a β-catenin-tcf complex in apc-/- colon carcinoma. *Science*, 275(5307):1784–1787, 1997.

[172] Tulay Koru-Sengul, Ana M Santander, Feng Miao, Lidia G Sanchez, Merce Jorda, Stefan Glück, Tan A Ince, Mehrad Nadji, Zhibin Chen, Manuel L Penichet, et al. Breast cancers from black women exhibit higher numbers of immunosuppressive macrophages with proliferative activity and of crown-like structures associated with lower survival compared to non-black latinas and caucasians. *Breast cancer research and treatment*, 158(1):113–126, 2016.

[173] Danai Koutra, Joshua T Vogelstein, and Christos Faloutsos. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM international conference on data mining*, pages 162–170. SIAM, 2013.

[174] Stephanie Kueng, Björn Hegemann, Beate H Peters, Jesse J Lipp, Alexander Schleiffer, Karl Mechtler, and Jan-Michael Peters. Wapl controls the dynamic association of cohesin with chromatin. *Cell*, 127(5):955–967, 2006.

[175] Andrew T Kwon, David J Arenillas, Rebecca Worsley Hunt, and Wyeth W Wasserman. opossum-3: advanced analysis of regulatory motif over-representation across genes or chip-seq datasets. *G3: Genes, Genomes, Genetics*, 2(9):987–1002, 2012.

[176] Julia Ladewig, Philipp Koch, and Oliver Brüstle. Leveling waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nature Reviews Molecular Cell Biology*, 14(4):225–236, 2013.

[177] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[178] Madeline A Lancaster, Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S Bicknell, Matthew E Hurles, Tessa Homfray, Josef M Penninger, Andrew P Jackson, and Juergen A Knoblich. Cerebral organoids model human brain development and microcephaly. *Nature*, 501(7467):373–379, 2013.

[179] Kinley Larntz and Michael D Perlman. A simple test for the equality of correlation matrices. *Rapport technique, Department of Statistics, University of Washington*, 141, 1985.

[180] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):1–17, 2014.

[181] François Le Dily, Davide Bau, Andy Pohl, Guillermo P Vicent, François Serra, Daniel Soronellas, Giancarlo Castellano, Roni HG Wright, Cecilia Ballare, Guillaume Filion, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & development*, 28(19):2151–2162, 2014.

[182] Christopher Lee, Meghan Robinson, and Stephanie M Willerth. Direct reprogramming of glioblastoma cells into neurons using small molecules. *ACS Chemical Neuroscience*, 9(12):3175–3185, 2018.

[183] Jung-Seob Lee, Byoung Soo Kim, Donghwan Seo, Jeong Hun Park, and Dong-Woo Cho. Three-dimensional cell printing of large-volume tissues: application to ear regeneration. *Tissue Engineering Part C: Methods*, 23(3):136–145, 2017.

[184] Minhyung Lee, Hyuna Sim, Hyunjun Ahn, Jeongmin Ha, Aruem Baek, Young-Joo Jeon, Mi-Young Son, and Janghwan Kim. Direct reprogramming to human induced neuronal progenitors from fibroblasts of familial and sporadic parkinson's disease patients. *International journal of stem cells*, 12(3):474, 2019.

[185] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature methods*, 11(11):1141–1143, 2014.

[186] Alona Levy-Jurgenson, Xavier Tekpli, Vessela N Kristensen, and Zohar Yakhini. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific reports*, 10(1):1–11, 2020.

[187] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1–16, 2011.

[188] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[189] Hsin-Kai Liao, Fumiyuki Hatanaka, Toshikazu Araoka, Pradeep Reddy, Min-Zu Wu, Yinghui Sui, Takayoshi Yamauchi, Masahiro Sakurai, David D O'Keefe, Estrella Núñez-Delicado, et al. In vivo target gene activation via crispr/cas9-mediated trans-epigenetic modulation. *Cell*, 171(7):1495–1507, 2017.

[190] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[191] Sergio Linares-Fernández, Céline Lacroix, Jean-Yves Exposito, and Bernard Verrier. Tailoring mrna vaccine to balance innate/adaptive immune response. *Trends in Molecular Medicine*, 26(3):311–323, 2020.

[192] Florian Lindner, Bailey Milne-Davies, Katja Langenfeld, Thorsten Stiewe, and Andreas Diepold. Litesec-t3ss-light-controlled protein delivery into eukaryotic cells with high spatial and temporal resolution. *Nature Communications*, 11(1):1–13, 2020.

[193] Stephen Lindsly, Can Chen, Sijia Liu, Scott Ronquist, Samuel Dilworth, Michael Perlman, and Indika Rajapakse. 4dnvestigator: time series genomic data analysis toolbox. *Nucleus*, 12(1):58–64, 2021.

[194] Stephen Lindsly, Wenlong Jia, Haiming Chen, Sijia Liu, Scott Ronquist, Can Chen, Xingzhao Wen, Cooper Stansbury, Gabrielle A Dotson, Charles Ryan, et al. Functional organization of the maternal and paternal human 4d nucleome. *Iscience*, 24(12):103452, 2021.

[195] Can Liu, Pinhao Li, Hui Li, Sicong Wang, Lifeng Ding, Hanbin Wang, Hui Ye, Yue Jin, Jinchao Hou, Xiangming Fang, et al. Trem2 regulates obesity-induced insulin resistance via adipose tissue remodeling in mice of high-fat feeding. *Journal of Translational Medicine*, 17(1):1–11, 2019.

[196] Lei Liu, Bokai Zhang, and Changbong Hyeon. Extracting multi-way chromatin contacts from hi-c data. *PLoS Computational Biology*, 17(12):e1009669, 2021.

[197] Ning Qing Liu, Michela Maresca, Teun van den Brand, Luca Braccioli, Marijne MGA Schijns, Hans Teunissen, Benoit G Bruneau, Elphge P Nora, and Elzo de Wit. Wapl maintains a cohesin loading cycle to preserve cell-type-specific distal gene regulation. *Nature genetics*, 53(1):100–109, 2021.

[198] Sijia Liu, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Pin-Yu Chen, Gerald Higgins, Pierre Baldi, Steve Smale, et al. Genome architecture mediates transcriptional control of human myogenic reprogramming. *Iscience*, 6:232–246, 2018.

[199] Tian-kun Liu, Yuan Pang, Zhen-zhen Zhou, Rui Yao, and Wei Sun. An integrated cell printing system for the construction of heterogeneous tissue models. *Acta Biomaterialia*, 95:245–257, 2019.

[200] Y Liu. Y., slotine. *J.–J. & Barabási, A.–L. Controllability of complex networks. Nature*, 473:167–176, 2011.

[201] Yen-Wen Liu, Billy Chen, Xiulan Yang, James A Fugate, Faith A Kalucki, Akiko Futakuchi-Tsuchida, Larry Couture, Keith W Vogel, Clifford A Astley, Audrey Baldessari, et al. Human embryonic stem cell–derived cardiomyocytes restore function in infarcted hearts of non-human primates. *Nature biotechnology*, 36(7):597–605, 2018.

[202] Haiping Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Mpca: Multilinear principal component analysis of tensor objects. *IEEE transactions on Neural Networks*, 19(1):18–39, 2008.

[203] Carey N Lumeng, Jennifer B DelProposto, Daniel J Westcott, and Alan R Saltiel. Phenotypic switching of adipose tissue macrophages with obesity is generated by spatiotemporal differences in macrophage subtypes. *Diabetes*, 57(12):3239–3246, 2008.

[204] Carey N Lumeng, Stephanie M DeYoung, Jennifer L Bodzin, and Alan R Saltiel. Increased inflammatory properties of adipose tissue macrophages recruited during diet-induced obesity. *Diabetes*, 56(1):16–23, 2007.

[205] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[206] Ben D MacArthur and Ihor R Lemischka. Statistical mechanics of pluripotency. *Cell*, 154(3):484–489, 2013.

[207] Ben D MacArthur, Avi Ma'ayan, and Ihor R Lemischka. Systems biology of stem cell fate and cellular reprogramming. *Nature reviews Molecular cell biology*, 10(10):672–681, 2009.

[208] Yoshinobu Maeda, Isao Tawara, Takanori Teshima, Chen Liu, Daigo Hashimoto, Ken-ichi Matsuoka, Mitsune Tanimoto, and Pavan Reddy. Lymphopenia-induced proliferation of donor t cells reduces their capacity for causing acute graft-versus-host disease. *Experimental hematology*, 35(2):274–286, 2007.

[209] Michiko Mandai, Akira Watanabe, Yasuo Kurimoto, Yasuhiko Hirami, Chikako Morinaga, Takashi Daimon, Masashi Fujihara, Hiroshi Akimaru, Noriko Sakai, Yumiko Shibata, et al. Autologous induced stem-cell–derived retinal cells for macular degeneration. *New England Journal of Medicine*, 376(11):1038–1046, 2017.

[210] Silas Maniatis, Tarmo Äijö, Sanja Vickovic, Catherine Braine, Kristy Kang, Annelie Mollbrink, Delphine Fagegaltier, Žaneta Andrusivová, Sami Saarenpää, Gonzalo Saiz-Castro, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435):89–93, 2019.

[211] Silas Maniatis, Tarmo Äijö, Sanja Vickovic, Catherine Braine, Kristy Kang, Annelie Mollbrink, Delphine Fagegaltier, Žaneta Andrusivová, Sami Saarenpää, Gonzalo Saiz-Castro, Miguel Cuevas, Aaron Watters, Joakim Lundeberg, Richard Bonneau, and Hemali Phatnani. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435):89–93, 2019.

[212] B Mann, M Gelos, A Siedow, ML Hanski, A Gratchev, M Ilyas, WF Bodmer, MP Moyer, EO Riecken, HJ Buhr, et al. Target genes of β-catenin–t cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas. *Proceedings of the National Academy of Sciences*, 96(4):1603–1608, 1999.

[213] Jessica Mariani, Gianfilippo Coppola, Ping Zhang, Alexej Abyzov, Lauren Provini, Livia Tomasini, Mariangela Amenduni, Anna Szekely, Dean Palejev, Michael Wilson, et al. Foxg1-dependent dysregulation of gaba/glutamate neuron differentiation in autism spectrum disorders. *Cell*, 162(2):375–390, 2015.

[214] Marc A Marti-Renom, Genevieve Almouzni, Wendy A Bickmore, Kerstin Bystricky, Giacomo Cavalli, Peter Fraser, Susan M Gasser, Luca Giorgetti, Edith Heard, Mario Nicodemi, et al. Challenges and guidelines toward 4d nucleome data and model standards. *Nature genetics*, 50(10):1352–1358, 2018.

[215] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[216] Julie Mathieu, Zhan Zhang, Wenyu Zhou, Amy J Wang, John M Heddleston, Claudia MA Pinna, Alexis Hubaud, Bradford Stadler, Michael Choi, Merav Bar, et al. Hif induces human embryonic stem cell markers in cancer cells. *Cancer research*, 71(13):4640–4652, 2011.

[217] Tracey McLaughlin, Shelley E Ackerman, Lei Shen, Edgar Engleman, et al. Role of innate and adaptive immunity in obesity-associated metabolic disease. *The Journal of clinical investigation*, 127(1):5–13, 2017.

[218] Giorgia Minello, Luca Rossi, and Andrea Torsello. On the von neumann entropy of graphs. *Journal of Complex Networks*, 7(4):491–514, 2019.

[219] Leonid A Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research*, 19(1):37–51, 2011.

[220] Tom Misteli. The self-organizing genome: Principles of genome architecture and function. *Cell*, 183:28–45, 2020.

[221] Ichiro Misumi, Joshua Starmer, Toru Uchimura, Melinda A. Beck, Terry Magnuson, and Jason K. Whitmire. Obesity expands a distinct population of t cells in adipose tissue and increases vulnerability to infection. *Cell Reports*, 27(2):514–524.e5, 2019.

[222] Norikatsu Miyoshi, Hideshi Ishii, Ken-ichi Nagai, Hiromitsu Hoshino, Koshi Mimori, Fumiaki Tanaka, Hiroaki Nagano, Mitsugu Sekimoto, Yuichiro Doki, and Masaki Mori. Defined factors induce reprogramming of gastrointestinal cancer cells. *Proceedings of the National Academy of Sciences*, 107(1):40–45, 2010.

[223] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C Devlin, Maayan Baron, Cristina H Hajdu, Diane M Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology*, 38(3):333–342, 2020.

[224] Lindsey Montefiori, Robert Wuerffel, Damian Roqueiro, Bryan Lajoie, Changying Guo, Tatiana Gerasimova, Supriyo De, William Wood, Kevin G Becker, Job Dekker, et al. Extremely long-range chromatin loops link topological domains to facilitate a diverse antibody repertoire. *Cell reports*, 14(4):896–906, 2016.

[225] Patrice J Morin, Andrew B Sparks, Vladimir Korinek, Nick Barker, Hans Clevers, Bert Vogelstein, and Kenneth W Kinzler. Activation of β-catenin-tcf signaling in colon cancer by mutations in β-catenin or apc. *Science*, 275(5307):1787–1790, 1997.

[226] David L Morris, Kae Won Cho, Jennifer L DelProposto, Kelsie E Oatmen, Lynn M Geletka, Gabriel Martinez-Santibanez, Kanakadurga Singer, and Carey N Lumeng. Adipose tissue macrophages function as antigen-presenting cells and regulate adipose tissue cd4+ t cells in mice. *Diabetes*, 62(8):2762–2772, 2013.

[227] Samantha A Morris, Patrick Cahan, Hu Li, Anna M Zhao, Adrianna K San Roman, Ramesh A Shivdasani, James J Collins, and George Q Daley. Dissecting engineered cell types and enhancing cell fate conversion via cellnet. *Cell*, 158(4):889–902, 2014.

[228] Samantha A Morris and George Q Daley. A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell Research*, 23(1):33–48, 2013.

[229] Christian Mosimann, George Hausmann, and Konrad Basler. Parafibromin/hyrax activates wnt/wg target gene transcription by direct association with β-catenin/armadillo. *Cell*, 125(2):327–341, 2006.

[230] Lindsey A Muir, Kae Won Cho, Lynn M Geletka, Nicki A Baker, Carmen G Flesher, Anne P Ehlers, Niko Kaciroti, Stephen Lindsly, Scott Ronquist, Indika Rajapakse, et al. Human cd206+ macrophages associate with diabetes and adipose tissue lymphoid clusters. *JCI insight*, 2022.

[231] Lindsey A Muir, Samadhi Kiridena, Cameron Griffin, Jennifer B DelProposto, Lynn Geletka, Gabriel Martinez-Santibañez, Brian F Zamarron, Hannah Lucas, Kanakadurga Singer, Robert W O'Rourke, et al. Frontline science: Rapid adipose tissue expansion triggers unique proliferation and lipid accumulation profiles in adipose tissue macrophages. *Journal of leukocyte biology*, 103(4):615–628, 2018.

[232] Maeve Mullooly, Hannah P Yang, Roni T Falk, Sarah J Nyante, Renata Cora, Ruth M Pfeiffer, Derek C Radisky, Daniel W Visscher, Lynn C Hartmann, Jodi M Carter, et al. Relationship between crown-like structures and sex-steroid hormones in breast adipose tissue and serum among postmenopausal breast cancer patients. *Breast Cancer Research*, 19(1):1–10, 2017.

[233] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012.

[234] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[235] Kim Nasmyth and Christian H Haering. Cohesin: its roles and mechanisms. *Annual review of genetics*, 43:525–558, 2009.

[236] Abdolrahman S Nateri, Bradley Spencer-Dene, and Axel Behrens. Interaction of phosphorylated c-jun with tcf4 regulates intestinal cancer development. *Nature*, 437(7056):281–285, 2005.

[237] Gioacchino Natoli. Maintaining cell identity through global control of genomic organization. *Immunity*, 33(1):12–24, 2010.

[238] Natalia Naumova, Maxim Imakaev, Geoffrey Fudenberg, Ye Zhan, Bryan R Lajoie, Leonid A Mirny, and Job Dekker. Organization of the mitotic chromosome. *Science*, 342(6161):948–953, 2013.

[239] Shane Neph, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286, 2012.

[240] MEJ Newman. Networks: An introduction. 2010 oxford.

[241] Guy Nir, Irene Farabella, Cynthia Pérez Estrada, Carl G Ebeling, Brian J Beliveau, Hiroshi M Sasaki, S Dean Lee, Son C Nguyen, Ruth B McCole, Shyamtanu Chattoraj, et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS genetics*, 14(12):e1007872, 2018.

[242] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L Van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[243] Pedro Olivares-Chauvet, Zohar Mukamel, Aviezer Lifshitz, Omer Schwartzman, Noa Oded Elkayam, Yaniv Lubling, Gintaras Deikus, Robert P Sebra, and Amos Tanay. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature*, 540(7632):296–300, 2016.

[244] Raquel A Oliveira and Kim Nasmyth. Cohesin cleavage is insufficient for centriole disengagement in drosophila. *Current Biology*, 23(14):R601–R603, 2013.

[245] Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12(1):1–10, 2011.

[246] Chin-Tong Ong and Victor G Corces. Ctcf: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4):234–246, 2014.

[247] Cameron S Osborne, Lyubomira Chakalova, Karen E Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A Mitchell, Susana Lopes, Wolf Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, 36(10):1065–1071, 2004.

[248] Cameron S Osborne and Christopher H Eskiw. Where shall we meet? a role for genome organisation and nuclear sub-compartments in mediating interchromosomal interactions. *Journal of cellular biochemistry*, 104(5):1553–1561, 2008.

[249] A Marieke Oudelaar, James OJ Davies, Lars LP Hanssen, Jelena M Telenius, Ron Schwessinger, Yu Liu, Jill M Brown, Damien J Downes, Andrea M Chiariello, Simona Bianco, et al. Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nature genetics*, 50(12):1744–1751, 2018.

[250] John F Ouyang, Uma S Kamaraj, Jose M Polo, Julian Gough, and Owen JL Rackham. Molecular interaction networks to select factors for cell conversion. In *Computational Stem Cell Biology*, pages 333–361. Springer, 2019.

[251] Robert-Jan Palstra, Bas Tolhuis, Erik Splinter, Rian Nijmeijer, Frank Grosveld, and Wouter de Laat. The β-globin nuclear compartment in development and erythroid differentiation. *Nature genetics*, 35(2):190–194, 2003.

[252] Vania Parelho, Suzana Hadjur, Mikhail Spivakov, Marion Leleu, Stephan Sauer, Heather C Gregson, Adam Jarmuz, Claudia Canzonetta, Zoe Webster, Tatyana Nesterova, et al. Cohesins functionally associate with ctcf on mammalian chromosome arms. *Cell*, 132(3):422–433, 2008.

[253] David S Parker, Yunyun Y Ni, Jinhee L Chang, Jiong Li, and Ken M Cadigan. Wingless signaling induces widespread chromatin remodeling of target loci. *Molecular and cellular biology*, 28(5):1815–1828, 2008.

[254] Filippo Passerini and Simone Severini. The von neumann entropy of networks. *arXiv:0812.2597*, 2008.

[255] Evan O Paull, Alvaro Aytes, Sunny J Jones, Prem S Subramaniam, Federico M Giorgi, Eugene F Douglass, Somnath Tagore, Brennan Chu, Alessandro Vasciaveo, Siyuan Zheng, et al. A modular master regulator landscape controls cancer transcriptional identity. *Cell*, 184(2):334–351, 2021.

[256] Jan-Michael Peters and Tomoko Nishiyama. Sister chromatid cohesion. *Cold Spring Harbor Perspectives in Biology*, 4(11):a011130, 2012.

[257] Jennifer E Phillips and Victor G Corces. Ctcf: master weaver of the genome. *Cell*, 137(7):1194–1211, 2009.

[258] Rob Phillips, Jane Kondev, Julie Theriot, Hernan G Garcia, and Nigel Orme. *Physical biology of the cell*. Garland Science, 2012.

[259] Jessica Piché, Patrick Piet Van Vliet, Michel Pucéat, and Gregor Andelfinger. The expanding phenotypes of cohesinopathies: one ring to rule them all! *Cell Cycle*, 18(21):2828–2848, 2019.

[260] Cara E Porsche, Jennifer B Delproposto, Lynn Geletka, Robert O'Rourke, and Carey N Lumeng. Obesity results in adipose tissue t cell exhaustion. *JCI insight*, 6(8), 2021.

[261] Cara E Porsche, Jennifer B Delproposto, Elise Patrick, Brian F Zamarron, and Carey N Lumeng. Adipose tissue dendritic cell signals are required to maintain t cell homeostasis and obesity-induced expansion. *Molecular and cellular endocrinology*, 505:110740, 2020.

[262] Elena A Pudova, Anna V Kudryavtseva, Maria S Fedorova, Andrew R Zaretsky, Dmitry S Shcherbo, Elena N Lukyanova, Anatoly Y Popov, Asiya F Sadritdinova, Ivan S Abramov, Sergey L Kharitonov, et al. Hk3 overexpression associated with epithelial-mesenchymal transition in colorectal cancer. *BMC genomics*, 19(3):5–13, 2018.

[263] Sofia A Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M Lai, Alexander A Shishkin, Prashant Bhat, Yodai Takei, et al. Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus. *Cell*, 174(3):744–757, 2018.

[264] Owen JL Rackham, Jaber Firas, Hai Fang, Matt E Oates, Melissa L Holmes, Anja S Knaupp, Harukazu Suzuki, Christian M Nefzger, Carsten O Daub, Jay W Shin, et al. A predictive computational framework for direct reprogramming between human cell types. *Nature Genetics*, 48(3):331, 2016.

[265] Arthur H Radley, Remy M Schwab, Yuqi Tan, Jeesoo Kim, Emily KW Lo, and Patrick Cahan. Assessment of engineered cells using cellnet and rna-seq. *Nature Protocols*, 12(5):1089–1102, 2017.

[266] Indika Rajapakse and Mark Groudine. On emerging nuclear order. *Journal of Cell Biology*, 192(5):711–721, 2011.

[267] Indika Rajapakse, Mark Groudine, and Mehran Mesbahi. Dynamics and control of state-dependent networks for probing genomic organization. *Proceedings of the National Academy of Sciences*, 108(42):17257–17262, 2011.

[268] Indika Rajapakse, Mark Groudine, and Mehran Mesbahi. What can systems theory of networks offer to biology? *PLoS Comput Biol*, 8(6):e1002543, 2012.

[269] Indika Rajapakse, Michael D Perlman, David Scalzo, Charles Kooperberg, Mark Groudine, and Steven T Kosak. The emergence of lineage-specific chromosomal topologies from coordinate gene regulation. *Proceedings of the National Academy of Sciences*, 106(16):6679–6684, 2009.

[270] Indika Rajapakse and Stephen Smale. Emergence of function from coordinated cells in a tissue. *Proceedings of the National Academy of Sciences*, 114(7):1462–1467, 2017.

[271] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14(3):263–266, 2017.

[272] FACL Ran, Le Cong, Winston X Yan, David A Scott, Jonathan S Gootenberg, Andrea J Kriz, Bernd Zetsche, Ophir Shalem, Xuebing Wu, Kira S Makarova, et al. In vivo genome editing using staphylococcus aureus cas9. *Nature*, 520(7546):186–191, 2015.

[273] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[274] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[275] Jason S Rawlings, Martina Gatzka, Paul G Thomas, and James N Ihle. Chromatin condensation via the condensin ii complex is required for peripheral t-cell quiescence. *The EMBO journal*, 30(2):263–276, 2011.

[276] Pavan Reddy, Robert Negrin, and Geoffrey R Hill. Mouse models of bone marrow transplantation. *Biology of Blood and Marrow Transplantation*, 14(1):129–135, 2008.

[277] Andreas M Reichmuth, Matthias A Oberli, Ana Jaklenec, Robert Langer, and Daniel Blankschtein. mrna vaccine delivery using lipid nanoparticles. *Therapeutic delivery*, 7(5):319–334, 2016.

[278] Sarah B Reiff, Andrew J Schroeder, Koray Kırlı, Andrea Cosolo, Clara Bakker, Soohyun Lee, Alexander D Veit, Alexander K Balashov, Carl Vitzthum, William Ronchetti, et al. The 4d nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nature communications*, 13(1):1–11, 2022.

[279] Shannon M Reilly and Alan R Saltiel. Adapting to obesity with adipose tissue inflammation. *Nature Reviews Endocrinology*, 13(11):633–643, 2017.

[280] Silvia Remeseiro, Ana Cuadrado, and Ana Losada. Cohesin in development and disease. *Development*, 140(18):3715–3718, 2013.

[281] Bing Ren, Hieu Cam, Yasuhiko Takahashi, Thomas Volkert, Jolyon Terragni, Richard A Young, and Brian David Dynlacht. E2f integrates cell cycle progression with dna repair, replication, and g2/m checkpoints. *Genes & development*, 16(2):245–256, 2002.

[282] Meiling Ren, Huanji Xu, Hongwei Xia, Qiulin Tang, and Feng Bi. Simultaneously targeting soat1 and cpt1a ameliorates hepatocellular carcinoma by disrupting lipid homeostasis. *Cell death discovery*, 7(1):1–15, 2021.

[283] Jonah Riddell, Roi Gazit, Brian S Garrison, Guoji Guo, Assieh Saadatpour, Pankaj K Mandal, Wataru Ebina, Pavel Volchkov, Guo-Cheng Yuan, Stuart H Orkin, et al. Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. *Cell*, 157(3):549–564, 2014.

[284] Thomas Ried and Indika Rajapakse. The 4d nucleome. *Methods (San Diego, Calif.)*, 123:1–2, 2017.

[285] Michael I Robson, I Jose, Rafal Czapiewski, Aishwarya Sivakumar, Alastair RW Kerr, and Eric C Schirmer. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during t-cell activation facilitates their association in chromosome compartments. *Genome research*, 27(7):1126–1138, 2017.

[286] Jeffrey J Roix, Philip G McQueen, Peter J Munson, Luis A Parada, and Tom Misteli. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nature genetics*, 34(3):287–291, 2003.

[287] Elizabeth A Rondini and James G Granneman. Single cell approaches to address adipose tissue stromal cell heterogeneity. *Biochemical Journal*, 477(3):583–600, 2020.

[288] Scott Ronquist, Geoff Patterson, Lindsey A Muir, Stephen Lindsly, Haiming Chen, Markus Brown, Max S Wicha, Anthony Bloch, Roger Brockett, and Indika Rajapakse. Algorithm for cellular reprogramming. *Proceedings of the National Academy of Sciences*, 114(45):11832–11837, 2017.

[289] Fábio F Rosa, Cristiana F Pires, Ilia Kurochkin, Alexandra G Ferreira, Andreia M Gomes, Luís G Palma, Kritika Shaiv, Laura Solanas, Cláudia Azenha, Dmitri Papatsenko, et al. Direct reprogramming of fibroblasts into antigen-presenting dendritic cells. *Science Immunology*, 3(30), 2018.

[290] JOSEPH F Ross and MICHAEL Orlowski. Growth-rate-dependent adjustment of ribosome function in chemostat-grown cells of the fungus mucor racemosus. *Journal of Bacteriology*, 149(2):650–653, 1982.

[291] AJ Rowan, H Lamlum, M Ilyas, J Wheeler, J Straub, A Papadopoulou, D Bicknell, WF Bodmer, and IPM Tomlinson. Apc mutations in sporadic colorectal tumors: a mutational "hotspot" and interdependence of the "two hits". *Proceedings of the National Academy of Sciences*, 97(7):3352–3357, 2000.

[292] Christoph Ruschil, Gisela Gabernet, Gildas Lepennetier, Simon Heumos, Miriam Kaminski, Zsuzsanna Hracsko, Martin Irmler, Johannes Beckers, Ulf Ziemann, Sven Nahnsen, et al. Specific induction of double negative b cells during protective and pathogenic immune responses. *Frontiers in immunology*, page 3304, 2020.

[293] Brendan E Russ, Moshe Olshanksy, Heather S Smallwood, Jasmine Li, Alice E Denton, Julia E Prier, Angus T Stock, Hayley A Croom, Jolie G Cullen, Michelle LT Nguyen, et al. Distinct epigenetic signatures delineate transcriptional programs during virus-specific cd8+ t cell differentiation. *Immunity*, 41(5):853–865, 2014.

[294] Lucia Russo and Carey N Lumeng. Properties and functions of adipose tissue macrophages in obesity. *Immunology*, 155(4):407–417, 2018.

[295] Wang-Shick Ryu. *Molecular virology of human pathogenic viruses*. Academic Press, 2016.

[296] Elias A Said, Nicolas Tremblay, Mohammed S Al-Balushi, Ali A Al-Jabri, and Daniel Lamarre. Viruses seen by our cells: the role of viral rna sensors. *Journal of Immunology Research*, 2018, 2018.

[297] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[298] Marc Santolini and Albert-László Barabási. Predicting perturbation patterns from the topology of biological networks. *Proceedings of the National Academy of Sciences*, 115(27):E6375–E6383, 2018.

[299] Abby Sarkar, Aaron J Huebner, Rita Sulahian, Anthony Anselmo, Xinsen Xu, Kyle Flattery, Niyati Desai, Carlos Sebastian, Mary Anna Yram, Katrin Arnold, et al. Sox2 suppresses gastric tumorigenesis in mice. *Cell reports*, 16(7):1929–1941, 2016.

[300] Toshiro Sato, Johan H Van Es, Hugo J Snippert, Daniel E Stange, Robert G Vries, Maaike Van Den Born, Nick Barker, Noah F Shroyer, Marc Van De Wetering, and Hans Clevers. Paneth cells constitute the niche for lgr5 stem cells in intestinal crypts. *Nature*, 469(7330):415–418, 2011.

[301] Alexandra Schebesta, Shane McManus, Giorgia Salvagiotto, Alessio Delogu, Georg A Busslinger, and Meinrad Busslinger. Transcription factor pax5 activates the chromatin of key genes involved in b cell signaling, adhesion, migration, and immune function. *Immunity*, 27(1):49–63, 2007.

[302] Thorsten M Schlaeger, Laurence Daheron, Thomas R Brickler, Samuel Entwisle, Karrie Chan, Amelia Cianci, Alexander DeVine, Andrew Ettenger, Kelly Fitzgerald, Michelle Godfrey, et al. A comparison of non-integrating reprogramming methods. *Nature Biotechnology*, 33(1):58–63, 2015.

[303] Stefan Schoenfelder, Tom Sexton, Lyubomira Chakalova, Nathan F Cope, Alice Horton, Simon Andrews, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Daniela S Dimitrova, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, 42(1):53–61, 2010.

[304] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A Fonseca, Wolfgang Huber, Christian H Haering, Leonid Mirny, et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51–56, 2017.

[305] Laura Seaman, Haiming Chen, Markus Brown, Darawalee Wangsa, Geoff Patterson, Jordi Camps, Gilbert S Omenn, Thomas Ried, and Indika Rajapakse. Nucleome analysis reveals structure–function relationships for colon cancer. *Molecular Cancer Research*, 15(7):821–830, 2017.

[306] Laura Seaman and Indika Rajapakse. 4d nucleome analysis toolbox: analysis of hi-c data with abnormal karyotype and time series capabilities. *Bioinformatics*, 34(1):104–106, 2018.

[307] Vlad C Seitan, Bingtao Hao, Kikuë Tachibana-Konwalski, Thais Lavagnolli, Hegias Mira-Bontenbal, Karen E Brown, Grace Teng, Tom Carroll, Anna Terry, Katie Horan, et al. A role for cohesin in t-cell-receptor rearrangement and thymocyte differentiation. *Nature*, 476(7361):467–471, 2011.

[308] Alexey Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*, page 060012, 2016.

[309] Omar Sharif, Julia Stefanie Brunner, Ana Korosec, Rui Martins, Alexander Jais, Berend Snijder, Andrea Vogel, Michael Caldera, Anastasiya Hladik, Karin Lakovits, et al. Beneficial metabolic effects of trem2 in obesity are uncoupled from its expression on macrophages. *Diabetes*, 70(9):2042–2057, 2021.

[310] Ruchi Sharma, Imke PM Smits, Laura De La Vega, Christopher Lee, and Stephanie M Willerth. 3d bioprinting pluripotent stem cell derived neural tissues using a novel fibrin

bioink containing drug releasing microspheres. *Frontiers in bioengineering and biotechnology*, 8:57, 2020.

[311] Yin Shen, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.

[312] Yanhong Shi, Haruhisa Inoue, Joseph C Wu, and Shinya Yamanaka. Induced pluripotent stem cell technology: a decade of progress. *Nature reviews Drug discovery*, 16(2):115–130, 2017.

[313] Mariana CC Silva, Sean Powell, Sabrina Ladstätter, Johanna Gassler, Roman Stocsits, Antonio Tedeschi, Jan-Michael Peters, and Kikuë Tachibana. Wapl releases scc1-cohesin and regulates chromosome structure and segregation in mouse oocytes. *Journal of Cell Biology*, 219(4), 2020.

[314] Rita Silvério-Alves, Andreia M Gomes, Ilia Kurochkin, Kateri A Moore, and Carlos-Filipe Pereira. Hemogenic reprogramming of human fibroblasts by enforced expression of transcription factors. *JoVE (Journal of Visualized Experiments)*, (153):e60112, 2019.

[315] Vijay Pratap Singh and Jennifer L Gerton. Cohesin and human disease: lessons from mouse models. *Current Opinion in Cell Biology*, 37:9–17, 2015.

[316] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[317] Charalampos G Spilianakis and Richard A Flavell. Long-range intrachromosomal interactions in the t helper type 2 cytokine locus. *Nature immunology*, 5(10):1017–1027, 2004.

[318] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613–626, 2012.

[319] Madhusudhan Srinivasan, Naomi J Petela, Johanna C Scheinost, James Collier, Menelaos Voulgaris, Maurici B Roig, Frederic Beckouët, Bin Hu, and Kim A Nasmyth. Scc2 counteracts a wapl-independent mechanism that releases cohesin from chromosomes during g1. *Elife*, 8:e44736, 2019.

[320] Deepak Srivastava and Natalie DeWitt. In vivo cellular reprogramming: the next generation. *Cell*, 166(6):1386–1396, 2016.

[321] Ralph Stadhouders, Guillaume J Filion, and Thomas Graf. Transcription factors and 3d genome conformation in cell-fate decisions. *Nature*, 569(7756):345–354, 2019.

[322] Ralph Stadhouders, Enrique Vidal, François Serra, Bruno Di Stefano, François Le Dily, Javier Quilez, Antonio Gomez, Samuel Collombet, Clara Berenguer, Yasmina Cuartero, et al. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature genetics*, 50(2):238–249, 2018.

[323] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[324] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59–64, 2017.

[325] Aaron F Straight, Amy Cheung, John Limouze, Irene Chen, Nick J Westwood, James R Sellers, and Timothy J Mitchison. Dissecting temporal and spatial control of cytokinesis with a myosin ii inhibitor. *Science*, 299(5613):1743–1747, 2003.

[326] Gilbert Strang. *Introduction to linear algebra*, volume 3. Wellesley-Cambridge Press Wellesley, MA, 1993.

[327] Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268–276, 2001.

[328] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[329] Takayoshi Suganami and Yoshihiro Ogawa. Adipose tissue macrophages: their role in adipose tissue remodeling. *Journal of leukocyte biology*, 88(1):33–39, 2010.

[330] Yaping Sun, Gabrielle A Dotson, Lindsey A Muir, Scott Ronquist, Katherine Oravecz-Wilson, Daniel Peltier, Keisuke Seike, Lu Li, Walter Meixner, Indika Rajapakse, et al. Rearrangement of t cell genome architecture regulates gvhd. *Available at SSRN 4022705*.

[331] Yaping Sun, Katherine Oravecz-Wilson, Sydney Bridges, Richard McEachin, Julia Wu, Stephanie H Kim, Austin Taylor, Cynthia Zajac, Hideaki Fujiwara, Daniel Christopher Peltier, et al. mir-142 controls metabolic reprogramming that regulates dendritic cell activation. *The Journal of clinical investigation*, 129(5):2029–2042, 2019.

[332] Yaping Sun, Katherine Oravecz-Wilson, Nathan Mathewson, Ying Wang, Richard McEachin, Chen Liu, Tomomi Toubai, Julia Wu, Corinne Rossi, Thomas Braun, et al. Mature t cell responses are controlled by microrna-142. *The Journal of clinical investigation*, 125(7):2825–2840, 2015.

[333] Yaping Sun, Isao Tawara, Meng Zhao, Zhaohui S Qin, Tomomi Toubai, Nathan Mathewson, Hiroya Tamaki, Evelyn Nieves, Arul M Chinnaiyan, Pavan Reddy, et al. Allogeneic t cell responses are regulated by a specific mirna-mrna network. *The Journal of clinical investigation*, 123(11):4739–4754, 2013.

[334] Amit Surana, Can Chen, and Indika Rajapakse. Hypergraph dissimilarity measures. *arXiv preprint arXiv:2106.08206*, 2021.

[335] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[336] Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. Principles of genome folding into topologically associating domains. *Science advances*, 5(4):eaaw1668, 2019.

[337] Mohammadsharif Tabebordbar, Kexian Zhu, Jason KW Cheng, Wei Leong Chew, Jeffrey J Widrick, Winston X Yan, Claire Maesner, Elizabeth Y Wu, Ru Xiao, F Ann Ran, et al. In vivo gene editing in dystrophic mouse muscle and muscle stem cells. *Science*, 351(6271):407–411, 2016.

[338] Masahito Tachibana, Paula Amato, Michelle Sparman, Nuria Marti Gutierrez, Rebecca Tippner-Hedges, Hong Ma, Eunju Kang, Alimujiang Fulati, Hyo-Sang Lee, Hathaitip Sritanaudomchai, et al. Human embryonic stem cells derived by somatic cell nuclear transfer. *Cell*, 153(6):1228–1238, 2013.

[339] Kazutoshi Takahashi. Cellular reprogramming–lowering gravity on waddington's epigenetic landscape. *Journal of Cell Science*, 125(11):2553–2560, 2012.

[340] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *cell*, 131(5):861–872, 2007.

[341] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

[342] Meiko Takahashi, Yusuke Nakamura, Kazutaka Obama, and Yoichi Furukawa. Identification of sp5 as a downstream gene of the β-catenin/tcf pathway and its enhanced expression in human colon cancer. *International journal of oncology*, 27(6):1483–1487, 2005.

[343] Hideyuki Tanabe, Felix A Habermann, Irina Solovei, Marion Cremer, and Thomas Cremer. Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 504(1-2):37–45, 2002.

[344] Antonio Tedeschi, Gordana Wutz, Sébastien Huet, Markus Jaritz, Annelie Wuensche, Erika Schirghuber, Iain Finley Davidson, Wen Tang, David A Cisneros, Venugopal Bhaskara, et al. Wapl is an essential regulator of chromatin structure and chromosome segregation. *Nature*, 501(7468):564–568, 2013.

[345] Osamu Tetsu and Frank McCormick. β-catenin regulates expression of cyclin d1 in colon carcinoma cells. *Nature*, 398(6726):422–426, 1999.

[346] Payal Tiwari, Ariane Blank, Chang Cui, Kelly Q Schoenfelt, Guolin Zhou, Yanfei Xu, Galina Khramtsova, Funmi Olopade, Ajay M Shah, Seema A Khan, et al. Metabolically activated adipose tissue macrophages link obesity to triple-negative breast cancer. *Journal of Experimental Medicine*, 216(6):1345–1358, 2019.

[347] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[348] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

[349] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[350] Larissa Traxler, Frank Edenhofer, and Jerome Mertens. Next-generation disease modeling with direct conversion: a new path to old neurons. *FEBS letters*, 593(23):3316–3337, 2019.

[351] Hui C Tsou, Xinhua Lee, Seong Pan Si, and Monica Peacocke. Regulation of retinoic acid receptor expression in dermal fibroblasts. *Experimental cell research*, 211(1):74–81, 1994.

[352] Hiromichi Tsuruga, Norikazu Yabuta, Katsuhito Hashizume, Masako Ikeda, Yuichi Endo, and Hiroshi Nojima. Expression, nuclear localization and interactions of human mcm/p1 proteins. *Biochemical and biophysical research communications*, 236(1):118–125, 1997.

[353] Alan Mathison Turing. The chemical basis of morphogenesis. *Bulletin of mathematical biology*, 52(1):153–197, 1990.

[354] AM Turing. The chemical basis of morphogenesis. *Philos. Trans. Roy. Soc.*, 1952.

[355] Netha Ulahannan, Matthew Pendleton, Aditya Deshpande, Stefan Schwenk, Julie M Behr, Xiaoguang Dai, Carly Tyer, Priyesh Rughani, Sarah Kudman, Emily Adney, et al. Nanopore sequencing of dna concatemers reveals higher-order features of chromatin structure. *bioRxiv*, page 833590, 2019.

[356] AG Uren, Jaap Kool, A Berns, and M Van Lohuizen. Retroviral insertional mutagenesis: past, present and future. *Oncogene*, 24(52):7656–7672, 2005.

[357] Jochen Utikal, Nimet Maherali, Warakorn Kulalert, and Konrad Hochedlinger. Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells. *Journal of cell science*, 122(19):3502–3510, 2009.

[358] Tomas Vacik, Jennifer L Stubbs, and Greg Lemke. A novel mechanism for the transcriptional regulation of wnt signaling in development. *Genes & development*, 25(17):1783–1795, 2011.

[359] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE transactions on visualization and computer graphics*, 27(1):1–13, 2019.

[360] Marc Van De Wetering, Elena Sancho, Cornelis Verweij, Wim De Lau, Irma Oving, Adam Hurlstone, Karin Van Der Horn, Eduard Batlle, Damien Coudreuse, Anna-Pavlina Haramis, et al. The β-catenin/tcf-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell*, 111(2):241–250, 2002.

[361] Susanne C van den Brink, Anna Alemany, Vincent van Batenburg, Naomi Moris, Marloes Blotenburg, Judith Vivié, Peter Baillie-Johnson, Jennifer Nichols, Katharina F Sonnen, Alfonso Martinez Arias, et al. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*, 582(7812):405–409, 2020.

[362] Johan H van Es, Andrea Haegebarth, Pekka Kujala, Shalev Itzkovitz, Bon-Kyoung Koo, Sylvia F Boj, Jeroen Korving, Maaike van den Born, Alexander van Oudenaarden, Sylvie Robine, et al. A critical role for the wnt effector tcf4 in adult intestinal homeostatic self-renewal. *Molecular and cellular biology*, 32(10):1918–1927, 2012.

[363] Phuc Van Pham, Ngoc Bich Vu, Thuy Thi-Thanh Dao, Ha Thi-Ngan Le, Lan Thi Phi, and Ngoc Kim Phan. Production of endothelial progenitor cells from skin fibroblasts by direct reprogramming for clinical usages. *In Vitro Cellular & Developmental Biology-Animal*, 53(3):207–216, 2017.

[364] Anne van Schoonhoven, Danny Huylebroeck, Rudi W Hendriks, and Ralph Stadhouders. 3d genome organization during lymphocyte development and activation. *Briefings in functional genomics*, 19(2):71–82, 2020.

[365] Jinchu Vijay, Marie-Frédérique Gauthier, Rebecca L Biswell, Daniel A Louiselle, Jeffrey J Johnston, Warren A Cheung, Bradley Belden, Albena Pramatarova, Laurent Biertho, Margaret Gibson, et al. Single-cell analysis of human adipose tissue identifies depot-and disease-specific cell types. *Nature metabolism*, 2(1):97–109, 2020.

[366] Campbell KH McWhir J Ritchie WA and I Wilmut. Sheep cloned by nuclear transfer from a cultured cell line nature 3806466. *Campbell, KH, McWhir, J., Ritchie, WA, and Wilmut, I.(1996). Sheep cloned by nuclear transfer from a cultured cell line. Nature*, 380:64–66, 1996.

[367] Conrad Hall Waddington. *The Strategy of the Genes, a Discussion of Some Aspects of Theoretical Biology*. G. Allen and Unwin, 1957.

[368] Todd Waldman. Emerging themes in cohesin cancer biology. *Nature Reviews Cancer*, 20(9):504–515, 2020.

[369] Michael L Wallace, Kee Wui Huang, Daniel Hochbaum, Minsuk Hyun, Gianna Radeljic, and Bernardo L Sabatini. Anatomical and single-cell transcriptional profiling of the murine habenular complex. *Elife*, 9:e51271, 2020.

[370] Chao Wang, Weiyi Liu, Yaohui Nie, Mulan Qaher, Hannah Elizabeth Horton, Feng Yue, Atsushi Asakura, and Shihuan Kuang. Loss of myod promotes fate transdifferentiation of myoblasts into brown adipocytes. *EBioMedicine*, 16:212–223, 2017.

[371] Jianglin Wang, Xueyan Jiang, Lixin Zhao, Shengjia Zuo, Xiantong Chen, Lingmin Zhang, Zhongxiao Lin, Xiaoya Zhao, Yuyan Qin, Xinke Zhou, et al. Lineage reprogramming of fibroblasts into induced cardiac progenitor cells by crispr/cas9-based transcriptional activators. *Acta Pharmaceutica Sinica B*, 10(2):313–326, 2020.

[372] Liguo Wang, Shengqin Wang, and Wei Li. Rseqc: quality control of rna-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012.

[373] Wenqi Wang, Vaneet Aggarwal, and Shuchin Aeron. Principal component analysis with tensor train subspace. *Pattern Recognition Letters*, 122:86–91, 2019.

[374] Luigi Warren and Cory Lin. mrna-based genetic reprogramming. *Molecular Therapy*, 27(4):729–734, 2019.

[375] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[376] Gilbert Weidinger, Chris J Thorpe, Katrin Wuennenberg-Stapleton, John Ngai, and Randall T Moon. The sp1-related transcription factors sp5 and sp5-like act downstream of wnt/β-catenin signaling in mesoderm and neuroectoderm patterning. *Current biology*, 15(6):489–500, 2005.

[377] Ada Weinstock, Emily J Brown, Michela L Garabedian, Stephanie Pena, Monika Sharma, Juan Lafaille, Kathryn J Moore, and Edward A Fisher. Single-cell rna sequencing of visceral adipose tissue leukocytes reveals that caloric restriction following obesity promotes the accumulation of a distinct macrophage population with features of phagocytic cells. *Immunometabolism*, 1, 2019.

[378] Harold Weintraub, Stephen J Tapscott, Robert L Davis, Mathew J Thayer, Mohammed A Adam, Andrew B Lassar, and A Dusty Miller. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of myod. *Proceedings of the National Academy of Sciences*, 86(14):5434–5438, 1989.

[379] Kerstin S Wendt, Keisuke Yoshida, Takehiko Itoh, Masashige Bando, Birgit Koch, Erika Schirghuber, Shuichi Tsutsumi, Genta Nagae, Ko Ishihara, Tsuyoshi Mishiro, et al. Cohesin mediates transcriptional insulation by ccctc-binding factor. *Nature*, 451(7180):796–801, 2008.

[380] Alex H Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I Ryu, Krishna V Shenoy, Mark Schnitzer, Tamara G Kolda, and Surya Ganguli. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98(6):1099–1115, 2018.

[381] Peter Wills and François G Meyer. Metrics for graph comparison: a practitioner's guide. *PloS One*, 15(2):e0228728, 2020.

[382] Daniel A Winer, Shawn Winer, Lei Shen, Persis P Wadia, Jason Yantha, Geoffrey Paltser, Hubert Tsui, Ping Wu, Matthew G Davidson, Michael N Alonso, et al. B cells promote insulin resistance through modulation of t cells and production of pathogenic igg antibodies. *Nature medicine*, 17(5):610–617, 2011.

[383] Michael M Wolf, Alicia M Klinvex, and Daniel M Dunlavy. Advantages to modeling relational data using hypergraphs versus graphs. In *2016 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2016.

[384] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):1–12, 2014.

[385] Shin-Rong Wu and Pavan Reddy. Tissue tolerance: a distinct concept to control acute gvhd severity. *Blood, The Journal of the American Society of Hematology*, 129(13):1747–1752, 2017.

[386] Gordana Wutz, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, Gregor Jessberger, Matthias Muhar, M Julius Hossain, et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by ctcf, wapl, and pds5 proteins. *The EMBO journal*, 36(24):3573–3599, 2017.

[387] Bingqing Xie, Da Sun, Yuanyuan Du, Jun Jia, Shicheng Sun, Jun Xu, Yifang Liu, Chengang Xiang, Sitong Chen, Huangfan Xie, et al. A two-step lineage reprogramming strategy to generate functionally competent human hepatocytes from fibroblasts. *Cell Research*, 29(9):696–710, 2019.

[388] Jun Xu, Yuanyuan Du, and Hongkui Deng. Direct lineage reprogramming: strategies, mechanisms, and applications. *Cell Stem Cell*, 16(2):119–134, 2015.

[389] Jia Xue, Susanne V Schmidt, Jil Sander, Astrid Draffehn, Wolfgang Krebs, Inga Quester, Dominic De Nardo, Trupti D Gohel, Martina Emde, Lisa Schmidleithner, et al. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity*, 40(2):274–288, 2014.

[390] Yosuke Yamada, Hironori Haga, and Yasuhiro Yamada. Concise review: dedifferentiation meets cancer development: proof of concept for epigenetic cancer. *Stem Cells Translational Medicine*, 3(10):1182–1187, 2014.

[391] Jing Yang, Amanda McGovern, Paul Martin, Kate Duffus, Xiangyu Ge, Peyman Zarrineh, Andrew P Morris, Antony Adamson, Peter Fraser, Magnus Rattray, et al. Analysis of chromatin organization and gene expression in t cells identifies functional genes for rheumatoid arthritis. *Nature communications*, 11(1):1–13, 2020.

[392] Jiyeon Yang, Lixiao Zhang, Caijia Yu, Xiao-Feng Yang, and Hong Wang. Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. *Biomarker research*, 2(1):1–9, 2014.

[393] Yaming Yang, Ruiguo Chen, Xianming Wu, Yannan Zhao, Yongheng Fan, Zhifeng Xiao, Jin Han, Le Sun, Xiaoqun Wang, and Jianwu Dai. Rapid and efficient conversion of human fibroblasts into functional neurons by small molecules. *Stem cell reports*, 13(5):862–876, 2019.

[394] Hiroki Yokota, Ger Van Den Engh, John E Hearst, Rainer K Sachs, and Barbara J Trask. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human g0/g1 interphase nucleus. *The Journal of cell biology*, 130(6):1239–1249, 1995.

[395] Kobe C Yuen and Jennifer L Gerton. Taking cohesin and condensin in context. *PLoS genetics*, 14(1):e1007118, 2018.

[396] Brian F Zamarron, Taleen A Mergian, Kae Won Cho, Gabriel Martinez-Santibanez, Danny Luan, Kanakadurga Singer, Jennifer L DelProposto, Lynn M Geletka, Lindsey A Muir, and Carey N Lumeng. Macrophage proliferation sustains adipose tissue inflammation in formerly obese mice. *Diabetes*, 66(2):392–406, 2017.

[397] Marco Zarbin, Ilene Sugino, and Ellen Townes-Anderson. Concise review: update on retinal pigment epithelium transplantation for age-related macular degeneration. *Stem cells translational medicine*, 8(5):466–477, 2019.

[398] Robert Zeiser and Bruce R Blazar. Acute graft-versus-host disease—biologic process, prevention, and therapy. *New England Journal of Medicine*, 377(22):2167–2179, 2017.

[399] Donghui Zhang and Wei Jiang. From one-cell to tissue: Reprogramming, cell differentiation and tissue engineering. *BioScience*, 65(5):468–475, 2015.

[400] Jingyu Zhang, Hengyu Chen, Ruoyan Li, David A Taft, Guang Yao, Fan Bai, and Jianhua Xing. Spatial clustering and common regulatory elements correlate with coordinated gene expression. *PLoS computational biology*, 15(3):e1006786, 2019.

[401] Ruochi Zhang and Jian Ma. Matcha: Probing multi-way chromatin interaction with hypergraph representation learning. *Cell systems*, 10(5):397–407, 2020.

[402] Xi Zhang, Filemon Dela Cruz, Melissa Terry, Fabrizio Remotti, and Igor Matushansky. Terminal differentiation and loss of tumorigenicity of human cancers via pluripotency-based reprogramming. *Oncogene*, 32(18):2249–2260, 2013.

[403] Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728, 2019.

[404] Yu Zhang, Rachel Patton McCord, Yu-Jui Ho, Bryan R Lajoie, Dominic G Hildebrand, Aline C Simon, Michael S Becker, Frederick W Alt, and Job Dekker. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148(5):908–921, 2012.

[405] Meizhen Zheng, Simon Zhongyuan Tian, Daniel Capurso, Minji Kim, Rahul Maurya, Byoungkoo Lee, Emaly Piecuch, Liang Gong, Jacqueline Jufen Zhu, Zhihui Li, et al. Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745):558–562, 2019.

[406] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in Neural Information Processing Systems*, 19:1601–1608, 2006.

[407] Linna Zhou, Anne C Wolfes, Yichen Li, Danny CW Chan, Ho Ko, Francis G Szele, and Hagan Bayley. Lipid-bilayer-supported 3d printing of human cerebral cortex cells reveals developmental interactions. *Advanced Materials*, 32(31):2002183, 2020.

[408] Iris Zhu, Wei Song, Ivan Ovcharenko, and David Landsman. A model of active transcription hubs that unifies the roles of active promoters and enhancers. *Nucleic acids research*, 49(8):4493–4505, 2021.

[409] Saiyong Zhu, Holger A Russ, Xiaojing Wang, Mingliang Zhang, Tianhua Ma, Tao Xu, Shibing Tang, Matthias Hebrok, and Sheng Ding. Human pancreatic beta-like cells converted from fibroblasts. *Nature communications*, 7(1):1–13, 2016.