MULTISCALE MODELING OF CELL FATE SWITCHING TO PREDICT PATIENT-SPECIFIC RESPONSE TO COMBINATION CANCER THERAPY

Lindsey R. Fernández

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

_____

Ravi Radhakrishnan, Ph.D., Professor of Bioengineering and Chemical and Biomolecular Engineering

Graduate Group Chairperson

_____

Yale E. Cohen, Professor of Otorhinolaryngology

Dissertation Committee

David F. Meaney, Professor of Bioengineering

James C. Gee, Professor of Radiologic Science

Lukasz Bugaj, Assistant Professor of Bioengineering

MULTISCALE MODELING OF CELL FATE SWITCHING TO PREDICT PATIENT-SPECIFIC RESPONSE TO COMBINATION CANCER THERAPY

To my mom, Susie

ACKNOWLEDGMENTS

I would first like to thank my adviser Dr. Ravi Radhakrishnan for his unwavering support and guidance throughout my PhD. Ravi is an incredible mentor and teacher, and it has been amazing to see how both current students and those that graduated years ago are so supported by him. While the past year brought unprecedented uncertainty and challenges, Ravi's kind and patient mentorship has been a constant and I am extremely grateful for it. My work with him has allowed me to enter the field of cancer systems modeling and through virtual conferences, workshops, and meetings, I have met wonderful people in this research community and feel so lucky to be a part of it. Additionally, I would like to thank my thesis committee Dr. David Meaney, Dr. James Gee, and Dr. Lukasz Bugaj as well as Dr. Yale Cohen and Dr. Arjun Raj for their invaluable support without which none of this would have been possible.

The work presented in this thesis was also made possible by encouragement, mentorship, and contributions of many lab mates. Through the support of Heidi Norton, Jon Beagan, Thomas Gilgenast, Mayuri Rege, and Ji Hun Kim, I made my first foray into the world of 3D genome folding. I would especially like to thank Thomas for his many enthusiastic explanations and from whom I learned many computational skills. As a junior grad student, I looked up to Heidi, who's wonderful presentations inspired me to explore this research area, and Mayuri, from whom I learned much about cell biology and who held the lab together with her calls for decorum. Heidi along with Jon and Linda Zhou were my greatest supporters throughout my time in the PhD and I was grateful to be able to navigate many late and challenging nights in lab with such wonderful people. I am also grateful to have worked alongside Zoltan Simandi, Harshini Chandrashekar, Ali Nikish, Wanfeng Gong, James Sun, Jackie Valeri, Harvey Huang, Shawn Srolovitz, and others who through

their camaraderie made the lab a wonderful place to be. Finally, I would like to thank all the members of the Radhakrishnan lab who have been so welcoming and supportive as I've made my way into multiscale and biophysical modeling. I would especially like to thank Reshma Kalyan Sundaram, Mengdi Tao, and Gabriela Witek for their help and guidance through many wonderful conversations during this time.

Additionally, I could not have made it through the PhD without the support of wonderful friends outside of lab. I am incredibly grateful to Lori Atlan who I looked up to starting from my days as a high school research intern at the Applied Physics Laboratory and who inspired me to pursue a PhD at Penn. Through my time at the medical school, I was lucky enough to meet Vivien Wong, Chukwuma Onyebeke, Alex Azan, Joseph Aicher, Bernadette Bucher, Michael Furdyna, and Rachel Mittelstaedt who helped me stay sane through many long days studying for preclinical and onwards through the research years. I also had the privilege of working alongside and befriending fellow classmates in the med school and engineering school Sonya Davey, Kent Grosh, Thulani Tsabedze, Vivek Nimgoankar, and Cody Cotner as we created SelfCerve, a cervical cancer screening device for use in low- and middle-income countries. I've also enjoyed being a part of the HHMI Interfaces Program and am grateful for the mentorship and support of other students in the program.

Finally, I would like to thank all of the friends and family who have encouraged me throughout my time in the PhD. I would especially like to thank my mom who decided to become an electrical engineer after one of her teachers told her that she would never be smart enough to do it. She has always supported my interests in engineering and science, and reminds me to take time to enjoy nature, prioritize health, and that no matter how dark the clouds today, the sun will always shine again.

ABSTRACT

MULTISCALE MODELING OF CELL FATE SWITCHING TO PREDICT PATIENT-SPECIFIC RESPONSE TO COMBINATION CANCER THERAPY

Lindsey R. Fernández

Ravi Radhakrishnan

All cells in the human body share the same DNA sequence, but differ in their functional identity, guided by a wide array of regulatory mechanisms controlling cellular lineage commitment and encoded in the unique epigenome of each cell type. Recent experimental studies with induced pluripotent stem cells have allowed researchers to investigate the dynamic nature of cell identity and relationships between gene regulation and differentiation. These studies have major implications for our understanding of not only human development, but also disease as cancers, and some neurological diseases, arise in part due to inappropriate persistence of cells in immature differentiation states. These studies have proliferated massive multi-omics databases as next generation sequencing (NGS) technologies are applied to extensively profile stem cells, *in vitro* differentiated cell populations, and cancer patient cohorts. As these data accumulate, important unanswered questions remain, including to what extent do physical states of genes change in development and disease, and how do these changes meaningful alter cell signaling pathways and clinically impact individual patients. To address these questions, new computational tools are needed to 1) rigorously assess epigenomic state changes captured with NGS

modalities in the course of lineage specification, and to 2) integrate models of patient (epi-) genomic states with those of disease-associated cell signaling pathways and clinical outcomes. This thesis describes the development and applications of computational methods to help address these needs. First described is a statistical tool for classifying long-range looping interactions that change across developmental models and disease states from data captured with NGS technologies. Its application is demonstrated for study of chromatin looping state changes in the course of neural lineage commitment and neuronal activation. Then a multiscale framework is described that integrates patient (epi-)genomic profiles with mechanistic models of signaling pathways critical to decision making to enter tumorigenic states. The framework is demonstrated in a clinical setting to predict patient-specific responses to different specific treatment combinations in nephroblastoma. Such methods have great potential to advance our understanding of the determinants of cellular identity and its loss in cancer, and in turn our ability to personalize patient care.

Contents

List of Tables

List of Illustrations

Part I: Introduction and Preliminary Concepts

1 Introduction

All cells in the human body share the same DNA sequence, but differ in their functional identity, guided by a wide array of regulatory mechanisms controlling cellular lineage commitment and encoded in the unique epigenome of each cell type. How a stem cell commits to one particular lineage or fate over another is determined by many interlinked regulatory layers and these must be maintained to achieve healthy cell, tissue, and organ function throughout life. [1] Cellular programming constitutes one such regulatory level and shapes cell fate decision making through transcriptional regulation, carried out by transcription factors (TFs) and chromatin regulators, and post-transcriptional regulation, carried out by microRNAs (miRNAs). [2] At a broader scale, cell signaling pathways enable cells to respond to stress, infection, and other external cues, and are necessarily interlinked with gene regulatory machinery to determine cellular identity. [2] Understanding the mechanistic relationships that establish and maintain a cell's functional identity is a pursuit at the heart of biology research. Unraveling these mysteries has major implications for our understanding of not only human development, but also disease. For instance, many cancers, and some neurological diseases, arise in part due to inappropriate persistence of cells in immature differentiation states or loss of the ability to keep cells in a previously established differentiation state. [3] [4] [5] [6] Determining the regulatory relationships and external cues that underlie specialized cellular functions and their loss in disease could provide a road map for creating therapies that optimally control cell fate decision making to reestablish healthy tissue function. [7]

Recent technological advances have fueled enormous progress in decoding cell fate decision making. Development of stem cell models and induced pluripotent stem (iPS) cell models have allowed researchers to investigate the dynamic nature of cell identity and relationships between gene regulation and differentiation. [8] [9] With these cellular models, researchers can follow changes in a differentiating cell's epigenomic state and signaling dynamics over time, identify key regulators, and investigate their role in determining cell fate. Tools have proliferated to allow researchers conduct these studies in different lineages, to perturb specific regulatory elements with genome editing or silencing, and to compare observations across developmental stages, lineages, perturbations, and disease states. These studies have proliferated massive multi-omics databases as next generation sequencing (NGS) technologies, such as whole genome sequencing, large scale RNA sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), and chromatin-conformation-capture sequencing (3C-seq), are applied to extensively profile stem cells, in vitro differentiated cell populations, and cancer patient cohorts. [10] [11] [12] [13] [14]

As these empirical data accumulate, it becomes increasingly clear that most diseases, including cancer, involve a large and diverse set of elements that interact via complex networks. [15] These networks complicate the work of designing therapies as cells often find alternative molecular routes when the action of individual target genes or molecules are perturbated. [14] Key to progressing in the era of large-scale biology research is the development of mathematical and computational tools for identifying bona fide biological state changes or molecular mechanisms often buried in the noise of large scale and/or genome-wide data sets. [16] [17] However, as new NGS techniques are developed so to must the tools used to process and interpret the resulting data evolve, and the development of

these tools has long represented a bottleneck in biomedical research. [14] Even these data are gathered, and processing techniques developed, it remains challenging to answer many fundamental questions in biomedical research, including to what extent do physical states of genes change in development and disease, and how do these changes meaningful alter cell signaling pathways and clinically impact patients?

In addition to these questions, there remains enormous need to determine which empirical observations are relevant for designing treatment plans for individual patients and to develop techniques that ensure treatments are optimally effective in individual patients. [18] [19] This stands out as particularly important for the development of chemoradiotherapy regimens; clinicians often struggle to balance treatment intensity against toxicity and often lack information to know if a particular drug or drug combination will be effective for an individual patient. [20] [21] Treatment design is also made more challenging by genomic instability common across many cancers, which contributes to high inter-tumor (tumor-by-tumor) heterogeneity and intra-tumor (within) heterogeneity in genotype and phenotype. [22] [23] These heterogeneities alter cell signaling and cell fate decisions, resulting in variable drug efficacy over time and across tumor-cell subpopulations, and ultimately development of therapeutic resistance. [24]

Answering these questions and needs requires the development of techniques to deal with large amounts of data and relationships between those datasets. [14] Mathematical modeling has emerged as a powerful tool to identify clinically relevant information from empirical data and predict patient-specific treatment outcomes. [25] As described in [25], examples of such clinical modeling approaches include statistical data-driven models which analyze clinical data from patients and predict probabilities of different recurrence scenarios,

pharmacokinetic-pharmacodynamic (PK-PD) models which determine patient-specific drug regimens, and mechanistic models of cellular processes and tumor progression derived from empirical data. The development of these modeling approaches represents a critical milestone in the path to the era of precision medicine, however challenges remain. To date, these modeling efforts have been disparate and most of them restricted to representation of phenomena of a specific length or time scale. [25] [26] Such approaches are inadequate for modeling of cell fate decision making, as well as diseases resulting from its dysregulation, which are guided by a complex hierarchy of mechanisms that span multiple scales in time and space (i.e. multiscale), and multiple interconnected physical, chemical, and biological processes (i.e. multiphysics). [27] [28] [29] [30] To understand these mechanistic relationships, I believe multiscale and multiphysics modeling techniques are needed to enable study these processes as *in silico* biology, built through integration of experimental observations with physical principles. [26] [31] These techniques will allow researchers to study otherwise technically infeasible parameters and to harness the power of ever-growing large scale -omics data sets. Such techniques have the potential to guide creation of the next generation of personalized cancer therapies, and to optimize and broaden access to more affordable cytotoxic treatment options.

My thesis work is to create multiscale mechanistic models of pathways critical to cell fate decision making between differentiative (healthy) cellular states and proliferative (cancerous) cellular states. Such multiscale models allow integration of mechanistic models of processes spanning vastly different time scales (e.g., coupling models of shorter time scale cell signaling and gene regulatory networks to models of far longer time scale processes like cell proliferation and apoptosis) by representing activities of model components as system

variables related by governing equations implemented as algorithms. The systems represented by these integrated models would be far too complex to solve analytically and intractable to fully investigate experimentally. These multiscale models thus have incredible value in allowing systematic investigation of complex networks of factors and signals contributing to patient outcomes and treatment responses and isolation of critical factors for experimental follow-up.

The overarching goal of my thesis was to create such a multiscale model and demonstrate its clinical value in designing personalized treatment plans for individual patients. In pursuit of this goal, I created computational tools that allow 1) rigorous assessment of epigenomic state changes captured with NGS modalities in the course of lineage specification, and 2) integration of models of patient (epi-)genomic states with those of disease-associated cell signaling pathways and clinical outcomes. The background, implementation, results, and future directions of this work are described across the following chapters, summarized below:

In **chapter two**, I discuss challenges in characterizing cell fate decisions and how this could benefit from a systems modeling approach. Transcriptional and post-transcriptional regulators are discussed. Cell signaling pathways critical to cell fate determination are detailed, specifically EGFR mediated Ras-MAPK pathway and p53 mediated cell cycle and damage response pathways, which are frequently altered in many cancers and for which single cell level mechanistic models have been developed.

In **chapter three**, I discuss the challenges of implementing such systems modeling approaches. Careful consideration is required to create a valuable representation of a complex system, including determination of the basic constituents of the system and

reasonable mathematical representation of their activity and interactions. These efforts are complicated in multiscale, multiphysics modeling as single scale, single physics models must be combined in a way that is representative of how signals are propagated across scales, however little empirical data may be available for defining such propagation and models spanning very different scales and physics are numerically challenging to solve, presenting computational challenges as well. These challenges and efforts to address them are discussed.

In chapters four through six, I describe the mathematical implementation single cell level models necessary to populate a multiscale model of cell fate decision making. In **chapter four**, I describe approaches for cellular level modeling, specifically detail efforts to represent previously mentioned EGFR and TP53 mediated cell signaling pathways via continuous-time ODE modeling and discrete logic-based systems modeling respectively. I also describe current modeling standards for combination chemotherapy. In **chapters five** and **six**, I detail a computational method for capturing epigenetic state changes across developmental models and disease states from data captured with NGS technologies, specifically 3C-seq data and its evaluation against chromatin immunoprecipitation (ChIP)-seq and RNA-seq data. Its application is demonstrated for study of chromatin looping state changes in the course of neural lineage commitment in chapter five and of neuronal activation in chapter six.

In **chapter seven**, I discuss how previously mentioned single scale models can be integrated in a multiscale modeling approach, detailing implementation of a model that modulates the activity of target proteins in a hybrid multiscale of signaling pathways critical to decision making to enter tumorigenic states according patient-specific (epi-)genetic profiles. I demonstrate this model for the prediction of patient-specific responses to

different chemoradiotherapy combinations for the treatment of nephroblastoma, or Wilms Tumor, a common pediatric tumor of the kidney.

Finally, in **chapter eight**, I discuss future work to expand this effort to account for tumor heterogeneity and microenvironmental cues with the implementation of agent-based model that represents the constituent cells of different lineages within a tumor with different cell level multiscale models. Motivations and challenges in implementing this agent-based approach, and I propose machine learning techniques to allow sufficient scale up of single cell level models to permit their simulation across multiple cells.

2 A Systems Biology Approach to Cell Fate Determination

Cell fate decision making refers to the ability of a cell respond to signals in its environment and process them to differentiate, proliferate, grow, or die as needed to retain the healthy function of the tissue and organ. Precise understanding of the processes that guide normal cellular decision making and how they are disrupted in disease is in turn critical to our ability to treat many diseases, to essentially modulate cell signaling dynamics, transcriptional controls, and environmental status as needed to restore normal cellular functional identity and overall organ function.

That many diseases fit this conceptual framework has long been understood. Cancer is in part a disease of mismanaged cell fate: over proliferation and avoidance of apoptosis, driven by disrupted DNA repair and abnormal cell signaling. Additionally, incomplete differentiation is thought to contribute to the pathogenesis of many cancers as mutations in developmentally important genes disrupt the balance between self-renewal and differentiation [32] [33]. Increasingly, neurological disease is being thought of from this perspective as well with low neural stem cell and progenitor cell populations being linked to Parkinson's Disease and certain epilepsies, and over-proliferation of those cell populations being associated with glioma and Huntington's Disease [34].

Based on this understanding, researchers have and continue to develop therapies to address different aspects of aberrant cell fate decision making. In the case of cancer, many successful cytotoxic drugs address over proliferation, including antimitogenics like Vincristine and DNA synthesis blocking drugs like Doxorubicin (see **Chapter 7** for modeling of combination application of these drugs). More recently, with the emergence of

genome wide sequencing, drug development has shifted to targeted therapies that aim to act upon biological changes specific to tumor cells while sparing normally functioning cells. Small molecules to induce cellular reprogramming to desired cell fates are also being explored, a treatment option that would potentially allow regeneration of tissues of all types, including for neuro-regeneration and cancer treatment [35].

Incredible successes have already been realized through therapies designed based on understanding of cell fate control, but significant limitations persist. Drug resistance remains a critical problem in cancer treatment, even for targeted therapies [36]. Attacking individual actors within complex signaling systems can be a futile effort as cells may rely on alternative pathways to retain aberrant behavior, especially in the context of cancer where signaling systems are constantly evolving as new subclonal populations arise. More broadly, the vast majority of candidate drugs subject to clinical trial are found to have little to no therapeutic benefit, highlighting the generally difficulty of anticipating the effects of drugs once administered to people [37]. Even for effective drugs, many sources of variability contribute to differences in drug response from person-to-person, site-to-site, and cell-to-cell.

To address these challenges, there has been great interest in studying the processes governing cell fate from a systems biology perspective, to understand the network of processes driving cell fate decision making in context rather than as individual elements to be pulled apart and characterized in isolation. Excitingly, with an ever-growing array of new sequencing and experimental techniques (e.g., single cell studies, gene silencing and editing experiments, light-activated protein studies, etc.), more data than ever are available to facilitate systems level study of cell fate regulation.

In the next few sections, I discuss the state of our systems level understanding of signaling pathways, signaling dynamics, transcriptional and post transcriptional controls, and the influence of cellular microenvironment as they contribute to cell fate decision making and therapy development. In the next chapter, I discuss multiscale modeling as a solution for synthesizing these new systems level insights to better characterize decision making and disease.

2.1 Cell signaling pathways

Cancer has long been understood to be a quintessential systems biology disease. Development of cancer is typically driven by multiple mutations leading to pathologic behavior of a complex network of interacting molecular processes and the loss of a cell's functional identity. The success of many cytotoxic drugs has hinged on influencing these networks to promote normal differentiative states over aberrant proliferative ones. In that pursuit, many signaling pathways critical to deciding between these states have been extensively studied and mapped, including pathways contributing to apoptosis, survival, the cell cycle, DNA repair, lineage commitment, and differentiation. Below, I describe pathways essential to decision making between differentiative (normal) and proliferative (cancerous) states that I ultimately represent in a multiscale modeling framework to aid in personalization and optimization of combination cancer therapy.

2.1.1 TP53 mediated cell cycle and damage response pathways

Uncontrolled proliferation driven by dysregulation of the cell cycle is one of the main traits of cancer. Crucially, as shown in ***Figure 2.1***, TP53 controls the phase transition from G1 to S and cells with mutations in TP53 may not proceed through that checkpoint or initiate

apoptosis. This has critical implications for cancer treatment as cytotoxic drugs depend the cell cycle and have greatest effect on actively proliferating cells [36]. For instance, as shown in ***Figure 2.1***, drugs that act on DNA synthesis damage cells during periods of DNA synthesis (S phase) while mitotic inhibitors produce cell kills through exposure during mitosis (M phase) [36]. These drugs may fail to produce cell kills in the presence of a *TP53* mutation and other mutations that keep the cell cycle from proceeding to these phases [36]. Additionally, therapies that act to reduce the integrity of DNA including many cytotoxic drugs and radiation may not produce cell kills in the absence of TP53, which controls the signaling machinery for detecting loss of DNA integrity and initiating apoptosis, resulting in therapeutic resistance [36]. Thus, in designing a model for simulation of combination chemoradiotherapy, it was critical to include representation of the TP53 mediated DNA damage response and cycle cell pathways.

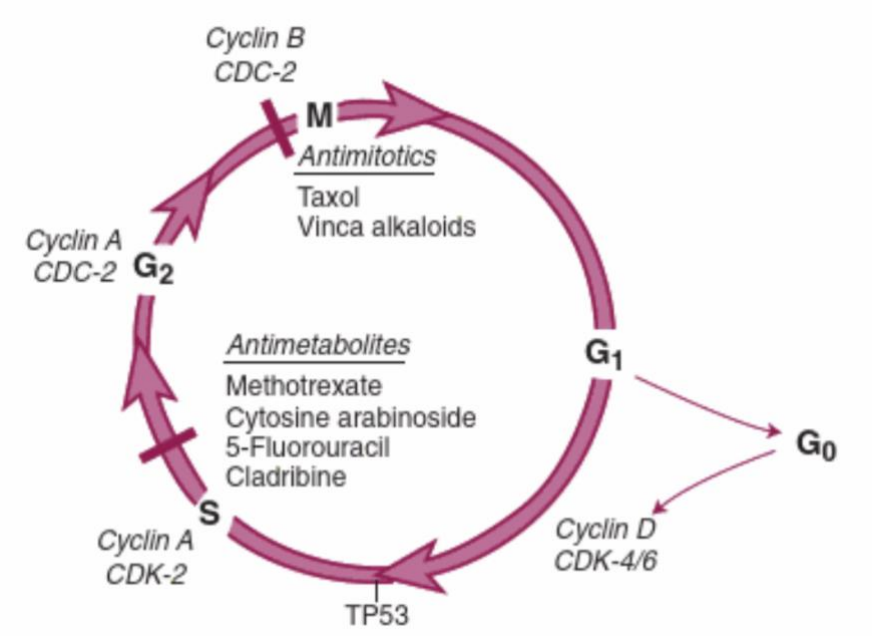


Figure 2.1. From Izar et al., the cell cycle, its controls and checkpoints as well as sites of action of cell-cycle specific cytotoxic drugs.

The cell cycle phases consist of non-dividing (G0), resting (G1), DNA synthesis (S), a gap between synthesis and mitosis (G2), and mitosis (M). Cyclin proteins activate cyclin-dependent kinases (CDKs) to control transition between phases (note, CDC-2 is also known as CDK-1). Additionally, TP53 monitors DNA integrity and controls passage through the G1/S transition. *[36]*

Though TP53 function has been under intense investigation for decades, the size and complexity of the regulatory network it acts upon make full experimental characterization of its kinetics technically infeasible. Integration of experimental results has led to the construction of simplified TP53 network models like that shown in **Figure *2.2*** and state-space analysis of such networks has provided insight into how TP53 dynamics are controlled by specific feedback loops and how perturbations of processes in these loops modulate those dynamics to alter cell fate [38]. As summarized in **Figure *2.2***, these feedback loops work to alter TP53 dynamics in response to DNA damage and cellular stress with specific dynamics mapping to specific cell fates of DNA repair, cell cycle arrest, senescence, or cell death [38].

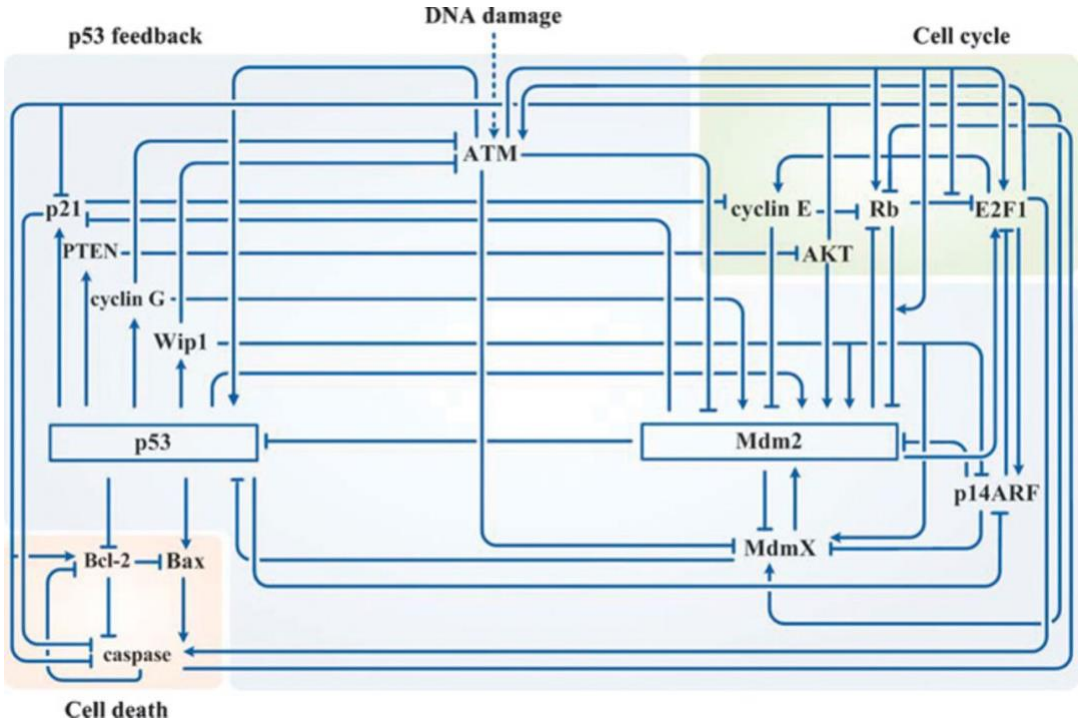


Figure 2.2. From Choi et al., a simplified schematic of the p53 regulatory network. (Note, TP53 is the human isoform of p53, but p53 and TP53 are typically used interchangeably to refer to that isoform). Based on combined literature data, Choi et al. constructed a p53 network containing 16 nodes as well as 160 negative and 218 positive feedback loops, the vast majority of which interact with p53. Arrows indicate activating processes and bars indicate inhibitory ones. In response to DNA damage, ATM activates, in turn activating p53 to turn on various feedback loops, with key loops shown in this schematic. The lower left orange area represents the cell death module, the upper right green area presents the cell cycle module, and the remaining blue area represents the p53 feedback module. *[38]*

2.1.2 EGFR mediated Ras-MAPK and PI3K/AKT pathways

Epidermal growth factor receptor (EGFR), also known as ErbB1, is part of the greater ErbB cell signaling network which is a major contributor to tumorigenesis and is under intense investigation for therapeutic targets. As shown in **Figure *2.3***, the network is comprised of many extracellular ligands and trans-membrane receptors like EGFR as well as many

enzymes, scaffolds, adaptors, and small molecules [39]. Signaling initiates when a ligand binds a receptor and causes the receptor to dimerize. This in turn activates the receptor's tyrosine kinase domain, which results in autophosphorylation of tyrosine residues [39]. In response, multiple proteins are recruited to the plasma membrane by binding phosphotyrosines and so a complex network of interactions between the activated receptors, recruited proteins, and plasma membrane molecules eventually culminates in the activation of multiple downstream effectors, including extracellular-signal-regulated kinase MAPK (originally known as ERK as it is listed in **Figure *2.3***) and protein kinase B/AKT, which are both implicated in the control of proliferation and survival [39].

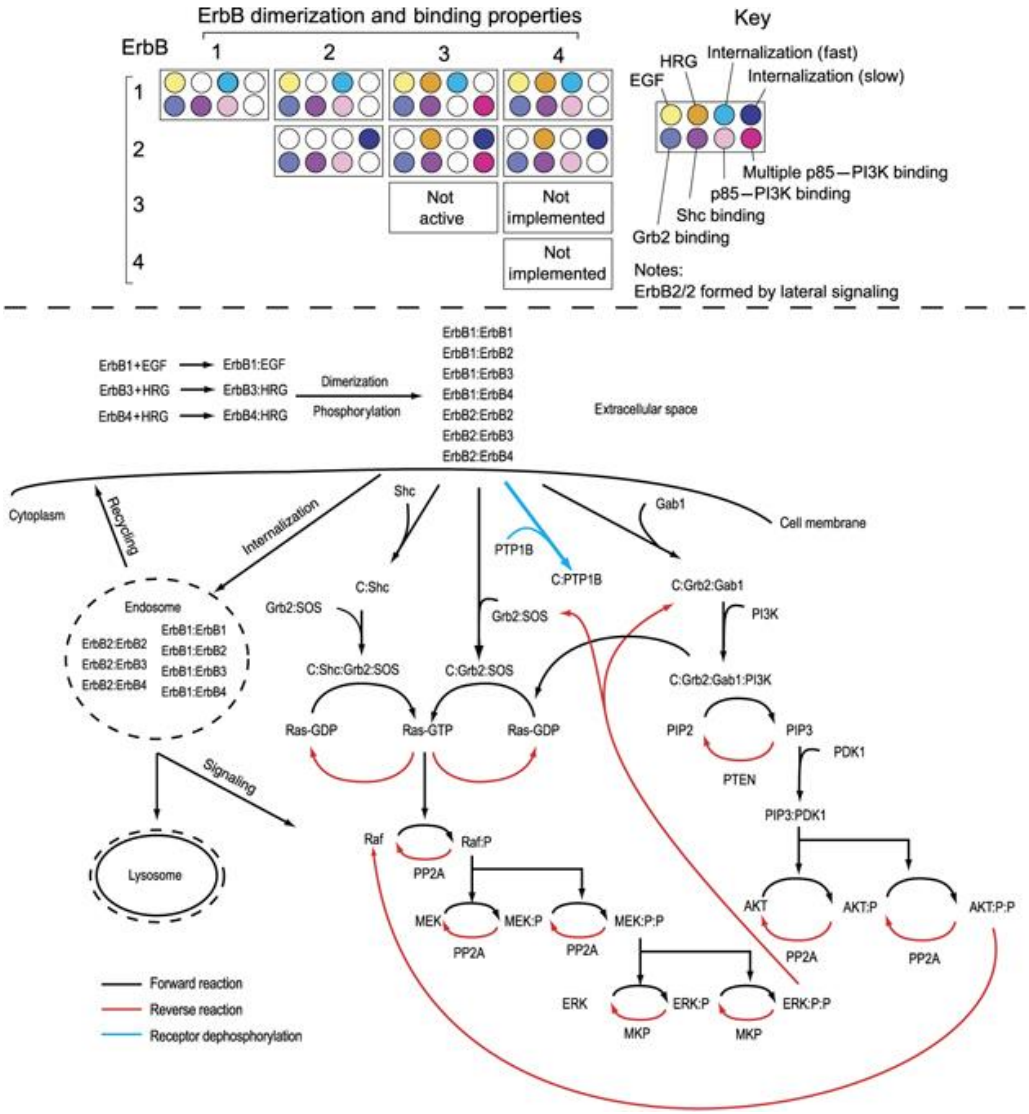


Figure 2.3. From Chen et al., a simplified schematic of ErbB regulatory pathways showing receptor activation, interaction, internalization, and recycling in the Ras-MAPK/ERK and PI3K/AKT cascades.

Forward reactions are shown with black arrows while negative feedback interactions are shown with red arrows. ErbB receptors dimerize upon activation with various ligands and the matrix in the top half of the figure summarizes the functional properties of observed dimers. Nodes with the prefix C indicate that one or more ErbB receptors form a complex with the species. *[40]*

Spatiotemporal dynamics of signaling are known to be critical to the ErbB network's control of cell fate, as different inputs stimulate different network kinetics and in turn lead to different cell fates [39]. The network as well as its downstream signaling cascades, including the Ras-MAPK pathway and the PI3K/AKT pathway, have been studied thoroughly at the molecular level resulting in clarification of its activation kinetics and well-defined systems models of their behavior based on those findings (see **Chapter 4**) [39] [40]. Additionally, many cancer therapies that interact with the ErbB network are known to dramatically vary in efficacy from patient-to-patient [40], further motivating the inclusion of these networks in patient specific modeling of combination therapy.

The Ras-MAPK pathway (also known as the MAPK/ERK or Ras-Raf-MEK-ERK pathway) acts in many ways to regulate cell cycle entry and control cellular proliferation and do so across many cell types. This pathway is responsible for integration of external cues from the presence of mitosis-triggering signals and growth factors into signaling cascades that support proliferation and growth, e.g., EGF binding EGFR leads to phosphorylation events in the MAPK cascade which ultimately activate the kinase activity of ERK, which must be present for cells to express genes necessary for cell cycle entry and to remove cell cycle blocks that allow cells to progress to synthesis (S phase of the cell cycle). Additionally, ERK signaling results in expression of c-Myc and other signals that control a downstream switch that prevents cells from returning to G1 after entering S phase. Importantly, this pathway is known to interact with TP53 signaling to control whether newly proliferated cells become quiescent (enter G0 phase) or immediate reenter the cell cycle. In addition to cancers, mutations in this pathway have been linked to multiple neuropsychiatric diseases.

The PI3K/AKT pathway (also referred to as the PI3K/AKT/mTOR pathway), is also critical for regulating the cell cycle, downregulating apoptosis, and the decision to proliferate rather than differentiate in stem cells. In cancer, this pathway is frequently overactive, driving proliferation and downregulating apoptosis. This pathway is a central contributor to many cancers and anti-cancer drug resistance. Oncogenic activation of the pathway can occur via many routes, including those that disrupt inhibition of the pathway through PTEN and those that cause overactivation through upstream overexpression of EGFR. The pathway is also critical to neural lineage cells, driving proliferation over quiescence in neural stem cells in response to sufficient glucose level, and so has an important role in neural development and plasticity, and is implicated in neural stem cell diseases.

Finally, these two EGFR mediated pathways are known to interact with the mentioned TP53 mediated DNA damage response and cell cycle control pathways through the linkages listed in **Figure *2.4*** below. As described in **Chapter 7**, these linkages will be used to allow propagation of molecular state information across a multiscale mechanistic model for simulation of combination chemoradiotherapy.

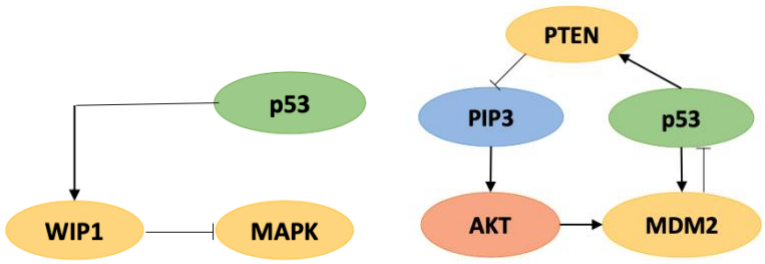


Figure 2.4. From Ghosh et al., the p53 regulatory network links to the Ras-MAPK pathway through the activity of WIP1 and to the PI3K/AKT pathway through the activity of PTEN and MDM2. *[25]*

2.2 Mechanisms of gene expression regulation

Cell signaling pathways act to control gene expression in response to internal and external cues. From a systems biology perspective, these signaling networks broadly represent one of multiple scales of biological processes that act to influence cell fate determination. At the next level below are the many layers of mechanisms through which these signals can act to modify gene expression to achieve versatile and precise expression dynamics and cell type specific function in mammalian cells. These can occur at any stage in the process of generating gene products (see ***Figure 2.5***), including epigenetic controls that enhance or limit access of transcriptional machinery to certain sections of the genome; transcriptional modifiers that determine when, how frequently, and which sections of a gene are transcribed; RNA expression controls that modify newly synthesized transcripts in the process of maturation to mRNA; translational regulators that determine how much protein is synthesized from a given mRNA, and post-translational modifications that edit existing proteins as needed to achieve precise spatiotemporal control of their activity.

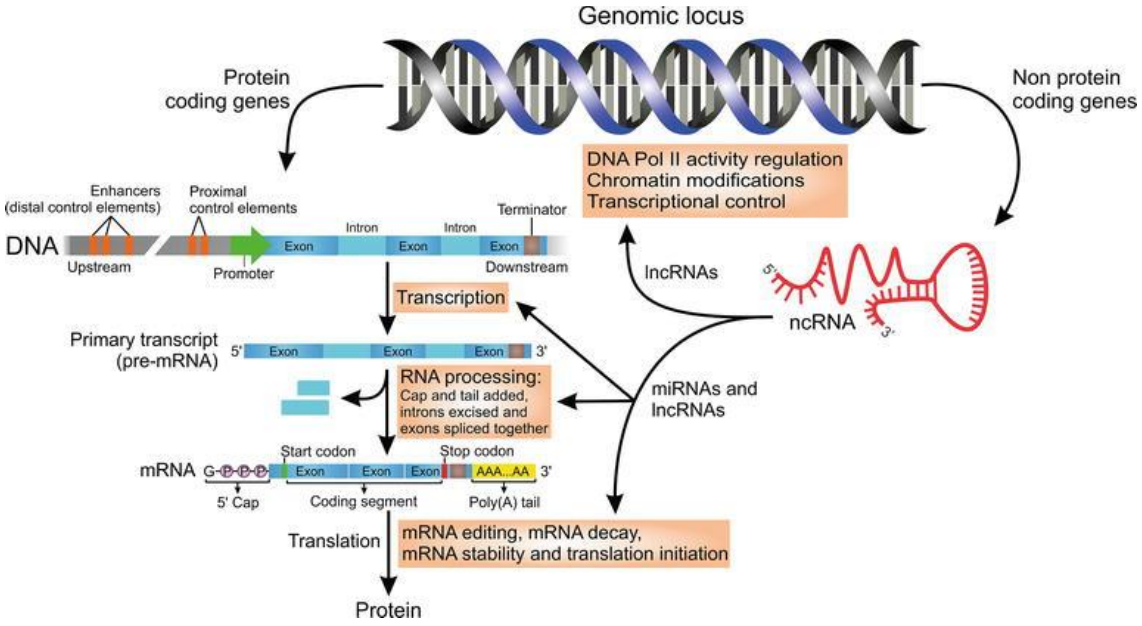


Figure 2.5. From DeOcesano-Pereira et al., a schematic representation of the roles of different RNA species in regulating mammalian gene expression.
A single genomic locus is depicted and at each step in the process of transcription and translation, multiple molecular mechanisms act to control the construction of the final gene product. Proximal control elements are located close to the promoter, while distal elements (enhancers) may be far away from a gene. Cis-acting regulatory elements, present in the pre-mRNA sequence, determine which exons are retained and which exons are spliced out, resulting in alternative transcript isoforms (alternative splicing). mRNA structure is stabilized in preparation for transport into the nucleus with the addition of 5' Cap and Poly(A) tail. Regulatory non-coding RNAs (ncRNAs) can act via multiple pathways to alter gene products in the course of transcription and translation. Long non-coding RNAs (lncRNAs) target protein complexes to specific genomic loci affecting transcription patterns (transcriptional interference), leading to chromatin modifications and DNA polymerase II activity. Advances in transcriptomics have resulted in the discovery of large numbers of ncRNAs (miRNAs and lncRNAs), many of which display the capacity to regulate gene expression at the levels of transcription (control of alternative splicing), post-transcription (mRNA editing, mRNA decay and mRNA stability) and translation (translation initiation) *[41]*.

The scale and complexity of these regulatory controls in mammalian gene expression has only been recently grasped with the advent of high-resolution genomic sequencing techniques and intense efforts remain underway to measure and map the dynamics created by these controls to higher order changes in cell signaling, cell fate, and cell function across different stages of development, in disease, and in response to experimental or therapeutic perturbations. Each regulatory layer presents a new lens through which to investigate the enormous breadth of genomic changes that occur in the course of cancer initiation and progression.

2.2.1 Transcriptional Regulation

Broadly, transcription is a process in which the enzyme RNA polymerase decodes the genetic information stored in the chromosomal DNA to produce an RNA transcript [42]. The resulting transcript may be one of five types: messenger RNA (mRNA, which become proteins; 1-2% of the total transcripts), ribosomal RNA (rRNA, required for translation; 80% of the total transcripts), transfer RNA (tRNA, required for translation), recently discovered microRNA (miRNA, post-transcriptional regulators of gene expression), and small interfering RNA (siRNA) [42]. While each type is synthesized according to a different set of regulatory mechanisms, in general, a transcriptionally active gene is controlled by a stretch of DNA typically located upstream of the transcription start site (−500 bp to −1000 bp) defined as a promoter which acts as a docking site for proteins known as transcription factors (TF) [42]. Transcription factors (TF) are the fundamental regulators of eukaryotic transcription; ubiquitous TFs each contain a specific DNA sequence binding motif through which it can recognize and act upon a specific genomic sequence [42]. Gene selective transcription factors connect to extra- or intracellular signaling pathways, which act as master regulators to switch a gene's expression on or off [42].

In addition to its promoter, a gene may be regulated by distal DNA sequences located several megabases away from its transcription start site, defined as enhancers. While distant in linear representation of the genome, enhancers are brought into spatial contact with the genes they act upon through three-dimensional folding of the genome. 3D genome structure is organized hierarchically across multiple spatial scales as shown in **Figure *2.6***. At the coarsest level, the genome is separated into chromosome territories (chromosomes

spatially separate in the nucleus) and each chromosome further separates into hubs of transcriptionally active and inactive chromatin (termed A and B compartments respectively).

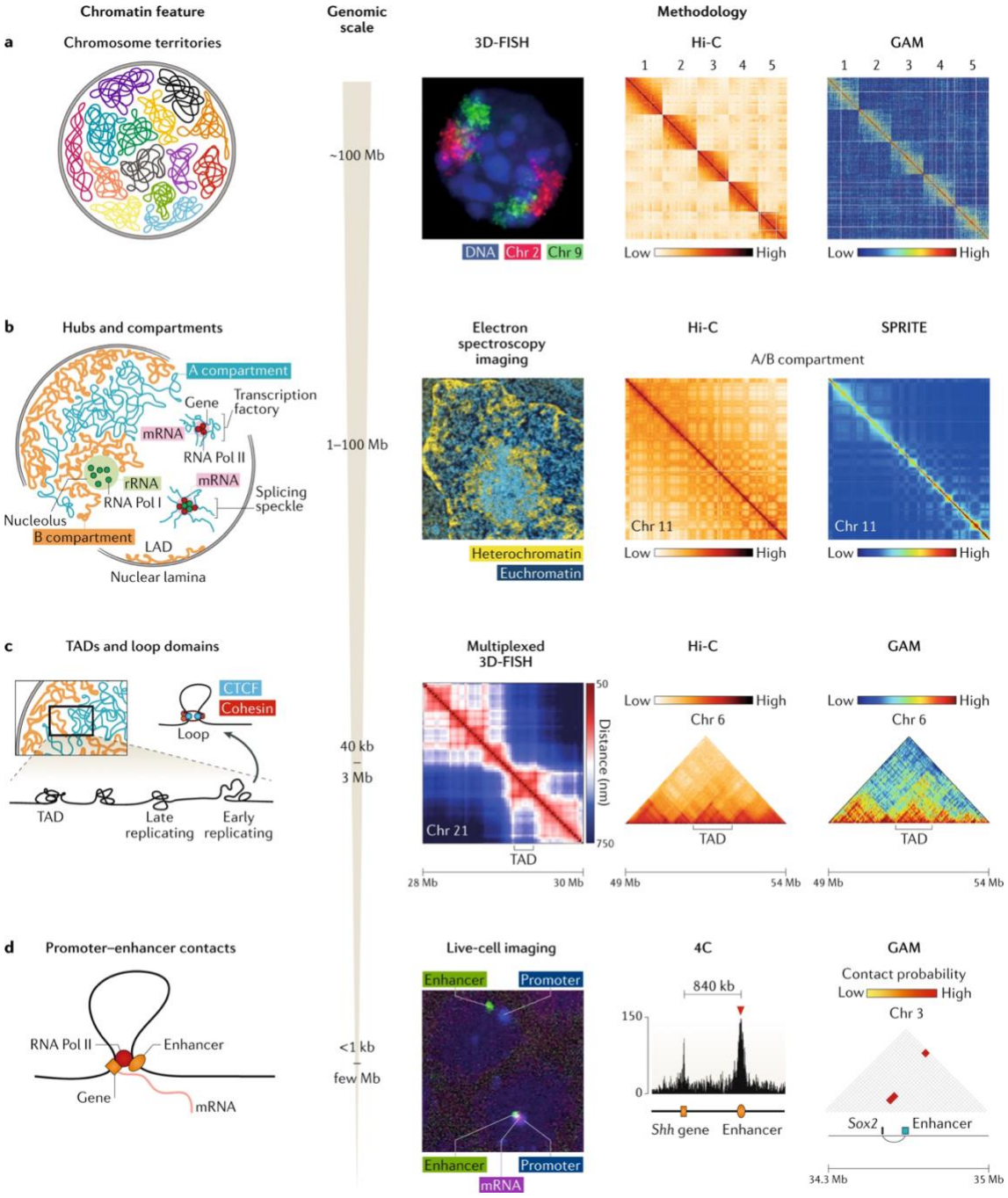


Figure 2.6. From Kempfer and Pombo, methods for studying the major features of 3D chromatin folding across different genomic scales.

a Chromosomes occupy discrete territories in the nucleus, which were first detected using imaging techniques. The 3D-fluorescence in situ hybridization (3D-FISH) image shows the positions of the chromosome territories of chromosome 2 (red) and chromosome 9 (green) within DAPI-stained nuclei (blue) from mouse embryonic stem cells (ESCs). Chromosome territories are also detected as regions of high-frequency intrachromosomal interactions on contact maps generated by chromosome conformation capture (3C)-based methods, such as Hi-C (high-throughput chromosome conformation capture), and ligation-free approaches, such as genome architecture mapping (GAM). **b** DNA inside the nucleus separates into hubs of active (A compartment) and inactive (B compartment) chromatin, clustering around the nucleolus, splicing speckles, transcription factories and other nuclear bodies not represented here. Electron spectroscopy imaging of the mouse epiblast shows the distribution of heterochromatin (yellow) around the nucleolus (light blue) and at the nuclear periphery. Decondensed euchromatin (dark blue) is positioned more centrally in the nucleus. Nucleic acid-based structures are stained yellow, protein-based structures blue. Hi-C and split-pool recognition of interactions by tag extension (SPRITE) contact maps of mouse chromosome 11 show the separation of chromatin into discrete contact hubs (A and B compartments), which are visible as checkerboard-like contact patterns. **c** At shorter genomic length scales, chromatin folds into topologically associating domains (TADs), which overlap with domains of early and late replication, and DNA loops, that arise from cohesin-mediated interactions between paired CTCF proteins. Multiplexed FISH of consecutive DNA segments in a 2-Mb region in the human genome shows the emergence of TADs in the population-average distance map. In Hi-C and GAM contact maps, TADs are represented by regions of high internal interaction frequencies and demarcated by a drop in local interactions at their boundaries. **d** Contacts between a gene and its cis-regulatory elements occur via loop formation between the enhancer bound by RNA polymerase II (Pol II) and the gene promoter. These contacts can be detected by live-cell imaging; shown are contacts between the enhancer (green) and promoter (blue) of the eve gene in a Drosophila melanogaster embryo, with simultaneous imaging of eve mRNA expression (red). The circular chromosome conformation capture (4C)-sequencing track shows the interactions between the Shh gene promoter and the ZRS (a limb-specific enhancer of

the Shh gene) in the anterior forelimb in mice. Contact maps can be processed using mathematic techniques to extract the most significant enhancer–promoter contacts from the data set, resulting in a contact matrix with only the high-probability interactions. The most significant interaction at the Sox2 locus can be found between the Sox2 gene and one of its well-studied enhancers. *[43]*

At shorter genomic length scales, chromatin folds into topologically associated domains (TADs) – genomic regions defined by their tendency to spatially aggregate and interact more with each other than with neighboring regions. [43] Through recently developed Chromosomal Conformation Capture (3C) techniques, finer scale folding could be measured, revealing that at finer genomic length scales, DNA loops arise from cohesin mediated interaction between CTCF proteins. These loops, also known as long-range looping interactions, form between enhancers and promoters, bringing genes in contact with their cis-regulatory elements.

At yet lower length scale, DNA wraps around nucleosomes to form what is termed as the chromatin fiber; tighter and looser wrapping around nucleosomes increases and decreases the accessibility of DNA by transcription machinery and a host of proteins act to alter this wrapping to control accessibility. Each nucleosome is composed of eight histone proteins and these are commonly modified by acetylation, phosphorylation, methylation, sumoylation and ubiquitination, which may be detected as epigenetic marks through a variety of NGS sequencing techniques. These marks are laid down by specific enzymes (termed writers), recognized by effector proteins (termed readers), and may be removed by other enzymes (termed erasers).

The epigenetic marks, genomic contacts, and bound proteins (TFs, RNA polymerases, CTCF, cohesin, etc.) can all be sequenced to yield genome-wide maps of their

locations. Methods to process and detect significant biologically significant features and patterns from these maps is an area of active development (see one such approach in **Chapter 5**). Using the results of such approaches, it is then possible recognize distinct epigenetic states of the genome and construct models of their dynamics in development and disease. These efforts work to understand transcriptional control from a systems biology perspective, understanding that developmentally important changes involve dynamics across a network of interacting genes and regulatory elements that can rarely be understood from the activity of an individual element.

2.2.2 Post-transcriptional regulation

In the early 1990s, a short sequence of non-coding RNA was discovered to be highly conserved across species and to regulate gene expression during translation [42]. This non-coding RNA came to be known as miRNA and since its discovery, the functional associations in gene expression as well as cell proliferation and differentiation of more than 2,000 different miRNAs have reported in literature [42]. Genes encoding miRNAs are located either between genes (intergenic), and transcribed by their own promoter, or within a gene (intragenic), and transcribed by that gene's promoter [42]. Both types of miRNAs begin as pre-miRNA that are later processed into their functionally active form, a short hair pin structure as the transcript folds back on itself.

While there is evidence that miRNAs may contribute to other regulatory mechanisms, their role in post-transcriptional modification is well established. miRNA is thought to form a complex with the Argonaut, an RNA binding protein, and then hybridize with a target mRNA to initiate processes resulting primarily in negative control of gene expression, including premature termination of translation, slowed elongation of translation,

ribosomal drop off, and recruitment of factors to degrade the mRNA [42]. Aberrant miRNA activity has been reported in many cancer types and they can function as either oncogenes or tumor suppressors [44]. Differential expression of miRNAs has been linked to differences in sensitivity to chemotherapy, cancer progression, and patient survival, and circulating miRNAs are under intense investigation for use as therapeutic targets or biomarkers for diagnosis or prognosis [44] [45]. However, many mechanisms guiding the activity of miRNAs are unknown, including those guiding secretion and transport of miRNAs into circulation, their potential role in cell-to-cell communication, their interaction with coding genes, and their interaction with cell signaling pathways [46]. These unknowns present a barrier to their optimization for use cancer detection and therapy. Systems level modeling could likely provide a route for investigating their role in modulating signaling and how differences in miRNA expression between patients manifest into clinically significant differences in the efficacy of therapies and disease progression.

2.3 Cues from the cellular microenvironment

External cues from neighboring cells and other external stimuli also act to influence fate in individual cells and in turn their behavior and contribution to tissue and organ function. The external neighborhood of cell, termed the cellular microenvironment, is a key determinant of cell functional identity as maintenance of tissue specific environmental cues has been shown to be critical for maintaining cell type specific differentiation [47]. The cellular microenvironment includes soluble factors, neighboring cells, the extracellular matrix (ECM), and biophysical fields providing stimulation in the form of structural stress and strain, temperature, and electrical stimulation [47]. The extracellular matrix (ECM) within the

cellular microenvironment provides a structural foundation for cell populations and regulates cell function through control of the distribution of soluble factors and propagation of mechanical and electrical fields [47]. In this manner, the cellular microenvironment provides heterogeneous yet structured cues guiding cell spreading and movement as well as cell fates of proliferation, differentiation, and apoptosis.

In the case of cancer, the tumor microenvironment stimulates the immense heterogeneity of cells within tumors as tumor cells hijack healthy cells through cell-cell communication and ECM interaction, forcing healthy neighbors to acquire new phenotypes that support tumor growth and invasion [48]. Increasing mechanical strain on cells as tumors expand into normal tissue is also thought to activate signaling cascades that disrupt normal tissue function, reactivating mechanosensitive developmental pathways to stimulate proliferation through control of the cell cycle, epithelial-to-mesenchymal transition, and cellular motility [49].

Given the critical role of the cell microenvironment in determining cell fate, many efforts to establish regenerative therapies are concerned with recreation of tissue specific environments in culture to grow correctly differentiated tissue and stimulation of aberrant cells to return to healthy phenotypes by restoring normal environmental cues. As more mechanisms of distant cell-to-cell communication are uncovered (e.g., exosomes, cell-free DNA, and apoptotic bodies), it becomes clear that at the scale of cell populations and tissue, control of cell fate is again regulated by large and highly complex network of interacting processes.

Beyond simple cell types, with the emergence of genomic sequencing, we have come to understand that the immense heterogeneity of cell fate emerges from many simultaneously occurring processes acting at spatially and temporally heterogeneous scales. While precise experimental work is necessary to advance understanding of these processes, we can augment these efforts through computational modeling to associate distinct molecular, genomic, and environmental profiles with distinct cellular states and explore the dynamics between those state transitions to better understand the drivers of cell fate. In the next chapter, I discuss multiscale modeling as a solution well suited for study of phenomena as complex as cell fate determination and the challenges of its implementation.

3 Challenges in Constructing Multiscale Models of Biological Systems

Over the past few decades, a revolution in data storage and computing has dramatically changed scientific research. Complexity is ever increasing both in terms of systems and processes studied and through the high-dimensional and heterogeneous data created to describe them. [31] Modeling and simulation are indispensable for tackling such problems, and as high-performance computing platforms and machine learning techniques become more powerful, the complexity and scale of systems and processes that can be feasibly studied with them will only increase. [31] Nonetheless, the growing data intensiveness of modern research problems poses an evolving challenge to researchers seeking to find the right tools to address these problems. [31]

These trends and challenges have replicated across biomedical research disciplines. Advances in high throughput experimental methodologies have led to the accumulation of enormous data sets describing processes at all levels of biological organization. [25] [50] A large body of research now focuses on the development of techniques to process data generated by these recently developed modalities to identify biologically meaningful signal and on relating data across different levels of organizational scale and experimental modalities. Multiscale modeling is well positioned to address these needs and provides a deep body of knowledge for constructing and connecting mathematical representations of processes occurring at divergent scales. [50] While multiscale modeling has had many successful applications in biomedical research, several common considerations must be

addressed to result in models that are accurate, predictive, and clinically impactful. I describe these challenges and solutions to them below.

3.1 Data availability and parameter estimation

Models are generally more likely to accurately describe observed behaviors when more empirical, quantitative observations are available to construct and constrain parameter values. [26] Quantifying parameters is a challenging task in developing many single-scale biological models and this problem must be addressed for the constituent models of a multiscale model. [26] Many parameters may not be experimentally available or measurable with current technologies, and instead must instead be estimated by comparing model results to empirical ones. [26] While there are many computational techniques for performing parameter estimation, special consideration must be taken to avoid over-fitting, i.e., having too many parameters to estimate relative to the data available, a common problem when constructing models of complex biological systems which may have an extremely large number of parameters. [25] [51] Over-fitting leads to inaccurate model outputs and erodes predictive value. [25] Instead, best practice for avoiding this issue is to ensure that model parameters are drawn as much as possible from direct experimental data from collaborators and experimental and theoretical biology literature. Retrieving values from literature may require careful consideration as recorded values may not have direct correspondence to model parameters and retrieval may be complicated by data access limitations. [26] The problems of data availability and parameter estimation may be exacerbated or relieved by model type, which also has significant impact on computation efficiency, as discussed in the next section.

3.2 Computational demand and feasibility

Another common challenge in creating computational models is that of composing the model in such a way that optimizes computational efficiency to allow increasingly complex systems to be represented, models to be run at finer resolution (i.e., more time steps), and, in the case of data-driven models, input of increasingly large-scale data. [26] [31] Multiscale models can be adapted to reduce the complexity of representation of subsystems within the larger model system to improve computational efficiency. While a large body of research is devoted to developing numerical methods or machine learning based implementations of multiscale models to achieve speed up, at a more fundamental level, researchers can decide between different underlying model types to reduce the number or resolution of variables represented as needed to adapt to compute resource limits.

For instance, in the case of spatial tumor models, discrete modeling represents each constituent cell of a tumor individually with its own internal state updated in the time course of the model according to a series of pre-defined rules informed by experimental findings and biophysical principles. [26] While these models are excellent for single cell *in silico* investigation, they require many parameters which may be difficult to obtain and computational demands scale directly with the size of the tumor-cell population modeled and model resolution. [26] By contrast, continuum models represent a tumor as a continuous block of tissue rather than a population of individual cells. These models cannot be used to investigate single cell dynamics, but rather overall tumor behaviors like growth and how they are impacted by bulk genetics, or microenvironment properties, but can be executed at dramatically lower computational cost as these models use only overall tumor properties for parameter values. [26] In multiscale modeling, discrete and continuous models can be

combined to lower the computational cost of representing a more complex system while investigating fine-scale dynamics by representing different constituent single scale systems with either discrete or continuous modeling as needed to achieve necessary speed up. [26] Additionally, information is allowed to propagate between constituent single scale models of a multiscale framework, providing constraint to parameters that would be otherwise completely unconstrained in single scale representations. [26]

In the case of study of cell fate decision making, especially in the context of cancer, many indicated cellular pathways have been well characterized via high throughput measurements of time-course changes and reaction kinetics, such as the Epidermal Receptor Growth Factor (EGFR) mediated Ras-MAPK pathway and PI3K/AKT pathway, and the TP53 mediated DNA damage response and cell cycle progression pathways. [40] [52] [53] When available data are sufficient to completely characterize the dynamics of a pathway, the pathway may be modeled using fine-grained methods such as continuous differential equations (aka. reaction rate equations), but when data are not sufficient to avoid over-fitting, parameters may be sufficiently reduced by using coarse-grained approaches such as logic-based modeling of which Boolean modeling is one example. [54] As mentioned previously, in a multiscale modeling approach, continuous and discrete models can be combined to lower the computational cost of representing a more complex system. This principle was applied for the multiscale model developed in thesis. The Ras-MAPK pathway is represented with a continuous model while the TP53 mediated DNA damage response and cell cycle progression pathways are represented with logic-based models within a hybrid multiscale modeling framework representing signaling pathways contributing to cell fate decision making between healthy (differentiative) or tumorigenic (proliferative) states. This

hybrid modeling approach both allowed avoidance of the parameter estimation challenge in the case of the TP53 mediated pathways and resulted in a model that well adapted to compute power limits, as is discussed in the next section. I describe my application of this type of multiscale modeling approach in **Chapter 7**.

By increasing the computational efficiency of models, researchers are not only able to represent increasingly complex systems on high performance computing systems, but to develop models that can be easily used in clinical contexts where time and compute resources are more limited. There is enormous demand for tools that allow clinicians to personalize therapeutic regimens for individual patients, but multiscale modeling remains largely absent from clinical usage despite its enormous potential, as discussed in the next section.

3.3 Clinical applications and personalized medicine

Systems biology approaches, like multiscale modeling, are poised to have incredible clinical impact, driven by the data and computing revolution. With more data and a much higher ceiling on compute power, models can be run at far higher levels of underlying complexity and in turn achieve greater accuracy and predictive ability. Systems models of disease, designed based on recent experimental insights and modulated by patient specific -omics profiles and clinical data, could be used to develop personalized therapies and help usher in the era of precision medicine. Additionally, systems models could be used to identify relationships between clinical outcomes and dynamics of variables of complex disease systems that would be difficult and costly to search for through empirical techniques, and so contribute to hypothesis-generation and testing, biomarker identification and validation, and development of targeted therapies.

These offerings are perhaps nowhere more relevant and urgently needed than they are for cancer therapy. For some cancers, there have been incredible gains in patient survival and survivor healthiness as a result of improvements in patient risk stratification and therapeutic offerings, but for nearly all cancers there is room for improvement. For every patient, a large and complex array of factors influence the progression of their disease, the efficacy of any therapy, and the potential for recurrence. Cytotoxic drugs and radiation therapy, alongside surgery, form the standard of care for most cancers. These, along with new therapies in development, are typical tested via large scale, randomized clinical trials. These trials determine whether a drug results in favorable outcomes on average but provide little insight into why a drug works or does not achieve that, and why treatment responses might vary widely between patients, that is why a drug might be effective in one person but not in another.

These challenges can be addressed through multiscale modeling. A mechanistic model of a cancer can be constructed based on insights from experimental data describing the cancer across every level of biological organization - from the organ and tumor levels, down to the cellular, genomic, epigenetic, and molecular levels. Patient-specific -omics profiles and clinical data can be used modulate the activity of species represented in the model to obtain patient-specific predictions and predictive value can be assessed by comparing these results to true outcomes. Finally, such models can be used to isolate variables and system dynamics that differ with treatment outcomes and thus uncover the mechanistic relationships that determine why a specific treatment is effective in a specific patient.

Such approaches have great utility to both accelerate the discovery of new therapeutic targets and to model the effects of therapeutics, and multiscale cancer modeling research increasing focuses on doing just that as recent advances in gene editing techniques drive excitement for targeted therapeutics. However, advances in targeted therapy for now have little impact on the vast majority of cancer patients who continue to be administered the standard of care, combination therapy of cytotoxic drugs and radiation therapy, also known as chemoradiotherapy.

Recent advances in systems modeling have largely overlooked these therapies as many are viewed as established technology. Still in spite of their common usage, clinicians often have little information as to which chemoradiotherapy regimen will be most effective for a specific patient. Clinical decision making for these therapies remains largely based on results from traditional pharmacokinetic-pharmacodynamic (PK-PD) models. These models are mostly phenomenological in nature, relating drug dosages and their duration of treatment to macroscopic parameters like tumor volume reduction. These models are parameterized using experimental values obtained through sources like medical imaging. Models like these provide no means to investigate the underlying mechanism for a drug's effect or to reason about why the drug is more effective in one person over another.

Clinicians could instead use a multiscale modeling tool to virtually test out therapy combinations on specific patients and so determine a patient's optimal therapy regimen to be maximally effective while minimizing dose to avoid adverse outcomes. A large segment of cancer patients is burdened by chronic severe health conditions secondary to their cancer treatment. Optimizing combination therapy through patient-specific modeling presents not

only the opportunity to improve therapeutic outcomes but to increase healthiness and quality of life for survivors.

With this motivation in mind, in the next chapter, I describe the mathematical implementation of cell level models necessary to create a multiscale model of cell fate decision making, emphasizing systems and models that I incorporated into my own work to create a framework for patient-specific modeling combination chemoradiotherapy.

Part II: Single-Scale Modeling

4 Formalisms in Cellular Modeling

Computational modeling of cellular processes has evolved to encompass the needs of researchers in diverse disciplines addressing diverse problems. As such, many different modeling formalisms been developed, including equation-based models, such as those based on ordinary differential equations (ODEs), and those based on graphs, like Boolean networks [55]. These and other commonly used model types are summarized in ***Figure 4.1*** below with toy examples. While each of these formalisms have wide applicability, the choice of model should be guided by the nature of the data available for design and constraint of it.

For instance, models based on differential equations are typically for modeling dynamical systems using the equations to describe the rate of change of system variables over time. Naturally, cell signaling networks fit this description and through these models, one can perform time-course simulations of these networks, predict outputs to different inputs, and design controllers of system behavior [55]. Creating these models however requires experimental data to estimate kinetic parameters, which have historically been difficult to produce for large scale networks. By contrast, Boolean networks are populated by Boolean variables that only represent a node (gene or molecule) as having two possible states, on/active and off/inactive. At each time step in the simulation, each node's state is determined by a logic rule which is a function of the state of its input nodes (its regulators) and every node in the network is updated synchronously [55]. While far less experimental data is needed for the construction of such models, exploring the full state space still may be infeasible for large networks as the number of possible states is 2^n for n network nodes. In cases where more than two node states must be represented but kinetic data is unavailable, a

Bayesian network could be employed. In these networks, nodes are discrete or continuous random variables linked by conditional dependencies, and the value at each node is determined by its own probability function which depends on the values of input nodes [55]. This approach is excellent for inferring parameters in the presence of incomplete data, but having no representation of time at baseline, cannot be readily used to model feedback loops [55]. Finally, all of these models are ill-suited for representation of processes described by spatial data, which could be more appropriately modeled with an agent-based model or cellular automata.

Additionally, model choice may be guided by factors such as the type of analysis one would like to perform across the network and the availability of previous work for testing, which has become increasing easy with the rise of community standardized formats such as Systems Biology Markup Language (SBML) that can be readily explored with visualization and simulation software like Copasi [56] [57].

These considerations came into play in construction of the multiscale model described in **Chapter 7** with a Boolean Network being used to describe the TP53 mediated DNA damage repair and cell cycle pathways, and an ODE model being used to describe the EGFR mediated Ras-MAPK and PI3K/AKT pathways [38] [40]. Patient miRNA profiles and drug effects were modeled as actors that could change the initial states of nodes in these networks as opposed as nodes or variables in their own right acting on other network elements in the course of simulation. These two key types of networks, Boolean and equation-based, are described below. The chapter ends with a brief overview of historically significant formalisms in combination chemotherapy modeling that are later used to evaluate the results of my model in **Chapter 7**. Additionally, considerations for adaptation of the

model to better represent intra-tumor heterogeneity and the influence of cellular microenvironment via agent-based modeling are introduced in **Chapter 8**.

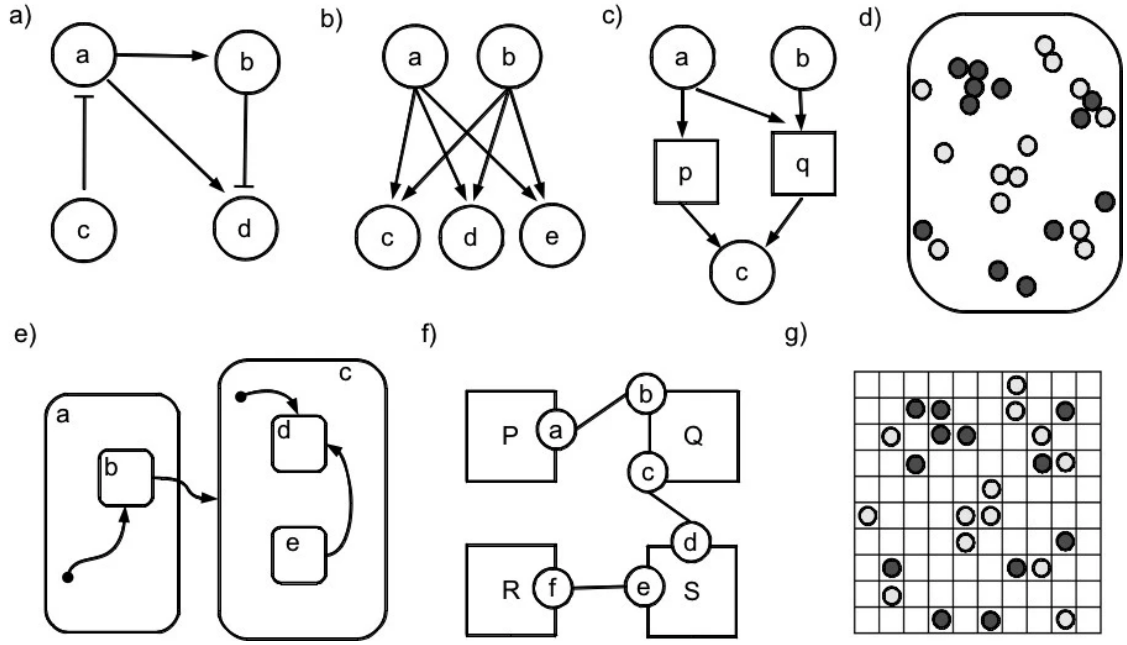


Figure 4.1. From Machado et al., visual representations of toy examples of popular systems modeling formalisms.

a Boolean network: genes are represented by nodes (a, b, c, d) and the arrows represent activation and repression **b** Bayesian network: the value of the output nodes (genes c, d, e) are given by a probability function that depends on the value of the input nodes (genes a and b) **c** Petri net: places represent substances (a, b, c), transitions represent reactions (p, q) and the arrows represent consumption and production **d** Agent based model: two types of agents representing two different kinds of cells (or molecules) can move freely and interact within the contained space **e** Interacting state machine: systems are represented by their state (a, b) where each state may contain one or more internal substates (b, d, e), arrows represent the transition between different system states **f** Rule-based model (represented by contact map): agents represent proteins (P, Q, R, S) which may contain different binding sites (a to f), the connections represent the rules for possible interactions (e.g. phosphorylation) **g** Cellular automata: a grid in which the value of each element can represent different kinds of cells (or molecules) that can change via interaction with their immediate neighbors *[55]*

4.1 Boolean Networks

The TP53 mediated DNA damage and cell cycle control pathways are large and involve numerous complex feedback loops, and in turn the dynamics of actors within it are not fully described with kinetic data. The Boolean network approach is well suited to representing such data and has been described by [38] to perform state attractor analysis of these TP53 mediated networks.

Defining the network, each node has a possible state of on or off determined as a function of values of its regulators, or input nodes [55]. The global state of the network is then defined as the state of all nodes, which are all updated synchronously at each time step in the simulation, such that the state of any node at time step t+1 is calculated from its input nodes' values at time t [55]. Each node integrates the values of its regulators via a Boolean function that can include combinations of Boolean operators such as AND, OR, and NOT. In the course of simulation runs, the network state can reach a steady state, in which node values do not change in subsequent time steps. The goal of state attractor analysis is to map initial network states to these steady states (also known as attractors) and determine how robust they are to changes input values or network structure.

4.2 Differential Equation Based Models

When rate laws and kinetic parameters are available, differential equation based models may be used to precisely represent a biological network. The system may be defined with different types of equations: ODEs are most commonly used to describe concentrations of species (genes, proteins, or molecules) as a function of time, partial differential equations to account for spatial distribution of species, stochastic differential equations to account for

stochastic components like noise, and piecewise-linear differential equations to integrate continuous features with discrete features (like threshold based switches) [55] [58]. Such a piecewise-linear differential equation model composed of ODEs is employed to represent the EGFR mediated Ras-MAPK and PI3K/AKT (see **Chapter 7**) for these reasons.

Generally, to set up an ODE based model, we can express the network by a set of equations with the species amount or activity level as a variable $\frac{dx_i}{dt} = f_i(x, u)$ where x is a n by 1 vector of the amount of each individual species in the network, u is an n by 1 vector of the external stimuli affecting each species (set to 0 if not including external inputs) and f_i is a continuous function [59]. Commonly, in representing gene regulatory networks, the effect of one species on another is not instantly realized (as intermediate mechanisms act to carry out the regulatory effect) and these discrete time delays may be accounted for using time-delayed ODE equations such as $\frac{dx_i}{dt} = f_i\left(x_1\left(t - \tau_{i,1}\right), x_2\left(t - \tau_{i,2}\right), \ldots, x_n\left(t - \tau_{i,n}\right), u\right) - \gamma_i x_i$, where γ_i

is the degradation rate constant for species i and $\tau_{i,j}$ is the delay in regulation of species i by species j [59]. Other functions like the Hill Equation can be used to guide the rate of a species regulation by another and these equations can be solved piecewise to eliminate nonlinearities that might make the system otherwise infeasible to solve.

4.3 Combination Chemotherapy Models

Additivity models have been historically been employed to mathematically predict dose-response relationships for combination chemotherapy from experimentally determined dose-response relationships of the individual therapies. Two such models have remained in

widespread use for decades: the Loewe additivity model and the Bliss additivity model, shown in ***Figure 4.2***.

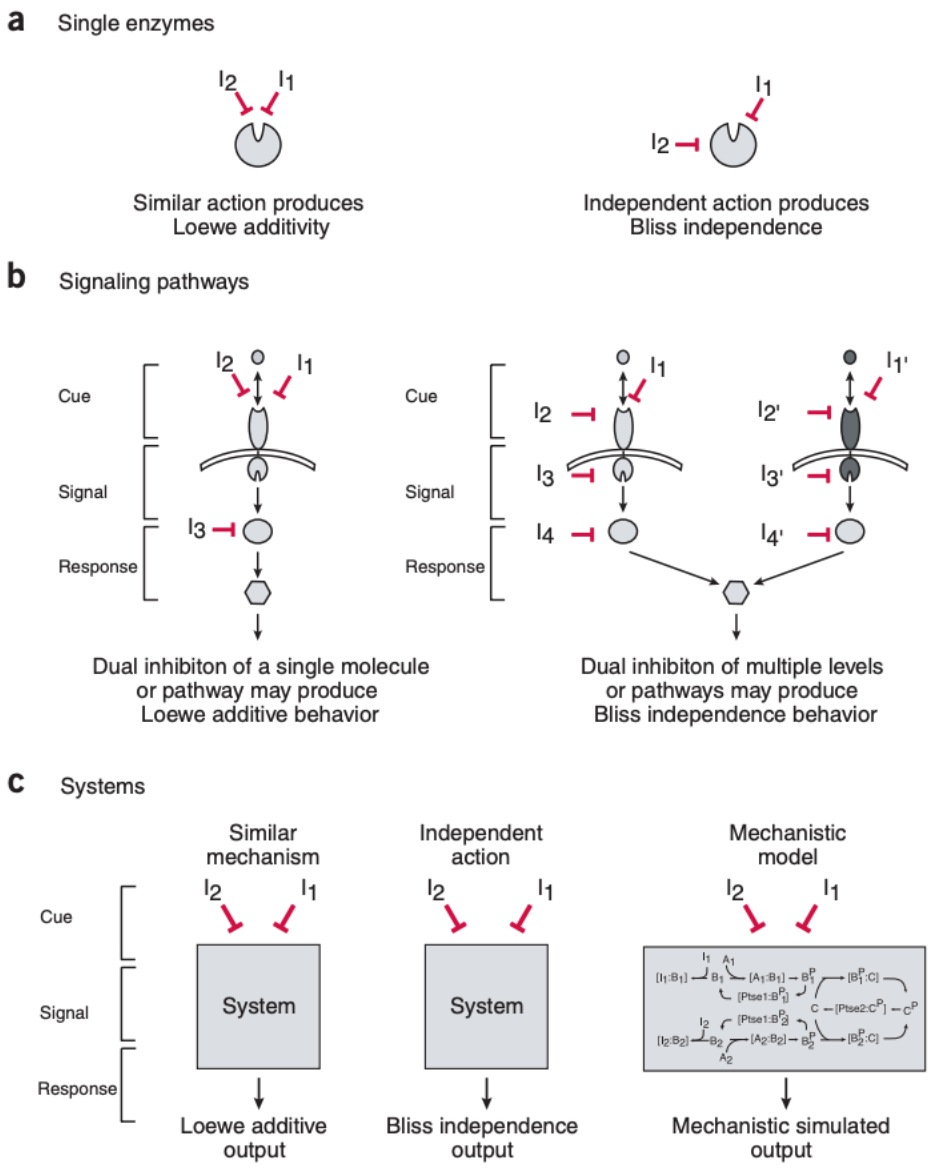


Figure 4.2. From Fitzgerald et al., a toy representation of Loewe and Bliss additivity models contrasted with that of a mechanistic modeling approach.

a Single enzymes: (Left) According to Loewe additivity, combinations of enzyme inhibitors act upon overlapping binding sites. (Right) According to Bliss independence, combinations of enzyme inhibitors act upon independent binding sites. **b** Application of Loewe additivity

and Bliss independence to signaling networks is unintuitive. (Left) Loewe additivity behavior could possibly be observed when inhibitor combinations act to inhibit the same pathway through similar action. (Right) Bliss independence behavior could be observed for inhibitor combinations that act independently at different sites on the same target, different levels in the same pathway, or upon different pathways. **c** Loewe additivity and Bliss independence do not account for the mechanisms of inhibitor interactions in complex systems, instead treating these systems as black boxes. Mechanistic models can capture complex signaling dynamics and so be used to compute how inhibitor combinations will perform. *[60]*

Loewe additivity assumes that two drugs act on a target through a similar mechanism, resulting in dose substitution and that to reduce cell survival by the same proportion X% achieved individually, the concentrations in combination can be calculated from the relationship $1 = \sum_{i=1}^{n} [\frac{[C_i]_{X\%}}{[I_i]_{X\%}}]$ where $[I_i]_{X\%}$ is the concentration of drug i needed to reduce cell survival by X% when administered individual and $[C_i]_{X\%}$ is the concentration of drug i needed to reduce cell survival by the same amount when administered in combination [60]. By contrast, the Bliss independence model assumes that drugs act on a target through independent mechanisms, resulting in effect multiplication; the effect of the combination therapy is predicted using the equation $F_{UA} = \prod_{i=1}^{n} F_{UA_i}$ where F_{UA} is the fraction of targets unaffected by combination therapy and F_{UA_i} as the fraction of targets unaffected during individual administration of drug i at the same dosage when used in combination therapy [60].

These models were originally designed to describe simple enzymatic interactions and do not adequately account for mechanisms underlying the interaction of actual chemotherapies with complex cell fate decision networks. Mechanistic models, like those constructed using the systems formalisms described early in this section, could represent the

dynamics of these networks to compute how these combinations will perform, more realistically than mechanism-agnostic additivity models.

5 A Statistical Tool for Detecting Epigenomic States in Lineage Commitment

Adapted from [61]

While multiscale, multiphysics models present a solution to the problem of synthesizing observations of biological phenomena across biological scales and heterogeneous data, new methods are also needed to pre-process empirical data generated with recently developed NGS technologies. As a solution to one such problem, I developed 3DeFDR, a statistical tool for detecting chromatin looping interactions that dynamically change across biological conditions. This tool was developed with the intent to answer the question of to what extend do long-range looping interactions change across developmental models, genetic perturbations, drug treatments, and disease states. Together with my co-author, I ultimately created a tool for identifying such dynamic loops from high-resolution Chromosome-Conformation-Capture-Carbon-Copy (5C) and Hi-C data. In this chapter, I demonstrate this method in analysis of data sets capturing chromatin looping states in the course of neural lineage commitment, including cross-reference of differential loop calls with RNA-seq and ChIP-seq results. I anticipate that this method could be used to help construct a more complete picture of epigenetic states in the course of lineage commitment and in turn, a more realistic multiscale model of lineage commitment.

5.1 Introduction

Chromosome-Conformation-Capture (3C)-based molecular techniques have recently been coupled with high-throughput sequencing to generate genome-wide maps of higher-order chromatin folding [62, 63, 64] . A number of massively parallel 3C-based technologies query

genome folding in a protein-independent manner, including Hi-C, 4C, 5C, and Capture-C [65, 66, 67, 68, 69, 70, 71]. All four techniques rely on proximity ligation and high-throughput sequencing to convert physically connected chromatin fragments into counts of specific interaction events. Briefly, chromatin is fixed in its native architectural state across a population of cells and then digested with a restriction enzyme. Restriction fragments are ligated to form billions of hybrid ligation junctions between two distal genomic loci. The two fragments in a given ligation junction can then be identified using high-throughput sequencing, and their frequency of ligation is proportional to their spatial proximity across a population of cells. Hi-C detects all chromatin interactions genome-wide using high-throughput sequencing, whereas 5C and Capture-C use tiled probes to selectively sequence large, megabase-scale subsets of the genome. 4C queries all genome-wide contacts involving a single chosen restriction fragment. Thus, the protein-independent 3C technologies of Hi-C, 5C, and Capture-C can be used to create high-resolution spatial maps of genome folding on the scale of a few megabases to genome-wide coverage.

Recently published 3C-based sequencing studies have revealed that the mammalian genome is folded into a hierarchy of distinct architectural features, including A/B compartments, lamina-associated domains (LADs), topologically associating domains (TADs), subTADs, and long-range looping interactions [67] [69] [71] [72] [73] [74] [75] [76] [77] [78] [79] Loops—groups of adjacent pixels which form a punctate focal increase in interaction frequency enriched above local TAD and subTAD structure—have been identified algorithmically in high-resolution Hi-C maps [72]. The highest resolution maps to date have enabled the detection of tens of thousands of looping interactions genome-wide [72] [80]. A subset of looping interactions occur at the corners of TADs/subTADs and are

known as "corner dots." A leading model for the mechanism of corner dot formation is that cohesin tracks along the chromatin fiber until it is blocked by the architectural protein CTCF, thus extruding out the intervening DNA [81] [82] [83] [84] [85] [86]. Corner dot TADs/subTADs anchored by CTCF are thought to demarcate the search space of enhancers for their target promoters [87] [88] [89] [90]. Moreover, enhancers can also connect directly to target genes via corner dots in a CTCF-dependent and CTCF-independent manner [91] [92] [93] [94]. Initial studies have suggested that specific subsets of looping interactions can reconfigure in development, disease, and in response to genetic perturbations [80] [89] [91] [92] [95] [96] [97] [98] [99] [100] [101]. Generally, however, it remains unknown to what extent loops are dynamically altered genome-wide as cells switch fate, due in part to the relative paucity of computational methods to evaluate statistically significant changes in interaction frequency across multiple biological conditions.

As high-resolution Hi-C and 5C chromatin folding maps begin to accumulate in developmentally relevant cellular models, there is an increasing need for methods to (1) precisely detect loops and clearly distinguish them from other classes of architectural features such as local TAD/subTAD structure and compartments and (2) rigorously classify loops by their dynamic behavior across cell types. A number of computational methods report the ability to identify loops in individual libraries generated by Hi-C. Forcato and colleagues performed a detailed comparison of Hi-C loop calling pipelines, including HiCCUPS [102], GOTHiC [103], HOMER (http://homer.ucsd.edu/homer/interactions/), diffHic [104], HIPPIE [105], and Fit-Hi-C [106]. The conclusion from this study was that loop calling methods in individual samples exhibit vastly different performance, with no clear gold standard emerging [107]. Importantly, most loop calling pipelines were developed

on low-resolution maps (40 kb up to 1 Mb bins) generated with the first-generation dilution Hi-C experimental procedure. More recently, Hi-C maps have achieved 1–5-kb resolution through higher read depth and markedly reduced spatial noise due to second generation in situ ligation and digestion techniques [72] [80]. We also note that active, unsynchronized extrusion events could create long-range interactions within TADs/subTADs that do not manifest as punctate loops in a 5C/Hi-C heatmap (i.e., transient loops in the making) [84]. Thus, it is likely that first generation loop calling algorithms show a wide dynamic range of performance because they were developed on lower resolution first-generation Hi-C maps and did not explicitly distinguish loops from general non-specific, long-range interactions. The emerging model from second-generation Hi-C studies is that quantitative loop detection in individual libraries requires rigorous modeling of local chromatin domain structure. HiCCUPS explicitly models and accounts for locus-specific TAD/subTADs [72], and accounting for local chromatin domain structure has therefore emerged as a leading candidate for identifying bona fide loop structures (i.e., persistent loops) in individual Hi-C maps. Building upon advances in Hi-C, similar statistical methodologies have been applied in lib5C to find loops in individual 5C maps [108].

To our knowledge, computational tools are not yet available to statistically test loops for their differential signal across two or three conditions in 5C data. Three tools (diffHic [104], FIND [109], and HiBrowse [110]) have been published to identify generally differential interactions between conditions in Hi-C data. All three methods in their published, first-generation form were not designed or verified to distinguish loops from higher-order folding patterns such as A/B compartments, TADs, subTADs, or non-specific long-range interactions. In the absence of accounting for these features, a large proportion

of the differential interactions identified may be due to cell type-specific fluctuations related to technical biases, local chromatin domains, extrusion lines, or higher-order compartments. Noteworthy, the diffHic manuscript indicates that modeling local chromatin domain structure would be essential to evaluate cell type-specific loops, suggesting that second-generation tools which accomplish this might be available in the future [104]. Computational tools have also been published to call within- and across-condition loops from libraries generated by Hi-ChIP and ChiA-PET assays [111] [112] [113] [114] [115] [116]. However, statistical frameworks built for protein-dependent 3C-methods cannot address the technical challenges unique to 5C and Hi-C data. Overall, a gold-standard statistical methodology for cell type differential loop detection in protein-independent proximity ligation data (both 5C and Hi-C) is an important unmet need.

Here, we present 3DeFDR, a new statistical method and software implementation for identifying cell type-specific looping interactions from genome-wide Hi-C (3DeFDR-HiC) and locus-specific 5C (3DeFDR-5C) data across two or three biological conditions. For locus-specific 5C matrices, 3DeFDR-5C computes an empirical false discovery rate (eFDR) by applying a thresholding scheme on the change in interaction score signal on real 5C libraries from multiple biological conditions and pseudo-replicates simulated from the same biological condition. We implement a controlling procedure in which we iterate thresholds to achieve an a priori determined eFDR under the assumption that all thresholded pseudo-replicate interactions simulated from the same condition are false positives. For genome-wide Hi-C matrices, 3DeFDR-HiC formulates a negative binomial likelihood ratio test parameterized with a Distance-Dispersion-Relationship (DDR) for every pixel engaged in persistent loops genome-wide. Cell type-specific loops called by 3DeFDR-5C have fewer

false positives and are more strongly enriched for chromatin modifications characteristic of the cellular state in which the loops are present compared to (i) an established ANOVA test and (ii) our own newly formulated parametric likelihood ratio test (3DLRT). We also benchmarked 3DeFDR-HiC against the leading published Hi-C non-specific differential interaction calling method diffHic and demonstrate superior performance. 3DeFDR-5C, 3DeFDR-HiC, and the parametric benchmarking test 3DLRT are freely available as Python packages to support the next wave of discoveries in cell type-specific looping.

5.2 Results

We set out to address a critical challenge in the analysis of looping interactions in 5C data: the paucity of methods for robustly classifying dynamic loops across multiple cellular conditions, a problem which becomes more challenging as the number of conditions increases. Our goal was to develop a statistical framework and software implementation to rigorously identify differential loops from 5C maps across two or three conditions using a target FDR to choose thresholds (***Fig. 1a***).

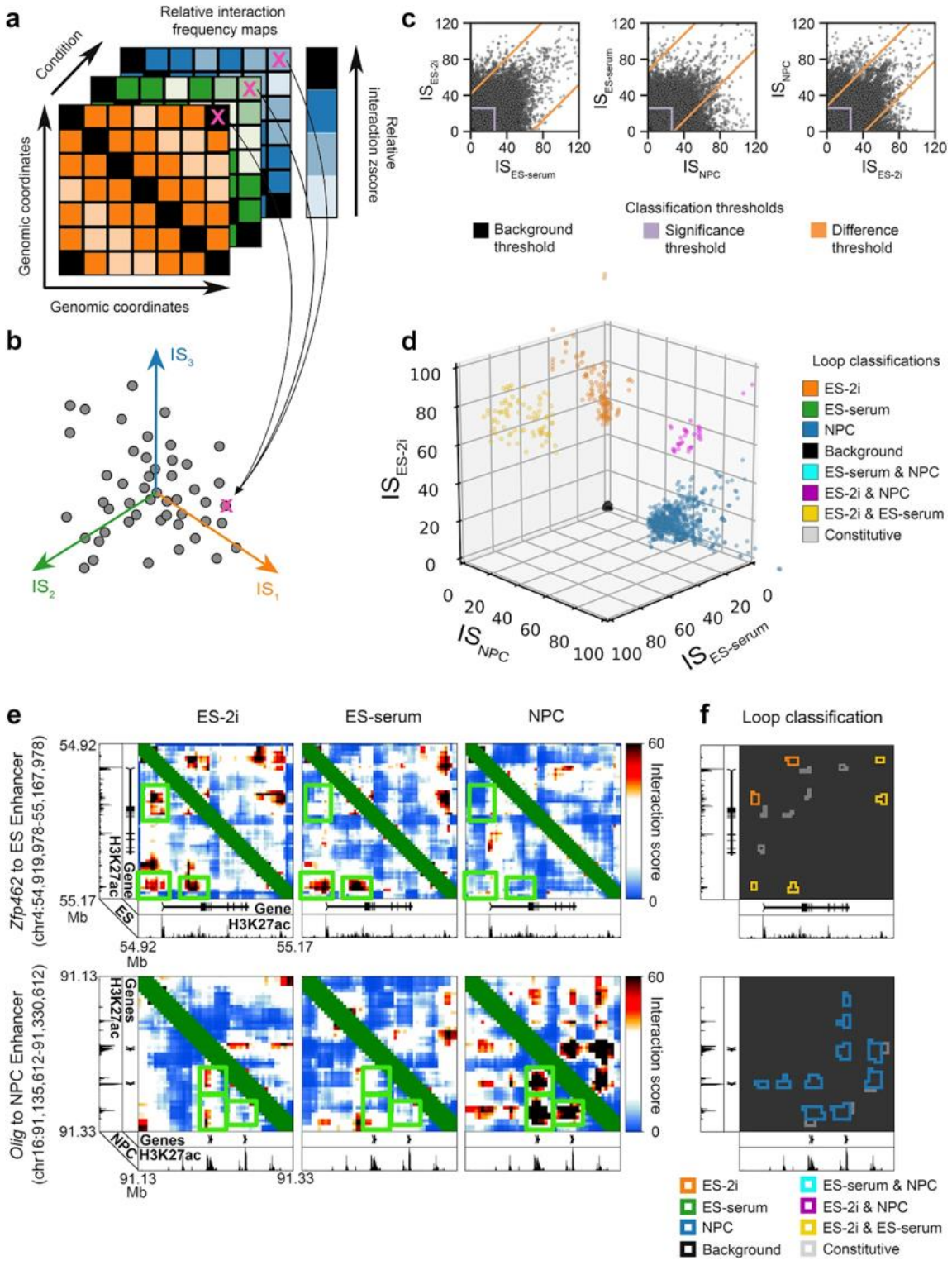

a
Relative interaction frequency maps
Condition
Genomic coordinates
Genomic coordinates
Relative interaction zscore
b
IS3
IS2
IS1
c
Classification thresholds
Background threshold
Significance threshold
Difference threshold
d
Loop classifications
ES-2i
ES-serum
NPC
Background
ES-serum & NPC
ES-2i & NPC
ES-2i & ES-serum
Constitutive
e
ES-2i
ES-serum
NPC
Zfp462 to ES Enhancer (chr4:54,919,978-55,167,978)
Olig to NPC Enhancer (chr16:91,135,612-91,330,612)
Interaction score
f
Loop classification

Figure 5.1. Overview of interaction score thresholding procedure for cell type-specific looping interaction classification.
a A 5C dataset is input as a set of interaction frequency matrices, with each matrix capturing the same set of genomic contacts under a different cellular condition. **b** Raw 5C counts are converted to interaction scores (IS) which reflect bias-corrected, sequencing depth normalized, local expected background signal normalized, and statistically modeled interaction frequency values that are comparable within and between conditions under the assumptions of our model (detailed in the "Methods" section and ***Fig. 4***). **c** Interaction scores are thresholded to allow detection and classification of looping interactions that are significantly differential across cellular conditions. **d** Seven looping interaction classes after a 3-way thresholding scheme on ES-2i, ES-serum, and NPC cellular states. **e** IS heatmaps at two selected genomic loci. Green boxes highlight regions of qualitatively apparent differences in looping signal. **f** Loop classification results after applying 3DeFDR-5C's 3-way IS thresholding procedure

First, we developed, applied, and benchmarked 3DeFDR-5C using 5C data across three distinct cellular states: mouse embryonic stem (ES) cells cultured in 2i media representing a naive pluripotent state, mouse ES cells cultured in LIF/serum representing the primed pluripotency state, and primary neural progenitors isolated from neonatal mice representing a multipotent adult stem cell state in the neuroectoderm lineage (Additional file 1: Table S1) [94]. These particular 5C datasets represent large-scale, 4-kb-resolution maps capturing 8 Mb of genomic sequence around key developmentally regulated genes. 5C relies on a primer-based hybrid capture step to selectively detect ligation junctions across specific genomic regions, thus enabling the creation of high-resolution matrices with a strikingly lower number of reads (~ 30–40 million per sample) compared to Hi-C (~ 3–6 billion per sample). We have recently determined that loops are markedly reconfigured during the transition from naive pluripotency to multipotency, thus making this dataset ideal for the testing and

development of our statistical framework. We tested and validated 3DeFDR-5C with a three cellular state experiment, but the statistical framework and code are also able to analyze a two cellular state experiment.

We first started by modeling and correcting biases, artifacts, and local chromatin domains in individual replicates. Despite their nuanced technical differences, data from protein-independent proximity ligation techniques share several common features, including: (1) distance-dependent background interaction signal in which non-specific interaction frequency is highest for the closest fragment-fragment pairs on the linear genome and decays as the distance separating the genomic fragments increases [67], (2) biases in ligation and amplification frequency caused by GC content and length of restriction fragments [117] [118] , (3) library complexity and sequencing depth differences across independent experiments for the same biological sample leading to nonlinear batch effects [108], and (4) highly locus-specific structure due to higher-order folding of chromatin into TADs, subTADs, and compartments [72]. One must model and address these features to ensure a rigorous analysis of looping interactions.

We reasoned that a differential loop calling method would have the most utility across protein-independent proximity ligation data if it started with a modified interaction score (IS) in which background signal as well as per-replicate and per-pixel confounding factors had been corrected. We recently discovered that sequence-related biases are not constant across cell types and replicates. Therefore, as is routinely done with Hi-C data, we used matrix balancing to correct for fragment-specific biases caused by GC content and restriction fragment length for every replicate individually (detailed in the "Methods" section). We also used conditional quantile normalization to normalize all replicates for non-

linear library complexity and sequencing depth differences (detailed in the "Methods" section). It is widely known that the distance-dependent background signal and local chromatin domain structure are widely variable across cell types and highly unique to each genomic region. Thus, we used the donut and lower left filters [72] [91] to model the distance-dependent and TAD/subTAD expected background signal for every interaction in the genome and every replicate individually (detailed in the "Methods" section). After bias correction, background normalization, and expected modeling, we assigned p values to every pixel in the 5C heatmap and computed an interaction score (IS) that allows for direct comparison of each bin-bin pair across replicates and conditions under the assumption that the replicates are similarly powered (***Fig. 1***). Moreover, the use of modeled IS as the random variable for differential testing allows 3DeFDR to have utility for matrices of any protein-independent 3C-based data that have been bias corrected, normalized, modeled, and transformed into p values using analysis techniques tailored to the specific method.

To identify differential looping interactions, we used a classification technique that relies on three-way thresholding on the difference in IS across cellular conditions (***Fig. 1***, Additional file 2: Fig. S1, Additional file 3: Table S2). For each biological replicate, we began with a framework in which IS is a square, symmetric matrix of interaction scores from a modeled and bias-corrected 5C experiment. The matrix IS has dimensions n by n, where n is the number of genomic bins in any particular genomic region, r. We use $IS_{ts,r,k,l}$ to refer to the interaction score between genomic bins k and l in region r as recorded for biological replicate s of condition t (detailed in the "Methods" section). We first identify potential looping interactions by parsing only bin-bin interactions with an $IS_{ts,r,k,l}$ greater than a specific significance threshold g for all replicates in at least one condition (purple lines, ***Fig.***

1). We then apply a series of thresholds (orange lines, ***Fig. 1c***) on the difference in $IS_{ts,r,k,l}$ across all three cellular conditions (Additional file 2: Fig. S1E-G, Additional file 3: Table 2S). To ensure the most conservative estimate of looping classes, we apply the thresholds on the minimum difference in IS across replicates of each condition. Thus, the end result is a preliminary set of seven classes of looping interactions: (1) ES-2i only, (2) ES-serum only, (3) NPC only, (4) ES-2i and ES-serum only, (5) ES-2i and NPC only, (6) ES-serum and NPC only, and (7) constitutive across all three cell types (***Fig. 1***). Examples of ES-2i only, ES-2i and ES-serum only, and NPC only interactions are illustrated in ***Fig. 1, f***.

We next used estimation and control of an empirical false discovery rate (eFDR) to guide the final placement of the difference thresholds for each looping class (orange lines, ***Fig. 1c***, detailed in the "Methods" section). The false discovery rate (FDR) is by definition FDR = E[V/R] where V is the number of false positives among tests declared significant and R is the total number of tests declared significant. Here, R is trivial to compute from our set of three conditions (T = {A, B, C} where A is ES-2i, B is ES-serum, and C is NPC) and six replicates (S = {A1, A2, B1, B2, C1, C2}) as the total number of significant bin-bin interactions in a given looping class (H = {{A}, {B}, {C}, {A, B}, {A, C}, {B, C}}). However, V is not known and requires a method for estimating the false-positive rate of our three-way thresholding procedure.

We hypothesized that V is approximately equal to the total number of interactions labeled as differential when applying 3DeFDR-5C to a set of biological samples with no true differential loops. We defined our null dataset as a set of samples that are replicates of a single cellular condition but are assigned a random set of labels matching conditions T. Our key assumption in formulating this approach is that that the false-positive rate (FPR) of calls

on the null dataset (FPR_{null}) is approximately equivalent to that of the experimental dataset (FPR_{exp}), such that $FPR_{null} \approx FPR_{exp}$. We computed and controlled an empirical false discovery rate (eFDR) as in Eq. 1:

$$eFDR = \frac{n_{null}}{n_{exp}} \approx \frac{V}{R} \quad \textbf{(1)}$$

where n_{exp} is the total number of interactions classified as significantly differential for a particular looping class using the experimental conditions T and n_{null} is the total number of interactions classified as significantly differential in the null dataset, which approximates $FPR_{exp.}$.

It is often cost prohibitive to generate six biological replicates of 5C data for each condition. Therefore, we generated 5C replicate simulations to populate the null sample set. We simulated 5C replicates of the same condition at the level of fragment-fragment ligation counts after conditional quantile normalization. Our rationale for this decision was that it would allow us to omit library complexity, batch effect, and sequencing depth terms in our count generating models. To construct our simulation generating model, we first computed the sample mean and sample variance for every interaction in every condition (Equations 2 and 3):

$$\mu_{t,r,i,j} = \frac{\sum_{s=1}^{n_t} C'_{t_s,r,i,j}}{n_t} \quad \textbf{(2)}$$

$$\sigma^2{}_{t,r,i,j} = \frac{\sum_{s=1}^{n_t} \left(C'_{t_s,r,i,j} - \mu_{t_s,r,i,j}\right)^2}{n_t - 1} \quad \textbf{(3)}$$

where n_t is the number of replicates of condition t and $C'_{t_s,r,i,j}$ is the conditional quantile normalized 5C counts of interaction (t,r,i,j) in the s^{th} replicate of condition t for every i^{th} and j^{th} fragment ligation in genomic region r.

Most genomics experiments suffer from poor parameter estimation due to the low number of replicates that are financially and logistically feasible to generate for every biological condition. To improve parameter estimates, we modeled the mean-variance relationship (MVR) between $\mu_{t,r,i,j}$ and ${\sigma^2}_{t,r,i,j}$ by pooling all interactions at similar interaction distances (***Fig. 2***). We stratified quantile normalized counts ${C'}_{t_s,r,i,j}$ for all regions by their linear genomic interaction distance using a dynamic size window (***Fig. 2a***). For distance regime 1 (0–150 kb), we stratified the interactions into fine-grained, 12-kb-sized sliding windows with a 4-kb step. For distance regime 2 (151–600 kb), we stratified the interactions into 24-kb-sized sliding windows with an 8-kb step. For distance regime 3 (601–1000 kb), we stratified the interactions into coarse-grained, 60-kb-sized sliding windows with a 24-kb step. We found that the variance was greater than the mean across all genomic distance scales, indicating that 5C counts data are overdispersed (***Fig. 2***). For each window in each distance regime, we modeled the MVR by fitting the function (Equation 4):

$$\hat{\sigma}^2{}_{t,r,i,j} = A_{t,w}\mu^2_{t,r,i,j} + \mu_{t,r,i,j} \quad \textbf{(4)}$$

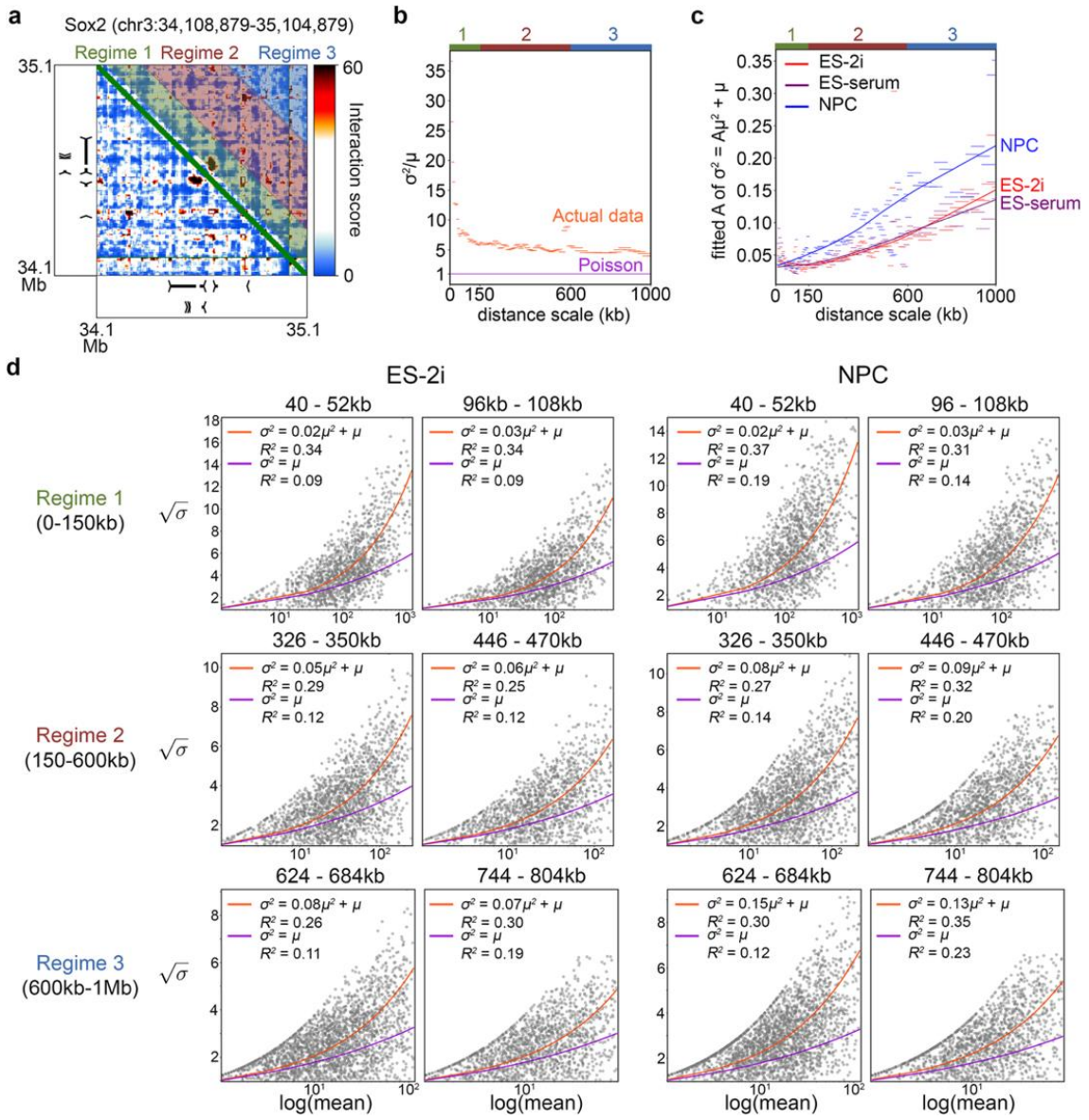


Figure 5.2. 5C counts are overdispersed and their mean-variance relationship varies as a function of linear genomic distance and cellular condition.

a Raw 5C contacts are stratified by genomic distance prior to characterization of their mean-variance relationship. In each of our three regimes, the width of the stratification windows is determined using a different binning scheme. **b** The coefficient of variation for raw 5C counts is plotted against the median genomic interaction distance for each sliding window. Each window captures counts from all genomic regions in the dataset in the ES-2i condition. **c** The dispersion parameter, A, for each distance scale window (short horizontal

lines) is computed by fitting sample means and variances to the function $\sigma^2 = A*\mu^2 + \mu$. Dispersion versus distance scale trends (solid smooth lines) were generated by Loess smoothing. **d** Mean-variance models for representative genomic distance windows from all three distance regimes. Fits of the Poisson mean-variance relationship ($\sigma^2 = \mu$) and the negative binomial mean-variance relationship ($\sigma^2 = A*\mu^2 + \mu$) are shown with their corresponding R^2 goodness of fit values

to all $\mu_{t,r,i,j}$ and $\sigma^2{}_{t,r,i,j}$ to find the overdispersion parameter, $A_{t,w}$, at each distance scale (detailed in the "Methods" section). We found that $A_{t,w}$ also varied as a function of distance and was unique to each cell type (***Fig. 2***). Together, these data demonstrate that 5C counts are overdispersed and that the overdispersion parameter varies as a function of distance and cellular state.

To generate simulated 5C libraries, we weighted the predicted variance $\hat{\sigma}^2{}_{t,r,i,j}$ against the original observed variance $\sigma^2{}_{t,r,i,j}$ to generate a final weighted variance $\underline{\sigma}^2_{t,r,i,j}$ for each interaction at each distance scale as in Equation 5 (detailed in the "Methods" section):

$$\underline{\sigma}^2_{t,r,i,j} = \alpha\hat{\sigma}^2{}_{t,r,i,j} + \beta\sigma^2{}_{t,r,i,j} \quad \textbf{(5)}$$

We used $\alpha = \beta = 0.5$ to achieve simulated 5C counts with pairwise correlations on par with that of real replicates while improving the quality of our variance estimate with the predicted contribution (Additional file 4: Table S3). Finally, we parameterized the negative binomial model for each $C'_{t,r,i,j}$ interaction and generated simulated counts from our models for each (t, r, i, j) interaction (Equation 6):

$$C'^{sim}_{t,r,i,j} \sim NB(\mu_{t,r,i,j}, \underline{\sigma}^2_{t,r,i,j}) \quad \textbf{(6)}$$

We created simulated replicates by filling in a simulated counts value for each (t, r, i, j) interaction with a random variable drawn from the negative binomial distribution

parameterized by $\mu_{t,r,i,j}$ and $\underline{\sigma}^2_{t,r,i,j}$. We then subjected the simulated 5C libraries, $C'^{sim}_{t,r,}$, to the same matrix balancing, binning, expected normalization, and modeling as the real 5C libraries (see the "Methods" section). Simulated 5C counts were highly similar to real 5C data in a qualitative comparison (***Fig. 3a–d***, Additional file 2: Fig. S2). Moreover, for the final predicted variance estimates (Equation 5 weighted at $\alpha = \beta = 0.5$), our simulated 5C libraries exhibit Spearman's correlations within and between conditions that are nearly equivalent to real replicates (***Fig. 3***). Together, these data show that 5C libraries can be simulated with a negative binomial distribution parameterized with an overdispersed distance-specific MVR.

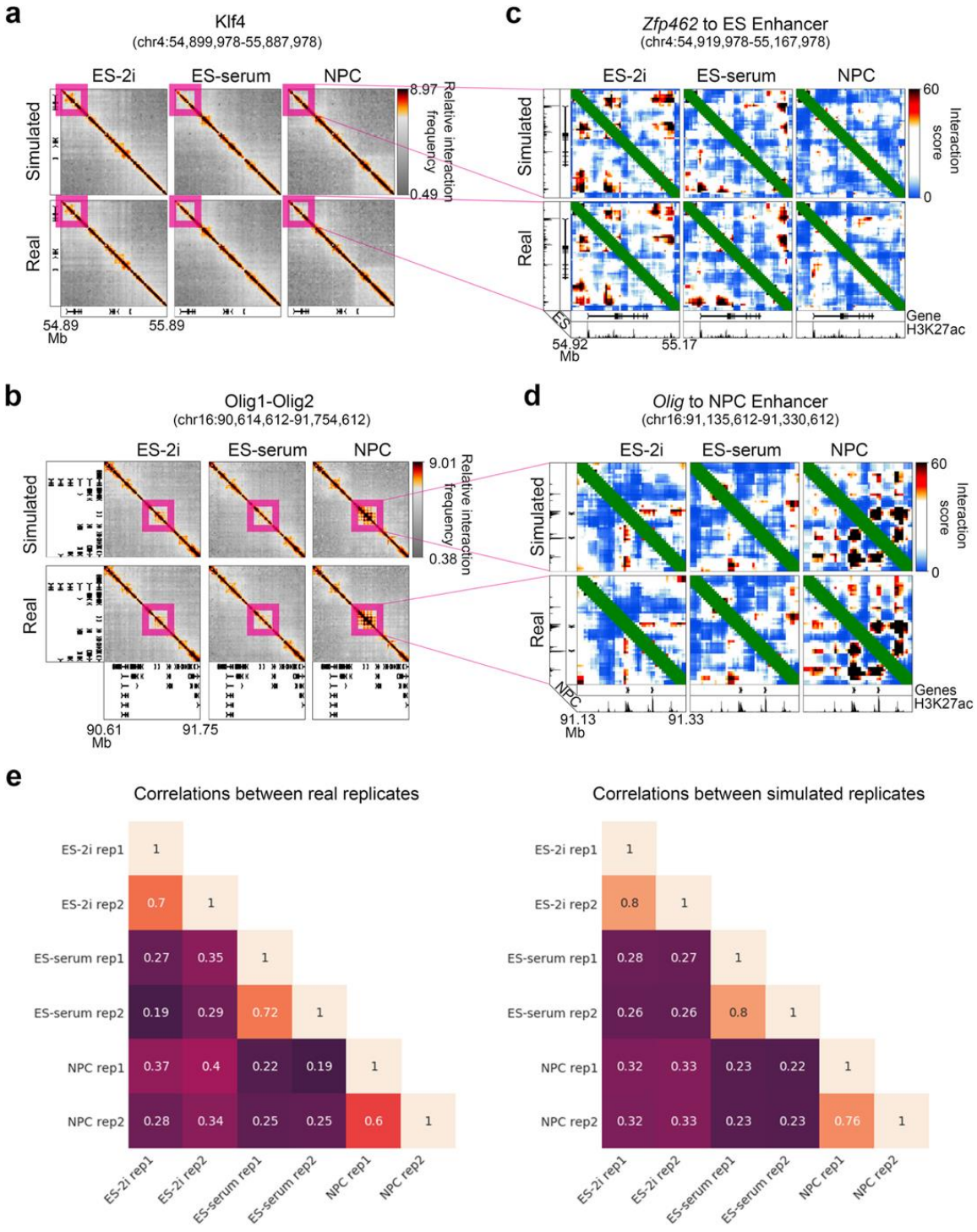


Figure 5.3. Simulated 5C datasets exhibit strong similarity to experimental 5C datasets. **a**, **b** Heatmaps of relative 5C interaction frequency in the genomic regions surrounding the **a** Klf4 and **b** Olig1/2 genes are shown for simulations and real experimental

data. **c**, **d** Heatmaps of interaction scores in the genomic regions surrounding the **c** Klf4 and **d** Olig1/2 genes are shown for simulations and real experimental data. **e** Matrices of pairwise Spearman's correlations between real and simulated 5C replicates after conditional quantile normalization (see the "Methods" section)

We next used simulated IS matrices (***Fig. 4a***) to compute an empirical FDR (eFDR) estimate for our looping classes across a sweep of IS difference thresholds applied to both real ($IS_{t_s,r,k,l}$) and simulated ($IS^{sim}_{t_s,r,k,l}$) values. For each loop classification, we computed eFDR estimates across a range of difference threshold values d, acquiring a difference threshold-to-eFDR mapping for each class, $eFDR_{d,h}$, as in Equation 7:

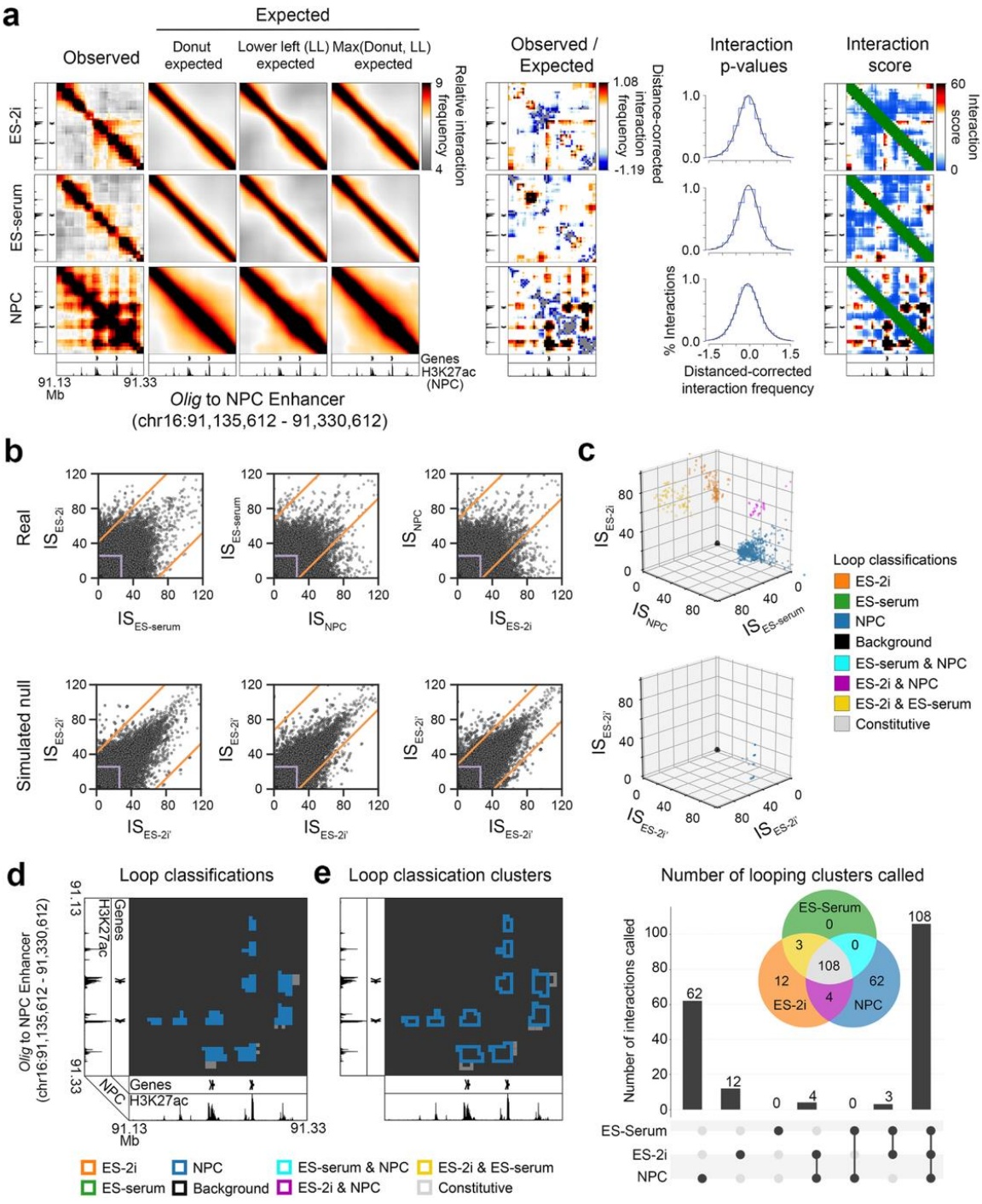


Figure 5.4. Application of 3DeFDR-5C to find cell type-specific looping interactions across three cellular states.

a Heatmaps representing binned, matrix balanced 5C counts (Observed) around a known looping interaction between the Olig1 gene and an NPC-specific enhancer (chr16:91,135,612-91,330,612). Observed counts are divided by the computed local expected signal to obtain background-normalized counts (Observed/Expected). These counts are fitted with a logistic distribution and the resulting p-values are transformed into interaction scores, where interaction score = − 10*log2(p value). **b** Interaction scores are thresholded to isolate contacts that are differentially looping across cellular conditions and whose signal meets a baseline requirement for significance. This thresholding procedure is applied to both real and simulated null replicate sets to compute an eFDR estimate. The dynamic thresholding procedure is applied with increasing stringency until a user-specified target false discovery rate is reached. **c** Loop classifications obtained with 3DeFDR-5C in real (top) and simulated null (bottom) replicate sets shown in an interaction scatterplot representation. **d**, **e** Heatmap of final loop classifications at **d** individual bin-bin pairs and **e** classified looping clusters after applying 3DeFDR-5C at a threshold of 2%. **f** UpsetR scalable Venn diagrams for differential looping clusters called by 3DeFDR-5C at a target eFDR of 2%

$$eFDR_{d,h} = \frac{card\left(\left\{(r,k,l) \in h_{null}^{d}\right\}\right)}{card\left(\left\{(r,k,l) \in h_{exp}^{d}\right\}\right)} \quad \textbf{(7)}$$

where h_{null}^{d} represents the set of interactions assigned to differential class h in the simulated null dataset at difference threshold d and h_{exp}^{d} represents the set of interactions assigned to the same class in the real experimental dataset at the same difference threshold d. We selected our final eFDR threshold τ as 2% (***Figs. 4 and 5***). We performed this eFDR controlling procedure for every differential looping class across our three cellular states to identify significantly differential bin-bin pairs (***Fig. 4***). We then clustered significantly differential bin-bin pairs of a similar looping class by spatial adjacency (see the "Methods" section); the end result was 108 constitutive, 12 ES-2i only, 62 NPC only, 3 ES-2i and ES-

Serum, and 4 ES-Serum and NPC looping clusters (***Fig. 4***). The 3DeFDR-5C algorithm is designed so that the user can tune the final looping classifications to a pre-determined target eFDR.

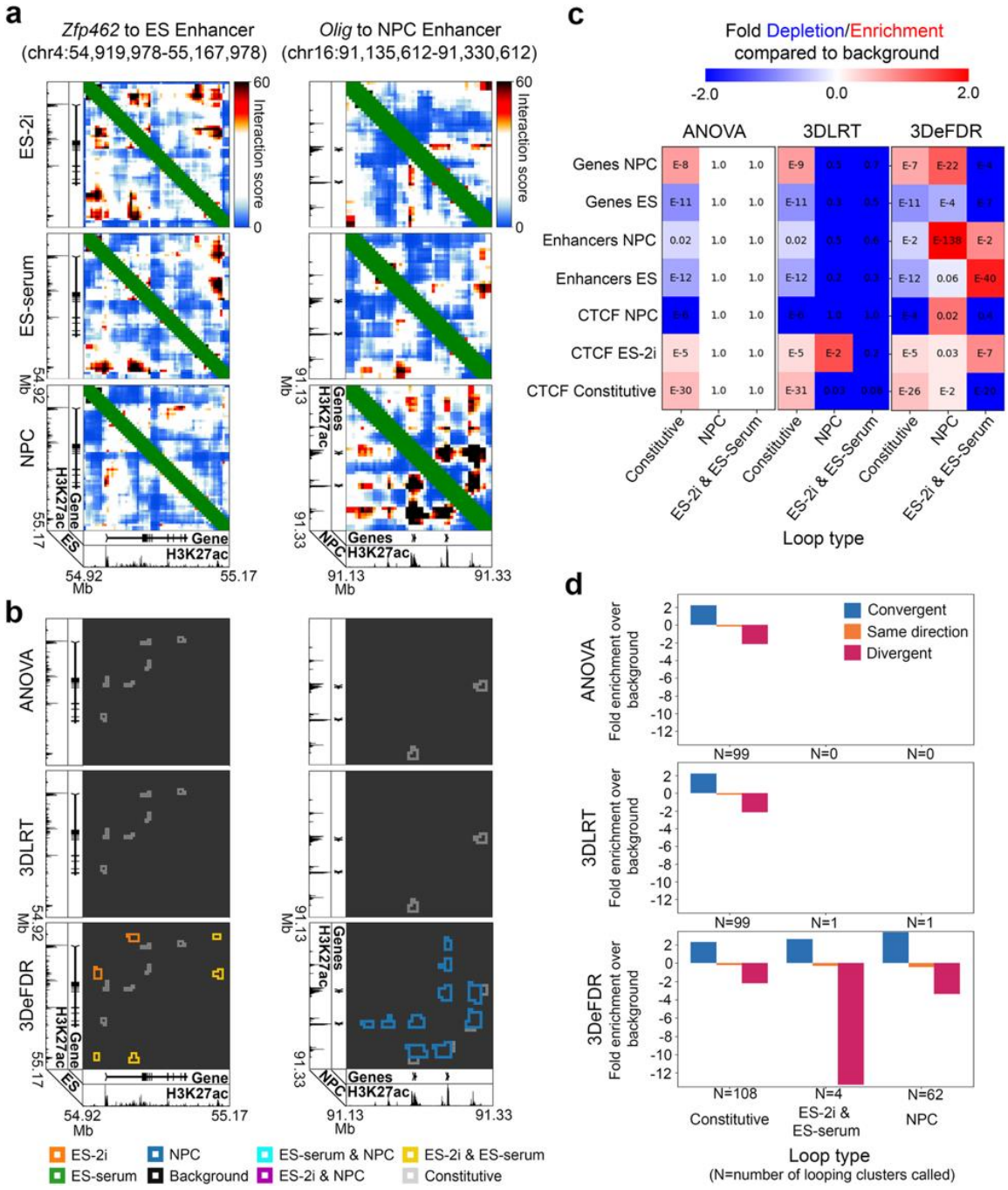


Figure 5.5. Dynamic 3D chromatin looping interactions identified using 3DeFDR-5C, 3DLRT, and ANOVA.

a Reference interaction score heatmaps for two sample loci. **b** Loop classification results achieved with each differential looping detection method at a target false discovery rate (FDR) of 2%. **c** Enrichment of cell-type specific markers in loops classified as NPC or ES-2i & ES-serum for each of the three methods at a target FDR of 2%. **d** Log fold-change in percent CTCF orientation among loops classified as constitutive, ES-2i & ES-serum, or NPC, over percent CTCF orientation among loops classified as background

To evaluate the performance of the 3DeFDR-5C pipeline, we implemented two additional methods for classifying differential looping interactions: ANOVA-BH and 3DLRT-BH (Additional file 2: Fig. S3). These methods use ANOVA and our newly formulated likelihood ratio test (3DLRT), respectively, to assign a differential looping *p* value to every bin-bin pair in an experimental dataset (detailed in the "Methods" section). In both approaches, output *p* values are then corrected for multiple testing using the Benjamini-Hochberg step-up procedure. When we compared ANOVA and 3DLRT benchmarking tests to 3DeFDR-5C, we found that the three different methods had different optimal FDR thresholds for identifying differential loops (Supplementary Figures 4–6, 8–10), with 3DeFDR-5C identifying the known, previously reported looping interactions at significantly lower FDR estimates than the other two approaches (***Fig. 5***). Thus, 3DeFDR-5C can identify known cell type-specific looping interactions with a lower estimated false discovery rate than ANOVA and 3DLRT benchmarking tests under the assumptions of our model.

To further understand the dynamic loops called by 3DeFDR-5C, we also compared them to chromatin modifications on the 1-D genome as well as to the performance of the leading non-specific differential interaction caller built for Hi-C data. We observed that

classes of differential loops identified by 3DeFDR-5C at an FDR of 2% were strongly enriched for genes and enhancers characteristic of cell types matching their differential loop class (***Fig. 5c***, Additional file 2: Figs. S7, S11). Moreover, we observed that convergently and divergently oriented CTCF motifs were over- and under-enriched, respectively, at the base of loops identified by 3DeFDR-5C (***Fig. 5***). Together, these data indicate that 3DeFDR-5C calls differential loops that exhibit the known hallmarks of cell type-specific looping interactions.

Finally, we formulated 3DeFDR-HiC to identify cell type-specific loops genome-wide in Hi-C data. To develop 3DeFDR-HiC, we relied on ultra-high-resolution Hi-C data from mouse ES cells and ES-derived NPCs [119]. We first identified loops genome-wide in each cell type individually (see the "Methods" section, ***Fig. 6a, b***). To identify which of the identified loops were ES- or NPC-specific, we formulated a negative binomial model parameterized by (i) the mean count per pixel across replicates for every biological condition, (ii) a distance-dependent scaling factor to normalize for sequencing depth (Additional file 2: Fig. S14), (iii) bias factors for every row in the raw Hi-C matrix, and (iv) an estimated dispersion per pixel across replicates for every biological condition (see the "Methods" section). We estimated the dispersion of loops at every 10-kb increment of genomic distance via a distance-dispersion-relationship (DDR) (see the "Methods" section, ***Fig. 6***). After fitting the parameters of our model to the data, we performed a likelihood ratio test to obtain *p* values against the null hypothesis that each interaction in a loop was not differential and applied the Benjamini-Hochberg step-up procedure to correct these *p* values for multiple testing. At an FDR threshold of 1% and a loop cluster size threshold of 3 (see the "Methods" section), we identified 818 ES-specific loops and 1435 NPC-specific loops

(***Fig. 6***), including the ES-specific loop connecting the *Sox2* gene to its ES-specific enhancer (box 1), and the longer-range ES-specific, NPC-specific, and constitutive loops around *Sox2* at this locus (box 2, box 3) (***Fig. 6 b, e***). Thus, we can identify cell type-specific looping interactions genome-wide in Hi-C data with 3DeFDR-HiC.

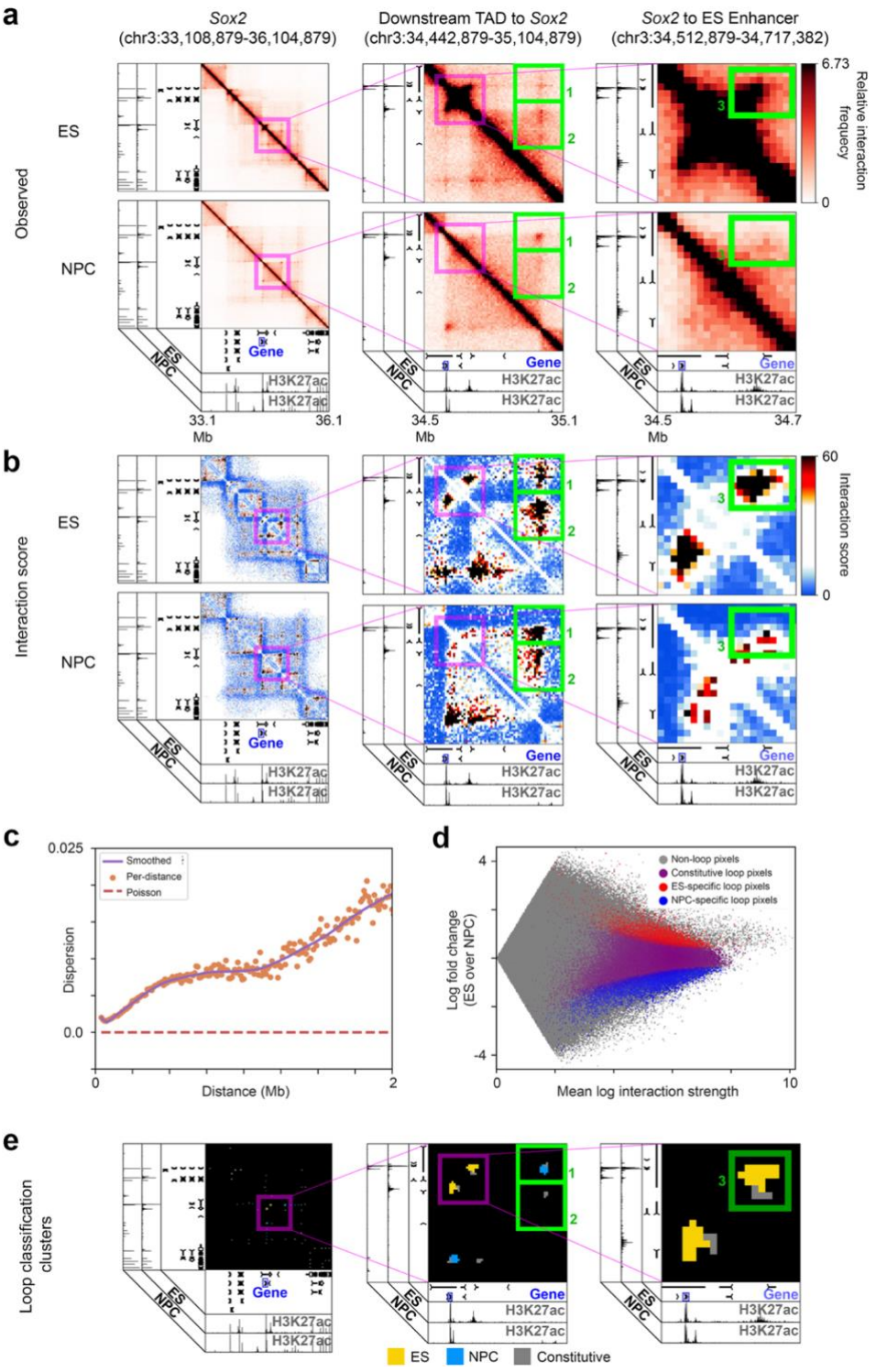


Figure 5.6. Cell-type specific looping interactions identified from Hi-C using 3DeFDR-HiC.

a Reference heatmaps of relative Hi-C interaction frequency (Observed) for the Sox2 region and two zoom-in views of loops involving the Sox2 gene. Boxes 1, 2, and 3 highlight areas of differential looping. **b** Reference interaction score heatmaps of the same genomic regions shown in **a**. **c** Distance-dispersion relationship in the ES condition in the Bonev et al. Hi-C dataset. The orange dots show the estimated negative binomial dispersion parameter at each distance scale. The purple line represents a LOWESS smoothing of the orange points. The red dashed line shows the effective dispersion of the Poisson distribution for comparison. **d** MA plot of the differential loop analysis comparing the ES and NPC conditions in the Bonev et al. Hi-C dataset. The x- and y-axes represent the average log interaction frequency and the log fold change across cell types, respectively, computed on observed Hi-C counts normalized for both locus specific biases and sequencing depth differences. The densities of non-loop, constitutive, and differential (called by our method at an FDR threshold of 1%) pixels are shown in different colors as indicated in the legend. **e** Heatmaps of final loop cluster classifications for each genomic region called by 3DeFDR-HiC at an FDR threshold of 1%

Our 3DeFDR-HiC method makes three critical assumptions: (1) the use of a negative binomial distribution is necessary to account for overdispersion in Hi-C data, (2) the model needs to account for the DDR, and (3) pooling dispersion or variance estimates is necessary to achieve good performance in the face of small numbers of available replicates. To test these three assumptions, we benchmarked the performance of 3DeFDR-HiC on simulated data against three alternative models that each dropped one of our three assumptions. These alternative models included a Poisson model (which assumes mean is equal to variance with no overdispersion), a "global negative binomial" model (which does not account for the DDR), and a "sample variance parameterized negative binomial" model (which does not pool dispersion or variance estimates and uses a sample variance computed for each pixel across replicates instead) (see the "Methods" section). We provide the intuition for how each

of the three models compares to our 3DeFDR-HiC method in Additional file 2: Fig. S13A. Our inspection of distributions of *p* values called on true null simulations revealed that the Poisson model failed to control type I error (Additional file 2: Fig. S13B). This failure to control type I error was also reflected in a failure to control FDR in simulations containing truly differential loops (Additional file 2: Fig. S13C). Next, we assessed the performance of the different approaches using receiver operator characteristic (ROC) curves, revealing that the "sample variance parameterized negative binomial" model resulted in inferior cell type-specific loop classification performance compared to 3DeFDR-HiC, which uses pooled dispersion estimates (Additional file 2: Fig. S13D). Finally, we assessed the bias of low *p* values in simulated null datasets with respect to distance (Additional file 2: Fig. S13E), revealing that the "global negative binomial" model is overly conservative at short distances, where it overestimates dispersion, and overly permissive at long distances, where it underestimates dispersion. Altogether, these results were used to formulate and justify the assumptions upon which we built our 3DeFDR-HiC model.

Finally, to benchmark 3DeFDR-HiC's performance, we applied diffHic [104] to the same Hi-C data. When comparing the two methods, we held constant either the FDR threshold or the total number of significant differential loops. In both the 'matched FDR' and 'matched loop number' benchmarking scenarios, we observed that diffHic called cell type-specific interactions throughout Hi-C data irrespective of whether or not the interactions were bona fide loops (Additional file 2: Fig. S12A,C). We also created simulated Hi-C maps containing pre-defined cell type dynamic looping interactions with a range of interaction strength effect sizes (see the "Methods" section, ***Fig. 7***). 3DeFDR-HiC markedly outperformed diffHic in the sensitivity and specificity of differential loops called on our

simulated datasets (Additional file 2: Fig. S12D). As expected, running 3DeFDR-HiC on simulations with stronger looping fold changes resulted in a higher number of differential loops called (***Fig. 7***). 3DeFDR-HiC exhibits strong sensitivity and specificity of loop detection which increases with increasing interaction frequency effect size (***Fig. 7***), as well as consistently strong FDR control at every tested interaction frequency effect size (***Fig. 7***). Our simulations can be used to perform power calculations at a variety of effect sizes (***Fig. 7***), providing estimates of the proportion of uncalled truly differential loops across a range of differential effect sizes. Together, these data characterize the performance of 3DeFDR-HiC and suggest that it outperforms the leading Hi-C interaction caller diffHiC.

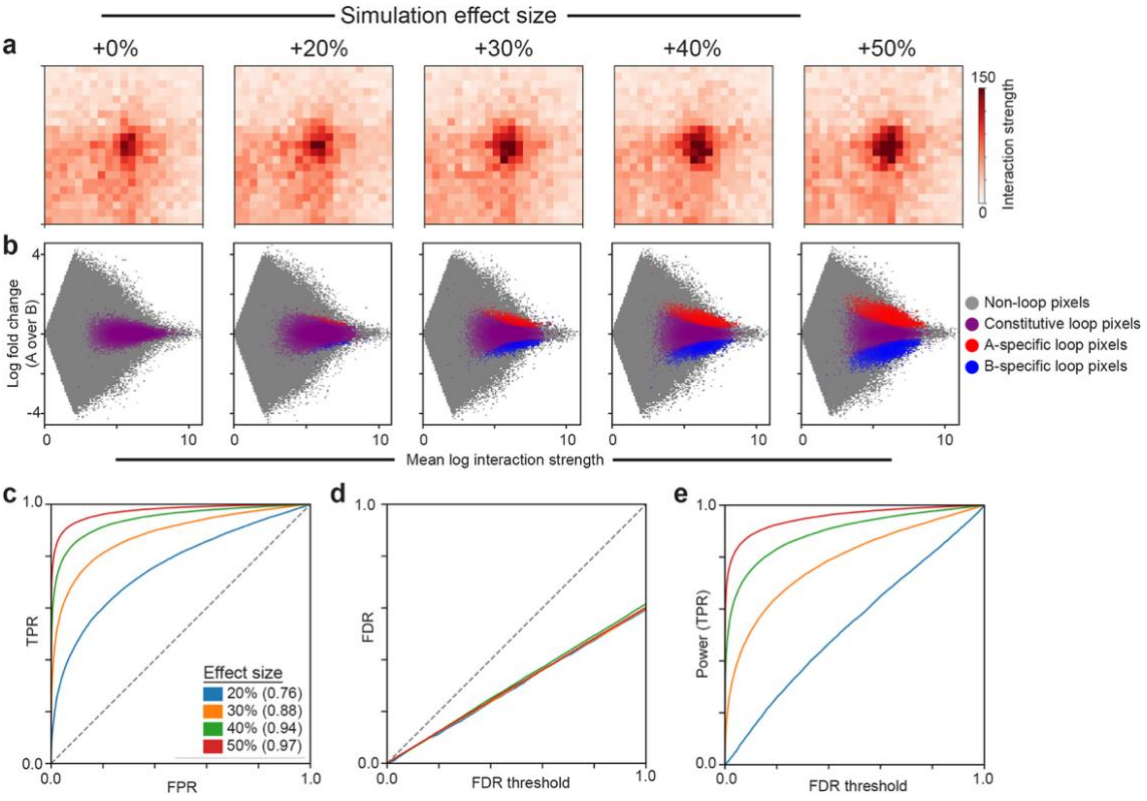


Figure 5.7. Characterization of performance of 3DeFDR-HiC method using simulated Hi-C data.

a Heatmaps showing a single example loop in simulations generated using varying effect sizes. The difference between any heatmap and the baseline loop strength shown in the far-left panel becomes more pronounced as effect size increases. **b** MA plots resulting from analysis of simulations of two artificial conditions ("A" and "B") generated using varying effect sizes, with red and blue points representing interactions called as differential by our method at a false discovery rate of 1%. No interactions are called differential when no loops are truly differential (effect size + 0%). The number of interactions called as differential increases with increasing effect size, though the true proportion of differential interactions remains fixed at 40% in the simulations shown here. **c** Receiver operating characteristic (ROC) curves showing performance of our method on simulations generated using varying effect sizes. Like in **b**, the true proportion of differential interactions remains fixed at 40%. The x-axis shows the false-positive rate (FPR), or one minus the specificity. The y-axis shows the true positive rate (TPR), or sensitivity. The area under the receiver operating characteristic curve (AUROC) for each curve is shown in parentheses in the legend. **d** False discovery rate (FDR) control curves showing FDR control characteristics of our method on simulations generated using varying effect sizes, colored as in (C). The x-axis shows a range of FDR thresholds, while the y-axis shows the actual FDR we observe in the differential calls made by our method at that FDR threshold. Methods that control FDR should stay below the dashed gray line. All FDR control curves should show an FDR of 60% at an FDR threshold of 100%, since only 40% of loops in each simulation are truly differential. **e** Power curves showing the proportion of truly differential interactions called differential by our method (y-axis) as a function of the FDR threshold used for thresholding (x-axis) in simulations generated using varying effect sizes, colored as in **c**

5.3 Discussion

Since the invention of 5C and Hi-C technologies, the field has been in need of statistical methods and computational tools for identifying differential long-range looping interactions among biological conditions. To date, there is a severe lack of differential loop calling methods available for analysis of 5C data by the scientific community. Moreover, although a

small number of "general differential interaction identification" methods have been published for Hi-C data, differential loop calling largely remains an open question because (1) currently available tools do not account for local distance-dependent background signal and TAD/subTAD/compartment structure to identify changes specifically at loops and (2) Hi-C datasets with the resolution necessary for looping interaction analysis have only very recently become available. We describe two variants of our method: 3DeFDR-5C, our original approach designed for identifying cell type-specific loops from 5C data, and 3DeFDR-HiC, a simplified and parallelized variant fast enough to identify differential loops in genome-wide Hi-C datasets.

It is important to acknowledge potential limitations in our methods. 3DeFDR-5C and 3DeFDR-HiC cannot in their current form detect global changes in looping due to a biological perturbation such as nuclear volume change which would lead to global shift in signal at a specific distance scale. We have created our code in a way that allows users to alter bias vectors and scaling parameters to account for their biological question. In cases of global changes, the normalization and correction of samples together would not be preferred. We also acknowledge that our work here represents one of the first in-depth studies of the problem of variance estimation in Hi-C data. To further enhance differential loop calling performance, newer modeling approaches will be needed to improve upon our dispersion estimates in the future. In an ideal scenario, Hi-C data for every condition would be obtained with a high number of biological replicates, thus facilitating the ability to estimate variance on a per-pixel basis and account for the local TAD/subTAD and compartment folding patterns that influence mean and variance estimates at each pixel. Here, we pool interaction frequencies by distance to create a DDR, but future studies may

reveal that dispersion is controlled by additional factors beyond distance and biological condition.

In this study, we analyze Hi-C datasets using a 10-kb bin resolution. In principle, our implementation of 3DeFDR-HiC is fast enough to call differential loops using smaller bin sizes; however, we have chosen to present results using 10 kb bins due to the scarcity of Hi-C datasets with sufficient read depth to reliably detect loops at bin sizes smaller than 10 kb. We expect that assessing the < 10-kb bin matrix resolution performance of 3DeFDR-HiC and other differential loop calling models will become an important area for future work as more ultra-high-resolution Hi-C datasets become available.

Our analyses thus far have suggested that variance estimation is not as critical for differential loop calling genome-wide in "C" data as it is for differential gene expression analyses in RNA-seq data. Our hypothesis for this discrepancy is that RNA-seq data has a much higher dynamic range of counts than "C" data and that the dispersion estimates matter most for modeling very highly expressed genes. Consistent with this idea, we do indeed observe that both of our methods (3DeFDR-5C and 3DeFDR-HiC) allow for more sensitive and specific loop detection in the case of ultra-short-range loops where the interaction frequencies have the highest mean. The advantage is, however, small compared to using per-pixel sample variances or a zero-dispersion Poisson model, and future studies will unravel how improved sensitivity/specificity in loop calling will aid in biological discovery in high-resolution Hi-C data. A systematic comparison of all differential looping models—including a more quantitative performance assessment for 5C differential loop calling—remains an important area for future work.

In conclusion, we provide 3DeFDR as a new statistical framework and computational tool for detecting and classifying differential looping interactions in high-resolution, multi-condition 5C and Hi-C datasets. We note that the performance of 3DeFDR is highly dependent on the quality of the input dataset and how effectively the raw sequencing counts of detected interactions have been processed to reduce batch effects, correct for bias, and account for distance-dependent and TAD/subTAD background signal. We provide 3DeFDR as a modular coding package that the user may integrate into their own 5C or Hi-C analysis pipeline. For the convenience of users, this package includes companion visualization tools for assessing 3DeFDR results to determine how effectively counts have been modeled for simulation, viewing differential loop calls as color-coded clusters, and computing the enrichment of classical epigenetic marks within classes of called loops.

5.4 Methods

5.4.1 5C Data

5C libraries generated with a single alternating primer design [91] in embryonic stem (ES) cells cultured in 2i media (ES-2i), ES cells cultured in serum/LIF (ES-Serum), and primary.

5.4.2 Hi-C data

Hi-C libraries were downloaded from GEO (Additional file 6: Table S5). Briefly, we used all raw Hi-C sequencing reads from the ES_1, ES_3, NPC_1, and NPC_2 replicates (representing the ES and NPC conditions), keeping the replicates separate.

5.4.3 5C data processing pipeline

5.4.3.1 Overview

Raw 5C counts were subjected to our previously published 5C count modeling methods [91] [108] [120] [121] [122]. The processing steps described briefly below ultimately resulted in the conversion of fragment-level, raw count matrices to a bias- and expected background-corrected contact matrices of interaction scores. Pre-processing steps were performed prior to the post-processing steps of matrix balancing, binning, mean-variance relationship modeling, and 5C replicate simulation. Binning and all subsequent normalization and modeling steps were performed on both experimental and simulated 5C replicates.

5.4.3.2 Data structure and pre-processing

We assembled sequencing counts from each 5C experiment t_s and each genomic region r into an $n_r \times n_r$ raw contact matrix, $C_{t_s,r}$, where n_r represents the total number of HindIII restriction fragments in each region r, $t \in$ {ES2i, ESserum, NPC} represents a cellular condition, and $s \in \{1, 2\}$ represents a biological replicate of the cellular condition t. Thus, $C_{t_s,r,i,j}$ is the number of reads that represent contacts between the ith and jth fragments in region r, where $i \in \{1, 2, 3, \ldots, n_r\}$ and $j \in \{1, 2, 3, \ldots, n_r\}$. Raw contact matrices were then normalized as described [91]. Briefly, the raw contact matrices $C_{t_s,r}$ were normalized for replicate biases due to batch effects, sequencing depth differences, and library complexity differences by conditional quantile normalization to create a normalized contact matrix $C'_{t_s,r}$.

5.4.3.3 Matrix balancing

Each normalized contact matrix $C'_{t_s,r}$ was then matrix balanced with joint express as described [91] [108] to correct for differences in fragment-specific biases, such as GC content, fragment length, and 5C primer-specific efficiency at each primer in region *r* to create a balanced contact matrix $C'_{t_s,r}$.

5.4.3.4 Contact matrix binning

Balanced contact matrices $C'_{t_s,r}$ were converted to binned interaction frequency matrices by binning at regular 4-kb intervals and smoothing at 16-kb intervals as described in [91] [108] [120]. The smoothing was performed because we developed the 3DeFDR-5C method on older 5C data from an alternating 5C primer design. 5C libraries made with double alternating designs do not require this smoothing step [121]. The resulting binned interaction frequency matrices $B_{t_s,r}$ have m_r by m_r elements where m_r is the total number of bins in region *r*. $B_{t_s,r,k,l}$ represents the arithmetic mean contact frequency between fragments in the *k*th and *l*th bins in genomic region *r* as recorded in replicate *s* under condition *t*. Binned interaction frequency matrices have reduced spatial noise relative to the original fragment-level matrices while preserving the underlying signal.

5.4.3.5 Distance dependence normalization

Following binning, expected values for each interaction in the binned interaction frequency matrices were computed using a modification of the local donut expected described by Aiden and colleagues that accounts for the local TAD/subTAD structure and the global distance-dependent background signal [72] [91] [108]. The binned interaction frequency values $B_{t_s,r,k,l}$ (Observed) were corrected by the maximum of expected donut

values $DE_{t_s,r,k,l}$ and expected lower left values $LLE_{t_s,r,k,l}$ to yield contact enrichments (Observed/Expected, or Obs/Exp) normalized for distance-dependent 5C count signal and local chromatin domain structure as detailed previously [91].

5.4.3.6 Probabilistic model fitting

As detailed previously [91], contact enrichment values (Obs/Exp) were modeled within each region by parameterizing a log-logistic distribution using maximum likelihood estimation, resulting in matrices of right-tailed *p* values $P_{t_s,r}$. *P* values were computed for each 5C genomic region separately.

5.4.3.7 Removal of interactions below distance limit

Interactions occurring between bins within 20 kb of each other on the linear chromatin fiber were removed from consideration and not included in further processing.

5.4.3.8 Interaction scores and z-scores

The final step of the post-processing pipeline is the conversion of modeled *p* values to interaction scores. We use $IS_{t_s,r}$ to refer to the matrix of interaction scores for region *r* and replicate *s* in condition *t*. For 3DeFDR-5C, *p* values were transformed to an interaction score of $-10 \times \log_2(p \text{ value})$. For benchmarking approaches, ANOVA, and 3DLRT (detailed below), *p* values were transformed to both an interaction score of $-10 \times \log_2(\text{pvalue})$ as well as a *z*-score computed using the standard normal quantile function (the inverse of the standard normal cumulative distribution function) (Equations 8 and 9):

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx \quad \textbf{(8)}$$

$$Z_{t_s,r,k,l} = \Phi^{-1}\left(1 - P_{t_s,r,k,l}\right) \quad \textbf{(9)}$$

where $P_{t_s,r,k,l}$ is the right-tail p value computed for the interaction between bins k and l in genomic region r as recorded in biological replicate s under condition t. We implemented the conversion of p values to z-scores using the stats.norm.isf function in the scipy Python library.

5.4.4 Hi-C data processing pipeline

Raw Hi-C data were aligned to the mm9 genome using bowtie2 (global parameters: --very-sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end --reorder; local parameters: --very-sensitive -L 20 --score-min L,-0.6,-0.2 --end-to-end --reorder) through the HiC-Pro software. Unmapped reads, non-uniquely mapped reads, and PCR duplicates were filtered out, and uniquely aligned reads were paired. *Cis* contact matrices were assembled by binning paired reads into uniform 10 kb bins.

5.4.5 3DeFDR-5C

5.4.5.1 Overview

3DeFDR-5C is designed to identify differential looping interactions across a set of 5C experiments containing either two or three cellular conditions with at least two replicates each. In this section, we describe the application of 3DeFDR-5C to three cellular conditions, referring to a set of three conditions $T = \{A, B, C\}$ and of six replicates as $S = \{A1, A2, B1, B2, C1, C2\}$.

5.4.5.2 Differential loop categories

In the 3DeFDR-5C framework, the set of possible classes of differential looping interactions is defined as all nonempty proper subsets H of the input condition set T (Equation 10):

$$H=\{\{A\},\{B\},\{C\},\{A,B\},\{A,C\},\{B,C\}\} \quad \textbf{(10)}$$

Interactions assigned to single-condition classes, e.g., $\{A\}$, $\{B\}$, or $\{C\}$, are considered to be interacting significantly higher in replicates of that specific condition than in those of the other two conditions (Additional file 2: Fig. S1E and Additional file 3: Table S2). Interactions assigned to dual condition classes, e.g., $\{A, B\}$, $\{A, C\}$, or $\{B, C\}$, are considered to be interacting significantly higher in replicates of the two specific conditions than in the remaining single condition (i.e., *C*, *B*, and *A*, respectively). If interaction scores for an interaction are sufficiently high in all conditions, that interaction is interpreted to be non-differential and labeled as a constitutive looping interaction. If interaction scores are sufficiently low in all replicates of all conditions, the interaction is not called a looping interaction; therefore, it is not tested for differential looping signal (Additional file 2: Fig. S1D). Points with very low interaction scores in all replicates are assigned to a background class, representing interactions that are very unlikely to be loops (Additional file 2: Fig. S1C).

5.4.5.3 Computing empirical false discovery rate

3DeFDR-5C controls an empirically estimated false discovery rate (eFDR) to classify loops as differentially interacting across cellular condition set *T*. By definition, $FDR = E\left[\frac{V}{R}\right]$ where *V* is the number of false positives among tests declared significant and *R* is the total number of tests declared significant. *R* is computed as the total number of pixels called as significantly differential in any differential class in *H* (Equation 10). By contrast, *V* is not trivially computed and requires a model for estimating what proportion of looping interactions in each class in *H* are false positives.

We hypothesized that V is approximately equal to the total number of interactions incorrectly labeled as differential when applying 3DeFDR-5C (holding all of its thresholds fixed) to a set of biological samples known to have no truly differential loops (i.e., a null biological sample set). We defined our null data set as a set of samples that are all replicates of a single cellular condition but are assigned a set of labels matching the different conditions in T. The key assumption of this approach is that that the false-positive rate (FPR) of calls on the null dataset (FPR_{null}) is approximately equivalent to that of the real experimental dataset (FPR_{exp}), such that $FPR_{null} \approx FPR_{exp}$. We computed and controlled an empirical false discovery rate (eFDR) (Equation 11):

$$eFDR = \frac{n_{null}}{n_{exp}} \approx \frac{V}{R} \quad \textbf{(11)}$$

where n_{exp} is the total number of interactions classified as significantly differential in the real experimental dataset and n_{null} is the total number of interactions classified as significantly differential in the null dataset.

We computed a piecewise interaction score thresholding scheme for each looping interaction class in the set of possible differential classifications H. To classify dynamic loops, 3DeFDR-5C applies a thresholding scheme based on the difference in interaction scores between conditions (***Fig. 1***). Using a sweep of IS difference thresholds d (see orange lines in ***Fig. 1***), 3DeFDR-5C assigns every pixel in a 5C data set either to one of the differential classes in H or to the background, constitutive, or "other" class (as described above and below) and computes a class-specific eFDR for each differential looping interaction class $h \in H$ as (Equation 12):

$$eFDR_{d,h} \approx \frac{n_{null}^{d,h}}{n_{exp}^{d,h}} \quad \textbf{(12)}$$

where $n_{exp}^{d,h}$ is the total number of interactions assigned to differential looping class *h* in the real experimental dataset at difference threshold *d* and $n_{null}^{d,h}$ is the total number of interactions assigned to differential looping class *h* in the simulated null dataset at the same difference threshold *d*. 3DeFDR-5C adapts the distance threshold for each differential looping class to maintain a user-specified target empirical FDR threshold τ across all differential looping classes. For each looping interaction class *h*, we determined the distance threshold *d* at which $eFDR_{d,h}$ is closest to τ while remaining less than τ. Thus, each differential looping class *h* will have a unique difference threshold *d* to reach the study-specific target eFDR threshold τ. Constitutive looping pixels are identified as those that are strong in all conditions and are not sufficiently differential to admit assignment to one of the differential classes (Additional File 3: Table S2). "Other" or "uncalled" pixels include those that pass the looping threshold but do not meet the requirements of any of the other classes (Additional File 3: Table S2). Overall, 3DeFDR-5C employs eFDR estimate control to guide the placement of IS thresholds to call differential looping classes.

We did not have access to an experimental dataset with enough replicates of the same cellular condition to create a null replicate set directly from real 5C libraries. To avoid the high costs and labor required to run additional experiments, we modeled and created simulations of our existing experimental replicates to create additional simulated replicates. We constructed a null dataset from six simulated replicates ($S_{null} = \{A1_{sim}, A2_{sim}, A3_{sim}, A4_{sim}, A5_{sim}, A6_{sim}\}$) all based on the same biological condition ($T_{null} = \{A, A, A\}$)).

5.4.5.4 Modeling and simulation of preprocessed replicates

We simulated 5C replicates of the same condition at the level of fragment-level counts after conditional quantile normalization. Our rationale for simulating quantile normalized counts rather than raw counts was that doing so would allow us to omit library complexity, batch effect, and sequencing depth terms in our count-generating models. We simulated fragment-resolution counts that have been quantile normalized but not balanced (resulting in simulated matrices comparable to $C'_{t_s,r}$) by parameterizing a different negative binomial distribution for each interaction as described below and then drawing a random variable from this distribution.

To begin constructing our simulation-generating model, we computed the sample mean and sample variance of the preprocessed sample counts of a single interaction across replicates of the same condition as in Equations 13 and 14:

$$\mu_{t,r,i,j} = \frac{\sum_{s=1}^{n_t} C'_{t_s,r,i,j}}{n_t} \quad \textbf{(13)}$$

$$\sigma^2_{t,r,i,j} = \frac{\sum_{s=1}^{n_t} \left(C'_{t_s,r,i,j} - \mu_{t_s,r,i,j}\right)^2}{n_t - 1} \quad \textbf{(14)}$$

where n_t is the number of replicates of condition t and $C'_{t_s,r,i,j}$ is the conditional quantile normalized 5C count value for the interaction between the i^{th} and j^{th} bins of region r in the s^{th} replicate of condition t.

Most genomics experiments suffer from poor parameter estimation due to the low number of replicates that are financially and logistically feasible to generate for every biological condition. We did not use $\mu_{t,r,i,j}$ and $\sigma^2_{t,r,i,j}$, computed from only $n_t = 2$ replicates, to directly parameterize the negative binomial (NB) counts models for each

quantile normalized interaction count $C'_{t_s,r,i,j}$. Instead, we modeled the mean-variance relationship (MVR) between $\mu_{t,r,i,j}$ and $\sigma^2_{t,r,i,j}$, thereby leveraging the high-dimensional nature of our data set to improve our variance estimates. We stratified quantile normalized counts, $C'_{t_s,r,i,j}$ for all regions by their linear genomic distance using overlapping stratification windows of different sizes depending on genomic distance. For distance regime 1 (0–150 kb), we stratified the interactions using fine-grained, 12-kb-sized sliding windows with a 4-kb step. For distance regime 2 (151–600 kb), we stratified the interactions into 24-kb-sized sliding windows with an 8-kb step. For distance regime 3 (601–1000 kb), we stratified the interactions into coarse-grained, 60-kb-sized sliding windows with a 24-kb step. For each window *w* in each distance regime, we modeled the MVR for each condition *t* by fitting the function $\sigma^2 = A_{t,w}\mu^2 + \mu$ to the $\mu_{t,r,i,j}$ and $\sigma^2_{t,r,i,j}$ values for all regions *r* and for all bin-bin pairs *i, j* whose linear genomic separation distance fell in window *w*. Prior to estimation of $A_{t,w}$, interactions with mean counts of one or less, or more than 2.5 standard deviations above the mean of mean counts for interactions in bin *w* were removed. The dispersion parameters, $A_{t,w}$, were then plotted as a function of genomic distance, and LOWESS smoothing with a smoothing fraction of 0.5 was used to compute the final dispersion estimates, $\bar{A}_{t,w}$. The predicted sample variance $\hat{\sigma}^2_{t,r,i,j}$ for each individual interaction was then computed using the LOWESS-smoothed dispersion estimate $\bar{A}_{t,w}$ appropriate for the window *w* corresponding to the interaction distance $|i-j|$ as in Equation 15:

$$\hat{\sigma}^2_{t,r,i,j} = \bar{A}_{t,w}\mu^2_{t,r,i,j} + \mu_{t,r,i,j} \quad \textbf{(15)}$$

We weighted the predicted variance value $\hat{\sigma}^2{}_{t,r,i,j}$ against the original observed variance of the interaction $\sigma^2{}_{t,r,i,j}$ to generate a final weighted variance $\bar{\sigma}^2_{t,r,i,j}$ for each interaction as in Equation 16:

$$\bar{\sigma}^2_{t,r,i,j} = \alpha\hat{\sigma}^2{}_{t,r,i,j} + \beta\sigma^2{}_{t,r,i,j} \quad \textbf{(16)}$$

We chose to use $a = \beta = 0.5$ to achieve pairwise correlations on par with that of real replicates while improving the quality of our variance estimate with the predicted contribution. As shown in Additional file 4: Table S3, increasing a led to higher pairwise correlation between simulated replicates. Finally, we parameterized a negative binomial distribution for each $C'_{t,r,i,j}$ interaction and generated simulated counts $C'^{sim}_{t,r,i,j}$ from it as in Equation 17:

$$C'^{sim}_{t,r,i,j} \sim NB\left(\mu_{t,r,i,j}, \bar{\sigma}^2_{t,r,i,j}\right) \quad \textbf{(17)}$$

5.4.5.5 Creating the null replicate set

Using the generative models described above, we created six simulated replicates of a chosen biological condition $t \in T$. For our results, we chose to use the NPC condition (which we will denote as condition $A \in T$), because this was the condition which showed the highest dispersion between replicates (***Fig. 2***) and would therefore result in the most conservative eFDR estimate. The simulated replicates that made up our null replicate set are shown in Equation 18:

$$S_{sim} = \{A1_{sim}, \ldots A6_{sim}, B1_{sim}, \ldots B6_{sim}, C1_{sim}, \ldots C6_{sim}\} \quad \textbf{(18)}$$

Users of 3DeFDR-5C may choose whichever condition they like when generating the null replicate set.

The simulated interaction counts $C'^{sim}_{t_s,r,i,j}$ were then subjected to the matrix balancing, binning, modeling, and *p* value transformation steps described above. 3DeFDR-5C takes as input the simulated replicate interaction scores $IS^{sim}_{t_s,r}$ and experimental replicate interaction scores $IS_{t_s,r}$ for each region *r*, replicate *s*, and condition *t*.

5.4.5.6 Identification of the background interaction set

Prior to the identification of differential looping interactions, we created a background null interaction set consisting of all interactions for which the interaction scores $IS_{t_s,r,k,l}$ of all replicates of every condition were less than a background threshold *b* as in Equation 19:

$$Background\ loops\ = \left\{(r,k,l): \max_{t_s\ \in\ S}\left(IS_{t_s,r,k,l}\right) <\ b\right\} \quad \textbf{(19)}$$

The exact threshold for background interactions that we used was $b = -10 \times \log_2(0.8)$, corresponding to a *p* value threshold of 0.8. Interactions not placed in this set were then passed on for further analysis for differential looping in the 3DeFDR-5C pipeline.

5.4.5.7 Preliminary classification of differential looping interactions

As outlined in Additional file 2: Fig. S1, to ultimately be classified as differential, a loop must pass thresholds for both baseline significance and IS difference across conditions.

5.4.5.8 Baseline significance filtering

To meet the criteria for differential looping for any differential classification *h*, an interaction must have $IS_{t_s,r,k,l}$ greater than a specific significance threshold *g* for all replicates in at least one condition in *T* as in Additional file 2: Fig. S1D and Equation 20:

$$Significant\ Loops\ = \left\{(r,k,l): \max_{t}\left[\min_{s}\left(IS_{t_s,r,k,l}\right)\right] >\ g\right\} \quad \textbf{(20)}$$

The threshold for a significant looping interaction used in the results presented in the main figures was $g = -10 \times \log_2(0.165)$, corresponding to a p value threshold of 0.165.

5.4.5.9 Thresholding interaction score differences across conditions

Starting with the subset of significant loops across conditions, we then set out to classify interactions according to how much their interaction scores changed across cellular conditions (Additional file 2: Fig. S1E and Additional file 3: Table S2). For each interaction (r, k, l) in the set of significant loops (Equation 20), we computed the difference in interaction score between each possible pair of replicates belonging to different conditions. We then computed initial looping class assignments (Equation 3) across a sweep of IS difference thresholds d as shown in Additional file 3: Table S2. Additionally, in Additional file 3: Table S2, we provide the exact set of thresholds applied to obtain each possible looping classification of a bin-bin pair in dataset capturing three conditions.

In 3DeFDR-5C, loop classifications are determined using this thresholding approach for each difference threshold across a sweep of all possible difference thresholds in a given data set. These classifications are considered preliminary prior to the application of the eFDR control procedure described in the next section.

5.4.5.10 Final loop classification via an adaptive eFDR control procedure

After obtaining preliminary classifications of each interaction across a sweep of IS difference thresholds, we determined each $IS_{r,k,l}$ interaction's final classification via the application of a classification-specific eFDR control procedure. For each possible loop classification $h \in H$, we computed its eFDR for every tested difference threshold d, acquiring a difference threshold-to-eFDR mapping for each class, $eFDR_{d,h}$, as in Equation 5. We next applied the

eFDR threshold τ to this mapping, identifying the difference threshold d at which $eFDR_{d,h}$ is closest to but still less than τ, and report loop calls of class h at this distance threshold. We perform the eFDR controlling procedure for every differential looping class $h \in H$, and the combined set of loop calls for each class constitutes our final set of differential classified loops.

Additionally, eFDR estimates can be computed as an average over a user-specified number, $N_{null-sets}$, of null replicate sets as in Equation 21:

$$eFDR_{d,h} = \frac{\frac{1}{N_{null-sets}} \sum_{m=1}^{N_{null-sets}} card\left(\left\{(r,k,l) \in h_{null}^{d}\right\}\right)}{card\left(\left\{(r,k,l) \in h_{exp}^{d}\right\}\right)} \quad \textbf{(21)}$$

The numerator is now the average number of loops called as class h in the null data sets at difference threshold d. The approach in Equation 21 can reduce variability in eFDR estimates due to random differences between different simulation sets generated from the same counts-generating model.

6 Epigenomic States of Neuronal Activity Response

Adapted from [123]

Beyond study of lineage commitment, I also developed a version of 3DeFDR to identify chromatin loops that dynamically change in response to neural activity. I include this chapter as a further demonstration of the performance of the tool and as a secondary example of how such methods can be used to study epigenetic state changes that contribute to cellular decision making. These observations could be used to develop mechanistic model of epigenetic programming and incorporated into a multiscale model of such decision making, allowing the creation of a more realistic model.

Neuronal activation induces rapid transcription of immediate early genes (IEGs) and longer-term chromatin remodeling around secondary response genes (SRGs). Here, we use high-resolution chromosome-conformation-capture carbon-copy sequencing (5C-seq) to elucidate the extent to which long-range chromatin loops are altered during short- and long-term changes in neural activity. We find that more than 10% of loops surrounding select IEGs, SRGs, and synaptic genes are induced de novo during cortical neuron activation. IEGs *Fos* and *Arc* connect to activity-dependent enhancers via singular short-range loops that form within 20 min after stimulation, prior to peak messenger RNA levels. By contrast, the SRG *Bdnf* engages in both pre-existing and activity-inducible loops that form within 1–6 h. We also show that common single-nucleotide variants that are associated with autism and schizophrenia are colocalized with distinct classes of activity-dependent, looped enhancers. Our data link architectural complexity to transcriptional kinetics and reveal the

rapid timescale by which higher-order chromatin architecture reconfigures during neuronal stimulation.

6.1 Introduction

Neurons have the remarkable ability to receive, transmit, and store information via a dynamic synaptic network. Experience-dependent neuronal activity regulates synaptic features such as dendritic outgrowth, maturation, elimination, and synaptic plasticity [124]. Neural activity governs synaptic structure and function via the upregulation of hundreds of activity response genes [125]. IEGs such as *Fos* (also known as c-*fos*) [126] [127] [128] and *Arc* (also known as *Arg3.1*) (refs. [129] [130]) are expressed within minutes after neuronal stimulation in a protein synthesis-independent manner, whereas SRGs are induced on the order of hours and require de novo protein synthesis [131] [132]. Enhancers—for example, synaptic activity responsive elements—have been identified using epigenetic signatures characteristic of noncoding regulatory activity and verified using reporter transgenes [133] [134] [135] [136] [137]. However, the precise genomic elements that are functionally linked to temporal expression patterns of each specific IEG and SRG remain elusive, in part because synaptic activity responsive elements are distributed across the genome in introns and noncoding regions and their specific target genes are generally unknown.

Chromosome conformation capture (3C) techniques have been used recently to demonstrate that the mammalian genome folds into a hierarchy of structurally and functionally distinct architectural features, including chromosome territories [138], A and B compartments [67] [72], topologically associating domains (TADs) [139] [74], nested

subTADs [72], and long-range looping interactions [72]. The highest resolution maps so far have enabled detection of tens of thousands of loops genome-wide across multiple mammalian cell types [72] [80]. However, little is known about three-dimensional (3D) genome dynamics during paradigms of synaptic plasticity, partly owing to the paucity of high-resolution architecture maps at key time points during neural circuit activation. Knockout of CCCTC-binding factor (CTCF), the primary architectural protein responsible for connecting loops, results in intellectual disability [140] [141] and severe synaptic and long-term potentiation defects in vivo [142]. Moreover, a recent study demonstrated that in vivo cohesin knockout in granule neurons disrupts the tactile startle response, which suggests that specific loops that are connected by cohesin may be required for learning [143]. Given the clear importance of chromatin architecture in brain function, there is a great need for studies that investigate how activity-dependent enhancers are temporally connected via long-range loops to regulate gene expression during a wide range of neuronal activity paradigms.

Here, we investigate the extent to which loops are altered during short- and long-term changes in neural activity, and to analyze the dynamic interplay between the 3D genome and the linear epigenome during the activity-dependent transcriptional response. We create high-resolution genome folding maps across more than 12 megabases (Mb) around *Arc*, *Bdnf*, *Fos*, *Nrxn1*, *Syt1*, and *Nlgn3* using 5C-seq [91] [120] and a double alternating primer design [121]. The 5C-seq approach enables us to create high-complexity, fine-scale architecture maps to explore genome folding dynamics without bias toward a particular chromatin feature across seven acute or chronic time points of neural activity inhibition and activation. We demonstrate that activity-inducible enhancers engage in either

pre-existing or de novo loops connected to genes that exhibit 1.3- and 24-fold activity-dependent increases in expression, respectively. We observe that IEGs *Fos* and *Arc* connect to activity-dependent enhancers via singular short-range loops that form within 20 min after stimulation, whereas the SRG *Bdnf* engages in both pre-existing and activity-inducible loops that form within 1–6 h. Genome-wide analyses confirm a model in which IEGs form fewer, shorter loops before maximum mRNA levels are reached, than the slower, more complex looping architectures formed by SRGs. We also identify a subclass of pre-existing loops that are anchored by enhancers decommissioned upon chronic, 24 h of neural activation. Unexpectedly, we find that common single-nucleotide variants (SNVs) linked to schizophrenia colocalize preferentially at genomic anchors of pre-existing loops connecting activity-decommissioned enhancers to activity-downregulated genes. By contrast, autism-associated SNVs preferentially colocalize with loop anchors that connect activity-inducible enhancers to upregulated genes. Together, our data link 3D genome architectural complexity to transcriptional kinetics and uncover distinct architectural motifs associated with neuropsychiatric disorders.

6.2 Results

3D genome maps of dynamic loops during cortical neuron inhibition and activation

We first created high-resolution maps of higher-order chromatin architecture after 24 h of pharmacologically induced low or high activity in primary neurons. We used an established in vitro model system in which murine cortical neurons were cultured for 15 d in vitro and then treated for 24 h with either 10 μM bicuculline (Bic) [144], which increases neuronal firing by blocking GABA (γ-amino butyric acid)-mediated inhibition, or 1 μM tetrodotoxin

(TTX) [145], a sodium channel blocker that inhibits neuronal firing (***Fig. 1a*** and Extended Data Fig. 1a–c). Chronic pharmacological induction of activity results in multiple forms of synaptic plasticity, including homeostatic changes in AMPA-type glutamate neurotransmitter receptor levels at synapses [146]. Our model system enabled us to interrogate the transcriptional, epigenomic, and architectural features of the mammalian genome in non-dividing, terminally differentiated cortical neurons across inactive (TTX-mediated activity inhibition), moderately active (Untreated), and highly active (Bic-mediated increased activity) states.

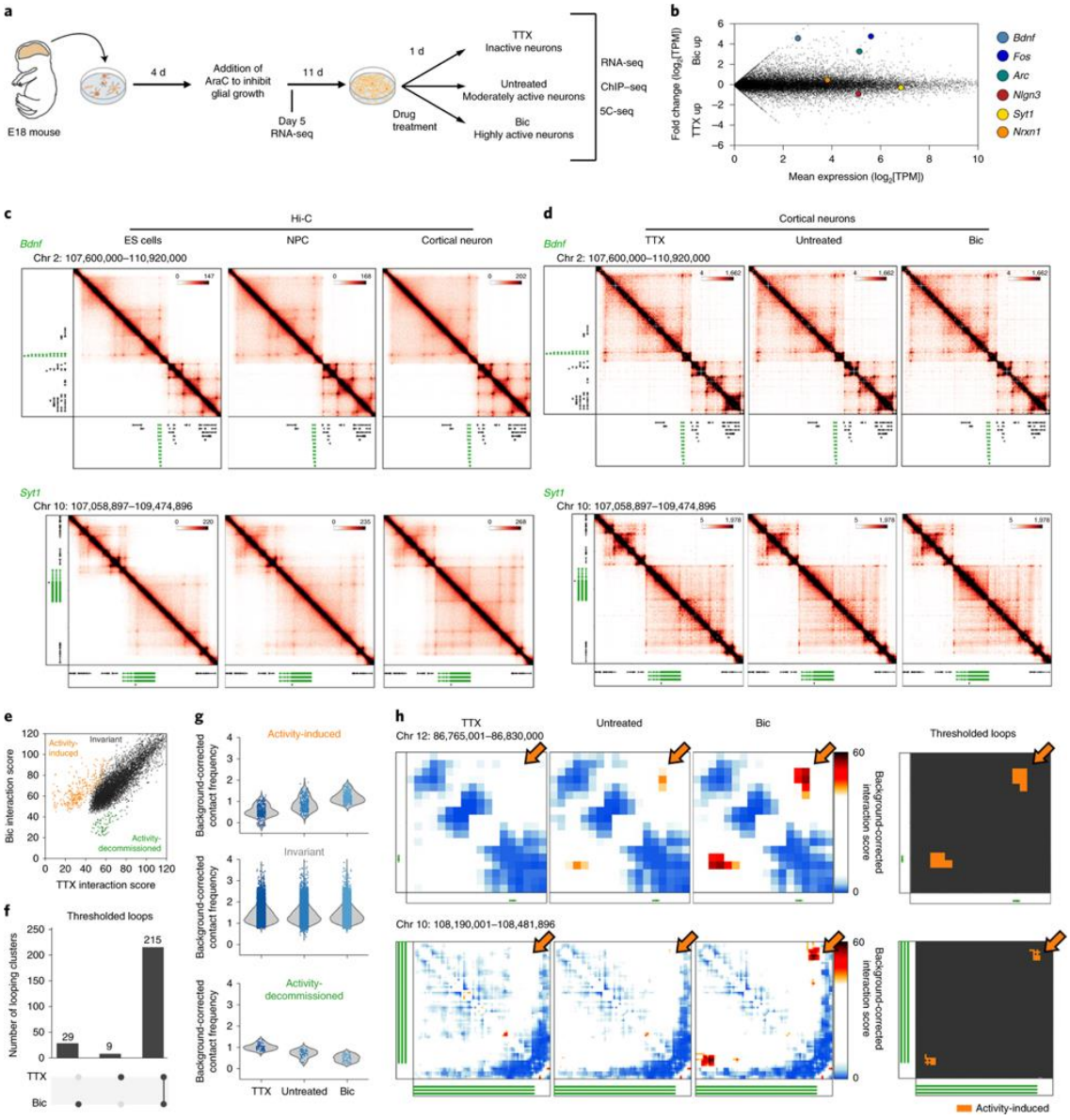


Figure 6.1. Identification of dynamic and invariant looping interactions across neuronal activity states.

a, Primary cultured cortical neuron preparation used to interrogate 3D genome changes during low, basal or high neuronal activity states. **b**, RNA-seq data in Bic and TTX conditions with selected genes highlighted by colored dots. **c**, Interaction frequency heatmaps of 1–3-Mb regions surrounding Bdnf and Syt1 genes (labeled in green) across ES cells, NPCs, and cortical neurons (data analyzed from ref. *[119]*). **d**, Interaction frequency

heatmaps of the regions presented in **c** across TTX-treated, untreated, and Bic-treated DIV16 cortical neurons. **e**, Scatterplot of the interaction scores of thresholded pixels in TTX and Bic conditions. **f**, Activity-inhibited (TTX-only), activity-induced (Bic-only), and activity-invariant (constitutive) loops after thresholding (Supplementary Methods). **g**, Background-corrected contact frequencies across the TTX, Untreated and Bic conditions for each looping class overlaid on kernel density estimate violin plots. n = 340 activity-induced interaction pixels, 7,992 constitutive interaction pixels and 81 activity-decommissioned interaction pixels as represented in **e**. **h**, Interaction score heatmaps and thresholded loops demonstrating activity-induced (Bic-only) loops created by Fos (top) and the Syt1 TSS (bottom).

We used 5C-seq and a double alternating primer design [121] to create high-resolution maps of genome folding in 12.2 Mb surrounding the IEGs *Arc* and *Fos*, the SRG *Bdnf*, the synaptic scaffold genes *Nrxn1* and *Nlgn3*, and the synaptic vesicle gene *Syt1* for a total of 157 unique transcripts (***Fig. 1***, Extended Data ***Fig. 1d,e***, and Supplementary Table 1). Our genome-wide RNA-seq data confirmed that *Arc*, *Fos*, and *Bdnf* were upregulated approximately 10- to 100-fold in Bic versus TTX conditions, whereas *Nrxn1, Nlgn3*, and *Syt1* were unchanged (***Fig. 1b*** and Supplementary Tables 2–4). As expected, under the Untreated (basal activity) condition we observed an intermediate level of *Arc*, *Fos*, and *Bdnf* expression between Bic (high activity) and TTX (inactive) conditions (Extended Data Fig. 1b,c). To confirm data quality, we compared the highest resolution Hi-C maps published so far in mouse embryonic stem (ES) cells, neural progenitor cells (NPCs), and in vitro differentiated cortical neurons (***Fig. 1c***, Extended Data Fig. 1d, and Supplementary Table 5) to our 5C maps (***Fig. 1d***, Extended Data ***Fig. 1e***, and Extended Data Fig. 2). 5C maps from our mature primary cortical neurons were highly correlated with and exhibited similar loops as published Hi-C maps from ES cell-derived cortical neurons (Extended Data Fig. 2c–e). We confirmed

high reproducibility of loops across four 5C replicates taken across two independent batches of neuronal cultures (Supplementary Table 6, Extended Data Fig. 3a,b, and Extended Data Fig. 4). Thus, we have created high-complexity, ultra-high-resolution maps of genome folding across three neuronal activity states.

We next set out to quantify the extent to which loops are altered across different activity states. We normalized the intrinsic biases in 5C data, binned maps to 4-kb matrix resolution, and applied our previously published modeling approaches to identify loops with statistically significant interaction frequency above the local distance-dependence and TAD or subTAD background [91] [120] [108] (Extended Data Fig. 5a and Supplementary Methods). We formulated a statistical method, 3DeFDR (ref. [61]), to stratify loops into invariant and activity-state-specific classes by using differences in interaction frequency across inactive and highly active neurons as thresholds (***Fig. 1e***, Supplementary Table 6, and Supplementary Methods), resulting in the sensitive detection of 215 activity-invariant, 29 activity-induced, and 9 activity-decommissioned loops within the 12.2 Mb of the genome queried (***Fig. 1f*** and Extended Data Fig. 5b). We observed that activity-invariant loops exhibited high interaction frequencies across Untreated, TTX, and Bic conditions (***Fig. 1g***). Importantly, activity-induced and activity-decommissioned loops showed two- to threefold upregulations or downregulations of interaction frequency, respectively, but were still lower in overall looping strength than the activity-invariant contacts (***Fig. 1g***). We confirmed that an enhancer–promoter loop that has been reported previously as activity-dependent at *Fos* via 3C-PCR (ref. [147]) was classified here as an activity-induced loop (***Fig. 1h***, top) and that additional activity-induced loops occurred across our 5C regions (***Fig. 1h***, bottom).

These data highlight that both activity-invariant and activity-dynamic loops encompass IEGs, SRGs, and synaptic genes.

Activity-dependent levels of gene expression are predicted by looping and enhancer acetylation

We quantified the relationship between activity-dependent changes in loop strength, enhancer acetylation, and gene expression. As the histone mark H3 lysine 27 acetylation (H3K27ac) correlates with enhancer and promoter activity, we conducted chromatin immunoprecipitation followed by sequencing (ChIP–seq) of H3K27ac to identify changes in putative noncoding enhancer elements genome-wide in neural activity states (Supplementary Methods and Supplementary Tables 7–10). We noticed a strong correlation between activity-dependent changes in promoter H3K27ac signal and gene expression (***Fig. 2a***), whereas the total sum interaction frequency made by each gene showed no correlation with gene expression (***Fig. 2b***). Instead of total interaction frequency, we next used only bona fide thresholded loops (***Fig. 1f***). We applied an adapted activity-by-contact (ABC) model [148] to identify the single loop or enhancer for each gene that displayed the maximum value of loop strength × enhancer H3K27ac signal (***Fig. 2c*** and Supplementary Methods). Importantly, at this subset of loops we observed a strong increase in interaction strength at the most strongly activity-upregulated genes (***Fig. 2d***), as well as a consistent increase in H3K27ac signal at enhancers that connected through these loops to activity-upregulated genes (***Fig. 2e***). These data indicate that the signal strength of epigenetic marks at distal regulatory elements and the interaction frequency of their long-range loops correlate with activity-dependent gene expression.

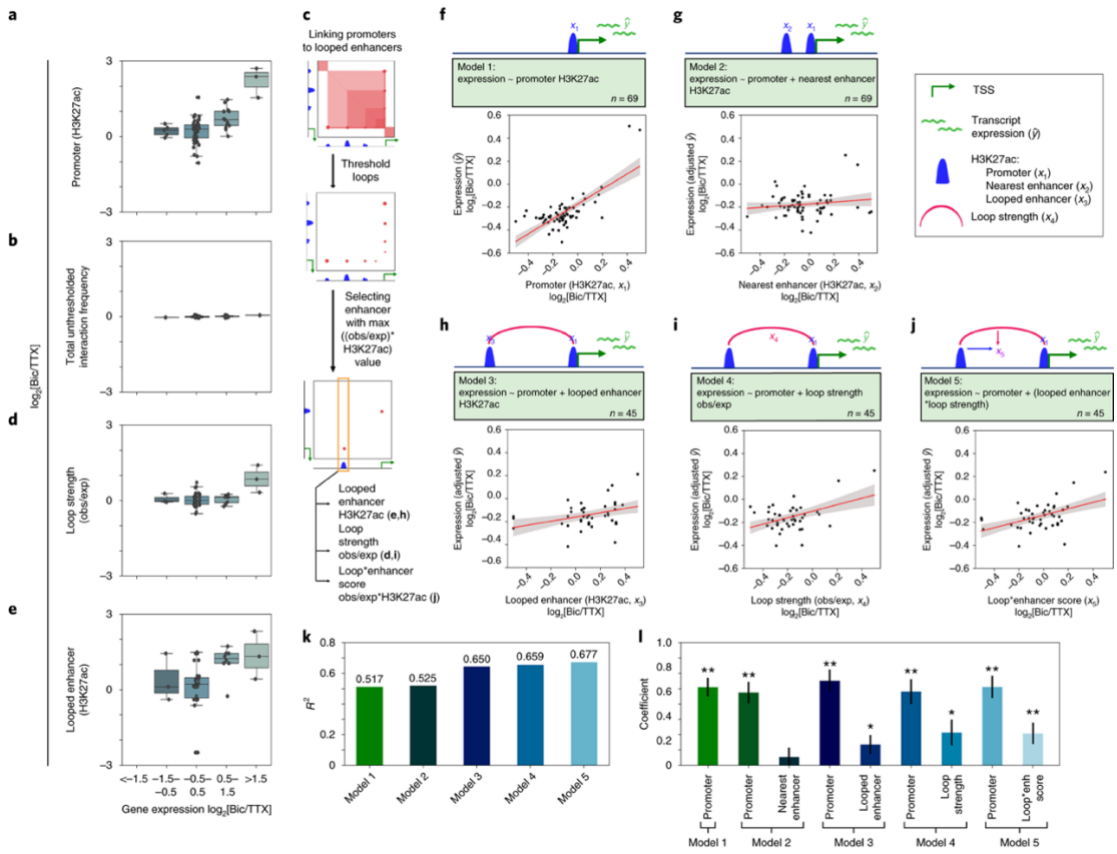


Figure 6.2. Activity-induced enhancers connected to distal target genes via looping interactions predict activity-stimulated expression.

a,**b**, Boxplots of the fold changes in promoter acetylation (**a**) and total interaction frequency (**b**) of genes grouped by fold change in expression. n = 69 independent genes. **c**, Schematic representation of the algorithm used to pair each gene with a single loop or enhancer that offered the highest predictive value. Only genes that formed such a loop (n = 45) were queried in the following models. obs/exp, observed/expected. **d**,**e**, Boxplots of the loop strength (**d**) and looped enhancer acetylation (**e**) after loops and enhancers are matched to genes using the schema presented in **c**. Boxes in **a**–**e** show the range from lower to upper quartiles, with the median line; whiskers extend to minimum and maximum data points within 1.5 times the interquartile range. **f**,**g**, Cartoon representations and scatterplots of the two 'null' models of the fold change in Bic-over-TTX (Bic/TTX) gene expression: promoter acetylation alone (model 1, **f**), promoter acetylation plus the acetylation of the nearest enhancer within 200 kb of the TSS (model 2, **g**). Fold change in expression is plotted on the

y axis, and the fold change in acetylation (of the promoter (**f**) and nearest enhancer (**g**) are plotted on the x axes. The fold change in expression in **g** has been adjusted to remove the values predicted by the promoter activity term in the model. Values have been min–max scaled to allow cross-model comparison. **h–j**, Cartoon representations and scatterplots of loop-containing models, plotted in the same manner as in **g**. In **f–j**, n represents the number of genes analyzed, the best fit line is shown in red, 95% confidence intervals are shown in gray. **k**, R^2 values for each of the three models. **l**, Barplot of explanatory variable coefficients from models 1–5. enh, enhancer. *P < 0.05, **P < 0.005 (two-tailed Student's t-test); error bars represent standard error of parameter elements.

Classic examples of activity-dependent enhancers, such as those for *Fos* and *Arc* [133] [134] [147], are relatively close (≤40 kb) to the promoters of these genes, but in many cases the nearest enhancers are insufficient to explain transcriptional regulation. We constructed multivariate linear models of activity-dependent gene expression (Supplementary Methods). Promoter H3K27ac alone explained 51.7% of the variance in gene expression after neuronal activation in our 5C regions (***Fig. 2f,k–l***). By adding the covariate of the H3K27ac signal at the nearest enhancer, we only marginally increased the performance of the model (***Fig. 2g,k–l***). We then built a third model with covariates of activity-dependent H3K27ac at (i) promoters and (ii) only distal enhancers engaged in maximum ABC-thresholded loops with their target genes (***Fig. 2c*** and Supplementary Methods). Our third, 'long-range enhancer' model markedly increased the variance of activity-dependent expression explained (***Fig. 2h,k–l***). Surprisingly, models that used loop strength (***Fig. 2i***) or the ABC value (loop strength × enhancer H3K27ac) between the selected enhancer and promoter (***Fig. 2j***) as covariates correlated similarly well with gene expression changes (***Fig. 2i–l***). These trends remained consistent when we analyzed the promoter and nearest enhancer models for genes

that only form long-range loops (Extended Data Fig. 6a–e). Together, these data indicate that long-range enhancers and loop strength can provide significant improvement in the prediction of activity-dependent expression compared to proximal, nearby enhancers.

Unique architectural motifs connect activity-dependent genes and enhancers

We next examined the extent to which looping reconfiguration occurred in parallel with activity-dependent enhancer changes or whether enhancers were pre-wired to their targets independent of their activation state (***Fig. 3a***). We first stratified H3K27ac peaks into activity-invariant (n = 14,424), activity-induced (n = 6014), and activity-decommissioned (n = 5402) putative enhancers (***Fig. 3b,c***, Supplementary Methods, Extended Data Fig. 6f–h, and Supplementary Tables 11–13). We quantified the degree of overlap between our enhancer classes and the anchors of our looping interactions. We identified three major architectural features for further exploration: (i) activity-induced loops anchored by activity-induced enhancers (n = 11) (class 1); (ii) activity-invariant loops pre-wired in inactive neurons and anchored by activity-induced enhancers (n = 41) (class 2); and (iii) activity-invariant loops pre-wired in inactive neurons and anchored by activity-decommissioned enhancers that lose their H3K27ac signal upon chronic neuronal activation (n = 15) (class 3) (***Fig. 3d,e***). These data reveal a complex long-range *cis*-regulatory landscape in which diverse loop classes might have unique roles in regulating activity-dependent gene expression.

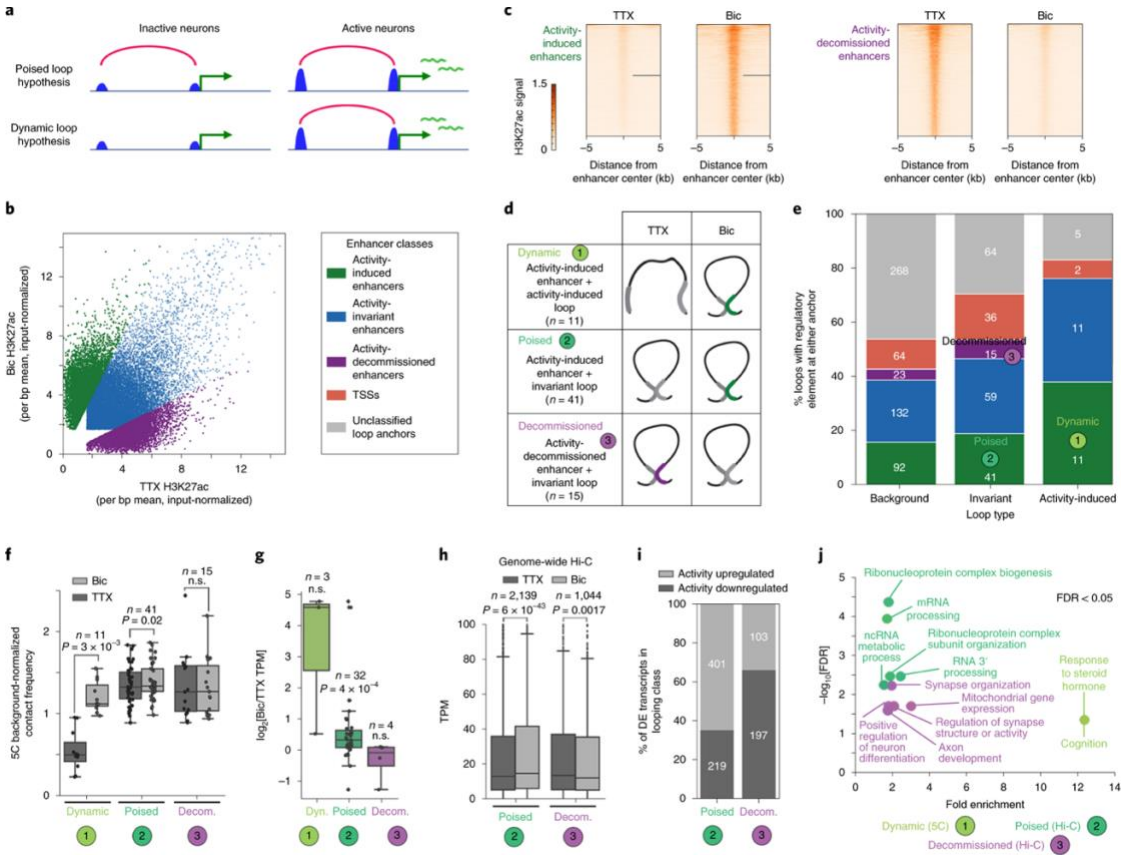


Figure 6.3. Unique topological motifs underlie the activity-dependent transcriptional response.

a, Cartoon representation of hypothesized models in which activity-induced enhancers operate to control gene expression via poised (top) or dynamic (bottom) loops. **b**, Scatterplot of enhancer acetylation across Bic and TTX conditions, thresholded by fold change in input-normalized signal and classified into activity-induced, activity-invariant and activity-decommissioned enhancers. **c**, Acetylation heatmaps of classified dynamic enhancers. **d**, Cartoon representations of the top three loop-enhancer classes of interest. Classified loop anchor colors match those in **b**, **c** and **e**. **e**, Stacked barplot displaying the percent of loops in each looping class with a classified enhancer at either of its anchors. A key of enhancer classes is shown in **b**. The number of loops in each subset is shown at the top of the bar. Loops could only be assigned to one enhancer class; the priority order of enhancer classes is from the bottom of the barplot (activity-induced enhancers, considered first) to the top (TSSs, considered last). **f**, Boxplots of background-normalized contact

frequencies for looping pixels in the five looping classes. Boxes in **f–h** show the range from lower to upper quartiles, with the median line; whiskers extend to minimum and maximum data points within 1.5 times the interquartile range. P values in **f–h** were calculated using the two-tailed Wilcoxon signed-rank test. The number of loops in each class is listed above boxes. n.s., not significant. **g**, Fold change in expression (log_2[Bic/TTX]) of the transcripts whose promoters intersect each looping class. The number of genes in each class is listed above boxes. **h**, Expression (TPM) of the genes whose promoters fall opposite activity-induced (class 2) and activity-decommissioned (class 3) enhancers in genome-wide cortical neuron loops (original data from ref *[119]*). The number of genes in each class is listed above boxes. **i**, Percent of differentially expressed (DE) genes (parsed using the Sleuth *[148]* Wald test, $q < 0.05$) in each genome-wide looping class that are upregulated in Bic compared to TTX (light gray) or downregulated in Bic compared to TTX (dark gray). n = number of genes in each set. **j**, Gene ontology enrichment calculated using Webgestalt for transcripts presented in **g** and **h**. Class 1 genes are from 5C regions only (**g**, n = 3); class 2 and 3 genes were parsed using genome-wide analyses (**h**, n = 2,139 class 2, n = 1,044 class 3). Only the top five terms for class 2 could be shown. See Extended Data Fig. 7e for the remaining terms at a false discovery rate (FDR) of < 0.05. ncRNA, noncoding RNA.

We then investigated the potential structural and functional properties of our three loop classes. We noticed that activity-induced loops anchored by activity-induced enhancers (class 1) underwent a 2.2-fold change in interaction frequency after 24 h of Bic treatment (***Fig. 3f***). Activity-invariant loops anchored by activity-decommissioned enhancers showed a strong and unchanged interaction frequency (class 3, ***Fig. 3f***). By contrast, interaction strength further strengthens after neuronal stimulation in the case of activity-invariant loops pre-wired to activity-induced enhancers (class 2, ***Fig. 3f***). Importantly, although class 1 loops are a rare occurrence, they corresponded to a 24-fold increase in activity-induced expression (***Fig. 3g*** and Extended Data Fig. 7a). Comparatively more genes engaged in class 2 loops but

on average displayed a modest 1.3-fold increase in expression in active neurons (***Fig. 3g*** and Extended Data Fig. 7a). These results suggest that, within our 5C regions, activity-induced loops are rare and connect to genes with large activity-dependent increases in expression, whereas pre-existing loops are more abundant but correlate with only minor gene expression changes.

To extend our findings genome-wide, we assessed the link between activity-invariant loop classes 2 and 3 and gene expression using the high-resolution Hi-C maps published in primary cortical neurons [119] and our activity-dependent RNA-seq and ChIP–seq data (***Fig. 3h*** and Extended Data Fig. 7b–d). We applied published methods [72] to identify 24,937 loops in cortical neurons (Extended Data Fig. 7c,d and Supplementary Tables 14–16) and stratify them into class 2 (n = 4,764) and class 3 (n = 3,259) groups (Supplementary Methods). Consistent with 5C loops, genes connected to activity-induced enhancers via activity-invariant loops (Class 2) displayed a modest but significant upregulation in expression after neuronal activation when we queried genome-wide loops (***Fig. 3h*** and Extended Data Fig. 7b). By contrast, genes looped to activity-decommissioned enhancers via activity-invariant loops (class 3) genome-wide exhibited a slight reduction in expression after neural activation (***Fig. 3h***). The majority of differentially expressed genes in class 2 versus class 3 loops were upregulated and downregulated, respectively, due to activity (***Fig. 3i***). Together, our data reveal that the genes connected to activity-induced enhancers via rare de novo loops show the largest effect size in activity-dependent expression. Genes can also exhibit modest but notable upregulation or downregulation when connected via pre-wired, activity-invariant loops to activity-induced (class 2) or activity-decommissioned (class 3)

enhancers, respectively. Pre-existing class 2 and class 3 loops are markedly more abundant in number than class 1 loops.

We investigated the ontology of the long-range target genes anchoring each looping class. Class 1 loops connect *Fos*, *Bdnf*, and *Tmed10* to activity-inducible enhancers, suggesting that the rapid upregulation of IEGs and SRGs involves the induction of de novo loops and de novo enhancers during neural activation (***Fig. 3j***). Class 2 pre-existing loops connect genes involved in several general cellular functions such as RNA processing to activity-induced enhancers, whereas class 3 pre-existing loops anchored by activity-decommissioned enhancers connect genes linked to synaptic organization and the regulation of synaptic activity (***Fig. 3j*** and Extended Data Fig. 7f). We were intrigued by the placement of synaptic genes in class 3 loops given that they connect to enhancers that are turned off during chronic (24 h) high activity levels. We therefore further stratified genes connected in class 3 loops by those (i) undergoing a 1.5-fold downregulation, (ii) undergoing a1.5-fold upregulation, and (iii) remaining unchanged after neural activity (Supplementary Methods). We found that the cohort of genes undergoing decreased expression in class 3 loops were predominantly involved in synapse organization and signaling, including *Gria1*, the main AMPA receptor subunit (***Fig. 3j*** and Extended Data Fig. 7f). These results reveal a potential mechanistic role for class 3 loops and activity-decommissioned enhancers in facilitating homeostatic plasticity during chronic high neural activity.

IEGs form shorter and less complex loops than SRGs

It is well established that IEGs are activated on the order of seconds to minutes in a translation-independent manner following neuronal activation, whereas SRGs are activated on the order of minutes to hours (ref. [125]). Consistent with this idea, we re-analyzed a

recently published RNA-seq time course during pharmacological neuronal activation [132] and found maximum activation of the IEGs *Fos* and *Arc* by 60 min, whereas maximum *Bdnf* upregulation occurred after 6 h (***Fig. 4a***). Visual inspection of the 5C heatmaps revealed two unexpected links between the kinetics of activity-dependent transcription and loop complexity (***Fig. 4b***). First, IEGs in our 5C regions form simple short-range loops with activity-dependent enhancers, and thus fall nearly exclusively in the class 1 category. For example, after 24 h of Bic treatment, *Fos* was upregulated more than 100-fold (***Fig. 4c***), but we identified only a single 40-kb-sized class 1 loop with an activity-induced enhancer (***Fig. 4d***). Similarly, *Arc* was upregulated more than 12-fold after neural activation (***Fig. 4c***) and also connected in a singular loop with an activity-induced enhancer (***Fig. 4e***). We note that the *Arc* interaction falls below our 30-kb distance threshold and therefore is not formally added to the class 1 loop list (***Fig. 3g–i***). By contrast, SRG *Bdnf* was upregulated 30-fold after neuronal activation (***Fig. 4c***) and connected into a complex network of multiple long-distance class 1 and class 2 loops (***Fig. 4f–i***), including: (i) at least two class 1 activity-induced loops anchored by activity-induced enhancers, but spanning longer distances (840 kb and 1,700 kb) than those formed with IEGs (***Fig. 4g,h***); and (ii) at least two class 2 activity-invariant loops anchored by activity-induced enhancers (***Fig. 4h,i***). The loops formed by *Bdnf* preferentially targeted its first promoter, from which we observed the highest level of transcription and strongest upregulation after 24 h of Bic-induced neuronal activation (Extended Data Fig. 8). Loops connected by *Bdnf* were significantly longer than those connected by *Fos* and *Arc* (***Fig. 4j***). These observations provide the basis for our working hypothesis that loop complexity and size underlie distinct epigenetic mechanisms governing IEG versus SRG upregulation in response to neuronal activation.

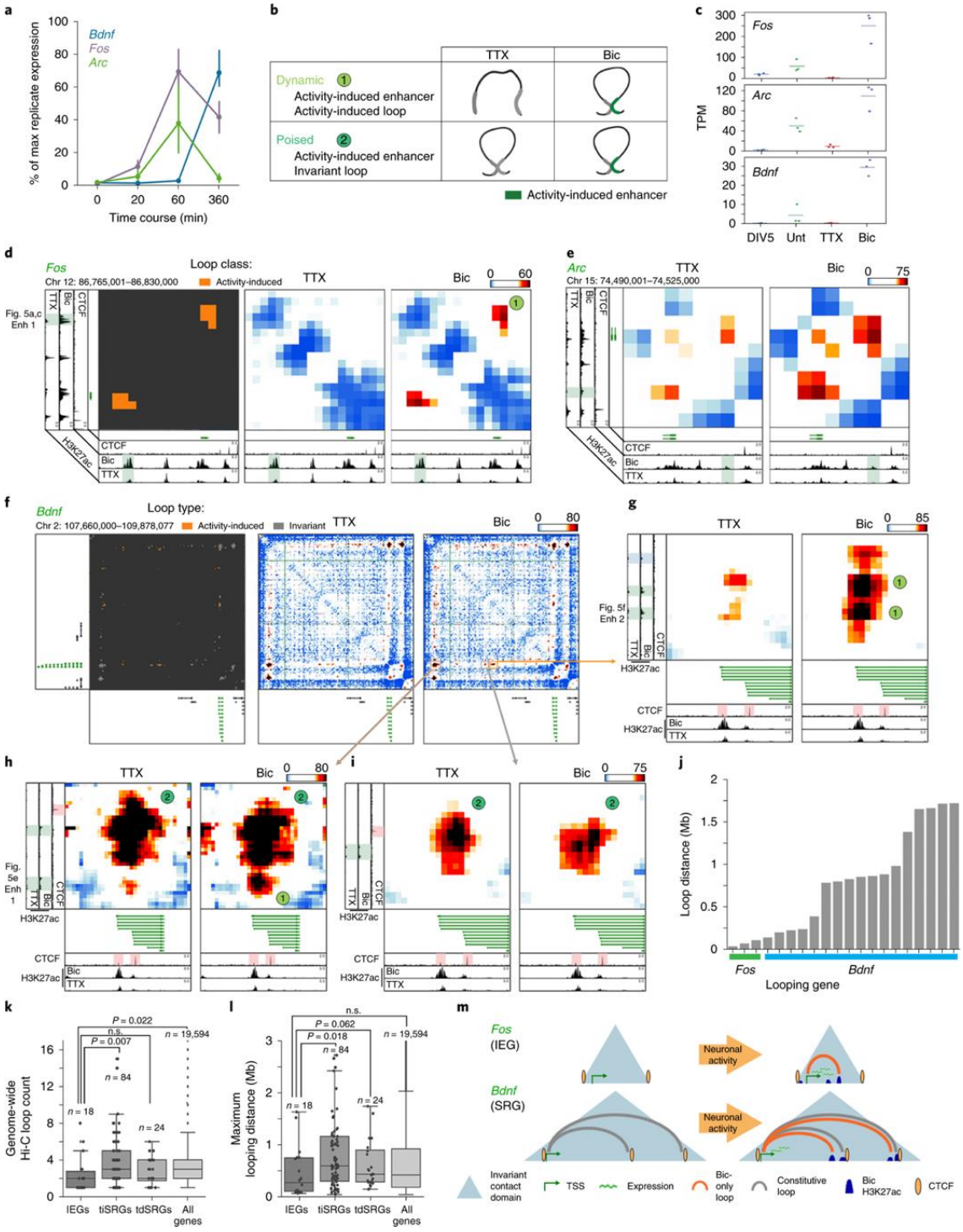


Figure 6.4. IEGs form shorter and less complex loops than SRGs.

A, Expression timing of Bdnf, Fos and Arc following the initiation of cortical neuron stimulation (from ref. *[133]*). N = 13 0-min, n = 4 20-min, n = 7 60-min, and n = 6 360-min replicates. The center line connects mean estimates, and error bars represent bootstrapped 95% confidence intervals. **B**, Cartoon representations of two loop classes identified in Fig. 3. **C**, Expression (TPM) of the Arc, Bdnf and Fos genes across the DIV5, untreated (Unt), TTX, and Bic conditions. N = 3, with mean lines plotted. **D**, Loop calls (left), TTX interaction score heatmap (middle) and Bic interaction score heatmap (right) of a ~65-kb region surrounding the Fos gene (green). Plotted beneath maps are cortical neuron CTCF (ref. *[119]*), Bic H3K27ac and TTX H3K27ac tracks. The Bic-specific enhancer underlying the Bic loop is highlighted in green. **E**, TTX interaction score heatmap (left) and Bic interaction score heatmap (right) of a ~35-kb region surrounding the Arc gene (green). **F)**, TTX interaction score heatmap (top), Bic interaction score heatmap (middle), and loop calls (bottom) of a ~2-Mb region surrounding the Bdnf gene (green). Bic loops are shown in orange and constitutive loops in gray. **G–I**, Interaction score heatmaps of three looping regions highlighted in **f** across TTX (left) and Bic (right) conditions. Plotted beneath maps are cortical neuron CTCF (ref. *[119]*), Bic H3K27ac and TTX H3K27ac tracks. Bic-specific enhancers are shown in orange and CTCF peaks highlighted in red. **J**, The genomic distance spanned by each loop formed by the Fos (n = 3) and Bdnf (n = 17) genes. **K,l**, Boxplots overlaid by stripplots of loop count (**k**) and maximum looping distance (**l**) for IEGs (defined as rPRGs in ref. *[133]*), translation-independent SRGs (tiSRGs, defined as dPRGs in ref. *[133]*), translation-dependent SRGs (tdSRGs), and all genes. P values are from two-sided Mann–Whitney rank tests comparing IEGs to other 3 classes. Boxes in **k,l** show the range from lower to upper quartiles, with the median line; whiskers extend to minimum and maximum data points within 1.5 times the interquartile range. N represents the number of genes in each class. **M**, Model representation of the distinct looping patterns of the Bdnf and Fos genes.

We next explored loop complexity genome-wide using published annotations of IEGs and SRGs [132] and the 24,937 loops from ES cell-derived mouse cortical neuron Hi-C maps (Extended Data Fig. 7c,d and Supplementary Tables 14–16). Published Hi-C data

represent only the untreated activity state, therefore we could not assess activity-induced loops (class 1) genome-wide. Nevertheless, we were able to integrate our data on genome-wide enhancers with cortical neuron Hi-C data to query the complexity of activity-invariant loops surrounding known activity-dependent genes genome-wide. Consistent with our locus-specific 5C results, we found that rapid response IEGs form significantly fewer loops (***Fig. 4k***), shorter loops (***Fig. 4l***), and connect to a lower number of activity-induced putative enhancers (Extended Data Fig. 8c) than both translation-independent and -dependent SRGs genome-wide. Together, these data are consistent with our working model in which SRGs engage in a complex network of long-range loops, whereas IEGs form simple, short-range loops to activity-induced enhancers to facilitate rapid activation independent of new protein synthesis (***Fig. 4m***).

Differential IEG and SRG looping kinetics after an acute neural activation time course

We next examined the kinetics of loop formation for IEGs and SRGs. We created 5C architecture maps in an acute time course of 0, 5, 20, 60, and 360 min of pharmacologically induced high activity in primary cultured mouse cortical neurons. To normalize baseline activity across different cultures, we pre-silenced our neural preparations by 24 h of TTX treatment before the addition of Bic (***Fig. 5*** and Supplementary Methods). We found that the class 1 loops surrounding *Fos* and *Arc* achieved peak contact frequency as quickly as 20 min after initiating stimulation (***Fig. 5a,b***). We also created total RNA-seq libraries at each time point and observed that the enhancer–promoter loop strength for IEGs peaks before maximum mRNA levels occur (60 min after stimulation) (***Fig. 5c,d***). Importantly, at early time points *Fos* interacted with an additional enhancer (***Fig. 5a***, 'Enhancer 2')

compared to its loop induced by 24 h of activity (***Fig. 4d***, 'Enhancer 1'), suggesting dynamic engagement with differential activity-induced enhancers over short time scales. We next measured enhancer activity dynamics by quantifying the RNA-seq signal that mapped to each enhancer (enhancer RNAs, eRNAs)12 (Supplementary Methods). We verified that our eRNA analysis approach produced activity-dependent dynamic patterns that resembled a previously published activity-induced eRNA data set12 and our own H3K27ac ChIP–seq data (Extended Data Fig. 9). The enhancers that loop to both *Fos* and *Arc* peak in activity 20 min after neuronal activity, and exhibit lower activity at all other time points (***Fig. 5c,d***). Although the extent to which loops causally drive gene expression is still under investigation, our observation that class 1 activity-induced enhancers and loops connect rapidly to IEGs before mRNA levels peak supports the assertion that the two are functionally linked.

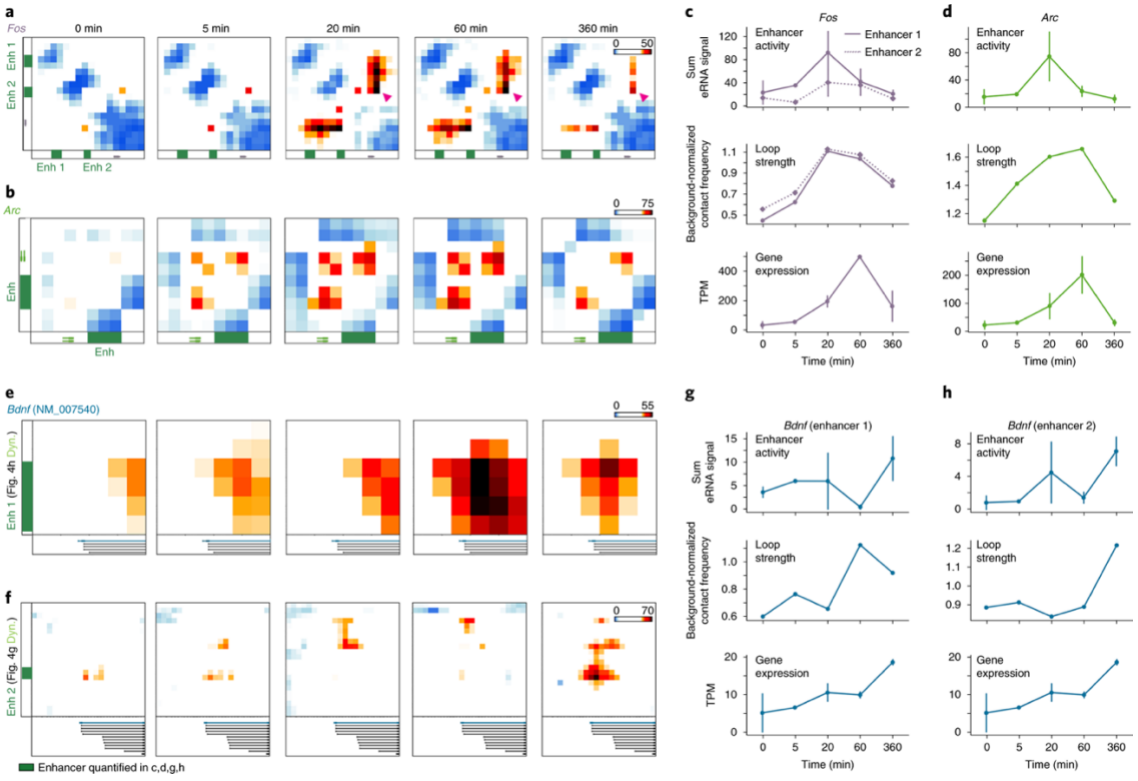


Figure 6.5. Activity-induced loops form before and persist after peak mRNA levels of IEGs.

a,**b**, Interaction score heatmaps surrounding Fos (**a**) and Arc (**b**) across 6 h of Bic treatment (preceded by 24 h of TTX silencing). Heatmap coordinates are identical to ***Fig. 4d*** (Fos) and ***Fig. 4e*** (Arc). Enhancers quantified in **c**,**d** are represented by green boxes. The magenta arrowhead denotes the Fos loop that is present only at early time points. **c**,**d**, Quantifications of Fos (**c**) and Arc (**d**) enhancer activity (top, quantified by eRNA signal), loop strength (middle, observed/expected 5C counts), and gene expression (bottom, TPM) across the activation time course. **e**,**f**, Interaction score heatmaps of activity-induced loops formed by the first Bdnf promoter. Heatmap coordinates in **f**, 'Enhancer 2', match those in ***Fig. 4g***. Heatmap coordinates in **e**, 'Enhancer 1', represent a zoomed-in subset of ***Fig. 4h*** to highlight an activity-induced loop. Enhancers quantified in **g**,**h** are represented by green boxes. **g**,**h**, Quantifications of Bdnf enhancer 1 (**g**) and enhancer 2 (**h**) activity (top) and loop strength (middle), coupled with the expression (bottom) of the Bdnf isoform with the strongest expression (see Extended Data Fig. 9). eRNA signal and gene expression are plotted as the mean of n = 2 RNA-seq replicates, error bars represent the 95% CI.

To test our hypothesis that loop dynamics contribute to the relatively delayed expression of SRGs (***Fig. 4k–n***), we quantified interaction frequency, enhancer activity, and mRNA levels for the class 1 loops formed by *Bdnf* (***Fig. 4g,h***). Consistent with our hypothesis, *Bdnf* class 1 loops did not interact until 60 (***Figs. 4h and 5e,g***, 'Enhancer 1') or 360 minutes (***Figs. 4g and 5f,h***, 'Enhancer 2') after stimulation. *Bdnf* enhancers and expression were upregulated in parallel with loops and did not reach maximum signal in our time course until 360 min of stimulated activity (***Fig. 5g,h***). Thus, *Bdnf* loop and enhancer dynamics are significantly delayed in comparison to those of *Fos* and *Arc*, corroborating our model that slower interaction kinetics may contribute to SRGs delayed expression.

6.3 Discussion

It has been known for decades that information storage in the brain requires de novo gene expression, but there is little consensus on whether and how specific epigenetic modifications maintain transcriptional signatures induced by neural activity. Here, we show that neuronal activity results in dynamic changes in the 3D genome that may inform precise temporal control of activity-dependent gene expression over short and long time scales.

Using chronic (24 h) neuronal activation and inhibition conditions, we demonstrate that activity-inducible enhancers engage in either de novo (class 1) or pre-existing (class 2) loops. Class 1 and class 2 loops connect to genes exhibiting a 24- and 1.3-fold activity-dependent increase in expression, respectively. Our 5C and genome-wide Hi-C results support our working model in which poised or pre-existing loops that connect to target genes in advance of activity-induced enhancer activation are abundant but have a modest effect on gene expression. Moreover, our 5C results suggest that loops that are induced by neural stimulation are relatively rare but exhibit a markedly higher effect on activity-dependent upregulation of distal target genes. The quantitative effect of these two looping classes on activity-dependent gene expression levels will be more precisely estimated in the future with genome-wide Hi-C and diverse activity-induction conditions. Future studies that focus on genome-wide detection of short-range class 1 architectural features, like *Arc* and *Fos*, will require maps with extremely high resolution using Micro-C [148] or high read-depth Hi-C created with restriction enzymes that cut 4-bp restriction sites.

A long-standing question in the transcription field is to what degree enhancer activation and/or looping strength are linked to gene expression. We used our loops and linear epigenetic data in chronic activity inhibition and induction conditions to create simple

linear models of activity-dependent expression changes. We find that H3K27ac signal at distal looped enhancers is a notably better predictor of activity-dependent target gene expression than nearest enhancers. The ability of our models to explain the variance of activity-dependent gene expression was achieved by building on a critical advance in the functional genomics field. In the ABC model, the multiplication of enhancer activity and 3D interaction frequency was the best predictor of enhancer–target gene pairs [148]. We used the ABC approach to choose a specific enhancer linked to each gene in our model, and this enabled us to prioritize and identify the looped enhancers that most significantly contributed to activity-dependent gene expression. Together, these data suggest that enhancer–target gene prediction would be facilitated by the use of chromatin architecture maps, instead of relying on the enhancer that is closest on the linear genome.

An important area of active research in neurobiology is elucidating the molecular mechanisms that regulate the unique temporal kinetics of IEGs and SRGs. Here, we unexpectedly observed that IEGs connect to enhancers via singular short-range loops that occur de novo after activation, whereas SRGs connect to multiple activity-inducible enhancers via a complex network of invariant and de novo loops. Consistent with our observations, another study reported—using H3K4me3 proximity ligation assisted ChIP–seq (PLAC-seq)—that the SRG *Nr4a3* engages in multiple long-range contacts over several hundred kb after neuronal stimulation [143]. These observations inspired our working hypothesis that looping complexity and distance are contributing factors to the kinetics of IEG and SRG expression (***Fig. 4m***). To critically assess this model, we induced acute pharmacological activation of neuronal activity and gathered looping and transcription data across multiple short time points. We observed striking differences in loop and enhancer

induction kinetics for IEGs and SRGs in our 5C regions. For example, enhancers and loops surrounding *Fos* and *Arc* peak in signal strength roughly 20 min after the induction of neuronal activity, before peak mRNA levels. By contrast, *Bdnf* loops and enhancers gain strength in parallel with mRNA levels over a longer time of sustained stimulation (360 min). We note that *Fos* engages in different short-range loops after 5 min, 20 min, and 24 h of neural activation, shifting interaction strength from a nearby enhancer to one that is more distal, suggesting that rapid activity-induced enhancer switching via alternative looping might be a mechanistic aspect of IEG upregulation (***Figs. 4d and 5a***). Together, these data form the basis of our working hypothesis that the complexity and size of long-range 3D interactions might functionally govern the kinetics of IEG and SRG expression with tight temporal precision during paradigms of synaptic plasticity.

We believe that greater understanding of how activity-dependent enhancers colocalize with daSNVs and connect over vast distances to distal target genes can provide critical new insights into the molecular mechanisms governing disease pathogenesis. Here, we identify a unique set of loops that are anchored by enhancers that decrease in activity during chronic stimulation. We speculate that enhancer decommissioning may be an epigenetic mechanism that is involved in homeostatic plasticity. Consistent with this hypothesis, we find that specific genes that are involved in homeostatic plasticity, such as *Gria1*, are connected in class 3 loops to activity-decommissioned enhancers and are downregulated during chronic high activity. We also observe that schizophrenia-associated SNVs are enriched at class 3 loops and are connected to downregulated genes after synaptic activity. By contrast, ASD-associated SNVs preferentially colocalize with class 2 loops that connect activity-inducible enhancers to activity-upregulated target genes. These results are striking as they suggest that

some disease-specific neuronal phenotypes may arise from noncoding SNVs that have different effects depending on the class of loops that they anchor (***Fig. 6e***). Moreover, the colocalization of schizophrenia-associated SNVs with class 3 loops suggests that defects in enhancer decommissioning might contribute to synaptic plasticity defects in neuropsychiatric diseases [149]. Genome misfolding has been reported in fragile X syndrome [122], the leading monogenic cause of ASD, as well as other human diseases [150], thus loop dysfunction could be possible owing to common SNVs in sporadic ASD and schizophrenia. Future work to build human activity-dependent loop maps and to dissect their functionality with genome editing will continue to refine our understanding of the functional role of distinct activity-dependent architectural features in neuropsychiatric disorders.

6.4 Methods

I include only sections describing my work to perform differential chromatin looping analysis using an adaptation of statistical tool I created, 3DeFDR. See [123] for complete methods.

5C interaction analysis

The adoption of the double alternating primer scheme and in situ 3C significantly improved 5C data quality (see ref. [121] for more details) such that some steps of our 5C analysis approach could be changed from those used previously [91] to more closely resemble those used for analyzing Hi-C [72]. Paired-end reads were aligned to the 5C primer pseudogenome using Bowtie, so that only reads with one unique alignment passed filtering. Only reads for which one paired end mapped to a forward or left-forward primer and the other end mapped to a reverse or left-reverse primer were tallied as true counts.

5C is subject to specific biases, such as primer GC content resulting in annealing or PCR biases, that methods such as Hi-C are not. This manifests in primer–primer pairs with mapped counts that are orders of magnitude higher than the neighboring primer–primer pairs. Such an extreme enrichment of single primer–primer pairs does not resemble the broader distribution of elevated counts, spanning clusters of neighboring primer–primer pairs, that exists at bona fide looping interactions across 5C and Hi-C data. Therefore, we decided to remove these biased primer–primer pairs before proceeding with interaction analysis. This was done by calculating for each primer–primer pair the median count of itself and the 24 primer–primer pairs nearest to the primer–primer pair in question (that is, a scipy.ndimagfor exampleeneric_filter window of size 5 was passed over the primer–primer pair matrix and the median of each window was recorded). If the count of one primer–primer pair was greater than eightfold higher than its neighborhood median then it was flagged as a high spatial outlier and removed. This process was performed for all primer–primer pairs, except for those in the 5C region surrounding the *Arc* gene, for which the eightfold threshold was found to be too stringent owing to low region complexity and therefore a 100-fold threshold was used instead.

After the removal of high outliers, primer–primer pair counts were quantile normalized across all 12 replicates (4 per condition) as described previously [108]. For plotting purposes, quantile-normalized counts were merged across replicates by summation, whereas for loop calling analysis all replicates were kept separate. Primer–primer pair counts were then converted to fragment–fragment interaction counts by averaging the primer–primer counts that mapped to each fragment–fragment pair (a maximum of two if both a forward or left-forward and a reverse or left-reverse primer were able to be designed to both

fragments and were not trimmed during outlier removal). We then divided our 5C regions into adjacent 4-kb bins and computed the relative interaction frequency of two bins (*i*,*j*) by summing the counts of all fragment–fragment interactions for which the coordinates of one of the constituent fragments overlapped (at least partially) a 12-kb window surrounding the center of the 4-kb *i*th bin and the other constituent fragment overlapped the 12-kb window surrounding the center if the *j*th bin. Binned count matrices were then matrix balanced using the ICE algorithm [108] [118], at which point we considered each entry (*i*,*j*) to represent the relative interaction frequency of the 4-kb bins *i* and *j*. Finally, the background contact domain 'expected' signal was calculated using the donut background model, as described previously [137], and used to normalize the relative interaction frequency data for the background interaction frequency present at each bin–bin pair. The resulting background-normalized interaction frequency (observed over expected) counts were fit with a logistic distribution from which *P* values were computed for each bin–bin pair and converted into background-corrected interaction scores (interaction score = $-10 \times \log_2[P \text{ value}]$) as described previously. Interaction scores have proven to be informatively comparable across replicates and conditions [120], and as such were used for most subsequent visualization analyses and all loop-calling analyses.

Quantitative 5C loop identification

We applied the 3DeFDR analysis package [61] to our data set to identify differential interactions across the TTX and Bic conditions (four replicates of each). In brief, 3DeFDR identifies differential interactions and estimates an empirical false discovery rate (eFDR) for each identified dynamic looping class. Interactions were considered for analysis only if the interaction scores of all eight replicates across both conditions surpassed a 'significance

threshold'. Interactions were classified as 'TTX-only' if all four interaction scores of the TTX replicates surpassed the interaction scores of the Bic replicates by more than a specified 'difference threshold'. 'Bic-only' interactions were classified in the same manner. Those interactions that passed the significance threshold but were not classified as Bic-only or TTX-only were classified as 'Constitutive'. Lastly, significant interactions that passed our thresholds were clustered based on spatial adjacency into 'loops'. Looping clusters that were smaller than 5 pixels were removed. The 3DeFDR package simulates null replicate sets (that is, eight replicates of the same cell type per condition) using a negative binomial counts generating function parameterized with mean-variance relationships computed from the real data. We compute an eFDR for each differential loop class as the total number of significant interactions called in that class on a simulated null replicate set divided by the total number of significant interactions called as that class with the original real replicate set.

We used the 'non-adaptive' functionality option of the 3DeFDR analysis package, which sweeps across a wide range of difference thresholds and calculates an eFDR for each loop class at each iteration. We generated 250 simulated null replicate sets of eight replicates based on mean-variance relationships underlying the real TTX replicates. We used the default 3DeFDR initialization parameters with the exception of 'bin_properties', which is a tunable parameter that specifies the distance scales over which fragment level interactions are stratified before fitting the negative binomial counts generating function to those interactions. We modified 'bin_properties' to capture the full extent of our regional matrices: (i) for close-range interactions (0–150 kb), we stratified the interactions using fine-grained, 12-kb sliding windows with a 4-kb step; (ii) for mid-range interactions (151–600 kb), we stratified the interactions into 24-kb sliding windows with an 8-kb step; and (iii) for longer-

range interactions (601–2,500 kb), we stratified the interactions into coarse-grained, 60-kb sliding windows with a 24-kb step. Through this approach we achieved an eFDR of 6.6% for Bic-only (activity-induced) loops using a difference threshold of 6.75, a significance threshold of $-10 \times \log_2(0.08)$ (that is, a *P* value of 0.08 resulting from the logistic fit to the observed over expected data), and a cluster size threshold of 5.

Part III: Multiscale Modeling

7 Modeling Patient-Specific Responses to Combination Cancer Therapies

Adapted from [151]

Recent advances in clinical cancer modeling have focused on discovery of new druggable targets and optimization of targeted therapeutics. Targeted therapeutics, however, are not available or affordable to the vast majority of cancer patients and instead these patients receive combination chemoradiotherapy in addition to surgery as standard of care. In clinic, modeling of combination therapy is typically performed via simple addition of the independent effects of these therapies, which are available in literature as population values without guidance for how they can be adapted for specific patients. Here, we present a mechanistic, multiscale modeling framework which can be applied in clinic to simulate combination therapy through simultaneous, rather than independent, application of component therapies to a patient-specific cell signaling model. This model is constructed through integration of the ErbB reception mediated Ras-MAPK and PI3K/AKT pathways with the TP53 mediated DNA damage response pathway and modulated based on patient-specific miRNA profiling. We anticipate that this model has utility in clinical decision making across many cancers for which combination therapy remains first line care. Here, we demonstrate this approach in one such cancer, describing its application to predict patient-specific responses to nephroblastoma, also known as Wilms' Tumor.

7.1 Introduction

Development, progression and metastasis of a wide variety of cancers have been attributed to activation of cellular signaling cascades through receptor proteins on the cell surface.

When activated by extracellular growth factors, these receptors typically dimerize and initialize a cascade of signals that propagate inside the cell eventually reaching the nucleus where they initiate various transcriptional programs that determines ultimate cell fate [152] [153]. Although there are considerable variations in the receptor type, signal duration and transcriptional program that determines cell fate in different cancers, they all share some common structural and functional features [154]. In most cell lines and in most common types of cancers, cell fate decisions are influenced by interaction of multiple processes operating at different time scales. Faster cell surface receptor mediated signaling pathways respond to suitable ligands to activate downstream proteins to transport into the nucleus. Once inside the nucleus, these proteins regulate slower processes guiding transcription and cell cycle progression [155]. In addition, the cell cycle is also influenced by DNA repair pathways which can be activated by cytotoxic drugs and radiation therapy. These repair pathways are often mediated by tumor suppressor gene TP53 [156] [157] which has been found to be frequently mutated in different cancers.

Apart from the events happening inside the cells, cellular outcome is affected in a profound manner by the heterogeneous microenvironment and in particular by its chemical and mechanical composition. Cells can detect and respond to changes in the microenvironment like stiff vs. soft extracellular matrix (ECM) or altered ligand composition through a complex interplay of receptor mediated signaling and reorganization of the matrix and cytoskeletal components [158] [159] [160].

Due to the complexity, multiscale nature, and sheer number of external factors that can influence outcomes, mathematical modeling is well positioned for investigating the processes which underlie cell fate decision making in the context of cancer. Such models

enable us to systematically vary input, explore parameter space and create testable predictions. They are useful not only from a basic science perspective to understand cellular behavior, and to motivate and design further experiments, but they also are of clinical value in determining effectiveness of personalized treatment strategies for specific patients and cancer types [40] [161] [162].

Despite of its great potential, mathematical modeling of biological processes in practice is a challenging task for multiple reasons:

- The amount of quantitative information available are often insufficient for detailed mathematical modeling. Many components and their interaction details are not known or been verified experimentally.
- Even if quantitative data are available, it is non-trivial to combine processes of multiple time scales.
- A single modeling paradigm is usually insufficient and there is a need to combine different models meaningfully.
- Uncertainties in the large number of free parameters can decrease the reliability of the model outcome unless special precautions are taken.

Here we have developed a heterogeneous and multi-scale modeling paradigm that can address these challenges, effectively combining models of different processes and timescales and generate testable and clinically useful predictions. Our aim was not to build an all-encompassing whole-cell model but rather provide a framework that can combine both existing and new models that were developed for different relevant processes using available experimental data possibly having different characteristic time scales. The models are

validated using both experimental and clinical information and are used to predict the effect of various single and combination therapies in patients of nephroblastoma (Wilms Tumor). We found that the patient response to single or combination chemotherapeutic drugs and radiation therapy is greatly influenced by the individual miRNA profiles of the patients. The model-predicted cell kill rates often differed from the literature cell kill rates of the chemotherapeutic drugs or cell kill rate calculated using empirical methods like Linear Quadratic model of radiation.

We believe that such an integrated modeling framework can be of great value in different cancers and help us understand the multiscale nature of cancer and design more effective treatment strategies using patient specific information and incorporate heterogeneity in tumor environment.

7.2 Challenges in Optimizing Clinical Combination Therapy

The central goal of all cancer therapy research is to identify and target properties of cancer cells that distinguish them from their normal counterparts [36]. Chief among these properties are overproliferation, defective DNA repair, reduced apoptosis, altered metabolism, increased angiogenesis, avoidance of immune surveillance, and invasion into neighboring tissues, which uninterrupted would lead to metastasis [36]. Many of these properties serve to tip the balance of cell fates from normal differentiative states to rapid proliferation of genetically unstable subclones that over time become increasingly adept at surviving hostile extracellular environments, as tumor expansion leads to lower nutrient availability and extracellular matrix (ECM) stiffening, and increasingly drug resistant. Understanding this, cancer therapies are developed to attack these properties with most

cytotoxic drugs acting to reduce proliferation and more recent targeted therapies addressing a wider variety of changes.

Increasingly effective cancer therapies have dramatically changed the outlook of cancer in the United States. When Sidney Farber introduced the first cytotoxic drug in 1944, the mean five-year survival rates across all cancers stood at 30% [36]. Today, the National Cancer Institute (NCI) reports a five-year survival rate of 69.3% for cancer patients for their most recent surveillance period of 2010-2016 [163]. These gains were achieved in part through optimization of chemotherapies in the clinical setting via pharmacokinetic studies, which profile how drug's plasma concentration changes over time, sometimes accompanied with information about how clearance changes with age, gender, drug interactions, and organ dysfunction, and to a far lesser extent, pharmacodynamic studies which correlate these properties to patient outcomes but are more challenging to conduct [36]. Importantly, these studies do not yield information about the mechanism of action of the drugs they evaluate, which historically has needed to be pieced together from cellular studies. These studies rarely evaluated combination therapy regimens with researchers instead reasoning that combination effects on cell kill rates would likely be additive [36] [164]. More recently, great effort has been made to characterize actions of targeted therapies and determine their pharmacokinetic-pharmacodynamic (PK-PD) correlates.

An unfortunate result of this is that PK-PD correlates have not been determined for the vast majority of combination therapies, which remains the standard of care for the majority of cancer patients, and that these patients have not benefitted from the same intensity of optimization efforts extended for targeted therapy. Instead, the practical endpoint for most combination therapy clinical studies has been to maximize dose per unit

time and in clinical practice, oncologists typically administer full doses of cytotoxic drugs with the goal of reaching reversible, well-tolerated toxicity as their endpoint [36]. Unfortunately, this is not always the case as many survivors are affected by chronic health conditions secondary to their care, and in extreme cases, patients may succumb to fatal toxicity. Without information in the literature for how to adjust regimens to individual patients, it remains extremely challenging for clinicians to determine what regimen could achieve best disease outcomes while minimizing dose toxicity.

In developing our multiscale framework, we hope to provide a tool that can help guide clinical decisions for combination therapy regimen design by providing cell kill rates that are patient specific and more representative of true simultaneous administration of multiple cytotoxic drugs than simple addition of population rates for individual drugs. Such guidance tools are highly sought after by clinicians and we describe this in the case of the nephroblastoma research community below.

7.3 Motivation for Application to Nephroblastoma

Nephroblastoma is a renal tumor affecting primarily children under ten years old. It is a relatively rare tumor at 7.7 cases per million children under 15 years old [165], but represents 7% of all childhood cancers [32]. This cancer is a treatment success story with progressively improved survival (85% overall [166]) and survivor healthiness due to advances in risk stratification, surgery, chemotherapy, and radiation therapy [20]. However, room for improvement remains with a cure rate of 95-99% for low stage tumors, and 66% for metastatic cases [167], and 25% survivors impacted by chronic conditions secondary to their treatment, including include "renal failure, infertility, cardiac toxicity, restrictive pulmonary

disease, and development of subsequent malignancies" according to Aldrink et al. 2019 [166] [168].

The overall goal of the nephroblastoma research community is to understand exactly what drives this cancer to initiate and recur in individual patients, and develop curative, minimally toxic therapy regimens optimized to an individual patient's genetics and tumor features. Current Wilms' Tumor management standards are defined primarily by two medical research groups, the Children's Oncology Group (COG) in the United States, and the International Society of Paediatric Oncology (SIOP), which is based in Europe. The groups generally agree on staging criteria but differ in treatment strategy; SIOP recommends pre-operative combination chemoradiotherapy while COG does not, instead exclusively applying these therapies after surgery [169] [170] [171]. Each approach has advantages, and both achieve strong outcomes, but lack of consensus suggests more information is needed to know which is truly optimal and highlights the need for new methods to inform treatment decision making.

Clinical decision making for treatment of nephroblastoma remains challenging and a significant portion of nephroblastoma research is dedicated addressing that through identification of biomarkers that are predictive of patient risk and treatment efficacy. In contrast with many pediatric tumors, nephroblastomas are not typically driven by single mutations, but a diverse, growing list of over forty cancer genes. [32] High genetic heterogeneity within tumors has also been observed for this cancer though it is rarely assessed in childhood tumors. [172] Additionally, Wilms' Tumor is a global disease with outcome disparities driven in part by unidentified genetic factors. [165] Amongst these factors perhaps are differences in transcriptional and post-transcriptional regulation of

targets within pathways critical to making cell fate decisions contributing to this cancer. For instance, aberrant expression of miRNAs is linked to nephroblastoma and are currently under investigation for use as biomarkers [173] [45] [174]. These issues are common across many cancers, though to a lesser extent in childhood cancers, and underscore an urgent need to personalize treatments for nephroblastoma to the individual genetics of patients.

Hoping to address these issues, SIOP-affiliated researchers across Europe formed a collaboration known as Computational Horizons In Cancer (CHIC) [175] to share clinical data and expertise necessary for developing *in silico* methods for modeling pediatric cancers and therapies. As part of the CHIC project, this lab gained access to sequencing, imaging, and clinical data necessary to validate predictions of Wilms' Tumor response to combination therapy. Here, we use this data to demonstrate our new offering to the research community in patient-specific mechanistic cancer modeling to support clinical decision making and therapy optimization.

7.4 Results

Our hybrid multiscale modeling framework can combine two or more biological processes and predict the time evolution and final steady state of the combination (please see the Methods section for more details). We combined the signaling pathways Ras-MAPK and PI3K/AKT mediated by Epidermal Growth Factor Receptor (EGFR) family with the tumor suppressor TP53 mediated DNA damage repair and cell cycle pathway, and we used this combined approach to predict the outcome of treatment for patients with nephroblastoma.

In the SIOP protocol, the standard treatment strategy is to either resect the tumor if it is less than a threshold size or prescribe chemotherapy to reduce the tumor size prior to

surgery. Across both the SIOP and COG protocols, varying combinations of the same standard chemotherapies are applied for most Wilms' Tumors, though some additional ones may be included to create a more intensive regimen for patients with additional risk factors, such as combined LOH at chromosomes 1p and 16q [176]. These common therapies include a) Doxorubicin b) Vincristine and c) Actinomycin D (also known as Dactinomycin) [177]. We use the model of the combined pathways to predict the cell kill rate for specific patients for different combinations of these three therapies.

One advantage of the SIOP protocol is that *in vivo* chemosensitivity of each child's tumor may be recorded as tumor volume change with preoperative therapy, and incorporated into risk stratification, which has allowed reduction of treatment intensity for patients across many stages of disease [178]. As a consequence, when provided with data from patients treated with this protocol, we received the following information:

- miRNA expression
- Prescribed chemotherapeutic drug dosage and schedule
- Tumor volume pre- and post-therapy

First, we used the model to predict probabilities of cell fates (death, growth, and senescence) for patients subjected to different cytotoxic therapy combinations. miRNA expression data was used to modulate target protein activity levels within our model to obtain patient-specific predicted cell fate probabilities. These probabilities were then used to compute tumor growth and compared against the actual growth rates. No mutations were part of the study. **Figure *7.1*** shows the cell kill probabilities calculated for a control patient (base values) and actual nephroblastoma patients with different miRNA expression profiles. The simulations were run for all possible combinations of three common chemotherapies used in

treatment of nephroblastoma and the subset of more commonly used combinations are shown in **Figure *7.1***. As expected, maximum cell death was achieved when all drugs were used and minimum when no drugs are used.

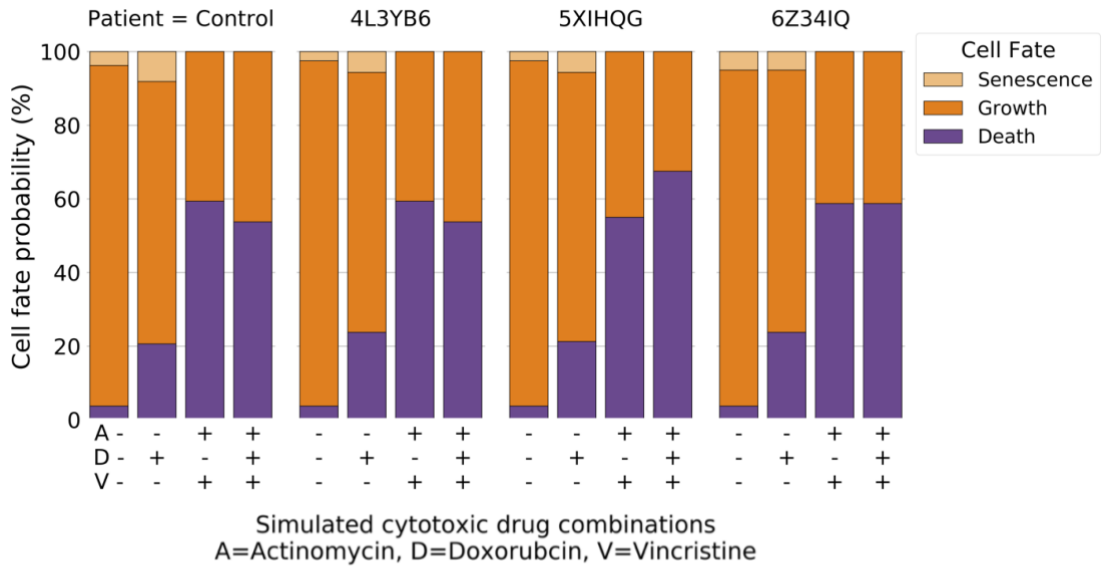


Figure 7.1. Predicted probabilities of cell death, growth, and senescence for a variety of possible cytotoxic therapy combinations.

Probabilities are computed as mean outcomes across an ensemble of models representing a range of initial starting values for unconstrained variables. Patient specific probabilities are obtained by modulating the initial activity or concentration levels of target proteins of a patient's most significantly differentially expressed miRNAs for targets present in the model. Such probabilities are shown for three nephroblastoma patients labeled with identification numbers. Control patient probabilities are obtained by running the simulation in the absence any modulation based on miRNA profiling. Drug specific probabilities are obtained by modeling each individual drug according to pharmacokinetic values in literature and modulating the activity levels of affected proteins in the heterogeneous, multiscale model.

In **Figure *7.2***, these data are summarized for all possible combinations of Actinomycin D, Doxorubicin, and Vincristine, using fixed dosages matching values used with actual patients (650 mg/m^2 Actinomycin D, 34 mg/m^2 Doxorubicin, and 1 mg/m^2 Vincristine).

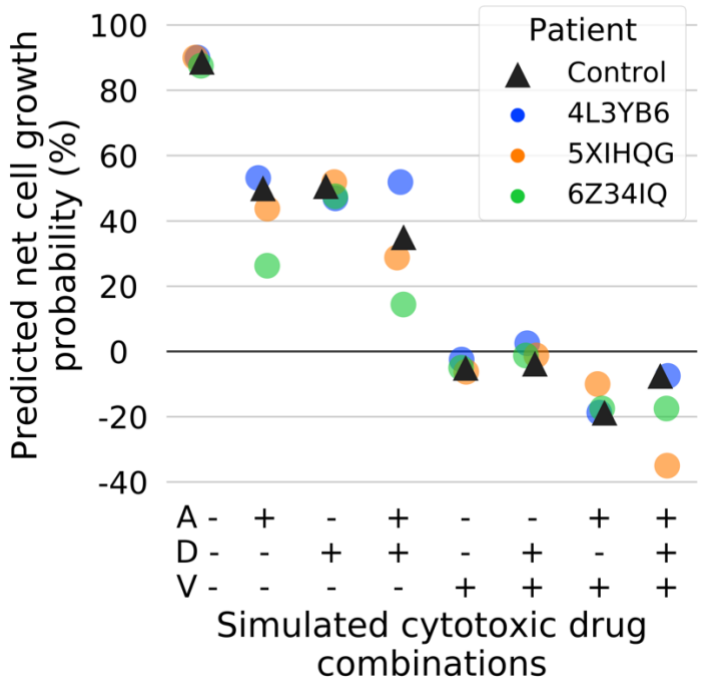
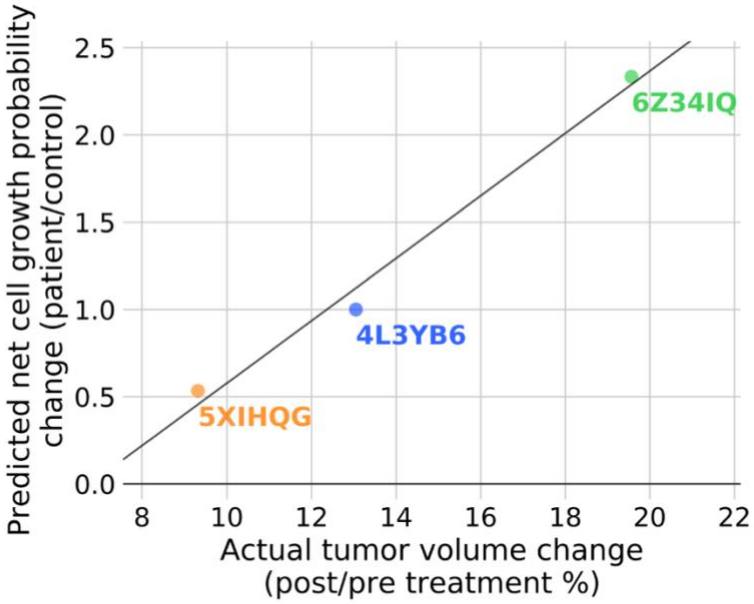


Figure 7.2. Predicted net cell growth probabilities computed as difference in cell death and cell growth probabilities for each patient and simulated drug combination.
Activity or concentration levels of target proteins are modulated in response to significantly differential patient miRNAs to obtain patient specific results, and in response to simulated drug activity derived from pharmacokinetic values for individual drugs in literature to obtain drug combination specific results. Simulated drugs are denoted as A = Actinomycin D, D = Doxorubicin, and V = Vincristine.

To determine whether patient specific cell fate probabilities were reflective of actual treatment effects observed, we compared observed change in tumor volume with treatment for each patient to the predicted change in net cell growth probability when the model was modulated using patient miRNA profiles and when it was not. As shown in **Figure *7.3***, we performed this patient for three patients for which we had tumor volume data pre- and post-chemotherapy as guided by the SIOP treatment protocol. While a constant percent reduction in tumor volume would have been predicted for patients treated with the same

chemotherapies based on literature values, we see that differences in volume reduction linearly correlate with patient-specific changes represented within our model. Based on these results, each patient potentially represents a different scenario of tumor response to generalized therapy: correspondence of treatment effect to literature values, underestimation of effect based on literature values, and overestimation of effect based on literature values.

a



b

Patient	Tumor volume (ml)		Administered chemotherapy		
	Pre	Post	Actinomycin D (ug/kg)	Doxorubicin (mg/m²)	Vincristine (mg/m²)
4L3YB6	287.20	37.48	650	34	1
5XIHQG	78.55	7.32	650	0	1
6Z34IQ	754.75	147.68	650	34	1

Figure 7.3. **a** Change in predicted net cell growth for three patients vs their observed changes in tumor volume.

Tumor volume change was computed as the percent of the post treatment tumor volume from the pretreatment tumor volume. Predicted net cell growth change was computed as the

ratio of net cell growth change of each patient to non-patient specific net cell growth change predicted for the corresponding therapy. **b** Pre and post treatment tumor volumes are listed for each patient alongside dosages of cytotoxic drugs administered.

In addition, we compared our model results to those obtained assuming additivity of constituent drug cell kill rates and separately, assuming additivity of rate constants, also known as the Bliss Independence model [164], which have both historically been used to assess therapy combinations. As shown in

Figure *7.4*, our results deviate strongly from those obtained assuming additivity of cell kill rates (listed as the sum modeling approach in the figure) as has been done for many cytotoxic drug combinations not formally assessed in pharmacodynamic studies. Comparatively, our model estimates deviate less strongly from those computed assuming additivity of rate constants, or drug independence according to the Bliss model, but are still more conservative for most patient and therapy combinations. This indicates that our model favors the assumption of additivity of rate constants over that of cell kill rates. We are interested in further investigating discrepancy between our results and those of the Bliss model, which is favored as a shorthand method for assessing drug combinations in clinical settings. When considered with the Bliss approach, our model might indicate more antagonistic behavior between drugs than would be expected, however, this type of Bliss assessment has been controversial in the research community with some indications that it is

misleading simplistic [164].

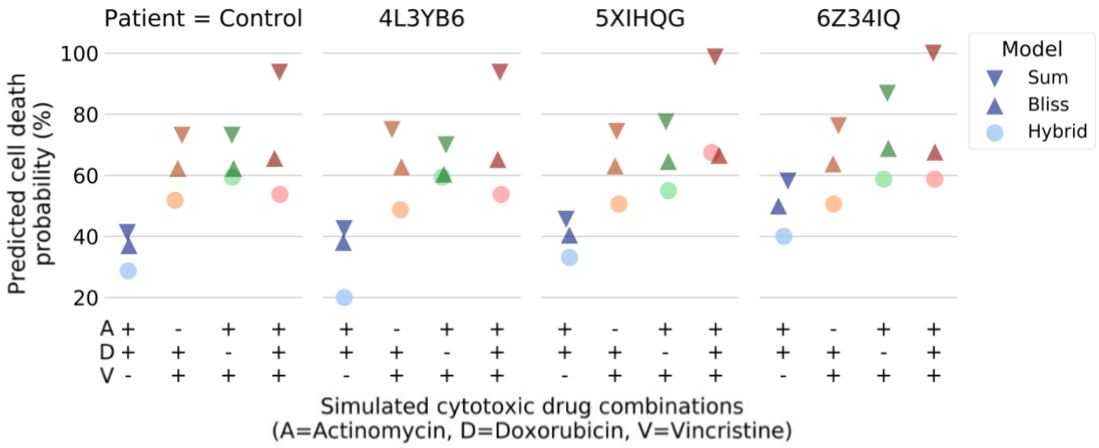


Figure 7.4. Comparison of predicted cell death probabilities obtained with the hybrid multiscale model vs the Bliss independence model and simple summation of individual drug rates.

Results are shown for patients listed by identification number and for simulated combinations of two or more of the cytotoxic drugs A = Actinomycin D, D = Doxorubicin, and V = Vincristine. Bliss model results and simple summation results were computed by applying them to cell death probabilities obtained from the hybrid model when simulating the administration of single drugs.

Next, we used the model to predict cell fate probabilities when subjecting each patient to different dosages of radiation at levels within range of that typical for treatment of nephroblastoma (14.4 Gy flank for intermediate risk disease and 25.2 Gy flank for high-risk disease for patient of moderate to high stage; patients may also receive a 10.8 Gy boost in both cases for lymph node involvement or gross disease [179]). In **Figure 7.5**, we show these results for four patients for which miRNA profiles were available through the CHIC project. We note that actual treatment dosages were not available for patient ECCOAH and that this patient had bilateral disease while the other patients had one affected kidney. As

shown in **Figure 7.5**, in the absence of chemotherapy, predicted net cell growth probability on average decreased nonlinearly with radiation dosage, perhaps reaching a plateau as radiation reached its maximal effect level in our model, and patient-specific probabilities were well matched to those predicted in the absence of modulation with patient miRNA profiles. However, in the presence of combination chemotherapy and radiation therapy, patient-specific predicted net growth probabilities differed considerably from generalized values and varied between patients and chemotherapy combinations.

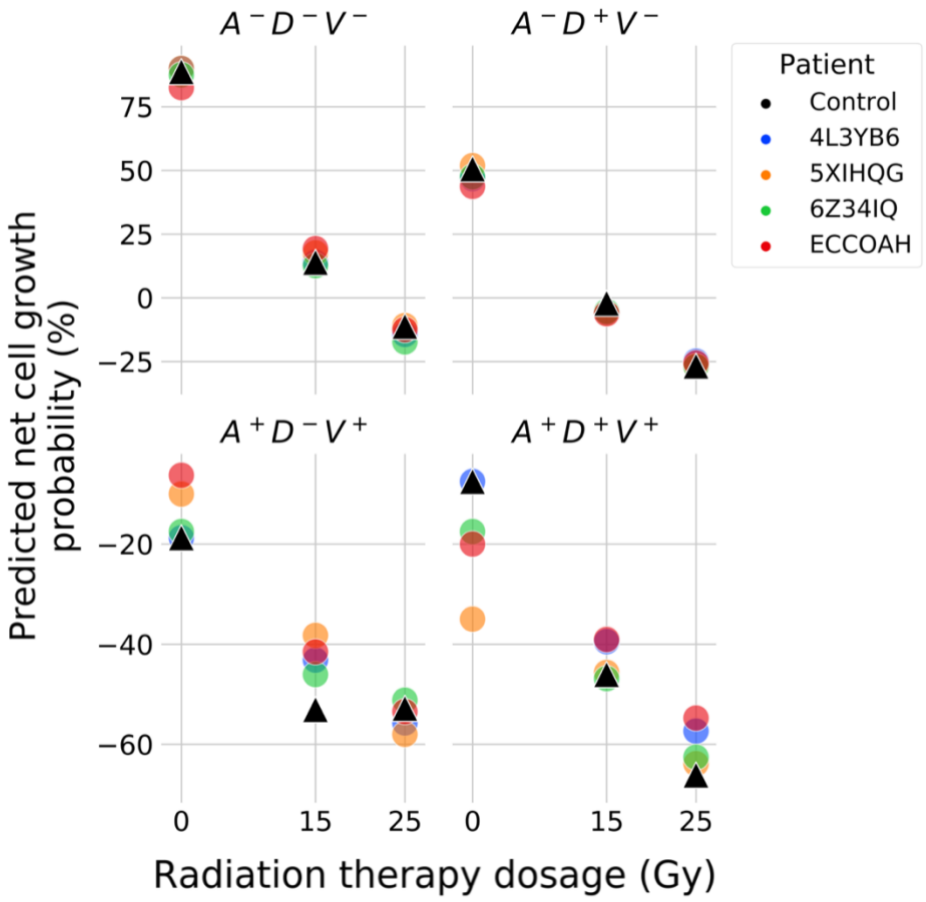


Figure 7.5. Predicted net cell growth probabilities for patients across a range of simulated radiation levels. Radiation is modeled using the standard linear quadratic model for dose response with model parameters (α,β) =(2.0e-2,5.1e-3) obtained from literature. We activate

ATM protein in TP53 pathway with a probability of $e^{-\alpha D}$ and obtain the cell survival fraction calculated only from double strand breaks via our hybrid mechanistic model. As described in **Methods**, we use this value to compute an adjusted patient-specific and treatment-specific estimate of cell kill and cell growth probability. Results are compared across four simulated cytotoxic drug combinations: no cytotoxic therapy ($A^- D^- V^-$), Doxorubicin alone ($A^- D^+ V^-$), Actinomycin D and Vincristine ($A^+ D^- V^+$), and all three drugs together ($A^+ D^+ V^+$). Patients are listed by identification number while the control patient refers to running of the model in the absence of patient specific modulation.

While combination chemoradiotherapy is common in treatment of nephroblastoma, we did not have data for tumor volume change as a result of radiation therapy or combined chemoradiotherapy in our patient data set with which to compare our findings. Instead, we observe that the patient-specific net cell growth probabilities predicted with our model trend closely with those yielded by the generalized linear-quadratic (LQ) dose response model often used to anticipate the effects of radiation therapy in clinical settings today [180] [181].

7.5 Methods

7.5.1 Description of modeling framework

We adapted a heterogenous multiscale modeling approach presented in [25] [182] to model nephroblastoma. In this framework, systems models representing distinct, but mechanistically linked biological processes are run simultaneously [25]. Each model in this framework represents each process at its operational time scale and time resolution, and these properties determine how models of different processes are integrated [25]. These of different processes are linked mechanistically according to what species occur in two or more constituent models and these links mediate the flow of information across models [25]. If two models do not have any species in common, they do not have direct interaction,

however, the interaction could still occur if a third model shared interfaces with both models [25].

To model Wilms' Tumor, we combined two modeling modules: an ErbB receptor mediated Ras-MAPK and PI3K/AKT signaling module and a TP53 mediated DNA damage response modeling [25]. The ErbB receptor models are implemented using continuous time and with a characteristic time scale of 6-8 hours [25]. By contrast, the TP53 mediated DNA damage response pathway is modeled in discrete time (Boolean model) and at a characteristic time scale of 24-48 hours [25]. A Boolean model is used in the case of the T53 mediated pathways as these pathways are challenging to precisely, quantitatively characterize because of their high complexity [25]. These models linked via common interfaces and their characteristic time scales are well separated, such that pseudo-steady state approximations may be used to combine these models as, when occurring at sufficiently separated time scales, from the perspective of the slower process, the faster process is at steady state [25].

7.5.2 Model interfaces and hybrid simulator algorithm

When processes occur at sufficiently distinct time scales, their models can be interfaced by evolving the slower process using steady state information from faster process [25]. Likewise, when multiple processes are modeled by different mathematical representations, e.g., continuous time ordinary differential equations and discrete time logical equations, their models can be mechanistically linked across common species by modifying the governing equations (e.g., ODE or Boolean rules) and initial state of one model by using information obtained by running the other model for a specified amount of time [25] [183]. The algorithm for the hybrid simulator is shown in the flowchart in **Figure *7.6*** and described in [182].

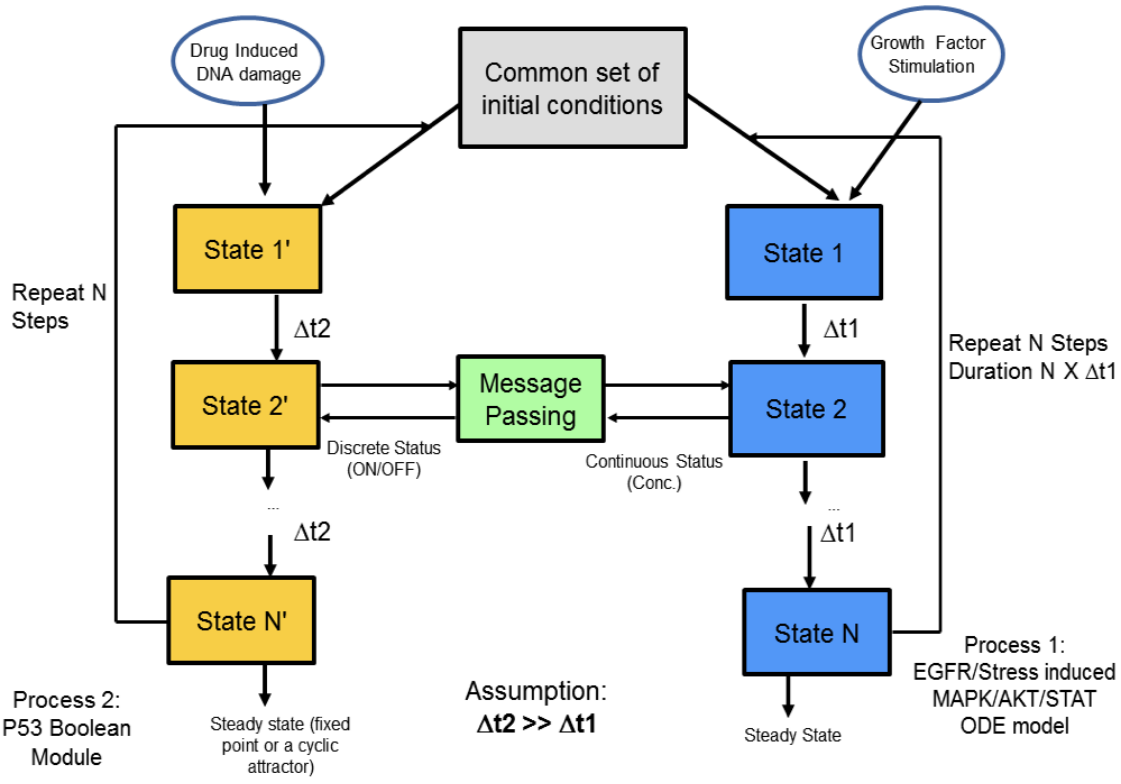


Figure 7.6. Flowchart representation of the hybrid simulator.

From Ghosh et al., yellow boxes represent states of the slower timescale processes and blue boxes represent models of faster timescale processes included in the framework. Each set of models has characteristic time scales, Δt1 and Δt2, and here Δt1 << Δt2. These models can have a common set of input conditions, such as patient-specific -omics profiles, therapy specific DNA damage or perturbation, or other variation of species for the purpose of *in silico* investigation of properties like cellular heterogeneity. Represented with the green box is the message passing interface that passes information across models using the common interface species. All constituent models are evolved this way until they reach a common steady state difference from control data. For Boolean models, in which only two states are possible, target nodes are constrained to an ON/OFF state dependent on their expression state for all members of the initial state space (see Methods for more detail). In this way, two instances of the model are initialized with two different gene and protein expression signatures which can be run to obtain the predictions for the patient and compared with

control results obtained when the simulation was run in the absence of patient specific data *[25]*.

7.5.3 Using miRNA expression data

MicroRNAs, or miRNAs, are short non-coding RNAs that act as post-transcriptional regulators of gene expression by either promoting mRNA degradation or inhibiting translation. Such miRNAs have been found to play a critical role in various forms of cancer, including nephroblastoma, and are under investigation for use as circulating serum biomarkers [44] [173]. From the CHIC project, tissue and serum miRNA expression profiles are available in the minml format for a group of nephroblastoma patients [45]. For each patient, we identified the specific mRNA targets of the 30 most significantly differentially expressed miRNAs in their profile using miRTarBase [183] [184]. Initial expression levels of these mRNA targets, if present in our model network, were constrained correspondingly before each model run. Hence the final outcomes were tailored to the expression profile of the patients to generate clinically useful outcomes. An example table is shown below indicating some top expressed miRNAs and their corresponding target mRNAs obtained from miRTarBase [182] [184].

miRNA	Sample mRNA Targets	Highly differentially expressed in patient			
		4L3YB6	5XIHQG	6Z34IQ	ECCOAH
hsa-miR-320b	IGF2, MAX				✓
hsa-miR-199a-5p	ERBB2, ERBB3				✓
hsa-miR-320d	IGF2, MAX				✓
hsa-miR-4284	MDM4			✓	
hsa-miR-125a-5p	CDKN1A, ERBB2, ERBB3, SIX1, TP53			✓	
hsa-miR-1260b	BRD7, CDKN1A			✓	✓
hsa-miR-1207-5p	ASXL1, MLLT1, TP53		✓		

hsa-miR-1275	ACTB		✓		
hsa-miR-574-5p	AMER1, CDKN1A		✓		
hsa-miR-4270	ASXL1, CDK2, ERBB2, MLLT1		✓		
hsa-miR-718	PTEN		✓		
hsa-miR-630	BCL2		✓		
hsa-miR-762	CDK2		✓		
hsa-miR-320c	IGF2, MAX		✓		
hsa-miR-4281	CDKN1A		✓		✓
hsa-miR-107	PTEN	✓		✓	
hsa-miR-4286	TP53	✓		✓	✓
hsa-miR-199a-3p	AKT1, MAP3K4	✓		✓	✓

Table 7.1. Sample list of top differentially expressed miRNAs based on miRNA profiles provided by the CHIC project for four patients.
Patients are listed by identification number. For each miRNA, sample mRNA targets are listed including those with species overlapping with those of our models and those linked to nephroblastoma. Targets were obtained from miRTarBase *[184]* and it should be noted that additional targets are associated with many of these miRNAs but as those are not relevant to our model, we do not list them here. Additionally, this list does not include miRNAs that were highly differentially expressed in these patients but did not have known targets listed on miRTarBase.

As an integrated clinical tool, our model takes miRNA profiles of a given patient, maps the most highly differential miRNAs to their target mRNAs, and then constrains these nodes in our network in order to capture the molecular effect of these miRNAs. Similarly, our model constrains nodes based on therapy interactions. In total, we consider Doxorubicin, Vincristine, and Actinomycin D as well as radiotherapy for nephroblastoma. The model is run based on the input miRNA profile and drug treatment and averages over several tissue conditions such as growth factor levels and receptor expression. An average as

well as a distribution of cell kill, cell senescence, and cell growth probabilities are obtained for a given patient, which are then passed on to the multi-modeler framework.

7.5.4 Model setup and simulations

Using the miRNA data, we identified the top 30 mRNA targets that are present in the core networks. The expression levels of these targets were increased or decreased by a preset value depending on the direction of miRNA expression changes. Additionally, we simulated the effects of administered chemotherapies, radiation therapy, and cellular heterogeneity. Thus, the basic workflow for simulation setup and run for each patient is as follows:

- Analyze miRNA profile and obtain a target set of proteins that will be activated or inhibited in the model
- Analyze patient treatment information (dosage of chemotherapeutic or radiation) and activate target DNA damage nodes
- Run and calculate cell fate probabilities for different growth factors to mimic cellular heterogeneity

For each type of chemotherapeutic drug, cell kill rates are available in literature and these generalized values are uniformly applied to patients in the absence of guidance for how to adjust them according to individual patient characteristics. To address this, we use our model to obtain an adjusted cell kill rate that accounts for patient specific genetic variation. We implemented this by first assuming that cytotoxic drugs influence cell survival according to Poisson distribution, such that the cell killed rate (CKR) may be obtained from:

$$CKR = 1 - e^{-kt}$$

where k is the rate constant and is proportional to the cell kill probability [25]. Using this relationship, we can obtain literature and adjusted cell kill rates as $CKR_{lit} = 1 - e^{-k_{lit}t}$ and $CKR_{adj} = 1 - e^{-k_{adj}t}$ respectively [25]. These two CKRs can then be related as follows:

$$\frac{k_{adj}}{k_{lit}} = \frac{\ln(1 - CKR_{adj})}{\ln(1 - CKR_{lit})}$$

The ratio $\frac{k_{adj}}{k_{lit}}$ is obtained from simulation for a patient and a control where control indicates no miRNA-based initialization of the model [25].

For a combination of multiple drugs, additivity of rate constants (probabilities) is assumed as opposed to additivity of cell kill fractions, as is commonly used in literature [25]. Given the cell kill rates of two different drugs $CKR_1$and CKR_2, additivity of rate constants yields:

$$\ln(1 - CKR_1) + \ln(1 - CKR_2) = -(k_1 + k_2)t = \ln(1 - CKR_{1+2})$$

Thus, assuming additivity of rate constants, the cell kill rate of chemotherapeutic drug combinations is the product rather than sum of the rates for the individual drugs [25]. We did find, as shown in

Figure *7.4* in the results section, that our model results do follow far more closely to results obtained assuming additivity of rate constants than assuming additivity of cell kill rates, however, our model results were more conservative than those computed directly from the equations above.

Next, to simulate the effects of radiation therapy, we referred to the linear quadratic model which predicts cell survival for dosage D Gy of radiation as

$$S = e^{-(\alpha D + \beta G D^2)} = e^{-\alpha D} e^{-\beta G D^2}$$

In the above, the parameter α is the proportionality constant for the number of DNA double strand breaks (DSB) to the dosage. In addition to DSB repair, binary mis-repair of pairs of DSBs can also produce lethal lesions which is proportional to the square of the dose. This part is not explicitly modeled and estimated directly from LQ model. G is Lea-Catchside factor which is calculated based on the dose duration and intervals. Using the value of β from literature, we activate ATM protein in TP53 pathway with a probability of $e^{-\alpha D}$and obtain the cell kill probability p_{ck}. Then the cell survival fraction calculated only from double strand breaks are $1 - p_{ck}$. Incorporating the mis-repair part from LQ model, the modified cell survival fraction is $(1 - p_{ck})e^{-\beta G D^2}$. Then the final adjusted cell kill probability is given by

$$p'_{ck} = 1 - (1 - p_{ck})e^{-\beta G D^2}$$

The results of this approach are shown in **Figure 7.5** and shows that cell kill probabilities assessed with this approach were more conservative when the model was modulated with patient-specific miRNA profiles than when it was used to obtain generalized estimates. These results were obtained using a = 2.0e-2 and b = 5.1e-3, obtained from literature. These values yield an a/b ratio of 3.92 Gy. The a, b parameters represent cells at different levels of radiosensitivity and a/b ratios vary with tissue, intra-tumor cell population, and cell cycle status with lower a/b ratios indicating higher radiosensitivity. The kidney is thought to have a/b ratios typically in the range of 3-5 Gy.

Finally, to mimic the variability of tumor microenvironment, simulations were run at multiple values of growth factor (EGF) concentration and across multiple values for randomly selected subsets of 15 unconstrained variables within our model. Final cell fate

probabilities are reported as averages across an ensemble set of model runs performed in this way.

7.6 Conclusions

To our knowledge, the described work represents the first multiscale mechanistic model of combination chemotherapy and radiation therapy. Here we employ a hybrid model framework representing interlinked growth factor mediated Ras-MAPK and PI3K/AKT pathways and TP53 mediated cell cycle and DNA damage response pathways to compute patient-specific and therapy-specific predictions of cell fate. Through this work, we demonstrate that a mechanistic approach to modeling combination therapy yields predictions of cell fate changes that differ widely from those of generalized black-box additivity models. We hope that this model can provide a mechanistic foundation for the development of tools applied in clinic to optimize combination therapy regimens for individual patients. In this work, we demonstrated this approach in the case of nephroblastoma and modulated through patient-specific miRNA profiling. In future work, we will expand this model into an agent-based framework in which tumor cell populations and surrounding tissue conditions can be represented. This will allow us to account for the effects of intratumor heterogeneity typically observed in nephroblastoma as well as those of tumor density, tumor microenvironment, and intercellular signaling on model predictions for treatment response. We will additionally conduct sensitivity analysis to identify key parameters driving model dynamics and finally, subject the framework to clinical validation.

8 Future Directions

I found it deeply gratifying to be able to translate computational modeling into methods that could be applied in the clinic to improve cancer treatment and want to continue contributing to such projects throughout my career. I am excited by multiple routes for building upon this work in the near future. Machine learning based enhancement of model performance represents one such route and allows one to obtain something of the best of both worlds between data-driven and mechanistic approaches. Mechanistic modeling enables us to obtain results that are explainable and insights that are transferrable across systems. I can apply machine learning to the parameters of mechanistic models to learn patterns of their dynamics that are predictive of model outcomes.

Refinement of my mechanistic representation of radiation therapy represents another promising route and there has been great interest in this segment of the project as the radiation oncology community seeks out tools to guide therapy planning through better prediction of patient-specific therapy benefit and toxicity, and I am very interested in expanding this model to represent therapy effects in healthy vs. tumor cells.

Finally, I hope to apply this modeling framework to predict effects of other cancer therapies, particularly immunotherapies through representation of tumor cell and microenvironment interactions with immune cells and estimation of their impact on anti-tumor immune suppression. Overall, I have been truly glad to be a part of this research community and look forward to other new opportunities to enhance the clinical actionability of multiscale cancer modeling.

References

[1] R. A. Nimmo, G. E. May and T. Enver, "Primed and ready: understanding lineage commitment through single cell analysis," *Trends in Cell Biology,*, vol. 25, no. 8, pp. 459-467, 2015.

[2] A. Marson, "Programming and Reprogramming Cellular Identity," *Dissertation ,* 2008.

[3] J. Holmberg and T. Perlmann, "Maintaining differentiated cellular identity," *Nature Reviews Genetics*, vol. 13, pp. 429-439, 2012.

[4] M. Jakovcevski and S. Akbarian, "Epigenetic mechanisms in neurological disease," *Nature Medicine,* vol. 18, pp. 1194-1204, 2012.

[5] A. B. Caldwell, Q. Liu, G. P. Schroth, D. R. Galasko, S. H. Yuan, S. L. Wagner and S. Subramaniam, "Dedifferentiation and neuronal repression define familial Alzheimer's disease," *Science Advances,* vol. 6, no. 46, p. eaba5933, 2020.

[6] D. Friedmann-Morvinski and I. M. Verma, "Dedifferentiation and reprogramming: origins of cancer stem cells," *EMBO Reports,* vol. 15, pp. 244-253, 2014.

[7] N. K. Lytle, A. G. Barber and T. Reya, "Stem cell fate in cancer growth, progression and therapy resistance," *Nature Reviews Cancer,* vol. 18, pp. 669-680, 2018.

[8] K. Takahashi and S. Yamanaka, "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors," *Cell*, vol. 126, no. 4, pp. 663-676, 2006.

[9] L. V. Nyugen, R. Vanner, P. Dirks and C. J. Eaves, "Cancer stem cells: an evolving concept," *Nature Reviews Cancer,* vol. 12, pp. 133-143, 2012.

[10] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander and J. M. Stuart, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, pp. 1113-1120, 2013.

[11] A. Conesa and S. Beck, "Making multi-omics data accessible to researchers," *Scientific Data,* vol. 6, p. 251, 2019.

[12] V. Marx, "The big challenges of big data," *Nature,* vol. 498, pp. 255-260, 2013.

[13] The GTex Consortium, "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *Science,* vol. 369, no. 6509, pp. 1318-1330, 2020.

[14] E. Wang, "A Roadmap of Cancer Systems Biology," in *Cancer Systems Biology*, Chapman & Hall / CRC Computational Biology, 2010, pp. 3-22.

[15] Data Integration in the Life Sciences: 8th Integration Conference, DILS 2012, College Park, MD, USA, June 28-29, 2012. Proceedings, Berlin, Heidelberg: Springer, 2012.

[16] B. Berger, J. Peng and M. Singh, "Computational solutions for omics data," *Nature Reviews Genetics,* vol. 14, pp. 333-346, 2013.

[17] W. Wen Bin Goh, W. Wang and L. Wong, "Why Batch Effects Matter in Omics Data, and How to Avoid Them," *Trends in Biotechnology,* vol. 35, no. 6, pp. 498-507, 2017.

[18] R. B. Mokhtari, T. S. Homayouni, N. Baluch and et al., "Combination therapy in combating cancer," *Oncotarget,* vol. 8, pp. 38022-38043, 2017.

[19] E. A. Ashley, "Towards precision medicine," *Nature Reviews Genetics,* vol. 17, pp. 507-522, 2016.

[20] J. Dome, E. J. Perlman and N. Graf, "Risk Stratification for Wilms Tumor: Current Approach and Future Directions," *American Society of Clnical Oncology Educational Book,* vol. 34, pp. 215-223, 2014.

[21] A. Hurria, K. Togawa, S. G. Mohile and et al., "Predicting Chemotherapy Toxicity in Older Adults With Cancer: A Prospective Multicenter Study," *Journal of Clinical Oncology,* vol. 29, no. 25, pp. 2457-3465, 2011.

[22] S. Ramón y Cajal, M. Sesé, C. Capdevila and et al., "Clinical implications of intratumor heterogeneity: challenges and opportunities," *Journal of Molecular Medicine,* vol. 98, pp. 161-177, 2020.

[23] A. A. Alizadeh, V. Aranda, A. Bardelli, C. Blanpain, C. Bock, C. Borowski, C. Caldas, A. Califano, M. Doherty, M. Elsner, M. Esteller, R. Fitzgerald, J. O. Korbel, P. Lichter, C. E. Mason, N. Navin, D. Pe'er, K. Polyak, C. W. M. Roberts, L. Siu, A. Snyder and Sto, "Toward understanding and exploiting tumor heterogeneity," *Nature Medicine,* vol. 21, pp. 846-853, 2015.

[24] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature Reviews Clinical Oncology,* vol. 15, pp. 81-94, 2017.

[25] A. K. Ghosh, "A Heterogeneous and Multiscale Modeling Framework to Develop Patient-Specific Pharmacodynamic Systems Models in Cancer," *Dissertation ,* 2020.

[26] T. S. Diesboeck, Z. Wang, P. Macklin and V. Cristini, "Multiscale Cancer Modeling," *Annual Review of Biomedical Engineering,* vol. 13, pp. 127-155, 2011.

[27] N. Moris, C. Pina and A. Martinez Arias, "Transition states and cell fate decisions in epigenetic landscapes," *Nature Reviews Genetics,* vol. 17, pp. 693-703, 2016.

[28] A. Brock, H. Chang and S. Huang, "Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours," *Nature Reviews Genetics,* vol. 10, pp. 336-342, 2009.

[29] A. Eldar and M. B. Elowitz, "Functional roles for noise in genetic circuits," *Nature,* vol. 467, pp. 167-173, 2010.

[30] R. Lopsick and C. Desplan, "Stochasticity and Cell Fate," *Science,* vol. 320, no. 5872, pp. 65-68, 2008.

[31] R. Radhakrishnan, "A survey of multiscale modeling: Foundations, historical milestones, current status, and future prospects," *AIChE Journal* , no. e17026, 2020.

[32] T. D. Treger, T. Chowdhury, K. Pritchard-Jones and S. Behjati, "The genetic changes of Wilms tumour," *Nature Reviews Nephrology* , vol. 15, pp. 240-251, 2019.

[33] A. T. Vessoni, E. C. Filippi-Chiela, G. Lenz and L. F. Z. Batista, "Tumor propagating cells: drivers of tumor plasticity, heterogeneity, and recurrence," *Oncogene,* vol. 39, pp. 2055-2068, 2019.

[34] D. A. Steindler, M. S. Okun and B. Scheffler, "Stem cell pathologies and neurological disease," *Modern Pathology,* vol. 25, pp. 157-162, 2011.

[35] D. Srivastava and N. DeWitt, "In Vivo Cellular Reprogramming: The Next Generation," *Cell,* vol. 166, no. 6, pp. 1386-1396, 2016.

[36] B. Izar, D. P. Ryan and B. A. Chabner, "Principles of Chemotherapy," in *Clinical Radiation Oncology*, Elsevier, 2016, pp. 171-185.

[37] C. H. Wong, K. W. Siah and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics,* vol. 20, no. 2, pp. 273-286, 2018.

[38] M. Choi, J. Shi, S. H. Jung and e. al., "Attractor Landscape Analysis Reveals Feedback Loops in the p53 Network That Control the Cellular Response to DNA Damage," *Science Signaling,* vol. 5, no. 251, p. ra83, 2012.

[39] M. R. Birtwistle, M. Hatakeyama, N. Yumoto and e. al., "Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses," *Molecular Systems Biology,* vol. 3, p. 144, 2007.

[40] W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger and P. K. Sorger, "Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data," *Molecular Systems Biology,* vol. 5, no. 1, p. 239, 2009.

[41] C. DeOcesano-Pereira, V. F. J, A. C. O. Carreira and e. al., "Post-Transcriptional Control of RNA Expression in Cancer," in *Gene Expression and Regulation in Mammalian Cells*, IntechOpen, 2018.

[42] P. P. Mitra, "Regulation of Mammalian Gene Expression," in *Gene Expression and Regulation in Mammalian Cells*, IntechOpen, 2018.

[43] R. Kempfer and A. Pombo, "Methods for mapping 3D chromsome architecture," *Nature Reviews Genetics,* vol. 21, pp. 207-226, 2019.

[44] Y. Peng and C. M. Croce, "The role of MicroRNAs in human cancer," *Signal Transduction and Targeted Therapy,* vol. 1, p. 15004, 2016.

[45] N. Ludwig, T. V. Werner, C. Backes and et al., "Combining miRNA and mRNA Expression Profiles in Wilms Tumor Subtypes," *International Journal of Molecular Science,* vol. 17, no. 4, p. 475, 2016.

[46] M. Cui, H. Wang, X. Yao and et al., "Circulating MicroRNAs in Cancer: Potential and Challenge," *Frontiers in Genetics,* vol. 10, p. 626, 2019.

[47] G. Huang, F. Li, Y. Ma and et al., "Functional and Biomimetic Materials for Engineering the Three-Dimensional Cell Microenvironment," *Chemical Reviews*, vol. 117, pp. 12764-12850, 2017.

[48] R. Baghban, L. Roshangar, R. Jahanban-Esfahlan and et al., "Tumor microenvironment complexity and therapeutic implications at a glance," *Cell Communication and Signaling* , vol. 18, p. 59, 2020.

[49] F. Broders-Bondon, T. H. N. Ho-Bouldoires, M.-E. Fernandez-Sanchez and e. al., "Mechanotransduction in tumor progression: The dark side of the force," *Journal of Cell Biology*, vol. 217, no. 5, pp. 1571-1587, 2018.

[50] J. Walpole, J. A. Papin and S. M. Peirce, "Multiscale Computational Models of Complex Biological Systems," *Annual Review of Biomedical Engineering*, vol. 15, pp. 137-154, 2013.

[51] E. D. Mitra and W. S. Hlavacek, "Parameter estimation and uncertainty quantification for systems biology models," *Current Opinions in Systems Biology*, vol. 18, pp. 9-18, 2019.

[52] F. Konrath, A. Mittermeier, E. Cristiano, J. Wolf and A. Loewer, "A systematic approach to decipher crosstalk in the p53 signaling pathway using single cell dynamics," *PLOS Computational Biology*, vol. 16, no. 6, 2020.

[53] T. Sun and J. Cui, "Dynamics of P53 in response to DNA damage: Mathematical modeling and perspective," *Progress in Biophysics and Molecular Biology*, vol. 119, no. 2, pp. 175-182, 2015.

[54] M. L. Wynn, N. Consul, S. D. Merajver and S. Schnell, "Logic-based models in systems biology: a predictive and parameter-free network analysis method†," *Integrative Biology*, vol. 4, no. 11, pp. 1323-1337, 2012.

[55] D. Machado, R. Costa, M. Rocha and e. al., "Modeling Formalisms in Systems Biology," *AMB Express*, vol. 1, p. 45, 2011.

[56] S. Hoops, S. Sahle and et al., "COPASI: a COmplex PAthway SImulator," *Bioinformatics*, vol. 22, pp. 3067-74, 2006.

[57] A. Finney, M. Hucka and et al., "Software Infrastructure for Effective Communication and Reuse of Computational Models," in *System Modeling in Cell Biology: From Concepts to Nuts and Bolts*, MIT Press, 2006, pp. 355-78.

[58] H. de Jong, J.-L. Gouze, C. Hernandez and e. al., "Qualitative simulation of genetic regulatory networks using piecewise-linear models," *Bulletin of Mathematical Biology*, vol. 66, no. 2, pp. 301-340, 2004.

[59] R.-S. Wang, "Ordinary Differential Equation (ODE) Model," in *Encyclopedia of Systems Biology*, 2013.

[60] J. B. Fitzgerald, B. Schoeberl, U. B. Nielsen and e. al., "Systems biology and combination therapy in the quest for clinical efficacy," *Nature Chemical Biology*, vol. 2, no. 9, pp. 458-466, 2006.

[61] L. R. Fernandez, T. G. Gilgenast and et al., "3DeFDR: statistical methods for identifying cell type-specific looping interactions in 5C and Hi-C data," *Genome Biology,* vol. 21, p. 219, 2020.

[62] W. de Laat and J. Dekker, "3C-based technologies to study the shape of the genome," *Methods,* vol. 58, pp. 189-191, 2012.

[63] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren and et al, "The 4D nucleome project," *Nature,* vol. 549, pp. 219-226, 2017.

[64] J. Dekker and L. Mirny, "Biological techniques: chromosomes captured one by one," *Nature,* vol. 502, pp. 45-46, 2013.

[65] J. Dekker, K. Rippe, M. Dekker and N. Kleckner, "Capturing chromosome conformation," *Science,* vol. 295, pp. 1306-1311, 2002.

[66] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum and et al., "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements," *Genome Research,* vol. 16, pp. 1299-1309, 2006.

[67] E. Lieberman-Aiden, N. L. van Berkum, L. Williams and et al., "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science,* vol. 326, p. 289–293, 2009.

[68] B. van Steensel and J. Dekker, "Genomics tools for unraveling chromosome architecture," *Nature Biotechnology,* vol. 28, pp. 1089-1095, 2010.

[69] B. Mifsud, F. Tavares-Cadete, A. N. Young and et al., "Mapping long-range promoter contacts in human cells with high-resolution capture hi-C," *Nature Genetics,* vol. 47, pp. 598-606, 2015.

[70] Z. Zhao, G. Tavoosidana, M. Sjolinder and et al., "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions," *Nature Genetics,* vol. 38, pp. 1341-1347, 2006.

[71] M. Simonis, P. Klous, E. Splinter and et al., "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)," *Nature Genetics,* vol. 38, pp. 1348-1354, 2006.

[72] S. S. P. Rao, M. H. Huntley, N. C. Durand and et al., "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell,* vol. 159, pp. 1665-1680, 2014.

[73] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie and et al., "Chromatin architecture reorganization during stem cell differentiation," *Nature,* vol. 518, pp. 331-336, 2015.

[74] J. R. Dixon, S. Selvaraj, F. Yue and et al., "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature,* vol. 485, pp. 376-380, 2012.

[75] L. Guelen, L. Pagie, E. Brasset and e. al., "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions," *Nature,* vol. 453, pp. 948-951, 2008.

[76] D. Peric-Hupkes, W. Meuleman, L. Pagie and et al., "Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation," *Molecular Cell,* vol. 38, pp. 603-613, 2010.

[77] J. O. Davies, J. M. Telenius, S. J. McGowan and e. al., "Multiplexed analysis of chromosome conformation at vastly improved sensitivity," *Nature Methods,* vol. 13, pp. 74-80, 2016.

[78] J. R. Hughes, N. Roberts, S. McGowan and et al., "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment," *Nature Genetics,* vol. 46, pp. 205-212, 2014.

[79] T. Nagano, Y. Lubling, T. J. Stevens and et al., "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure," *Nature,* vol. 502, pp. 59-64, 2013.

[80] D. H. Phanstiel, K. Van Bortle, D. Spacek and et al., "Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development," *Molecular Cell,* vol. 67, no. e1036, pp. 1037-1048, 2017.

[81] K. Nasmyth, "Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis," *Annual Review of Genetics,* vol. 35, pp. 673-745, 2001.

[82] A. D. Riggs, "DNA methylation and late replication probably aid cell memory, and type I DNA reeling could aid chromosome folding and enhancer function," *Philosophical Transactions of The Royal Society B Biological Sciences,* vol. 326, pp. 285-297, 1990.

[83] E. Alipour and J. F. Marko, "Self-organization of domain structures by DNA-loop-extruding enzymes," *Nucleic Acids Research,* vol. 40, pp. 11202-11212, 2012.

[84] I. M. L. C. G. A. A. N. M. L. Fudenberg G, "Formation of chromosomal domains by loop extrusion," *Cell Reports,* vol. 15, pp. 2038-2049, 2016.

[85] A. L. Sanborn, S. S. P. Rao, S. C. Huang and et al., "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes," *Proceedings of the National Academy of Sciences USA,* vol. 112, pp. E6456-65, 2015.

[86] A. Goloborodko, J. F. Marko and L. A. Mirny, "Chromosome compaction by active loop extrusion," *Biophysics Journal,* vol. 110, pp. 2162-8, 2016.

[87] J. M. Dowen, Z. P. Fan, D. Hnisz and et al., "Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes," *Cell,* vol. 159, pp. 374-387, 2014.

[88] E. P. Nora, A. Goloborodko, A. L. Valton and et al., "Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization," *Cell,* vol. 169, no. e922, pp. 930-944, 2017.

[89] V. Narendra, P. P. Rocha, D. An and et al., "Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation," *Science,* vol. 347, pp. 1017-1021, 2015.

[90] W. A. Flavahan, Y. Drier, B. B. Liau and et al., "Insulator dysfunction and oncogene activation in IDH mutant gliomas," *Nature,* vol. 529, pp. 110-114, 2016.

[91] J. A. Beagan, M. T. Duong and et al., "YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment.," *Genome Research,* vol. 27, pp. 1139-1152, 2017.

[92] A. S. Weintraub, C. H. Li, A. V. Zamudio and et al., "YY1 is a structural regulator of enhancer-promoter loops," *Cell,* vol. 171, no. e1528, pp. 1574-1588, 2017.

[93] S. S. P. Rao, S. C. Huang, B. Glenn St Hilaire and et al., "Cohesin loss eliminates all loop domains," *Cell,* vol. 171, no. e324, pp. 305-320, 2017.

[94] W. Schwarzer, N. Abdennur, A. Goloborodko and et al., "Two independent modes of chromatin organization revealed by cohesin removal," *Nature,* vol. 551, pp. 51-56, 2017.

[95] D. Hnisz, A. S. Weintraub, D. S. Day and et al., "Activation of proto-oncogenes by disruption of chromosome neighborhoods," *Science,* vol. 351, pp. 1454-1458, 2016.

[96] M. Franke, D. M. Ibrahim, G. Andrey and et al., "Formation of new chromatin domains determines pathogenicity of genomic duplications," *Nature,* vol. 538, pp. 265-269, 2016.

[97] D. G. Lupianez, K. Kraft, V. Heinrich and et al., "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions," *Cell,* vol. 161, pp. 1012-1025, 2015.

[98] A. Sanyal, B. R. Lajoie, G. Jain and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature,* vol. 489, pp. 109-113, 2012.

[99] E. M. Smith, B. R. Lajoie, G. Jain and J. Dekker, "Invariant TAD boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the CFTR locus," *American Journal of Human Genetics,* vol. 98, pp. 185-201, 2016.

[100] B. M. Javierre, O. S. Burren, S. P. Wilder and et al., "Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters," *Cell,* vol. 167, no. e1319, pp. 1369-84, 2016.

[101] J. Onkar, S. Y. Wang, T. Kuznetsova and et al., "Dynamic reorganization of extremely long-range promoter-promoter interactions between two states of pluripotency," *Cell Stem Cell,* vol. 17, pp. 748-757, 2015.

[102] N. C. Durand, M. S. Shamim, I. Machol and et. al., "Juicer provides a one-click system for analyzing loop-resolution hi-C experiments," *Cell Systems,* vol. 3, pp. 95-98, 2016.

[103] B. Mifsud, I. Martincorena, E. Darbo and et al., "GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data," *PLoS One,* vol. 12, p. e0174744, 2017.

[104] A. T. Lun and G. K. Smyth, "diffHic: a bioconductor package to detect differential genomic interactions in Hi-C data," *BMC Bioinformatics,* vol. 16, p. 258, 2015.

[105] Y. C. Hwang, C. F. Lin, O. Valladares and et al., "HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements," *Bioinformatics,* vol. 31, pp. 1290-1292, 2015.

[106] F. Ay, T. L. Bailey and W. S. Noble, "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts," *Genome Research,* vol. 24, pp. 999-1011, 2014.

[107] M. Forcato, C. Nicoletti, K. Pal and e. al., "Comparison of computational methods for Hi-C data analysis," *Nature Methods,* vol. 14, pp. 679-685, 2017.

[108] T. G. Gilgenast and e. al., "Systematic evaluation of statistical methods for identifying looping interactions in 5C data," *Cell Systems,* vol. 8, no. e113, pp. 197-211, 2019.

[109] M. N. Djekidel, Y. Chen and M. Q. Zhang, "FIND: difFerential chromatin INteractions Detection using a spatial Poisson process [published online ahead of print, 2018 Feb 12]," *Genome Research,* vol. 28, no. 3, pp. 412-422, 2018.

[110] J. Paulsen, G. K. Sandve, S. Gundersen, T. G. Lien, K. Trengereid and E. Hovig, "HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization," *Bioinformatics,* vol. 30, pp. 1620-1622, 2014.

[111] C. A. Lareau and M. J. Aryee, "hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data," *Nature Methods,* vol. 15, pp. 155-156, 2018.

[112] C. A. Lareau, M. J. Aryee and B. Berger, "diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data," *Bioinformatics,* vol. 34, pp. 672-674, 2018.

[113] G. Li, Y. Chen, M. P. Snyder and M. Q. Zhang, "ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis," *Nucleic Acids Resesearch,* vol. 45, p. e4, 2017.

[114] C. He, M. Zhang and X. Wang, "MICC: an R package for identifying chromatin interactions from ChIA-PET data," *Bioinformatics,* vol. 31, pp. 3832-3834, 2015.

[115] D. H. Phanstiel, A. P. Boyle, N. Heidari and M. P. Snyder, "Mango: a bias-correcting ChIA-PET analysis pipeline," *Bioinformatics,* vol. 31, pp. 3092-3098, 2015.

[116] N. Harmston, E. Ing-Simmons, M. Perry, A. Baresic and B. Lenhard, "GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data," *BMC Genomics,* vol. 16, p. 963, 2015.

[117] E. Yaffe and A. Tanay, "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture," *Nature Genetics,* vol. 43, pp. 1059-1065, 2011.

[118] M. Imakaev, G. Fudenberg, R. P. McCord and et al., "Iterative correction of Hi-C data reveals hallmarks of chromosome organization," *Nature Methods,* vol. 9, pp. 999-1003, 2012.

[119] B. Bonev, N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J. Hugnot, A. Tanay and G. Cavalli, "Multiscale 3D genome rewiring during mouse neural development," *Cell,* vol. 171, no. e524, pp. 557-572, 2017.

[120] J. A. Beagan, T. Gilgenast and et al., "Local genome topology can exhibit an incompletely rewired 3D-folding state during somatic cell reprogramming," *Cell Stem Cell,* vol. 18, pp. 611-24, 2016.

[121] J. H. Kim and et al., "5C-ID: increased resolution Chromosome-Conformation-Capture-Carbon-Copy with in situ 3C and double alternating primer design," *Methods,* vol. 142, pp. 39-46, 2018.

[122] J. H. Sun, L. Zhou and e. al., "Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries," *Cell,* vol. 175, no. 1, pp. 224-238.e15, 2018.

[123] J. A. Beagan, E. D. Pastuzyn, L. R. Fernandez, M. H. Guo, K. Feng, K. R. Titus, H. Chandrashekar, J. D. Shepard and J. E. Phillips-Cremins, "Three-dimensional genome restructuring across timescales of activity-induced neuronal gene expression," *Nature Neuroscience,* vol. 23, pp. 707-717, 2020.

[124] S. W. Flavell and M. E. Greenberg, "Signaling mechanisms linking neuronal activity to gene expression and plasticity of the nervous system," *Annual Review of Neuroscience,* vol. 31, pp. 563-590, 2008.

[125] E. L. Yap and M. E. Greenberg, "Activity-regulated transcription: bridging the gap between neural activity and behavior," *Neuron,* vol. 100, pp. 330-348, 2018.

[126] M. E. Greenberg and E. B. Ziff, "Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene," *Nature,* vol. 311, pp. 433-438, 1984.

[127] R. Muller, R. Bravo, J. Burckhardt and T. Curran, "Induction of c-fos gene and protein by growth factors precedes activation of c-myc," *Nature,* vol. 312, pp. 716-720, 1984.

[128] T. Curran and J. I. Morgan, "Superinduction of c-fos by nerve growth factor in the presence of peripherally active benzodiazepines," *Science,* vol. 229, pp. 1265-1268, 1985.

[129] W. Link and et al., "Somatodendritic expression of an immediate early gene is regulated by synaptic activity," *Proceedings of the National Academy of Science USA,* vol. 92, pp. 5734-5738, 1995.

[130] G. L. Lyford and et al., "Arc, a growth factor and activity-regulated gene, encodes a novel cytoskeleton-associated protein that is enriched in neuronal dendrites," *Neuron,* vol. 14, pp. 433-445, 1995.

[131] T. Fowler, R. Sen and A. L. Roy, "Regulation of primary response genes," *Molecular Cell,* vol. 44, pp. 348-360, 2011.

[132] K. M. Tyssowski and et al., "Different neuronal activity patterns induce different gene expression programs," *Neuron,* vol. 98, pp. 530-546, 2018.

[133] T. Kawashima and et al., "Synaptic activity-responsive element in the Arc/Arg3.1 promoter essential for synapse-to-nucleus signaling in activated neurons," *Proceedings of the National Academy of Science USA,* vol. 106, pp. 316-321, 2009.

[134] S. A. Pintchovski, C. L. Peebles, H. J. Kim, E. Verdin and S. Finkbeiner, "The serum response factor and a putative novel transcription factor regulate expression of the immediate-early gene Arc/Arg3.1 in neurons," *Journal of Neuroscience,* vol. 29, pp. 1525-1537, 2009.

[135] T. K. Kim and et al., "Widespread transcription at neuronal activity-regulated enhancers," *Nature,* vol. 465, pp. 182-187, 2010.

[136] A. N. Malik and et al., "Genome-wide identification and characterization of functional neuronal activity-dependent enhancers," *Nature Neuroscience,* vol. 17, pp. 1330-1339, 2014.

[137] Y. Su and et al., "Neuronal activity modifies the chromatin accessibility landscape in the adult brain," *Nature Neuroscience,* vol. 20, pp. 476-483, 2017.

[138] M. Schardin, T. Cremer, H. D. Hager and M. Lang, "Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories," *Human Genetics,* vol. 71, pp. 281-287, 1985.

[139] E. P. Nora and et al., "Spatial partitioning of the regulatory landscape of the X-inactivation centre," *Nature,* vol. 485, pp. 381-385, 2012.

[140] M. Gabriele and et al., "YY1 Haploinsufficiency causes an intellectual disability syndrome featuring transcriptional and chromatin dysfunction," *American Journal of Human Genetics,* vol. 100, pp. 907-925, 2017.

[141] A. Gregor and et al., "De novo mutations in the genome organizer CTCF cause intellectual disability," *American Journal of Human Genetics,* vol. 93, pp. 124-131, 2013.

[142] T. Hirayama and et al., "CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons," *Cell Reports,* vol. 2, pp. 345-357, 2012.

[143] T. Yamada and et al., "Sensory experience remodels genome architecture in neural circuit to drive motor learning," *Nature ,* vol. 569, pp. 708-713, 2019.

[144] D. W. Straughan and et al., "Evaluation of bicuculline as a GABA antagonist," *Nature,* vol. 233, pp. 352-354, 1971.

[145] T. Narahashi, J. W. Moore and W. R. Scott, "Tetrodotoxin blockage of sodium conductance increase in lobster giant axons," *Journal of General Physiology,* vol. 47, pp. 965-974, 1964.

[146] J. D. Shepherd and R. L. Huganir, "The cell biology of synaptic plasticity: AMPA receptor trafficking," *Annual Review of Cell Developmental Biology,* vol. 23, pp. 613-643, 2007.

[147] J. Y. Joo and et al., "Stimulus-specific combinatorial functionality of neuronal c-fos enhancers," *Nature Neuroscience,* vol. 19, pp. 75-83, 2016.

[148] C. P. Fulco and et al., "Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations," *Nature Genetics,* vol. 51, pp. 1664-1669, 2019.

[149] C. Luscher and R. C. Malenka, "NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD)," *Cold Spring Harbor Perspectives in Biology,* vol. 4, p. a005710, 2012.

[150] H. K. Norton and et al., "Crossed wires: 3D genome misfolding in human disease," *Journal of Cell Biology,* vol. 216, pp. 3441-3452, 2017.

[151] L. R. Fernandez, A. K. Ghosh and R. Radhakrishnan, "Patient-specific Molecular Modeling of Nephroblastoma," *(To be submitted),* 2021.

[152] J. Schlessinger, "Cell Signaling by Receptor Tyrosine Kinases," *Cell,* vol. 103, no. 2, pp. 211-225, 2000.

[153] N. E. Hynes and H. A. Lane, "ERBB receptors and cancer: the complexity of targeted inhibitors," *Nature Reviews Cancer,* vol. 5, pp. 341-354, 2005.

[154] Y. Yarden and M. X. Sliwkowski, "Untangling the ErbB Signalling Network," *Nature Reviews Molecular Cell Biology,* vol. 2, pp. 127-137, 2001.

[155] A. Citri and Y. Yarden, "EGF-ERBB signalling: towards the systems level," *Nature Reviews Molecular Cell Biology,* vol. 7, pp. 505-516, 2006.

[156] S. Banin, L. Moyal, S. Shieh and e. al., "Enhanced phosphorylation of p53 by ATM in response to DNA damage," *Science,* vol. 281, no. 5383, pp. 1674-1677, 1998.

[157] Y. Ziv, D. Bielopolski, Y. Galanty and et al., "Chromatin relaxation in response to DNA double-strand breaks is modulated by a novel ATM- and KAP-1 dependent pathway," *Nature Cell Biology,* vol. 8, pp. 870-876, 2006.

[158] M. J. Paszek, N. Zahir, K. R. Johnson and et al., "Tensional homeostasis and the malignant phenotype," *Cancer Cell,* vol. 8, no. 3, pp. 241-254, 2005.

[159] K. R. Levental, H. Yu, L. Kass and et al., "Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling," *Cell,* vol. 139, no. 5, pp. 891-906, 2009.

[160] J. Swift, I. L. Ivanovska, A. Buxboim and et al., "Nuclear Lamin-A Scales with Tissue Stiffness and Enhances Matrix-Directed Differentiation," *Science,* vol. 341, no. 6149, p. 1240104, 2013.

[161] M. J. Lee, A. S. Ye, A. K. Gardino and et al., "Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks," *Cell,* vol. 149, no. 4, pp. 780-794, 2012.

[162] B. N. Kholodenko, "Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades," *European Journal of Biochemistry,* vol. 267, no. 6, pp. 1583-1588, 2000.

[163] N. C. Institute, "Cancer Query System: SEER Survival Statistics, SEER 9 Relative Survival," 2016. [Online]. Available: https://seer.cancer.gov/canques/survival.html. [Accessed 2020].

[164] W. Zhao, K. Sachsenmeier, L. Zhang, E. Sult, R. E. Hollingsworth and H. Yang, "A New Bliss Indepedence Model to Analyze Drug Combination Data," *Journal of Biomolecular Screening,* vol. 19, no. 5, pp. 817-821, 2014.

[165] M. E. Cunningham, T. D. Klug, J. G. Nuchtem and et al., "Global Disparities in Wilms Tumor," *Journal of Surgical Research,* vol. 247, pp. 34-51, 2020.

[166] J. H. Aldrink, T. E. Heaton, R. Dasgupta and et al., "Update on Wilms tumor," *Journal of Pediatric Surgery,* vol. 54, no. 3, pp. 390-397, 2019.

[167] T. Koshinaga, T. T, T. Oue and et al., "Outcome of renal tumors registered in Japan Wilms Tumor Study-2 (JWiTS-2): A report from the Japan Children's Cancer Group (JCCG)," *Pediatric Blood & Cancer,* vol. 65, no. 7, 2018.

[168] A. M. Termuhlen, J. M. Tersak, Q. Liu and et al., "Twenty-five year follow-up of childhood Wilms tumor: A report from the Childhood Cancer Survivor Study," *Pediatric Blood & Cancer,* vol. 57, no. 7, 2011.

[169] G. J. D'Angio, "Pre- or postoperative therapy for Wilms' tumor?," *Journal of Clinical Oncology,* vol. 26, no. 25, p. 4055, 2008.

[170] N. Graf and R. Furtwangler, "Preoperative chemotherapy and local stage III in nephroblastoma," *Translational Pediatrics,* vol. 3, no. 1, pp. 4-11, 2014.

[171] C. t. f. W. t. C. a. S. standards, "Wang, J; Li, M; Tang, D; et al.,," *World Journal of Pediatric Surgery* , vol. 2, no. 3, p. e000038, 2019.

[172] G. D. Cresswell, J. R. Apps, T. Chagtai and et al., "Intra-Tumor Genetic Heterogeneity in Wilms Tumor: Clonal Evolution and Clinical Implications," *EbioMedicine,* vol. 9, pp. 11-12, 2016.

[173] N. Ludwig, N. Nourkamai-Tutdibi, C. Backes and et al., "Circulating serum mirnas as po- tential biomarkers for nephroblastoma.," *Pediatric Blood and Cancer,* vol. 62, no. 8, pp. 1360-1367, 2015.

[174] F. J. Pérez-Linares, M. Pérezpeña-Diazconti, J. García-Quintana and et al., "MicroRNA Profiling in Wilms Tumor: Identification of Potential Biomarkers," *Frontiers in Pediatrics,* vol. 8, p. 337, 2020.

[175] G. Stamatakos, D. Dionysiou, F. Misichroni and et al., "Computational horizons in cancer (CHIC): Developing meta- and hyper-multiscale models and repositories for in Silico Oncology - A brief technical outline of the project," in *International Advanced Research Workshop on In Silico Oncology and Cancer Investigation*, 2014.

[176] D. B. Dix, C. V. Fernandez, Y. Y. Chi and et al., "Augmentation of therapy for favorable-histology Wilms Tumor with combined loss of heterozygosity of chromosomes 1p and 16q: A report from the Children's Oncology Group studies AREN0532 and AREN0533.," *Journal of Clinical Oncology,* vol. 33, no. 15, p. 10009, 2015.

[177] M. Chintagumpala, "Treatment and prognosis of Wilms tumor," UpToDate, Nov 2020. [Online]. Available: https://www.uptodate.com/contents/treatment-and-prognosis-of-wilms-tumor?search=wilms%20tumor&source=search_result&selectedTitle=2~125&usage_type=default&display_rank=2#H1. [Accessed Nov 2020].

[178] K. Pritchard-Jones, C. Bergeron, B. de Camargo and et al., "Omission of doxorubicin from the treatment of stage II-III, intermediate-risk Wilms' tumour (SIOP WT 2001): an open-label, non-inferiority, randomised controlled trial," *Lancet,* vol. 386, no. 9999, p. 1156, 2015.

[179] J. S. Dome, N. Graf, J. I. Geller and e. al., "Advances in Wilms tumor treatment and biology: Progress through international collaboration," *Journal of Clinical Oncology,* vol. 33, p. 2999, 2015.

[180] S. J. McMahon, "The linear quadtratic model: usage, interpretation and challenges," *Physics in Medicine & Biology,* vol. 64, p. 01TR01, 2018.

[181] C. M. van Leeuwan, A. L. Oei, J. Crezee and et al., "The alfa and beta of tumours: a review of parameters of the linear-quadratic model, derived from clinical radiotherapy studies," *Radiation Oncology,* vol. 13, p. 96, 2018.

[182] A. Ghosh and R. Radhakrishnan, "Heterogeneous multi-scale framework for cancer systems models and clinical applications," in *Proceedings of Mathematical Oncology Meeting*, Portland, OR, 2019.

[183] J. R. Karr, J. C. Sanghvi, D. N. Macklin and et al., "A whole-cell computational model predicts phenotype from genotype," *Cell,* vol. 150, no. 2, pp. 389-401, 2012.

[184] H.-Y. Huang, Y.-C.-D. Lin, J. Li and et al., "miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database," *Nucleic Acids Research,* vol. 48, no. D1, pp. D148-D154, 2020.

[185] V. Cristini and J. Lowengrub, Multiscale Modeling of Cancer, Cambridge University Press, 2010.

[186] H. Pimentel and et al., "Differential analysis of RNA-seq incorporating quantification uncertainty," *Nature Methods* , vol. 14, pp. 687-690, 2017.

[187] N. Krietenstein and et al., "Ultrastructural details of mammalian chromosome architecture," *Preprint at bioRxiv,* 2019.

[188] M. I. Love, W. Huber and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology,* vol. 15, p. 550, 2014.

[189] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics,* vol. 9, pp. 321-332, 2008.

[190] E. Weinen, B. Engquist and Z. Huang, "Heterogeneous multiscale method: A general methodology for multiscale modeling," *Physical Review B,* vol. 67, p. 092101, 2003.

[191] C. M. van Leeuwan, J. Crezee, A. Bel and e. al., "The alfa and beta of tumours: a review of parameters of the linear-quadratic model, derived from clinical radiotherapy studies," *Radiation Oncology,* vol. 13, p. 96, 2018.

[192] H. D. Thames, K. K. Ang, F. A. Stewart and e. al., "Does incomplement repair explain the apparent failure of the basic LQ model to predict spinal cord and kidney responses to low doses per fraction?," *International Journal of Radiation Biology,* vol. 54, no. 1, pp. 13-19, 1988.

[193] J. W. Schneider, Z. Gao, S. Li and e. al., "Small-molecule activation of neuronal cell fate," *Nature Chemical Biology ,* vol. 4, pp. 408-410, 2008.

[194] A. Bauer-Mehren, L. I. Furlong, M. Rautschka and F. Sanz, "From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways," *BMC Bioinformatics,* vol. 10, p. S6, 2009.