Methods for Analyzing the 4D Nucleome, with Application to Cellular Reprogramming

by

Scott Ronquist

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2019

Doctoral Committee:

Associate Professor Indika Rajapakse, Chair
Professor Daniel M. Burns
Associate Professor Ryan E. Mills
Affiliated Professor Roger S. Newton
Professor Max S. Wicha

Scott Ronquist

scotronq@umich.edu

ORCID iD: 0000-0002-6514-9194

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor Indika Rajapakse for always believing in me. Your passion for science is inspiring and reassures me that I made the right decision pursuing a PhD. I would also like to thank my thesis committee members Max Wicha, Ryan Mills, Daniel Burns, and Roger Newton for their scientific guidance during my graduate work.

I owe a deep debt of gratitude to the Rajapakse lab members, former and present, for our scientific discussions that guided my work: Walter Meixner, Haiming Chen, Lindsey Muir, Jie Chen, Laura Seaman, Geoff Patterson, Sijia Liu, Stephen Lindsly, Gabrielle Dotson, Can Chen, Wenlong Jia, and Emily Crosette. I would also like to thank my collaborators: Thomas Ried, Markus Brown, Rudiger Meyer, Darawalee Wangsa, John Snyder, Roger Brockett, Erdogan Gulari, and Alnawaz Rehemtulla. Without you all none of this would be possible. Additionally, I would like to thank the Bioinformatics graduate program coordinators and staff for keeping me on track to graduate.

Finally, I would like to thank my parents, Ronald and Maureen Ronquist, for their love and support throughout my graduate studies. Our regular conversations helped keep me level during periods of heavy stress. I am so incredibly fortunate to have both of you as my parents.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS

2D two-dimensional

3D three-dimensional

4DN 4D Nucleome

BAC bacterial artificial chromosome

CAGE-seq capped analysis of gene expression sequencing

ChIP-seq chromatin immunoprecipitation sequencing

CSC cancer stem cell

CT chromosome territory

DGC data guided control

DOF degrees of freedom

ESC embryonic stem cell

FISH fluorescent in situ hybridization

FPKM fragments per kilobase of transcript per million mapped reads

GO gene ontology

GRN gene regulatory network

HCEC human colonic epithelial cell

Hi-C genome wide chromosome conformation capture

iPSC induced pluripotent stem cell

kb kilobase

Mb megabase

MCD mean closest distance

MVE minimum volume ellipsoid

PCA principal component analysis

RE restriction enzyme

RNA-seq RNA sequencing

RPKM reads per kilobase of transcript per million mapped reads

SKY spectral karyotyping

TAD topologically associating domain

TF transcription factor

TFBS transcription factor binding site

VNE von Neumann entropy

ABSTRACT

The dynamical relationship between chromatin structure, gene transcription, and cellular phenotype is referred to as the 4D Nucleome (4DN) [141]. 4DN data analysis has gained significant research attention in recent years for its ability to illuminate cell regulation principles [47]. Despite its benefits, 4DN analysis can be difficult: data sets are large, analysis methods are underdeveloped, and data analysis requires a high-level understanding of biology, computer science, and mathematics. In this dissertation, I present novel analysis methods that address these issues. Using these methods, I help uncover 4DN relationships in the cell cycle, the circadian rhythm cycle, and cellular reprogramming.

First, I investigate methods for analyzing time series cellular images. In this work, I demonstrate how the detection of a "canonical framework," or a consistent genomic coordinate system both within and between time points, can help researchers elucidate latent patterns within cellular imaging data. Next, I examine 4DN data collected on proliferating human fibroblasts. Dynamical structure-function relationships between genes are uncovered in known gene modules, including cell cycle, circadian rhythm, and wound healing gene networks. Building upon this data set, I helped develop an algorithm for cellular reprogramming. This algorithm models the dynamics of human fibroblast proliferation and determines where and when to input control using transcription factors (TFs) to achieve optimal reprogramming efficiency. TFs known to successfully reprogram between cell types are recovered using this algorithm, thus validating the predictions. This includes the prediction of MYOD1 for fibroblast to

muscle cell reprogramming. Next, I analyze 4DN data collected on fibroblasts undergoing MYOD1-mediated reprogramming to muscle cells to reveal previously unknown relationships between structure, function, and biological processes. A connection between MYOD1 and the core circadian clock gene network is also uncovered in this work. I continue with a 4DN analysis of cancer cells, including colorectal and breast cancer cells, where changes in structure and function show a clear relationship with their disease state. I conclude with the description of a software toolbox containing novel methods for the analysis of 4DN data. Collectively, this work provides a comprehensive guide for the analysis of 4DN data, spanning a range of experimental methods and applications.

CHAPTER I

Introduction

1.1 Research overview

This chapter is an overview of the content covered within this dissertation, and a review of background information pertinent to this work.

The 4D Nucleome (4DN) refers to the relationship between genome structure, genome function, and cellular phenotype over time [141]. 4DN data analysis has gained significant research attention in recent years for its ability to illuminate cell regulation principles [47]. Despite its benefits, 4DN analysis can be difficult: data sets are often large, analysis methods are underdeveloped, and data analysis requires a high-level understanding of biology, computer science, and mathematics.

There are many techniques currently available now to observe genome structure and function. One of the first methods developed to observe nuclear structures was cell imaging. Fluorescent *in situ* Hybridization (FISH) imaging is a nucleus imaging technique used to determine the location of genomic loci in the nucleus [148]. Advances in the field of microscopy and improvements in cell imaging protocols allow for the detection of multiple genomic regions in the same nucleus at high resolution. Additionally, whole chromosomes can be "painted" to determine chromosome shape and position within the nucleus [142]. A biochemical technique, genome-wide chromosome conformation capture (Hi-C) allows researchers to determine how often any

two genomic loci contact each other in three-dimensional (3D) space [102]. Hi-C data analysis can uncover structural phenomena on many scales: from promoter-enhancer loops to chromatin compartments and translocations [94, 154]. Pairing this analysis with cell imaging can validate these findings. Furthermore, combining these structural measurements with functional measures helps in understanding cell phenotype. Popular methods used to analyze genome function include RNA sequencing (RNA-seq) and proteomics, which quantify the abundance of RNA and protein, respectfully [118]. Finally, since cells are a dynamical system, a time course of study must be employed to perform 4DN analysis and fully understand how a cell operates.

In this dissertation, I present my work towards advancing the field of 4DN research. This includes the analysis of novel 4DN data sets, and the description of novel methods to perform this analysis. The remainder of this research overview will summarize the concepts and methods to be discussed in subsequent chapters.

Chapter II, "Insight into Dynamic Genome Imaging," focuses on time series FISH imaging data and methods to determine a consistent genomic coordinate system for all cells in all time points [146]. This consistent coordinate system is known as a "canonical framework." When cells are imaged by a microscope, their orientation relative to the microscope lens is random. This makes it difficult to determine how genes move relative to other genes, and relative to the nucleus edge, over time. The movement of genes within the nucleus has been found to affect cell phenotype, thus characterizing this movement is important for understanding how a cell functions [31, 134, 180]. By fitting an ellipsoid to the nucleus shape and determining the position of genes relative to the ellipsoid, a modified Procrustes analysis can be used to determine the optimal translation, rotation, and reflection of the nucleus to minimize the distance between genes in each nucleus. These changes in nucleus orientation (e.g. translation, rotation, and reflection) place the nuclei in the most likely canonical framework, making time series image analysis much easier to interpret. To demonstrate this,

data is analyzed from two separate experiments: proliferating human fibroblasts and human fibroblasts undergoing MYOD1-mediated reprogramming. This analysis uncovers previously unknown movements in gene positioning over time. Computer code is provided to perform this analysis, and extensions to 3D image analysis are discussed within this chapter.

Chapter III, "Functional Organization of the Human 4D Nucleome," examines time series Hi-C, FISH, and RNA-seq data collected on proliferating human fibroblasts [31]. Prior to data collection, cells are synchronized in their circadian rhythm and cell cycle phase. The circadian rhythm cycle is a major orchestrator of cell function, affecting up 70% of genes in some cell types [190]. Analysis of circadian gene expression and gene positioning over time reveals a pattern between the genes *PER2* and *CLOCK*, where expression of each correlates with their relative distance between each other. Furthermore, examination of transcription factor binding sites (TFBS) near genes with correlated structure and function over time reveals novel gene networks. These networks help us understand how genes communicate within the genome, and can assist in determining the downstream consequences of targeting certain transcription factors (TF) by drugs. This work demonstrates the power of 4DN analysis methods for uncovering latent genome regulation principles.

Chapter IV, "Algorithm for Cellular Reprogramming," describes an algorithm for determining TFs that can be used for cellular reprogramming [147]. This work has broad ranging applications in medicine, such as autologous transplantation and personalized medicine drug screening. First, the natural dynamics of a cell type is modelled using 4DN data (Hi-C and RNA-seq). Here we use the same data set collected and analyzed in Chapter III. Then, TFBS information and chromatin accessibility data is used to determine where a TF can influence the initial cell type's function. Using control theory equations, we determine the optimal timing, combination, and relative amount of TFs that can catalyze reprogramming from the initial

cell type towards a target cell type. This algorithm correctly predicts known reprogramming TFs, thus bypassing expensive and time-consuming experimental methods for discovery.

Chapter V, "MYOD1-mediated Fibroblast to Muscle Reprogramming," revisits Dr. Harold Weintraub's original cellular reprogramming experiment, MYOD1-mediated human fibroblast to muscle cell reprogramming [184]. Hi-C and RNA-seq data is collected every 8 hours post-MYOD1 introduction into the nucleus. From these data, substantial changes in genome structure are seen prior to changes in function [103]. Whether genome structure or function changes first has been speculated upon, but was previously unknown. Analysis of muscle specific super-enhancer regions reveals a significant increase in Hi-C contacts prior to an increase in proximal muscle specific gene expression. Additionally, robust synchronization of core circadian rhythm genes is found post-MYOD1 introduction, revealing a novel relationship between MYOD1 and circadian rhythms. This is the most comprehensive high-resolution data set available for understanding cellular reprogramming principles, and the analysis of this data set demonstrates the power of this experimental approach.

Chapter VI, "The 4DN of Cancer," covers methods for analyzing 4DN data collected from cancer cells. Two distinct cancer cell lines are analyzed here: colorectal cancer cells and breast cancer cells. In the colorectal cancer cell analysis, healthy and pre-cancerous colorectal cells (carrying an extra copy of chromosome 7) are compared using Hi-C, RNA-seq, and cell imaging. Striking structural and functional differences can be observed genome-wide between the two samples, specifically in regions that contain genes related to cancer. From this analysis, we theorize that specific changes in the HGF/MET signaling pathway are affected early on in the progression of colorectal cancer. In the breast cancer cell analysis, cells are sorted into Cancer Stem Cell (CSC) and non-CSC subpopulations based on the expression of *ALDH1A1*. Hi-C

and RNA-seq data is extracted and analyzed from these samples revealing genome-wide differences between them. The characterization of these subpopulations has implications for cancer treatment strategies.

Chapter VII, "4DNvestigator," describes the 4DNvestigator, an user-friendly toolbox for the analysis of 4DN data. With the increase in 4DN data sets, the development of methods to properly analyze these data is imperative. Given time series Hi-C and RNA-seq data, the 4DNvestigator can determine genomic regions that change significantly over time, using both established and novel methods. The toolbox introduces new methods for comparing Hi-C matrices, providing statistical measures to determine the significance of the Hi-C matrix differences. This work allows both novice and experienced researchers to analyze 4DN data in a principled manner.

Together, the experiments and methods discussed within this dissertation provide a framework for comprehensively analyzing 4DN data, with applications for controlling cell function. This work has implications in cell and tissue regeneration, personalized medicine drug screening, and cancer treatment strategies.

1.2 Hi-C and RNA-seq

Chromatin folding and orientation, or "genome structure," can affect genome function [47]. The relationship between genome structure and function can be observed on many scales. Enhancer and promoter regions, ranging from a few kilobases (kb) to over 1 megabase (Mb) in distance, contact each other in 3D space to control gene expression [117]. Chromosomes occupy distinct regions within the nucleus, referred to as chromosome territories (CT), and have a preferences for how close they are to the nucleus periphery [40]. A gene's distance to the nuclear periphery is correlated with expression changes; often, genes close to the nuclear periphery are transcriptionally silenced [176]. Considering the size ($>$ 3 billion bp) of the human genome, observing structure genome-wide is a difficult task. Only in the past decade have researchers

been able to explore genome structure extensively through genome-wide chromosome conformation capture, or Hi-C.

Hi-C measures whether two genomic loci come close together in 3D space [102]. This technique gets its name based on its predecessor methods: chromosome conformation capture (3C), chromosome conformation capture-on-chip (4C), and chromosome conformation capture carbon copy (5C). The original method developed in 2002, 3C, was able to determine chromosome contacts between two pre-specified regions [46]. Each subsequent advancement added more range to the number of regions observable within an experiment, culminating in Hi-C which could capture and quantify nearly all genomic contacts within a cell. Hi-C, first developed in 2009 on a population of cells, is obtained via the following steps: cross-link regions of proximal chromatin using formaldehyde fixation, fragment the genome using restriction enzyme (RE) digestion, re-ligate cross-linked loci, then sequence each end of the re-ligated DNA via high-throughput sequencing. The sequenced reads can then be mapped back to a reference genome to determine their genomic origin. Since its original conception, several methodological improvements have been made to reduce the number of input cells, increase the number of genomic contact incidences, and improve the spatial genomic resolution [139]. Recent advances also allow for single cell Hi-C matrices to be obtained [138]. A typical population-level Hi-C experiment will result in over 100 million paired-end sequence reads, where the majority of identified "contacts" come from loci that are close together in the 1D genome, but many significant long-distance contacts can be observed. These data can be constructed into a contact matrix, $\mathbf{A}$, where the number of contacts between locus i and locus j resides in element a_{ij}. This contact matrix can be constructed at many resolutions (typically 5 kb-1 Mb).

Analysis of Hi-C contact matrices has revealed principles of higher-order genome structure. The first principal component of the intra-chromosomal correlation matrix of the contact matrix was shown to segregate the genome into two distinct compart-

ments: compartments A (generally active, euchromatin) and B (generally inactive, heterochromatin) [102]. Loci within a compartment are in contact more with each other than with loci in the alternate compartment. Further analysis revealed regions in the genome that have a high number of contacts and tend to have similar expression, called "topologically associating domains" (TADs), for which boundaries are relatively cell-type invariant [31, 53]. Breakdown of TAD boundaries has been linked to phenotypic defects in gene function, and chromosomal aberrations can be determined with high precision through Hi-C [26, 108].

RNA-seq measures the RNA abundance within a sample. This technique was first discovered in 2008, following the advancement of next generation sequencing (NGS) machines [118]. NGS refers to technology that can sequence millions of DNA fragments in parallel. A typical RNA-seq experiment will isolate and collect RNA from a population of cells, reverse transcribe the RNA to DNA (cDNA), sequence the cDNA, then map the sequenced reads to a reference genome. RNA isolation is achieved via DNase digestion, which degrades DNA, but leaves RNA sequences intact. Reverse transcription is used to convert the remaining RNA to its DNA counterpart, which is a generally more stable molecule. These cDNA sequences are fragmented so that they can be passed to a high-thoughput sequencer. RNA-seq can be performed using 50 bp single-end sequencing, which contain enough information to map unique genomic regions, such as gene coding sequences. Paired-end sequencing can be employed to more easily distinguish gene splice variants, which can perform distinctly different functions within the cells [173]. RNA-seq has also recently been developed for single cell resolution. Quantifying the number of RNA fragments derived from each gene creates a functional signature for the cell, and cell type specific genes can be used to identify the cell type.

1.3 The 4D Nucleome

The field of 4DN research was formalized in 2013 to bring together experts from a wide range of backgrounds, including: biology, computer science, and mathematics [141]. Before the conception of the 4DN, few concerted efforts had been made to combine the analysis of genome structure, function, and phenotype over time. This was understandable, as many researchers devote their entire career to comprehending just one subsection of this field. A genuine challenge in 4DN analysis is familiarizing researchers to the ideology and terminology of alternate fields. To support this goal and advance 4DN research, the NIH created the "4D Nucleome" as a Common Fund initiative [47].

Since the 4DN's conception, many cell regulation principles have been discovered. A general framework for the analysis of genome structure and function over time was laid out in Chen *et al.* [31]. The use of phase planes is a natural mathematical way to visualize structure and function changes over time. Pioneering work by Lupiáñez *et al.* [108] discovered a link between gene sequence, TAD boundaries, nearby gene expression, and disease. In this work, a genetic mutation in a TAD boundary on chromosome 2 caused misregulation of nearby gene expression, leading to limb malformations such as F-syndrome, polydactyly, and brachydactyly. Dixon *et al.* advanced our understanding of early stage stem cell differentiation in a 2015 paper [52]. The analysis of Hi-C and RNA-seq for six distinct cell states, embryonic stem cells (ESC) and five lineage specific stem cells and progenitors, revealed that significant structural and functional changes can be observed in genomic regions that define the cell lineage. Cancer research has also benefited from 4DN analysis. Work by Seaman *et al.* revealed that the correlation between chromatin structure and function increases at the site of translocation in colorectal cancer cells. Without the combined analysis of both structure and function over time, these discoveries would not be possible.

1.4 Cell imaging

Cell imaging is a relatively inexpensive way to observe aspects of cell structure on a single cell level. This field can be broadly divided into two subcategories: light microscopy and fluorescence microscopy. Light microscopy uses visible light and lenses to enlarge images. This technique is straightforward to perform and can characterize cell morphology and count healthy cells in a culture to determine if cells are surviving culture conditions. Fluorescence microscopy uses high intensity light to excite fluorescent material within the object you are studying. For instance, immunofluorescent (IF) imaging determines the location of specific biomolecules within the cell, after fluorescent antibodies have bound to an antigen [150]. Distinct from IF, more general "fluorescent stains" bind to non-protein cellular molecules. The popular fluorescent stain DAPI binds to AT-rich regions of DNA, and is used to illuminate the cell nucleus. AT-rich regions of DNA are often gene-poor heterochromatin regions, so the fluorescent intensity of DAPI signal in a region can be used to infer chromatin accessibility.

The Fluorescent *in situ* Hybridization (FISH) imaging technique targets specific nucleic acid sequences (e.g. genomic loci or RNA) within the cell, and was first described in 1977 [148]. To target these sequences, short fluorescent DNA fragments, which have complementary sequences to the targeted region, are introduced into the cell. The cell is then subjected to high temperature conditions which causes DNA denaturing, allowing the probes to bind to the target regions within the genome. Excess probes are washed away before imaging, leaving only probes that have successfully bound to their targets. Probes are designed to target sequences of DNA which are unique within the genome to ensure specificity of binding. The targeted region must have a high concentration of unique DNA sequences to allow multiple probes to bind within the regions, resulting in a strong fluorescent signal. There is a trade-off between the percentage of unique DNA within the target region and the tar-

get region length; if the intended target region has a low percentage of unique DNA, the region is often extended to allow for more probes to bind nearby. 3D images can be constructed by taking "z-stack images," where multiple images are taken while incrementally moving the microscope lens in the vertical axis.

FISH analysis has its advantages and disadvantages. One advantage is that FISH is inherently single cell resolution, which can provide insight into determining gene position variability. However, a large number of FISH images may need to be collected to see real trends in the data. Analyzing this imaging data brings its own set of difficulties: large data sets, experimental noise, and variable fluorescent signal are common issues for researchers [153]. Automated analysis programs are necessary to explore a large number of cells and metrics (such as distance between genomic regions) in a reasonable amount of time. It is often difficult to tune parameters correctly to extract the right features from inherently noisy microscopy data. Extending this analysis to 3D images increases the time needed for collection and the potential for experimental noise.

1.5 Cellular reprogramming

Cellular reprogramming is the ability to convert an initial cell type into an alternate target cell type. The first official demonstration of cellular reprogramming was performed by Dr. John Gurdon in 1962, when Dr. Gurdon demonstrated that differentiated cells (cells that are in their final cell type) could be reprogrammed to a pluripotent state (a cell type with the ability to differentiate into alternate cell types) by transferring the nucleus of a differentiated cell type into an enucleated egg (an egg with the nucleus removed) [70]. This process produced only a few reprogrammed cells, since each cell needed to be handled individually, and the process was marred by reprogramming inefficiencies.

This idea was expanded upon in 1989 by Dr. Harold Weintraub, who used TFs to

mediate cell reprogramming. In this experiment, Dr. Weintraub converted a human fibroblast into a muscle cell via addition of the TF MYOD1 [184]. MYOD1 was a well-known regulator of the myogenic lineage, only active in muscle cells, but its ability to completely take over the function of an unrelated cell type was surprising. MYOD1 was able to reprogram several alternate cell types into muscle cells and became one of the first known "master regulators" of cell fate. Since Dr. Weintraub's discovery, a number of alternate cell reprogramming experiments were discovered, but the integrity of reprogramming, and whether the reprogrammed cells could really function like their natural counterpart, was unknown.

Cellular reprogramming encountered a paradigm shift in 2007, when Dr. Shinya Yamanaka converted a human fibroblast into an ESC-like state [166]. This cell type, named an induced pluripotent stem cell (iPSC), is nearly indistinguishable from an ESC. Additionally, iPSCs have the ability to differentiate into alternate cell types, much like ESCs. To achieve this conversion, Dr. Yamanaka transduced fibroblasts with a combination of four TFs: OCT4, SOX2, KLF4, and MYC. These factors have collectively come to be known as the "Yamanaka factors." This discovery was especially timely, since the use of natural ESCs for therapeutic research was banned by many countries, and this discovery circumvented these restrictions. Dr. Yamanaka, as well as the cellular reprogramming pioneer Dr. Gurdon, received a Nobel Prize for this work in 2012.

Since these discoveries, the field of cellular reprogramming has expanded rapidly. Cellular reprogramming experiments have been discovered for many target cell types such as neuron, hepatocyte, and cardiac progenitor [81]. Cellular reprogramming has been performed *in vivo*, by reprogramming cardiac fibroblasts into cardiac myocytes in mice [132]. The first human clinical trial using iPSCs was approved in 2012 for use in macular degeneration treatment [111]. Though results from this trial showed limited improvement to condition, no conditions were worsened because of this treatment.

While recent progress within the field has been substantial, predicting which TFs will be successful for reprogramming is a limiting factor. Many papers within this field are the result of years of "guess-and-check" experimentation, where researchers try multiple different TFs on cell populations and report on which methods work. Researchers can make educated guesses for which TFs will work based on cell type specific TF activity in the target cell type, but this does not always result in a successful reprogramming experiment. For example in Dr. Yamanaka's original experiment, he started with 24 genes that are highly over-expressed in ESCs, transduced combinations of the 24 genes into a population of cells, then began removing unnecessary genes in subsequent experiments until arriving at the minimum necessary gene set [165]. Considering that iPSC generation takes $\sim$3 months to create in each experiment, this is painstaking work that drains valuable time and resources [164].

To circumvent these problems, researchers have begun to create algorithms that can predict successful reprogramming TFs from data. Many of these algorithms rely on methods for quantifying RNA-seq differential expression between cell types, such as mutual information and Jensen-Shannon divergence [24, 44]. More recent algorithms explore TF binding networks and CAGE-seq data to refine their predictions [133]. These algorithms greatly reduce the cost of discovering TFs for reprogramming, but no algorithm is 100% accurate. Additional genomic features, such as structure, epigenetics, and dynamics should be considered to achieve higher prediction accuracy.

1.6 Mathematics of 4DN

4DN analysis is challenging due to the large amount of data that is generated for each experiment. Hi-C samples often contain over 100 million contacts that define the interactions between 3 billion bps (in human samples). These data sets are often >2 Gb in size once processed. RNA-seq data is one-dimensional and usually smaller than Hi-C data, but still typically requires >20 million sequenced reads.

Imaging data sets are notoriously bulky, and many z-stack microscope images must be collected to achieve a good sample size. Additionally, all these metrics scale linearly with additional time points, replicates, and samples. To make sense of these high-dimensional data sets, rigorous mathematical techniques must be developed to quantify our findings.

A mathematical operator useful for the analysis of Hi-C data is the graph Laplacian. The Laplacian quantifies the diffusion between interacting entities. As noted in Section 1.2, Hi-C data can be viewed as an adjacency matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$. Similarly, Hi-C can be viewed as a graph, $\mathcal{G}(V, E)$, where the nodes, V, refer to the n genomic bins in the adjacency matrix and the edges, E, are given by the number contacts, $a_{i,j}$. Given $\mathbf{A}$, the unnormalized Laplacian matrix is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A},$$

where $\mathbf{D}$ is the degree matrix, $\mathbf{D} = diag(d_1, \ldots, d_n)$ and $d_i = \sum_{i=1}^{n} a_{i,j}$. The normalized Laplacian matrix is defined as

$$\overline{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}},$$

where the eigenvalues λ of $\overline{\mathbf{L}}$ are bounded by $0 = \lambda_1 \leq \cdots \leq \lambda_n \leq 2$. Both the normalized Laplacian and unnormalized Laplacian have properties that are useful for spectral clustering and graph partitioning [179]. Specifically, the normalized Laplacian of the Hi-C adjacency matrix has been used to partition the genome into biologically relevant clusters. The second smallest eigenvalue of $\overline{\mathbf{L}}$ is called the Fiedler number. The eigenvector associated with the Fiedler number is called the Fielder vector. The Fielder vector can be used to partition the graph based on the sign $(+/-)$ of its elements [161]. This partitioning method finds a balance between minimizing the weights of the edges that are "cut" to create the partitioning, and maximizing

the weights of the edges within each cluster [179]. From a biological perspective, this partitioning correlates strongly with heterochromatin and euchromatin regions [31]. Analysis of how the Fielder vector changes over time identifies regions that change from heterochromatin to euchromatin, or vice versa. Additional partitioning of sub-regions that have a large Fielder number can be used to detect TAD boundaries [32].

A graph perspective for Hi-C data analysis is also useful for the extraction of centrality measures. Centrality measures quantify node importance in a graph. Several centrality measures exist, each of which quantifies a different type of nodal influence. For example, the degree centrality of a node is measured by summing the edge weights of all edges connected to a node; recall the degree matrix $\mathbf{D}$, where $d_i = \sum_{i=1}^{n} a_{i,j}$. This measure summarizes the local connectedness of a node. Alternative centrality measures focus on global network properties, such as how often a node is present along the shortest path between any two other nodes (betweenness centrality). These measurements, when applied to Hi-C matrices, have been linked with biological phenomena [103].

With an understanding of how the genome changes in structure and function over time, we can develop methods to control the genome [137]. From a control theory perspective, we can formalize the problem as such: the system to be controlled is the cell, the sensors are experimental data such as Hi-C and RNA-seq, and the controllers are chemicals that affect genome structure or function in predictable ways. A natural controller of the genome is a TF, since TFs bind to specific genomic regions and change nearby structure and function. Given information on where TFs bind and how they affect local genomic regions, TFs can be used to control the genome and steer the cell state towards a desired target phenotype.

Another useful measurement for Hi-C and RNA-seq data analysis is entropy. Entropy can be thought of as measuring "uncertainty" in a system by quantifying the

number of possible states a system (cell) can be in, and determining the probability that the system is in any given state. Several fields have found this measurement useful, from thermodynamics to information theory [38]. In the context of cell biology, the cell state (e.g. cell type, cell cycle phase, etc.) can be defined using RNA-seq, and entropy measures our uncertainty in knowing which genes will be expressed. Since ESCs have highly variable gene expression, we are more uncertain in its expression signature. One specific measure of entropy, Shannon entropy S, is defined as

$$S = -\sum_i p_i \log(p_i)$$

where p_i represents the probability that the system (cell or group of cells) is in state i. Recently, RNA-seq and modified entropy equations were used to quantify stemness [69, 171].

Related to gene expression variability, another hallmark of stem cells is their relatively high percentage of euchromatin compared to differentiated cells [22]. Euchromatin is less compact and therefore has more freedom to move within the nucleus. Freedom of chromatin movement allows for more choices in chromatin contacts in Hi-C, and therefore more "uncertainty" in cell state, which can be quantified by entropy. Since Hi-C is a multivariate analysis measurement (each contact coincidence involves two variables, the two loci), we use Von Neumann entropy (VNE) as follows:

$$\text{VNE} = -\sum_i \lambda_i \log(\lambda_i)$$

In this application, the eigenvalues are derived from either the correlation matrix or the Laplacian matrix so that all eigenvalues are positive. Furthermore, the eigenvalues are normalized to sum to 1. VNE summarizes the eigenspectrum of the Hi-C matrix, which tells us the matrix complexity; for matrices with a single large eigenvalue, we are more "certain" in its conformation.

In the context of clinical applications, there is an important connection to made between entropy and controllability. In control theory, systems that have high entropy are generally thought to be more controllable [137]. Worded a different way, systems that are more uncertain in their structure are more easily influenced by external forces. This means that given the right inputs that can affect the cell state, such as small molecules or TFs, cells with high entropy/stemness could be influenced while affecting differentiated cells to a lesser degree. Since at its core cancer is a disease of aberrant reprogramming, targeted re-reprogramming of cancer cells, specifically CSCs, may be key for treatment. Considering the significant mutations and abnormal karyotype of nearly all cancer cells, a reasonable target state for reprogramming would be toward apoptosis. The issue now becomes determining appropriate inputs for a given cell population. Prediction of inputs that can control cell fate is an active research field, and this work may find significant application for reprogramming cancer stem cells to a vulnerable state [133, 147].

CHAPTER II

Insight into Dynamic Genome Imaging

This chapter is based on a paper by Scott Ronquist, Walter Meixner, Indika Rajapakse, and John Snyder [146]. I have summarized my contributions to this paper here.

2.1 Abstract

The dynamic nature of the human genome complicates FISH imaging analysis. Time series FISH imaging has enhanced our understanding of genome structure, but due to cell state and experimental variability, this data is often noisy and difficult to analyze. Furthermore, computational analysis techniques are needed for homolog discrimination and canonical framework detection, in the case of time-series images. In this paper, we introduce novel ideas for nucleus imaging analysis, present findings extracted using dynamic genome imaging, and propose an objective algorithm for high-throughput time-series FISH imaging. While a canonical framework could not be detected beyond statistical significance in the analyzed data set, a mathematical framework for detection has been outlined with extension to 3D image analysis.

2.2 Introduction

The human genome can be seen as the blueprint for cellular activity, encoding nearly all the genes that human cells use to function. While researchers have developed tools to decode the genome's linear sequence, there are aspects that remain incompletely understood, such as location, organization and dynamic behavior. The genome is packed into a nucleus averaging $374\mu m^3$ in volume, though this measurement is highly variable depending on cell-type and cell cycle stage. Within this nucleus, the genome contains 3.2 billion base pairs of DNA totaling 2m in end-to-end length in G_0 and G_1 and (6.4 billion bps with 4m in late S-phase through G_2) [182]. This packing exhibits organized spatial structure. Chromosomes are confined to CTs which occupy particular locations in relation to the nucleus edge and to each other [39, 136]. At a smaller scale, genes within each chromosome also inhabit preferred locations. Gene locations vary as a function of cell type; forced relocation leads to misregulation of gene expression [115, 134, 156]. The system is also dynamic; changes in gene position over time correlate with gene expression [31].

Microscopy techniques have been developed to image the genome and interrogate its structure. Specifically, fluorescent in situ hybridization, or "FISH", is used to locate specific genetic material within the nucleus. FISH works by introducing fluorescent nucleotide sequences to the cell (a "probe"), allowing these probes to bind to the cell chromatin through hybridization, and imaging the probe's binding location through laser excitation. If the probe nucleotide sequence has a unique complementary sequence within the genome, binding specificity can be assured. Since its invention in 1980, FISH has become a standard method for researchers and clinicians alike [7, 99, 130]. Recently, FISH has been enhanced to visualize multiple DNA and RNA locations simultaneously, improve loci imaging specificity, and increase image resolution [120].

For probe design, the main techniques are bacterial artificial chromosomes (BACs)

and oligoprobes [9, 159]. To construct a BAC probe, the desired loci is extracted from a species-specific cell, integrated and exponentially duplicated within a bacterial culture, and eventually extracted and nick-translated with a fluorescent molecule [72]. These probes are often relatively large (150 kb), but thereby have high binding specificity. Oligoprobes are designed based on the species reference genome, synthetically created, and then PCR amplified [9]. The clear advantage of this technique is that shorter fragments can be imaged, as long as the sequences are still unique. Often multiple $\sim$40 bp probes will be designed to cover the full length of a loci of interest. For both BACs and oligoprobes, the number of different fluorescent molecules available and microscope laser lines has continued to increase allowing for visualization of up to 5 distinct loci per image in standard FISH, with more points distinguishable through more advanced techniques such as spectral barcoding [98]. The image resolution is inherently diffraction limited (typically $\sim 0.2-0.25\mu$m x-y, $\sim 0.4\mu$m z), without the help of emerging super-resolution image restoration technology [63].

While recent experimental advances such as super-resolution and RNA-FISH have increased the power of FISH assays, extracting relevant information is complicated and computational analysis is needed to make sense of the large amount of data. FISH imaging is often performed on a population of cells to mitigate the effects of undesired variables, e.g. in cell orientation, cell cycle state, and growth conditions, as well as from imaging noise. Manual analysis compounds errors and is prohibitively slow when hundreds of cells are needed to achieve statistical significance. As an attempt to solve these problems, researchers have employed computer programs to automatically analyze FISH images and obtain fast, unbiased results. Fortunately, many image processing techniques have been developed in the field of computer science to aid in the examination of this noisy data, but for best results these programs must have knowledge of cell biology and image collection methods [144]. Unfortunately, no one program has emerged as robust enough to handle any given nucleus imaging experi-

ment, while also being high-throughput in the number of cells detected. This is partly a product of the large number of different FISH assays available, exploring different genomic areas of different cell types. For each specific FISH assay image, additional biological information such as gene copy number in each cell-cycle state (to know how many points to look for), nucleus and probe size (to determine the size of objects to recognize), and wavelength spectrum of fluorescent probes (to help discriminate between overlapping fluorescent spectra) can greatly enhance FISH analysis. These programs can then be used to extract useful information on gene location, number of genes detected per nuclei, gene-gene distances, and nucleus shape fitting.

Automatic nucleus image analysis was first attempted in 1996, by counting the number of dots per cell nucleus on acquired images [124]. This algorithm was broken down into 4 main processes: extract nucleus-containing regions, determine nucleus size and shape, find dots within the image, and count the number of dots. Statistics on nucleus shape distance between points, and number of points compared favorably with manual analysis. The objective image processing techniques used, such as top-hat transforms and Laplacian filters, set the framework for high-throughput image analysis programs. Successive algorithms included distance transforms with watershed segmentation [45], 3D images [92], and neural network learning algorithms [97]. As some algorithms consistently matched expert manual analysis, they have been used for automated disease diagnosis based on the number of spots detected [135].

Today a large number of high-throughput image analysis programs have become available. Smart 3D FISH and its subsequent graphical user interface package, NEMO, uses a top-hat transform, and local maxima for spot determination to determine Euclidean distances between points and distance from nucleus center [68, 78]. FISH Finder was developed in 2011, using a similar general algorithm with an available user interface, though FISH Finder uses a Bayesian classifier segmentation method [158]. TANGO was developed for more intense analysis of individual nucleus

features [126]. Users can define over 30 different segmentation algorithms for object recognition and an included 3D graphical interface display's identified features. FISH-quant is designed specifically for mRNA-FISH and counting dense signal images through gaussian mixture model point detection [119]. HIP-Map was developed as a full start to finish experimental protocol with an automated image analysis algorithm [156]. By defining the experimental conditions for hybridization, HIP-map is best set up to analyze the output images. While all programs are well-suited for specific types of analysis, no single program has emerged as the gold standard for FISH analysis.

Here we discuss possible solutions to lingering issues in nucleus image analysis, propose ideas for further investigation, and present novel findings related to genome structure along with our developed algorithm for time series FISH image analysis. These discussions revolve around largely overlooked aspects in nucleus image analysis such as 3D to 2D projection, canonical framework detection, and time series analysis. Novel findings are based on 2 distinct time series FISH experiments that analyze direct reprogramming and circadian rhythm cycles. The algorithm presented here was developed specifically to investigate these data sets and ideas, with emphasis on time-series images and the discovery of a consistent canonical framework. This algorithm allows for analysis of any channel separated image, of any size and number of nuclei. It is robust in its selection of probe locations within normal diploid nuclei, automatically selecting 2-4 points based on fluorescent intensity in accordance with natural gene copy numbers, though this parameter can be manually adjusted to fit the scope of a given project. We constructed our algorithm for analysis of nuclei with multiple fluorescent probes, as it can calculate spectral overlap between fluorescent channels and use this information to extract true signals within a channel. An algorithm for geometric transformation optimization analysis is used to plot nuclei in a canonical framework allowing all nuclei to be compared easily from a consistent viewing angle orientation, both within and between time points. Although a consistent canonical

framework could not be detected beyond statistical significance in the data sets explored here, we discuss reasons why this may be obscured in the analyzed data and extend this outline to 3D. This algorithm has been packaged as a MATLAB program available for download.

2.3 Re-thinking nucleus image analysis

Nucleus imaging has been instrumental in uncovering significant biological findings, but the full potential of this technique is yet unrealized and subtle aspects of nucleus image analysis have been overlooked. Here we highlight some potential issues with current methods of analysis and provide suggestions for improvement.

2.3.1 Cell to cell alignment: canonical framework

There is mounting evidence in the field of cellular biology that certain aspects of a cell genome are fixed in relation to the nucleus, and in relation to other chromatin [12, 136]. While it has long been known that DNA is separated into distinct chromosomes, FISH analysis has shown that the relative location of these chromosomes in relation to the nuclear edge can be non-random for a given cell type, termed chromosome territories [13, 15], and the location of certain genes can be predicted similarly [156]. This gives credence to the idea that there may exist a consistent canonical framework for genome structure in all cells with the same state. A canonical framework here refers to the idea that at a given cell state (cell-type, cell cycle stage, circadian time, etc.) a cell may have a preferred 3D organization of its genetic material. This potential canonical framework may be obscured in current methods of analysis for a variety of reasons: FISH nuclei images have a preset viewing angle determined by the microscope lens, nuclei are subjected to unnatural monolayer growth conditions that distort nuclear shape, analysis is largely performed on maximum projection 2D images, and system noise may cause some positional variability.

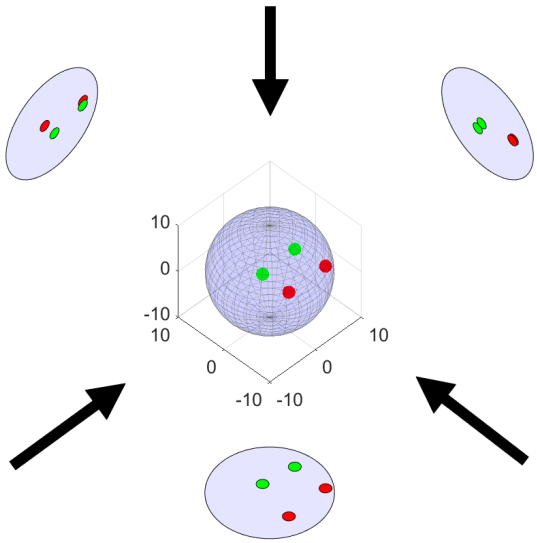


Figure 2.1: 3D to 2D image projection. Example on how viewing angle projection of a 3D object to a 2D image can distort probe location analysis. This figure was taken from Ronquist *et al.* [147].

To accurately analyze nuclei in FISH images, ideally nuclei should be viewed from the same viewing angle relative to this potential canonical framework. This information could come from 3D z-stack images and with cells that are grown in culture suspension to best preserve nucleus structure and reflect *in vivo* growth conditions (unless monolayer culture better mimics *in vivo* for a certain cell type) [30]. Figure 2.1 depicts the observation distortion issue that can arise when analyzing a nucleus from a set viewing angle. Transformation optimization analysis can be used to account for part of this issue, where the image can be translated, rotated, scaled and reflected to best match a global reference shape, in an attempt to set a consistent viewing angle (See Section Canonical framework detection - 2D).

2.3.2 Distance to nuclear edge: 3D imaging and ratio

A consistently used metric in analyzing gene positions is distance to nuclear edge. This is measured as the shortest distance of a FISH probe to the edge of a DAPI stained nucleus, from a maximum projection image. This measure of a gene or chro-

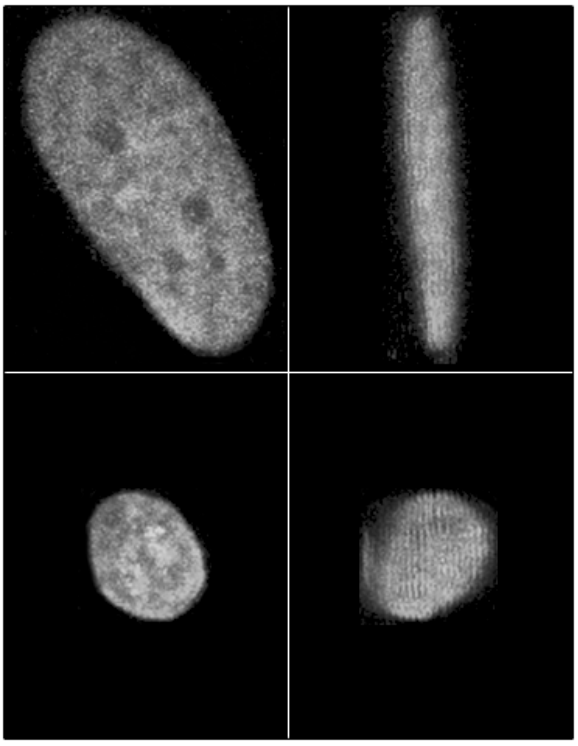


Figure 2.2: Spheroid vs monolayer growth conditions for human fibroblast cells. z dimension is defined as parallel to microscope viewing angle. Top Left: x-y monolayer. Top Right: x-z monolayer. Bottom Left: x-y spheroid. Bottom Right: x-z spheroid. This figure was taken from Ronquist *et al.* [147].

mosome location can be used to predict whether the DNA is activated or repressed, as genes closer to the nuclear edge have been shown to be less transcriptionally active on average [136].

There are two potential issues with this measurement as commonly performed: the measurement often ignores the z-dimension and it ignores the nucleus size and shape. The neglect of the z-dimension is often rationalized by comparing the measurements from maximum projection 2D images to z-stack 3D images on the same nucleus, then showing that the difference in measurement is negligible for certain cell types. What this analysis fails to take into account is that cells are often unnaturally flattened when grown in a monolayer culture flask, then subjected to coverslip fixation, potentially "squishing" the z-dimension. This idea is supported by recent studies showing the plasticity of the cell nuclei [23], and comparing cells grown in monolayer cultures

to the same cells grown in "spheroids," where cells have more freedom to grow in all directions [30]. The minimal change in distance between 2D and 3D may be an artifact of the environment in which it is grown. This can be seen in Figure 2.2 where nucleus shape for the same cell line varies dramatically depending on growth conditions. While measures of distance to nucleus edge can be informing even on 2D images, researchers should be aware of possible misinterpretation. Additionally, the cell shape and size can vary as a product of both its growth environment and of its cell cycle stage. This can lead to the distance to nuclear edge changing if a cell is in G_1 vs G_2 [87]. To account for this size discrepancy, a potentially more significant measurement is the relative distance to nuclear edge. This can be calculated as the ratio of the distance from the nucleus center to the FISH signal, over the distance from the nucleus center to the nucleus edge along the same trajectory.

2.3.3 Discrimination of chromosome copies

The human genome, and nearly all eukaryote genomes, are composed of diploid chromosomes, meaning each chromosome has a highly similar copy, corresponding to one chromosome from a father and one chromosome from a mother. While the sequence of each chromosome is largely preserved, the function and relative location of each chromosome in relation to the nucleus can be highly variable. An acute example of this difference in chromosomal function and position is highlighted in X chromosome inactivation. In this process, one X chromosome homolog is inactivated leading to highly reduced transcription, a shrinkage in chromosome territory size, and a change in chromosome territory position towards the nucleus edge [33, 160, 170]. Though this is a special case, the idea that different chromosome copies play different roles in both transcription and position is of importance and should be highlighted when exploring chromosome and gene positioning, especially when analyzing genes that are monoallelic in gene expression. Only recently have advanced FISH techniques

allowed for direct homolog discrimination through SNP differences, and this technique is limited in that the target genome SNPs must be known prior to analysis [8]. Often, the two different copies are lumped together when exploring "distance to nucleus edge," blurring the true nature of the information.

2.4 Time series data analysis - 2D images

Another important aspect of nucleus imaging to consider is the dynamics of the genome. Many chemical interactions happen simultaneously within a cell nucleus every second, resulting in cellular processes such as gene transcription, protein binding, and nucleotide replication and repair. As a consequence of these processes, the DNA within the cell is in constant motion. Large scale changes in chromatin positioning can be observed as the cell progresses through the cell cycle, most notably during mitosis as chromatin condenses, aligns, and the cell divides. Subtler changes in gene positioning exist as well, such as circadian genes moving in slight oscillations with 24 hour periodicity, and genes involved in differentiation moving towards or away from the nucleus edge in parallel with their expression [31, 134]. These changes can only be observed through time series FISH images taken on a population-level of synchronized cells, or through live cell imaging.

With time series data we can attempt to identify probe location changes over time. Since time-series images fix the cell, thus ceasing all cellular processes, each time point must be collected on a new population of cells that may have slight system variation, and different distinct viewing angles in relation to a potential canonical framework. Time series imaging is specifically important when making observations on genomic structure for cells that are proliferating of differentiating. Both conditions imply changing the cell state, either permanently during differentiation or cyclically during proliferation. To demonstrate this importance, we have analyzed the gene position changes over time for 2 conditions; cell-cycle and circadian rhythm synchro-

nized proliferating fibroblast cells, and fibroblasts being directly reprogrammed into myotubes [31]. Similar analysis was performed on both populations of collected cell images.

We note here that analysis was performed on the FISH location of 2 points within each nucleus, in accordance with the normal diploid human fibroblast cell line used. While some cells are expected to be in cell cycle phase G_2 and therefore have 4 gene locations (as seen in imaging), the sister chromatids are often found to be close enough together to consider the location as a single point, or the brightest 2 fluorescent points are chosen in cases of ambiguity.

2.4.1 Circadian rhythm analysis

A circadian rhythm refers to any biological process that oscillates in a 24 hr period. Many of these circadian processes have been related back to a core set of genes that control this rhythm, and oscillate in gene transcription in a 24 hr period. As many as 43% of all protein coding genes have been shown to be within, or regulated by, a network of circadian genes [190].

For this experiment, well-known circadian genes *PER2*, *CRY1*, *ARNTL*, and *CLOCK* were FISH imaged in a cell culture population of fibroblasts over time (Figure 2.3a). Images were collected 8 hours apart, over 16 time points, encompassing 3 circadian rhythm cycles. The distance between all genes in each nucleus and the ratio distance to the nucleus edge was calculated for all nuclei over all time points. We note here that these measurements do not depend on homolog assignment, and that image analysis was performed using 2D images. Genes have a preferred location in relation to the nucleus edge as *PER2* is found to be closer to the nucleus edge then *CLOCK* in all time points. Interestingly, some genes show a sinusoidal period in their distance between homologs, most noticeable in *PER2* and *ARNTL* (Figure 2.3a), in line with their circadian gene expression.

2.4.2 Direct reprogramming

While fully differentiated cells have a static phenotype *in vivo*, all cells within an organism have the same genetic information and therefore may have the capacity to become any other cell type, given the right control inputs. The first example of this was shown in 1989, when researchers reprogrammed fibroblast cells into myotubes by transfecting the fibroblasts with the master regulator protein *MYOD1* [169]. Since then, a number of additional reprogramming factors have been discovered to convert between cell types. Though cell reprogramming is often marred by inefficiency, fibroblast to myotube conversion has been performed at greater than 90% efficiency [96].

For this experiment, the myogenesis related genes *MYOD1* and *MYOG* were FISH imaged in a cell culture population of fibroblasts over time, along with the circadian genes *PER2* and *CLOCK* (Figure 2.3b). Images were collected once per day over 7 days, along with a control (day -1) and a day 10 sample, encompassing the entire reprogramming process. The distance between all genes in each nucleus and the ratio distance to the nucleus edge was calculated for all nuclei over all time points. Our measurements did not attempt to distinguish between homologs for this analysis. All imaged genes tended towards the center of the nucleus over time, which has been shown to be correlated with increased transcriptional activity (Figure 2.3b) [136].

2.4.3 Canonical framework detection - 2D

To search for a canonical framework it is best to view all cells in the same orientation, both within and between time points. This framework is not obvious as cells may have variability in shape, size, and loci positioning. To try and reconstruct this framework, we can first set a coordinate system based on the nucleus shape. For analysis of 2D nuclei, an ellipse can be fit as a reasonable mask for most cell-types, with the major and minor axes acting as a coordinate system for each nucleus. All

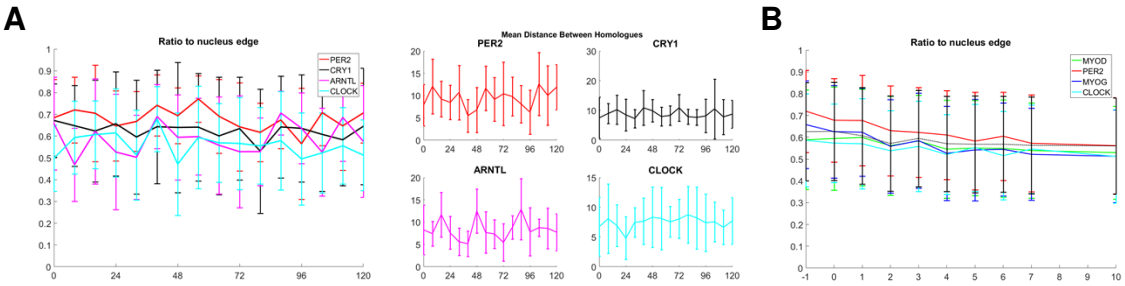


Figure 2.3: Analysis of gene position changes over time. A) Circadian genes in synchronized fibroblasts, time units in hours after start of synchronization. Left: Ratio to nucleus edge over time. Right: Mean distance (μm) between homologs over time. Time units are in hours after synchronization. B) Ratio to nucleus edge over time for genes in fibroblasts being directly reprogrammed. Ratios are computed as distance to nucleus center divided by distance from nucleus center to nucleus edge along the same trajectory. Dashed black line shows ratio to nucleus edge over time for random point locations over the same nucleus shapes, used as a control. Time units are in days after exogenous *MYOD1* addition. This figure was taken from Ronquist *et al.* [147].

nuclei can then be rescaled to the average ellipse mask size at a given time point by calculating the average major and minor axis and scaling each nucleus accordingly, with labeled points scaled as well. These nuclei can then be reflected along the x axis, reflected along the y axis, or reflected along both the x and y axis (rotated 180°), to minimize the distance between points within each nucleus (See Figure 2.4).

For example, first consider the problem where all nuclei are collected from a single time point, with the same assumed state (cell-type, cell cycle phase, etc.). Let's assume we have some template for the homolog locations. For a nucleus target we're analyzing, we can try all transformations (we have a set of 4; no transformation, reflect x, reflect y, reflect x and y) and all possible homolog assignments for each gene (2 assignments for each diploid gene) in order to minimize the sum of distances from target to template. We can perform these operations for all imaged nuclei, and get optimal transformation and homolog assignment for each. We can then try each of the imaged nuclei as the template for all the other nuclei, calculate distances of errors over all nuclei, and pick the one yielding the lowest sum of distances. This concept could

also be applied without nucleus size scaling and without the constraint of nucleus ellipse alignment, at which point this becomes a classic Procrustes analysis problem based strictly on gene locations relative to themselves. We include the constraint of nucleus ellipse alignment based on previously referenced evidence that gene and chromosome locations have been shown to have a preferred nuclear location [13, 15, 136, 156]. This analysis attempts to determine if this preferred location is a result of a consistent canonical framework.

With this framework set, we can use the same idea for the computation of a template between time points. For each time point, we determine the average position of each homolog in reference to the average nucleus shape and input these positions into the same algorithm. This calculation will provide our best guess for homolog assignment both within and between time points. We term this collective algorithm "2D Transformation Optimization."

Analysis of *MYOD1* and *MYOG* loci location over time during fibroblast to myotube reprogramming shows dynamic loci movement within the nucleus. The 2D transformation optimization algorithm explained above is used for this analysis. The last time point (day 10) is used as our template between time points because it is collected on fully transformed myotubes, therefore these nuclei may exert the lowest amount of structure variability. *MYOD1* and *MYOG* positions have been shown above in Figure 2.3 to move towards the nucleus center, and upon further analysis their preferred location appears to be more constrained as well. Point clouds corresponding to loci location for 100 nucleus samples become more condensed over time, with the average distance to the gene position centroids decreasing as the cell becomes more differentiated (Figure 2.5).

This analysis was performed on *MYOG*, *MYOD1*, *MYOG* and *MYOD1* together, and randomly distributed points (1 gene case). We note that when analyzing multiple genes as in Figure 2.5c, there is a distinct possibility that different clustering

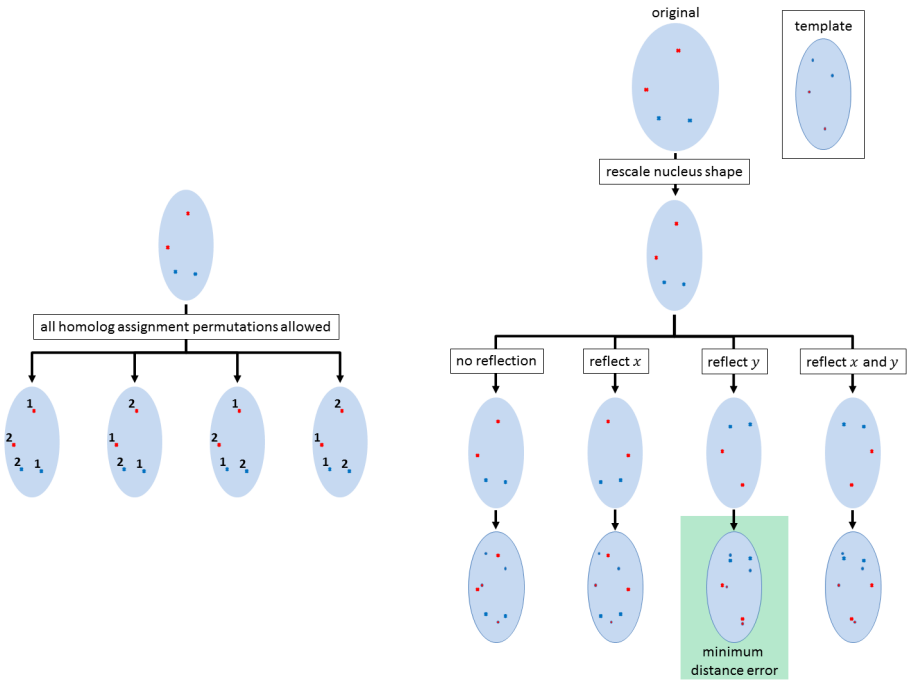


Figure 2.4: 2D Transformation Optimization Analysis. Example depicting 2D Transformation Optimization Analysis permutations and transformations. Left: Example of the homolog assignments allowed. Right: Example of the transformations allowed with correct selection. This figure was taken from Ronquist *et al.* [147].

results will be found compared to single gene analysis as there are more operational constraints (i.e. transformation must minimize sum of more distances). Results here suggest that gene positioning relative to nucleus may be highly variable since random gene location permutations leads to similar mean distances to computed centroids at each time point, a measure used to determine gene positional variability post-Transformation Optimization. An alternative possible conclusion is that the canonical framework is obscured by the confounding effects of 2D imaging and/or desynchronization between nuclei in each time point. As an initial attempt at canonical framework detect we believe this method is biologically and computationally reasonable and can be applied to any time series FISH imaging experiment for future investigation.

2.4.4 Canonical framework detection - 3D

High-throughput nucleus image analysis often relies on 2D maximum intensity projection images. There are many reasons for this: 3D image stacks are much larger and harder to analyze, the z dimension has limited resolution compared to x and y, and cells grown in monoculture are often flattened in that dimension. These issues can be overcome through super-resolution measurement or spheroid growth conditions, but these procedures increase the experimental difficulty and cost.

Assuming a consistent configuration of gene positions between cells in the same state and a sufficient number of measured cells, it is possible to recover 3D information from 2D images. For this recovery, Procrustes-like analysis can be used to extract z coordinates of gene positions. Procrustes analysis optimally aligns sets of points to minimize error between corresponding points in each set through global transformations. These transformations include rotation, translation, scaling, and reflection.

For example, given a set of four uniquely labeled genes within two similar nuclei

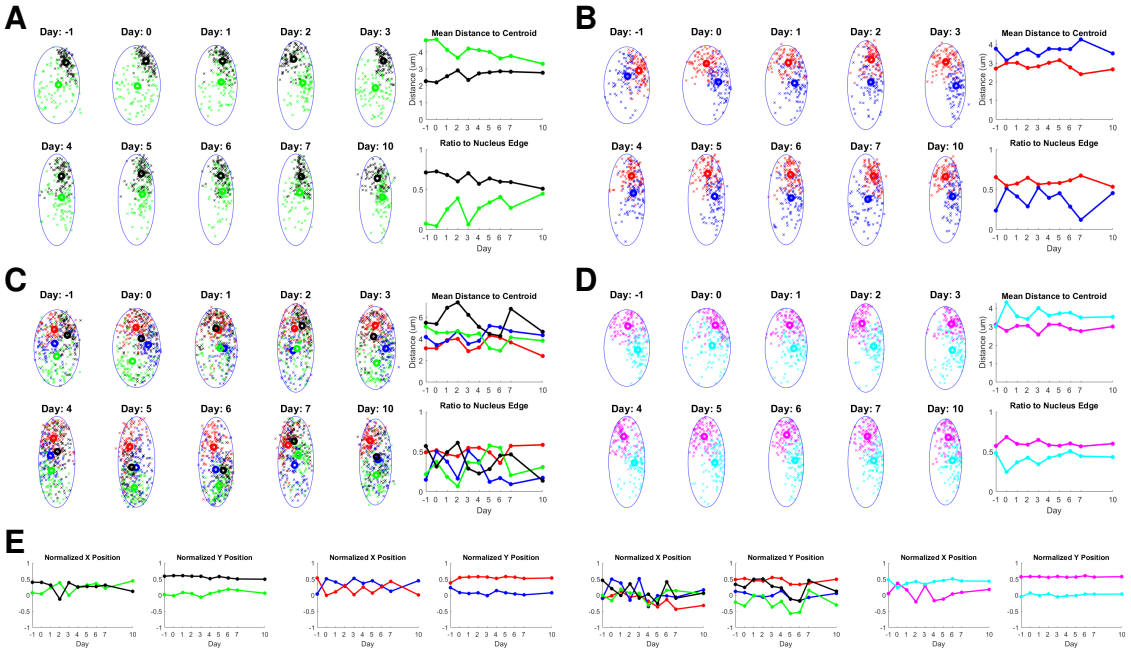


Figure 2.5: 2D Transformation Optimization Analysis - Overview, synchronized fibroblasts and MYOD1-mediated reprogramming. A-D) Gene position over time during fibroblast to myotube reprogramming, in relation to the average nucleus shape, after 2D Transformation Optimization analysis. Days are labeled relative to the addition of TF *MYOD1*. 100 nuclei are taken per time point, with 10 distinct time points sampled. Plot shows the mean of distances from all points to their cluster centroid over time, and ratio to nucleus edge of average homolog position. Within each nucleus, 'x' shows real gene locations, 'o' shows computed centroid or average position of a given homolog. A) *MYOG* (green and black). B) *MYOD1* (red and blue). C) *MYOD1* and *MYOG*. D) Random gene positions for 1 gene (cyan and magenta). E) X and Y positions of gene centroids normalized by maximum axis length. This figure was taken from Ronquist *et al.* [147].

with xy locations given, we can perform a search (e.g. Nelder-Mead optimization) to determine which z values minimize error. Since there are more constraints [$8x + 8y = 16$], than degrees of freedom (DOFs) [$8z$ + 3 rotational DOFs + 3 translational DOFs = 14], a unique solution is expected to minimize error between corresponding points. Adding more points and more nuclei can give higher confidence in the results, if the error remains small. As with our 2D analysis, homolog uncertainty complicates the problem: we can't unambiguously determine which FISH point corresponds to each homolog between sets. Homolog assignment information can be reasonably predicted by trying all possible homolog assignments and selecting the assignment that achieves lowest error.

Demonstrating the hypothesis that genes have a consistent, if time-dependent, 3D location is theoretically possible with 4 distinct FISH points (as analyzed in this experiment) if the observed locations are known precisely; clearly the reconstruction degrades as the number of gene points decreases or their positional measurement error increases. In future study, better supporting our hypothesis may require better synchronization of the cell population and true 3D rather than 2D analysis, as well as more gene points with lower positional error.

2.5 FISH image analysis algorithm

The general process workflow for our automated, high throughput program is depicted in Figure 2.6 and is detailed below for 2D nucleus image analysis.

Pre-Process: 2D maximum projection images are input, with the nucleus image channel and the number of channels corresponding to fluorescent probes specified. 2D images can be any dimension (length, width, and channel number), with any number of non-overlapping nuclei per image. Multiple images can be input and analyzed in the same run and images corresponding to different time points can be specified for

time series analysis for subsequent Transformation Optimization.

Nuclei Extraction: The first step in the image processing algorithm is to determine the number and location of nuclei within the image. This information is extracted through a series of image processing techniques. The nucleus image channel is first converted to grayscale with nuclei assigned higher pixel values. In order to segment nuclei from background, the gradient magnitude is calculated in both the x and y direction. This is used to determine nuclei borders. Foreground objects are determined through "opening-by-reconstruction" to eliminate noise within the image. A mild Gaussian filter is used to ensure any remaining noise within each nucleus is removed before determining a pixel value threshold for separation of foreground and background objects. This threshold is computed through Otsu's method to binarize the image. The gradient segmentation is superimposed on the binarized image to only allow local maxima within defined regions. Finally, a distance transform followed by watershed segmentation is then applied on the binary image to distinguish between slightly overlapping nuclei, and to determine nucleus centers. Nucleus area is computed, and outliers are discarded to prevent M-phase nuclei and non-segmented overlapping nuclei from further analysis.

Individual Nucleus Analysis: Each nucleus is fitted with an ellipse mask, and major and minor axes are determined. The validity of an ellipse fit for nuclear shape is reasonable for many cell types, but should be determine by the user on a cell-type specific case-by-case basis. This fitting in necessary for downstream Transformation Optimization, but is inconsequential for computed distance between points and distance to nucleus edge. The major and minor axes can set the Cartesian coordinate system for the nucleus to allow for comparison to other nuclei in later analysis. The fluorescent channels are then linear unmixed to elicit true signal from overlapping fluorescent signals. Local maxima within each nucleus are found for each channel to

determine the probe locations, which are then projected onto the Cartesian coordinate system. This program has been optimized to analyze normal karyotyped diploid cell gene positions, and as a result the number of points detected typically varies from 2 to 4 based on fluorescent intensity of signals. For Diploid cells, nuclei in cell cycle phase $G_{0/1}$ should have 2 copies, in S they should have 2-4, and G_2-M should have 4. To detect gene copy numbers, our program extracts the top 2 FISH points within each nucleus mask based on fluorescent intensity within a square pixel window, and then determines subsequent points based on comparative fluorescence. Points selected beyond the first 2 are included in further analysis if their fluorescence is within some optimize percentage of the brightest point in the channel image, though this parameter can be altered by the user to extract more/less points for a given experiment. This eliminates the need for the cell stage to be known *a priori* and allows our program to best select true probe signal in potentially noisy channel images. Distance between all probes within each nucleus are computed along with nucleus size and shape, distance from probe to nucleus edge, and the ratio from nucleus edge to nucleus center.

Plane Orientation: Once the nucleus has been ellipse fitted, mapped to a Cartesian coordinate plane, and probe points are selected, the collective coordinate planes are oriented using 2D Transformation Optimization analysis to align the axis origin in a best fit manner. This reflects and scales the coordinate plane set by the major/minor axes of the ellipse to minimize the distance between points for all channels and all nuclei at a given time point. This creates a best-fit orientation for easy comparison of all nuclei within the same time point. This analysis is also applied between time points in time-series FISH images, thus constructing a consistent canonical framework showing how a loci moves within the nucleus over time.

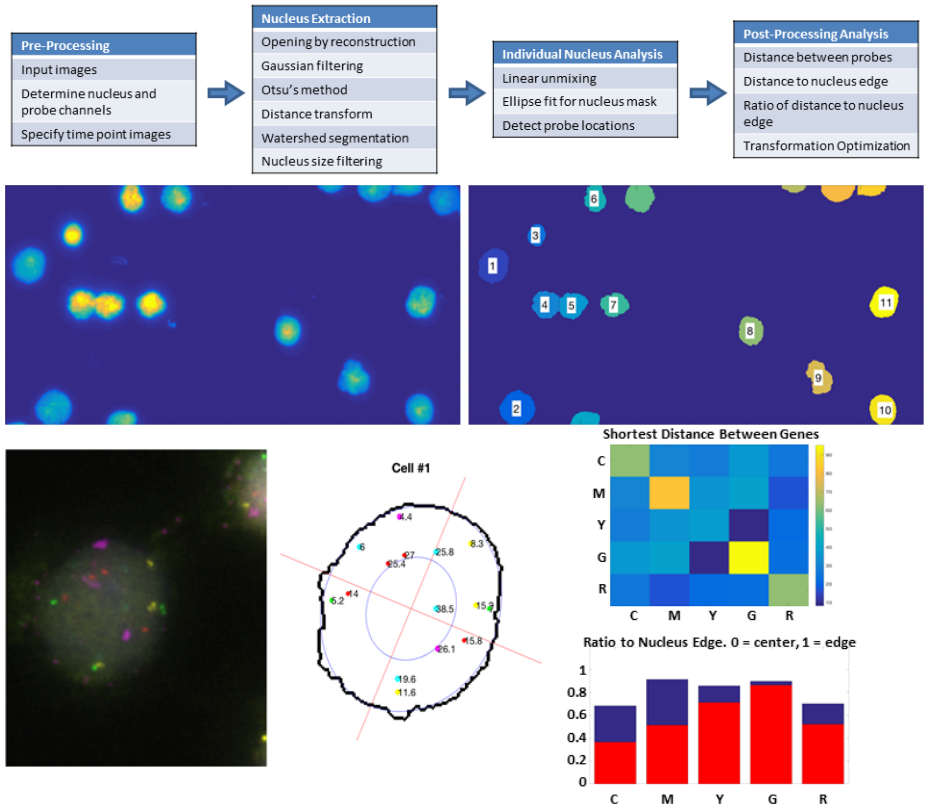


Figure 2.6: Automated time series FISH image analysis. Top row: General algorithm workflow. Example images from different points along algorithm workflow. Middle row: DAPI channel image of nuclei and nuclei segmentation results. Bottom row: Individual nucleus analysis, distance matrix between probe locations for individual nuclei (letters correspond to fluorescent color), ratio to nucleus edge of points within individual nuclei. This figure was taken from Ronquist *et al.* [147].

2.6 Discussion

Before beginning any nucleus imaging project, a balance must be determined between FISH labeling technique specificity and resources invested. While it would be convenient for analysis if all experiments used cells grown in *in vivo*-like conditions, with direct FISH homolog distinction and super resolution microscopy, few labs have the funding, personnel, equipment, and time to perform these techniques, and furthermore they may not be necessary to test a given hypothesis. Researchers should be aware of the pitfalls from any given FISH imaging technique. As discussed, 2D culture-grown cell images can be misleading in analysis and homolog distinction should be noted, especially when analyzing monoallelic expressed genes.

For the experiments conducted and analyzed in the present study, a number of image resolution limitations should be noted. Cell images in the present study were analyzed on a confocal microscope, allowing image resolution up to 200nm in x-y plane and 400nm in the z plane. This resolution is sufficient for precise gene location determination in relation to the general nuclear shape, but overlooks some of the intricacies of nucleus architecture that have been shown to play a role in gene regulation, such as distance to nucleolus and nuclear pores [131]. Integration of this information with gene function is considered a promising area of future work.

For studying genes that are known to be dynamic in either transcription or location over time, it is necessary to utilize time-series imaging to determine if the dynamic transcription is correlated with positioning. Examples of gene position changes related to circadian rhythms and cell reprogramming have been shown using time-series imaging, where genes with circadian expression have corresponding circadian gene movement, and fibroblast to myotube reprogramming indicates movement of expressed genes towards the nucleus center over time. For this analysis we have proposed using 2D Transformation Optimization to extract information on how labeled genes may move in relation to other genes, and in relation to the nucleus shape over

time. This algorithm is designed with canonical framework detection considered, and thus can be used to provide a reasonable best guess at homolog identification. Though results using this technique on fibroblasts being reprogrammed were shown to be inconclusive when compared to random point locations, it is unknown if this reflects true positional genome randomness, poor synchronization within imaged cells, or confounding 2D imaging artifacts coupled with high positional variability. Nonetheless, this technique can be applied to alternate data sets to test for canonical frameworks in future experiments. Additionally, an extension to 3D image analysis for determination of a consistent canonical framework has been elucidated.

All analysis performed can be found in our available MATLAB scripts (http://bionetworks.ccmb.med.umich.edu/). This algorithm can be used for any given FISH image experiment and is best suited for time series images, with the extension for Transformation Optimization following analysis.

CHAPTER III

Functional Organization of the Human 4D Nucleome

This chapter is based on a paper by Haiming Chen, Jie Chen, Lindsey A. Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse [31]. I have summarized my contributions to this paper here.

3.1 Abstract

Cells alter their genome structure and function to mediate proliferation. For example, cyclin gene expression fluctuates to control cell cycle progression and chromosomes condense, align, and divide during mitosis. However, the interplay between genome structure and function during cell proliferation is not well understood. Here, we analyze time series Hi-C, RNA-seq, and FISH imaging data to characterize the 4DN of proliferating human fibroblasts. In this chapter, we present a framework for the identification of potentially co-regulated genes. Furthermore, our analysis reveals a correlation between circadian gene positioning and expression over time. The analysis of this comprehensive data set highlights the benefits of a 4DN approach.

3.2 Introduction

A comprehensive understanding of the dynamical genome structure-function (S-F) relationships is necessary to fully understand how a cell operates. Clear examples where changes in genome topology (structure) can affect gene expression (function) have been found at many scales [28, 136]. One such example is in cancer, where chromosomal alterations have a profound effect on gene expression and cell regulation, driving the cell towards an uncontrolled proliferative state. Additionally, distinct S-F relationships have been observed through the cell cycle [121] and in different cell types [52]. However, most experiments either focus on specific genomic regions or do not explore all 4DN modalities, thus a comprehensive picture of the 4DN is still lacking.

Here, we present the analysis of a complete 4DN data set, collected on proliferating human fibroblasts. Cells are cell cycle- and circadian rhythm-synchronized to capture dynamics from a population of cells [6, 134]. Time series Hi-C and RNA-seq samples are collected to observe genome structure and function, respectively. Given this data, along with information on TF binding, we present methods for the extraction of potentially co-regulated genomic regions. Specific biological modules are analyzed using this method to extract larger potentially co-regulated gene networks. Furthermore, a S-F relationship between the core circadian genes *PER2* and *CLOCK* is uncovered, where gene position correlates with gene expression. The analysis presented within this chapter furthers our understanding of the 4DN, and the methods are applicable to any 4DN data set.

3.3 Results

3.3.1 Dynamical S-F correlations

We have used concepts from the theory of networks to evaluate genome-wide S-F dynamical correlations. For structure, gene boundaries were defined by the tran-

scribed region plus 2 kb upstream of the transcription start site and 2 kb downstream of the polyadenylation site [27]. In method A, we used gene dynamics: the time-dependent variation in structure and function within each gene. In method B, we used gene network dynamics. Network analyses were performed using two methods. In method A, we inferred networks from gene dynamics, that is, by constructing the interaction or the edge based on the correlation between gene expression and the correlation between the structures of each gene. In method B, we constructed edges based on the correlation between gene expression and Hi-C contacts (Figure 3.1). In both methods, we surveyed regulatory regions of the identified correlated gene pairs for common TF binding sites and determined whether these common TFs were expressed in our RNA-seq dataset. Gene network dynamics therefore facilitate identification of gene pairs or clusters with high potential for coregulated expression that is consistent with the transcription factory model.

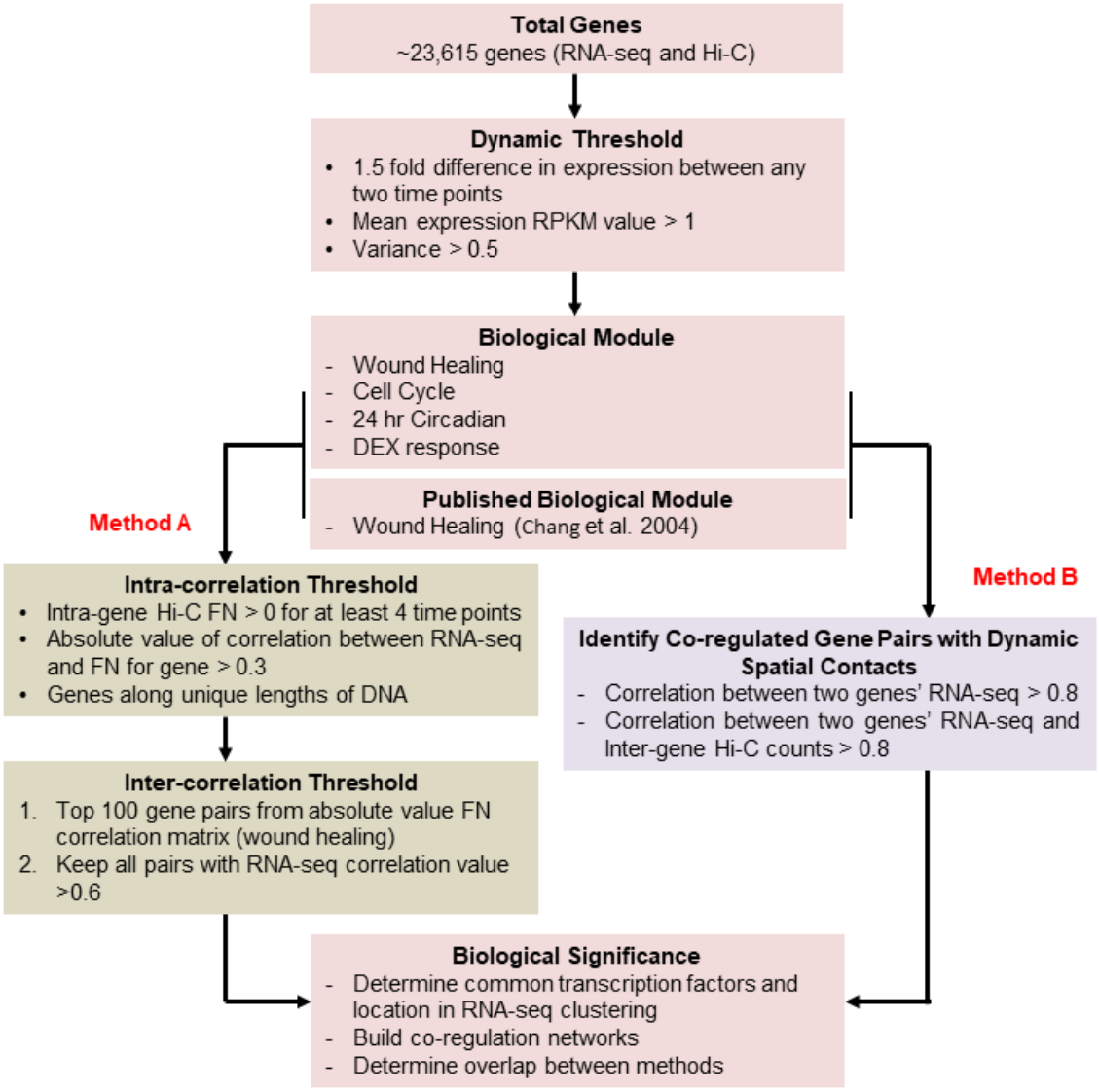


Figure 3.1: Algorithm for extraction of Dynamic Form Function correlated gene pairs. This figure was taken from Chen *et al.* [31].

3.3.2 Gene dynamics

To identify the genes with S-F correlations, we studied the 7,786 genes that significantly varied in expression and then four biological modules: wound healing, cell cycle, 24-h circadian clock, and dex response. We defined gene dynamics using the dynamical correlation between gene expression and gene structure (Appendix A.1). We identified a set of 2,574 genes from the 7,786 significant genes, using the following

criteria: (i) a gene's Hi-C Fiedler number >0 for at least four time points (a value of 0 was considered artifactual), (ii) the absolute value of the correlation between RNA-seq and a Fiedler number >0.3, and (iii) genes were along a unique length of DNA.

3.3.3 Gene network dynamics

We first constructed networks by considering the correlation between gene structures represented by Fiedler number. We performed the analysis (method A above, Figure 3.1) on the above-mentioned 2,574 genes, chromosome by chromosome. A total of 986 gene pairs were identified (Dataset S10 in Chen *et al.* [31]). We report the identified gene pairs and the constructed networks for chromosome 14 in Figure 3.2. We then identified common binding sites of expressed TFs for these gene pairs. We found that gene pairs shared more binding sites than randomly expected in 16 of 22 chromosomes, suggesting that transcription may be coordinated in these structure- and function-correlated gene pairs. We also examined the mean contact over time between all gene pairs of the extracted 2,574 genes. We found that if two genes have a high Fiedler number correlation, they are more likely to have contacts between them.

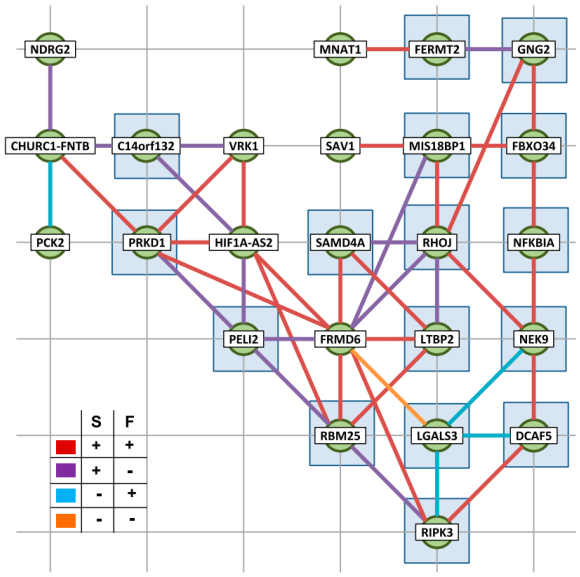


Figure 3.2: Networks of dynamic intracorrelated and intercorrelated S-F gene pairs on Chr 14. Green nodes represent genes, and thick edges between pairs of genes represent a correlation. (Inset) Colors of edges show how the two genes are correlated (color key). Genes with TFs in common with all other genes that share edges are denoted by shaded blue squares. TFs associated with gene pairs are given in Dataset S10 in Chen *et al.* [31]. This figure was taken from Chen *et al.* [31].

We applied the above analysis to the four biological modules and found 35 wound healing, 49 cell cycle, 52 circadian clock, and 49 dex response gene pairs that were highly correlated between structure and function (Dataset S10 in Chen *et al.* [31]). We also mapped common binding sites for all these gene pairs (Dataset S10 in Chen *et al.* [31]).

3.3.4 Periodicity in spatial movement in core circadian genes

We used multicolor 3D-FISH to examine spatial dynamics of the core circadian genes *CLOCK*, aryl hydrocarbon receptor nuclear translocator-like (*ARNTL*), cryptochrome 1 (*CRY1*), and period 2 (*PER2*) [21], which have well-studied transcriptional periodicity. Notably, although we found no Hi-C contacts between *CLOCK* and *PER2*, we found that *CLOCK* S-F dynamics are negatively correlated, whereas they are positively correlated for *PER2*. Because transcriptional activity that is con-

current with movement in the nucleus has been reported [134, 136], we hypothesized that *CLOCK* and *PER2*, which have antiphase transcriptional periodicity [21], would show distinct spatial dynamics. We therefore used multicolor 3D-FISH to obtain the allele locations of the genes for each of 16 time points simultaneously (Figure 3.3A).

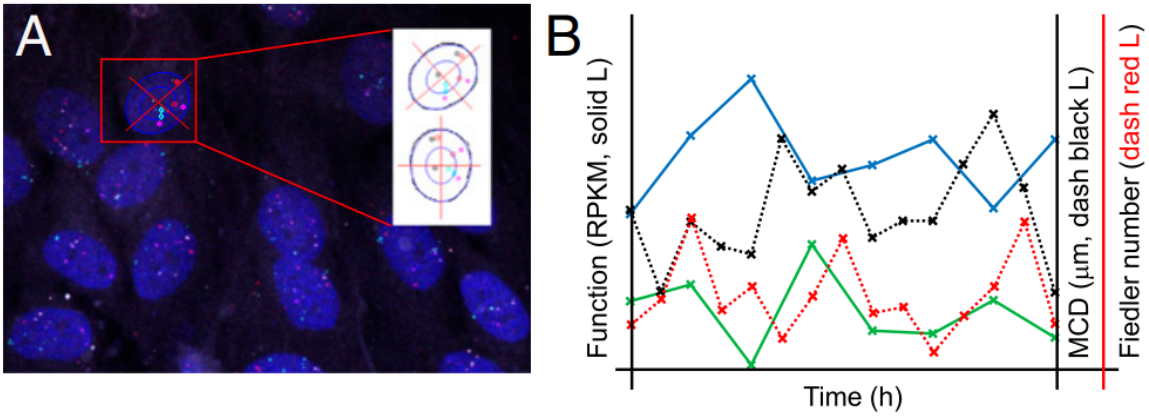


Figure 3.3: Processing of 3D-FISH raw data maximum projection images (MPIs). A) Cartesian coordinate system is superimposed after fitting nuclei to an ellipse. Red, cyan, white, and magenta points represent probe signals for *PER2*, *CRY1*, *ARNTL*, and *CLOCK*, respectively. B) RNA-seq data over time are plotted on the left y axis for *CLOCK* [solid blue line (L)] and *PER2* (solid green L) in RPKM. MCD in micrometers (dashed black L) and Fiedler number (dashed red L) between *CLOCK* and *PER2* over time are plotted on the right y axis. This figure was taken from Chen *et al.* [31].

We used two measures to quantify the dynamics of the relative distances between gene pairs for all four circadian genes. The first measure was the mean closest distance (MCD; Appendix A.1) between each gene pair (relative distance curve), and the second measure was the Fiedler number of the Euclidean distance matrix among the four genes (stability curve). We then correlated these measures with transcription data. We found that the relative distance and stability curves for *CLOCK* and *PER2* showed periodicity, and followed a 24-h circadian rhythm (Figure 3.3B). When the MCD between *PER2* and *CLOCK* was at its minimum, *PER2* transcription was minimal and *CLOCK* transcription was maximal, and vice versa. Collectively, we observed that the MCD, the Fiedler number, and expression levels over time between

PER2 and *CLOCK* were all within approximately 6-h phase shifts of one another. This *CLOCK*/*PER2* system had the highest Fiedler numbers when *CLOCK* and *PER2* had the largest relative distance between them, which may hint at a particular state or time for which genome topology has particular significance in circadian gene dynamics. A schematic of this process is outlined in Figure 3.4. These observations provide insight into circadian gene modules, although the mechanisms driving these S-F dynamics require further investigation.

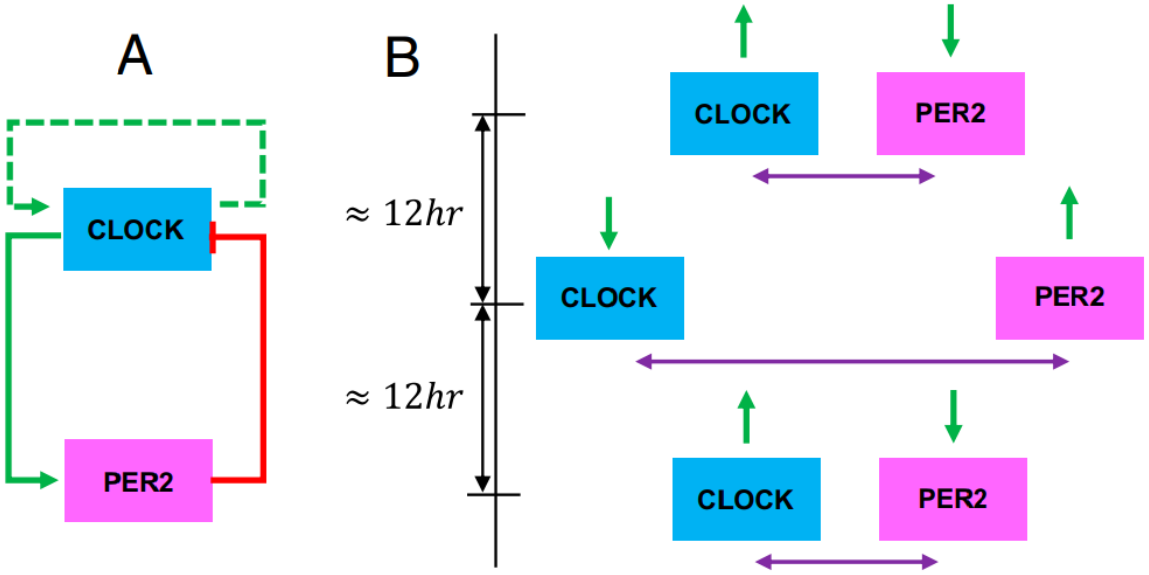


Figure 3.4: *CLOCK*/*PER2* circuit. A) Proposed feedback circuit for *CLOCK* and *PER2* expression, where *CLOCK* may self-activate. B) Relative expression of *CLOCK* and *PER2* (green arrows) at given relative Euclidean distances (purple arrows). This figure was taken from Chen *et al.* [31].

3.4 Discussion

In proliferating human fibroblasts, the 7,786 genes with highly dynamic expression may be considered a wound healing module that responds to proliferation cues, such as serum stimulation [82]; indeed, we found that 76% of known wound healing genes were within this highly dynamic set (Appendix A.1, Dataset S3 in Chen *et al.* [31]). A large number of genes in this highly dynamic set have highly correlated S-F dynamics,

but the significance of this property and whether it is found in other cell types are not yet known. Interestingly, the wound healing response overlaps with cancer metastasis [29, 82]. Thus, a better understanding of the wound healing module may provide insight into cellular functions that are active in metastatic cancer cells.

For S-F data collected over time, we have developed an algorithm for genome-wide identification of gene pairs or networks and candidate TFs that may be involved in coordinated gene expression. Criteria used in this identification are consistent with expression via transcription factories, although it is unknown whether genes would be actively recruited into or spontaneously self-organize in shared transcriptional space. Our algorithm could be used in any system to identify novel or key interacting genes, and varying correlation thresholds provides additional flexibility in restricting analyses to gene pairs with particular S-F properties (Appendix A.1, Figure 3.1 and Datasets S10 and S11 in Chen *et al.* [31]). The observations of *CLOCK* and *PER2* transcription and genomic movements in 3D space provide a geometric picture of gene regulation in the context of circadian clocks, one that may give insight into the mechanisms regulating biological time. These studies also suggest that important spatial relationships may be too distant in Euclidean space for capture by Hi-C.

CHAPTER IV

Algorithm for Cellular Reprogramming

This chapter is based on a paper by Scott Ronquist, Geoff Patterson, Lindsey A. Muir, Stephen Lindsly, Haiming Chen, Markus Brown, Max S. Wicha, Anthony Bloch, Roger Brockett, and Indika Rajapakse [147]. I have summarized my contributions to this paper here.

4.1 Abstract

Cellular reprogramming is the process of converting one cell type into an alternate cell type. To determine methods for reprogramming, researchers often select TFs that are overexpressed in the target cell type, input the selected TFs to the initial cell type, then check whether the experiment was successful. This "guess-and-check" method is time-consuming, inefficient, and ignores the dynamic nature of the cell. Using the comprehensive 4DN data set presented in the previous chapter, I helped model the dynamics of the cell cycle to create an algorithm for cellular reprogramming predictions. By modelling the natural dynamics of the initial cell type and determining where TFs can bind within the genome, we determine the optimal timing, combination and relative amount of TFs that can catalyze reprogramming from the initial cell type towards a target cell type. This algorithm correctly predicts known reprogramming TFs, thus bypassing expensive and time-consuming experimental methods

for discovery.

4.2 Introduction

All cells in a human body are derived from a single zygote, a totipotent cell that can proliferate and differentiate into all possible cell types. This means that all the different cell types in a human body contain (roughly) the same genetic information, and therefore have the potential to operate as any possible cell type. However, for many years researchers believed that cell types could only develop through unidirectional cellular differentiation, and that these differentiation decisions were largely controlled by lineage-specific TFs. This belief was challenged in 1989, when Dr. Harold Weintraub reprogrammed human fibroblasts into muscle cells via over-expression of the TF MYOD1 [184]. The limits to cellular reprogramming were pushed further in 2007, when Dr. Shinya Yamanaka reprogrammed human fibroblasts into embryonic-stem-cell-like cells using a combination of TFs [166]. Since these pioneering discoveries, many additional cellular reprogramming experiments have been demonstrated and researchers are beginning to test how reprogrammed cells could be used in a clinical setting.

Though cellular reprogramming has immense therapeutic potential, the methods used for the discovery of reprogramming TFs are time-consuming and inefficient. Often, researchers try multiple combinations of TFs that are overexpressed in the target cell type before finding the right solution. Recently, researchers have developed methods to predict TFs for reprogramming the cell state using NGS data [24, 44, 114, 133]. For example, Rackham *et al.* devised a predictive method based on differential expression, as well as gene and protein network data [133]. These methods, while useful, mainly rely on differential expression and largely ignore genome structure and dynamics.

Here, we take a 4DN approach for predicting TFs that can be used for cellular

reprogramming. By integrating time series Hi-C and RNA-seq data collected on a population of synchronized human fibroblasts, we model the dynamics of the genome to determine where and when to input control. Our model is based on control theory equations, where Hi-C and RNA-seq define the cell state and TF binding sites determine where TFs can influence the cell state [20]. Furthermore, to decrease model complexity, we use TADs to reduce the dimension of the cell state. With this model, we can investigate the efficiency, timing, and optimality (minimizing the number and concentration of TFs) for all possible TF-mediated cell type conversions. Our method identifies TFs previously found to reprogram human fibroblasts into embryonic stem cell-like cells, muscle cells and many additional target cell types.

The true dynamics within a cell are undoubtedly non-linear, but given the constraints of the available data, we use a linear control equation [5],

$$x_{k+1} = A_k x_k + B u_k. \tag{4.1}$$

In this case, the three items listed above correspond respectively to the value of the state x_k at time k, the time dependent state transition matrix A_k, and the input matrix B (along with the input function u_k).

4.3 Methods

4.3.1 Genome state representation and dimension reduction: $\mathbf{x}_k$

In order to use the linear control equation given in Eq. 4.1, we must decide which measurements define the cell state, x. To fully represent the state of a cell, a high number of measurements would need to be taken, including gene expression, protein level, chromatin conformation, and epigenetic measurements. As a simplification, we assume that the gene expression profile is a sufficient representation of the cell state.

Gene expression for a given cell is dependent on a number of factors, including (but

not limited to): cell type, cell cycle stage, circadian rhythm stage, and growth conditions. In order to best capture the natural fibroblast dynamics from population-level data, time series RNA-seq was performed on cells that were cell cycle and circadian rhythm synchronized in normal growth medium conditions (See Appendix A.2). Prior to data collection, all cells were temporarily held in the first stage of the cell cycle, G_0/G_1, via serum starvation. Upon release into the cell cycle, the population was observed every $\Delta t = 8$ hours (h) for 56 h, yielding 8 time points (at 0, 8, 16, ..., 56 h). Let $g_{i,k}$ be the measured activity of gene $i = 1, \ldots, N$ at measurement time $k = 1, \ldots, 8$, where N is the total number of human genes observed (22,083). Analysis of cell-cycle marker genes indicated that the synchronized fibroblasts took between 32-40 h to complete one cell cycle post growth medium introduction. Because of this, we define $K = 5$ to be the total number of time points used for this model.

Using g to represent x would result in model with over 20,000 variables. To reduce model complexity, we looked for methods to reduce the dimension of x. Conveniently, the genome offers a natural dimension reduction based on TADs. TADs are inherent structural units of chromosomes: contiguous segments of the 1-D genome for which empirical physical interactions can be observed [31]. Moreover, genes within a TAD tend to exhibit similar activity, and TAD boundaries have been found to be largely cell-type invariant [31, 51]. TADs group structurally and functionally similar genes, serving as a natural dimension reduction that preserves important genomic properties. Figure 4.1 depicts an overview of this concept. We computed TAD boundaries from Hi-C data via an algorithm that uses Fielder vector partitioning, described in Chen *et al.* (See Appendix A.2) [32].

Let $tad(i) := j$ if gene i is contained within TAD j. We define each state variable $x_{j,k}$ to be the expression level of TAD $j = 1, \ldots, \tilde{N}$ at time k, where $\tilde{N} = 2,245$ is the total number of TADs that contain genes. Specifically, $x_{j,k}$ is defined as the sum of the expression levels of all genes within the TAD, measured in reads per kilobase

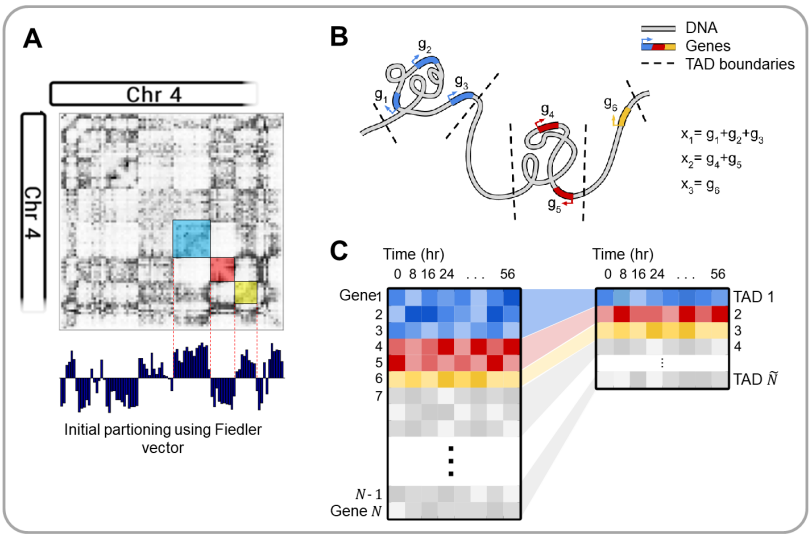


Figure 4.1: Overview of TAD dimension reduction. A) Partitioning the Hi-C matrix based on the Fiedler vector. B) Cartoon depiction of TAD genomic structure. C) TAD dimension reduction summary. This figure was created by Geoff Patterson and Scott Ronquist, and was taken from Ronquist *et al.* [147].

of transcript per million (RPKM), i.e.

$$x_{j,k} := \sum_{\substack{i \text{ s.t.} \\ tad(i)=j}} g_{i,k}. \tag{4.2}$$

The vector of all TAD activities at measurement k is denoted with a single subscript $x_k \in \mathbb{R}^{\tilde{N}\times 1}, k = 1, \ldots, K$.

4.3.2 State transition matrix: $\mathbf{A}_k$

To capture the system dynamics, we define a time dependent state transition matrix A_k as follows:

$$A_k := I_{\tilde{N}} + \frac{(x_{k+1} - x_k)x_k^T}{x_k^T x_k} \in \mathbb{R}^{\tilde{N}\times\tilde{N}}, \qquad k = 1, 2, 3, 4, 5 \tag{4.3}$$

where $I_{\tilde{N}}$ is the $\tilde{N} \times \tilde{N}$ identity matrix. Let the measured values of the state of the unforced evolution be $x_1, x_2, \cdots, x_5$; let the controls be labeled $u_1, u_2, \cdots, u_5$; let the

values of the state with the controls acting be $z_2, z_3, \cdots, z_6$. Letting z denote the deviation from the cell cycle average, we have

$$z_{k+1} = \left(I + \frac{(x_{k+1} - x_k)x_k^T}{x_k^T x_k} \right) z_k + Bu_k$$

where A_k is as above. Solving this difference equation, we have

$$z_k = \prod_{i=1}^{k-1} A_i x_1 + \sum_{i=1}^{k-1} \prod_{j=i}^{k-1} A_{j-1} B u_i$$

with the understanding that $A_0 = I$.

4.3.3 Input matrix and input signal: $\mathbf{B}, \mathbf{u}_k$

With the natural TAD-level dynamics established in the context of our control Eq. 4.1, we turn our attention to quantifying methods for control.

A TF can regulate a gene positively or negatively by binding to a specific DNA sequence near a gene and encouraging or discouraging transcription. The degree to which a TF activates or represses gene expression depends on the specific TF-gene interaction, which is influenced by a variety of factors that are difficult to quantify. Let $w_{i,m}$ be the theoretical *regulation weight* of TF m on gene i, where $w_{i,m} > 0$ ($w_{i,m} < 0$) if TF m activates (represses) gene i, and $m = 1, \ldots, M$, where M is the total number of well-characterized TFs. Weights that are bigger in absolute value, $|w_{i,m}| \gg 0$, indicate stronger transcriptional influence, and weights equal to zero, $w_{i,m} = 0$, indicate no influence.

Extensive TF perturbation experiments would be needed to determine $w_{i,m}$ for each TF m on each gene i. Instead, we propose a simplified method to approximate $w_{i,m}$ from existing, publicly available data for TFBSs, gene accessibility, and average activator/repressor activity. To determine the number of possible binding sites a TF m recognizes near gene i, the reference genome was scanned for the locations of

potential TFBSs following methods outlined by Neph *et al.* (See Appendix A.2) [123]. Position frequency matrices (PFMs), which give information on TF-DNA binding probability, were obtained for 547 TFs from publicly available sources ($\therefore M = 547$). Let $c_{i,m}$ be the number of TF m TFBSs found within ±5kb of the transcriptional start site (TSS) of gene i.

Although many TFs can do both in the right circumstances, most TFs have tendency toward either activator or repressor activity [57]. That is, if TF m is known to activate (repress) most genes, we can say with some confidence that TF m is an activator (repressor), so $w_{i,m} \geq 0$ ($w_{i,m} \leq 0$) for all i. To determine a TF's function, we performed a literature search for all 547 TFs and labeled 299 as activators and 124 as repressors (See Appendix A.2). The remaining TFs were labeled unknown for lack of conclusive evidence and were evaluated as both an activator and a repressor in separate calculations. Here, we define a_m as the activity of TF m, with 1 and -1 denoting activator and repressor, respectively.

TFBSs are cell-type invariant since they are based strictly on the linear genome. However, it is known that for a given cell type, certain areas of the genome may be opened or closed depending on epigenetic aspects. To capture cell type specific regulatory information, we obtained publicly available gene accessibility data (DNase-seq) on human fibroblasts (GSM1014531). DNase-seq extracts cell type specific chromatin accessibility information genome-wide by testing the genome's sensitivity to the endonuclease DNase I, and sequencing the non-digested genome fragments. These data are used for our initial cell type to determine which genes are available to be controlled by TFs [172]. Here, we define s_i to be the DNase I sensitivity information (accessibility; open/close) of gene i in the initial state, with 1 and 0 denoting accessible and inaccessible, respectively (See Appendix A.2).

We approximate $w_{i,m}$ as

$$w_{i,m} := a_m s_i c_{i,m}, \tag{4.4}$$

so that the magnitude of influence is equal to the number of observed consensus motifs $c_{i,m}$, except when the gene is inaccessible ($s_i = 0$) in which case $w_{i,m} = 0$.

Since we are working off a TAD-dimensional model, our input matrix B must match this dimension. Let b_m be a 2,245-dimensional vector, where the j^{th} component is

$$b_{j,m} := \sum_{\substack{i \text{ s.t.} \\ tad(i)=j}} w_{i,m} \tag{4.5}$$

and define a matrix $B = \begin{bmatrix} b_1 \; b_2 \; \cdots b_M \end{bmatrix}$.

The amount of control input is captured in u_k, which is a $\mathbb{R}^{M \times 1}$ vector representing the quantity of the external TFs we are inputting to the system (cell) at time k. This can be controlled by the researcher experimentally through manipulation of the TF concentration [19]. In this light, we restrict our analysis to $u_k \geq 0$ for all k, as TFs cannot be subtracted from the cell. $u_{m,k}$ is defined as the amount of TF m to be added at time point k.

With all variables of our control Eq. 4.1 defined, we can now attempt to predict which TFs will most efficiently achieve cellular reprogramming from some x_I (initial state; fibroblast in our setting) to x_T (target state; any human cell type for which compatible RNA-seq data is available) through manipulation of u_k. An overview of our DGC framework is given in Figure 4.2.

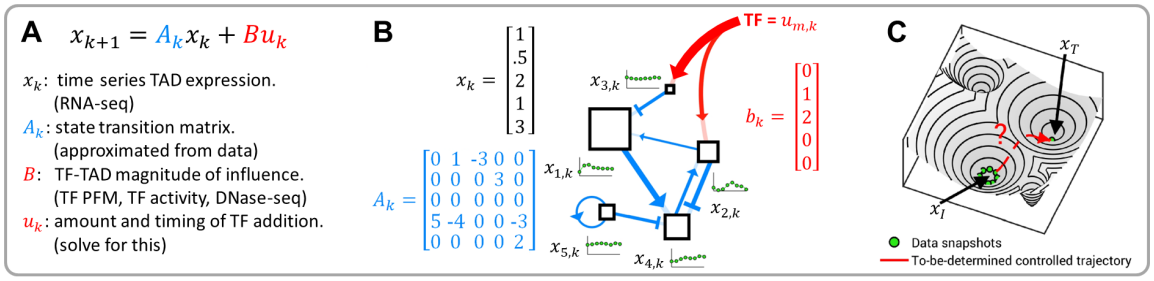


Figure 4.2: Data-guided control overview. A) Summary of control equation variables. B) Each TAD is a node in a dynamic network. The blue connections represent the edges of the network and are determined from time series fibroblast RNA-seq data. The green plots represent the expression of each TAD changing over time. The red arrows indicate additional regulation imposed by exogenous TFs. C) A conceptual illustration of the problem: can we determine TFs to push the cell state from one basin to another? This figure was created by Geoff Patterson and Scott Ronquist, and was taken from Ronquist *et al.* [147].

4.3.4 Selection of TFs

Our general procedure for scoring TFs is explained as follows. Eq. 4.1 has an explicit solution that is given below. The first few terms are

$$
\begin{aligned}
z_2 &= A_1 x_1 + B u_1 \\
z_3 &= A_2 A_1 x_1 + A_2 B u_1 + B u_2 \\
z_4 &= A_3 A_2 A_1 x_1 + A_3 A_2 B u_1 + A_3 B u_2 + B u_3 \\
&\vdots
\end{aligned}
$$

This shows how z_4 depends on u_1, u_2, and u_3.

If x_T is a target condition, then the Euclidean distance $\|\cdot\|$ can be used to measure how close a state is to the target state. We define

$$
d = \|x_T - z_6(u)\|, \tag{4.6}
$$

where the notation $z_6(u)$ is used to emphasize the dependence of z_6 on u. Considering all possible input signals, one can compute the optimal control that finds the minimum

distance for a given initial and target cell type. Let u_* denote the optimal u used to minimize d, and d_* denotes this minimum distance value.

When appropriate, we write $z_6(u)$ to emphasize the fact that the final state depends on the input. The Euclidean distance $\|\cdot\|$ can be used to measure how close a given state is to the target. If there were no restrictions on the u terms, the control that minimizes the distance between z_6 and the target could be computed without difficulty. However, there are reasons for restricting the number of different TFs used in any one trial. Transfection of cells with too many TFs can lower the efficiency of transfection and even lead to cell death. Moreover, many confirmed direct reprogramming experiments use $\leq$4 TFs to achieve reprogramming. For these reasons, we modify the optimization problem by adding the constraint that there are no more than a fixed number TFs (components of u) used in a given trial.

Let $\hat{p}$ be a set of integers that identifies the subset of the components of u (read: TFs) that are allowed to be non-zero. For example, $\hat{p} = \{1, 4, 7\}$ refers to TFs 1, 4, and 7. Let p be the number of elements in $\hat{p}$. Given a set of TFs, $\hat{p}$, we determine the quantity and timing of TF input, u_{*k}, that minimizes the difference between x_6 and the target cell state, x_T. Mathematically, this can be written as

$$\begin{aligned} \underset{u}{\text{minimize}} \quad & \|x_T - z_6(u)\| \\ \text{subject to} \quad & \begin{cases} u_{m,k} \geq 0, & k = 1, ..., 5 \\ u_{m,k} = 0, & \text{if } m \notin \hat{p} \\ u_{m,k+1} \geq u_{m,k} \end{cases} \end{aligned} \tag{4.7}$$

We use MATLAB's *lsqnonneg* function to solve Eq. 4.7, which gives u_{*k} and d_*.

Let $d_0 := \|x_T - x_0\|$, be the distance between the final state and target state with no control input. Define a score $\mu := d_0 - d_*$, which can be interpreted as the

improvement provided by a particular choice of u. This can be calculated for each $\hat{p}$ and sorted (high to low) to determine which TF or TF combination is the best candidate for direct reprogramming between x_0 and x_T.

We consider different scenarios for the type of input regime in the results. The first assumes the input signal is constant $u_1 = u_k = \bar{u}$, intended to mimic empirical regimes where TFs are given at a single time point. Later, we also consider inputting TFs at different times $\hat{k}$, which can be viewed mathematically as requiring $u_{m,k} = 0$ for all $k < \hat{k}$, and $u_{m,k}$ is a constant value for all $k \geq \hat{k}$. This is intended to mimic inputting a TF at time $\hat{k}$, which will continue to express at a constant level until time point $k = 6$.

Remark: Subsets of TFs were chosen for each calculation based on the following criteria: $\geq$10-fold expression increase in target state compared to initial state, and $\geq$10 RPKM in target state. These criteria are used to select differentially expressed TFs and TFs that are sufficiently active in the target state.

4.4 Results

4.4.1 Quantitative measure between cell types

To predict TFs for reprogramming, compatible data on target cell types must be collected. For this, we explore a number of publicly available databases where RNA-seq has been collected, along with RNA-seq data collected in our lab. The ENCODE Consortium has provided data on myotubes and ESCs (See Appendix A.2) [36]. The GTEx portal provides RNA-seq data on a large variety of different human tissue types [106]. Although each GTEx experiment is performed on tissue samples, thus containing multiple different cell types, we use these data as more general cell state targets.

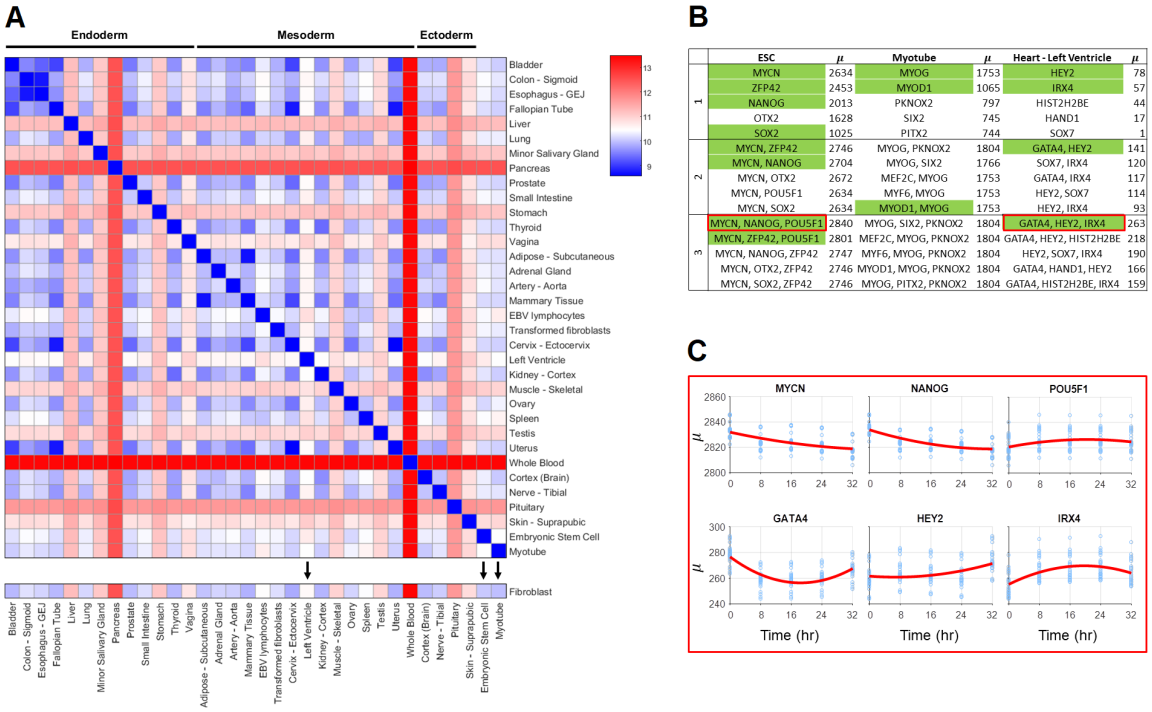
To give a numerical structure to cell type differences, conceptually similar to

Waddington's epigenetic landscape, we calculate d_0 between all cell types collected. Figure 4.3*A* shows d_0 values for 32 tissue samples collected from the GTEx portal, along with ESC, myotube, and our fibroblast data (additional cell type d_0 values shown in Appendix A.2). GTEx RNA-seq data is scaled to keep total RPKM difference between time series fibroblast and GTEx fibroblast RNA-seq minimal (See Appendix A.2).

4.4.2 TF scores

To assess our method's predictive power, a subset of target cell types are presented here that have either validated TF reprogramming methods or TFs highly associated with the target cell type. Additional predicted TFs for reprogramming are included in Appendix A.2. We note that though experimentally validated TFs provide the best current standard for comparison, we believe experimental validation with our predicted TFs may provide more efficient and comprehensive reprogramming results. For all reprogramming regimes presented in this section, fibroblast is used as the initial cell type due to the availability of synchronized time series data, and all TFs are introduced at $k = 1$ [31].

For conversion of fibroblast to myotubes, the top predicted single input TFs are MYOG and MYOD1, both of which are known to be crucial for myogenesis. While MYOD1 is the classic master regulator reprogramming TF for myotube conversion, activation of downstream factor MYOG is necessary for full conversion [183]. For fibroblast to ESC conversion, a number of TFs known to be necessary for pluripotency are predicted, including MYCN, ZFP42, NANOG, and SOX2 [166]. With the knowledge that no single TF has been shown to fully reprogram a fibroblast to an embryonic state, combinations of TFs are more informative for this analysis. The top scoring combination of 3 TFs is MYCN, NANOG, and POU5F1 - three well-known markers for pluripotency [166]. Interestingly, POU5F1 scores poorly when input in-



	ESC	μ	Myotube	μ	Heart - Left Ventricle	μ
1	MYCN	2634	MYOG	1753	HEY2	78
	ZFP42	2453	MYOD1	1065	IRX4	57
	NANOG	2013	PKNOX2	797	HIST2H2BE	44
	OTX2	1628	SIX2	745	HAND1	17
	SOX2	1025	PITX2	744	SOX7	1
2	MYCN, ZFP42	2746	MYOG, PKNOX2	1804	GATA4, HEY2	141
	MYCN, NANOG	2704	MYOG, SIX2	1766	SOX7, IRX4	120
	MYCN, OTX2	2672	MEF2C, MYOG	1753	GATA4, IRX4	117
	MYCN, POU5F1	2634	MYF6, MYOG	1753	HEY2, SOX7	114
	MYCN, SOX2	2634	MYOD1, MYOG	1753	HEY2, IRX4	93
3	MYCN, NANOG, POU5F1	2840	MYOG, SIX2, PKNOX2	1804	GATA4, HEY2, IRX4	263
	MYCN, ZFP42, POU5F1	2801	MEF2C, MYOG, PKNOX2	1804	GATA4, HEY2, HIST2H2BE	218
	MYCN, NANOG, ZFP42	2747	MYF6, MYOG, PKNOX2	1804	HEY2, SOX7, IRX4	190
	MYCN, OTX2, ZFP42	2746	MYOD1, MYOG, PKNOX2	1804	GATA4, HAND1, HEY2	166
	MYCN, SOX2, ZFP42	2746	MYOG, PITX2, PKNOX2	1804	GATA4, HIST2H2BE, IRX4	159

Figure 4.3: Quantitative measure between cell types and TF scores. A) d_0 values between GTEx tissue types and ESC, myotube, and fibroblast. Tissue types and cell types with black arrows have predicted TFs for reprogramming from fibroblasts shown in Figure 4.3B. B) Table of predicted TFs for a subset of cell and tissue types. Top 5 TFs for combinations of 1-3 shown. Green labeled TFs are either highly associated with the differentiation process of the target cell type and/or validated for reprogramming. These TFs are discussed in the main text. C) Time-dependent scores for selected combinations of 3 TFs for fibroblast to ESC and fibroblast to "Heart - Left ventricle." x-axis refers to time of TF addition, y-axis refers to μ. This figure was taken from Ronquist *et al.* [147].

dividually, but is within the top set of 3 TFs when used in combination with MYCN and NANOG. Left ventricle reprogramming includes TFs that are known to be necessary for natural differentiation in the top score for all 1-3 combinations. These include GATA4 (a known TF in fibroblast to cardiomyocyte reprogramming), HEY2, and IRX4 [59, 79, 122].

4.4.3 Time-dependent TF addition

Fibroblast to ESC conversion was of particular interest in our analysis as this is a well-studied regime with a number of validated TFs (with a variety of reported efficiencies), and this conversion is promising for its regenerative medicine application. High scoring TFs yield many that are known markers for pluripotency, but the top combination of 3, MYCN, NANOG, and POU5F1, has not been used specifically together, to our knowledge. Here, we analyzed how the TF combination would score if input at different points throughout the cell cycle.

Time-dependent analysis of the top scoring ESC TFs reveals that scores vary widely depending on the time of input. MYCN and NANOG show a strong preference for input at the beginning of the cell cycle, while POU5F1 shows a slight preference for input towards the end of the cell cycle, with the highest score achieved when MYCN and NANOG are input at 0 h and POU5F1 is input at 32 h. Analysis on how the time of input control affects μ is shown in Figure 4.3*C*. Time-dependent analysis was also conducted for the top combination of 3 TFs for fibroblast to left ventricle. This analysis predicted that the best reprogramming results would occur if GATA4 is given immediately (0 h), with IRX4 and HEY2 given later (24 and 32 h, respectively).

4.5 Discussion

The results from this algorithm show promise in their prediction of known reprogramming TFs, and demonstrate the importance of including time series data for

gene network dynamics. Time of input control has shown to have an impact on the end cell state, in line with what has been shown in natural differentiation [104].

While we believe that this is the best model currently available for predicting TFs for reprogramming, we are aware of its limitations and assumptions. TAD-based dimension reduction is based on the observation that genes within them correlate in expression over time, though we lack definitive proof of regulation by shared transcriptional machinery [31]. This assumption was deemed necessary for dimension reduction in the context of deriving transition matrix A_k. With finer time steps in RNA-seq data, the assumption may not be necessary for TF prediction, at the cost of increased computation time. Additionally, a 5kb window flanking the TSS of each gene was used to ensure that all potential regulators are found, at the cost of potential inclusion of false positive motifs.

Although this program can score TFs relative to other TFs in a given reprogramming regime, it is difficult to predict a μ threshold that would guarantee conversion. Additionally, rigorous experimental testing will be required to validate these findings and determine how our u vector translates to TF concentration. This is a product of the large number of assumptions that we have made to develop the initial framework for a reprogramming algorithm. With finer resolution in the time series gene expression, more subtle aspects of the genomic network may be observed, allowing for better prediction.

Our proposed data-guided control framework successfully identified known TFs for fibroblast to ESC and fibroblast to muscle cell reprogramming regimes. We employ a biologically-inspired dimension reduction via TADs, a natural partitioning of the genome. This comprehensive state representation was the foundation of our framework, and the success of our methods motivates further investigation of the importance of TADs as functional units to control the genome.

A dynamical systems view of the genome allows for analysis of timing, efficiency,

and optimality in the context of reprogramming. Our framework is the first step toward this view. The successful implementation of time-varying reprogramming regimes would open new avenues for direct reprogramming. Experimental verification of predicted regimes and development of methods to identify optimal sets of TFs are planned for the near future. This template can be used to develop regimes for changing any cell into any other cell, for applications that include reprogramming cancer cells and controlling the immune system. Our DGC framework is well equipped for designing personalized cellular reprogramming regimes. Finally, this framework can serve as a general technique for investigating the controllability of networks strictly from data.

CHAPTER V

MYOD1-mediated Fibroblast to Muscle Reprogramming

This chapter is based on a paper by Sijia Liu, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Pin-Yu Chen, Gerald Higgins, Pierre Baldi, Steve Smale, Alfred Hero, Lindsey A. Muir, and Indika Rajapakse [103]. I have summarized my contributions to this paper here.

5.1 Abstract

In the first cellular reprogramming experiment, Dr. Harold Weintraub reprogrammed fibroblast cells to the muscle lineage using the master TF MYOD1. Though many cellular reprogramming transitions have been discovered since this revelation, the genomic changes that facilitate these conversions are not well understood. Here we collect and analyze 4DN data derived from a population of human fibroblasts that are undergoing MYOD1-mediated cellular reprogramming to the muscle cell lineage. Time series Hi-C and RNA-seq samples are collected to analyze genome structure and function, respectively. From these data, we find that structural changes precede functional changes during this course of reprogramming. Additionally, we find that the transduction of MYOD1 robustly synchronizes the circadian rhythm cycle

of these cells. This work helps explain how cells can transition between cell lineages and uncovers a role for MYOD1 in the core circadian gene network.

5.2 Introduction

During cellular reprogramming, TFs that are introduced into the initial cell type bind to the genome, change nearby genome structure and expression, and steer the cell state towards a desired cell type. While the start and end states of cellular reprogramming have been characterized in some experiments, the process by which the TFs bring about this change is not well understood [65, 166, 183]. To observe these changes, high resolution data sets which measure genome structure and function must be created from cells undergoing reprogramming. Furthermore, since cellular reprogramming is a dynamic process, these data set must have a high temporal resolution as well. Fortunately, recent experimental techniques for probing genome structure and function have been discovered. Hi-C can detect over 100 million genomic interactions in a population of cells, allowing researchers to observe genome architecture genome-wide. To observe genome function, RNA-seq can be used to accurately quantify the transcript abundance of a cell population. Yet, few efforts have been made to characterize the relationship between genome structure, function, and time during cellular reprogramming. Furthermore, the affect that cellular reprogramming has on cell type invariant biological rhythms has not been characterized as well. Thus, a concerted effort to study all the modalities that comprise the 4DN movement (i.e. structure, function, phenotype, and time) during cellular reprogramming will be invaluable to research community [31, 52, 60, 93].

In this work we examined the dynamical interactions between genome structural features and transcription in human fibroblasts undergoing MYOD1-mediated reprogramming into the myogenic lineage. Sampling across a time course during reprogramming, we captured structure by Hi-C, and transcription by RNA-seq. To better

understand the features of genome structure and expression in a dynamical setting, we adopt a network point of view. Nodes of the network correspond to genomic loci that can be partitioned at different scales, for example into larger scale 1 Mb regions or smaller scale gene level regions. The edges of the network indicate contact between two genomic loci, with edge weights given by Hi-C entries. From the network perspective, A/B compartments are identified as distinct connected nodes of a network.

To further reveal chromatin spatial organization, we use network centrality measures. Using network centrality enables identification of nodes that play influential topological roles in the network [125]. Several centrality measures exist, each specialized to a particular type of nodal influence. For example, degree centrality characterizes the local connectedness of a node as measured by the number of edges connecting to this node, while betweenness centrality is a global connectedness measure that quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Eigenvector centrality is a neighborhood connectedness property in which a node has high centrality if many of its neighbors also have high centrality. In other words, a node is important if it is connected to other important nodes. For reference, Google's PageRank algorithm uses a variant of eigenvector centrality [105].

By examining different centrality measures we have discovered important features in Hi-C data largely overlooked in previous studies. We also found that cells undergoing reprogramming have significant architectural reorganization prior to changes in transcription and subsequently show potent activation of the myogenic program that ties into regulation of biological rhythms.

5.3 Results

5.3.1 Myogenic reprogramming of human fibroblasts

We converted primary human fibroblasts into the myogenic lineage using the TF and master regulator MYOD1, following Weintraub's method for myogenic reprogramming [183]. Fibroblasts were transduced with a lentiviral construct that expressed human MYOD1 fused with a tamoxifen-inducible ER(T) domain (L-MYOD1) [89]. With 4-hydroxytamoxifen (4-OHT) treatment, transduced cells showed nuclear translocation of L-MYOD1, morphological changes consistent with expression of key myogenic genes downstream of *MYOD1* (*MYOG* and *MYH1*), and myogenic differentiation. These data demonstrate conversion of fibroblasts into the myogenic lineage by L-MYOD1.

We used this system to delineate the dynamics of genome structure and transcription underlying direct cellular reprogramming. Analyses were carried out on transduced, 4-OHT treated cells, sampling at 8-hour (hr) intervals for RNA-seq (three replicates per time point, small RNA-seq, and Hi-C (single replicates per time point), and at 24-hr intervals for proteomics (Figure 5.1A).

We evaluated up to 16 time points (-48, 0, ..., 112 hrs) for genome structure (form) through Hi-C and transcription (function) through RNA-seq. The resulting time series data were studied at different scales (Figure 5.1B). Scale was based on units of length along the linear genome (1 Mb, 100 kb) or by structurally/functionally defined units of the genome, such as TADs or individual genes.

5.3.2 Architectural changes precede activation of the myogenic program

Given the cell state trajectory, it was unclear whether MYOD1-mediated reprogramming induced rewiring of genome architecture prior to the role of MYOD1 in mediating muscle gene transcription, or vice versa [91, 136]. To answer this question, we

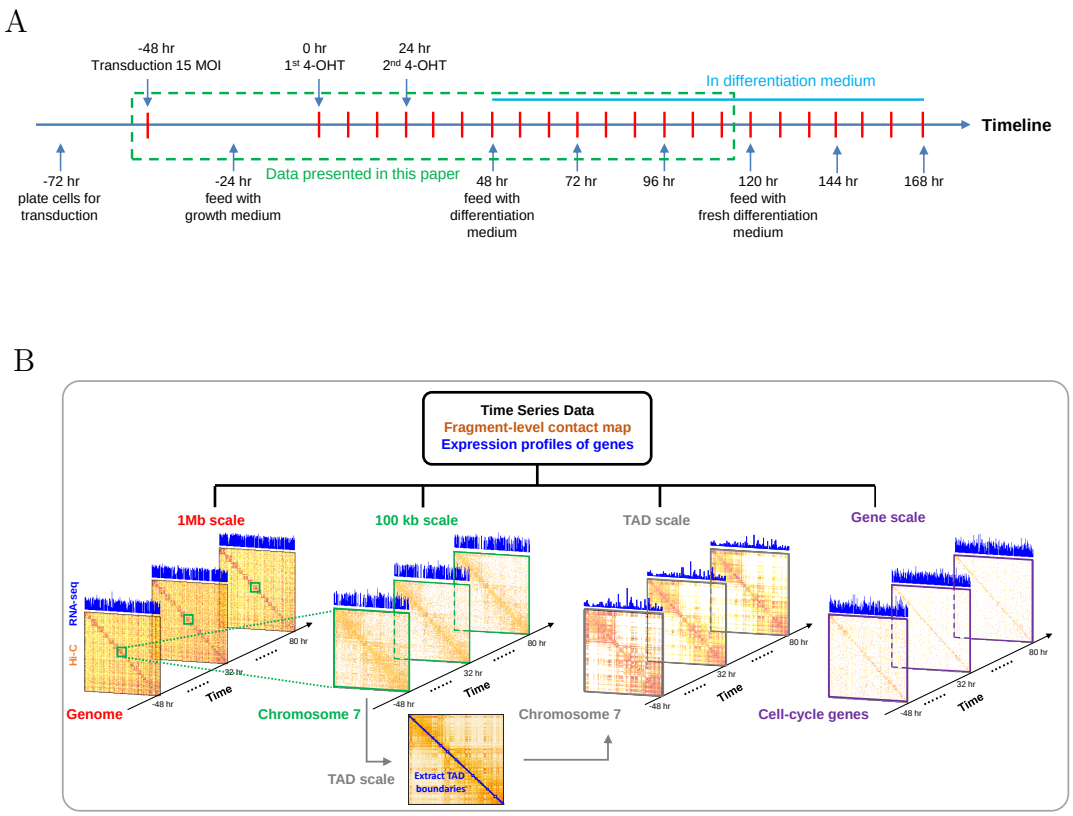


Figure 5.1: Myogenic reprogramming of human fibroblasts. A) Time course of MYOD1-mediated reprogramming. The time window outlined in green corresponds to time points at which both genome structure and transcription were captured by Hi-C (single replicates) and RNA-seq (in triplicate). B) Scale-adaptive Hi-C matrices and gene expression. The considered scales include 1 Mb, 100 kb, TAD and gene-level. This figure was created by Sijia Liu and Scott Ronquist, and was taken from Liu *et al.* [103].

focused on form and function dynamics of 22083 genes genome-wide, where the form is depicted by inter-gene Hi-C contact maps (See Appendix A.3), and the function corresponds to RNA-seq FPKM values (Figure 5.2A). The form-function evolution is then evaluated by determining the difference in network centrality features (extracted from inter-gene contact maps) and gene expression between successive time points. We refer to this measure as temporal difference score (TDS; See Appendix A.3). Based on TDS at successive time points (Figure 5.2B), we found that a significant form change at 8 hrs preceded a significant function change at 16 hrs.

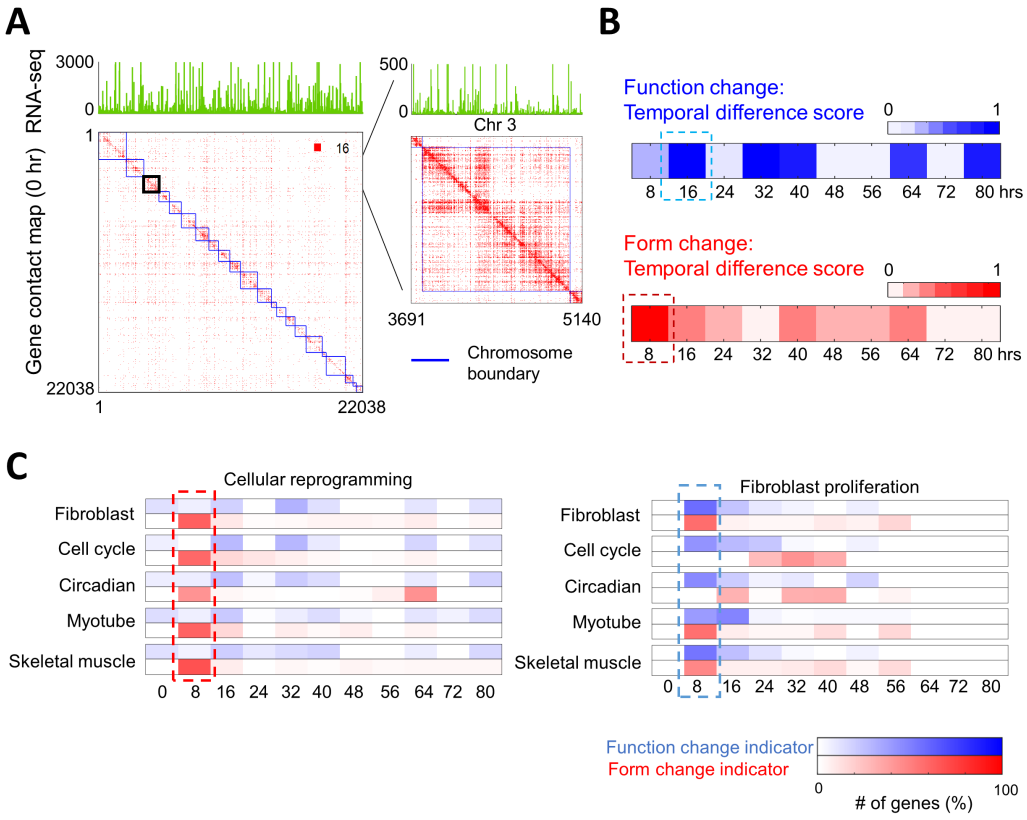


Figure 5.2: Changes in genome architecture precede activation of the myogenic program. A) Genomic structure (form) and gene expression (function) given by a Hi-C contact map and RNA-seq. Hi-C and RNA-seq are constructed at gene-level resolution. B) Function and form change at successive time points evaluated by temporal difference score (TDS; See Appendix A.3) of RNA-seq and network centrality features of Hi-C data, respectively. The significant form change (at 8 hrs) occurs prior to the function change (at 16 hrs). C) Form-function change indicators for gene modules of interest during cellular reprogramming (left) and fibroblast proliferation (right), respectively. Here each row represents one gene module of interest, each column represents a time step, and the amount of change, as a percentage of total change over time for each module, is depicted by color. Percentage is determined by finding the number of genes with significant form-function change for each module and time step, and dividing this number by the total number of significant gene changes for each module over time (row). This figure was created by Sijia Liu and Scott Ronquist, and was taken from Liu *et al.* [103].

We then contrasted our reprogramming data with data on human fibroblast proliferation. Data on proliferating human fibroblasts were previously obtained using similar methods over a time course [31] after cell cycle and circadian rhythm synchronization, with collection of RNA-seq and Hi-C every 8 hours. We found that

the pattern of form-function evolution during reprogramming is quite different from fibroblast proliferation (Figure 5.2C). Consistent with findings represented in Figure 5.2B, the effects of nuclear reorganization were detectable prior to transcription changes, that is, form preceded function. Given these results, we propose that chromatin architectural changes facilitate the orchestrated activation of transcriptional networks associated with adoption of a new cell identity.

5.3.3 Early stage chromatin remodeling

We additionally sought to understand regulatory dynamics during reprogramming, including early-stage gene expression dynamics related to chromatin remodeling, super enhancer dynamics, and microRNA expression. Examination of early stage RNA-seq data [-48, 16] (hrs) revealed endogenous mechanisms relevant to *MYOD1* transcriptional activation including muscle stage-specific markers and chromatin remodeling factors (See Figure 5.3A). At 16 hrs, the combined upregulation of *DES*, *MYL4*, *TNNT1* and *TNNT2* suggests myogenic differentiation [62, 152]. *EZH2* has been associated with both "safe-guarding" the transcriptional identity of skeletal muscle stem cells and with terminal differentiation of myoblasts into mature muscle [88]. *ARID5A*, a regulator of the myotube BAF47 chromatin remodeling complex, is significantly upregulated at 8 hours ($P = 7.2 \times 10^{-5}$) and may act to enhance MYOD1 binding to target promoters [86]. *NR4A3*, *MEF2D*, *SIX4*, *SIX1*, and *SOX4* expression are also increased at 8 hr, all of which have important regulatory functions during differentiation in the myogenic lineage (See Figure 5.3B) [11, 58, 84].

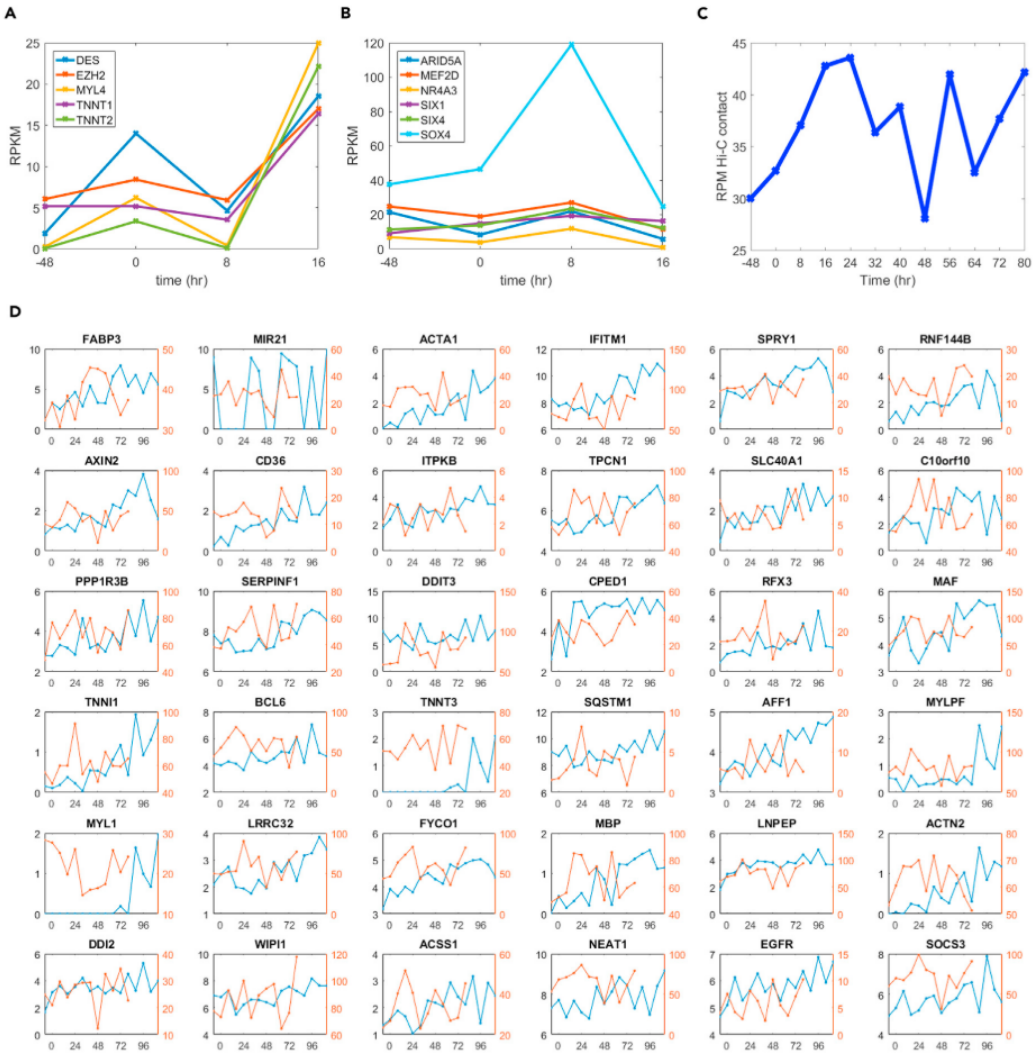


Figure 5.3: Increased genomic contacts among myogenic regulatory elements set the stage for reprogramming. A) Early-phase expression dynamics of genes related to muscle cell terminal differentiation and chromatin remodeling, including components of the contractile apparatus (DES, MYL4, TNNT1, TNN2), and EZH2, a repressor that is involved in myogenesis. B) Expression of chromatin remodeling factors and master TFs in the early phases of cellular reprogramming. These factors include ARID5A, part of the BAF47 muscle remodeling complex which acts in cooperation with MYOD1, MEF2D, which drives differentiation of myotubes to skeletal and cardiac muscle, NR4A3 (aka NOR1) involved in differentiation of myotubes into smooth muscle, and SIX1, SIX4 and SOX4, which control the differentiation of myotubes into muscle cells. C) structure and function of super enhancers and associated genes over time. Average Hi-C RPM contact between potential super enhancer and associated gene TSS regions over time, as defined by [75]. D) Top upregulated SE-P genes, log2(FPKM) (Blue) and SE-P Hi-C normalized contact (Red; see Appendix A.3) over time. This figure was taken from Liu *et al.* [103].

We also investigated how muscle-related super enhancer-promoter (SE-P) interactions change over time throughout MYOD1-mediated reprogramming. To capture these dynamics, we extracted the Hi-C contact between skeletal muscle super enhancer regions and associated genes transcription start site (TSS) (±1kb), as determined by [75] (618 SE-P regions; See Appendix A.3). We observed that for these skeletal muscle SE-P Hi-C regions, the strongest amount of contact occurred relatively early in the reprogramming process, peaking 16-24 hrs post-L-MYOD1 addition to the nucleus ($P = 4.17 \times 10^{-9}$, Figure 5.3C; See Appendix A.3). Exact SE-P contact vs function trends were variable, but a number of important myogenesis genes, such as *TNNI1*, *MYLPF*, *ACTN2*, and *TNNT3* show strong upregulation in function over time, with an increase in SE-P contact post-MYOD1 activation. Contact vs. function trends for the top 36 upregulated genes are shown in Figure 5.3D (See Appendix A.3).

5.3.4 Linking myogenic genes with entrainment of biological rhythms

Several studies have explored the link between MYOD1 and circadian genes ARNTL and CLOCK, revealing that ARNTL and CLOCK bind to the core enhancer of the *MYOD1* promoter and subsequently induce rhythmic expression of *MYOD1* [3, 191]. Here we discovered that upon *MYOD1* activation, circadian genes exhibited robust synchronization in gene expression, suggesting MYOD1 feedback onto the circadian gene network. Further inspection showed that core circadian genes (Table S3 in Liu *et al.* [103]) that contain E-boxes displayed the most profound synchronization initially, starting with an uptick in gene expression just after MYOD1 activation (Figure 5.4A-D). Analysis using JTK_CYCLE [77] confirmed our observation; all E-box circadian genes were found to have a synchronized period of 24 hrs, with a maximum lag of 4 hrs between genes, with the exception of *CRY1* (Table S6 in Liu *et al.* [103]).

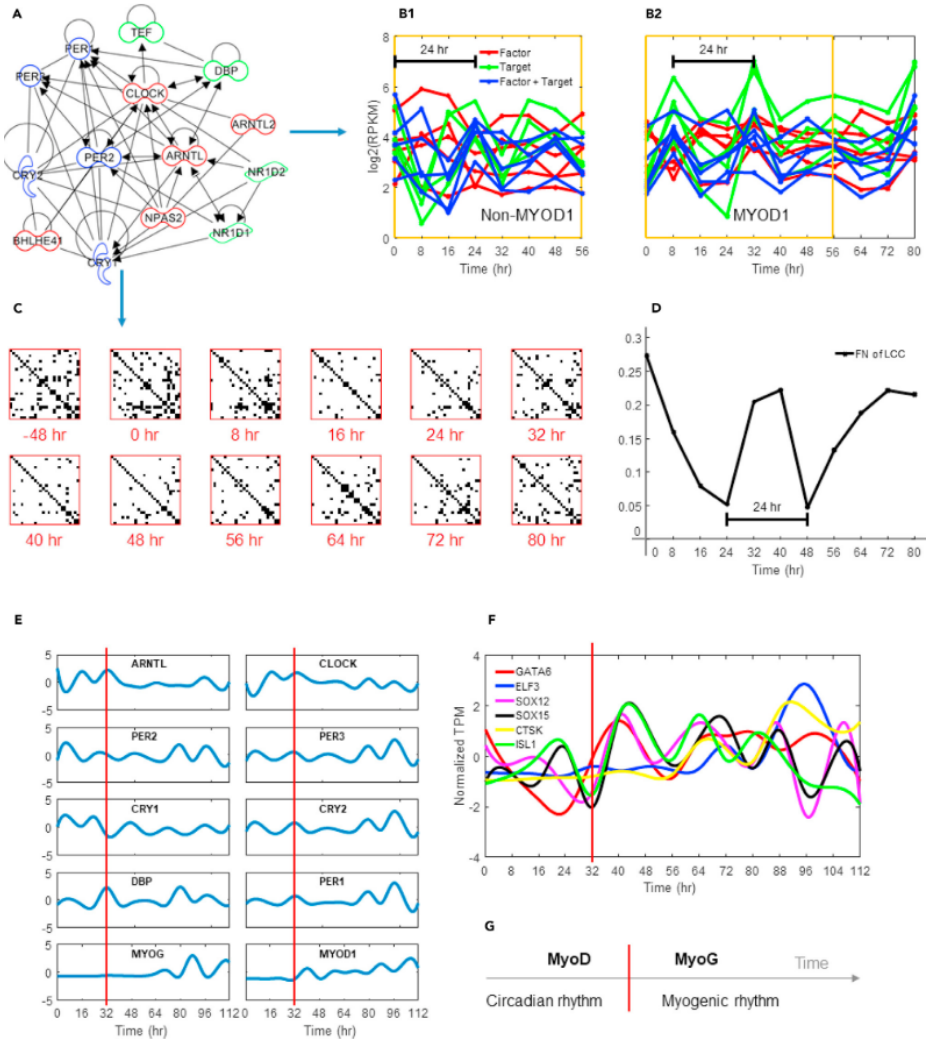


Figure 5.4: Myogenic genes participate in entrainment of biological rhythms. A) Gene network interactions between circadian E-box genes, derived from Ingenuity Pathway Analysis. B) Core circadian gene expression over time. B1) Dexamethasone synchronization. B2) L-MYOD1 synchronization. Target and factor correspond to genes with E-box targets and TFs that bind to E-box genes, respectively. C) Hi-C contacts between 26 core circadian genes over time (See Table S3 in Liu *et al.* [103]). Rows and columns correspond to core circadian genes, contacts are binary. D) Network connectivity of the largest connected component of the studied Hi-C contact maps at different time points. E) Normalized gene expression (FPKM, cubic spline) highlighting oscillation dampening after 32 hrs (red line) and the switch to differentiation medium for select core circadian genes; MYOD1 and MYOG also shown. F) Normalized TPM of TFs that are targeted by *MYOG* or *MYOD1* (*ELF3*) and that only showed oscillation after 32 hrs (red line). G) Conceptual diagram of biological rhythm entrainment during MYOD1-mediated reprogramming, where the red line signifies the bifurcation event. This figure was taken from Liu *et al.* [103].

The subset of transcripts with oscillatory behavior was different before and after the 32 hr time point. Endogenous *MYOD1* and *MYOG* expression began close to 32 hrs and both transcripts displayed oscillatory expression. Additionally, circadian transcript oscillations dampened at 40 hrs, coinciding with the switch to low-serum differentiation medium (Figure 5.4E). To determine which newly oscillating transcripts were potential targets of MYOD1 and MYOG, we further investigated which transcripts have MYOD1 or MYOG binding motifs in their promoters using MotifMap [43], and which were synchronized in expression with MYOD1 and MYOG. Among the oscillating transcripts that fit these criteria, we found six TFs that were oscillatory only after the 32 hr critical transition point, have upstream MYOG binding sites, and were synchronized in expression with MYOG. Of these six TFs, only *ELF3* was found to have binding motifs for *MYOD1*, as well as synchronized expression with *MYOD1* (Figure 5.4F). Several of the six oscillatory TFs targeted by MYOG or MYOD1 are associated with muscle developmental and differentiation processes, including SOX15 [112], GATA6 [186], ISL1 [128], and ELF3 [14].

Robust synchronization in expression of circadian genes that are downstream targets of MYOD1 suggests MYOD1 feedback onto circadian gene circuits. After the 32 hr critical transition point, MYOG was associated with synchronized expression of a subset of important myogenic TFs. These findings support regulatory roles for MYOD1 and MYOG in entraining circadian and cell type-specific biological rhythms.

5.4 Discussion

In this study, we analyzed MYOD1-mediated reprogramming of human fibroblasts into the myogenic lineage from a dynamical network perspective. Distinct from previous studies, we generated an enriched time-series data set including Hi-C, RNA-seq, miRNA, and proteomics data. This provides a comprehensive genome-wide form-function description over time, and allows us to detect early stage cell-fate commit-

ment changes during cellular reprogramming. Capturing these dynamics may help us identify genes that are key players in other reprogramming settings, and develop a more universal understanding of the process and requirements for reprogramming between any two cell types.

A number of studies have explored the link between *MYOD1* and circadian genes *ARNTL* and *CLOCK*, revealing that ARNTL and CLOCK bind to the core enhancer of the *MYOD1* promoter and subsequently induce rhythmic expression of *MYOD1* [3, 191]. We found that upon activation of L-MYOD1, the population of cells exhibits robust synchronization in circadian E-box gene expression. Among these E-box targets are the *PER* and *CRY* gene family, whose protein products are known to repress CLOCK-ARNTL function, thus repressing their own transcription. Additionally, E-box target NR1D1, which is synchronized upon addition of L-MYOD1, competes with ROR proteins to repress *ARNTL* transcription directly. This adds another gene network connection under MYOD1 influence, indirectly acting to repress *ARNTL*, leading us to posit that MYOD1 can affect CLOCK-ARNTL function through E-Box elements, in addition to CLOCK-ARNTL's established activation effect on MYOD1. Furthermore, these oscillations dampen just after the 32 hr point, after which MYOG entrains the oscillations of a distinct subset of myogenic TFs. Therefore, MYOD1-mediated reprogramming and circadian synchronization are mutually coupled, consistent with other systems that modulate cell fate [49, 174].

Our proposed bio- and computational technologies shed light on the hypothesis that nuclear reorganization occurs at the time of cell specification and both precedes and facilitates activation of the transcriptional program associated with differentiation (or reprogramming), i.e. form precedes function [136]. The alternative hypothesis is that function precedes form, that is nuclear reorganization occurs because of differential transcription and is a consequence of, rather than a regulator of, differentiation programs [91]. Our findings support that nuclear reorganization occurs prior to gene

transcription during cellular reprogramming, i.e., form precedes function, and that dynamical nuclear reorganization plays a key role in defining cell identity. Our data do not establish a causal relationship, and for this, additional experiments will be necessary. For example, Hi-C and RNA-seq can be supplemented using MYOD1 ChIP-seq to identify the regions of greatest adjacency differences between cell types that correlate with transcription and/or MYOD1 binding.

Understanding the dynamical process of cellular reprogramming is critical in regenerative medicine. Comprehensive 4DN studies can improve our ability to guide cells toward repair and regeneration of tissue in injury and disease. Furthermore, identifying a structural function for TFs that is distinct from transcription would define a new molecular function with as yet an unknown role in development and disease.

CHAPTER VI

The 4DN of Cancer

This chapter combines the analysis of two 4DN data sets derived from cancer cell lines. Additionally, Section 6.2, "The 4DN of single chromosome aneuploidy," is based on a paper by Rüdiger Braun, Scott Ronquist, Darawalee Wangsa, Haiming Chen, Lena Anthuber, Timo Gemoll, Danny Wangsa, Vishal Koparde, Cynthia Hunn, Jens K. Habermann, Kerstin Heselmeyer-Haddad, Indika Rajapakse, and Thomas Ried [18]. I have summarized my contributions to this paper here.

6.1 Abstract

Chromosomal aneuploidy and genetic mutations have a direct effect on the 4DN of cells. For example, gene copy number variations can cause an increase in the gene's RNA abundance and this misregulation can propagate through the gene's regulatory network, steering the system towards an uncontrolled proliferative state. A real a example of this is observed in colorectal cancer. Most sporadic colorectal carcinomas carry extra copies of chromosome 7, an aneuploidy that emerges in premalignant adenomas, and is maintained throughout tumor progression and in derived cell lines. A comprehensive understanding on how chromosomal aneuploidy affects nuclear organization and gene expression, i.e., the nucleome, remains elusive. Additionally, chromosome aberrations can steer the system to a more pluripotent state. CSCs are

a subpopulation of cancer cells that exhibit "stem–like" properties. These cells have upregulated expression of pluripotency genes and have the ability to recapitulate the entire tumor microenvironment. This stemness property is believed to make the cells more resistant to treatment, as CSCs are much more plastic in their cell state. Furthermore, CSCs are believed to be more quiescent, and thus traditional therapeutic strategies that target proliferating cells may be ineffective for this population. In this chapter, we explore the 4DN of two distinct cancer data sets. One data set is collected on colorectal cells, pre- and post- addition of an extra copy of chromosome 7. The other data set is collected on breast cancer cells that have been sorted to isolate the CSC subpopulation.

In the first study, we analyzed a cell line established from healthy colon mucosa with a normal karyotype (46,XY) and its isogenic derived cell line that acquired an extra copy of chromosome 7 as its sole anomaly (47,XY,+7). We studied structure/function relationships consequent to aneuploidization using Hi-C, RNA sequencing and protein profiling. The gain of chromosome 7 resulted in an increase of transcript levels of resident genes as well as genome-wide gene and protein expression changes. The Hi-C analysis showed that the extra copy of chromosome 7 is reflected in more interchromosomal contacts between the triploid chromosomes. Chromatin organization changes are observed genome-wide, as determined by changes in A/B compartmentalization and TAD boundaries. Most notably, chromosome 4 shows a profound loss of chromatin organization, and chromosome 14 contains a large A/B compartment switch region, concurrent with resident gene expression changes. No changes to the nuclear position of the additional chromosome 7 territory were observed when measuring distances of chromosome painting probes by interphase FISH. Genome and protein data showed enrichment in signaling pathways crucial for malignant transformation, such as the HGF/MET-axis. We conclude that a specific chromosomal aneuploidy has profound impact on nuclear structure and function,

both locally and genome-wide. Our study provides a benchmark for the analysis of cancer nucleomes with complex karyotypes.

In the second study, we present preliminary results characterizing the differences between CSCs and non-CSCs. The cells analyzed here are derived from a breast cancer cell line, SUM-159. These preliminary results show that the CSC population has a different chromatin architecture in specific genomic regions. Additionally, targeting CSCs via up-regulation of the stem cell marker ALDH1A1 may enrich for a CSC subpopulation with a distinct karyotype.

6.2 The 4DN of single chromosome aneuploidy

Chromosomal aneuploidy is a hallmark of many cancers [140, 181]. Specifically, colorectal cancers often show a gain in chromosome 7 (trisomy 7) early in cancer development [16, 64, 71, 143]. This genomic imbalance undoubtedly affects the 4DN of these cells, but it is difficult to characterize how this specific early-stage event leads to an aggressive cancer phenotype at later time points. Recently, methods to create a cell population with trisomy 7 as the only detectable abnormality were developed for the human colonic epithelial cell line (HCEC) [109, 145]. Long term propagation ($\sim$ 40 population doublings) of HCEC cells under serum-free culture conditions gave rise to cells with trisomy 7 as the only detectable genetic abnormality (referred to here as HCEC+7).

A comprehensive understanding of complex biological systems depends on recognizing its structure-function relationships from molecules to the entire system. Such efforts form the basis for the NIH Common Fund Initiative, the 4D Nucleome (https://commonfund.nih.gov/4Dnucleome), which is aimed to elucidate how nuclear organization affects the cellular transcriptome and the phenotype of cells [141]. We have previously used Hi-C and RNA-seq to study structure/function relationships in the Colorectal carcinomas (CRC) cell line HT-29, a cell line that harbors multiple

structural and numerical chromosomal aberrations which led to profound changes in genome structure and function. We could show that chromosome conformation capture identifies chromosomal aberrations at high resolution, and that these aberrations alter the relationship between structure and function [154]. We now use a model system that consists of a matched pair of isogenic HCEC and HCEC+7 cells to dissect the consequences of a single chromosomal aneuploidy on genome architecture and function. We analyzed the genome using Hi-C, RNA-seq and imaging. We chose trisomy of chromosome 7 as a model, because it occurs with high frequency in colorectal adenomas and must be considered as a "point of no return" in the progression towards malignancy.

6.2.1 Trisomy 7 results in specific alterations to nuclear organization as measured by Hi-C

Changes to nuclear organization as a consequence of aneuploidy were determined by Hi-C. We show that three copies of chromosome 7 resulted in increased inter- and intrachromosomal contacts of the aneuploid chromosomes which is shown genome-wide in Figure 6.1A. Figure 6.1B displays chromosome 7 specific Hi-C maps, and confirms that extra copies of chromosome indeed results in increased intrachromosomal contacts. This is consistent with previous results from our group analyzing the colorectal cancer cell line HT-29, where we demonstrated that copy number changes resulted in increased contacts [154].

The global Hi-C maps revealed additional, aneuploidy specific alterations to nuclear organization. Chromatin is organized as active and inactive domains, also referred to as A and B compartments. We observed a clear change in this compartmentalization on chromosome 14. Here, we determined a clear switch in compartmentalization from A in HCEC to B in HCEC+7 in a region spanning chr14:62.4Mb-63.8Mb (Figure 6.2A). This region contains few proteins coding genes, many of which change

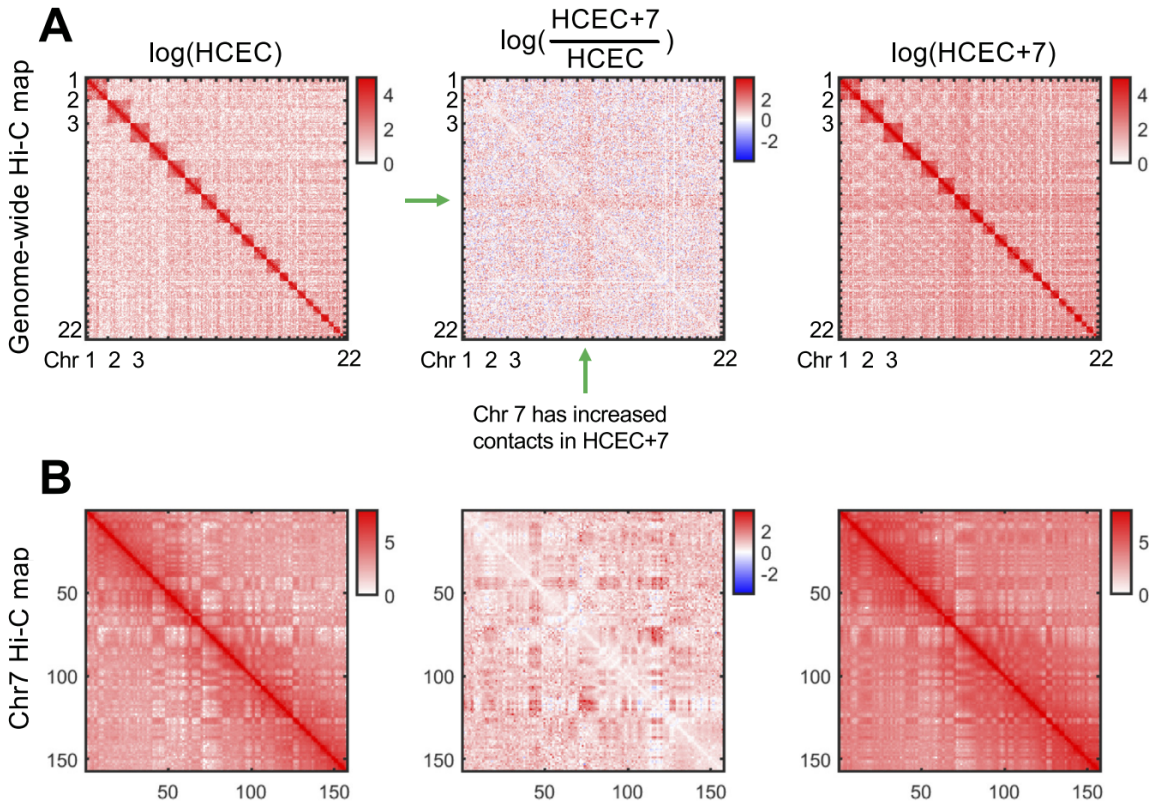


Figure 6.1: Hi-C maps show an increase in chromosome 7 contacts genome-wide. A) Genome-wide Hi-C contact maps for HCEC (left), HCEC+7 (right), and the difference between the two samples (center). Matrices are shown at 1Mb resolution and log-scale. Red regions in the middle matrix are enriched in HCEC+7, blue regions in the middle matrix are enriched in HCEC. A clear increase in HCEC+7 contacts is observed in regions involving chromosome 7 (green arrow). B) Chromosome 7 Hi-C contact maps for HCEC (left), HCEC+7 (right), and the difference between the 2 samples (center). Matrices are shown at 1Mb resolution and log-scale. Red regions in the middle matrix are enriched in HCEC+7, blue regions in the middle matrix are enriched in HCEC. A clear increase in HCEC+7 contacts is observed in chromosome 7. This figure was taken from Braun *et al.* [18].

significantly in expression. The most significant change in expression is observed in KCNH5 and RHOJ, both of which are down regulated and completely repressed in HCEC+7 (adjusted p-value 1.7824e-17 and 2.9776e-06, respectively), in line with their change from compartment A to B. KCNH5 hypermethylation has been observed in a number of cancers, which may explain the compartment switch from A to B [73].

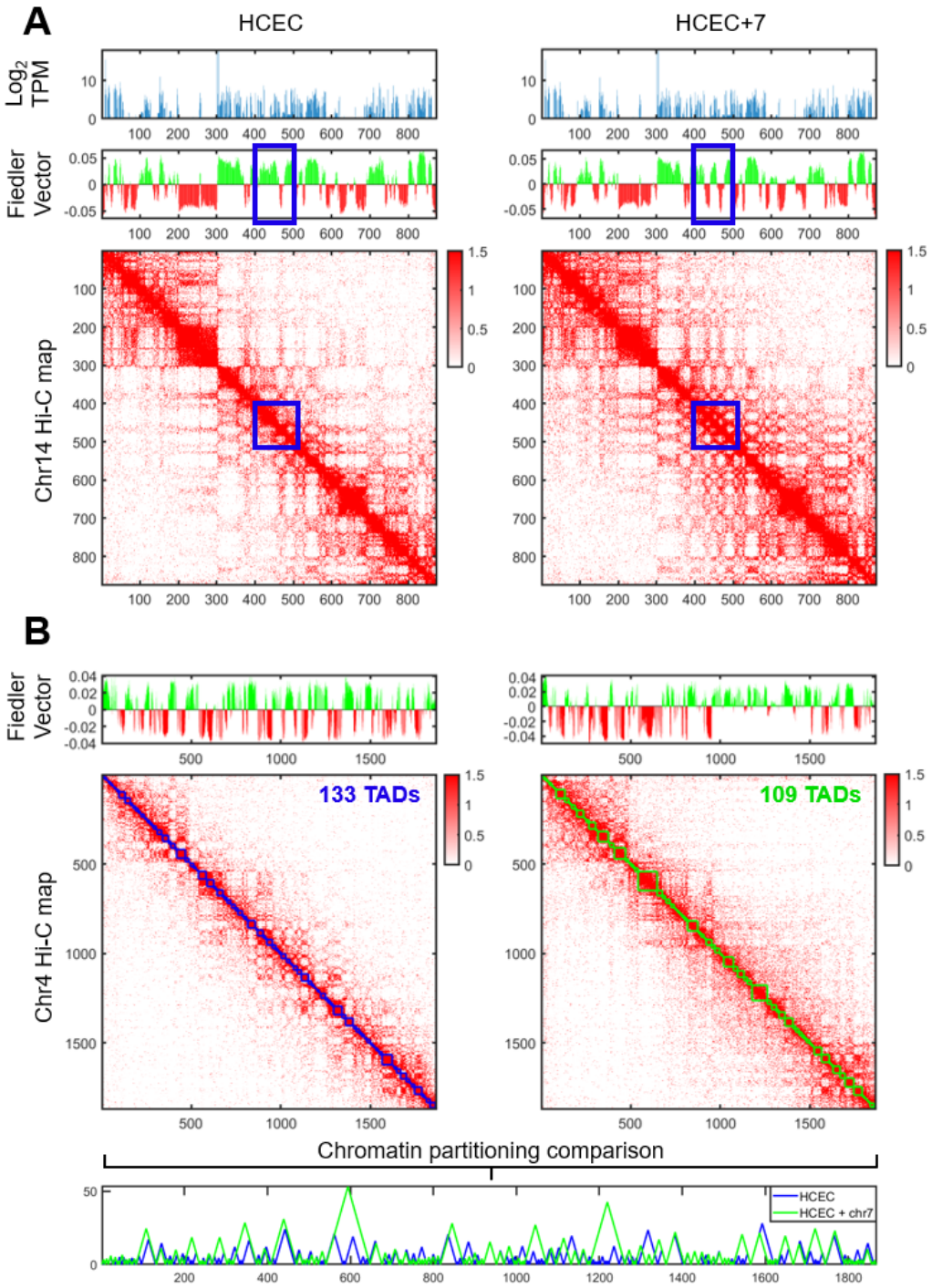


Figure 6.2: HCEC+7 results in genome-wide structural changes...
continues on next page

Figure 6.2: continued from previous page
....A) A change in chromatin partitioning is observed in chromosome 14. The left and right side show HCEC and HCEC+7, respectively. The top row shows the average gene expression for each 100kb bin. The middle row shows the Fiedler vector partitioning. A clear change is observed between 425 and 450 (blue square). The bottom row is the Hi-C contact map for chromosome 14, shown at 100kb resolution, log-scale. B) HCEC+7 shows a clear change in chromosome 4 patterning. The left and right side show HCEC and HCEC+7, respectively. The top row shows the Fiedler vector partitioning. The middle row is the Hi-C contact map for chromosome 4, shown at 100kb resolution, log-scale. Squares (green and blue) along the diagonal of the Hi-C matrix depict the TAD structure. A direct comparison of the TAD structure between samples is shown below. This figure was taken from Braun *et al.* [18].

Next, we observed a visible loss of chromosome 4 patterning, with clear changes in A/B compartmentalization as well as TAD structure chromosome-wide (Figure 6.2B). Genes within regions that change from A in HCEC to B in HCEC+7 are generally down-regulated, as expected. 8.8% of genes within these regions are significantly down-regulated, with 4.4% of genes significantly up-regulated (adjusted p-value $<$ 0.05). Genes within regions that change from B in HCEC to A in HCEC+7 are up-regulated to a larger extent. 22.3% of genes within these regions are significantly up-regulated, with 1% of genes significantly down-regulated (adjusted p-value < 0.05). HCEC revealed 133 TADs on chromosome 4, which were reduced to 109 TADs as a consequence of chromosome 7 aneuploidy. Change was also observed in TAD structure: the segregation of the domains was less pronounced in the HCEC+7 cells. This effect was not observed on other chromosomes.

6.2.2 Trisomy 7 affects chromosomal organization in the interphase nucleus as measured by 3D-FISH

We next determined whether an extra copy of chromosome 7 affects the positioning of that chromosome in the interphase nucleus using an independent method. Therefore, we performed dual-color 3D-FISH on morphologically preserved HCEC

and HCEC+7 nuclei. In addition to chromosome 7, we also determined the positioning of chromosome 19 as a control. Chromosome 19 has the highest gene density of all human chromosomes and is positioned closer towards the center of the interphase nucleus [42]. To objectively evaluate CT positioning and to enable statistical comparison between HCEC and HCEC+7 cells, 3D images were processed via a MATLAB script (See Appendix A.4). The shape of the nuclei was detected and analyzed individually, the number of CTs was counted, and the size and distance between CTs and the nuclear periphery was calculated. An example of the image reconstruction and the position measurements is shown in Figure 6.3. The size of the CTs of chromosome 7 were slightly smaller in the HCEC+7 cells (HCEC: 5.96 $\pm$ 1.62 μm^3; HCEC+7: 4.92 $\pm$ 1.54 μm^3) although this trend did not reach statistical significance. In turn, the size of the chromosome territories of chromosome 19 tended to be slightly bigger in the HCEC+7 cells compared to the diploid cells (HCEC: 6.26 $\pm$ 1.70 μm^3; HCEC+7: 7.54 $\pm$ 2.56 μm^3) without any statistical significance. As expected the nuclear position of chromosome 19 was more internal than chromosome 7, which is gene poorer, regardless of an extra copy of chromosome 7. However, the positioning of the CTs of chromosome 7 was more variable in the HCEC+7 cells as measured by the distance to the edge of the nucleus to the centroid of the CT (HCEC: 2.10 $\pm$ 0.96 μm; HCEC+7: 2.37 $\pm$ 1.34 μm). The nuclei of the HCEC+7 cells were significantly bigger compared to the nuclei of the diploid cells (1CT: 139.06 $\pm$ 35.24 μm^3; 1CT+7: 173.23 $\pm$ 48.82 μm^3, $p<0.00085$).

6.2.3 Trisomy 7 results in global gene expression changes

To determine the consequences of nuclear structural changes as a result of trisomy 7, we performed transcriptome profiling by RNA sequencing. The acquisition of an extra copy of chromosome 7 resulted in global gene expression changes. Chromosome 7 had the highest proportion of genes differentially expressed, as would be expected

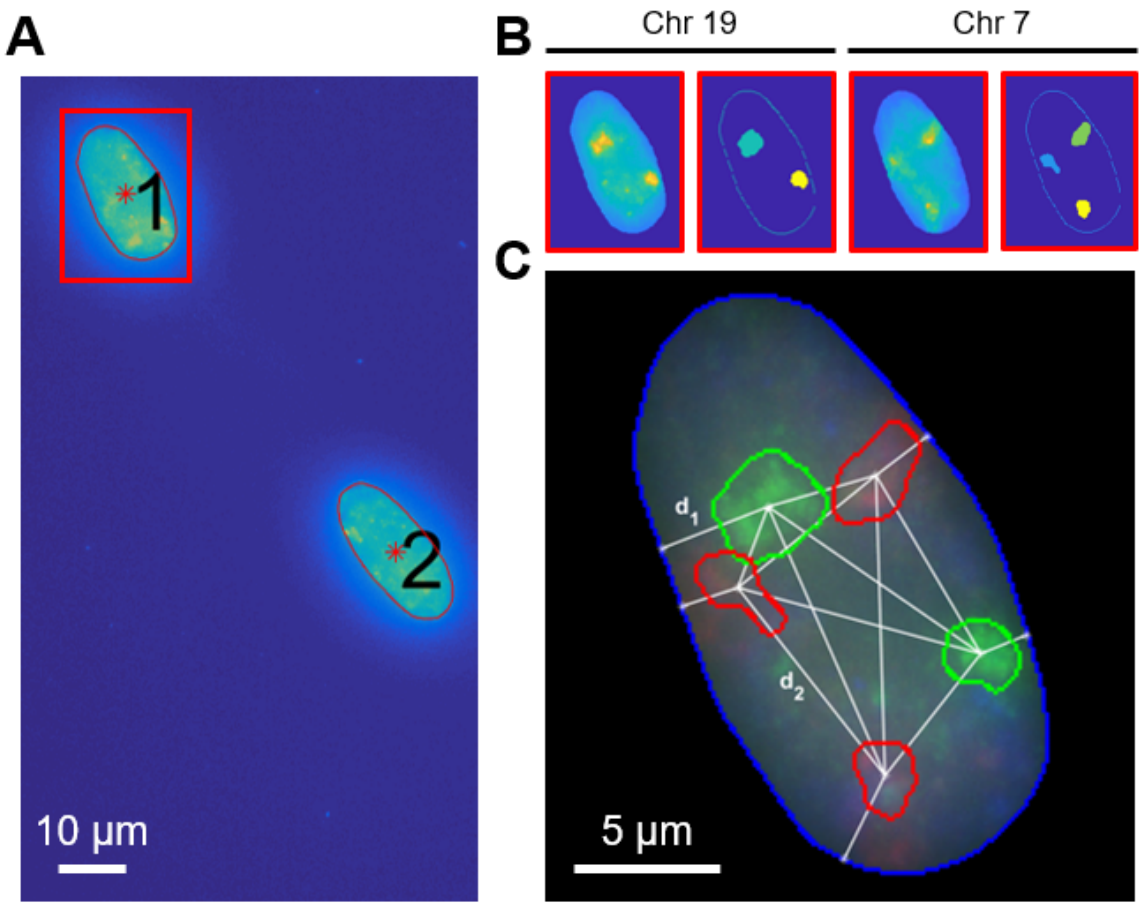


Figure 6.3: 3D-FISH image analysis of CTs. A) Cell nuclei were automatically selected from DAPI images. B) Fluorescent channels corresponding to chromosome 7 and 19 were analyzed, and CTs were automatically selected. C) CT size, distance from CT to nucleus edge (d_1), and distance between CTs (d_2) was extracted. This figure was taken from Braun *et al.* [18].

with additional gene copies, though not all genes were upregulated.

Next, we wanted to interrogate the functional space of gene expression changes. This was performed using gene set association analysis of RNA-seq data (GSAASeqSP) (Figure 6.4A). From this analysis, we identified 2 signaling networks that were significantly enriched in the HCEC+7 cells with an FDR q-value < 0.005, and others that are strongly associated with the phenotype. The top 4 gene sets are crucial networks in colorectal carcinogenesis, namely, P53-pathway (0.003), MYC-targets (0.003), TNF-α via NF-κB (0.126), and KRAS signaling up (0.160).

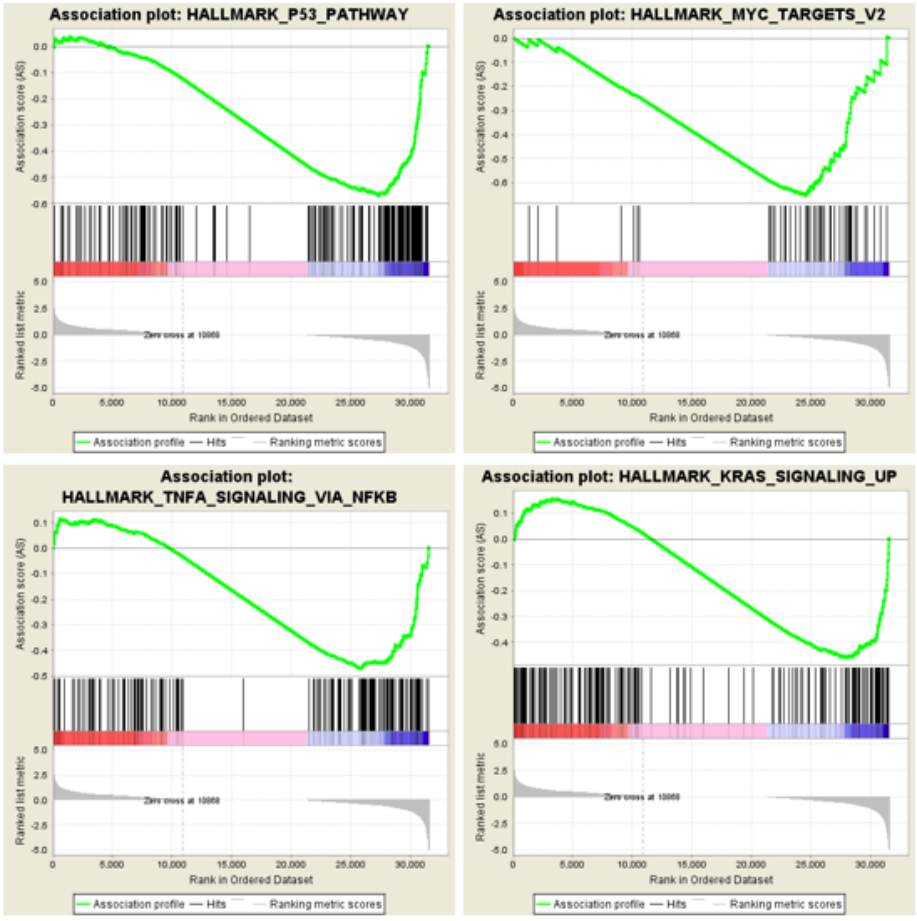


Figure 6.4: Identification of differentially regulated pathways. GSAA analysis. The top four signaling networks that were enriched in the HCEC+7 cells are shown; P53-pathway (0.003), MYC-targets (0.003), TNF-α via NF-κB (0.126), KRAS signaling up (0.160). This figure was taken from Braun *et al.* [18].

More detailed analysis of genes coded on chromosome 7, which are linked to

colorectal carcinogenesis, revealed that HGF was significantly down-regulated in HCEC+7 cells, whereas MET was significantly up-regulated. In line, GSAA analysis showed an up-regulation of KRAS signaling, which is downstream of the HGF/MET axis. Though neither HGF nor MET showed a change in A/B compartment, MET is located within the region that shows a clear loss of chromatin patterning.

6.2.4 Discussion

Tissue-specific chromosomal aneuploidies emerge at early stages of tumorigenesis [16, 74, 143]. During the transformation of normal colon to polyps and eventually invasive carcinoma, cells frequently acquire extra copies of chromosome 7 at early stages of malignant transformation [16, 64, 71, 143]. We therefore attempted to understand the global consequences of trisomy 7 in colon cells on both nuclear organization and function.

Hi-C identified changes specific to chromosome 7, however, additional chromosomes were affected by distinct chromatin organization modifications. The increase in direct contacts are a result of an additional copy of chromosome 7. An increase in chromosome 7 Hi-C counts was observed genome-wide, inferring that the extra copy of chromosome 7 had no detectable preference in its position relative to other chromosomes. This is in line with our CT reconstruction analysis using 3D-FISH which did not show changes in chromosomal positioning. These results are consistent with previous results from our laboratories reporting comprehensive Hi-C maps of the colorectal cancer cell line HT-29 [154]. In addition, we showed that the Hi-C maps faithfully recapitulate 2D changes determined by high resolution molecular cytogenetic analyses, i.e., SKY, aCGH and FISH to chromosome number and structure.

Changes of the nucleome of the HCEC+7 cells were not restricted to chromosome 7. Genome-wide analysis revealed many regions with differences in structure and function, with the strongest changes observed on chromosome 4 and 14. On both

of these chromosomes, A/B compartment changes were observed. On chromosome 4 we not only observed switches in the A/B compartments, but also changes in TAD boundaries. The number of TADs decreased from 133 in the diploid cells to 109 in the HCEC+7 cells. In addition, we observed structural changes. The mechanisms for these changes are presently not known. However, from our results it appears obvious that understanding the consequences of chromosomal trisomies requires genome wide analyses.

Changes in the nuclear structure resulted in a change in function defined by gene expression levels. Genes that switched from A to B regions were generally down-regulated, whereas genes that switch from B to A regions were generally upregulated. One example is the switch of a region on chromosome 14 (chr14:62.4Mb-63.8Mb) that changed from the A to B which is possibly reflective of hypermethylation. Hypermethylation is associated with heterochromatin and B compartmentalization. Furthermore, KCNH5 has been shown to hypermethylated previously in cancers [73, 127]. How this region is targeted after the trisomy 7 event is unknown. The loss of compartment structure observed on chromosome 4 extended from 91 Mb to 158.9 Mb and was generally associated with A to B transition. Again, the compartment switch resulted in a change in gene expression in a directionality that would be intuitive. The mechanism for this targeted loss of chromatin structure is unknown, however, we surmise that global gene expression changes have a structural correlate.

We found a significant down-regulation of HGF-mRNA expression in HCEC+7 cells, while MET was significantly up-regulated. HGF and MET are both located in a region on chromosome 7 that showed a loss of chromatin patterning, however, we did not observe a switch in A/B compartment. The hepatocyte growth factor (HGF) specifically binds to the receptor tyrosine kinase "mesenchymal to epithelial transition" (MET), which is a proto-oncogene [17, 129]. MET-mRNA quantification in primary CRC suggested that MET overexpression plays an important role in the

development of loco-regional invasiveness in early stage of CRC development [167]. Consistent to changes in HGF and MET expression, we found an up-regulation of KRAS signaling by GSAA gene enrichment analysis, although it did not reach statistical significance. Besides KRAS signaling, trisomy 7 in normal colon cells results in dysregulation of further signaling pathways crucial for CRC genesis such as the P53-pathway and TNF-α via NF-κB.

Our data demonstrate that structural nuclear changes caused by an extra chromosome 7 entails the dysregulation of several pathways associated with colorectal carcinogenesis. Interestingly, trisomy 7 by itself is not carcinogenic as demonstrated here by us using murine xenograft models in nu/nu mice and others [109]. As trisomy 7 is commonly found in early stages of carcinogenesis in the colon such as adenomas, we conclude from our results that this chromosomal aberration is a perquisite facilitating subsequent malignant transformation. We show that this specific chromosomal aneuploidy has a profound impact on nuclear structure and function, both, locally and genome wide.

6.3 The 4DN of cancer stem cells

The cancer stem cell hypothesis has garnered significant research attention in recent years. This hypothesis states that there exists a population of CSCs in many tumors that have been reprogrammed to a pluripotent state, and that these cells have the ability to recapitulate the tumor microenvironment and drive metastasis [185]. These cells can be identified by their upregulation of pluripotency genes. Specifically, ALDH1A1 is a marker of CSCs in many cancer cell types, and flow sorting on this marker can isolate the CSC population [76]. Furthermore, knock-down of SPEN significantly decreases the expression of *ALDH1A1* in breast cancer cells [192]. These findings have advanced our understanding of how CSCs may operate, but a complete characterization of genome structure and function in these cell populations has not

been performed.

In this work, we characterize the structural and function differences between four distinct cell populations. These populations represent varying levels of potential cell stemness: SUM-159 cells (SUM-159), ALDH1A1 positive SUM-159 cells (ALDH+), ALDH1A1 negative SUM-159 cells (ALDH-), and SUM-159 cells with SPEN knock-down (SPEN-KD). For all cell populations, Hi-C and RNA-seq data is collected and analyzed. Differences in each of these measures are detected and described in the following sections.

6.3.1 ALDH1A1 dependent chromosome translocations

Visual inspection of the ALDH+ and ALDH- Hi-C matrices led us to believe there may be differences in the karyotype between these populations. To identify translocations in a principled manner, we developed a method for the detection of translocation differences between samples (See Section A.4). Briefly, our method detects large genomic regions in the inter-chromosomal Hi-C matrices where the number of contacts in the region far exceeds the expectation.

Our analysis detected multiple translocation differences between the two populations. Overall, there were more translocations in the ALDH+ population than the ALDH- population (Figure 6.5A). An unbalanced translocation between the q arm of chromosome 3 and the q arm of chromosome 7 is clearly detected in the ALDH+ sample, but not the ALDH- sample (Figure 6.5B). Another unbalanced translocation is detectable between the q arm of chromosome 10 and the q arm of chromosome 13 in the ALDH+ sample, but not the ALDH- sample (Figure 6.5C). Additionally, a translocation between the chromosome 2 and chromosome 11 is detectable in both populations, but appears to be much stronger in the ALDH- samples. Since Hi-C data is derived from a population of cells, this may indicate that a higher percentage of cells in the ALDH- population have the translocation between chromosome 2 and

chromosome 11.

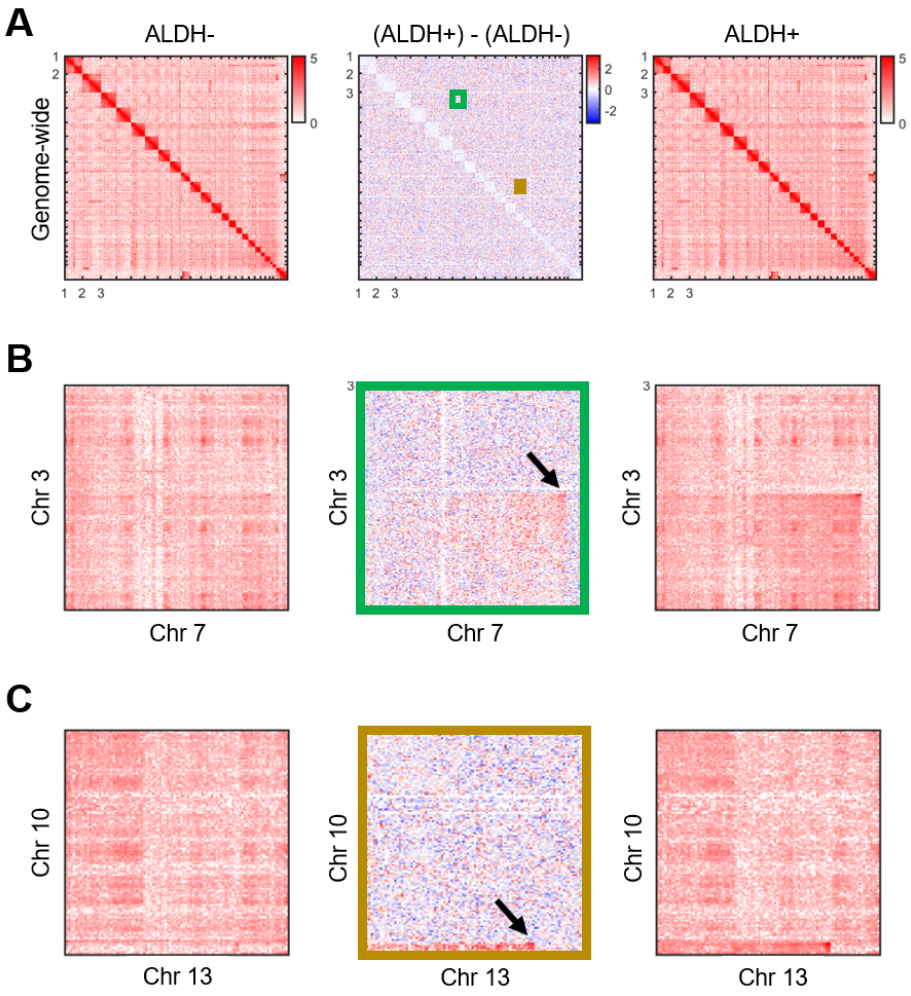


Figure 6.5: CSC vs non-CSC translocation differences. A) Genome-wide Hi-C for ALDH- (left), the difference between ALDH+ and ALDH- matrices (middle), and ALDH+ (right). The green and brown boxes denote regions where translocation differences were detected. B) Inter-chromosome 3 and 7 Hi-C for ALDH- (left), the difference between ALDH+ and ALDH- matrices (middle), and ALDH+ (right). The black arrow denotes where the translocation difference can be observed. C) Inter-chromosome 10 and 13 Hi-C for ALDH- (left), the difference between ALDH+ and ALDH- matrices (middle), and ALDH+ (right). The black arrow denotes where the translocation difference can be observed.

To validate these findings, we performed spectral karyotyping (SKY) on the ALDH+ and ALDH- populations. In total, 15 cells from the ALDH- population and 19 cells from the ALDH+ population were karyotyped via SKY (Figure 6.6). Surprisingly, SKY revealed extreme heterogeneity in each population, with nearly

every cell in each population carrying a distinct karyotype. This finding leads us to believe that the translocations detected via Hi-C exist in a small population of cells, and a larger number of cells will need to be karyotyped to see these differences.

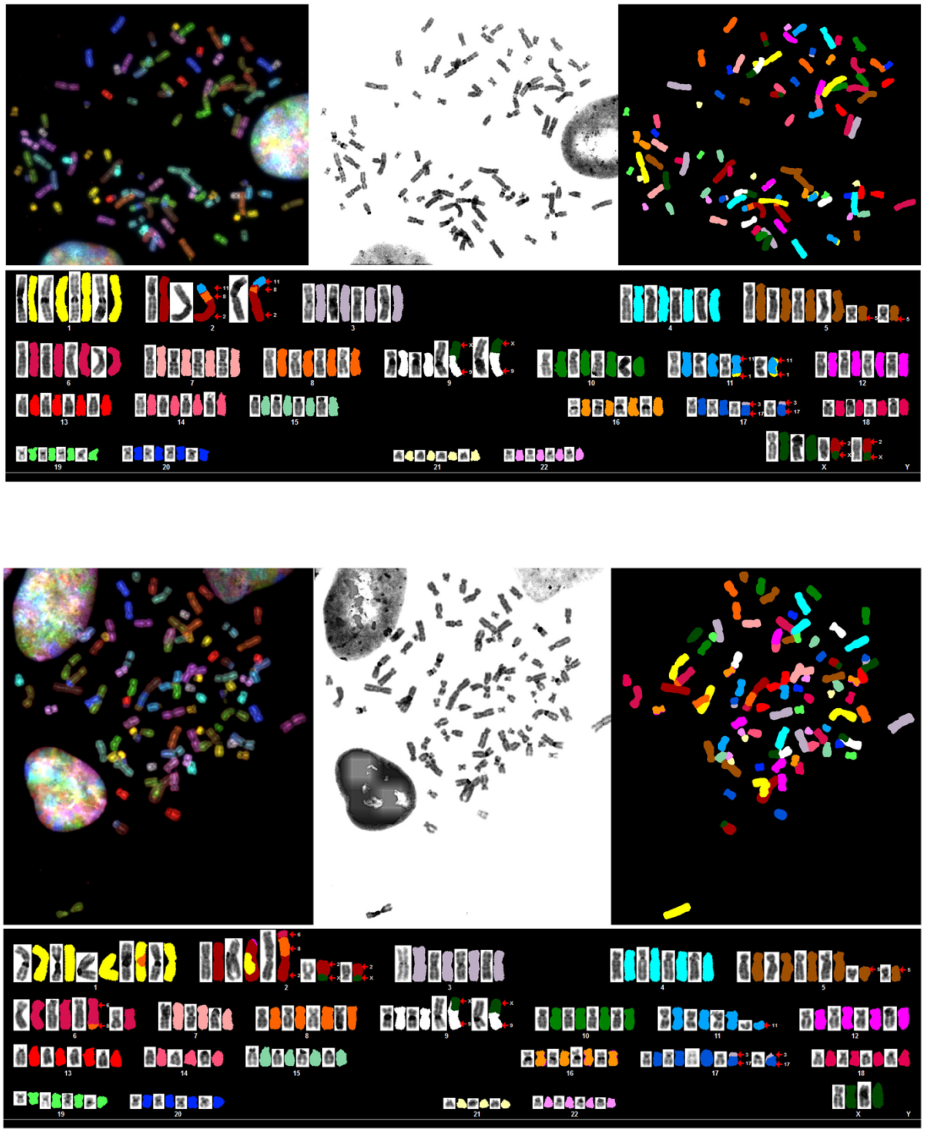


Figure 6.6: Spectral karyotyping of CSC and non-CSC populations. The top image is SKY derived from the ALDH- population. The bottom image is SKY derived from the ALDH+ population. Clear translocation differences can be observed between the two samples. Images obtained from Darawalee Wangsa, NIH.

6.3.2 *ALDH1A1* chromatin accessibility

In order to determine the chromatin accessibility of the *ALDH1A1* locus, the Fiedler vector was calculated for a region surrounding this locus (See Section A.4). From this analysis we can clearly see that the locus is in the A compartment in the ALDH+ sample, and at least partially in the B compartment for all other samples (Figure 6.7).

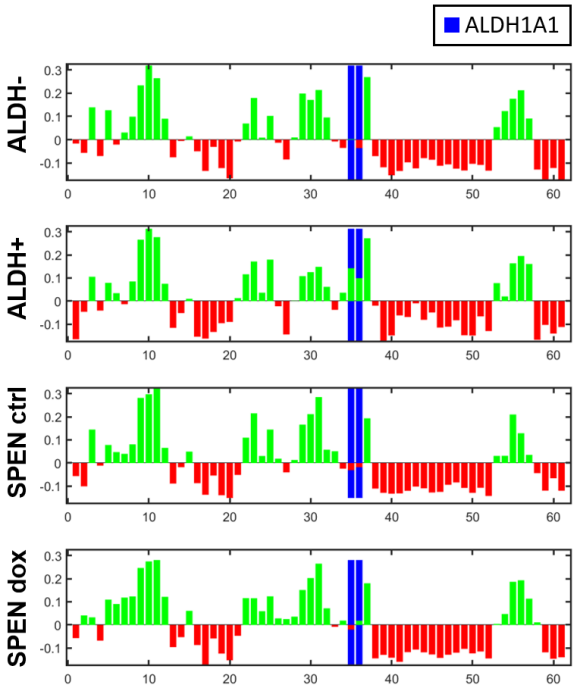


Figure 6.7: *ALDH1A1* chromatin accessibility. The Fielder vector of a sub-regions surrounding the *ALDH1A1* locus is plotted for each sample. Green bars denoted positive Fiedler vector elements (A compartment) and red bars denoted negative Fiedler vector elements (B compartment). Blue background denotes genomic regions that contain the *ALDH1A1* locus. The *ALDH1A1* locus is clearly in the A compartment in the ALDH+ sample, and at least partially in the B compartment for all other samples. Details described in Section A.4.

6.3.3 Low dimensional projection of chromatin structure and function

To determine genome-wide differences in structure and function, we projected measurements of each to a low dimensional space (See Section A.4). Briefly, PCA was applied to normalized RNA-seq and chromatin accessibility, via the Fiedler vector.

From this analysis, we observe that PC1 clearly separates the ALDH1A1 sorted samples (ALDH+ and ALDH-) from the non-sorted samples (SUM-159 and SPEN-KD) in both the structural and function PCA plots (See Figure 6.8). Interestingly, while the ALDH1A1 sorted samples appear to be very similar in function genome-wide, they are clearly separated in the structural PCA plot. This suggests that identification of CSCs via structural measurements may be more effective than identification via functional measurements.

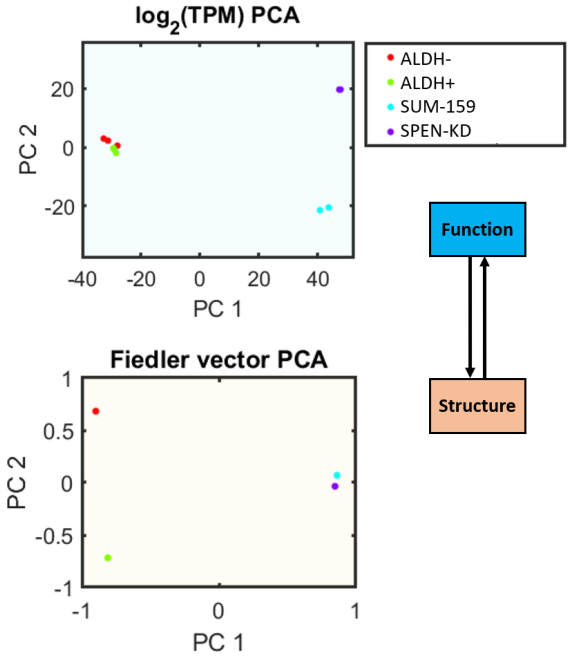


Figure 6.8: Low dimensional projection of chromatin structure and function. Functional (Top) and structural (Bottom) genome-wide measurements derived from RNA-seq and Hi-C, respectively, are projected to a low dimensional space via PCA. Details described in Section A.4.

CHAPTER VII

4DNvestigator

This chapter is based on a paper by Scott Ronquist, Sijia Liu, Michael Perlman, and Indika Rajapakse (under review). I have summarized my contributions to this paper here.

7.1 Abstract

The combined analysis of genome structure and function over time, and how these changes affect cellular phenotype, is referred to as the 4DN. 4DN analysis is necessary to fully understand how a cell operates, but 4DN analysis tools are currently underdeveloped. We present the "4DNvestigator," a user-friendly toolbox for the analysis of time-series genome structure, measured by Hi-C, and genome function, measured by RNA-seq.

Availability: https://github.com/scotronq/4DNvestigator

7.2 Introduction

The genome is a dynamical system where changes in genome structure and function over time affect the cell's phenotype. This dynamical relationship between genome structure, transcriptional landscapes, and cellular phenotypes is referred to

as the 4DN [141]. The 4DN is an emerging field of study, wherein a significant investment is being made by research institutes to enhance our understanding of how changes in nuclear organization affects normal development and disease states [47].

To analyze the 4DN, Hi-C and RNA-seq are often used to observe genome structure and function, respectively. As high throughput sequencing costs decline, the availability and volume of Hi-C and RNA-seq data sets is expected to increase. With this increase in data, the development of methods to properly analyze these data sets is imperative. Furthermore, analyzing these high-dimensional data sets is non-trivial and requires a high-level understanding of biology, mathematics, and computer science.

To aid researchers in the analysis of 4DN data, we present the 4DNvestigator. The 4DNvestigator is a MATLAB toolbox that loads time series Hi-C and RNA-seq data, extracts important structural and functional features, and is compatible with established Hi-C and RNA-seq processing programs and formats. This toolbox includes both established and novel 4DN data analysis methods, and displays important findings in simple visualizations. The 4DNvestigator takes a network-based approach to this analysis. This approach allows for the computation of network centrality and network entropy, both of which have been shown to reflect biological phenomena [103, 110]. Additionally, we present a novel statistical method for comparing two or more Hi-C matrices. We compare this method against established Hi-C comparison methods within this text. We believe adoption by the community at large of analytical tools such as the one described in our paper is critical to advance the field of 4DN research.

7.3 Materials and methods

An overview of the 4DNvestigator workflow is depicted in Figure 7.1A. The 4DNvestigator takes processed Hi-C and RNA-seq data as input, along with a meta-

data file which describes the sample and time point for each input Hi-C/RNA-seq file (See A.5). A graphical user interface (GUI) is provided for ease of use (Figure 7.1B). A number of novel methods for analyzing 4DN data are included within the 4DNvestigator and are described below. A list of accepted input file formats is provided at https://github.com/scotronq/4DNvestigator.

7.3.1 4DN feature analyzer

The "4DN feature analyzer" quantifies and visualizes how much a genomic region changes in structure and function over time. To achieve this, we adopt a network point of view, where genomic regions are nodes and interactions between regions are edges. Edge weights are set relative to the number of contacts between regions in Hi-C. With this network point of view, we can quantify the importance of each node using network centrality. Network centrality identifies nodes that play an influential topological role in the network [125]. There are several different centrality measures, each of which identifies a particular type of nodal influence. For example, *degree centrality* measures the connectedness of a node locally, calculated by the summation of edge weights for all edges connected to the node. In contrast, *betweenness centrality* is a more global connectedness measure, which counts the number of times a given node is on the shortest path between all other pairs of nodes. Centrality measures derived from Hi-C have been shown to have biological meaning, reflecting the chromatin accessibility of each region and predicting A/B compartment switch locations [103]. In the 4DN feature analyzer, centrality measures and RNA-seq define the state of each genomic region at each time point. The 4DN feature analyzer quantifies and visualizes how each genomic region changes in this state over time. Genomic regions can be defined at many scales, such as gene-level, 100 kb-level, or TAD-level.

Hi-C matrices are balanced using the Knight-Ruiz (KR) algorithm and are observed over expected (O/E) normalized using Juicer Tools [56]. We define Hi-C ma-

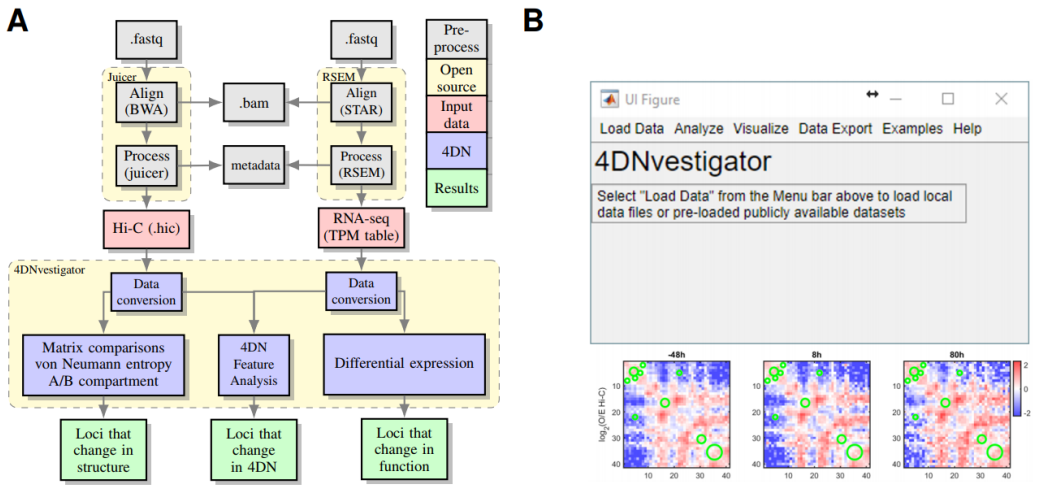


Figure 7.1: Overview of the 4DNvestigator. A) Data processing pipeline for the 4DNvestigator. Within this diagram, 4DN refers to the 4DNvestigator. B) Overview of the 4DNvestigator GUI, with a comparison between time series Hi-C samples at a 4Mb region surrounding the gene locus *MYH1*. Data is obtained from Liu *et al.*, time points -48, 8, 80h [103]. The procedure described by Larntz and Perlman (See section "Hi-C matrix comparison") detects regions that are most significantly different between samples, and rejects the null hypothesis that these matrices are equal. Green circles highlight regions with the largest difference between samples. Example codes to recalculate these results are included within the 4DNvestigator.

trices of this form as $\mathbf{A}^{(m)}$, where m denotes the time point (or sample). The following network centrality measures are extracted from $\mathbf{A}^{(m)}$: degree, eigenvector, betweenness, and closeness [103]. The centrality measures are combined with the RNA-seq expression for the corresponding genomic regions to form a "feature" matrix that defines the state of each genomic region at each time point. RNA-seq data is measured by the $\log_2$ transformation of Transcripts Per Million (TPM). For regions containing more than one gene, the mean $\log_2$(TPM) is computed. We define RNA-seq data of this form as $\mathbf{r}^{(m)}$. The z-score for each feature is computed to normalize the data and to put features on the same relative scale. Feature matrices for each time point are then projected to a common low dimensional space (2D or 3D) via PCA, Laplacian Eigenmaps, or t-SNE [10, 175]. This allows the genomic regions to be visualized in 2D. Genomic regions that move significantly within this low dimensional space also change the most in structure and function over time. Points that correspond to the

same genomic region at different time points are fit with a Minimum Volume Ellipsoid (MVE). The MVE volume quantifies the "4DN variance" of each genomic region. A Minimum Area Ellipse (MAE) is calculated if the points lie on a plane. Methods to perform this analysis are described in Algorithm 1, following methods outlined in Liu *et al.* [103].

Algorithm 1: 4DN feature analyzer

Input: Hi-C matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{n \times n}$, and RNA-seq vectors $\mathbf{r}^{(m)} \in \mathbb{R}^{n \times 1}$, $m = 1, \ldots, k$

Output: low dimensional space $\mathbf{Y}^{(m)}$, and 4DN variance $\mathbf{v}$

1 Compute degree, eigenvector, betweenness, and closeness centrality of $\mathbf{A}^{(m)}$, and define as $\mathbf{b}_{deg}^{(m)}$, $\mathbf{b}_{eig}^{(m)}$, $\mathbf{b}_{bet}^{(m)}$, $\mathbf{b}_{close}^{(m)}$, respectively, where each $\mathbf{b}^{(m)} \in \mathbb{R}^{n \times 1}$

2 Form the feature matrices $\mathbf{X}^{(m)} = [\mathbf{b}_{deg}^{(m)}, \mathbf{b}_{eig}^{(m)}, \mathbf{b}_{bet}^{(m)}, \mathbf{b}_{close}^{(m)}, \mathbf{r}^{(m)}]$, where $\mathbf{X}^{(m)} \in \mathbb{R}^{n \times 5}$

3 Normalize the columns of $\mathbf{X}^{(m)}$

4 Compute the common low dimensional space $\mathbf{Y}^{(m)}$ using PCA, Laplacian Eigenmaps, or t-SNE

5 Compute the 4DN variance for each genomic region $\mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^{n \times 1}$

Return: $\mathbf{Y}^{(m)}$ and $\mathbf{v}$

7.3.2 Network entropy

Entropy measures the order within a system, where higher entropy corresponds to more disorder [38]. Here, we apply this measure to Hi-C data to quantify the order in chromatin structure. Biologically, genomic regions with high entropy likely correlate with high proportions of euchromatin, as euchromatin is more structurally permissive than heterochromatin [110]. Furthermore, entropy can be used to quantify stemness, since cells with high pluripotency are less defined in their chromatin structure [113]. Since Hi-C is a multivariate analysis measurement (each contact coincidence involves two variables, the two loci), we use multivariate entropy, VNE. The algorithm to compute VNE is given in Algorithm 2.

Algorithm 2: VNE Computation

Input: Hi-C matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$

Output: VNE

1 Compute the correlation matrix $\mathbf{C} = \text{corr}(\log_2(\mathbf{A}))$

2 Compute the eigendecomposition of $\mathbf{C}$, where $\lambda_1 \leq \lambda_i \leq \lambda_n$ are the eigenvalues of $\mathbf{C}$

3 Normalize the eigenvalues: $\bar{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$

4 Compute VNE: $\text{VNE} = -\sum_{i=1}^{n} \bar{\lambda}_i \ln(\bar{\lambda}_i)$

Return: VNE

7.3.3 Hi-C matrix comparison

A number of methods have been developed in the Hi-C research community to compare Hi-C matrices. These methods can be broadly grouped into 2 categories: Hi-C reproducibility methods (e.g. HiCRep, HiC-spector), and Differential Chromatin Interactions (DCI) methods (e.g. SELFISH, HiCCompare, FIND, diffHic) [4, 54, 107, 162, 188, 189]. Hi-C reproducibility methods are useful for assessing the equality of Hi-C matrices genome-wide, and detecting technical bias between samples. DCI methods determine which loci have a significantly different number of Hi-C contacts between samples. All methods listed above are comparisons between only two matrices.

Here we pose a different, but related, question: Is the chromatin structure, within a genomic region, equivalent between samples? To answer this question, we take a multivariate statistical approach. Elements within Hi-C matrices have been shown to be normally distributed when the $\log_2$ transformation is applied to $\mathbf{A}^{(m)}$ (See Figure 7.2B) [25]. For data of this form, a method for testing the equality of correlation matrices was proposed by Larntz and Perlman in 1988 [95]. Here, we apply this technique to Hi-C correlation matrices to determine the statistical difference between multiple Hi-C samples at genomic regions of interest. Furthermore, we extend this method to determine where the matrices are most significantly different between samples. We refer to this method herein as the "LP" method. We recommend

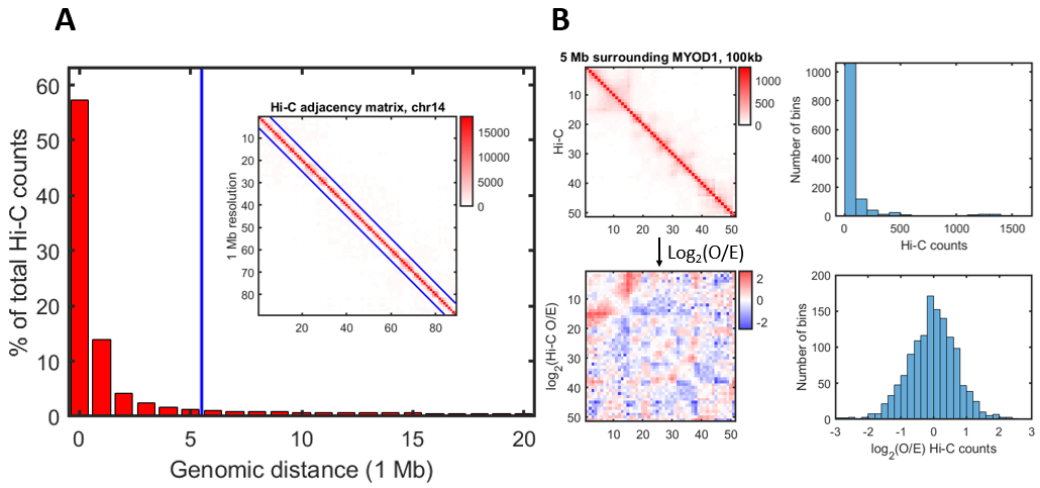


Figure 7.2: Hi-C normalization for the LP method. A) Percentage of total Hi-C contacts by genomic distance. Inset figure shows a typical Hi-C intra-chromosome adjacency matrix at 1 Mb resolution. The blue line denotes where the genomic distance exceeds 5 Mb. B) A 5Mb Hi-C matrix (100 kb resolution, top left) and a histogram of the counts within the matrix (top right). When the $\log_2$(O/E) transformation is computed on the Hi-C matrix (bottom left), elements within the matrix are normally distributed (bottom right).

that the LP method is performed at Hi-C resolutions $\leq$ 100 kb and that regions do not extend >5 Mb, as signal (counts) often become sparse as the genomic distance between loci increases (See Figure 7.2A). The full algorithm is given in Algorithm 3. Algorithm 3 outputs a P-value for the equality of the Hi-C matrices and a matrix $\mathbf{S}$ where the largest values in $\mathbf{S}$ correspond to the genomic regions that are most different between samples. An example of this analysis is shown in Figure 7.1B, which can be recreated using codes given at https://github.com/scotronq/4DNvestigator.

7.3.4 Chromatin partitioning and differential expression

The 4DNvestigator includes a suite of previously developed Hi-C and RNA-seq analysis methods. Hi-C A/B compartments can be extracted using previously defined methods [31, 102]. Regions that change compartments between samples are automatically identified. The 4DNvestigator also utilizes developed MATLAB scripts for

Algorithm 3: LP method

Input: Hi-C matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{n \times n}$, $m = 1, \ldots, k$

Output: P-value p, and test statistic $\mathbf{S}$

1 Compute the correlation matrix $\mathbf{C}^{(m)} = \text{corr}(\log_2(\mathbf{A}^{(m)}))$. Define the corresponding population correlation matrices as $\mathbf{P}^{(m)}$

2 Define the null hypothesis

$$H_0 : \mathbf{P}^{(1)} = \cdots = \mathbf{P}^{(k)}$$

3 Compute the Fisher z-transformation $\mathbf{Z}^{(m)}$. Elements in $\mathbf{Z}^{(m)}$ and $\mathbf{C}^{(m)}$ are denoted as $z_{ij}^{(m)}$ and $c_{ij}^{(m)}$, respectively, and

$$z_{ij}^{(m)} = \frac{1}{2} \ln \left[\frac{1 + c_{ij}^{(m)}}{1 - c_{ij}^{(m)}} \right]$$

4 Form $\mathbf{S}$, where elements in $\mathbf{S}$ are denoted as s_{ij} and

$$s_{ij} = (n-3) \sum_{m=1}^{k} (z_{ij}^{(m)} - \bar{z}_{ij})^2, \ \bar{z}_{ij} = k^{-1} \sum_{m=1}^{k} z_{ij}^{(m)}$$

5 Calculate the test statistic $T = \max_{1 \le i < j \le n} s_{ij}$

6 Reject H_0 at level α if $T > \chi^2_{k-1,\epsilon(\alpha)}$, where $\chi^2_{k-1,\epsilon(\alpha)}$ is the chi-squared distribution with $k-1$ degrees of freedom and $\epsilon(\alpha) = (1-\alpha)^{2/n(n-1)}$ is the Šidák correction

7 Calculate p, the level α at which $T > \chi^2_{k-1,\epsilon(\alpha)}$

Return: p and $\mathbf{S}$

differential gene expression that follow methods outlined in Anders *et al.* [2]. The 4DNvestigator takes results obtained from the methods described above and identifies regions that change in both Hi-C and RNA-seq.

The first method for determining A/B compartments is based on the Fielder vector of the Hi-C matrix, following methods outlined in Chen *et al.* [32]. The second method for determining A/B compartments is based on the first principal component of the Hi-C matrix, following methods outlined in Lieberman-Aiden *et al.* [102]. The A/B compartment partitioning for every chromosome and sample is plotted in a figure, where regions that have a large change in sign and magnitude are highlighted.

7.3.5 Simulated Hi-C data

Simulated Hi-C data was created to compare the LP method against alternative Hi-C comparison methods. Two distinct simulated data sets were created to have: (1) changes in chromatin loop structure and (2) changes in chromatin compartment structure. Chromatin loops and chromatin compartments are two features that have been used to characterize Hi-C structure [102, 139]. Both simulated data sets are created by perturbing data from real Hi-C matrices: a 400 kb region (10 kb resolution) is used for the chromatin loop data set and a 2Mb region (50 kb resolution) is used for the chromatin compartment data set.

One additional simulated matrix was also created by adding a small amount of random noise to the 400 kb region (10 kb resolution) Hi-C matrix. A random number sampled from a normal distribution, $\mathcal{N}(0,\, 0.05)$, is added to each element in $\log_2(\mathbf{A})$. This matrix is used to determine how robust Hi-C matrix comparison methods are to small amounts of noise.

For each simulated data set, 10 matrices were created that are incrementally more divergent from the original Hi-C matrix. For changes in loop structure, counts were added to a specific off-diagonal region, following a 2D gaussian distribution ($\sigma = 1$),

to model a chromatin loop structure. For changes in compartment structure, counts aligned to a specific genomic region (bin) were changed to decrease the correlation coefficient between the specified bin in the simulated Hi-C $\log_2(\mathbf{A})$ matrix and the specified bin in the original Hi-C $\log_2(\mathbf{A})$ matrix. These changes reflect what would be observed if the specified bin was changing its compartment structure. Methods to recreate the simulated Hi-C matrices are provided within the 4DNvestigator.

7.4 Results

7.4.1 4DN feature analyzer

We demonstrate how the 4DN feature analyzer makes use of time series information by analyzing time series Hi-C and RNA-seq data collected on human dermal fibroblasts that are being reprogrammed to the muscle lineage, mediated by the addition of master TF MYOD1 [103]. Three time points are analyzed here, corresponding to 100 kb resolution Hi-C and RNA-seq data on samples collected: 48 h prior to MYOD1 addition ("-48 h"; control), 8 h post-MYOD1 addition, and 80 h post-MYOD1 addition. Figure 7.3 shows the results from the 4DN feature analyzer for chromosome 11. Each time point is labeled with a unique color to show how the regions change in 4DN over time. Regions that move the most in this low dimensional space are highlighted with red ellipses, and the genes within these regions are shown within the figure. Data and code to perform this analysis are provided within the toolbox "examples4DNvestigator.m".

7.4.2 Network entropy

To demonstrate how VNE can be used to quantify disorder in chromatin structure, we have calculated the VNE of two distinct cell types: a differentiated cell type, HFFc6, and an undifferentiated cell type, H1-hESC. It is clear from the Hi-C matrices

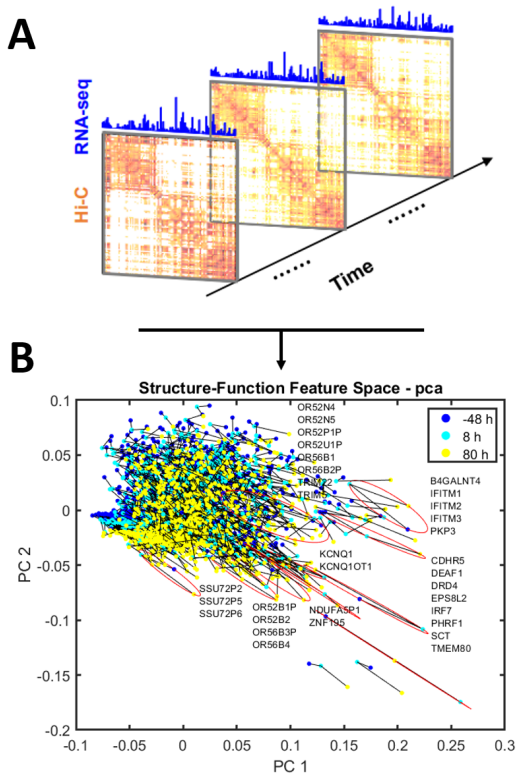


Figure 7.3: 4DN feature analyzer. (A) Time series Hi-C and RNA-seq data. (B) 4DN feature analyzer example. Points correspond to genomic loci (100 kb). Loci that change the most in structure and function are highlighted with red minimum volume ellipses, and genes within these regions are listed within the figure.

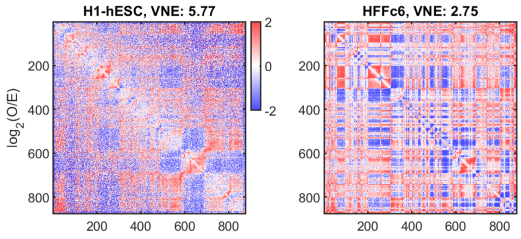


Figure 7.4: VNE difference between cell types. The VNE for an undifferentiated cell type (embryonic stem cell, H1-hESC) is much higher than the more differentiated cell type (fibroblast, HFFc6). Intra-chromosome $\log_2(\mathbf{A})$ Hi-C matrices are shown at 100 kb resolution for chromosome 14.

that the differentiated cell type is more ordered in its chromatin structure (See Figure 7.4). VNE reflects this observation, as the VNE for the differentiated cell type HFFc6 is much lower (2.87) than the undifferentiated cell type H1-hESC (5.74).

7.4.3 Hi-C matrix comparison

Within the 4DNvestigator we present a novel statistical method for detecting differences in Hi-C matrices, the LP method. Specifically, we test the null hypothesis that the Hi-C correlation matrices, derived from different samples, are equivalent (See Section Hi-C matrix comparison). To assess the LP method's ability to detect differences in Hi-C matrices between samples, we have compared the LP method against alternate Hi-C comparison methods: HiCRep, HiC-spector, and SELFISH [4, 188, 189]. HiCRep, HiC-spector are Hi-C reproducibility metrics, while SELFISH is a DCI method that was recently shown to outperform all prior methods for DCI detection. We note that there are no methods, that we are aware of, that are a direct comparison with our method. All of the alternate methods we are comparing here were designed to address related, but fundamentally different, questions: Hi-C reproducibility metrics for genome-wide comparisons, and DCI for loci interaction differences (not overall structural differences within a region). Nevertheless, we compare them here to highlight where our method is advantageous.

There are many ways in which Hi-C matrices can be different between samples. Two well established Hi-C features are loops and compartments [102, 139]. To assess how the LP method performs when these features are different between samples, we have created simulated Hi-C data sets with incremental changes in these features. We then determined the point at which the LP method, as well as alternate methods, detect differences between the matrices (See Section Simulated Hi-C data).

For Hi-C matrices with changes in loop structure as the only differences between samples, the LP method does not detect that the matrices are different until the loop

interaction contained four times the mean number of contacts for loci at the given genomic distance (See Figure 7.5A). In this situation, SELFISH does the best job of detecting differences between samples. HiC-spector shows a consistent decrease in its reproducibility score, while HiCRep performs similar to the LP method. We note that the LP method, as well as HiCRep, relies on changes in the correlation between samples, and thus changes to a small number of elements within the Hi-C matrices do not affect the measurement significantly.

For Hi-C matrices with changes in the compartment structure of a single bin (genomic region) as the only differences between samples, the LP method performs very well (See Figure 7.5B). SELFISH detects differences between all samples, but also detects differences when only a small amount of noise is added to the original matrix. The LP method is robust to noise and shows a trend consistent with the amount of change created. The HiC-spector reproducibility measurement shows no clear trend as the matrices diverge, while HiCRep begins to detect differences between the matrices much later than the LP method. Compartment changes are often observed in time series Hi-C matrices as cells transition between cell states, either through reprogramming or differentiation [52].

7.5 Discussion

We believe the 4DN research community will find the 4DNvestigator suite of tools useful for the analysis time series Hi-C and RNA-seq data. The 4DN feature analyzer methods have already been used to detect important 4DN changes in times series Hi-C and RNA-seq data sets [103]. Providing the standardized code here within the 4DNvestigator will allow a wider audience of researchers to apply this technique to their own work. Methods to systematically quantify stemness and plasticity via VNE will help identify cells in specific states. Such methods may be especially useful in the context of cancer, as identifying and targeting cells with high stemness is an attractive

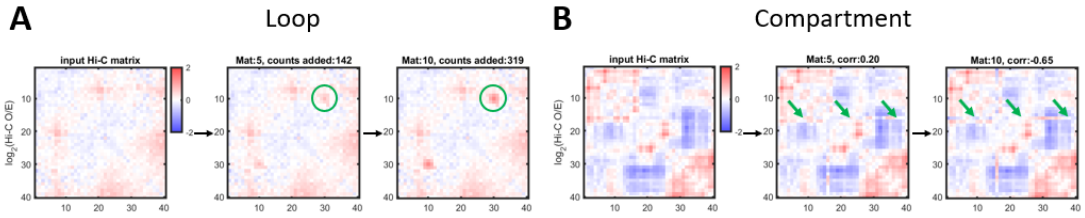


Figure 7.5: Simulated Hi-C data. A) Simulated Hi-C matrices for loop changes (See Table 7.1). The original Hi-C matrix is shown on the left, with matrices that are increasingly more divergent from left to right. Green circles indicate where the matrix has been perturbed to change the chromatin loop structure. B) Simulated Hi-C matrices for compartment changes (See Table 7.1). The original Hi-C matrix is shown on the left, with matrices that are increasingly more divergent from left to right. Green arrows indicate where the matrix has been perturbed to change the chromatin compartment structure.

	Counts added					Correlation					
Method	35	106	177	248	319	1 (noise)	0.87	0.57	0.2	-0.26	-0.65
LP (P-value)	1.0	1.0	1.0	1.0	0.86	1.0	1.0	2.1E-03	2.5E-07	1.7E-13	0.0
SELFISH (P-value)	0.0	0.0	0.0	0.0	0.0	0.02	2.9E-05	1.4E-08	1.3E-10	3.5E-13	2.1E-07
HiCRep (SCC)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.98	0.97	0.96
HiC-spector (Q)	1.0	0.99	0.97	0.96	0.95	0.39	0.43	0.33	0.48	0.44	0.50

Table 7.1: Hi-C matrix comparison methods. "Counts added" refers to the number of counts added to the simulated Hi-C matrix, at the specified location (See Section "Simulated Hi-C data"). "Correlation" refers to the correlation between the selected column in the original Hi-C matrix, and the same column in the simulated Hi-C matrix (See Section "Simulated Hi-C data"). We note here that each method outputs a different unit, as specified within the table. The lowest P-value output from SELFISH is given here. This table is related to Figure 7.5.

clinical strategy [168].

The LP method for comparing Hi-C matrices will be extremely valuable to the Hi-C research community as it fulfills a number of unmet needs. The LP method can be used to analyze a range of Hi-C resolutions, from 5 kb resolution to 100 kb chromatin compartment changes, and all resolutions in between. Many Hi-C analysis techniques were developed to detect specific biological phenomena within Hi-C matrices, such as loops or compartments, and are therefore less robust for detecting overall changes in chromatin structure. Also, the LP method outputs a p-value for the simple statistical test that entire chromatin structure is different between the samples being compared, not just specific chromatin interactions, making the LP technique more robust. Additionally, the LP method is built for the analysis of multiple Hi-C samples, accommodating time series data which often include more than two time points.

7.6 Conclusion

The 4DNvestigator provides methods to analyze time series Hi-C and RNA-seq data in a rigorous yet automated manner. The combined analysis of network centrality and RNA-seq over time can be easily performed using the 4DN feature analyzer. This analysis outputs a simple 2D plot representing how the genome changes in 4DN, with large changes highlighted. Network entropy is a simple metric which can be used to characterize the structural disorder in a region. Finally, the 4DNvestigator introduces a novel statistical method for comparing Hi-C matrices, the LP method. The LP method is distinct from established Hi-C matrix comparison methods, as it takes a statistical approach to test for matrix equality, and allows for the comparison of ≥ 2 matrices simultaneously. The 4DNvestigator can be applied to any time series Hi-C and RNA-seq data set, and many methods are applicable to non-time series data samples as well.

CHAPTER VIII

Concluding Remarks

The field of genomics has advanced rapidly in the 21st century, and shows no sign of slowing down in the coming years. The completion of the human reference genome, less than 20 years ago, substantially advanced our understanding of the rules of life [80]. With this information, we have learned how mutations can lead to disease states, how preventative medications will react in individuals with specific mutations, how the human species evolved, and much more. The creation of the human reference genome catalyzed the invention of next generation sequencing technologies, which can read millions of DNA sequences in parallel. The combination of these advancements has led to experimental techniques that can measure many aspects of genome structure and function, including RNA-seq, Hi-C, ChIP-seq, and ATAC-seq. Furthermore, with the recent discovery of CRISPR-Cas9 gene editing, we not only have the genome mapped, we have the tools to manipulate it as well [35].

With this information as a foundation, the next frontier in genomic research is the 4D Nucleome. While genome function receives plenty of attention in the research community, genome structure and dynamics, and how these aspects affect genome function and cell phenotype, are underappreciated. Fortunately, this type of research has gained momentum in recent years from the founding of the 4D Nucleome project, where researchers within the field can contribute to a public database of 4DN data

[47].

This disparity in research attention is understandable; examining structure and dynamics is often difficult, both in terms of data collection and analysis. Each additional time point in an experiment adds a proportional amount of time and resources to the project. Genome structure data collection can be burdensome, often requiring an order of magnitude increase in both the budget and data storage capacity. Cell images, while inexpensive, are often stored in bulky data files. Additionally, Hi-C data sets are larger, more expensive, and more difficult to collect than RNA-seq data sets. Furthermore, analysis tools are underdeveloped for 4DN data sets, discouraging many researchers from collecting these data.

However, the benefits that a 4DN approach provides are undeniable. 4DN analysis explains the mechanisms for how cells operate. For example, a mutation in a TAD boundary region that affects nearby gene function would be difficult to explain prior to Hi-C. With Hi-C, we can see that this region may act as a conformational insulator between the affected gene and an enhancer element. This scenario is not just speculative; this situation has been observed in real data [108]. Furthermore, the genome is not a static object. Analysis of structure and function data obtained from a population of proliferating cells is muddled by cell cycle dependent changes; G1 phase cells display drastically different structure and function than M phase cells, as would be expected. Even in cell cycle synchronized populations, chromatin has been shown to move up to $0.5\mu m$ in less than 10 seconds [50]. This means that Hi-C samples collected seconds apart from each other could contain distinctly different information. While data collection on this time resolution scale is not yet possible for Hi-C, larger scale changes can be clearly observed using time series Hi-C.

To improve cellular reprogramming, we need 4DN analysis. Cellular reprogramming is not a magical one-step process where cells instantly transform into the target cell type after introduction of TFs; there exists a continuum of change up to the point

where a reprogrammed cell finally functions as the target cell type. 4DN analysis helps us understand the roadblocks along the way to reprogramming. For example, 4DN data analysis has revealed that chromatin changes occur prior to functional changes in many experiments, which is why the addition of chromatin remodeling proteins often increases reprogramming efficiency [90, 103, 149]. With this enhanced understanding of the reprogramming process, we can design better experimental strategies.

I hope that the work presented within this dissertation will advance 4DN and cellular reprogramming research and motivate others to continue this work. 4DN analysis and cellular reprogramming have the potential revolutionize medicine in many ways, and we have just begun to scratch the surface of their potential.

Chapter II, which outlines methods for analyzing time series FISH images, could be used as a clinical diagnostic tool. Compared to high-throughput genomics, cellular imaging is an easy, fast, and inexpensive tool for analyzing genome structure. This makes it an appealing measure for diagnostics where time is of the essence. Cellular imaging has been used to diagnose diseases for many years, but often images are taken as single snapshots. With the methods outlined within this chapter, clinicians could use time series images taken from the patient throughout treatment to determine if cells are responding as intended.

The methods outlined in Chapter III provide a framework for 4DN analysis. This was the first paper to collect and analyze time series RNA-seq, Hi-C, and FISH data all derived from the same synchronized cell population. The methods for extracting gene regulatory networks from Hi-C and RNA-seq over time are novel, and can be used to determine network stability over time. Unstable networks are often more susceptible to outside influence, making this timing crucial for controlling cell fate. This analysis could be applied to a diseased population of cells, and the time point of lowest stability would be the best time to administer treatment. This work also highlights the importance of knowing a gene's location in the nucleus for understanding its

function. The circadian rhythm cycle, one of the most dominant biological pathways, displays a clear correlation of gene positioning and function. This could potentially explain sleep pattern disorders if this interplay between structure and function breaks down.

The work discussed in Chapter IV builds upon the 4DN framework given in Chapter III to create an algorithm for cellular reprogramming. Within Chapter III, TADs are identified as an inherent functional unit within the genome, where expression of genes within TADs display correlated expression over time. This idea is consistent with the concept of transcription factories, where nearby segments of chromatin are found to be transcribed at the same time, by the same transcriptional machinery [37]. This work demonstrates the tangible benefits of 4DN analysis by predicting TFs that are known to achieve reprogramming strictly from data. With this power, cellular reprogramming experiments can be performed more efficiently and cost effectively, bringing us one step closer to the day where these techniques can be translated to clinical therapies. People suffering from organ failure may be able to replace the dysfunctioning organ with cells derived from their own body.

Following from the successful reprogramming predictions made in Chapter IV, Chapter V takes an in-depth look at one of the simplest known reprogramming experiments, MYOD1-mediated cellular reprogramming. Through analysis of time series Hi-C, RNA-seq, and FISH imaging data, we uncovered principles of cellular reprogramming that inform our algorithm predictions. The discovery that genome structure changes significantly before function motivates the use of chromatin remodeling proteins in reprogramming predictions. DNA and histone demethylating proteins could open up heterochromatin regions, making these regions susceptible to control. The mutual feedback discovered between MYOD1 and the circadian gene network highlights the unpredictability of gene networks. The idea that MYOD1, a regulator of the muscle lineage, would have such a profound influence on circadian

rhythms was unexpected. All of these findings supplement our goal of creating an algorithm for predicting reprogramming factors, which could have a wide range of clinical applications.

A natural application area for 4DN analysis is in cancer research. Since a hallmark of cancer progression is the development of aneuploidy and genetic mutations, the cell's 4DN will clearly be affected. In Chapter VI, the 4DN of two distinct cancer cell types are explored, colorectal cancer cells and breast cancer cells. In the colorectal cancer analysis, cells with a premalignant karyotype (trisomy 7) are compared to their health counterpart. Changes in the cell's 4DN are observable genome-wide and specific genomic pathways that are linked to cell proliferation are already significantly affected in the premalignant sample. In the breast cancer analysis, we focus on characterizing the 4DN of CSCs, which play and important role in mediating tumor metastasis and treatment resistance. Surprisingly, we observed high heterogeneity in both CSC and non-CSC subpopulations, and a clear preference for specific translocations between the populations. To effectively treat cancer, a complete understanding of where and when to administer treatment is necessary. Together, these findings further our understanding of the 4DN of cancer, which we hope will lead to more effective treatment strategies.

Finally, we package the tools we have developed through our years of 4DN research into the 4DNvestigator, as discussed in Chapter VII, alleviating the current barrier to entry in this field that requires a high level understanding of biology, computer science, and mathematics to make sense of the large complex data sets. By providing researchers with user-friendly codes that automatically perform complex analysis methods, anyone with computer and internet connection can analyze 4DN data. With standardized tools for 4DN analysis, 4DN research will be much more accessible to researchers in the future.

APPENDIX

APPENDIX A

Supplemental Materials

A.1 Functional organization of the human 4D Nucleome supplement

For a complete description of methods, please refer to Chen *et al.* SI Appendix [31].

Criteria for extracting dynamic genes

We establish several criteria to extract genes that have significant dynamics of transcription over time. These genes can be considered as genes that are highly expressed in Fibroblast cells and have response to serum stimulation. The criteria are listed here below:

- At least 1.5 fold difference in expression between any two time points (q value < 0.01).
- Mean expression over time RPKM value larger than 1.
- Variance over time larger than 0.5.

- Ratio between variance over time and variance within each time point replicate larger than 1.5.

Measures used for characterizing gene locations in FISH data

- **Mean closest distance (MCD) between two genes**: MCD is defined as the shortest distance between two different gene locations (given that there are two homologous genes per target, there are 4 different "distances" between two gene locations; take the shortest distance). MCD is meant to be used as a measure of relative spatial interaction between two genes at a given time point, with a lower distance correlating to a higher likelihood of genomic contact.

- **Distance Matrix**: The matrix with entries denoting the distance in Euclidean space between two genes.

- **Index of Stability**: Derived from the Fiedler number of normalized Laplacian of the distance matrix. This can be computed for any size distance matrix (i.e. number of genes) to characterize relative structure. Our analysis used the normalized Laplacian to create a scale invariant measure of stability.

Rationale for MCD

MCD was developed as a simple measurement of inter-gene distance. Genes measured with MCD in our analysis are biallelic, meaning each gene has a copy that also is transcriptionally active. As a result of these gene copies, when measuring the distances between two genes there are 4 possible combinations. The shortest distance derived from the 4 possible combinations, MCD, does not take into account the gene copies, but was chosen to give a general picture of how close two genes may be at certain time point.

A.2 Algorithm for cellular reprogramming supplement

For a complete description of methods, please refer to Ronquist *et al.* SI Appendix [147].

Data

The fibroblast (FIB) data (Hi-C and RNA-seq) used for this application was originally collected and published in a paper by Chen *et al.* [31]. We refer the reader to this paper for a full description of technical protocols. ESC and myotube (MT) data was downloaded from NCBI-GEO (GSE23316 ENCODE Caltech RNA-seq and GSE52529) [36]. 53 different tissue RNA-seq samples were downloaded from GTEx portal [106]. 51 different immune cell type RNA-seq samples were obtained from the BLUEPRINT Epigenome project [1].

Hi-C and construction of TADs

We computed TAD boundaries from genome-wide chromosome conformation capture (Hi-C) data using an algorithm described in Chen *et al.* [32]. The algorithm was applied to averaged time series Hi-C data from proliferating human fibroblast (FIB) at 100 kilo-base pair (kb) resolution, which identified 2,562 TADs across all autosomal chromosomes (i.e. excluding Chromosomes X and Y). Of the 2,562 TADs, 317 contained no genes and were excluded from our analysis, leaving 2,245 TADs. These TADs ranged in size from a few hundred kb to several Mb, and contained on average 9-10 genes (standard deviation of 18 genes); one gene at minimum, and 249 genes maximum.

Construction of B matrix

TF binding site position frequency matrix (PFM) information was obtained from Neph *et al.* and MotifDB, which is a collection of publicly available PFM databases

such as, JASPAR, Jolma *et al.* cispb_1.02, stamlab, hPDI, UniPROBE [123, 157]. TRANSFAC PFM information was included as well. Motif scanning of the human reference genome (hg19) was performed using FIMO of the MEME suite, in line with methods established by Neph *et al.* [123]. DNase-seq information for human fibroblasts was derived from ENCODE for fibroblast (GSM1014531). If a narrow peak is found within the ±5kb of a gene TSS, the region is classified as open. TF function information was determined through an extensive literature search.

Scaling of RNA-seq

Due to differences in data collection procedures, the RNA-seq RPKM values obtained from the GTEx portal were of lower value, on average, compared to our fibroblast data set, thus favoring repressor TFs for μ scoring. In order to account for this in our model, we scaled all GTEx RNA-seq data by a factor that solves the equation

$$\underset{\alpha}{\text{minimize}} \quad \|\mathbf{g}_{FIB,UM} - \alpha \mathbf{g}_{FIB,GTEX}\| \tag{A.1}$$

where $g_{FIB,UM}$ is the gene-level RNA-seq vector average of our fibroblast data, $g_{FIB,GTEX}$ is the gene-level RNA-seq vector of "Cells - Transformed fibroblasts" from the GTEx portal, and α is a scalar that solves this equation. For this data, $\alpha = 2.6113$ and all GTEx data used as a target state was scaled by this factor.

Removal of microRNA

MicroRNA were removed from this analysis due to their high variance in RPKM values and unpredictable function.

TF scores - additional GTEx data

For fibroblast to Adipose-Subcutaneous, the highest scoring factor is EBF1, a known maintainer of brown adipocyte identity, and a known promoter of adipogenesis in fibroblasts [85]. The 2nd highest scoring marker, PPARG, has been shown to be involved in adipose differentiation, and can be used individually to achieve reprogramming from fibroblasts [67]. Curiously, ATF3 is implicated here as being useful for adipocyte differentiation although its function has been shown to repress PPARG and stymie cell proliferation [83]. Upon further research using time dependent addition, ATF3 addition scores best when added towards the end of reprogramming process.

Two Brain tissue samples, Cerebellum and Hippocampus, both predict TFs necessary for natural differentiation. Interestingly, our algorithm selects different TFs for each conversion, with factors linked specifically to each tissue. For Cerebellum, NEUROD1, has been shown to be required for granule cell differentiation, while ZIC1 and ZIC4 are both known to promote cellebular-specific neuronal function [61, 116]. The top scoring combination of 3 TFs are all similarly known to be important in neurogenesis (NEUROD1, ZBTB18, UNCX) [34, 151]. Hippocampus TF scoring includes FOXG1 as the top predicted factor, a factor specifically needed in hippocampus development. OLIG2, FOXG1, and GPD1 are the top scoring set for hippocampus reprogramming, all of which have been shown to been necessary for hippocampus function.

Colon TF scoring finds known differentiation factor in natural colon secretory linage development, ATOH1, as the highest scoring individual factor [178]. The top scoring combination of 2 TFs includes ATOH1 along with CDX2, another known factor necessary for full differentiation of colon cells, specifically small intestine maturation [41]. Liver cell reprogramming similarly finds known factors for differentiation in the top score of all 3 combinations: HNF4A, CUX2, PROX1 [48, 155, 177]. All

TFs play a role in correct development of hepatic progenitor cell-types and hepatic stem cells, the cell types just above in lineage differentiation.

Algorithm for cellular reprogramming data sources

A summary of the data used for this algorithm is shown below, with citations and accession numbers or website link, where applicable.

- Gene Expression
  - Fibroblast: Chen *et al.* [31]
  - GTEx: `https://www.gtexportal.org/home/` [106]
  - ESC: GSE23316 [36]
  - Myotube: GSE52529 [36]
  - BluePrint Epigenome: `http://www.blueprint-epigenome.eu/` [1]
- DNase-seq
  - Fibroblast: GSM1014531 [172]
- Hi-C TAD boundaries
  - Fibroblast: Chen *et al.* [31]
- TF PWM
  - Neph *et al.* [123]
  - MotifDB + FIMO [66, 157]

A.3 MYOD1-mediated fibroblast to muscle reprogramming supplement

For a complete description of methods, please refer to Liu *et al.* SUPPLEMENTAL INFORMATION [103].

Identification of genes of interest

Genes of interest (GOIs) are mainly extracted through Gene Ontology (GO), with a few GOI subsets curated through other means. GO-extracted lists include myotube, myoblast, skeletal muscle, fibroblast, and circadian. "Muscle" genes are the union of myoblast, myotube, and skeletal muscle genes. Additional circadian related subsets were extracted from JTK analysis and literature reviews (core circadian), and additional cell cycle subsets were extracted from literature reviews (Table S3 in Liu *et al.* [103]).

Super enhancer-promoter region dynamics

SE-P regions for skeletal muscles were downloaded from [75] (BI_Skeletal_Muscle). The Hi-C contacts between the SE and the associated gene TSS ($\pm$1kb) were extracted over time. SE-P contacts were normalized by dividing by the total number of contacts per sample, then multiplying by 100,000,000 (arbitrary scalar to best show trends). To determine the top upregulated genes, the linear regression slope of log2(FPKM) over time was calculated and sorted for each gene, high to low. To determine significance, we first normalized the contacts by dividing by the total number of contacts for each SE-P region over time (so that all SE-P regions are on the same relative scale). We then performed a t-test between 16-24 hr and -48,0-8 hr normalized contacts.

A.4 The 4DN of cancer supplement

Trisomy 7 data generation and methods

For a complete description of methods, please refer to Braun *et al.* [18].

FISH image analysis

Nuclei were extracted and analyzed semi-automatically through our MATLAB image processing pipeline. Nuclei boundaries were detected via Marker-Controlled Watershed Segmentation, following methods modified from (`https://www.mathworks.com/help/images/marker-controlled-watershed-segmentation.html`). The z-dimension provided little additional spatial resolution, so maximum projection of the images was analyzed. DAPI channel images were converted to grayscale and the image gradient magnitude was computed to determine nuclei borders. The original grayscale images were eroded then dilated to remove noise and create a flat regional maximum within nuclei. Gradient boundaries enclosing regional maxima determined where nuclei were found. The convex hull of the gradient boundaries defined the nuclei shape.

Each nucleus was analyzed individually, and chromosome territories (CTs) were automatically extracted using a method similar to that of nuclei extraction, but within each defined nucleus region. For each fluorescent channel (containing CT 19 and 7), the image was passed through a Gaussian filter and binarized. Images were then dilated and eroded to reduce noise. Connected components reflecting CTs were manually selected. The CT area, CT centroid distance to nuclear periphery and the CT centroid distance to other centroids were then calculated automatically. For HCEC+7, only nuclei that had 3 copies of chromosome 7 were kept. 42 cells were used for HCEC, 41 cells were used for HCEC+7.

RNA-seq data analysis

Reads were aligned to reference genome hg19 using STAR [55]. Gene expression was quantified using RSEM [100]. Alignment parameters were set based on RSEM default parameters, "rsem-calculate-expression" "–star". Transcripts per million (TPM) was used to quantify gene expression. TPM expression was binned into 100kb regions for comparison with Hi-C matrix dimensions. To achieve this, the sum of all gene-level TPMs was calculated for all genes within each 100kb region. 100kb regions that contain only a proportion of a gene (i.e. the gene spans multiple 100kb regions) received a TPM value proportional to how much the gene overlaps that region.

Differential expression was determined following methods described in [2]. p-values for differential expression were adjusted for false discovery following methods outlined in [163]. GSAASeqSP was used for Gene Set Analysis of RNA-seq data [187]. Raw counts were input to GSAASeqSP in the form of a .gct file, and the MSigDB hallmark gene sets were analyzed [101]. The following GSAASeqSP parameters were used: "Permutation type: gene_set", "Metric for gene set analysis: Weighted_KS", "Metric for differential expression analysis: Signal2Noise", "Association statistic: weighted", and "P value threshold: 0.05".

Generation and analysis of Hi-C matrices

Paired end reads were processed using the juicer pipeline with default parameters [56]. Reads were mapped to reference genome hg19, with "-s" (site parameter) MboI. Reads with MAPQ > 30 were kept for further analysis. Data was extracted and input to MATLAB using Juicebox tools command "dump". Knight-Ruiz (KR) normalization was applied to all matrices, observed over expected (O/E) matrices were used for A/B compartmentalization and identification of topologically associating domains (TADs). Rows and columns for which more than 10% of entries had zeros were removed from the matrix. A/B compartmentalization was determined using the Fiedler

vector, as previously described [31]. TAD structure was determined using previously described methods [32], with $\lambda_{thr} = 0.8$, and minimum TAD size defined as 300 kb. This technique is derived from spectral graph theory, and is used for graph segmentation. Here, the parameters are set so that 300kb is the minimum domain size and 0.8 is the minimum Fiedler number of each domain. The Fiedler number is a measure of graph connectivity, with higher values implying stronger connectivity.

CSC data generation and methods

Sample preparation

The breast cancer cell line SUM-159 was cultured in 2D cell culture, then flow cytometry was used to sort cells with the top 10% and bottom 10% expression of *ALDH1A1*, a previously established method for distinguishing CSC-like and nonCSC-like cellular populations. Hi-C and RNA-seq samples were collected and processed as described in Seaman *et al.* [154].

Hi-C and RNA-seq processing

Hi-C and RNA-seq data generation were performed as described in [18]. Paired end reads were processed using the juicer pipeline with default parameters [56]. Reads were mapped to reference genome hg19, with "-s" (site parameter) MboI. Reads with MAPQ > 30 were kept for further analysis. Data was extracted and input to MATLAB using Juicebox tools command "dump". Knight-Ruiz (KR) normalization was applied to all matrices, observed over expected (O/E) matrices were used for A/B compartmentalization. Rows and columns for which more than 10% of entries had zeros were removed from the matrix. A/B compartmentalization was determined using the Fiedler vector, as previously described [31].

Reads were aligned to reference genome hg19 using STAR [55]. Gene expression was quantified using RSEM [100]. Alignment parameters were set based on RSEM de-

fault parameters, "rsem-calculate-expression" "–star". Transcripts per million (TPM) was used to quantify gene expression. TPM expression was binned into 100kb regions for comparison with Hi-C matrix dimensions. To achieve this, the sum of all gene-level TPMs was calculated for all genes within each 100kb region. 100kb regions that contain only a proportion of a gene (i.e. the gene spans multiple 100kb regions) received a TPM value proportional to how much the gene overlaps that region.

Identification of chromosome tranlocations

Differences in genomic translocations were detected via the following method. Genome-wide Hi-C contact maps were generated at 1 Mb resolution for the ALDH+ and ALDH- samples. Matrices were KR balanced and O/E normalized. Each matrix was median filtered to smooth the signal and reduce experimental noise. The "difference matrix" was determined by calculating the absolute value of the difference between the matrices. A threshold was set equal to five times the mean of all elements in the difference matrix. This threshold was used to binarized the difference matrix (i.e. all elements above the threshold are set to 1, all elements below the threshold are set to 0). Large regions in this binarized matrix are determined to be regions were tranlocation differences exist between the two matrices.

Determination of ALDH1A1 chromatin accessibility

Chromatin accessibility is often determined via the Fielder vector, derived from 100 kb resolution intr-chromosome Hi-C. This analysis can be compromised on chromosomes with translocations and large copy number variations. Chromosome 9 in SUM-159 cells contains a regions on the p-arm that obfuscates Fiedler vector partitioning. To determine the chromatin accessibility of the ALDH1A1 locus (located on chromosome 9), we performed Fiedler vector partitioning on a sub-region surrounding the locus ($\pm$3 Mb).

Low dimensional projection of chromatin structure and function

To identify differences between samples in structure and function we projected measurements of each to a low dimensional space using principal component analysis (PCA). For function, we applied PCA to the $\log_2$(TPM) vector for all samples. This outputs the location of four points in a 2 dimensional space, where each points represents a sample and the distance between samples approximates how different the samples are in function genome-wide. For structure, we first concatenated the Fiedler vectors derived from each chromosome at 100 kb resolution for each sample. We then applied PCA to this vector for all samples. This outputs the location of four points in a 2 dimensional space, where each points represents a sample and the distance between samples approximates how different the samples are in structure genome-wide.

A.5 4DNvestigator supplement

Data pre-processing

The 4DNvestigator requires a metadata table, referred to as an "Index File," to describe the data. Rows correspond to sequencing data samples (RNA-seq or Hi-C), and the columns correspond to data descriptors. The Index File must have 4 columns with the following headers:

- "path" defines the computer path to the sequencing data
- "dataType" defines the type of sequencing data, either "hic" or "rnaseq"
- "sample" defines the sample, (e.g. "treatment" or "control")
- "timePoint" defines the time point of sample, (e.g. 0 or 24)

Index Files can be .csv, .tsv, or .xls. An example of an Index File is available here.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology* **30,** 224–226 (2012).

2. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11,** 1–12 (2010).

3. Andrews, J. L. *et al.* CLOCK and BMAL1 regulate MyoD and are necessary for maintenance of skeletal muscle phenotype and function. *Proceedings of the National Academy of Sciences* **107,** 19090–19095 (2010).

4. Ardakany, A. R., Ay, F. & Lonardi, S. Selfish: Discovery of Differential Chromatin Interactions via a Self-Similarity Measure. *bioRxiv,* 540708 (2019).

5. Aström, K. J. & Murray, R. M. *Feedback systems: an introduction for scientists and engineers* (Princeton university press, 2010).

6. Balsalobre, A. *et al.* Resetting of circadian time in peripheral tissues by glucocorticoid signaling. *Science* **289,** 2344–2347 (2000).

7. Bauman, J. G., Wiegant, J., Borst, P. & Van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Experimental cell research* **128,** 485–490 (1980).

8. Beliveau, B. J. *et al.* Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nature communications* **6** (2015).

9. Beliveau, B. J. *et al.* Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences* **109,** 21301–21306 (2012).

10. Belkin, M. & Niyogi, P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Nips* **14,** 585–591 (2001).

11. Bentzinger, F. C., Wang, Y. X. & Rudnicki, M. A. Building Muscle : Molecular Regulation of Myogenesis. *Cold Spring Harbor perspectives in biology* **4,** 1–16 (2012).

12. Berger, A. B. *et al.* High-resolution statistical mapping reveals gene territories in live yeast. *Nature Methods* **5,** 1031–1037 (2008).

13. Bickmore, W. A. The spatial organization of the human genome. *Annual review of genomics and human genetics* **14,** 67–84 (2013).

14. Böck, M. *et al.* Identification of ELF3 as an early transcriptional regulator of human urothelium. *Developmental biology* **386,** 321–330 (2014).

15. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biology* **3,** 0826–0842 (2005).

16. Bomme, L. *et al.* Assessments of clonal composition of colorectal adenomas by FISH analysis of chromosomes 1, 7, 13 and 20. *International journal of cancer* **92,** 816–823 (2001).

17. Bottaro, D. P. *et al.* Identification of the hepatocyte growth factor receptor as the c-met proto-oncogene product. *Science* **251,** 802–804 (1991).

18. Braun, R. *et al.* Single Chromosome Aneuploidy Induces Genome-Wide Perturbation of Nuclear Organization and Gene Expression. *Neoplasia (United States)* **21,** 401–412 (2019).

19. Brewster, R. C. *et al.* The transcription factor titration effect dictates level of gene expression. *Cell* **156,** 1312–1323 (2014).

20. Brockett, R. W. *Finite Dimensional Linear Systems* (John Wiley & Sons, Inc., New York, USA, 1970).

21. Buhr, E. D. & Takahashi, J. S. in *Circadian clocks* 3–27 (Springer, 2013).

22. Bulut-Karslioglu, A. *et al.* The Transcriptionally Permissive Chromatin State of Embryonic Stem Cells Is Acutely Tuned to Translational Output. *Cell Stem Cell* **22,** 369–383 (2018).

23. Bustin, M. & Misteli, T. Nongenetic functions of the genome. *Science* **352,** aad6933 (2016).

24. Cahan, P. *et al.* CellNet: Network biology applied to stem cell engineering. *Cell* **158,** 903–915 (2014).

25. Calandrelli, R., Wu, Q., Guan, J. & Zhong, S. GITAR: An Open Source Tool for Analysis and Visualization of Hi-C Data. *Genomics, Proteomics and Bioinformatics* **16,** 365–372 (2018).

26. Canela, A. *et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170,** 507–521 (2017).

27. Cao, Y. *et al.* Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming. *Developmental Cell* **18,** 662–674 (2010).

28. Cavalli, G. & Misteli, T. Functional implications of genome topology. *Nature structural & molecular biology* **20,** 290 (2013).

29. Chang, H. Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biology* **2,** 206–214 (2004).

30. Chen, H. *et al.* Chromosome conformation of human fibroblasts grown in 3-dimensional spheroids. *Nucleus* **6,** 55–65 (2015).

31. Chen, H. *et al.* Functional organization of the human 4D Nucleome. *Proceedings of the National Academy of Sciences* **112,** 8002–8007 (2015).

32. Chen, J., Hero, A. & Rajapakse, I. Spectral Identification of Topological Domains. *Bioinformatics* **32,** 2151–2158 (2016).

33. Clemson, C. M., Hall, L. L., Byron, M., McNeil, J. & Lawrence, J. B. The X chromosome is organized into a gene-rich outer rim and an internal core containing silenced nongenic sequences. *Proceedings of the National Academy of Sciences* **103,** 7688–7693 (2006).

34. Cohen, J. *et al.* Further evidence that de novo missense and truncating variants in ZBTB18 cause intellectual disability with variable features. *Clinical genetics* **91,** 697–707 (2017).

35. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science,* 1231143 (2013).

36. Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

37. Cook, P. R. The organization of replication and transcription. *Science* **284,** 1790–1795 (1999).

38. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

39. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics* **2,** 292–301 (2001).

40. Cremer, T. & Cremer, M. Chromosome Territories. **147,** 13–24 (1999).

41. Crissey, M. A. S. *et al.* Cdx2 levels modulate intestinal epithelium maturity and Paneth cell development. *Gastroenterology* **140,** 517–528 (2011).

42. Croft, J. A. *et al.* Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of cell biology* **145,** 1119–1131 (1999).

43. Daily, K., Patel, V. R., Rigor, P., Xie, X. & Baldi, P. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC bioinformatics* **12,** 495 (2011).

44. D'Alessio, A. C. *et al.* A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports* **5,** 763–775 (2015).

45. De Solórzano, C. O., Santos, A., Vallcorba, I., Garcia-Sagredo, J.-M. & del Pozo, F. Automated FISH spot counting in interphase nuclei: Statistical validation and data correction. *Cytometry* **31,** 93–99 (1998).

46. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).

47. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549,** 219–226 (2017).

48. DeLaForest, A. *et al.* HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development* **138,** 4143–4153 (2011).

49. Dierickx, P., Van Laake, L. W. & Geijsen, N. Circadian clocks: from stem cells to tissue homeostasis and regeneration. *EMBO reports* **19,** 18–28 (2018).

50. Dion, V. & Gasser, S. M. Chromatin movement in the maintenance of genome stability. *Cell* **152,** 1355–1364 (2013).

51. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Molecular cell* **62,** 668–680 (2016).

52. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518,** 331–336 (Feb. 2015).

53. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

54. Djekidel, M. N., Chen, Y. & Zhang, M. Q. FIND: DifFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research* **28,** 412–422 (2018).

55. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

56. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3,** 95–98 (2016).

57. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology* **34,** 1180–1190 (2016).

58. Ferrán, B. *et al.* The nuclear receptor NOR-1 regulates the small muscle protein, X-linked (SMPX) and myotube differentiation. *Scientific reports* **6,** 1–11 (2016).

59. Fischer, A., Schumacher, N., Maier, M., Sendtner, M. & Gessler, M. The Notch target genes Hey1 and Hey2 are required for embryonic vascular development. *Genes & development* **18,** 901–911 (2004).

60. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome biology* **16,** 180 (2015).

61. Frank, C. L. *et al.* Regulation of chromatin accessibility and Zic binding at enhancers in the developing cerebellum. *Nature neuroscience* **18,** 647–656 (2015).

62. Gard, D. L. & Lazarides, E. The synthesis and distribution of desmin and vimentin during myogenesis in vitro. *Cell* **19,** 263–275 (1980).

63. Godin, A. G., Lounis, B. & Cognet, L. Super-resolution microscopy approaches for live cell imaging. *Biophysical journal* **107,** 1777–1784 (2014).

64. Grade, M., Becker, H., Liersch, T., Ried, T. & Ghadimi, B. M. Molecular cytogenetics: genomic imbalances in colorectal cancer and their clinical impact. *Analytical Cellular Pathology* **28,** 71–84 (2006).

65. Graf, T. Historical origins of transdifferentiation and reprogramming. *Cell stem cell* **9,** 504–516 (2011).

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

67. Gregoire, F. M., Smas, C. M. & Sul, H. S. Understanding adipocyte differentiation. *Physiological reviews* **78,** 783–809 (1998).

68. Gué, M., Messaoudi, C., Sun, J. S. & Boudier, T. Smart 3D-FISH: Automation of distance analysis in nuclei of interphase cells by image processing. *Cytometry Part A* **67,** 18–26 (2005).

69. Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A. & Xu, Y. SLICE: Determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Research* **45,** 1–14 (2017).

70. Gurdon, J. B. The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. *Journal of Embryology and Experimental Morphology* **10,** 622–640 (1962).

71. Habermann, J. K. *et al.* Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes, Chromosomes and Cancer* **46,** 10–26 (2007).

72. Hanson, R. E. *et al.* Fluorescent in situ hybridization of a bacterial artificial chromosome. *Genome* **38,** 646–651 (1995).

73. Hawes, S. E. *et al.* DNA hypermethylation of tumors from non-small cell lung cancer (NSCLC) patients is associated with gender and histologic type. *Lung cancer* **69,** 172–179 (2010).

74. Heselmeyer, K. *et al.* Gain of chromosome 3q defines the transition from severe dysplasia to invasive carcinoma of the uterine cervix. *Proceedings of the National Academy of Sciences* **93,** 479–484 (1996).

75. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).

76. Huang, E. H. *et al.* Aldehyde dehydrogenase 1 is a marker for normal and malignant human colonic stem cells (SC) and tracks SC overpopulation during colon tumorigenesis. *Cancer research* **69,** 3382–3389 (2009).

77. Hughes, M. E., Hogenesch, J. B. & Kornacker, K. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of biological rhythms* **25,** 372–380 (2010).

78. Iannuccelli, E. *et al.* NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments. *Bioinformatics* **26,** 696–697 (2010).

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

80. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

81. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes and Development* **28,** 2679–2692 (2014).

82. Iyer, V. R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283,** 83–87 (1999).

83. Jang, M. K., Kim, C. H., Seong, J. K. & Jung, M. H. ATF3 inhibits adipocyte differentiation of 3T3-L1 cells. *Biochemical and biophysical research communications* **421,** 38–43 (2012).

84. Jang, S. *et al.* KAT5-mediated SOX4 acetylation orchestrates chromatin remodeling during myoblast differentiation. *Cell death & disease* **6,** 1–11 (2015).

85. Jimenez, M. A., Åkerblad, P., Sigvardsson, M. & Rosen, E. D. Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade. *Molecular and cellular biology* **27,** 743–757 (2007).

86. Joliot, V. *et al.* The SWI/SNF subunit/tumor suppressor BAF47/INI1 is essential in cell cycle arrest upon skeletal muscle terminal differentiation. *PloS one* **9,** 1–11 (2014).

87. Jorgensen, P. *et al.* The size of the nucleus increases as yeast cells grow. *Molecular biology of the cell* **18,** 3523–3532 (2007).

88. Juan, A. H. *et al.* Polycomb EZH2 controls self-renewal and safeguards the transcriptional identity of skeletal muscle stem cells. *Genes & development* **25,** 789–794 (2011).

89. Kimura, E. *et al.* Cell-lineage regulated myogenesis for dystrophin replacement: a novel therapeutic approach for treatment of muscular dystrophy. *Human molecular genetics* **17,** 2507–2517 (2008).

90. Knaupp, A. S. *et al.* Transient and Permanent Reconfiguration of Chromatin and Transcription Factor Occupancy Drive Reprogramming. *Cell Stem Cell* **21,** 834–845 (2017).

91. Kosak, S. T. & Groudine, M. Form follows function: the genomic organization of cellular differentiation. *Genes & development* **18,** 1371–1384 (2004).

92. Kozubek, M. *et al.* High-resolution cytometry of FISH dots in interphase cell nuclei. *Cytometry* **36,** 279–293 (1999).

93. Krijger, P. H. L. *et al.* Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. *Cell Stem Cell* **18,** 597–610 (2016).

94. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* **72,** 65–75 (2015).

95. Larntz, K. & Perlman, M. D. A simple test for equality of correlation matrices. *Statistical Decision Theory and Related Topics IV,* 289–298 (1988).

96. Lassar, A. B., Paterson, B. M. & Weintraub, H. Transfection of a DNA locus that mediates the conversion of 10T12 fibroblasts to myoblasts. *Cell* **47,** 649–656 (1986).

97. Lerner, B., Clocksin, W. F., Dhanjal, S., Hultén, M. A. & Bishop, C. M. Feature representation and signal classification in fluorescence in-situ hybridization image analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **31,** 655–665 (2001).

98. Levsky, J. M., Shenoy, S. M., Pezo, R. C. & Singer, R. H. Single-cell gene expression profiling. *Science* **297,** 836–840 (2002).

99. Levsky, J. M. & Singer, R. H. Fluorescence in situ hybridization: past, present and future. *Journal of cell science* **116,** 2833–2838 (2003).

100. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12** (2011).

101. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1,** 417–425 (2015).

102. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–294 (2009).

103. Liu, S. *et al.* Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming. *iScience* **6,** 232–246 (2018).

104. Loh, K. M. *et al.* Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* **166,** 451–467 (2016).

105. Lohmann, G. *et al.* Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PloS one* **5,** 1–8 (2010).

106. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45,** 580–585 (2013).

107. Lun, A. T. & Smyth, G. K. diffHic: A Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16,** 1–11 (2015).

108. Lupi{\'a}{\˜n}ez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–1025 (2015).

109. Ly, P. *et al.* Characterization of aneuploid populations with trisomy 7 and 20 derived from diploid human colonic epithelial cells. *Neoplasia* **13,** 348–IN17 (2011).

110. Macarthur, B. D. & Lemischka, I. R. Statistical mechanics of pluripotency. *Cell* **154,** 484–489 (2013).

111. Mandai, M. *et al.* Autologous Induced Stem-Cell–Derived Retinal Cells for Macular Degeneration. *New England Journal of Medicine* **376,** 1038–1046 (2017).

112. Meeson, A. P. *et al.* Sox15 and Fhl3 transcriptionally coactivate Foxk1 and regulate myogenic progenitor cells. *The EMBO journal* **26,** 1902–1912 (2007).

113. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature reviews Molecular cell biology* **7,** 540 (2006).

114. Michael, D. G. *et al.* Model-based Transcriptome Engineering Promotes a Fermentative Transcriptional State in Yeast. *Proceedings of the National Academy of Sciences* **113,** E7428–E7437 (2016).

115. Misteli, T. Self-organization in the genome. *Proceedings of the National Academy of Sciences* **106,** 6885–6886 (2009).

116. Miyata, T., Maeda, T. & Lee, J. E. NeuroD is required for differentiation of the granule cells in the cerebellum and hippocampus. *Genes & development* **13,** 1647–1652 (1999).

117. Mora, A., Sandve, G. K., Gabrielsen, O. S. & Eskeland, R. In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics* **17,** bbv097 (2015).

118. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).

119. Mueller, F. *et al.* FISH-quant: automatic counting of transcripts in 3D FISH images. *Nature Methods* **10,** 277–278 (2013).

120. Murgha, Y. E., Rouillard, J.-M. & Gulari, E. Methods for the preparation of large quantities of complex single-stranded oligonucleotide libraries. *PLoS One* **9,** e94752 (2014).

121. Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342,** 948–953 (2013).

122. Nelson, D. O., Jin, D. X., Downs, K. M., Kamp, T. J. & Lyons, G. E. Irx4 identifies a chamber-specific cell population that contributes to ventricular myocardium development. *Developmental Dynamics* **243,** 381–392 (2014).

123. Neph, S. *et al.* Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell* **150,** 1274–1286 (2012).

124. Netten, H. *et al.* FISH and chips: automation of fluorescent dot counting in interphase cell nuclei. *Cytometry* **28,** 1–10 (1997).

125. Newman, M. *Networks: an introduction* (Oxford university press, 2010).

126. Ollion, J., Cochennec, J., Loll, F., Escudé, C. & Boudier, T. TANGO: a generic tool for high-throughput 3D image analysis for studying nuclear organization. *Bioinformatics,* btt276 (2013).

127. Ouadid-Ahidouch, H., Rodat-Despoix, L., Matifat, F., Morin, G. & Ahidouch, A. DNA methylation of channel-related genes in cancers. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1848,** 2621–2628 (2015).

128. Pacheco-Leyva, I. *et al.* CITED2 Cooperates with ISL1 and Promotes Cardiac Differentiation of Mouse Embryonic Stem Cells. *Stem Cell Reports* **7,** 1037–1049 (2016).

129. Park, M. *et al.* Sequence of MET protooncogene cDNA has features characteristic of the tyrosine kinase family of growth-factor receptors. *Proceedings of the National Academy of Sciences* **84,** 6379–6383 (1987).

130. Pinkel, D., Straume, T. & Gray, J. W. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proceedings of the National Academy of Sciences* **83,** 2934–2938 (1986).

131. Popken, J. *et al.* Remodeling of the nuclear envelope and lamina during bovine preimplantation development and its functional implications. *PloS one* **10** (2015).

132. Qian, L. *et al.* In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature* **485,** 593–598 (2012).

133. Rackham, O. J. L. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nature genetics* **48,** 331–5 (2016).

134. Ragoczy, T., Bender, M. A., Telling, A., Byron, R. & Groudine, M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes and Development* **20,** 1447–1457 (2006).

135. Raimondo, F. *et al.* Automated evaluation of Her-2/neu status in breast tissue from fluorescent in situ hybridization images. *IEEE Transactions on Image Processing* **14,** 1288–1299 (2005).

136. Rajapakse, I. & Groudine, M. On emerging nuclear order. *Journal of Cell Biology* **192,** 711–721 (2011).

137. Rajapakse, I., Groudine, M. & Mesbahi, M. Dynamics and control of state-dependent networks for probing genomic organization. *Proceedings of the National Academy of Sciences* **108,** 17257–17262 (2011).

138. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nature Methods* **14,** 263–266 (2017).

139. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680 (2014).

140. Ried, T. Homage to Theodor Boveri (1862–1915): Boveri's theory of cancer as a disease of the chromosomes, and the landscape of genomic imbalances in human carcinomas. *Environmental and molecular mutagenesis* **50,** 593–601 (2009).

141. Ried, T. & Rajapakse, I. The 4D Nucleome. *Methods* **123,** 1–2 (2017).

142. Ried, T., Schröck, E., Ning, Y. & Wienberg, J. Chromosome painting: a useful art. *Human molecular genetics* **7,** 1619–1626 (1998).

143. Ried, T. *et al.* Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes, chromosomes and cancer* **15,** 234–245 (1996).

144. Rohr, K. *et al.* Tracking and quantitative analysis of dynamic movements of cells and particles. *Cold Spring Harbor Protocols* **2010,** pdb–top80 (2010).

145. Roig, A. I. *et al.* Immortalized epithelial cells derived from human colon biopsies express stem cell markers and differentiate in vitro. *Gastroenterology* **138,** 1012–1021 (2010).

146. Ronquist, S., Meixner, W., Rajapakse, I. & Snyder, J. Insight into dynamic genome imaging: Canonical framework identification and high-throughput analysis. *Methods* **123,** 119–127 (2017).

147. Ronquist, S. *et al.* Algorithm for cellular reprogramming. *Proceedings of the National Academy of Sciences* **114,** 11832–11837 (2017).

148. Rudkin, G. T. & Stollar, B. D. High resolution detection of DNA–RNA hybrids in situ by indirect immunofluorescence. *Nature* **270,** 572 (1977).

149. Ruetz, T. *et al.* Constitutively Active SMAD2/3 Are Broad-Scope Potentiators of Transcription-Factor-Mediated Cellular Reprogramming. *Cell Stem Cell* **21,** 791–805 (2017).

150. Sainte-Marie, G. A paraffin embedding technique for studies employing immunofluorescence. *Journal of Histochemistry & Cytochemistry* **10,** 250–256 (1962).

151. Sammeta, N., Hardin, D. L. & McClintock, T. S. Uncx regulates proliferation of neural progenitor cells and neuronal survival in the olfactory epithelium. *Molecular and Cellular Neuroscience* **45,** 398–407 (2010).

152. Schiaffino, S., Rossi, A. C., Smerdu, V., Leinwand, L. A. & Reggiani, C. Developmental myosins: expression patterns and functional significance. *Skeletal muscle* **5,** 22 (2015).

153. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9,** 671–675 (2012).

154. Seaman, L. *et al.* Nucleome Analysis Reveals Structure–Function Relationships for Colon Cancer. *Molecular Cancer Research* **15,** 821–830 (2017).

155. Seth, A. *et al.* Prox1 ablation in hepatic progenitors causes defective hepatocyte specification and increases biliary cell commitment. *Development* **141,** 538–547 (2014).

156. Shachar, S., Pegoraro, G. & Misteli, T. HIPMap: A High-Throughput Imaging Method for Mapping Spatial Gene Positions. *Cold Spring Harbor symposia on quantitative biology* **LXXX,** 73–81 (2015).

157. Shannon, P. MotifDb: An annotated collection of protein-DNA binding sequence motifs. *R package version* **1** (2014).

158. Shirley, J. W., Ty, S., Takebayashi, S. I., Liu, X. & Gilbert, D. M. FISH finder: A high-throughput tool for analyzing FISH images. *Bioinformatics* **27,** 933–938 (2011).

159. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences* **89,** 8794–8797 (1992).

160. Smeets, D. *et al.* Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. *Epigenetics & chromatin* **7,** 8 (2014).

161. Spielman, D. A. & Teng, S. H. Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and Its Applications* **421,** 284–305 (2007).

162. Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I. & Dozmorov, M. G. HiCcompare: An R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics* **19,** 13–16 (2018).

163. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100,** 9440–9445 (2003).

164. Takahashi, K. & Yamanaka, S. Induced pluripotent stem cells in medicine and biology. *Reproductive Medicine and Biology* **12,** 39–46 (2013).

165. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126,** 663–676 (2006).

166. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131,** 861–872 (2007).

167. Takeuchi, H. *et al.* c-MET expression level in primary colon cancer: a predictor of tumor invasion and lymph node metastases. *Clinical Cancer Research* **9,** 1480–1488 (2003).

168. Tang, C., Ang, B. T. & Pervaiz, S. Cancer stem cell: target for anti-cancer therapy. *The FASEB Journal* **21,** 3777–3785 (2007).

169. Tapscott, S. J. *et al.* MyoD1: a nuclear phosphoprotein requiring a Myc homology region to convert fibroblasts to myoblasts. *Science* **242,** 405 (1988).

170. Teller, K. *et al.* A top-down analysis of Xa-and Xi-territories reveals differences of higher order structure at $>$ 20 Mb genomic length scales. *Nucleus* **2,** 465–477 (2011).

171. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications* **8,** 1–15 (2017).

172. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).

173. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).

174. Umemura, Y. *et al.* Transcriptional program of Kpna2/Importin-α2 regulates cellular differentiation-coupled circadian clock development in mammalian cells. *Proceedings of the National Academy of Sciences* **111,** 5039–5048 (2014).

175. Van Der Maaten, L. & Hinton, G. E. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

176. Van Steensel, B. & Belmont, A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169,** 780–791 (2017).

177. Vanden Heuvel, G. B. *et al.* Hepatomegaly in transgenic mice expressing the homeobox gene Cux-1. *Molecular carcinogenesis* **43,** 18–30 (2005).

178. VanDussen, K. L. & Samuelson, L. C. Mouse atonal homolog 1 directs intestinal progenitors to secretory cell rather than absorptive cell fate. *Developmental biology* **346,** 215–223 (2010).

179. Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **17,** 395–416 (2007).

180. Wang, H. *et al.* CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. *Cell,* 1–13 (2018).

181. Weaver, B. A. & Cleveland, D. W. Aneuploidy: instigator and inhibitor of tumorigenesis. *Cancer research* **67,** 10103–10105 (2007).

182. Webster, M., Witkin, K. L. & Cohen-Fix, O. Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *Journal of cell science* **122,** 1477–1486 (2009).

183. Weintraub, H. The MyoD family and myogenesis: Redundancy, networks, and thresholds. *Cell* **75,** 1241–1244 (1993).

184. Weintraub, H. *et al.* Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proceedings of the National Academy of Sciences* **86,** 5434–5438 (1989).

185. Wicha, M. S., Liu, S. & Dontu, G. Cancer stem cells: an old idea—a paradigm shift. *Cancer research* **66,** 1883–1890 (2006).

186. Xie, Y. *et al.* Phosphorylation of GATA-6 is required for vascular smooth muscle cell differentiation after mTORC1 inhibition. *Science signaling* **8,** 1–27 (2015).

187. Xiong, Q., Mukherjee, S. & Furey, T. S. GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. *Scientific reports* **4,** 6347 (2014).

188. Yan, K. K., Yardlmcl, G. G., Yan, C., Noble, W. S. & Gerstein, M. HiC-spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33,** 2199–2201 (2017).

189. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient. *Genome Research,* gr.220640.117 (2017).

190. Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences* **111,** 16219–16224 (2014).

191. Zhang, X. *et al.* A non-canonical E-box within the MyoD core enhancer is necessary for circadian expression in skeletal muscle. *Nucleic acids research* **40,** 3419–3430 (2012).

192. Zhu, Y. *et al. A SHARP/Xist complex regulates breast cancer stem cells* 2017.