4D Nucleome of Cancer

by

Laura A Seaman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2017

Doctoral Committee:

Assistant Professor Indika Rajapakse, Chair
Professor Brian D. Athey
Professor Daniel M. Burns, Jr.
Professor Alfred O. Hero III
Professor Thomas Ried, National Cancer Institute
Professor Max S. Wicha

Laura A Seaman

laseaman@umich.edu

ORCID iD: 0000-0001-9111-9776

ACKNOWLEDGEMENTS

Without the support and eternal optimism of my advisor, Indika Rajapakse, this work literally would not have been possible. Thank you to the rest of my committee members as well, Al Hero, Brian Athey, Dan Burns, Thomas Ried, and Max Wicha as well as former members Vijay Nair and Colin Duckett whose advice and enthusiasm kept me going. I would like to thank my lab mates for helping in so many ways including answering many, many questions: Haiming Chen, Scott Ronquist, Walter Meixner, Sijia Liu, Geoff Patterson, and Jie Chen. Thank you to all of my collaborators and research associates including Thomas Ried, Markus Brown, Darawalee Wangsa, Jordi Camps, Gilbert Omenn, John Snyder, Max Wicha, Yongyou Zhu, Ryan Mills, Rich McEachin, and Alexy Nesvizhskii, Brandon Govindarajoo, Tony Chun, Daysha Ferrer-Torres, Stephen Lindsly, Cyrus Najarian, Nicholas Comment, Teal Guidici, Shweta Ramadas, and anyone else I may have missed. Special shout out to Julia Eussen provided constant guidance and kept everything moving.

Thank you to all of my fellow graduate students for their support and at times much needed distractions. In particular, I would like to thank Teal Guidici and Shweta Ramadas for joining me during our (very roughly) weekly study meet-ups over the last two years that helped keep me sane. Thank you Taylor Pratt for all of your love and support over the last year. You have helped me stay both focused and distracted as I have needed it.

Finally, I would like to thank my family: Karen, Claude, Charlie, and Katie. You are all amazing. Mom, thank you for being my rock and constantly listening.

Dad, thank you for constantly supporting me to the point of putting my published papers in your office where no one understands them. Charlie, thank you for being an amazing big brother. I'm finally doing something before you! Katie, thanks for being the best sister ever. I wouldn't trade our year as roomies for anything and I'm so proud of you. Danielle, Allie, and Cara, thank you for your support as well. The steady supply of facetime chats, pictures, and videos never fail to make me smile.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

2D two dimensional

3D three dimensional

4D four dimensional

ANOVA analysis of variance

BAC Bacterial artificial chromosome

BAM binary sequence alignment map

bp base pairs

Bru-seq bromouridine sequencing

CGH comparative genomic hybridization

CRC colorectal cancer

CSC cancer stem cell

DAVID database for annotation, visualization, and integrated discovery

DE differentially expressed

EASE expression analysis systematic explorer

FC fold change

FDR false discovery rate

FISH fluorescent in situ hybridization

FPKM fragments per kilobase of transcript per million mapped reads

FPM fragments per million

GO gene ontology

Hi-C genome wide chromosome conformation capture

HMM hidden markov model

HSR homogeneously staining region

ICE iterative correction and eigenvector decomposition

kb kilobase

lncRNA long non-coding RNA

MAE monoallelic expression

Mb megabase

MSE mean squared error

NAT 4D Nucleome Analysis Toolbox

PC principle component

PCA principle component analysis

PSNR peak signal-to-noise ratio

RE restriction enzyme

RNA-seq deep sequencing of RNA transcripts

SAM sequence alignment map

SKY spectral karyotyping

SNP single nucleotide polymorphism

TAD topologically associating domains

TFBS Transcription factor binding site

WGS Whole genome sequencing

ABSTRACT

Chromosomal translocations and aneuploidy are hallmarks of cancer genomes; however, the impact of these aberrations on the nucleome (i.e., nuclear structure and gene expression) are not yet understood. This dissertation aims to understand the changes in nuclear structure and function that occur as a result of cancer, i.e., the 4D nucleome of cancer. Understanding of nuclear shape and organization and how it changes over time in both healthy cells as well as cancer cells is an area of exploration through the 4D nucleome project.

First, I explore healthy cells including periodic changes in nuclear shape as fibroblasts cells grow and divide. Shape and volume changed significantly over the time series including a periodic frequency consistent with the cell cycle. Next, combined analysis of genome wide chromosome conformation capture and RNA-sequencing data identified regions with different expression or interactions in cells grown in 2D or 3D cell culture. Next, I elucidate how chromosomal aberrations affect the nucleome of cancer cells. A high copy number region is studied, and we show that around sites of translocation, chromatin accessibility more directly reflects transcription. The methods developed, including a new copy number based normalization method, were released in the 4D nucleome analysis toolbox (NAT), a publicly available MATLAB toolbox allowing others to use the tools for assessment of the nucleome.

Finally, I describe continuing projects. By comparing cancer stem cells to non-stem cell like cancer cells, a bin on chromosome 8 was identified that includes two stem cell related transcription factors, *POU5F1B* and *MYC*. Then tools for evaluating allele

specific expression are developed and used to measure how allele specific structure and function varies through the cell cycle. This work creates a foundation for robust analysis of chromosome conformation and provides insight into the effect of nuclear organization in cancer.

CHAPTER I

Introduction

1.1 Research overview

Cancer is the second most common cause of death in the United States behind only heart disease [1]. 1.6 million cases of cancer will be diagnosed this year and almost $600,000$ people will die from cancer in the United States alone [2]. Research has shown that cancer is caused by the combination of a number of hallmarks, including genetic and epigenetic changes that cause unregulated cell proliferation [47]. The availability of high throughput sequencing has led to extensive characterization of the mutations that lead to this deregulation of cell proliferation.

The four dimensional (4D) nucleome is studied by integrating dynamical features of three dimensional (3D) architecture with the dynamical gene expression landscape and consequent changes in cellular differentiation and disease. One of the most powerful tools for studying the nucleome is genome wide chromosome conformation capture (Hi-C), which was originally described in 2009 [73]. Since that time, research has shown that the genome can be partitioned into active, i.e. euchromatic regions, and inactive, i.e. heterochromatic regions. Additionally, the genome can be further divided into megabase sized domains called topologically associating domains (TAD)s separated by small boundary regions [31]. These ideas have been used to explore dynamical changes in the nucleome that occur because of the cell cycle [16] and how to

control the genome and therefor the cell by understanding these processes [105].

Whole genome sequencing (WGS) has provided new a opportunity for detecting translocations and mutations, both of which provide insight into how cancers occur and potential therapeutic opportunities using personalized medicine. Similarly to how WGS advanced the understanding of genetic mutations underlying cancer, Hi-C provides an opportunity to learn about nuclear organization of cancer cells and its biological relevance. This dissertation aims to lay out some key findings and many tools for exploring the biological structure and function of cancer.

This chapter provides an overview of the work covered in this dissertation and a literature review of relevant work that has provided a foundation for this research.

Chapter II studies the unperturbed shape of the nucleus over a 75 hour time course in cell-cycle synchronized primary human fibroblasts. By modeling the nucleus as an ellipsoid, we derived simple time-varying shape properties that were fit to a range of frequencies to extract the primary oscillations. We found two peak frequencies one of which was consistent with the cell cycle. This work provides a statistical framework for analyzing populations of fixed cells and shows that a single sample in time provides an incomplete picture of nuclear shape.

Chapter III examines genome structure and gene expression of fibroblasts grown in two dimensional (2D) and 3D cell culture conditions. The combined analysis of Hi-C and deep sequencing of RNA transcripts (RNA-seq) datasets showed a large number of differentially expressed genes many of which are localized in genomic regions that displayed structural changes. We also find that gene expression of 3D cultured cells more closely resembles native tissue than 2D cultured cells for a set of skin-specific genes. This confirms previous observations that 3D cell culture more closely resembles native tissues. This work shows that nuclear structure and function depend on the cellular environment including cell culture conditions.

In Chapter IV, this analysis is extended from normal cellular populations to can-

cer. We explore the 4D nucleome of cancer by analyzing nuclear structure and function of HT-29, a human colorectal cancer cell line, grown in 2D and 3D culture for two different time points. A new copy number-based normalization method for Hi-C data was developed and used to determine that around sites of translocation, chromatin accessibility more directly reflects transcription. Additionally, a high copy number region containing the oncogene *MYC* is composed of open chromatin and interacts strongly with an amplified region containing the oncogene *STARD7*. The methods described can be used to asses the nucleome of any cell type regardless of karyotype. 4D Nucleome Analysis Toolbox (NAT), a MATLAB package was released containing tools for loading data, normalization, defining topological domains, exploring translocations, and analyzing time series datasets.

Chapter V covers results from ongoing projects focusing on cancer stem cells. We explore the structural and functional signature of cancer stem cell (CSC)s by comparing genetically identical cellular populations. A number of regions with CSC specific Hi-C structural interactions are identified including a region containing *MYC* and *POUF51B*, two stem cell related transcription factors. These results show that cellular subpopulations have unique structures. Learning more about these unique structures can help elucidate more effective ways to target difficult to kill CSCs that often allow cancer to evade treatments.

One of the limitations of most Hi-C analysis is that different copies of a region, including the two copies of each chromosome, cannot be distinguished. Chapter VI presents initial analysis of Hi-C, RNA-seq, and bromouridine sequencing (Bru-seq) of a genotyped cell line, allowing assignment of reads covering a single nucleotide polymorphism (SNP) to be assigned to the correct parent of origin. We compared the maternal and paternal expression within a cell and analyzed how they varied through the cell cycle. The methods developed provide the opportunity to identify gene expression and nascent transcription that are specific to a single allele.

1.2 Nuclear structure in cancer

All cancers have chromosomal aberrations. They can be structural (translocations, insertions, deletions, inversions) or numerical (aneuploidy) [47, 41]. These aberrations are a hallmark of cancer and change nuclear structure by disrupting the normal patterns of folding and organization. The aberrations cause cancers by activating tumor-promoting pathways or inactivating tumor-suppressing signaling pathways [47]. However, the interplay between chromosomal aberrations (structure) and gene expression (function) is not fully understood [48, 40, 80, 101].

One method for measuring chromosomal aberrations is spectral karyotyping (SKY), which uses chromosome specific probes to stain each chromosome with a different fluorophore [112]. SKY allows visual identification of large scale genetic changes and estimation of cellular heterogeneity by quantifying observed changes in multiple cells. To estimate where the chromosomal alterations occur, SKY relies on chromosome bands. Chromosome bands are naturally occurring patterns of genomic regions that lead to consistent coloring differences that can be used to estimate genomic location at low resolution [124, 63].

Techniques that rely on the genetic sequence to determine genetic location are much higher resolution. Array based comparative genomic hybridization (CGH) is an array based technique for estimating copy number. CGH hybridizes the total genomic DNA content of both test and reference cellular populations to an array then uses fluorescent detection of the relative abundance of the probes to estimate copy number [92]. With as many as $20,000$ loci per array, CGH gives a much more quantitative measure of the relative abundance of genomic loci as well as the boundaries of the amplified regions.

Another genomic technique for characterizing cancer is WGS which uses sequencing to determine genetic changes [82]. Historically, WGS has focused on characterization of mutations and small insertions and deletions ($\leq$ 100 base pairs (bp)) in

cancer samples. More recently, WGS has been used to detect larger alterations like translocations or copy number alterations with the use of very deep sequencing [19]. Copy number alterations can be detected by measuring the relative number of reads mapped to a region in test and control samples in essentially the same way fluorescent differences indicate copy number alternations in CGH [19]. Unlike CGH, WGS can detect balanced translocations if reads are sequenced that span the junction. One difficulty of WGS is that it requires large amounts of sequencing and as a result can be quite expensive, especially if the goal is to gather enough reads to detect translocations.

1.3 Genome wide chromosome conformation capture

The development of chromosome conformation capture techniques provide unprecedented insight into spatial chromatin organization and long-range chromatin interactions in the interphase nucleus [73]. Hi-C generates matrices that reflect chromatin interactions by using proximity-based ligation followed by sequence analysis as shown in Figure 1.1A [73]. More specifically, Hi-C requires crosslinking cells to fix the DNA, then cutting DNA with a restriction enzyme. After marking exposed ends with biotin, ligation is used to create chimeric molecules in which DNA that was in physical proximity is now linearly connected. The DNA is then sheared, biotin marked pieces are purified, and paired end sequencing is used to identify pieces of DNA that were in physical proximity.

These pairs are then compiled into matrices of predetermined resolution (often 100 kilobase (kb) or 1 megabase (Mb)) by adding up the number of reads that were sequenced between each pair of loci. The first step in making the matrices is to align the reads to a reference genome for which multiple tools are available [69, 71]. During alignment the paired end nature of the reads are ignored since paired end aligners assume a fixed insertion size and linear proximity of the sequences, which is

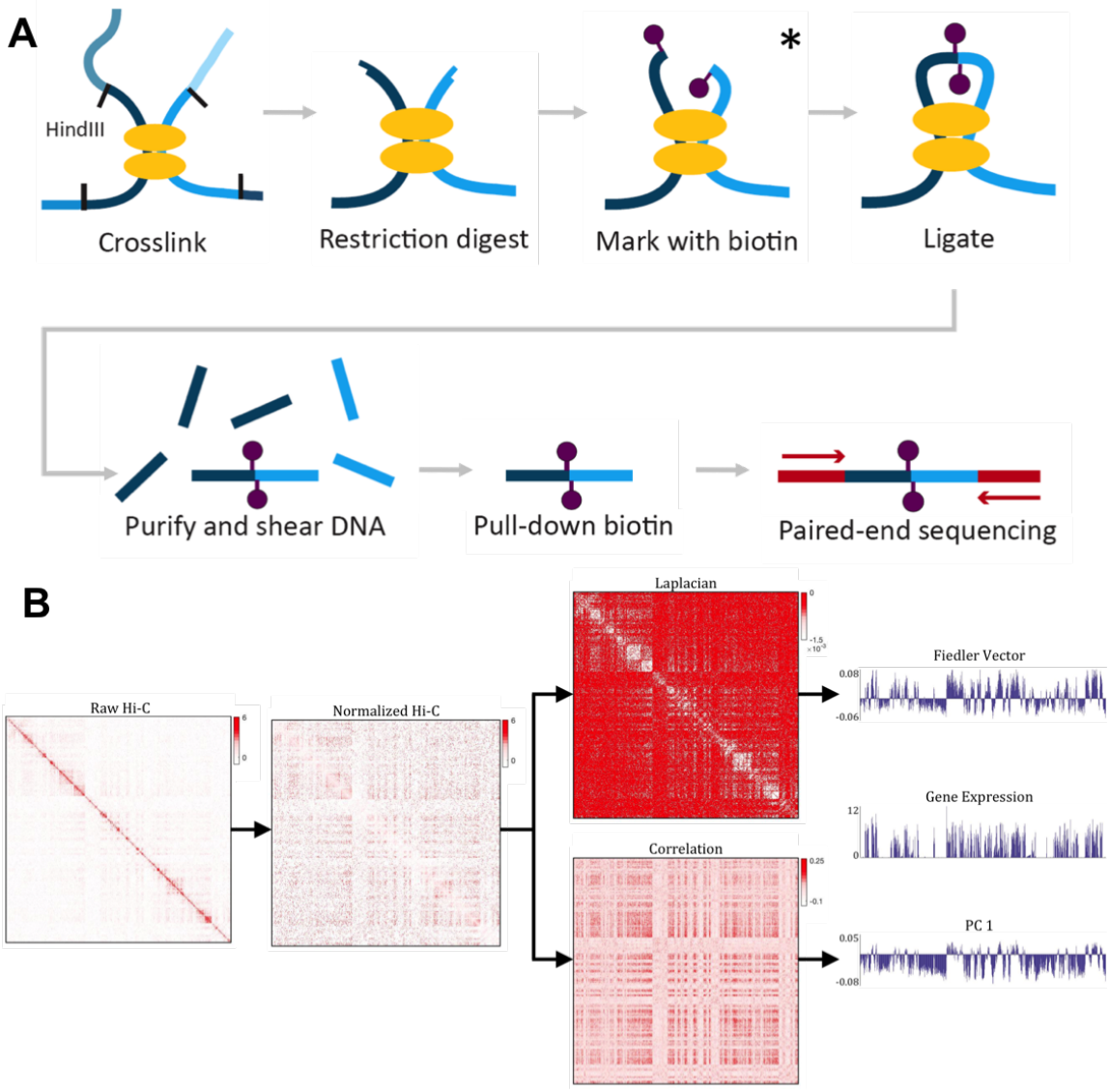

Figure 1.1: Hi-C experimental computational methods. A) Experimentally, Hi-C requires crosslinking DNA and digesting it with a restriction enzyme. The hanging ends are then labeled with biotin, then ligated at low concentration so that they form connections with other sequences that are nearby in 3D space. Next, the DNA is purified and sheared before biotin is pulled down and paired end sequencing is used to measure pairs of loci in physical proximity. B) After initial normalization, Hi-C can be compared to RNA-seq by calculating the Laplacian and from that the Fiedler vector (top) or by calculating the correlation matrix then the first principle component (bottom).

not present in Hi-C data. Next, Homer is used to match the separately aligned reads with their pair, organize them by chromosome and locus, and count the number of interactions in each pair of loci for the resolution selected [57]. The text files Homer creates can then be loaded into other software and downstream analysis can begin.

Because of the linear order of DNA and the limitations on folding that imposes, there are far more interactions observed between neighboring DNA sequences than there are between opposite ends of a single chromosome. This results in raw Hi-C matrices that are highly diagonally dominant [73]. Since the goal is to study interactions at a wide range of distances, as opposed to just short range interactions, this diagonal dominance needs to be adjusted. A number of methods have been develop to do so. One of the first methods developed was iterative correction and eigenvector decomposition (ICE) which assumes that all genomic regions should have equal visibility and uses an iterative method to correct the matrix creating a map of the relative probabilities between pairs of loci [55]. Another method, called Toeplitz normalization, assumes that the expected value for any two loci is monotonic with distance [16]. As a result, normalization is performed by dividing the diagonal and parallels of the diagonal by the average of the non-zero elements of the parallel. The assumptions used by both of these methods are violated by the chromosomal aberrations present in cancer cells and thus a newly developed normalization method is described in Chapter IV.

Chapter IV focuses on using Hi-C and RNA-seq to understand the cancer nucleome by exploring the relationship between its structure and function. Previous studies of cancer genomes using Hi-C showed long range interactions between known risk loci for the development of colorectal cancer (CRC) and regulatory regions [58], demonstrated proto-oncogene activation by disruption of chromosome neighborhoods [51], determined changes in inter-chromosomal interaction frequency in breast cancer [7], and showed that changes in genomic copy number subdivide the domain structure

of chromosomes [119]. In Chapter IV, I extended this work through a comprehensive analysis of the CRC cell line HT-29 including analyzing how chromosomal aberrations affect the nucleome by integrating Hi-C and RNA-seq analyses.

1.4 Comparing nuclear structure and function

To measure how nuclear structure and organization affect the cell, Hi-C data must be compared to a measure of cellular function. Often, gene expression, measured by RNA-seq, is used for this purpose. RNA-seq uses sequencing to measure what transcripts are present in a cell and estimate the abundances [122, 127]. Since transcription is the most direct output of DNA, comparing Hi-C to RNA-seq provides a method for determining the biological relevance of changes in structure. To compare the two data types, they must be converted to represent the same genomic units. Traditionally, expression is measured for each gene while structure is measured for bins of a fixed size. One common approach to comparing the two data sets is to summarize the expression within each bin. This can be done by adding up the expression of all of the genes in a bin, and when a large gene spans multiple bins, dividing its output among the bins proportionally to the amount of the gene in each bin [16].

Once Hi-C and RNA-seq have been calculated for the same genomic units, correlation can be used to compare them. In order to compare the structure measured by the Hi-C matrix (two dimensional), to DNaseI hypersensitivity or gene expression (one-dimensional), Hi-C data is converted to a vector. In the original Hi-C publication, these data types were compared by first calculating the correlation matrix of the normalized Hi-C data, which describes the correlation between each pair of genomic regions. Eigendecomposition was then used to extract the first principal component, which identifies the vector that best approximates the matrix. The correlation between this vector and the gene expression vector is a measure of the strength of the

structure-function relationship. Additionally, regions within the first principle component that have the same sign (positive or negative values) were defined as A and B compartments. These compartments divide the genome and correlate with the presence of open or closed chromatin as measured by DNaseI hypersensitivity and active or repressed gene expression, respectively [73]. Another method for relating nuclear structure is to use the Laplacian framework as described below.

1.5 The Laplacian framework

The key mathematical operator in our analysis of Hi-C data is the graph Laplacian. The goal is to uncover partitioning within Hi-C matrices and to correlate that with function. The Laplacian represents diffusion or consensus among a discrete number of entities and has been used when discrete entities reach a consensus without direction [88]. Examples of this includes movements of groups of animals including flocking birds and emergence of language in primitive civilizations [26].

In the case of Hi-C data, the normalized Hi-C matrix can be interpreted as an adjacency matrix, $\mathbf{A}$, in which the nodes represent genomic segments and the edges are weighted by the number of interactions seen between them. Within the adjacency matrix,

$$(\mathbf{A})_{i,j} = w(n_i, n_j) \tag{1.1}$$

where the weight function w, must be symmetric, i.e. $w(n_i, n_j) = w(n_j, n_i)$, and non-negative, $w(n_i, n_j) \geq 0$. The Laplacian is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{1.2}$$

,

where the degree matrix is the sum of the weights surrounding each node, i.e. $\mathbf{D} =$

$\text{diag}(d_1, d_2, ..., d_k)$ and $d_1 = \Sigma_{j=i}^{k} a_{i,j}$. The normalized Laplacian is

$$\bar{\mathbf{L}} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-1/2} \tag{1.3}$$

,

and has the added benefit of being scale invariant. The second smallest eigenvalue of this matrix is called the Fiedler number and the corresponding eigenvector is the Fiedler vector.

The sign of the Fiedler vector partitions the graph into two components, so that the cost of the cuts required to separate the two components is minimized [22]. In the case of Hi-C matrices, this partitioning defines A/B compartments similarly to how the first principle component of the correlation matrix can also be used to define A/B compartments [16]. In fact, the Fiedler vector is very similar to the first principle component of the correlation matrix and both are strongly correlated with gene expression. The parallels between the two methods are shown in Figure 1.1B. In both cases the normalized Hi-C matrix is used to calculate the an additional matrix that captures the connectivity within the graph: the correlation matrix and the Laplacian matrix for the Lieberman-Aiden et. al. and Chen et. al. methods respectively [73, 16]. Then these matrices are summarized with a single vector that captures A/B compartments. Both the principle component (PC) and Fiedler vector relate strongly to function as measured with RNA-seq as shown in Figure 1.1B ($r = 0.67$ and 0.64 for the PC and Fiedler vector, respectively).

One benefit of the Laplacian framework is that in addition to using the Fiedler vector to partition the genome, the Fiedler number can be used to understand the connectivity and compare regions or samples. The Fiedler number is maximized in a fully connected graph and falls to zero in an unconnected graph [22]. In the context of understanding Hi-C data, the Fiedler number reflects the underlying stability of the topology of the genomic region for which it is calculated, at any scale.

In addition to Fiedler number, to understand genomic stability the entropy of a Hi-C matrix can be calculated. Traditionally, Shannon entropy is defined as

$$-\Sigma p_i \log p_i \tag{1.4}$$

,

where p_i are probabilities [91]. However, this has been extended to matrices through Von Neumann entropy or

$$-\Sigma \lambda_i \log \lambda_i \tag{1.5}$$

,

where λ_i are the eigenvalues of a matrix [91]. Entropy measures the amount of disorder in a system so the higher the Von Neumann entropy, the more disordered the matrix. In the context of understanding the 4D nucleome, high Von Neumann Entropy reflects large amounts of disorder in the system indicating that the structure is unstable.

1.6 Detecting topologically associating domains

TADs are linear chromatin domains within the genome that show increased interactions within the domain and decreased interactions with neighboring domains [31]. TADs vary in size from approximately 200 kb up to a few Mb and have been shown to be consistent across cell types [18]. Several methods have been developed to define TADs in Hi-C data. The first definition used a directionality index based on the χ^2 statistic to quantify the proportion of a loci's interactions that are either upstream or downstream of the region [31]. An hidden markov model (HMM) was used to identify genomic regions whose directionality index indicated groups of loci that interact with each other. These regions can be separated by small boundary regions that have few

interactions on either side and within which the directionality index switches sign. The model was used to predict the states underlying the genome as TAD, boundary, or neither and identified $2,200$ TADs in the mouse genome that averaged 880 kb in size and covered approximately 91% of the mouse genome [31]. One benefit of this method is that it allows prediction of boundaries of different sizes as well as unorganized regions. However, HMMs are computationally inefficient compared to newer methods.

Computationally efficient methods include a community detection based algorithm and an iterative Laplacian based method. The community detection based method formulates the identification as a maximization problem by maximizing the total reads within a domain using a scaled density of the subgraph between two potential TAD boundaries [36]. This methods uses a dynamic programming algorithm to solve the problem with a given scale parameter [36]. The community detection method also has the benefit that the size of the domains can be tuned which helps identify the hierarchical domain structure of the genome by comparing domains found with different scale parameters.

Finally, the iterative Laplacian algorithm works by using the Laplacian framework to identify domain structures. The iterative Laplacian algorithm assumes the genome can be understood as a network whose adjacency matrix is the normalized Hi-C matrix [18]. The Laplacian algorithm starts by initializing TADs as regions the Fiedler vector that have a continuous sign, i.e. contiguous regions of A/B compartments [18]. The regions are then subdivided to maximize the connectivity of each domain. The algorithm subdivides regions using the sign of the Fiedler vector of each region to define increasingly small domains until the Fiedler number of each domain is greater than a tunable threshold [18]. This method has the benefit of being computational efficient since it relies on calculating a single eigenvector and eigenvalue for each domain for which efficient algorithms have been developed. Additionally, as with

the community detection algorithm, by changing a tuning parameter, in this case selecting a smaller or larger Fiedler number threshold, domains of a variety of sizes can by explored.

1.7 Changes in nuclear shape

The shape of the nucleus is tightly regulated and changes as a cell differentiates, generally starting out spherical and ending more oblong. Abnormal shape and size are linked with a number of diseases such as cancer and progeria [59]. Misshapen or lobulated nuclei are used to identify cancerous tissue and estimate cancer grade [21]. Lamin gene mutations, called laminopathies, lead to misshapen nuclei and cause muscular dystrophy or premature aging in the case of progeria by disrupting the structural network lamin forms around the nuclear periphery [27].

In addition to changes in shape related over time due to cell cycle or other factors, cell culture itself causes massive morphological changes in nuclear shape. Growing mammalian cells *in vitro* is an indispensable technique for cell biology and biomedical research. Conventionally, human cells have been derived to grow in defined medium either in suspension or as an adherent monolayer. Adherent monolayer (2D) cell cultures do not resemble the natural 3D structures of body tissues, and as a result cells grown in 2D may have considerable discordances in cellular morphology, physiology, pathology, cell-cell interaction and communication compared with natural tissues.

Increasing evidence shows that *in vitro* 3D culture captures natural tissue complexity better than 2D cultures [46, 3, 20, 108]. Advances in 3D culture techniques open new avenues for *in vitro* modeling of human organ development, tissue morphogenesis, pathogenesis of diseases, cellular response to drugs or other perturbations, and screening for novel therapeutics [108, 23]. Human cell-based 3D models in pharmaceutical research can complement animal models, which often fail to predict the efficacy and toxicities of new drugs. 3D human-cell models may also provide

more effective and economical screening of new drugs than the use of animal models [117, 128]. Furthermore, *in vitro* 3D modeling of native tissue provides tools for regenerative medicine. However, understanding the fundamental cell biology is critical in translating *in vitro* discoveries into clinical applications, e.g., functional replacement of damaged tissue.

Tissue-specific gene expression is the molecular basis of cellular function. It is not fully established how closely *in vitro* 3D tissue culture mimics native tissue. We hypothesize that the interplay between genome structure and function, i.e., the nucleome, is the key component of tissue-specific gene expression. Chapter III studies how the nucleome changes between 3D- and 2D- grown cells. We previously observed chromosome conformation changes between human fibroblasts grown as spheroids vs. monolayer cultures [17].

CHAPTER II

Periodicity of nuclear morphology in human fibroblasts

2.1 Abstract

Motivation: Morphology of the cell nucleus has been used as a key indicator of disease state and prognosis, but typically without quantitative rigor. It is also not well understood how nuclear morphology varies with time across different genetic backgrounds in healthy cells. To help answer these questions we measured the size and shape of nuclei in cell-cycle-synchronized primary human fibroblasts from six different individuals at 32 time points over a 75 hour period.

Results: The nucleus was modeled as an ellipsoid and its dynamics analyzed. Shape and volume changed significantly over this time. Two prominent frequencies were found in the six individuals: a 17 hour period consistent with the cell cycle and a 26 hour period. Our findings suggest that the shape of the nucleus changes over time and thus any time-invariant shape property may provide a misleading characterization of cellular populations at different phases of the cell cycle. The proposed methodology provides a general method to analyze morphological change using multiple time points even for non-live-cell experiments.

2.2 Introduction

Whether form follows function or function follows form is an ongoing debate in biology. Nuclear shape is known to play a role in mechanotransduction, in which cells convert physical forces into chemical signals through connections between the cytoskeleton, nuclear envelope and lamina [27]. Recent results by Rangamani et. al. suggest that changes in cell shape induce local gradients of receptors or signaling molecules which amplify signals during differentiation [97]. Nuclear shape may play a similar signaling role in transcription and other cellular processes. To confirm and understand such phenomena, the first step is studying how nuclear morphology changes over time.

The shape of the nucleus is tightly regulated and changes as a cell differentiates, generally starting out spherical and ending more oblong. Abnormal shape and size are linked with a number of diseases such as cancer and progeria [59]. Misshapen or lobulated nuclei are used to identify cancerous tissue and estimate cancer grade [21]. Lamin gene mutations, called laminopathies, lead to misshapen nuclei and cause muscular dystrophy or premature aging in the case of progeria by disrupting the structural network they form around the nuclear periphery [27].

Several studies have compared nuclear shape under different conditions, such as diseased versus healthy or differentiated versus undifferentiated [35, 44]. None have examined how nuclear morphology normally varies over time within a cellular state (e.g. healthy fibroblasts). Our paper considers primary human fibroblasts synchronized to start at G1 in the cell cycle. We created a program which fits ellipsoids to data from confocal image stacks, then analyzed the resulting shape properties to test the statistical significance of their time variation and extract periodic behavior.

The cell cycle is an obvious explanation for changes in nuclear shape: as the cell grows leading to replication and then cell division, the volume of the nucleus might be expected to increase and then decrease. It is also known that the cell rounds

during mitosis as the spindle poles form, allowing chromosomes to line up for division [66]. For primary human fibroblasts, the cell cycle lasts between 16 and 28 hours with a mean of 20 hours [121]. Recent exploration of transcription factories and the hypothesis that genes physically move into and out of these regions as their expression levels change suggests that cyclically-expressed genes might be an additional cause for changes in nuclear shape [99]. One such set are clock genes controlling circadian rhythm for which a 24 hour or longer cycle would be expected [116]. Another set are genes related to ultradian rhythms for which we would expect to shorter periods often of $8 - 10$ hours [110].

In this study, we probed the unperturbed shape of the nucleus over 75 hours in cell-cycle synchronized primary human fibroblasts from six different individuals. Fibroblasts were chosen because of their applications to cellular reprogramming, wound healing, and ease of access [28, 131, 77]. Nuclei were stained with DAPI at 32 time points and then captured with 3D confocal microscopy. By modeling the nucleus as an ellipsoid, we derived simple time-varying shape properties including volume and eccentricity, i.e., roundness or flatness. This is an extension from most current methods that only calculate volume without other shape parameters. The resulting data was then fit to a range of frequencies to extract its primary oscillations. We found two peak frequencies one of which was consistent with the cell cycle.

2.3 Algorithms

2.3.1 Ellipsoidal modeling

We built a general analysis tool which fits ellipsoids to 3D volumetric data. Although the tool can handle multiple ellipsoids, in this case each volumetric dataset is cropped to contain just a single nucleus, to which a single ellipsoid is fit. As well as allowing visualization in 3D of a nuclei and its fitted ellipse, shown in Figure 2.1,

the tool's outputs are the lengths and directions of the ellipsoid's three primary axis, from which other properties such as volume and eccentricity are easily calculated. To reduce background noise, we clamped all pixels with an intensity less than 1% of the maximum to 0. The set of images from each nucleus was considered as a three-dimensional volumetric distribution, denoted $P(p)$, with a fluorescent intensity associated with each 3D position p. Point $p_i = (x_i, y_i, z_i)^T$ represents a sample, i.e., pixel in a single z-slice image, dot in right side of Figure 2.1B) in the volumetric grid with corresponding intensity w_i.

Our method is based on the data's first- and second-order moments (mean and covariance). It is similar to alternatives such as Gaussian mixture models, but models intensity using a quadratic rather than exponential function. In brief, our method first computes the quadratic distance function best representing the falloff in measured intensities over the nuclear volume, and then optimally thresholds this distance to yield an approximating 3D ellipsoid.

Squared distance at an arbitrary point p is defined as

$$D^2(p, Q) = (p - o)^T Q^{-1} (p - o) \tag{2.1}$$

where o is the distance origin and Q is a symmetric, positive definite 3×3 matrix. Expressing Q in an eigen-decomposition yields

$$Q = R_Q \begin{bmatrix} a_Q^2 & 0 & 0 \\ 0 & b_Q^2 & 0 \\ 0 & 0 & c_Q^2 \end{bmatrix} R_Q^T \tag{2.2}$$

where a_Q^2, b_Q^2, and c_Q^2 are the eigenvalues (representing axis scale factors), and R_Q is a 3×3 rotation matrix, representing axis directions. $\mathbf{Q}$ (in boldface) denotes the set of all parameters determining the anisotropic distance metric: $\mathbf{Q} = \langle \mathbf{o}, \mathbf{Q} \rangle = \langle \mathbf{o}, \mathbf{R_Q}, \mathbf{a_Q}, \mathbf{b_Q}, \mathbf{c_Q} \rangle$.

Unconstrained minimization of equation 2.1 causes a_Q, b_Q, and c_Q to increase without bound. We therefore constrain the anisotropy so that the sum of axis scale factors raised to some power equals the dimensionality:

$$a_Q^\gamma + b_Q^\gamma + c_Q^\gamma = 3. \tag{2.3}$$

Note that the identity transformation ($Q = I$) satisfies this constraint. As $\gamma \to 0$, anisotropy is unconstrained; highly eccentric shapes like needles or pancakes are freely permitted. As $\gamma \to \infty$, the ellipsoid is forced to be completely spherical. We used the normalization power $\gamma = 1$ to balance between these extremes for robust model fitting.

We then seek the $\mathbf{Q}$ yielding minimal sum of intensity-weighted squared distances over P

$$\mathbf{Q}_*(\mathbf{P}) = \arg\min_{\mathbf{Q}} \left(\sum_{\mathbf{p_i} \in \mathbf{P}} \mathbf{w_i}\, \mathbf{D^2}(\mathbf{p_i}, \mathbf{Q}) \right) \tag{2.4}$$

subject to the constraint in equation 2.3. It can be calculated in terms of the volumetric datasets mean vector, $\bar{p}(P)$, a weighted average of the dataset, and covariance matrix

$$C(P) = \frac{\sum_{p_i \in P} w_i \, (p_i - \bar{p}) \otimes (p_i - \bar{p})}{\sum_{p_i \in P} w_i}, \tag{2.5}$$

See A.1 for the detailed derivation of optimal ellipsoid, which we summarize in the following.

The optimal origin in $\mathbf{Q}_*$ is given by the mean of the dataset: $o_* = \bar{p}(P)$. Let the covariance matrix, $C(P)$, be decomposed into its eigenvectors and eigenvalues via

$$C(P) = R_C \begin{bmatrix} a_C^2 & 0 & 0 \\ 0 & b_C^2 & 0 \\ 0 & 0 & c_C^2 \end{bmatrix} R_C^T \tag{2.6}$$

The method of Lagrange multipliers can then be used to show that the optimal $\mathbf{Q}_*$ has rotation identical to the covariance's eigen-rotation; that is, $R_* = R_C$. The optimal scale factors a_*^2, b_*^2, c_*^2 are proportional to the exponentiated eigenvalues of $C(P)$ via

$$\begin{aligned} a_*^2 &= \alpha\, a_C^{4/(\gamma+2)}, \\ b_*^2 &= \alpha\, b_C^{4/(\gamma+2)}, \\ c_*^2 &= \alpha\, c_C^{4/(\gamma+2)}, \end{aligned} \tag{2.7}$$

where

$$\alpha = \left(\frac{3}{a_C^{2\gamma/(\gamma+2)} + b_C^{2\gamma/(\gamma+2)} + c_C^{2\gamma/(\gamma+2)}} \right)^{2/\gamma} \tag{2.8}$$

and, as mentioned earlier, we fix $\gamma = 1$. An approximating ellipsoid can then be computed from this anisotropic distance metric by thresholding squared distance via $D^2(p, \mathbf{Q}_*) \leq \mathbf{d^2}$, for some appropriate threshold d. It is computed so that an arbitrarily-scaled version of the binary function

$$B(p, \mathbf{Q}_*, \mathbf{d}) = \begin{cases} 1 & : \; D^2(p, \mathbf{Q}_*) \leq \mathbf{d^2}, \\ 0 & : \; \text{otherwise}, \end{cases} \tag{2.9}$$

has least squared error compared to the actual volume of intensities w_i. More precisely,

$$d_* = \arg\min_d \sum_{p_i \in P} \left(\tau\, B(p_i, \mathbf{Q}_*, \mathbf{d}) - \mathbf{w_i} \right)^2, \tag{2.10}$$

where the optimal scale factor τ is given by

$$\tau = \frac{\sum_{p_i \in P} B(p_i, \mathbf{Q}_*, \mathbf{d})\, \mathbf{w_i}}{\sum_{p_i \in P} B(p_i, \mathbf{Q}_*, \mathbf{d})}. \tag{2.11}$$

Note that we can remove the square in the denominator above usually present in least-squares projection because B is a binary function, so $B^2 = B$. Finally, the lengths of the approximating ellipsoid's three axes are given by

$$\begin{aligned} a &= d_* \, a_*^{-1}, \\ b &= d_* \, b_*^{-1}, \\ c &= d_* \, c_*^{-1}. \end{aligned} \tag{2.12}$$

Figure 2.1 shows an example.

2.3.2 Period estimation

We examined six measures of nuclear shape including volume from ellipsoid fitting, $V = \frac{4\pi}{3} a\, b\, c$, volume by direct counting of voxels whose intensity exceeds a threshold, and eccentricity, $\epsilon = \sqrt{1 - c^2/a^2}$, where c is the shortest and a the longest ellipsoid axis. Eccentricity reflects the roundness or flatness of a shape; a sphere has an eccentricity of 0 while a needle or pancake shape has an eccentricity close to 1. Finally, we measured the three axis lengths themselves, yielding a total of six different shape properties. Sampling each property over time for one of the six individuals yields a time series f_i, $i = 1, 2, \ldots, n$, sampled at $n = 32$ different time points denoted t_i. These times were not sampled uniformly over the 75 hours (see Table 2.1), complicating spectral analysis. The non-uniform sampling and limited number of time points made Fourier analysis ineffective. Each time series value, f_i, was calculated by averaging over the 20 cell nuclei sampled per time point.

To extract prominent frequencies in these time series, we fit them to a single-frequency pair of harmonic basis functions using least squares, considering frequency as a continuous parameter. We first centered the data by subtracting the mean of each time series, giving $\tilde{f}_i$, $i = 1, 2, \ldots, n$. Given a frequency ω, we minimized the mean squared error (MSE) between the resulting data, $\tilde{f}_i$, and the basis functions, $\alpha \sin \omega t_i + \beta \cos \omega t_i$, by solving a 2×2 linear system in the coefficients α and β:

$$\begin{aligned} \alpha \sum_{i=1}^{n} \sin^2 \omega t_i + \beta \sum_{i=1}^{n} \sin \omega t_i \cos \omega t_i &= \sum_{i=1}^{n} \tilde{f}_i \sin \omega t_i, \\ \alpha \sum_{i=1}^{n} \sin \omega t_i \cos \omega t_i + \beta \sum_{i=1}^{n} \cos^2 \omega t_i &= \sum_{i=1}^{n} \tilde{f}_i \cos \omega t_i. \end{aligned} \tag{2.13}$$

In addition to looking at prominent frequencies, the phase of the basis function, $\theta = \tan^{-1}(\alpha/\beta)$, reflects the relative timing.

The spectrum was represented by the amount of squared energy accounted for by this fit; that is, the difference between the squared signal energy in the original time series data and the squared residual (unfit) energy, given by

$$F(\omega) = \sum_{i=1}^{n} \tilde{f}_i^2 - (\tilde{f}_i - \alpha \sin \omega t_i - \beta \cos \omega t_i)^2. \tag{2.14}$$

We then swept the basis function frequency ω and looked for peaks in $F(\omega)$, indicating a relatively good fit at that frequency. To reduce noise and identify frequencies prevalent across individuals, we also plotted the average fit power, F, over all six individuals.

Our method is similar to the Lomb periodogram [76, 93], a standard procedure for analyzing periodicities in an irregularly-sampled time series. It is a simple extension of the related least-squares spectral analysis [78] which unlike Lomb analysis keeps the phase information of the basis functions. Our method differs by directly evaluating

the squared signal energy represented by the basis rather than its squared coefficients, $\alpha^2(\omega) + \beta^2(\omega)$. [1]

2.4 Methods

2.4.1 Sample preparation

Human primary fibroblasts from 6 normal male newborns (discarded foreskin tissue, passage 3) were cultured in complete media: MEM medium (Life Technologies, $10370 - 088$) supplemented with 10% fetal bovine serum (VWR, SH30071.03), 2 mM L-glutamine (Life Technologies, 25030081), and 1× Antibiotic-Antimycotic (Life Technologies, $15240 - 062$), at 37°C with 5% CO_2. On the day before the experiment, cells were trypsinized, 5×105 of them re-suspended in 15 ml complete media as described above, and seeded into T75 flasks [67]. Inspecting cells under the microscope 24 hours later, confluency was found to range from $30 - 50\%$. (We wanted to avoid 100% confluency with its likely inhibition of cell division.)

Cells were washed with 15 ml pre-warmed PBS twice, and serum-free MEM medium (2 mM L-glutamine, and 1× Antibiotic-Antimycotic) was added to each flask to begin cell synchronization. Cells were incubated at 37°C with 5% CO_2. After 24 hrs, each flask of cells were re-suspended with 2 mls 0.25% trypsin-EDTA, followed by 10 mls of complete media to inactivate the trypsin treatment, centrifuged at 750 rpm for 5 minutes, and each cell set re-suspended in 12 mls of serum-free media. Cell counts were performed, and 150μls (∼30K cells) placed on Fisher superfrost slides in petri dishes. Cells were allowed to settle for 2 hrs at 37°C with 5% CO_2. 15 mls of complete media were then added to each petri dish, sample slides taken at the time shown in Table 2.1, rinsed briefly in PBS, fixed in 4% paraformaldehyde for 8 minutes,

[1] Note that the two basis functions are not in general orthogonal over the irregularly-sampled time series: $\sum_{i=1}^{n} \sin \omega t_i \cos \omega t_i \neq 0$. Thus the energy fit by each of the two basis functions should not be considered to be independent, as is implicit in the measure $\alpha^2(\omega) + \beta^2(\omega)$.

and rinsed 3×5 minutes in PBS. 15μl of Prolong Gold (p36941 Life Technologies) with DAPI was placed on each slide, an 18×18 mm coverslip applied, sealed, and stored at -20°C until imaging. The sample from individual 3 at time point 5 was unusable due to a lack of cells. All imaging was completed on a Zeiss LSM 710 Microscope with a 63× Oil DIC objective, 0.2μm x and y resolution, 0.5μm z resolution, an oversampled pixel size of 0.132 μm x and y and 0.320 μm z, and 24 μm pinhole. Excitation was by a 405 nm laser with an emission collection band from 411 to 486 nms.

A volumetric dataset for each individual nucleus was formed by cropping a z-stack of images from confocal microscopic measurements. Ellipsoidal approximation (described in the Algorithms section) was performed on 20 nuclei from each individual at each time point.

2.4.2 Volume verification by thresholding

We used MATLAB to calculate the volume of nuclei by thresholding to validate our ellipsoid fitting. The 3D images were loaded into MATLAB, and thresholded with a cutoff of 5% of the maximum. Any holes that could not be reached from the outside of the image were filled to prevent nucleoli or other internal structures from being missed. The volume of all selected pixels was then integrated over all images in the z-stack. Images from 20 nuclei for each individual and time point were analyzed as was done for the ellipsoid fitting. To verify that the choice of threshold did not determine the periodic results, we tried multiple thresholds $(0.04, 0.045, 0.05, 0.055, 0.06, 0.065)$ and calculated spectrums for each.

2.4.3 Bootstrapping

To see how consistent our periodicity results were, we took 100 random samples with replacement (bootstrapping) of half the data (10 out of 20 nuclei from each time point and individual) and used this limited data set to rerun the periodicity

calculations described in 2.3.2.

2.4.4 Variance analysis over time

We used analysis of variance (ANOVA) to show that changes over time in volume or eccentricity cannot adequately be explained by sampling from a single distribution. Each time point was considered a category with 20 observations (nuclei) and the test was run separately on each of the six individuals. A p-value of .05 was used to test the null hypothesis that all of the variability seen between time points was due to random chance and that the eccentricity and volume were each drawn from a single distribution. The alternate hypothesis was that the distributions changed over time. Bonferroni correction was used to account for multiple tests: 6 individuals $\times 2$ properties [9].

2.5 Results and Discussion

2.5.1 Ellipsoid model captures variability in nuclear shape

Nuclear shape and volume of primary human fibroblasts that had been cell-cycle synchronized by two-days serum starvation was analyzed. Nuclear shape and volume were analyzed at 32 time points over 75 hours (sampling regime in Table 1) for cells from six different individuals. For each time point and individual, confocal microscopy was used to get 3D volumetric distributions of 20 nuclei stained by DAPI. An analysis and visualization tool performed the ellipsoidal approximation for each nucleus and calculated the three lengths of the ellipsoid axes as shown in Figure 2.1. [2]

The lengths of the three axes, a, b, and c, were then used to calculate each nuclei's volume, $V = \frac{4\pi}{3} a\, b\, c$, and eccentricity, $\epsilon = \sqrt{1 - c^2/a^2}$. We also calculated volume

[2]This tool was originally developed to spatially approximate homologous chromosome territories in the nucleus, where a pair of such territories was expected to be present simultaneously in the volumetric data. We applied the same tool to approximate an entire nucleus as a single ellipsoid.

independently by counting non-zero voxels after thresholding the images in MATLAB, yielding a total of six different shape indicators.

In Figure 2.2A, xy and xz projections of the time course for each individual and the average of the individuals show fluctuation in the nuclear shape over time. The ellipsoid volume, threshold volume and ellipsoid eccentricity are also shown over the time course in Figure 2.2B, 2.2C, and 2.2D respectively.

2.6 Nuclear shape changes over time

To show that the fluctuations in nuclear size and shape, seen in Figure 2.2, can not be explained by random chance, we performed ANOVA on the data for three shape properties: ellipsoidal volume, thresholded volume, and eccentricity. ANOVA tests whether data from different categories, in this case time points, can be explained by a single distribution or whether it requires different per-category distributions. Box plots of the distributions for each individual are shown in Supplemental Figures S1, S2, and S3 respectively. The null hypothesis was that a single distribution explains all of the variability. We were able to reject the null hypothesis at a $p \leq 0.05$ level for all three shape properties (eccentricity, ellipsoidal volume, and thresholded volume) and for all individuals.

2.6.1 Periodicity of the nuclear shape matches cell cycle and circadian rhythm timing.

By fitting a set of single-frequency basis functions to mean-centered ellipsoid volume, threshold volume, eccentricity, and three axis lengths, as described in 2.3.2, we computed the extent to which each was fit by a range of different frequencies. Figure 2.3B shows how well each parameter was fit by the basis functions for each individual. Results varied significantly over individuals, but a few frequencies were seen consistently across the 6 individuals and 6 nuclear shape indicators. The most prominent

peak, marked with red dots on Figure 2.3A, has a mean of 17.3 hours (min 16.5, max 18.3) and is the highest peak in the spectrums for four of the six measures. For the two measures of shortest axis length and eccentricity, it was the second highest peak.

Best fit basis functions for each shape property are shown in Figure 2.4 A-F. Supplemental Table S1 reports the basis function parameters we calculated, along with normalized MSE (MSE divided by the mean squared energy of the original signal), and peak signal-to-noise ratio (PSNR). The fit for eccentricity, shown in Figure 2.4B, has a 17 hour period and matches the data fairly well, yielding normalized MSE of 0.0045 and PSNR of 12.7. Such a low value for normalized MSE and high value for PSNR indicates that the fit captures much of the variability in the time series, and supports the hypothesis that this prominent peak reflects the cell cycle.

Cells become rounder during mitosis so we expect a dip in eccentricity as mitosis begins and an increase after the cells finish dividing [13]. In addition, nuclear volume is expected to increase through the cell cycle leading up to division. We looked at the phase of the top fits (all those included in Supplemental Tables 2.1 and 2.2) that had periods according with the cell cycle ($15 - 22$ hours) and found that the eccentricity had an average phase of 0.6645 rad, meaning it peaked 1.5 hours after serum was returned to the cells (and every $\simeq$ 17 hours after that). The average phase of the volume was -0.1143 rad, meaning it peaked about 10 hours after serum was re-added to the cells. Both observations are consistent with an initial stalling of the cells in the G0/G1 phase due to a lack of serum, followed by later attainment of maximal volume as the cells prepare to divide.

Across individuals and shape features, we also observed a second period at roughly 26 hours (min 23.5, max 28.9). This peak was weaker than the 17.3 hour cycle in all shape measures except for the shortest axis length. It could be a result of the circadian rhythm that controls humans' internal clock and sleep schedule. Although not as prevalent as the cell cycle period across all individuals and shape features,

it yielded fits with a median normalized MSE of 0.0041 and PSNR of 15.5 for the individuals and shape features for which it was one of the top two peaks, indicating an even better fit there than the first (cell cycle) period.

The spectrum for thresholded volume, based on a threshold of 5% of the dataset's maximum intensity, is mostly consistent with the other ellipsoid-based shape parameters and includes peaks at both of the above frequencies. However, the spectrum also includes a peak (in fact its tallest) not seen in the other spectrums at 11.3 hours. Oscillation at a single frequency explains the observed time variation only imperfectly, suggesting that multiple complex traits affect nuclear shape. In addition, although the cells are initially synchronized to the same place in the cell cycle, natural variation in cell division time leads to progressively less synchronization over the 75 hours. Different individuals are not necessarily in the same phase; see supplementary Tables 2.1 and 2.2. They are almost all within the same half cycle, but each peak at different times. This may be due to a combination of 1) differences across individuals in the time needed for cells in serum to return to growth and therefore begin dividing, and 2) progressively degrading synchronization over the roughly 4 cell cycles within our observation window.

We also tried other volume thresholds $(0.04, 0.045, 0.055, 0.06, 0.065)$ and found they agreed better with periods extracted from other shape properties. We observed peaks at the same three frequencies in all cases (see Supplemental Figures S6 and S7). In fact, the 0.05 threshold was the only one in which the 11.3 hour peak was tallest. In four of the alternate thresholds, the 11 hour peak was second tallest after the 17 hour peak, and in one it fell below both the 17 and 26 hour peaks.

To verify that these periods were not accidental, we used bootstrapping (random sampling with replacement of 10 out of 20 nuclei per time point) and extracted dominant periodicities from this data subset. After doing this 100 times, we made histograms of the top two peaks seen in each sample for each of the six shape features.

Five examples of the dynamics and spectrums of these random samples are shown in Supplemental Figures S4 and S5. As seen in Figure 2.4 G-L, the histograms all have strong peaks in the $14-18$ hour bin as well as another split between the $22-26$ and $26-30$ hour bins depending on the feature. These peaks are weakest in panel H, corresponding to thresholded volume, as it is dominated by a peak in the $10-14$ hour bin, consistent with the peak at 11.3 hours observed across the full dataset.

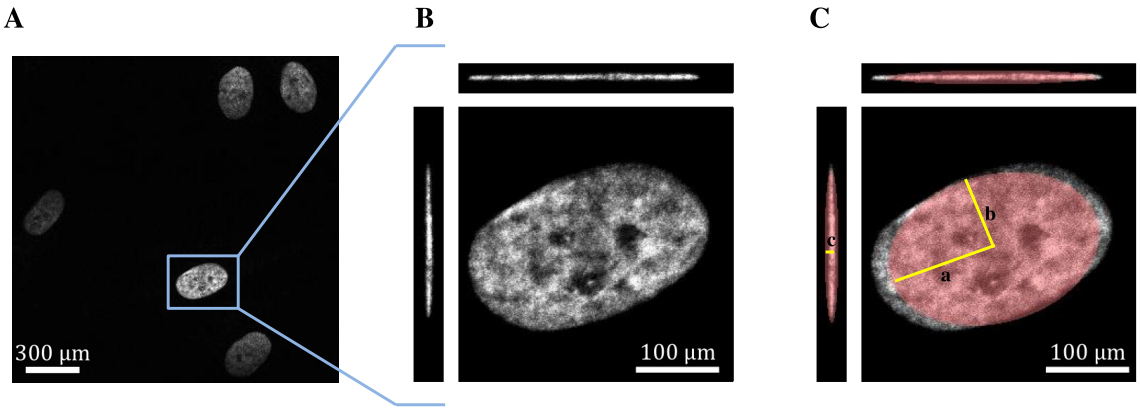

Figure 2.1: Image segmentation and ellipsoid fitting. A) A single xy slice from a raw z-stack containing multiple DAPI-stained nuclei. B) An xy slice with xz and yz projections after segmentation into an individual nucleus. C) Ellipsoidal fit described by the lengths of three axes. Fits of 20 nuclei for each individual and time point were then analyzed.

2.7 Conclusion

Using a simple model of nuclear shape in which the nucleus is modeled as an ellipsoid represented by its three axis lengths and derived from DAPI-stained images, we find that both the eccentricity and volume of primary human fibroblast nuclei change significantly over time. A single sample in time provides an incomplete picture. This result has significant impact for studies comparing cell populations, where normal time variation can be conflated with differences between cell types. Observations at multiple time points seem to be necessary to establish that any size or shape differences are due to intrinsic differences rather than natural oscillations. By comparing

Table 2.1: Sampling Schedule. The time points imaged and analyzed for each of the six individuals.

time index i	time from start t_i (hr)	interval $t_i - t_{i-1}$ (hr)
1	5	5
2	8	3
3	11	3
4	14	3
5	17	3
6	20	3
7	23	3
8	26	3
9	29	3
10	31	2
11	33	2
12	35	2
13	37	2
14	39	2
15	41	2
16	43	2
17	45	2
18	47	2
19	49	2
20	51	2
21	53	2
22	55	2
23	57	2
24	59	2
25	61	2
26	63	2
27	65	2
28	67	2
29	69	2
30	71	2
31	73	2
32	75	2

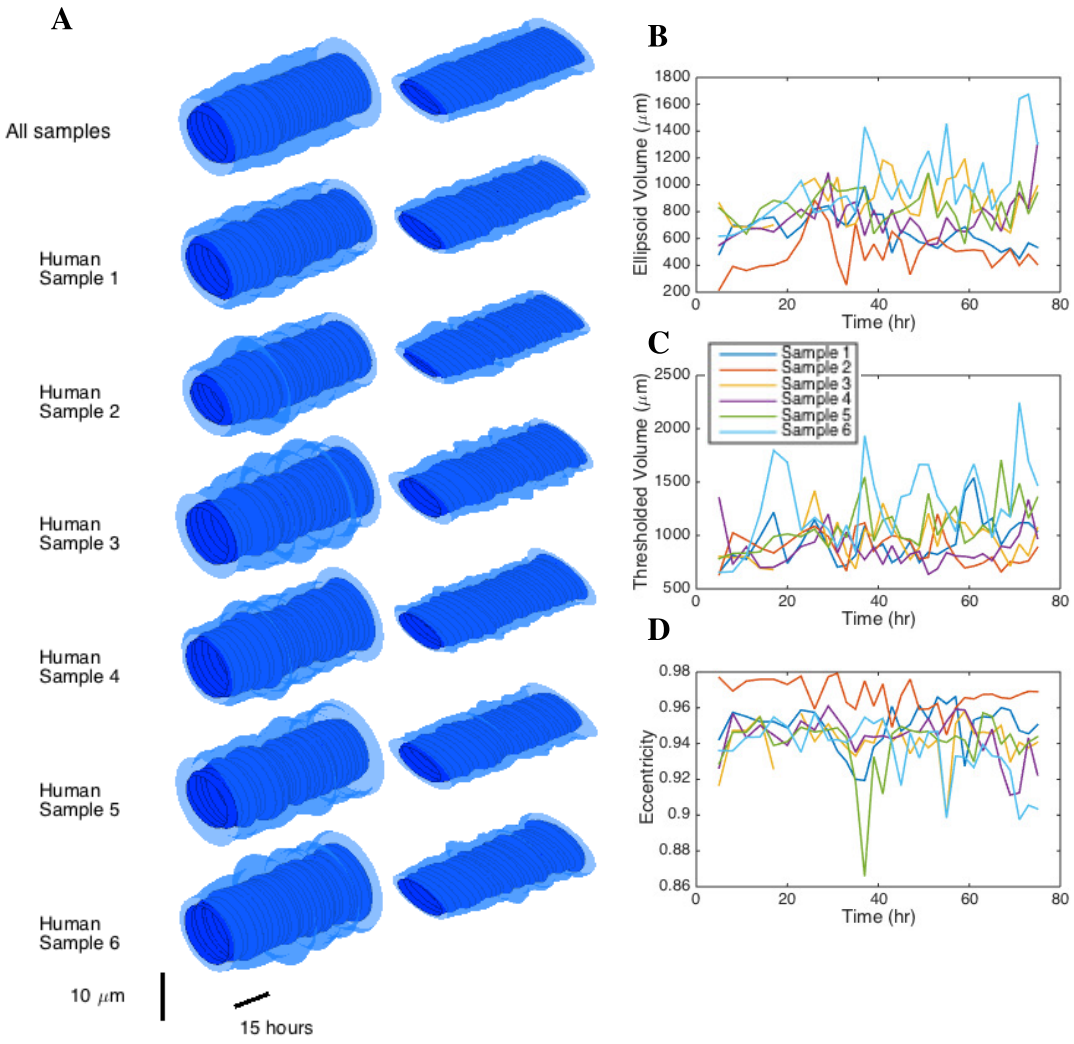

Figure 2.2: Nuclear shape dynamics. A) xy (left) and xz (right) projections, shown averaged over all individuals (top row) and for each individual (bottom six rows). The long axis is time with each cross-section forming an ellipse defined by the average length plus standard deviation for the outer and average length minus standard deviation for the inner ring, at each time point. B) Ellipsoidal volume, C) threshold volume, and D) eccentricity over time, separated by individual.

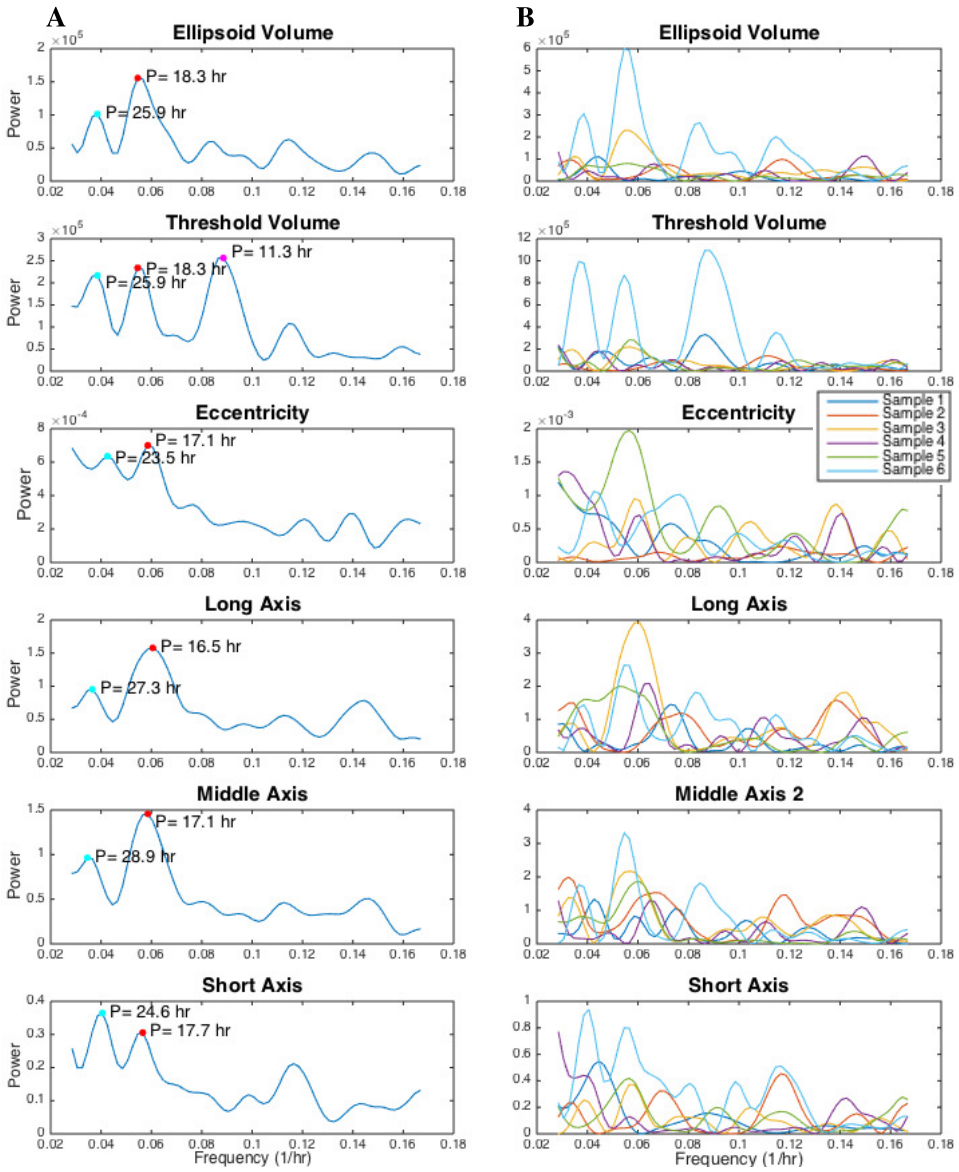

Figure 2.3: Frequency spectrums for nuclear shape. A) The average amount of squared signal energy fit from the six individuals for shape properties representing volume, threshold volume, eccentricity, and lengths of the three ellipsoid axes. Red dots mark the (generally strongest) peak around 17 hours. Blue dots mark the (somewhat weaker) first peak around 26 hours. B) The spectrums for all individuals.

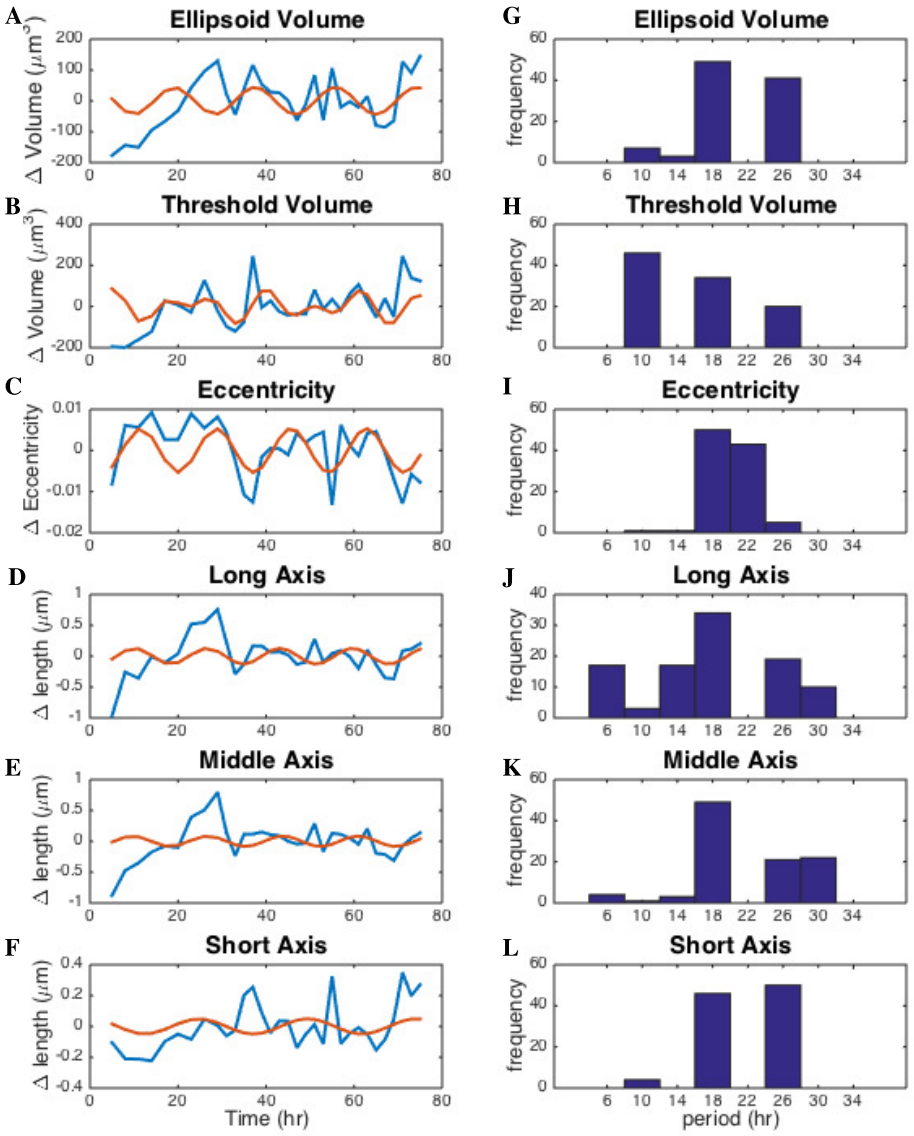

Figure 2.4: Optimal fits over all individuals The blue lines are the average of the mean-centered data for A) volume, B) threshold volume C) eccentricity, D) longest, E) middle, and F) shortest axis lengths. The red lines show the fit at the strongest detected frequencies of $18.3, 17.1, 16.5, 17.1$, and 24.6 hours, respectively. G-L show histograms recording the top two peaks from each of 50 random samples of half the data for G) volume, H) threshold volume, I) eccentricity, J) longest, K) middle, and L) shortest axis lengths, respectively.

Table 2.2: Periodic fit results. Average ± standard deviation for each feature and individual.

	Threshold Volume (μm^3)	Ellipsoid Volume (μm^3)	Eccentricity
Sample 1	951 ± 387	649 ± 352	$0.949 \pm 2.47e-4$
Sample 2	877 ± 299	492 ± 301	$0.967 \pm 9.20e-5$
Sample 3	960 ± 464	883 ± 519	$0.941 \pm 8.37e-5$
Sample 4	872 ± 332	748 ± 305	$0.943 \pm 6.91e-5$
Sample 5	1090 ± 488	836 ± 406	$0.940 \pm 1.19e-4$
Sample 6	1300 ± 608	1010 ± 497	$0.936 \pm 1.17e-4$
All Samples	1010 ± 466	769 ± 438	$0.946 \pm 1.49e-4$

	Axis 1 (μm)	Axis 2 (μm)	Axis 3 (μm)
Sample 1	2.55 ± 0.691	6.68 ± 1.32	8.56 ± 1.20
Sample 2	2.02 ± 0.620	6.42 ± 1.37	8.36 ± 1.32
Sample 3	2.91 ± 0.397	7.61 ± 1.56	8.87 ± 1.70
Sample 4	2.78 ± 0.376	7.22 ± 1.09	8.55 ± 1.20
Sample 5	2.91 ± 0.400	7.40 ± 1.34	8.76 ± 1.48
Sample 6	3.17 ± 0.606	7.83 ± 1.34	9.15 ± 1.40
All Samples	2.72 ± 0.644	7.19 ± 1.43	8.71 ± 1.41

multiple nuclei sampled at different time points, conclusions can be drawn about the dynamics of nuclear shape without measuring it as a continuous property in a single cell.

The methodologies provided in this paper are straightforward and simple to apply. Our statistical methodology has the benefit of being applicable to non-live, i.e., sacrificial, protocols. Although the initial steps in data analysis required some user input, they can easily be fully automated to simplify future studies or translational work.

The strongest oscillatory signal, as shown through basis fitting, is at 17 hours and is consistent with the length of the cell cycle of primary human fibroblasts. Additionally, the phasing of volume and eccentricity is consistent with serum starvation synchronization. Further studies with live cell imaging would help clarify the exact causes of the oscillations by monitoring nuclear volume and shape in single cells as

they grow and divide. Another weaker signal was seen at 26 hours. This signal is hypothesized to relate to circadian rhythm. A future study including RNA sequencing and nuclear shape measurements could clarify its role by looking for correlation between the changes in nuclear shape and expression of CLOCK, other circadian rhythm genes, or any other cyclically expressed genes.

Much previous work on the shape and structure of the nucleus has examined its pathology in diseases such as progeria and cancer. Looking at the dynamics of these systems and how their time course differs from healthy cell dynamics can provide more insight into the role of structure and shape in these diseases. Additionally, the interplay between nuclear shape and chromatin organization and dynamics can be further explored by studying how these shape changes correlate with chromatin conformation as observed through fluorescent in situ hybridization (FISH) or Hi-C. Rangimini et. al. showed that changes in cellular shape might lead to local chemical gradients and thus to amplification of signals including transcriptional regulation at a cellular level. At the nuclear level, a similar mechanism might be at work where changes in nuclear shape influence the distribution of chemicals at different times in the cell cycle, leading to transcriptional changes.

Acknowledgement

We would like to thank Haiming Chen for experimental support, Stephen Lindsly and Alicia Kalisi for help processing the data, and Scott Ronquist for valuable feedback on the drafts.

CHAPTER III

Chromosome conformation and gene expression patterns differ profoundly in human fibroblasts grown in spheroids versus monolayers

3.1 Abstract

Human cells derived for *in vitro* cultures are conventionally grown as adherent monolayers (2D) which do not resemble the natural 3D tissue architecture. We examined genome structure with chromosome conformation capture and gene expression with RNA-seq in fibroblasts derived from human foreskin grown in 2D and 3D conditions. Our combined analysis of Hi-C and RNA-seq data shows a large number of differentially expressed genes between 2D and 3D cells, and that these changes are localized in genomic regions that displayed structural changes. We also find a trend of expression in a subset of skin-specific genes in fibroblast cells grown in 3D that resembles those in native tissue.

3.2 Introduction

Growing mammalian cells *in vitro* is an indispensable technique for cell biology and biomedical research. Conventionally, human cells have been derived to grow in

defined medium either in suspension or as an adherent monolayer. For examples, lymphoblastoid cells derived from human blood are grown in suspension while fibroblasts derived from human skin and many cancer cell lines are grown in monolayers. Adherent monolayer (2D) cell cultures do not resemble the natural 3D structures of body tissues, and as a result cells grown in 2D may have considerable discordances in cellular morphology, physiology, pathology, cell-cell interaction and communication compared with natural tissues.

Increasing evidence shows that *in vitro* 3D culture captures natural tissue complexity better than 2D cultures [46, 3, 20, 108]. Advances in 3D culture techniques open new avenues for *in vitro* modeling of human organ development, tissue morphogenesis, pathogenesis of diseases, cellular response to drugs or other perturbations, and screening for novel therapeutics [108, 23]. Modeling organogenesis and development has been advanced by generating human micro-tissues *in vitro* [65]. For example, human pluripotent stem cells can differentiate into a midbrain-like structure in 3D cultures consisting of neurons expressing midbrain markers such as neuromelanin, and producing dopamine [60]. Alzheimer disease pathology has been recapitulated in 3D neural culture, which demonstrated a more matured neuronal and glial differentiation, and increased expression of adult tau isoform protein levels in 3D culture compared with 2D culture [20]. Human cell-based 3D models in pharmaceutical research can complement animal models, which often fail to predict the efficacy and toxicities of new drugs. 3D human-cell models may also provide more effective and economical screening of new drugs than the use of animal models [117, 128]. Furthermore, *in vitro* 3D modeling of native tissue provides tools for regenerative medicine. However, understanding the fundamental cell biology is critical in translating *in vitro* discoveries into clinical applications, e.g., functional replacement of damaged tissue.

Tissue-specific gene expression is the molecular basis of cellular function. It is not fully established how closely *in vitro* 3D tissue culture mimics native tissue. We

hypothesize that the interplay between genome structure and function, i.e., the nucleome (https://commonfund.nih.gov/4Dnucleome/index), is the key component of tissue-specific gene expression. Hi-C provides a tool to study genome structure by allowing measurement of genomic regions that are physically close together in cell nuclei [73]. Analysis of Hi-C data suggests that mammalian chromatin is partitioned into two compartments, corresponding to transcriptionally active euchromatin and inactive heterochromatin regions [73]. In addition, Hi-C analysis identified that mammalian chromosomes are organized into local chromatin interaction domains, called TADs [31]. The nucleome of a cell type can be investigated by combining analysis of Hi-C with RNA-seq [16]. We are interested in studying how the nucleome changes between 3D- and 2D- grown cells. We previously observed chromosome conformation changes between human fibroblasts grown as spheroids vs. monolayer cultures [17]. Here we extend our investigation into how genome conformation (structure) changes affect changes in genome- wide transcription (function). We focus on the nucleome of human fibroblasts grown in 3D and 2D cultures for 48 hours. We find that more than three thousand genes change expression levels greater than 2-fold (false discovery rate (FDR) 0.05) between 2D and 3D cultures without other perturbations. Analysis of Hi-C data shows that these genes are localized in genomic regions with different spatial configuration between cells grown in 3D and 2D cultures.

3.3 Results

3.3.1 Differentially expressed genes between 3D and 2D cell cultures

We analyzed the expression profiles between 3D and 2D cultures with the edgeR software [104], and identified 3297 genes that changed expression levels greater than 2-fold between the two groups (FDR ≤ 0.05). Among these changes, 1253 genes showed increased expression levels, and 2044 genes showed decreased expression levels in the

3D group relative to the 2D samples (Figure 3.1, Table A.4). We identified biologic themes from the lists of up- and downregulated genes using the expression analysis systematic explorer (EASE) software for gene ontology (GO) annotation [53]. We used a FDR threshold ≤ 0.05 to call significant gene set enrichment under any GO term.

Among the genes with increased expression levels in the 3D samples, we identified functional gene sets that significantly clustered under 113 GO terms (Table A.4). These functional gene sets are part of several important biologic processes, including those for chromosome structure/chromatin assembly; transcription or regulation of transcription; apoptosis; responses to stress, defense, inflammatory, or wound healing; responses to unfold protein or protein stimulus; signal transduction; and cytokine-cytokine receptor interaction. In addition, several gene sets are identified under GO terms in the "Cellular Component" system, including genes whose protein products are localized in cellular subcompartments, i.e., enriched under GO terms of nucleus, chromosome, chromatin, nucleosome, and extracellular space (Table A.4). The preferential cellular component localization suggests that the upregulated genes are non-randomly distributed in cellular sub-compartments. Two examples of the coordinated expression of these functionally related genes follow.

First, we looked at the 131 genes clustered under the GO term "transcription" (Table A.4). For example, more than 21 genes encode DNA binding zinc finger transcription factors; 11 genes (*GTF2A1, GTF2B, NR1D1, NR2C2, NR4A2, NR4A3, POLR2H, PPARA, TAF13, TAF7, TBP*) encode factors involved in transcription initiation or transcription elongation from RNA polymerase II promoters; 9 genes (*AHR, ARNTL, ATF4, CRY1, CREM, NPAS2, NR1D1, PPARA, RELB*) encode transcription factors that are known components critical for circadian regulation of gene expression. Second, in a cluster of 111 genes under the GO term "cell differentiation" many of them are likely to be regulated by the transcription factors from

the "transcription" cluster described above. The "cell differentiation" related genes were expressed at higher levels in the 3D samples relative to 2D samples (Table A.4). For instance, 14 of these genes encode cytokines or growth factors and are secreted into the extracellular space; 12 genes encode for proteins participating in signaling pathways, such as the TNF, NF-kappa B signaling, and cytokine-cytokine receptor pathways, likely leading to increased activity of these pathways.

We separately performed GO annotation for the downregulated genes in 3D cells relative to 2D cells. We identified gene clusters significantly enriched under 116 GO terms (Table A.4). The main biologic themes extracted from the downregulated genes include cell cycle control; cell growth regulation; and cytoskeleton organization and biogenesis. For example, we found 102 genes significantly clustered under the GO term of "cell cycle". To name a few, genes encoding cyclins (*CCNA2, CCNB1*, and *CCNE1*) and cyclin dependent kinase 6 (CDK6) are significantly downregulated. The expression of these genes is cell cycle regulated, and promotes G1 progression, G1/S and G2/M phase transitions. As another example, we found 228 genes clustered under the *GO* term of "anatomic structure development" (Table A.4). Among the 228 genes, for instance, there are sub-clusters encoding signal peptides (89 genes), secreted proteins (62 genes), glycoproteins (88 genes), or proteins for extracellular matrix organization (22 genes), or extracellular space (57 genes). In addition, from the list of downregulated genes we found that *GO* terms in the "Cellular Component" system enriched with genes whose protein products were predominantly localized outside the 170 nucleus, and formed significant clusters for basement membrane, cytoskeleton, extracellular matrix, intracellular membrane-bound organelles, mitochondrion, and cytoplasm. These cellular sub-component distributions are different from those upregulated in 3D cells.

3.3.2 Validation of RNA-seq results with TaqMan assays

We tested the expression levels of 8 genes with TaqMan assays [52] for validating differential gene expression between 2D and 3D cells identified from edgeR analysis. We found that all the genes tested were differentially expressed as shown in the RNA-seq result (3.1). The $\log_2$ fold change (FC)s between TaqMan and RNA-seq are highly correlated ($r = 0.997$, $p \leq 3.467E-8$). This analysis confirmed the list of differentially expressed genes identified from our RNA-seq experiment.

Taken together, we show that the upregulated genes in 3D cells compared with 2D cells whose products are mostly transcription factors, growth factors, signaling proteins, or proteins involved in chromosome assembly. The downregulated genes are related to cell cycle control, cytoskeleton organization and cellular structure morphogenesis, formation of extracellular matrix, or they are signaling peptides. The coordinated expression of a large number of genes suggests that the nucleome is reconfigured in 3D samples to adapt to the dense growing environment in spheroids.

Previous results suggest that 3D cultures are closer to native tissues [46, 3, 20, 108]. In our experiments we analyzed gene expression in human foreskin fibroblasts. Therefore, the nearest native tissue to compare is human skin. A recent study of gene expression profiles in human tissues by Edqvist et. al. identified 106 skin-specific genes known to be involved in skin development and differentiation [34]. Comparing the top 50 skin-specific genes available from this publication[36], we found that in both 3D and 2D samples 37 of them were not detectable at the current sequencing depth, 30 genes were expressed at low levels (fragments per kilobase of transcript per million mapped reads (FPKM) < 1) either in 3D or 2D cells, and 3 genes were called expressed. All 3 expressed genes (*ASPRV1, KRT10, and SERPINB7*) showed increased expression levels in 3D cells relative to 2D cells. Among the 30 low level expression genes, 20 showed higher levels in 3D cells (Table A.4). This trend of higher expression of skin-specific genes in 3D cells suggests that 3D cultures are closer to

native tissues.

3.3.3 Relationship between chromosome conformation and gene expression level changes

To gain insights into how genome structure affects gene expression patterns observed in 3D cells, we explored chromosome conformation changes from Hi-C data for the respective culturing conditions. First, we calculated the Fiedler number for each of the differentially expressed genes [16]. In the context of Hi-C analysis, the magnitude of the Fiedler number is a measure of the underlying stability of the topology of the genomic region, in this case a gene with defined linear sequence coordinates. A high Fiedler number suggests a high conformational stability, i.e., few alterations between chromatin states that may be important for regulation of gene expression. We found that the Fiedler number changes in 91% of the differentially expressed genes, while for the entire genome, this number changed in 86% of the genes ($p \leq 0.001$). Figure S8 shows the interaction matrices for 4 sets of genes clustered under GO terms transcription (131 genes), cell differentiation (111 genes), anatomic structure development (228 genes), and cell cycle (102 genes) for the 3D and 2D samples, as well as the difference between them. These plots show that the connections within a set of related genes change between 3D and 2D growth. This, in combination with the observation that the Fiedler number of these regions changes, shows that differentially expressed genes also undergo structural changes between 3D and 2D culture.

We also wanted to explore more generally how changes in structure are related to changes in expression. It is known that the genome is partitioned into transcriptional active or inactive regions [73], and further organized into TADs [31]. We found $2,487$ TADs in the 3D sample and $3,018$ TADs in the 2D sample (Table A.4, also see supplemental method). Three quarters of the TAD boundaries defined in the 3D samples were also present in the 2D samples. Interestingly, the TADs on chromosomes 18

were the most consistent between the samples while the TADs on chromosome 19 changed the most between 3D and 2D culture (also $3, 6, 11$, and 21). It has previously been shows that chromosome 18 had increased intra-chromosomal interactions while the chromosomes whose TADs changed the most, including 19, had decreased intra-chromosomal interactions [17]. Additionally, chromosome 19 has the highest gene density while chromosome 18 has the lowest, which further suggests that gene expression and chromosomal structure are tightly coupled. Figure 3.2 shows the gene expression and Fiedler vectors for chromosomes 18 and 19, as well as a portion of the Hi-C matrix with the TAD boundaries overlaid and the strong interactions within the region in both 3D and 2D growth conditions. The small number of bins whose Fiedler vector flips sign in chromosome 18 compared with the large number that change in chromosome 19 indicates that the very gene poor chromosome does not change structure nearly as much as the very gene rich chromosome between 3D and 2D cells. This is consistent with the interaction plots (Figure 3.2D and H) in which chromosome 19 had far fewer connections that did not change between samples than chromosome 18 (14 and 125, respectively).

In summary, we present here a comprehensive comparison of both genome structure, as measured using Hi-C, and function as established by RNA-seq. Our results show massive changes between 3D and 2D cultured isogenic cells in both structure and function and we conclude that 3D cultures more faithfully recapitulate patterns observed in primary tissues.

3.4 Discussion

We report here a larger number of genes that are differentially expressed between 3D and 2D cells due to a simple difference in the growth condition of a flat surface or spheroids. Among the 1253 genes that increase expression levels in 3D cells, gene ontology annotation shows clusters of genes significantly enriched under GO terms

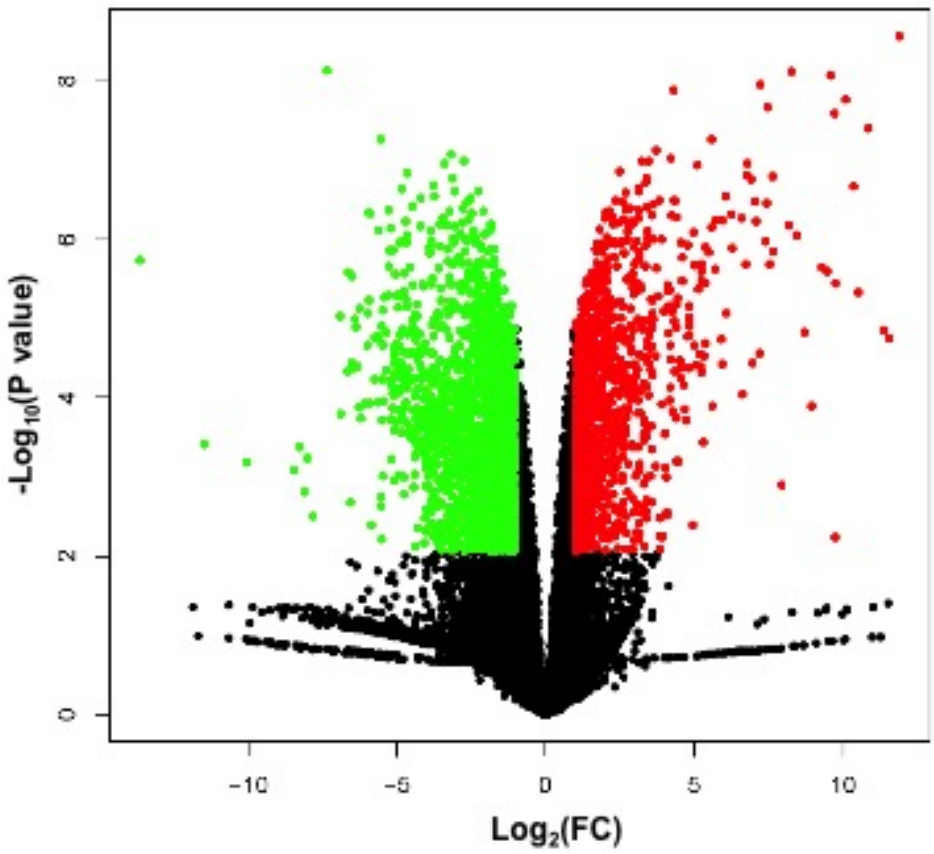

Figure 3.1: Volcano plot of gene expression changes. A volcano plot shows the up-regulated genes (red dots) and downregulated genes (green dots) in 3D cells relative to 2D cells. The X-axis shows $\log_2$ FC, and Y-axis indicates $\log_{10}$ P value.

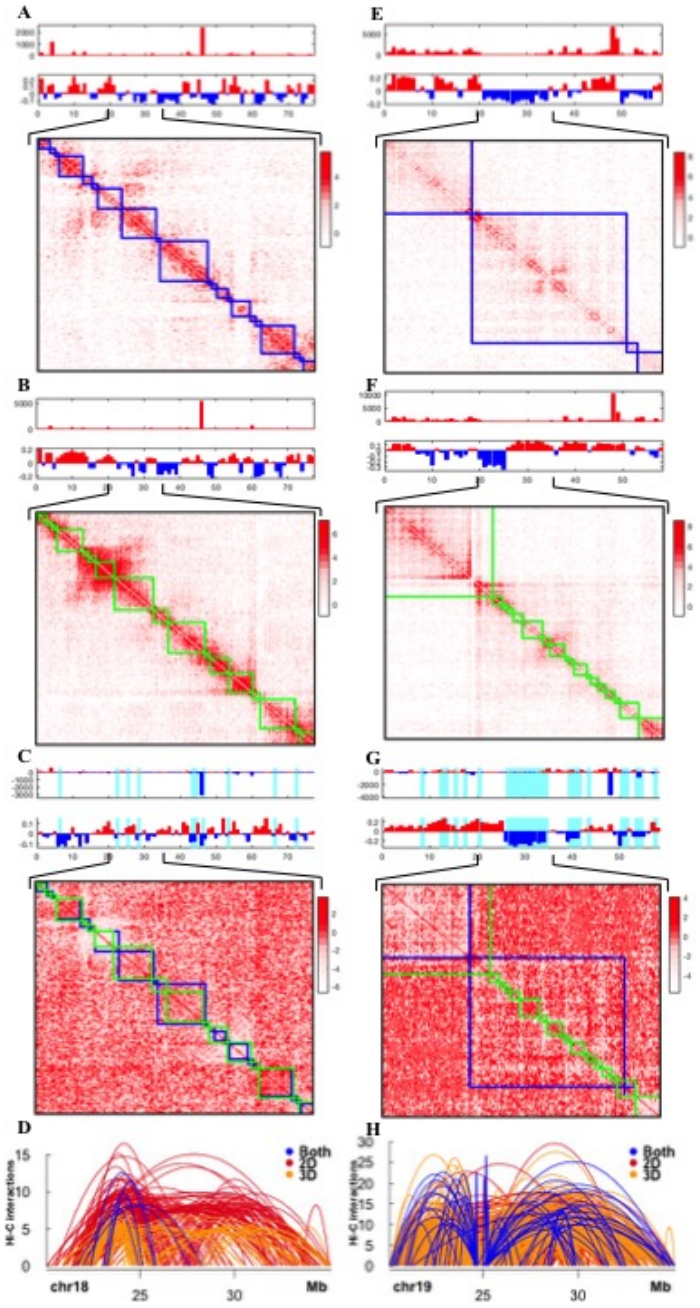

Figure 3.2: Differences in structure and function across chromosomes. The 1 Mb gene expression (top) and Fiedler vectors (middle), and part of the 100 kb Hi-C matrix with TAD boundaries (green) for chromosome 18 in (A) 2D culture, (B) 3D culture, and (C) the difference between the cultures. (D) Strong interactions within the same region in the 2D sample (red), 3D sample (yellow), and both samples (blue). The same data for a different region in chromosome 19 in E) 2D culture F) 3D culture, (G) the difference, and (H) the strong interactions in each.

Table 3.1: Comparison of TaqMan assay-based real-time quantitative PCR and RNA-seq analyses of 8 genes differentially expressed between 2D and 3D cells. TaqMan P is the t-test p value from TaqMan data analysis for each gene. The $\log_2$ FC correlation coefficient between TaqMan and RNA-seq is 0.997 ($P \leq 3.467 \times 10^{-8}$).

Gene Gene	RNA-seq $\log_2$FC	TaqMan mean $\log_2$FC	TaqMan p Bonferroni
ATP5O	−1.02	−0.83	1.14×10^{-2}
BDH2	−2.28	−2.80	2.88×10^{-3}
COL5A2	−3.58	−4.06	4.12×10^{-6}
DEPDC1	−5.09	−5.09	7.72×10^{-5}
FKBP8	−1.02	−1.49	3.51×10^{-4}
IL6	7.55	7.18	1.73×10^{-5}
NFIL3	2.77	2.73	4.03×10^{-5}
TFRC	1.60	1.63	5.29×10^{-5}

related to transcription, chromosome assembly, and signaling pathways. There are also 2044 downregulated genes whose protein products are primarily localized in the cytoplasm, extracellular matrix, extracellular space, and are related to cell cycle and cellular signaling. We validated a subset of 8 genes using the TaqMan method [52].

Our gene expression data show increased expression of genes (e.g., *CDKN1C, CCNT1*, and *CCNT2*) inhibiting G1 progression, G1/S and G2/M transition in the cell cycle, or decreased expression of genes (e.g., *CCNA2, CCNB1, CCNE1*, and *CDK6*) promoting proliferation. This suggests the 3D cells may have reduced proliferation rates compared with 2D cells. It is currently undetermined whether cells grown in 3D are quiescent or senescent. However, the increased expression of 111 genes related to cell differentiation suggests fibroblasts grown in 3D may transition toward a more differentiated state compared with the more proliferating state in 2D. A comparison to the top 50 skin-specific genes from previously published work [34] showed a trend of higher expression of skin-specific genes (23 out of 50) in 3D cells relative to 2D cells. For the remaining 27 genes, 17 were not detected in our samples, and 10 showed lower expression levels in the 3D samples. This discrepancy might be

explained due to the relatively short period of tissue culture (48 hour). At this early stage the 3D spheroid is immature and has not developed into a skin-like structure. Nevertheless, the fact that more skin-specific genes show higher expression levels in the 3D samples suggests that 3D cultures might be closer to native tissues.

We also compared these changes in gene expression to changes in the structure of the genome as measured by Hi-C. We found that differentially expressed genes were significantly more likely to have changes in their structural stability, as measured by Fiedler number, than expected from a random change. Of the differentially expressed genes 71% showed decreases and 18% showed increases in Fiedler number from 2D to 3D. This indicates that the genes that change functionally, i.e., expression levels, also have corresponding changes in their chromatin organization.

In our analysis of Hi-C data to infer chromosome conformation, we use the Fiedler vector for chromatin compartment partition and TAD identification [16, 18]. This method performs equally well compared with other methods [73, 31, 36]. In general, we observed TAD boundaries changing and Fiedler vector sign switching between 3D and 2D cells genome-wide. These observations suggest that chromosome conformation is reconfigured in 3D cells when 2D cells were used as the baseline. Interestingly, the most gene dense chromosome, chromosome 19, has one of the greatest changes in structure while the least gene dense chromosome, chromosome 18, has the least change in structure between 2D and 3D culture. This may be due to the fact that chromosome 19 is gene rich and transcriptionally active, therefore significant changes in structure are required for the changes in gene expression between 3D and 2D growth. Chromosome 18 is gene poor, and transcriptionally inactive, thus might not need to undergo as many structural changes.

We notice that TADs identified by our method do not exactly match those from other studies [31, 36]. See detailed comparison in our previous publication by Chen et. al. [18]. However, the majority of TADs are approximately within the same

genomic regions given a boundary between TADs in sizes from 40 kb to 400 kb [31]. It is possible that a TAD found by another method might decompose into several TADs obtained by our method. This is not surprising, since we take into account the connectivity of Hi-C while finding TADs. To be specific, if one TAD defined by other methods does not meet our connectivity criterion (namely, greater than λ_0), [18] it would further split into TADs of reduced size in our approach. We feel that our method is reasonable since one can adjust the parameter λ_0 to find TADs of proper size [18], and a high connectivity indicates a large modularity of community structure in Hi-C [87].

A Hi-C matrix naturally associates a graph to the genome, where nodes are defined by binned loci in the genome, and the edge weight between a pair of loci is proportional to their contact frequency. Consequently, a topological domain (or a community structure) is a compact region that can often be visually distinguished as a diagonal block in the Hi-C matrix [18]. We emphasize that our proposed topological domains are strongly connected graph components having strong intra-connections and weak inter-connections, which could be sub-regions of the commonly-used TADs. We are aware of the fact that no standard criteria are applicable to the selection of significant genes from genome-scale expression analyses. We believe that the use of FC ≥ 2 plus FDR ≤ 0.05 is a reasonable control to compensate for false positives.

In summary, we find a large number of differentially expressed genes between cells grown in 3D and 2D. Genes that show significantly increased expression levels in 3D cells are responsible for the regulation of transcription, for chromatin assembly, and for the production of cytokines and growth factors. Those that are significantly decreased in 3D cells are enriched in cell cycle control, proliferation, cytoskeleton organization and cellular morphogenesis. We observed that genes that changed expression levels were co-localized in genomic regions with structural changes as seen in sign switching in the Fiedler vectors and in changing of TAD boundaries between 3D and 2D cells. In

addition, our data add evidence to previous observations that 3D cultures recapitulate the environment of native tissues more faithfully than 2D cultures.

3.5 Materials and methods

3.5.1 Hi-C and RNA-seq data collection

We grew human foreskin fibroblasts (BJ, ATCC number CRL−2522) in 150mm dishes (2D) and in hanging drops in a 96-well PERFECTA3D plate (3D) (3D Biomatrix, Ann Arbor, MI). After 48-hours of growth, we sampled the cells for Hi-C and RNA-seq analyses. Hi-C libraries were constructed from 20 million cells for each culturing condition as described by Chen et. al. [17]. Briefly, we used the HindIII restriction enzyme (RE) for chromatin digestion. RE created DNA fragment ends were marked with biotin-dCTP (Cat# $19518-018$, Life Technologies) and re-ligated. After reverse cross-linking, the DNA is fragmented for paired-end sequencing on the Illumina HiSeq2500 platform. Meanwhile, 3 biologic replicates were collected from 2D and 3D culture conditions for RNA-seq analysis as described by Chen et. al. [16].

3.5.2 RNA-seq data analysis

We used Tophat (version 2.0.9)[122] and Bowtie (version 2.1.0.0) [69] to align the RNA-seq reads to the reference transcriptome (HG19). The average number of sequence reads generated from each sample is 35.6 million, and the average read genome alignment rate is 83.51%. We generated quantification counts from RNA-seq reads for a set of 23599 unique transcripts of RefSeq definition by NCBI. FPKM values were calculated for each gene. We used an average FKPM value ≤ 1 in either the 2D or the 3D group to call a gene as expressed, which identifies a set of 13907 genes for subsequent analysis. We used the edgeR software package [104] to identify differentially expressed genes between 2D and 3D cells. A gene is called differentially

expressed given an absolute FC ≥ 2 with FDR ≤ 0.05. We performed functional annotation of significant genes identified using the EASE software package [53].

3.5.3 Validation of differentially expressed genes identified with edgeR

We performed real-time quantitative polymerase chain reaction (RT-qPCR) using the TaqMan method [52] to verify a subset of differentially expressed genes. Eight TaqMan assays were purchased from Thermo Fisher (Cat # 4331182). All TaqMan assays were performed using a 2-step procedure according to the supplier manual (Part Number 4454239 Rev. A). First, we performed single-stranded cDNA synthesis from total RNA SuperScript®III First-Strand Synthesis System (cat # 18080051, Thermo Fisher). Second, we performed TaqMan RTqPCR assays according to the manufacturer's recommended conditions (ABI) on a 7900HT Fast Real-Time PCR System (ABI). We used SDS2.2.1 software (ABI) for quantification analysis in conjunction with the $2^{-\Delta\Delta Ct}$ method [75] using GAPDH as the reference control for normalization. The same biologic replicates for 2D and 3D RNA-seq analysis were used for TaqMan assays. The $\log_2$FC was derived from 3 Taq-Man replicates for each biologic sample in each group. For significance testing, we performed 2-tailed unpaired t test and adjusted the p-values using Bonferroni correction.

3.5.4 Hi-C analysis

Initial processing and normalization were performed as described by Chen et. al. [16]. Genome-wide TADs were defined using the iterative methods of maximizing the Fiedler number of Hi-C matrices as described by Chen et. al. [18]. A boundary was considered unchanged if it moved by less than two bins to account for uncertainty in the boundaries based on previous work that allowed variation in the boundary size [31]. At gene level analysis, an adjacency matrix for a gene was generated by the method described by Chen et. al. [18], and the Fiedler number corresponding to

each gene matrix was derived. The Fiedler number is a graph theory based measure of how well connected a graph is, with a more connected graph leading to a higher Fiedler number. Interaction matrices for the gene sets were extracted from the genome wide 1 Mb resolution Hi-C map by picking the rows and columns with the relevant differentially expressed genes in them. In line with Hi-C 1 Mb resolution maps, RNA-seq data are combined into the corresponding 1 Mb regions along a chromosome, and the gene expression level of each bin is the sum of FPKM values for all the genes in a bin Strong Hi-C interactions are those above the 95th percentile of all interactions on that chromosome.

Acknowledgments

We thank the University of Michigan Sequencing Core members for outstanding data generation.

CHAPTER IV

Nucleome Analysis Reveals Structure-function Relationships for Colon Cancer

4.1 Abstract

Chromosomal translocations and aneuploidy are hallmarks of cancer genomes; however, the impact of these aberrations on the nucleome (i.e., nuclear structure and gene expression) are not yet understood. Here, the nucleome of the CRC cell line HT-29 was analyzed using Hi-C to study genome structure, complemented by RNA-seq to determine consequent changes in genome function. Importantly, translocations and copy number changes were identified at high resolution from Hi-C data and the structure-function relationships present in normal cells were maintained in cancer. In addition, a new copy number-based normalization method for Hi-C data was developed to analyze the effect of chromosomal aberrations on local chromatin structure. The data demonstrate that at the site of translocations the correlation between chromatin organization and gene expression increases; thus, chromatin accessibility more directly reflects transcription. Additionally, the homogeneously staining region of chromosome band 8q24 of HT-29, which includes the *MYC* oncogene, interacts with various loci throughout the genome and is composed of open chromatin. The methods described herein, can be applied to the assessment of the nucleome in other cell

types with chromosomal aberrations. These tools created are packaged as NAT,a user-friendly and powerful MATLAB toolbox for time series analysis of Hi-C data and RNA-seq data. NAT can load and normalize data, define topologically associating domains, analyze translocations, produce visualization, and study time course data.

Implications: Findings show that chromosome conformation capture identifies chromosomal abnormalities at high resolution in cancer cells and that these abnormalities alter the relationship between structure and function.

4.2 Introduction

All cancers have chromosomal aberrations. These aberrations can be structural (translocations, insertions, deletions, inversions) or numerical (aneuploidy) [47, 41]. Such aberrations may activate tumor-promoting or inactivate tumor-suppressing signaling pathways [47]. However, the interplay between chromosomal aberrations (structure) and gene expression (function) is not fully understood [48, 40, 80, 101]. The development of chromosome conformation capture techniques provides unprecedented insights into spatial chromatin organization and long-range chromatin interactions in the interphase nucleus [73]. Hi-C generates matrices that reflect chromatin interactions by using proximity-based ligation followed by sequence analysis [73]. Hi-C data confirmed that the human genome is partitioned into regions of open and closed chromatin [73]. The first step in identifying these regions is to calculate the correlation matrix of the normalized Hi-C data, which describes the correlation between each pair of genomic regions. In order to compare the structure measured by the Hi-C matrix (two dimensional), to DNase I hypersensitivity or gene expression (one-dimensional), Hi-C data are converted to a vector using eigendecomposition (Table 4.1) to extract the first principal component, which identifies the vector that best approximates the matrix. Lieberman-Aiden et. al. showed that the sign of the first principal com-

ponent (positive and negative regions) divides the genome into two compartments that correlate with the presence of open or closed chromatin as measured by DNase I hypersensitivity and active or repressed gene expression, respectively [73].

Previous studies of cancer genomes using Hi-C showed long range interactions between known risk loci for the development of CRC and regulatory regions [58], demonstrated proto-oncogene activation by disruption of chromosome neighborhoods [51], determined changes in inter-chromosomal interaction frequency in breast cancer [7], and showed that changes in genomic copy number subdivide the domain structure of chromosomes [119]. We have extended this work through a comprehensive analysis of the CRC cell line HT-29 to analyze how chromosomal aberrations affect nuclear structure and gene expression, i.e., the nucleome, by integrating Hi-C and RNA-seq analyses.

4.3 Methods

4.3.1 Experimental protocols

Hi-C, RNA-seq, and FISH data were collected from human fibroblasts and the CRC cell line, HT-29, cell lines as described by Chen et. al. [16]. Cell culture in 2D and 3D growth was performed as described by Chen et. al. [17]. Extended protocols for RNA-seq, Hi-C and FISH are in the supplemental methods.

4.3.2 Normalization of Hi-C matrices

The method of Toeplitz normalization used by Chen et. al. was adapted to account for uneven genomic copy number [18]. The method, outlined in Figure S9, includes using the total number of reads in each bin of the Hi-C intrachromosomal region as a measure of the genomic copy number. A band-pass filter (Butterworth, order 4, 10^{-6} resolution) was applied to remove the high frequency noise. Breakpoints

were defined as changes in the signal greater than a threshold, and separated by at least 1/8th the chromosome length (Table S4). Each matrix was divided into submatrices based on these breakpoints and independently normalized as described by Chen et. al. [18]. The submatrices were put back together to form the final, normalized matrix. The Toeplitz normalization and iterative correction methods described by Chen et. al. [18] and Wu et. al. [134] were implemented for comparison.

The normalized matrix is used to calculate TADs, as previously described by Chen et. al., using the Fiedler vector [18]. Briefly, the Fiedler vector is the second smallest eigenvalue of the normalized Laplacian, $\bar{\mathbf{L}} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-1/2}$, where $\mathbf{A}$ is the adjacency matrix, in this case a chromosome's normalized Hi-C matrix, and $\mathbf{D}$ is the degree matrix [22]. The Fiedler vector divides the chromosome into two regions, one mostly active, the other mostly inactive and was used to calculate structure-function correlations. These regions are then subdivided into TADs by calculating the Fiedler vector of the submatrix including one of these regions until the Fiedler number of the submatrix was above the threshold of 0.6. All figures showing raw matrices are on $\log_2$ scale.

4.3.3 Hi-C matrices for translocated chromosomes

To create Hi-C matrices for translocated chromosomes, interchromosomal Hi-C matrices were visualized to identify where translocations occurred; then the exact location was refined using the read level data (Figures S10-S24). To construct Hi-C matrices for the translocated chromosomes, the information organized according to chromosome number and the traditional reference chromosomes (hg19) needs to be rearranged. To do this, each chromosome is viewed as a matrix that can be decomposed into submatrices. Based on the location of a translocation, four submatrices are created as diagrammed in the top of Figure S15. Any two intrachromosomal matrices, A and B, can be represented as

$$C_A = \begin{bmatrix} C_A(1,1)_{m\times m} & C_A(1,2)_{m\times n} \\ C_A(2,1)_{n\times m} & C_A(2,2)_{n\times n} \end{bmatrix} \tag{4.1}$$

$$C_B = \begin{bmatrix} C_B(1,1)_{r\times r} & C_B(1,2)_{r\times s} \\ C_B(2,1)_{s\times r} & C_B(2,2)_{s\times s} \end{bmatrix} \tag{4.2}$$

where m is the location of the translocation (in number of 100 kb bins) and n is the length between the translocation and the end, such that m+n is the length of chromosome A. Similarly, chromosome B has a translocation at r and total length $r + s$. Their interchromosomal space can be written as follows:

$$C_{AB} = \begin{bmatrix} C_{AB}(1,1)_{m\times r} & C_{AB}(1,2)_{m\times s} \\ C_{AB}(2,1)_{n\times r} & C_{AB}(2,2)_{n\times s} \end{bmatrix} \tag{4.3}$$

From these definitions, the Hi-C matrix for the translocated chromosome AB can be pulled out:

$$T_{AB} = \begin{bmatrix} C_A(1,1)_{m\times m} & C_{AB}(1,2)_{m\times s} \\ C_{AB}(2,1)_{s\times m} & C_B(2,2)_{s\times s} \end{bmatrix} \tag{4.4}$$

Note, due to the symmetry of the Hi-C matrix, $C_{AB}(2,1)_{s\times m} = C_{AB}(1,2)^T_{m\times s}$. The same notation can be used on more complex translocations like T_{3-12}, which includes three pieces with breaks at both of the translocations. Matrices were normalized with a forced breakpoint at the translocation location.

Gene expression and banding structures were created for each translocated chromosome by piecing together the relevant parts of each chromosome. Neighborhoods were defined as submatrices centered on the translocation of a given size defined by other criteria (either 300 kb, TAD encompassing, or gene encompassing). TAD encompassing were defined as the maximum across the samples of the smallest size

required to cover a TAD boundary with a single size being chosen for each translocation. Gene encompassing was the minimum size required to include a gene on both sides of the breakpoint. The significance of the change in structure-function correlation was calculated by randomly selecting 1000 sets of locations and constructing matrices representing fake-translocations. The observed change in correlation for the real translocations was compared to the average change in correlation for the same sized regions of the randomly placed fake-translocations. The von Neumann Entropy of these regions was calculated as $\sum_{i=1}^{d} \lambda_i \log_2 \lambda_i$, where d is the sizes of the matrix and λ_i are the eigenvalues calculated from the submatrix describing the neighborhood of the normalized Hi-C matrix [91].

4.3.4 Two-way ANOVA

Two-way ANOVA analysis was performed to identify genes with expression level change between 2D and 3D cultures, between 12-hour and 5 day cultures. GO analysis was performed using database for annotation, visualization, and integrated discovery (DAVID) [54] with official gene symbols and the default background set for human analysis. The statistical test comparing the sample was performed as described by Chen et. al. [17].

4.4 Results

4.4.1 Interpretation of Hi-C with aberrant cancer genomes

We analyze the nucleome of the CRC cell line HT-29 using Hi-C to characterize chromatin organization and RNA-seq to understand consequent changes to the cellular transcriptome. Hi-C and RNA-seq datasets were generated for HT-29 cells grown on a flat surface (2D) or as spheroids (3D) for 12 hours or 5 days (indexed as 2D12hr, 2D5day, 3D12hr, and 3D5day). The time-points and culture conditions

were chosen to assess how growth conditions and cell density affect genome structure and gene expression. In normal cells, chromosomes are organized as distinct territories, which appear as diagonal dominant blocks for each chromosome in the genome-wide Hi-C contact matrix (Figure S16). However, in solid tumors, chromosomal aberrations are common, which is evident from the SKY of HT-29 (Figure 4.1A). The genome of HT-29 is near-triploid (mean = 70 chromosomes). Structural aberrations, such as translocations, deletions and inversions, rearrange the chromosomes. Such aberrations are readily visible on the Hi-C matrix as distinctive L or X shaped patterns for unbalanced and balanced translocations, respectively (Figure 4.1B). The translocations found by SKY (Figure 4.1A), e.g., the balanced translocation between chromosomes 6 and 14 (black arrow in Figure 4.1B) and the insertion of chromosome 12 material into the p arm of chromosome 3 (black arrow), are clearly visible. Enlarged representations for each of the translocated chromosomes are shown in Figure S17. Cytogenetic analysis detected a homogeneously staining region (HSR) on chromosome 8q that contains the *MYC* oncogene [101, 14]. The blue arrows in Figure 4.1B mark increased interactions between the whole genome and high copy number regions like the HSR and a smaller amplification on chromosome 2 (Figure 4.1B). The smaller amplification on chromosome 2 was confirmed by interphase FISH analysis. The recapitulation of cytogenetic changes in interphase Hi-C maps is reflected in the contact maps of chromosome 3, where the short arm ($\sim$ 2 copies) has fewer contacts than the long arm ($\sim$ 5 copies).

We compare genomic copy number with the total number of Hi-C reads for each gene and gene expression patterns based on the observation that high copy number regions have more contacts in the genome wide Hi-C matrix. Figure 4.1C shows the copy numbers, $\log_2$-FC gene expressions (relative to fibroblasts), and total Hi-C reads (FPKM) for each gene averaged for each chromosome arm (Table A.6). Genomic copy number and gene expression exhibit a strong correlation (Pearson $r = 0.65, p \leq 10$),

consistent with previous work [101]. The correlation between copy number and the total Hi-C contacts is 0.81 ($p \leq 10$), indicating that the total number of reads per bin can be used as an approximation for copy number. Therefore, the numbers of reads are a direct reflection of the likelihood that chromosome contacts occur in the interphase nucleus.

4.4.2 A novel copy number based normalization method

In cancer cells, the interpretation of Hi-C data is complicated by the presence of copy number alterations, which can affect read frequencies. Therefore, we developed a novel approach for the normalization of Hi-C data for cell lines with complex karyotypes. We validated our new approach by comparing it to high resolution molecular cytogenetic analyses of HT-29 by SKY, FISH, and array-based CGH. For instance, in HT-29 the *MYC* oncogene is present in multiple copies in an HSR at the distal end of the q-arm of chromosome 8 (Figure 4.2). Figure 4.2B shows the total number of raw Hi-C reads per bin as well as the genomic copy number as measured by CGH. Genomic copy number directly and strongly influences the total number of Hi-C reads per bin ($r = 0.77$). Based on this finding, the total number of reads per bin was used to create a new normalization method in which the Hi-C matrix was divided into sub-matrices with a constant genomic copy number (blocks in Figure 4.2C). The blocks were normalized independently as described by Chen et. al. [18], then combined to form the normalized matrix as shown in Figure 4.2C. To verify our method, we compared it to previously published methods: Toeplitz normalization and ICE [18, 134]. We found that the correlation between structure and function was highest after copy number based normalization (Figure S18, $r = 0.60, 0.53$ and 0.17 for copy number, Toeplitz and iterative, respectively). Additionally, the method performed well on all of the HT-29 samples ($r = 0.59, 0.59, 0.63$, and 0.54 for 2D12hr, 2D5day, 3D12hr, and 3D5day). We also tested the method on chromosome 20 from

the myelogenous leukemia cell line K562 [98, 24] and again found that the correlation between structure and function was highest after copy number based normalization (Figure S19, $r = 0.63, 0.59, and 0.20$ for copy number, Toeplitz and iterative, respectively). Similar to Wu and Michor [134], our method can be used for any Hi-C matrix without requiring copy number information through other approaches such as CGH.

4.4.3 Structure and function of the HSR

To further explore the effect of genomic copy number changes on genome structure, we focused our analysis on the HSR, a highly amplified 27 Mb region on chromosome band 8q24 containing *MYC* and 107 other genes that appears as a bright red band in the Hi-C matrix (Figure 4.1B). If we assume the two normal copies of the region in the HSR provide about the same number of reads as the two copies of the first third of chromosome 8, then of the reads coming from the amplified region of chromosome 8, 86% derive from the HSR, while 14% are accounted for by the unamplified copy of the region. The HSR is visualized on metaphase and interphase cells by FISH with a genomic probe for *MYC* in Figure 4.3A. We calculated the volume of the chromosome 8 territories using 3D-FISH; the chromosomes with the HSR were 2.5 times larger than the normal copies (Figure 4.3A top insert, S20, Table S6).

To quantify genome organization, we define the adjacency matrix (a sub-matrix of the normalized Hi-C matrix) for a region of interest. Then we calculate the eigenvalues of the adjacency matrix to quantify genome organization through approximating the entropy (a measure of the distribution of chromatin state) in chromosomal regions. A similar approach has previously been used to show that, during differentiation, entropy initially increases before a progressive decline as the cell approaches its differentiated state [95]. Here, we use eigenvalues of the Hi-C matrix to estimate the entropy of chromosomal regions, which is inversely proportional to order. Since chromosomal aberrations disrupt the baseline distribution of the local chromatin state,

we hypothesize that the entropy near these alterations should increase.

To quantify the structure of the HSR from Hi-C data, we calculated the entropy of the adjacency matrix that represents the contacts in the HSR. We define entropy as $\sum_{i=1}^{d} \lambda_i \log_2 \lambda_i$, where d is the sizes of the matrix and λ_i are the eigenvalues of the adjacency matrix [91]. The degree of entropy reflects the frequency with which chromatin states change in a given region. Only 6% of randomly selected regions in the HT-29 genome have lower entropy than the HSR, indicating that the HSR is highly interconnected and ordered. Since the structural conformation of a DNA region is dictated by the sequence, we expect consistency in conformation across the amplified region [91, 4]. Thus, the HSR, which contains multiple copies of the sequence, has a highly ordered structure. We measured the structure-function relationships by calculating the correlation between gene expression and chromatin state, i.e., heterochromatin or euchromatin, using the Fiedler vector. The sign of the Fiedler vector (positive or negative values) divides the genome into regions of heterochromatin and euchromatin [16]. The structure-function correlation of the HSR is greater than in 60% of the rest of the genome. In summary, our analysis showed that the chromosome containing the HSR is larger than the normal chromosome as seen with 3D-FISH (Figure 4.3A). The HSR itself is highly organized, i.e., is less entropic, and has a strong structure-function relationship.

We next explored how the HSR interacts with the rest of the genome. At 1 Mb resolution, we analyzed genome-wide interactions and interactions within the HSR (Figure 4.3B). We identified a single region in chromosome 2 that interacts strongly with the HSR in all of the HT-29 samples ($p \leq 10$ in all samples, Figure S21). The region includes six genes (*STARD7, TMEM127, CIAO1, SNRNP200, ITPRIPL1, LOC285033*) and its interactions with the HSR were verified with FISH (Figure S14). *STARD7* has been previously implicated in choriocarcinoma, CRC, breast and lung cancers [38]. This strong interaction between the amplified regions on chromosomes

2 and 8 had not been recognized before.

4.4.4 Hi-C provides high resolution maps of translocations

In addition to understanding how numerical aberrations affect chromatin organization, we explored the consequences of chromosomal translocations. As shown in Figure 4.1B, translocations generate L or X shaped patterns in the genome wide Hi-C. Hi-C allowed identification of translocations too small to be detected by molecular cytogenetic techniques, e.g., the unbalanced translocation between chromosomes 2 and 15 shown in Figure 4.4A. We confirmed this aberration using FISH (Figure 4.4B). Additionally, the balanced translocation t(6; 14) is clearly visible in the 100 kb matrix (Figure 4.4C). By viewing the translocation in the read level data, the resolution at which the breakpoint was identified increased to 1 kb. Figure 4.4D shows a single break in chromosome 14 as well as two breaks in chromosome 6. The top right shows many reads connecting the portion of chromosome 14 proximal to the breakpoint to the portion of chromosome 6 distal to the breakpoint. The bottom left portion shows reads where one of the pairs mapped to the portion of chromosome 6 proximal to the breakpoint, while the other mapped to portion of chromosome 14 distal to the breakpoint. Since there is a single horizontal line dividing the locations of the reads on chromosome 14, there is a single breakpoint, as expected for a balanced translocation. However, along chromosome 6 there is a 65 kb region between the two vertical lines that is contained in both translocated chromosomes thus it interacts with both the distal and proximal portions of chromosome 14. This was confirmed using FISH with Bacterial artificial chromosome (BAC) clones that hybridize to the translocation (Figure 4.4E). Hence, this seemingly balanced translocation is in fact unbalanced.

4.4.5 Translocations increase entropy

In order to explore the structure and function of translocated chromosomes, we constructed the Hi-C matrices representing the translocated chromosomes (Figure S15). Hi-C matrices representing the seven translocated chromosomes in HT-29 (Table 4.2) and the normal chromosomes from which they originate were constructed from raw reads (Figure S10-S14). The insertion ins(17; 22) was not used due to the presence of at least eight different configurations with unknown frequencies and pairings (Figure S23).

After constructing the Hi-C matrices for the translocated chromosomes, we analyzed the neighborhoods surrounding the breakpoints. Figure 4.5 shows the regions surrounding the breakpoints for the translocation t(6; 14). Additional translocations are presented in Figure S24-S30. Each Hi-C matrix shows a 6 Mb region centered on the translocation breakpoint with the natural domain structure of the genome, i.e., TADs) overlaid. The two plots below the Hi-C matrices show three different neighborhoods, the gene expression for a region that contains three TADs. Each neighborhood represents a different region surrounding the breakpoint: the smallest possible neighborhood, a TAD encompassing neighborhood sized to include a TAD boundary, and a gene encompassing neighborhood sized to encompass one gene on both sides of the translocation. The last two neighborhoods vary in size for the analyzed translocations, with TAD encompassing neighborhoods varying from 700 kb to 1.7 Mb. The entropy was calculated for the TAD encompassing neighborhood for each translocation (Table 4.2). Unlike in the HSR, the entropy in the region surrounding translocations was higher than at the same regions of the wild type chromosomes for 5 of the 7 translocations, including t(6; 14) (avg 1.89 and 2.03 for wild type and translocated chromosomes, respectively). This suggests that the translocations reduce local stability.

To explore whether the results were specific to HT-29, or a reflection of more

general phenomena in tumors, we analyzed publicly available Hi-C and RNA-seq data from the myelogenous leukemia cell line K562 [98, 24]. We constructed Hi-C matrices for the six translocated chromosomes in K562 (Table SS7). Figure S31-S37 shows the Hi-C map, neighborhoods, and gene expression for each of the translocations and the normal chromosomes from which they were derived. The average entropy of the neighborhoods surrounding the breakpoints on the normal chromosomes in K562 is 1.68, while the entropy of the seven translocated chromosomes averages 1.93, a 15% increase, and each translocation has higher entropy than either of the normal chromosomes (Table SS7). These results suggest a general pattern in cancer cells.

In addition, we analyzed the structure-function relationship, i.e., the correlation between Fiedler vector and gene expression, for TAD-sized neighborhoods surrounding the translocations. For HT-29, the structure-function correlation is slightly greater in the translocated regions (Table 4.2, $r = 0.36$ and 0.34, respectively). The same applies for K562 (Table SS7, $r = 0.43$ and 0.32, respectively). Compared to random locations, this is a greater increase in correlation than expected ($p < 0.09$). In conclusion, our results indicate that translocations both increase entropy and the strength of the structure-function relationship.

4.4.6 Sample differences

We previously observed differential chromatin interactions of human fibroblasts cultured in 2D or 3D conditions [17], and now explore whether such differences can be observed in HT-29, as well as differences between the time points. The percent of intrachromosomal reads that fall along the diagonal is 89% for the 3D5day sample whereas it is 72% or less for the others. Additionally, the 3D5day sample had only 48% of its total reads as intrachromosomal, whereas for the other samples 55% or more of their reads were intrachromosomal. One explanation is that the Hi-C reads are distributed differently due to changes in the cell cycle. This indicates that the

3D5day sample is far more diagonally dominated but less intrachromosomal than the 12 hr samples. This is consistent with previous results showing the same patterns in fibroblasts grown in 2D and 3D cultures [17].

Two-way ANOVA (see Methods) was performed on the RNA-seq data and showed that 287 genes were significantly differentially expressed between 2D and 3D cultures ($p \leq 0.05$, Table S8). We also explored whether cell density influences gene expression. We found 661 genes that changed between the 12 hr and 5 day time points (Table S9), of which 178 also change with growth conditions. DAVID analysis [54] of these data sets identified a number of significantly enriched GO terms including cell cycle processes, cell cycle phase, cell cycle checkpoints, regulation of cell cycle, DNA repair, and DNA-dependent DNA replication, suggesting the changes in expression are mostly related to the cell cycle (Tables S10, S11).

Table 4.1: Glossary of Terms

Term	Definition
Adjacency Matrix	square matrix with $(\boldsymbol{A})_{ij} = w_{(n_i,n_j)}$. If there is an edge between nodes i and j, the entry is the edge's weight, otherwise it is 0.
Aneuploidy	an abnormal number of chromosomes, i.e., different from 46 chromosomes for human cells.
Degree Matrix	a diagonal matrix $(\boldsymbol{D})_{ij} = \sum_{j=1}^{k} (\boldsymbol{A})_{i,j}$, the total number of edges attached to each node.
Eigenvalues and Eigenvectors	a set of numbers associated with linear systems. The decomposition into eigenvalues and eigenvectors is called as eigendecomposition. Eigenvalues are represented by λ and eigenvectors by x. $\boldsymbol{A}x = \lambda x$ with $x \neq 0$.
Entropy	a measure of uncertainty or disorder, $\sum \lambda_i \log_2 \lambda_i$ where λ_i are eigenvalues.
Fiedler number and vector	the Fiedler number is the second smallest eigenvalue of the Laplacian and a measure of the connectivity of a graph. The corresponding eigenvector is the Fiedler vector, whose sign can be used to divide a graph into two regions.
Karyotype	the number and appearance of the chromosomes in a cell.
Laplacian	a symmetric matrix, $\mathbf{L} = \mathbf{D} - \mathbf{A}$, normalized as $\tilde{\mathbf{L}} = \mathbf{D^{-1/2}}(\mathbf{D} - \mathbf{A})\mathbf{D^{-1/2}}$.
TAD	a region of a chromosome with increased local contacts and decreased contacts with its neighbors.

4.5 Discussion

We investigated how genome structure and function are altered by chromosomal aberrations in cancer cells by analyzing Hi-C and RNA-seq data from the colorectal cancer cell line HT-29. Cells were grown in 2D and 3D conditions for 12 hrs and 5 days. We showed that Hi-C captures chromosomal aberrations, including genomic copy number changes and chromosomal translocations, some of which were previously unknown. Next, we mapped the translocations using read level Hi-C data and identified the breakpoints at kb resolution. This allowed us to describe a previously

Table 4.2: The first column indicates which chromosome the translocations are on. Read Loc tells the best estimate of the location of the translocation. Ent Fib and Ent HT-29 report the von Neumann Entropy of the TAD-encapsulating neighborhoods from the 100 kb Hi-C data centered on the translocation in fibroblasts and HT-29 (average), respectively. After translocation the entropy increases on average and for 5 of the 7 translocations. S-F Fib and S-F HT-29 show the correlation between the structure (Fiedler vector) and function (RNA-seq) for TAD encompassing neighborhoods for the average of the HT-29 samples and the fibroblast sample, respectively. No t avg and t avg refer to the average of all of the translocated and non-translocated chromosomes respectively.

Chr	Read Loc	Ent Fib	Ent HT-29	S-F Fib	S-F HT-29
$2-15$			2.52		0.45
2	236760000	2.73	2.65	0.43	0.43
15	96682000	3.03	2.39	0.29	0.29
$3-12$p			1.86		0.03
3	83410000	2.04	1.64	0.00	0.00
12	34435000	1.64	1.65	0.26	0.38
$3-12$q			1.84		0.95
12	21057000	1.91	1.78	0.08	0.24
3	89440000	2.13	1.64	0.62	0.94
$5-6$			2.26		0.30
5	546620000	2.59	2.03	0.67	0.58
6	162295000	2.44	1.96	0.28	0.15
$6-14$			1.93		0.26
6	13285000	2.13	1.78	0.02	0.15
14	36508800	1.94	1.87	0.50	0.71
$14-6$			2.04		0.30
14	36508800	1.94	1.87	0.50	0.71
6	132890000	2.13	1.78	0.02	0.15
$19-17$			1.79		0.22
19	24600000	1.89	1.83	0.45	0.38
17	22253300	1.98	1.57	0.03	0.06
No t Avg		2.18	1.89	0.32	0.34
t Avg			2.03		0.36

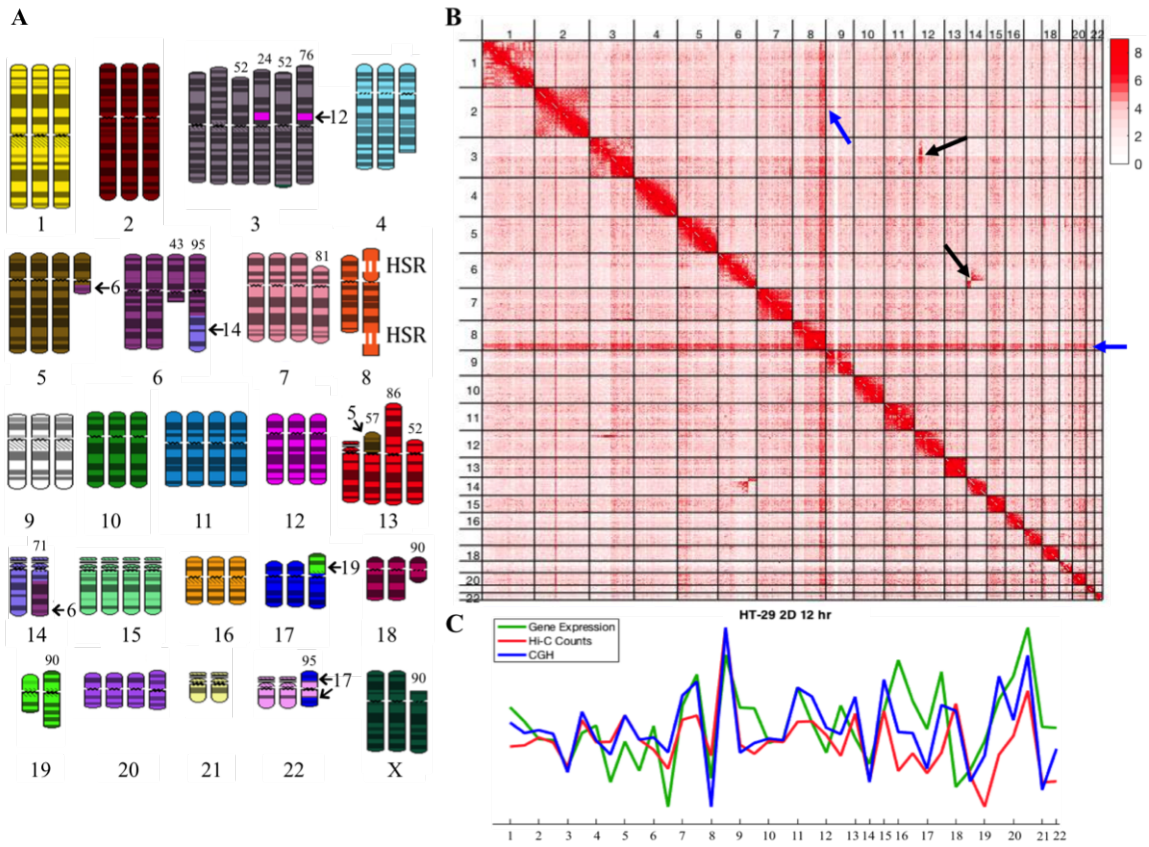

Figure 4.1: Chromosomal aberrations in Hi-C data. A) HT-29 karyotype adapted from Knutsen et. al. [64]. The numbers above chromosomes are the percent of the 21 analyzed cells in which that chromosome was seen. HT-29 averages 70 chromosomes per cell. B) Genome-wide Hi-C matrix for 2D12hr HT-29 cells at 1 Mb resolution. The X pattern marking the t(6; 14) translocation and the region of high contacts marking ins(3; 12) are identified by black arrows (see Figure S17 all translocations). The blue arrows identify amplified regions including the HSR on chromosome 8q and the amplification of a small region on chromosome 2 that interacts strongly with the HSR. The uneven copy number between the p and q arms of chromosome 3 (2 p arms, 5 q arms) can also be seen by the fact that the first half of chromosome 3 in the acHi-C matrix are a lighter red than the second half. C) The average $\log_2$ FC gene expression (green), change in Hi-C reads (red), and genomic copy number (blue) for each chromosome arm (p-arm first), also in Table SA.6.

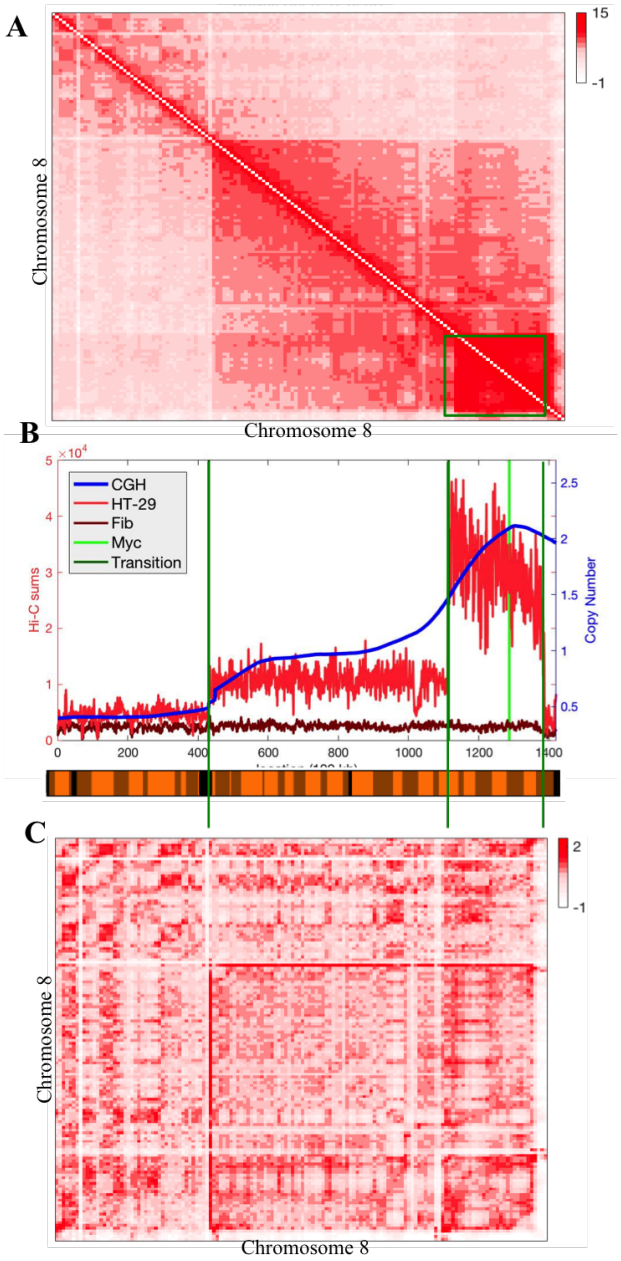

Figure 4.2: Normalization accounting for copy number changes. A) The raw chromosome 8 matrix in which regions of different genomic copy number can be seen by the differences in brightness. The HSR the box at the bottom right. B) These changes are measured by the changes in the total reads in each bin of the Hi-C matrix (bright red, HT-29 12hr2D), which follow closely the genomic copy number measured by CGH (blue). The normalization breakpoints are shown in dark green and the location of *MYC* is shown in bright green. The total reads in each bin for chromosome 8 in fibroblasts are shown in dark red. C) Each block created by the transitions between copy number regions was normalized independently then pieced back together to create the normalized matrix.

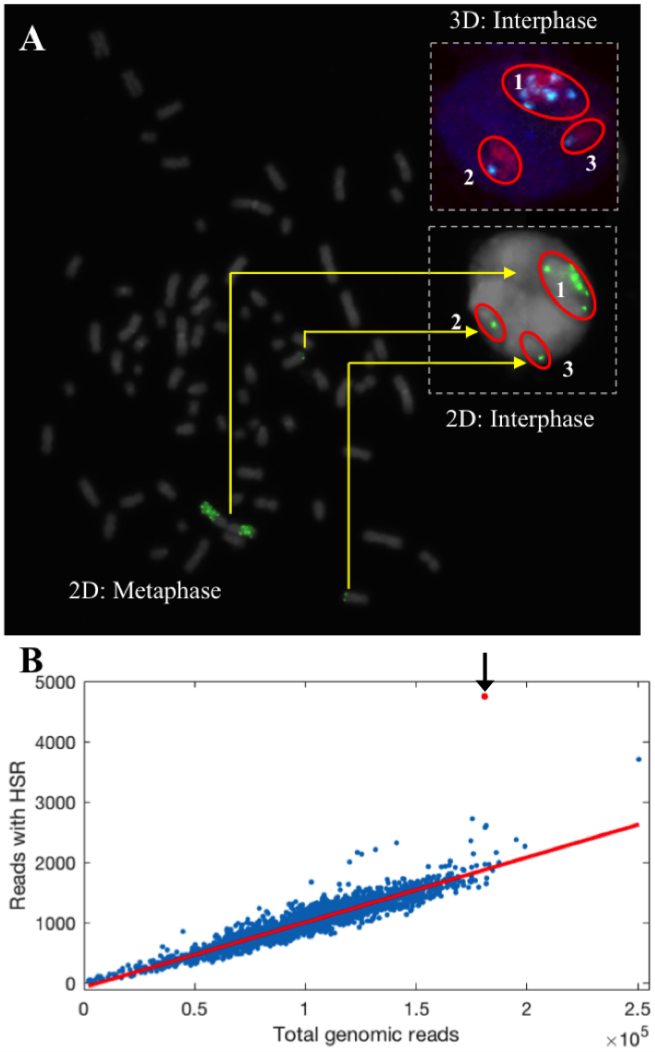

Figure 4.3: Genome wide HSR interactions. A) A metaphase spread and 2D interphase nucleus (lower inset) with *MYC* labeled in green allowing visualization of the two normal copies of the gene as well as multiple copies in the HSR. The 3D interphase image (upper inset) was used to calculate the volume of the chromosome 8 territories. B) A graph of the total genomic interactions for each interchromosomal bin against their interactions with just the HSR for 2D12hr. The red line shows the best-fit line for a region's interactions with the HSR. The red point is the amplified region on chromosome 2 that interacts more strongly than any other region in all HT-29 samples (Figure S21).

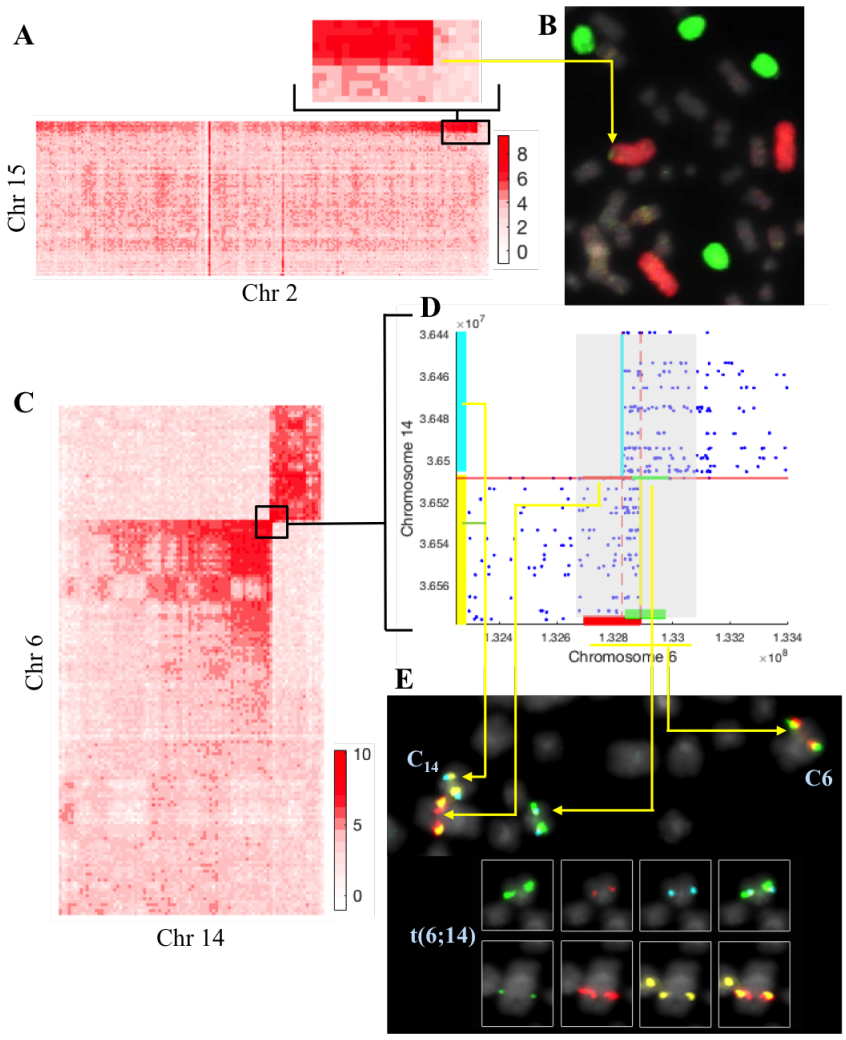

Figure 4.4: Translocations in Hi-C. A) An unbalanced translocation in which the end of chromosome 15 was added to the end of chromosome 2 at 1 Mb resolution. The inset zooms in on the relevant region. B) The translocation (yellow arrow), was verified by chromosomal painting. Because of the small size, this translocation was previously unidentified. Chromosome 2 is red while chromosome 15 is green. No chromosome 15 contains chromosome 2 material, verifying the translocation is unbalanced. C) A seemingly balanced translocation between chromosomes 6 and 14 shown at 100 kb resolution. D) The read level Hi-C data for the translocation, showing the breakpoint in chromosome 14 and two breakpoints in chromosome 6 marked by red lines. The hybridization locations of the probes are shown around the perimeter. The cyan and yellow probes mark chromosome 14 before and after the translocation. Parts of both the red and green probes on chromosome 6 are in the duplicated region. E) FISH verification of the translocation location, which is different than previously published via SKY. Because of the duplication of a 65 kb segment, both translocated chromosomes contain parts of both probes.

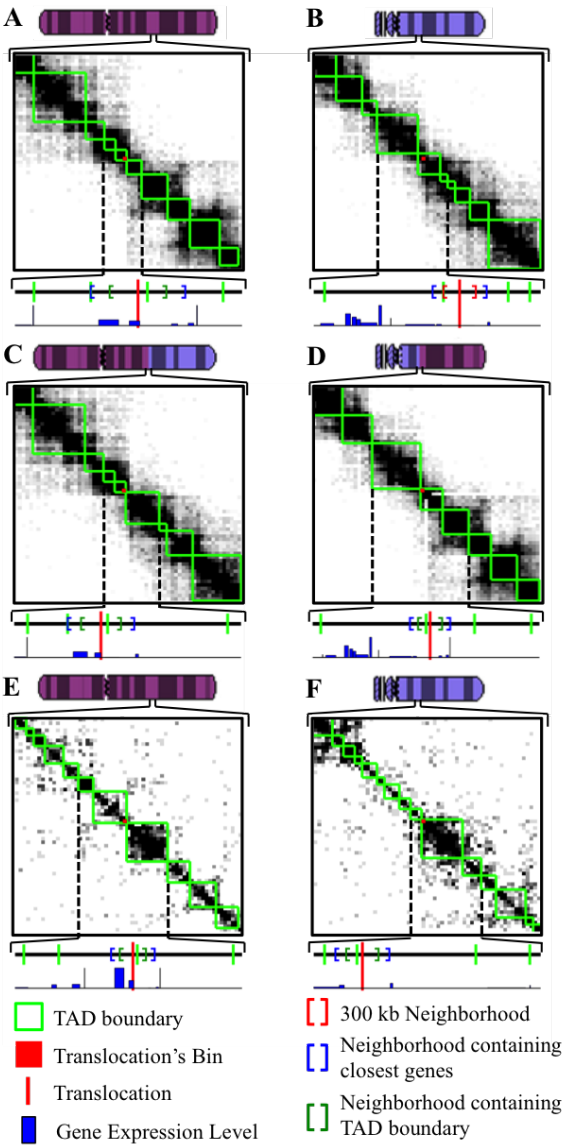

Figure 4.5: TADs on chromosomes affected by translocation t(6; 14). For each region around the translocation, the Hi-C matrix and TAD boundaries (green) are shown. The matrices show a 6 Mb region with the site of translocation at the center. The line below shows the site of the translocation and three different neighborhoods: a small 300 kb neighborhood, a neighborhood that contains a TAD boundary, and the neighborhood that contains a gene on each side of the translocation. The bar plot shows gene expression. These features are shown for A) chromosome 6 and b) chromosome 14 in the 2D12hr sample, C) the translocated chromosome containing the beginning of chromosome 6 and the end of chromosome 14 in the 2D12hr sample, D) the translocated chromosome containing the beginning of chromosome 14 and the end of chromosome 6 in the 2D12hr sample as well as E) chromosome 6 and F) chromosome 14 in the healthy fibroblast sample.

unknown unbalanced translocation, der(2), which was too small to be identified by SKY or high-resolution aCGH. Additionally, we refined the location of the seemingly balanced translocation t(6; 14) and showed that it is in fact unbalanced [14]. Therefore, in addition to providing information on the 3D organization and local chromatin stability, we showed that Hi-C can identify translocations with unprecedented resolution. It is remarkable to see the extent to which structural and numerical chromosomal aberrations are recapitulated in the interphase nucleus.

We found that the HSR on chromosome 8q interacts with many other genomic regions and is highly organized, i.e., less entropic. Entropy reflects the frequency with which chromatin states change in a given region. The HSR has a stronger relationship between structure and function, i.e., gene expression, than other regions in the genome, indicating that chromatin accessibility more directly reflects transcription. The HSR consists of open chromatin, making it conducive for transcription. We identified a small amplified region in chromosome 2 that interacts very strongly with the HSR. This finding was confirmed using FISH. The previously unidentified region contains *STARD7*, which has been previously implicated in cancers [38]. One limitation of Hi-C and RNA-seq is that different alleles of the same region cannot be distinguished. For the HSR analysis, reads from the unamplified copy of the region cannot be differentiated from those originating from the HSR. Of the reads coming from the amplified region of chromosome 8q, 86% derive from the HSR, while 14% are accounted for by the unamplified copy of the region. Thus, we expect properties of the HSR to dominate the analysis.

We analyzed local chromatin stability at translocation breakpoints in HT-29 and K562 neighborhoods and showed that regions around translocations have increased entropy compared to the corresponding regions on the normal chromosomes. This increase in entropy near breakpoints suggests that translocations decrease the local stability of adjacent neighborhoods around the translocation. The entropy in fibrob-

lasts was higher than the entropy in HT-29 or K562, which could be a reflection of the non-terminal differentiation status of fibroblasts. We also found that translocations increase the structure-function relationship in the neighborhood flanking the break-point compared to the equivalent regions on normal chromosomes. This might be a reflection of their role in tumorigenesis. We analyzed the *BCR-ABL* translocation between chromosomes 9 and 22 present in K562. Like most translocations it showed increased entropy as compared to the non-translocated regions. Unlike most of the translocations we analyzed in HT-29 and K562, the structure function correlation for the *BCR-ABL* translocation is lower than either of the normal chromosomes it comes from. This might be due to the fact that the *BCR-ABL* fusion protein exhibits constitutive activity and therefore does not require increased expression. We submit that decreasing local stability and increasing the structure-function relationship is a common phenomenon of translocations in cancer cells.

Finally, we characterized the differences between 2D and 3D cell growth and 12 hr and 5 day time points. We found that genes differentially expressed between 2D and 3D growth were primarily related to cell cycle regulation and DNA repair. We also found that the 3D5day sample was different from the other HT-29 Hi-C matrices as measured by the correlation of interchromosomal reads. The 3D samples had a higher percentage of intrachromosomal reads that fell on the diagonal and lower percentage of all reads that were intrachromosomal. Change in the distribution of counts in Hi-C matrices is consistent with previous results showing the same patterns in fibroblasts grown in 2D and 3D cultures [17]. In contrast to the 12 hr samples, the 5 day samples were completely confluent. The reason the 5 day samples are more intrachromosomal and less diagonally dominant than the 12 hr sample could be because the cells in the 12 hr sample did not have enough time to complete nuclear reorganization into a 3D growth pattern. Previous results have shown that mitotic cells lead to purely diagonal matrices since the chromosomes are organized in tight rods during mitosis

[86]. This suggests some of the differences may be related to cell cycle, as supported by the strong significance of cell cycle and mitosis related GO terms.

In summary, our analysis identifies undetected chromosomal aberrations and provides novel insight into the nucleome of cancer cells.

Acknowledgements

We would like to think Daysha Torres and Greg Farnum for invaluable feedback and Walter Meixner for imaging assistance. This study was supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, the University of Michigan Rackham Merit Fellowship program, the DARPA Biochronicity Program, and the NIH ES017885 (to G.O.). J.C. is supported by the European Commission (COLONGEVA), and received a travel grant from the Instituto de Salud Carlos III and was co-funded by the European Regional Development Fund (ERDF) (MV15/00026).

4.6 4D Nucleome analysis toolbox

4.6.1 Introduction

Understanding 3D genomic organization and how it changes over time (4th dimension) will lead to increased understanding of development and disease as supported by the 4D Nucleome project (http://www.4dnucleome.org/). A popular tool for probing genome organization is Hi-C [73]. A limiting factor in the use of Hi-C is the complex analysis required to produce biologically meaningful results.

Understanding the nucleome requires powerful and user-friendly analysis tools. Most currently available tools focus on alignment and initial read processing [32, 70, 115] or visualization in the form of browsers [33, 70, 111]. Some of these tools [33, 70] can perform some analysis including defining domains, however, they are not

easily customized or extended to perform additional analysis. Here we introduce 4D Nucleome Analysis Toolbox (NAT), a MATLAB toolbox for analysis and visualization of Hi-C data. Unlike previously available packages, we include methods capable of processing time series experiments [16] and data from abnormally karyotyped cells, which are common in cancer [113]. Additionally, since NAT is written in MATLAB, it can be readily extended to include customized analysis based on the questions of interest in any dataset.

4.6.2 Methods

We created NAT, an open-source MATLAB toolbox (https://github.com/ laseaman/ 4D_Nucleome _Analysis_Toolbox) for Hi-C data analysis including time series and karyotypically abnormal cell types. The toolbox includes functions necessary to 1) load Hi-C matrices into MATLAB, 2) normalize data using three different methods, 3) define TADs using three different methods, 4) visualize translocations, and 5) analyze time series data.

NAT can load Hi-C matrices produced by Homer into MATLAB [57]. Provided functions read data into MATLAB while simultaneously detecting chromosome boundaries. Loading data allows users to save data in .mat format, speeding up further analysis. Three different methods can then be used to normalize matrices (example in Load_Normalize.m). ICE normalization works under the assumption that genomic regions (bins) should have equal coverage and therefore the same total number of reads [55]. Toeplitz normalization reduces the diagonal dominance of Hi-C matrices to create an adjacency matrix for graph-based analysis [18]. Copy number normalization extends Toeplitz normalization for use on karyotypically abnormal cell lines by normalizing constant copy-number submatrices defined from Hi-C data [113].

After normalization, one of three different methods can be used to define TADs. TADs are genomic regions with many intra-domain interactions defined from Hi-C

matrices and are sub-regions of A/B compartments [31, 73]. First, the directionality index uses an HMM to define domains and boundaries [31]. Second, dynamic programming is used to define TADs whose scale can be easily adjusted though a tuning parameter [36]. Third, the iterative method is based on graph partitioning using network properties. TAD_methods.m includes examples of all three methods on chromosome 22 from human fibroblasts.

For cells with abnormal karyotypes, translocations can be visualized and analyzed (examples in TranslocationAnalysis). Sites of translocations are found by plotting the interchromosomal matrix at 100 kb and read resolution (Figure 4.6 A-B). The X pattern indicates the site of translocation and shows the interactions between normally separate chromosome arms. Combining portions of the chromosome 6 and 14 matrices creates a matrix representing the translocated chromosome [113] (Figure 4.6C). Additional visualization tools plot the chromosome contact matrices with TAD boundaries overlaid, the A/B compartments, and gene expression (Figure 4.6D).

NAT also includes functions to visualize time series structure and function data (Figure 4.6E) and analyze the dynamics of structure and function simultaneously using a phase plane (Figure 4.6F) at multiple genomic scales (chromosome, TAD, and gene level) [16]. In the phase plane, the X-axis represents structure and the Y-axis represents function. The Hi-C matrix and RNA-seq vector from a single time point in Figure 4.6E becomes a single point in Figure 4.6F. The x-coordinate shows the structure as is measured by Fiedler number or vonn Neumann entropy of Hi-C matrices, while the y-axis shows function as measured by gene expression. Figure 4.6F is a phase plane for Chromosome 22 using time series data from human fibroblasts (PhasePlane.m). PhasePlane.m also includes an example loading RNA-seq data and and converting it to bins, making it easy to compare to Hi-C data.

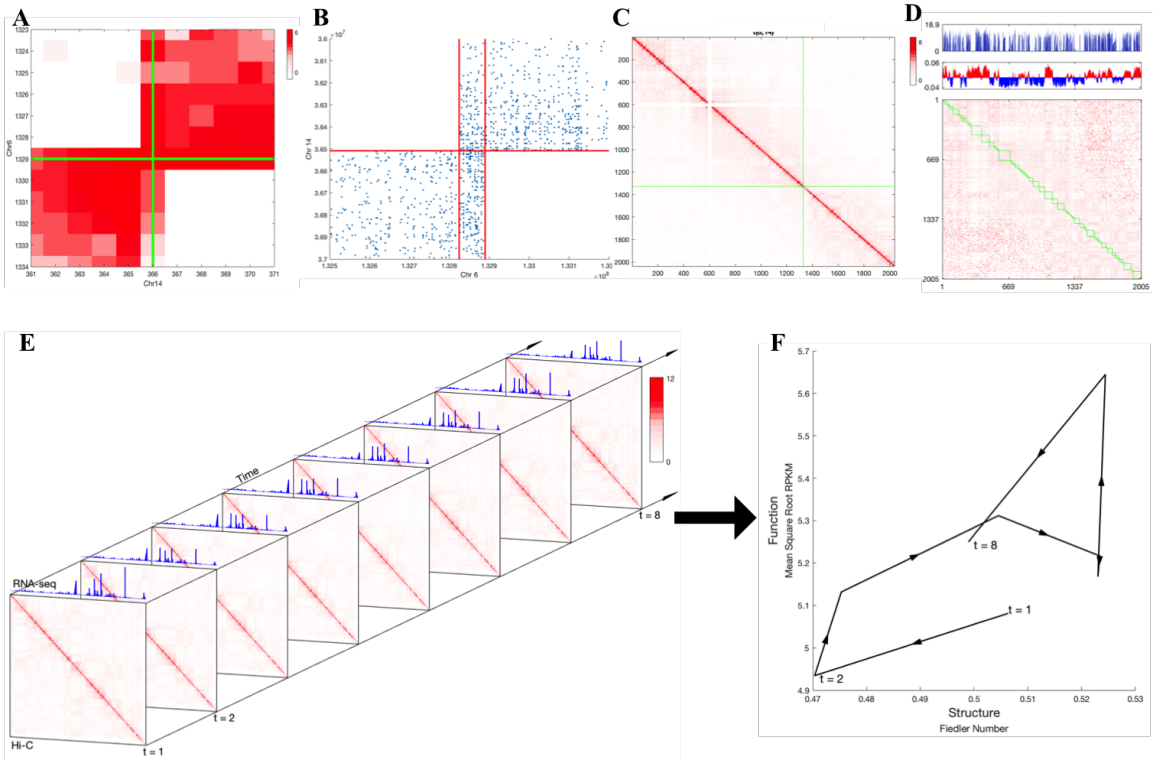

Figure 4.6: A) A portion of the interchromosomal matrix showing the interactions between chromosome 6 and 14 focused on a translocation in the colorectal cancer cell line HT-29 at 100 kb resolution. B) Read level data showing t(6; 14) at high resolution. Each dot represents a read pair and red lines indicate translocation breakpoints. C) The Hi-C matrix representing t(6; 14) at 100 kb resolution with green lines indicating the translocation sites. D) Hi-C matrix with TAD overlay (green boxes), A/B compartments (red and blue bars), and gene expression (top bars) for t(6; 14). E) Chromosome 22 time series Hi-C and RNA-seq data at 100kb resolution from a fibroblast time series [16]. F) Chromosome 22 phase plane showing how structure and function, measured with Fiedler number and gene expression, vary dynamically [16].

4.6.3 Conclusions

NAT provides easy to use and flexible functions to help with analysis of Hi-C datasets and other genomic features including RNA-seq. NAT can form a foundation for analyzing Hi-C time series data from any species or cell type, including those with altered karyotypes. The availability of robust and easy to use analysis tools, like NAT, helps move the field forward.

Acknowledgements

We would like to thank Scott Ronquist, Haiming Chen, and the winter 2017 Writing 630 class for feedback on the manuscript. Thank you to Scott Ronquist and Jie Chen for coding assistance. This work was supported in part by the Rackham Merit Fellowship program and the DARPA Biochronicity Program.

CHAPTER V

Cancer stem cell nucleome

5.1 Abstract

Cancer stem cells are a subpopulation of cancer cells with distinct properties that are thought to be particularly important to therapeutic resistance and metastasis. This chapter covers an ongoing project that extends the analysis of nuclear structure and function to cancer stem cells. These preliminary results show nucleome analysis of CSCs identifies CSC specific nuclear interactions illuminating how CSCs are distinct from the general cancer cell population.

5.2 Introduction

CSCs are a subpopulation of cancer cells with the abilities of stem cells including the ability to self-renew and differentiate eventually recreating a heterogeneous tumor composition [83, 136, 132]. Because of this, CSCs are believed to be responsible for metastasis which is the ultimate cause of 90% of cancer related deaths [81]. Additionally, many CSCs have the ability to become quiescent, i.e. enter a dormant state, allowing them to avoid therapeutics that target rapidly dividing cells [83]. Recent studies have also shown that a tumor's similarity to normal stem cells is predictive of outcome [102] further validating the importance of cancer stem cells to survival.

Previous work has used expression of specific markers, like ALDH1A1 or CD44 to distinguish a stem cell like subpopulation within a tumor sample [94, 15]. In this study we collected Hi-C and RNA-seq from subpopulations of SUM-159, a breast carcinoma cell line, selected for high and low expression of ALDH1A1, a marker of stemness. The development of chromosome conformation capture techniques, including Hi-C, provides unprecedented insights into spatial chromatin organization and long-range chromatin interactions in the interphase nucleus [73]. By measuring the nuclear structure in CSCs and genetically matched non-CSCs, we can directly compare the datasets and identify interactions that are unique to or missing in CSC nuclei.

5.3 Methods

5.3.1 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 non-CSC-like cellular populations. Hi-C and RNA-seq samples were collected and processed as described in [113].

5.3.2 Hi-C and RNA-seq processing

Hi-C and RNA-seq analysis were performed as described in Seaman et. al. [113]. Briefly, for RNA-seq, Tophat was used to align the reads to the reference transcriptome (HG19) with parameter settings: "–b2-very-sensitive", "–no-coverage-search", and "–no-novel-juncs" [122]. Cufflinks/Cuffdiff was used for expression quantification and differential expression analysis with parameter settings: "–multi-read-correct" and "–upper-quartile-norm" [123]. A locally developed R script using CummeRbund

was used to format the Cufflinks output [43]. Gene level analysis was performed using FPKM and $\log_2$ FC with pseudocounts, i.e. $\log_2$ FC $= \log_2(\text{CSC}+10^{-20}) - \log_2(\text{non-CSC}+10^{-20})$, for comparisons of samples and properties. Bin level gene expression vectors were calculated by adding up the raw counts for all the genes in each bin then normalizing by million reads to convert to fragments per million (FPM).

For Hi-C analysis, we used a standardized in house pipeline. Paired-end reads were mapped to the reference human genome (HG19) using Bowtie2 [68], with "-very-sensitive-local", which produced a sequence alignment map (SAM) formatted file for each member of the read pair (R1 and R2). HOMER was used to develop the contact matrix with "makeTagDirectory", "tbp 1". Then analyzeHiC is used with the "-raw" and "-res 1000000" settings to produce the raw contact matrix at 1Mb resolution, or with the "-res 100000" settings to produce contact matrix at 100kb resolution.

5.3.3 Normalization and TAD identification

Due to the abnormal karyotype of SUM-159, copy number based normalization [113] was used. For 1 Mb matrices, a threshold of 4000 was used on all chromosomes except 5, 9, 10, and 17 which used thresholds of 3600, 3000, 2000, and 1500, respectively. For 100 kb matrices, a threshold of 120 was used on all chromosomes except 1, 9, 16, and 17 and which used thresholds of 410, 70, 140 and 50, respectively. Interchromosomal matrices were normalized by dividing all entries by the mean of non-zero entries for that chromosome pair. TADs were defined using the iterative method described by Chen et. al. [18] with an algebraic connectivity threshold of 0.6. TADs are considered active if the sign of the Fiedler vector for that region is positive. The Fiedler vector is standardized to so that the correlation with gene expression is positive.

5.3.4 Quantification of structural changes

To find CSC specific genomic interactions, we identified regions where structure changes between CSC and non-CSC samples. First, the normalized 1 Mb whole genome non-CSC matrix was subtracted from the CSC matrix. The difference matrix was then filtered to reduce noise and help identify regions with large changes rather than single points that are likely to be outliers. For intra-chromosomal regions, any bin whose absolute value was above the threshold of 0.6 ($>$ 4 standard deviations above average) were selected as significantly changed (Figure 5.1). For inter-chromosomal regions, large structural variations like translocations and copy number increases were masked, then regions whose absolute value was above 4 were selected ($>$ 6 standard deviations above average). For statistical purposes, random sets of the same number of regions were chosen from across the genome $10,000$ times. During random selection the sizes of the regions and the numbers of intra- and inter-chromosomal interactions were kept consistent.

Transcription factor binding site (TFBS) were identified by scanning the genome for known motifs as previously described [105]. Briefly, TFBS were scanned across the genome using FIMO [45] to look for binding sites for transcription factors found in a number of databases [79, 106, 129, 103, 126, 114]. By looking for binding sites within 5 kb of gene transcription start sites, the genes that can be bound by each transcription factor were identified. A Hi-C interaction is bound by a transcription factor if at least one gene in each of the genomic regions that forms the interaction has a binding site for a given transcription factor.

The regions whose interactions changed were ranked according to a combination of their Hi-C, RNA-seq, long non-coding RNA (lncRNA), and TFBS as described in Table 5.1. Points were given for large changes in absolute as well as $\log_2$ fold change Hi-C and RNA-seq to balance the focus on large fold changes while wanting to avoid focusing on regions with small absolute expression (or Hi-C) levels. For a

region to get higher points for RNA-seq, the genes in both regions must change (as opposed to one changing while the other has no expression). Points were also given for having many TFBS as well as having binding sites for differentially expressed (DE) transcription factors since these are most likely to be biologically important to the differences between CSCs and normal cancer cells. Finally, points were awarded for regions that containing DE lncRNA since they are also thought to be important to CSC properties. Combining the points from all of these measures leads to a maximal possible score of 22.

5.3.5 Centrality and principle component analysis

Centrality measures how important or central each node is within a graph. A number of different measures for this have been developed 8 different types of which were calculated for the Hi-C data (see /refsup:cent for more information). For each node a total of 14 different centralities were calculated: eigenvector centrality, degree centrality, local Fiedler vector centrality, betweenness, closeness, local clustering coefficient centrality, $1-5$ hop walk centralities, and 3 distance to reference node centralities. These measures of centrality can be written as an matrix (nbins $\times$ nCentrality+1) when combined with the expression vector [74]. principle component analysis (PCA) was used to reduce the dimension of the data since many of the measure of centrality are correlated. The distance moved by each genomic bin was calculated as the Euclidean or straight line distance, $d = \sqrt{x^2 + y^2}$.

5.4 Results

5.4.1 Identifying changes in structure

We collected Hi-C and RNA-seq datasets from stem cell like (CSC) and non-stem cell like (non-CSC) breast cancer cells by sorting SUM-159 cells by their ALDH1A1

abundance. Analysis of RNA-seq data identified 1660 genes differentially expressed between CSC-like and non-CSC SUM-159 samples.

Overall, the Hi-C matrices appear very similar and the same translocations and copy number changes are visible as is expected since the samples are genetically identical (Figure S38). However, there are some interactions that are visible in only one of the samples (Figure 5.1 A-B). In order to identify these regions systematically genome wide, we first subtracted the normalized 1 Mb Hi-C non-CSC matrix from the CSC matrix (Figure 5.1C). We next median filtered the difference then selected regions above a threshold as regions whose interactions change (Figure 5.1 D-E).

This process identified 99 intra-chromosomal and 128 inter-chromosomal regions that change interaction between CSC and non-CSC nuclei. Figure 5.2 shows where the regions occur, which sample had the stronger interaction, and how the gene expression of the region changed. Of the 227 regions, 183 have stronger interaction in CSC than non-CSC while 44 have weaker interactions. 45 and 47 regions have increased and decreased expression in CSC relative to non-CSC, respectively. Chromosomes $20, 21$, and 2 are the most over-represented while chromosomes 4 and 14 are the most under-represented.

In order to further characterize these regions, we used an iterative method to define TADs [18] and characterized their activity level based on the sign of the Fiedler vector. 49% of changing regions are in active TADs, 43% are in inactive TADs, and 7% of the changing regions split TAD boundaries. There are statistically significantly fewer regions that split a boundary than expected if the regions were randomly distributed across the genome. This is consistent with the idea that TADs have increased interactions within the TAD and long-range interactions are likely to be between pairs of TADs. Additionally, the 217 regions whose interactions change include 39 DE lncRNA genes and 233 other DE genes which is significantly more than are found in random regions of the same size and number ($p \leq 0.001$ and 0.031 respectively).

In order to determine which regions change the most and are most biologically interesting, we developed a ranking system that scores each region based on the size of the change in Hi-C and RNA-seq as well as the presence of shared TFBS and DE lncRNAs. The maximum possible score was 22 while the average actually seen was 7. A total of 6 regions had a score greater than 15 and they are marked with stars around the perimeter of Figure 5.2. Of these regions, 5 have increasing Hi-C and RNA-seq while the sixth has decreasing Hi-C and RNA-seq. Interestingly, 4 of the top 6 regions involve either the beginning of chromosome 6 or the end of chromosome 11, including the top region which involves both. The region involving chromosome 11 and contains the cell cycle gene *ATM*. The region on chromosome 6 contains many DE histone genes (*HIST1H2AG, HIST1H2AH, HIST1H2AI, HIST1H2AJ, HIST1H2AK, HIST1H2AL, HIST1H2AM, HIST1H2BJ, HIST1H2BK, HIST1H2BL, HIST1H2BM, HIST1H2BN, HIST1H2BO, HIST1H3H, HIST1H3J, HIST1H*). Histone genes have previously been shown to be important for stem cell properties in cancer [61], making their change in structure potentially biologically important. Additional work will explore how these regions and their interactions are important for CSC properties.

5.4.2 Changes in centrality

Centrality is used to determine how important or central to a graph a node in a network is (See Appendix A.7.5). In the case of Hi-C data, we used several different measures of centrality to determine how well connected each 1 Mb bin is in the network. The centrality measures were combined with gene expression and used in PCA to reduce the dimension of the data while maintaining as much of the variability as possible. Projections of each bin onto the first two PCs are show in Figure 5.3.

In order to quantify which genomic regions change most, the distance between a bin's projection in PC space for the CSC and non-CSC samples was calculated. The bins whose centrality changed the most are highlighted with green lines in Figure

5.3. Of these 11 bins, 8 of them were also selected as having changing interactions using the filter and threshold based method described above. This indicates that although these methods take very different approaches to finding changes in genomic structure, both successfully find large changes since they have such a strong overlap ($p < 0.0002$).

Further validating this finding, two of the bins whose centrality changes most are on the end of chromosome 11 in the region whose interaction with chromosome 6 changed more than any other. This suggests that further work should explore the importance of cell cycle in distinguishing CSCs from non-CSCs and in particular focus on *ATM*'s role in the process. Previous work has shown that CSCs can enter a quiescent or not actively dividing state which is part of what makes them resistant to traditional cancer therapies like chemotherapy [83].

Another interesting region in the list of those whose centrality changes a lot are two bins on chromosome 8 that contain a number of genes known to be related to cancer including *MYC, POU5F1B, PCAT1*, and *PVT1*. The presence of *MYC* and *POU5F1B* among these are particularly interesting as they have been related to the function of stem cells in a healthy population. *MYC* is one of the four transcription factors shown to cause healthy differentiated cells to revert to a stem cell state [120]. *POU5F1B* is a pseudogene for *OCT4*. *OCT4* is another of the four stem cell reprogramming transcription factors. Additionally, the pseudogene *POU5F1B* has been shown to be amplified and expressed in some gastric cancers [50]. As reflected by selection of this bin as one whose centrality changes significantly, the degree of the bin changes significantly more than other regions in the genome using either the change in the normalized Hi-C matrix or the $\log_2$ fold change ($p \leq 0.0172$ and 0.0041 respectively).

5.5 Discussion

By sorting SUM-159 cells based on their expression of *ALDH1A1*, we were able to collect Hi-C and RNA-seq from CSC-like and non-CSC samples of a breast cancer cell line. Analysis of RNA-seq data identified 1660 genes differentially expressed between CSC-like and non-CSC SUM-159 samples. Analysis of Hi-C data identified 217 interactions that changes significantly between CSC and non-CSC samples as well as 11 genomic regions whose centrality changed significantly.

Within these regions are a number of genes whose function and relationship to stem cell properties and cancer stem cells in particular needs to be further explored. A region on chromosome 6 that contains 16 histone genes was found to have a changing interaction with a region on chromosome 11 that contains the cell cycle checkpoint gene *ATM*. *ATM* was also found in one of the regions whose centrality changed a lot between the CSC and non-CSC samples. Also in regions whose centrality changed the most were two bins on chromosome 8 that contain 4 genes previously associated with cancer (*MYC, POU5F1B, PCAT1*, and *PVT1*), the first two of which are known to be related to stem cell properties in normal cells including the ability to reprogram healthy cells into stem cells.

This work shows that there are important differences between the genomic structure and function of cancer stem cells and cancer cells. Further work will help elucidate the role of these differences in the CSC phenotype.

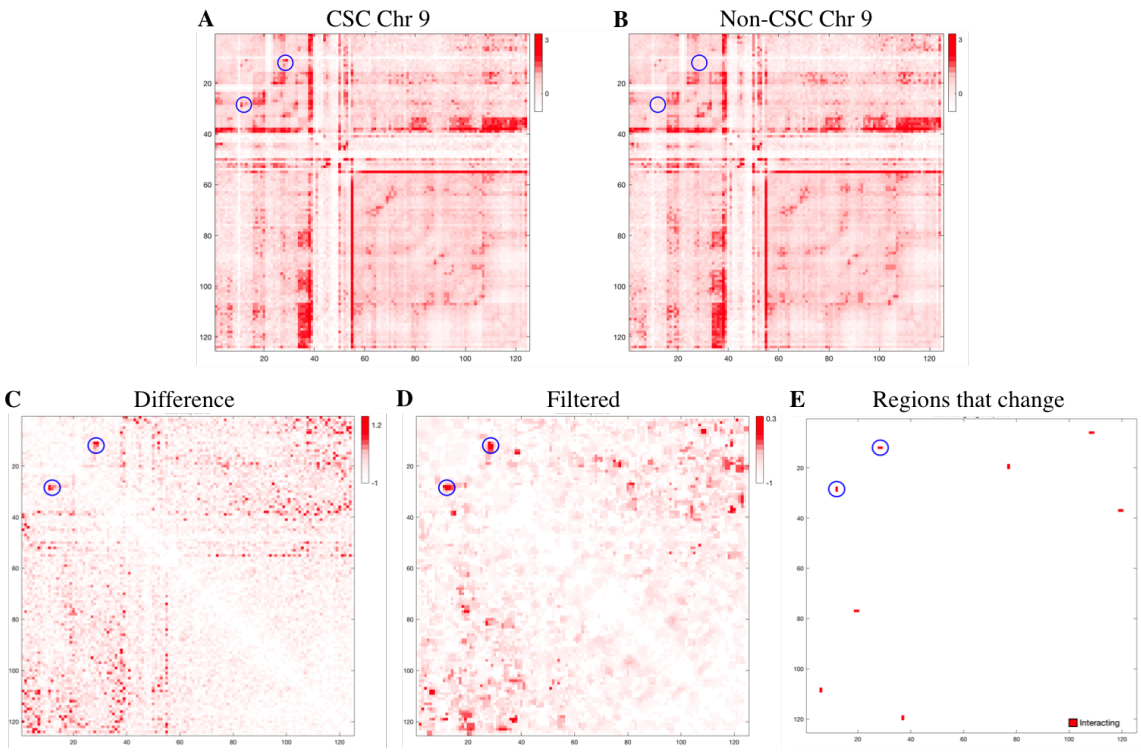

Figure 5.1: Selection of changing regions. The normalized 1 Mb resolution Hi-C matrix of chromosome 9 from the A) CSC and B) non-CSC samples. The blue circles show an interaction present in the CSC sample that is not present in the non-CSC sample. Regions that change interactions were found by C) subtracting the normalized matrices, D) median filtering the matrix to remove noisy changes, then E) selecting regions above the threshold.

Table 5.1: Thresholds used for scoring regions whose Hi-C interactions changed. Each row indicates a different category that was scored. 0 points were given if the actual number for a region falls below the first value listed in the thresholds column, 1 point was given if it is between the first and second values, continuing up to the maximal points if the true value was larger than the last listed threshold.

Max Points	Thresholds	Category
4	Intra: $.6, 1.5, 3.5, 7$ Inter: $8, 25, 60, 120$	Absolute change in Hi-C
4	$.6, .75, .9, 1.4$	$\log_2$ FC Hi-C
2	$80, 1200$	Absolute change in RNA-seq
2	$0.04, 0.25$	$\log_2$ FC RNA-seq
2	$1, 30$	DE TFBS
2	$1, 300$	total TFBS
2	$1, 2$	DE lncRNA

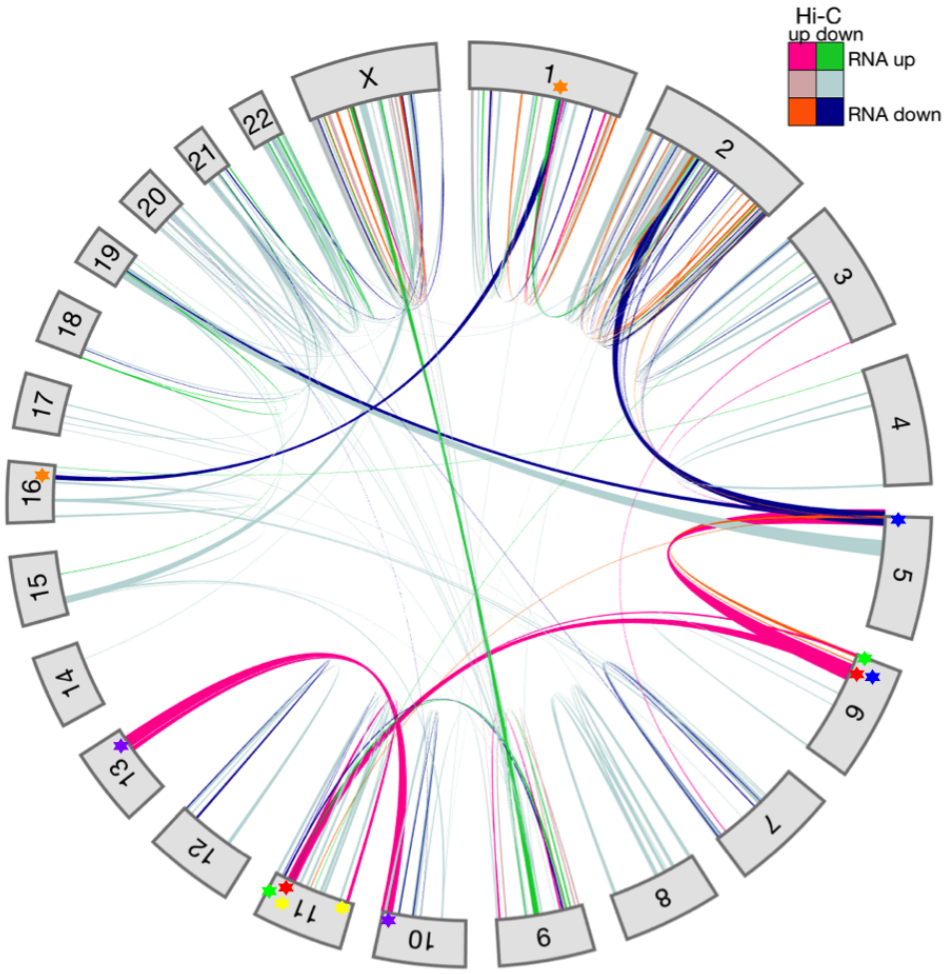

Figure 5.2: Regions that change. The outer circle shows the 22 chromosomes analyzed and each line represents an interaction that changed between the CSC and non-CSC samples. The thickness indicates the magnitude of the change while the color indicates weather the interaction was up or down in CSC relative to non-CSC and if RNA-seq is up, minimally changing, or down in CSC relative to non-CSC. Stars mark the interactions of interest based on a high score

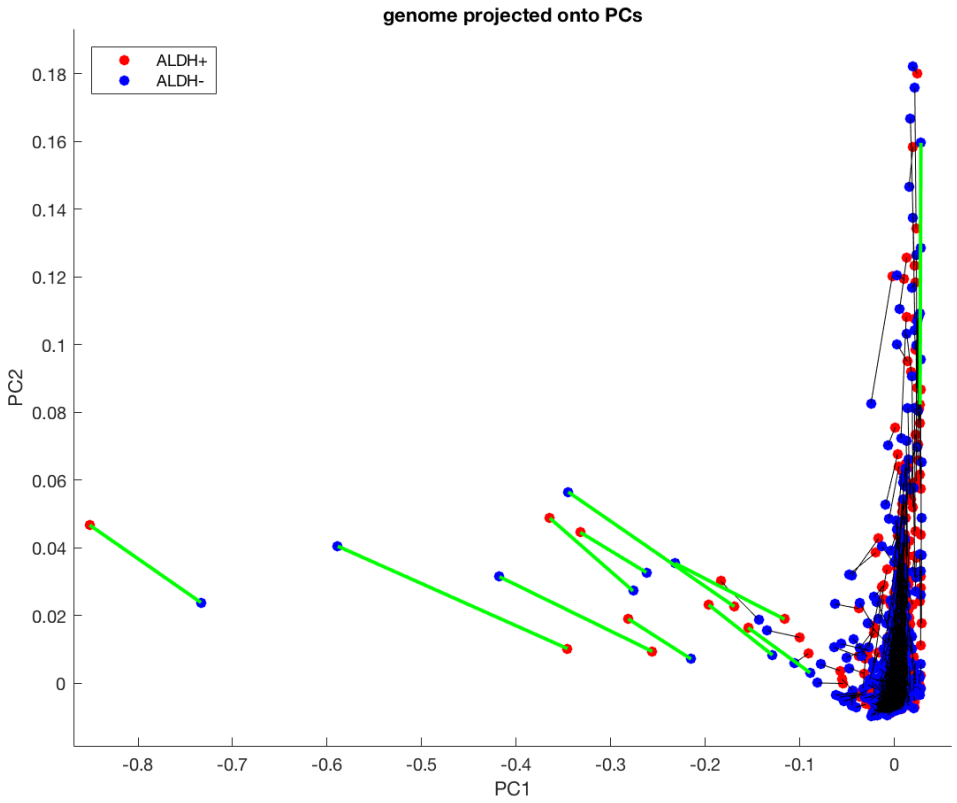

Figure 5.3: Centrality and Principle Component Analysis of the whole genome. D) Projection of all 1 Mb bins onto first two PCs calculated from a feature matrix measuring centrality and gene expression of all bins in the genome. Black lines connect the same bin in the two different samples. Green lines mark the 11 bins with the largest change in centrality.

CHAPTER VI

Allele specific structure and function

6.1 Abstract

Every normal human somatic cell contains two copies of each chromosome folded inside the nucleus yet almost all studies of genomic folding ignore this, analyzing the combination of the chromosomes instead. The presence of single nucleotide differences between the chromosomes, called SNPs, allows assignment of reads to the maternal or paternal copy of each chromosome. To identify monoallelic structures and gene expression within the nucleus, we collected Hi-C, RNA-seq, and Bru-seq from B-lymphocytes in which all of the SNPs had been sequenced and phased. This chapter presents preliminary analysis of this data focusing on characterizing monoallelic expression measured with RNA-seq and Bru-seq.

6.2 Introduction

Microscopic studies of the interphase nucleus reveal that individual chromosomes are spatially confined in chromosome territories [25]. Hi-C analysis suggests that chromosomes compartmentalize into regions of euchromatin and heterochromatin [73], and further organize into TADs that are cell type invariant and conserved in vertebrates [31].

One limitation of most Hi-C analysis is that it is unable to detect which copy of a chromosome reads come from. Because of this, most currently available Hi-C matrices represent chromatin interactions from both homologous chromosomes. It is not clear whether both homologues have the same spatial conformation, or contribute equally to functional output measured by the abundance of gene transcripts.

Alleles of parental autosomal genes presumably contribute compatibly to the abundance of each correspondent transcript in human diploid cells. However, there are exceptions to this assumption leading to monoallelic expression (MAE). One of the exceptions is the MAE of X-linked genes due to inactivation of one of the X chromosomes in females [133]. The mechanism of X inactivation is known to be driven by the transcript of *XIST* exclusively expressed from the inactivated chromosome [12, 11]. Another exception is the discovery of genomic imprinting that sets forward a clear mechanism of parental-specific and epigenetically inheritable MAE [100]. There are a total of 215 human genes experimentally detected or predicted to be expressed under the imprinting mechanism [5, 6, 107]. In addition, genome wide transcriptomics analyses reveal that approximately 20% of human genes experience MAE [85, 109]. MAE appears to be stochastic and independent of parental origin in single cells [10, 29].

Although a large number of genes have been identified or predicted as MAE, there is not any comprehensive analysis of such genes throughout the cell cycle. A proliferating cell goes through the several stages of the cell cycle during which its genome is replicated and divided into two daughter cells. It is currently not clear how is MAE maintained in the cell cycle phases G1, S, and G2/M when the cell is growing, replicating DNA, and dividing, respectively. While genetic and epigenetic mechanisms have been identified as underlying mechanisms in controlling MAE [62, 85, 89], it is also important to explore the role of three dimensional chromatin organization in controlling of gene expression including MAE.

To see if regions with MAE also have allele specific structures and interactions, we collected data about genomic structure using Hi-C, gene expression using RNA-seq, and nascent gene expression using Bru-seq. Bru-seq is a technique for identifying recently transcribed sequences by tagging new transcripts with bromouridine then isolating the new transcripts from the wider population before sequencing [90]. Here I present preliminary analysis of the RNA-seq and Bru-seq data including detailing the bioinformatics methods developed for analysis.

6.3 Methods

6.3.1 Experimental methods

All experiments use cells at cell cycle phases G1, S, and G2/M. We grow the NA12878 cells for live cell flow cytometry sorting to obtain cell fractions at these phases. Fractions of the sorted live cells are used for RNA-seq and acBru-seq analyses. Subsequently, RNA-seq library construction was carried out in the sequencing core facility, and sequence reads of 50-base in length were generated on an Illumina HiSeq 2500 station.

For Bru-seq, we performed 5′-bromouridine incorporation in live cells for 30 minutes, and the bromouridine-labeled cells were then subjected to flow cytometry sorting to isolated G1, S, and G2/M phase cells. We isolated total RNA for bromouridine-labeled transcripts pulldown with an anti-bromouridine antibody [90], and generated sequence reads at 125-base length.

6.3.2 Allele specific RNA-seq and Bru-seq methods

The pipeline developed for estimating allele specific expression from RNA-seq and Bru-seq is outlined in Figure 6.1. The left side shows the normal flow of a non-allele specific RNA-seq or Bru-seq pipeline which is combined at the end with results from

the allele specific portion of the pipeline to get abundance estimates for the maternal and paternal copy of each gene.

The normal RNA-seq and Bru-seq analysis were performed as previously described [113, 90]. As shown on the left side of Figure 6.1, Bru-seq reads were aligned using Tophat (v1.3.2) without de novo slice junction calling after checking quality with FastQC. A custom gene annotation file was used in which introns are included but preference to overlapping genes is given on the basis of exon locations and stranding where possible (See [90] for full details). Similarly, in the RNA-seq data processing, the raw reads were checked with FastQC (version 0.10.1). Tophat (version 2.0.11) and Bowtie (version 2.1.0.0) were used to align the reads to the reference transcriptome (HG19). Cufflinks/Cuffdiff (version 2.2.1) was used for expression quantification and differential expression analysis, using UCSC hg19.fa and hg19.gtf as the reference genome and transcriptome. A locally developed R script using CummeRbund was used to format the Cufflinks output.

GSNAP was used to align reads without biasing against SNP positions for the allele specific portion Bru-seq and RNA-seq data analysis as indicated along the right column of Figure 6.1. The gene annotation file was used to create the files for mapping to splice sites, with the $-$s option. Optional inputs to perform SNP aware alignment were included. Specifically, $-$v was used to include the list of heterozygous SNPs (ftp://platgene_ro@ussd-ftp.illumina.com/2016-1.0/hg19/ small_variants/ NA12878/ NA12878_variants/NA12878/NA12878.vcf.gz) and –use $-$ sarray $=$ 0 was used to prevent bias against non-reference alleles.

After alignment, the output SAM files were converted to binary sequence alignment map (BAM) files, sorted and indexed using SAMTOOLs [72]. The number of each base that were observed at each of the heterozygous SNP locations was quantified using bam-readcounter (D. Larson et. al., https://github.com/genome/bam-readcount). The statistical significance of allele specific expression for each SNP was

then quantified by a binomial test with a null probability of 0.5.

Gene allele specificity was then assessed by combining all of the SNPs in each gene. For RNA-seq the total number of maternal and paternal alleles were calculated by adding up all of the allele of each type in all exonic SNPs within a gene. For Bru-seq exonic SNPs were counted first, and then non-exonic SNPs were counted if they were in a gene's intronic region. Paternal and maternal abundance of each gene were calculated by multiplying the overall abundance estimate by the fraction of the SNP-covering reads that were paternal and maternal, respectively. Genes with less than 5 reads containing SNPs per sample were not used for allele specific estimation of expression or testing of MAE since they do not have enough data to estimate the allele specific abundances accurately.

Gene level significance of MAE for each gene in each cell cycle stage was evaluated using a negative binomial model. Variance estimation is improved through a local regression relating variance to the mean (https://www.mathworks.com/help/bioinfo/ref/ nbintest.html). The same method was used to determine differential gene expression between the overall abundance in different cell cycle stages. ANOVA was used on the $\log_2$ FPKM values to determine what genes changed over the cell cycles as well as between maternal and paternal alleles.

RNA-seq and Bru-seq were binned into 100 kb and 1 Mb bins to match the resolution of the Hi-C data. This was done separately using the maternal and paternal expression estimates by adding the expression of the genes in a bin and when necessary dividing a genes counts according the proportion of the bin in each gene. About 75% of genes could not be assessed for allele specific expression due to a lack of SNPs in the gene body. When binning RNA-seq and Bru-seq, an assumption of 50% maternal and paternal expression was made for genes lacking SNPs to avoid losing that data.

6.3.3 TAD analysis

In order to test if there is any statistical evidence that genes in the same TAD, tend to express the same allele, genes were randomly generated to have paternal, maternal, or biallelic expression with the same frequencies observed in the data. The genes were then grouped into TADs to match the distribution of genes per TAD. The number of real TADs with strong paternal expression was compared to the number of random TADs with strong paternal expression. A TAD was considered strongly paternally expressed if over 90% of the genes in the TAD were paternally expressed. The same was done for maternal and biallelic expression of TADs as well. TADs on chromosome X were excluded due to the strong preference for all of genes on chromosome X to show strong paternal expression regardless of TAD.

6.4 Results

6.4.1 Allele specific RNA-seq

Differential expression analysis of the FPKM normalized RNA-seq data with the software package edgeR [103] identified 480 genes that were differentially expressed between G1, S, or G2/M (FDR < 0.05) regardless of MAE. Functional annotation of the 480 genes showed that genes were significantly enriched under GO terms exclusively related to the cell cycle. Since there is no other perturbation to the cells except cell cycle based sorting, it is expected that the changes in gene expression between the cell cycle phases are cell cycle related. The cell cycle stage-specific expression of correspondent genes confirms that the cells isolated from flow cytometry sorting are indeed in G1, S, and G2/M (Figure 6.2A). For the comparison between maternal and paternal alleles, only $5,080$ genes in which maternal and paternal expression could be separately estimated were included. For comparing cell cycle stages regardless of MAE, all $19,267$ genes were used. The results are summarized in the figure 6.2B-C

(FDR < 0.05).

As an alternative approach, we also performed two-way ANOVA on the $\log_2$ FPKM values. Using ANOVA with an FDR < 0.05 cut off, $1,762$ genes showed a significant change in expression between the maternal and paternal alleles (914 paternal higher, 848 maternal higher). Additionally, $1,789$ genes showed a significant change across the cell cycle stages (regardless of MAE). In total, $2,838$ genes showed significant change with either allele or cell cycle stage, and 713 genes showed significant change with both. The ANOVA test is far less conservative than the negative binomial based test which could be counteracted through a more stringent FDR threshold. The chromosome distribution of the $1,762$ MAE genes identified from the ANOVA is shown in Figure 6.2D.

6.4.2 Allele specific Bru-seq

Bru-seq is a technique that gives a short-term view of active progressive gene transcription [90]. Bru-seq results were obtained for NA12878 cells at G1, S, and G2/M in collaboration with Dr. Ljungman. Non-allelic pair-wise comparisons, S vs G1, G2/M vs G1, and G2/M vs S revealed large numbers of genes significantly changed nascent transcription. In the S vs G1 comparison, 568 significant genes were identified (FDR < 0.05) among which 492 genes increased expression and 76 decreased expression levels in S phase. The G2/M vs G1 comparison resulted in 417 significant genes (FDR < 0.05), of which 348 genes were up-regulated, and 69 were down-regulated in G2/M phase. In the G2/M vs S comparison, we identified 34 significant genes, of which only one (*CCNB1*) was upregulated and rest were down-regulated in G2/M phase. As expected, DAVID analysis shows that many of the genes are related to changes in cell cycle stage.

Analysis of allele specific Bru-seq results showed that because Bru-seq includes introns, many more SNPs have some coverage from the reads sequenced. In fact,

there were 266,899 SNPs with at least 5 reads across all of the samples in the Bru-seq data while there were only 65,676 SNPs in RNA-seq data. Despite this, many of the SNPs have too low read depth to be statistically evaluated. Figure 6.3A shows that Bru-seq has far more SNPs with very low coverage.

When we look only at SNPs with at least 5 reads per sample, then RNA-seq and Bru-seq had very similar numbers, with $19,394$ and $19,998$ SNPs, respectively. Because of this, the number of genes with enough reads to analyze allele specific expression (> 5 reads) for Bru-seq is only slightly larger than for RNA-seq with $6,168$ and $5,065$ genes, respectively.

A single gene, *MTRNR2L2*, was a strong outlier and was removed from analysis. The correlation matrix shown in figure 6.3 is calculated without the outlier showing that when the outlier is removed the replicates and samples are all highly correlated as expected. The first replicate, especially in the G2 sample, is different from the other replicates. Statistical tests were run with and without the replicate but there were minimal differences in gene level significance. The reduction in statistical power from losing a replicate was comparable to that gained from the reduced variance. With all replicates included, two way ANOVA identified 393 genes with MAE and 247 genes whose expression changed through the cell cycle.

Pairwise comparisons (G1 maternal versus paternal for example) using the negative binomial test introduced in the RNA-seq analysis revealed a very small number of genes that changed (≤ 10 for each pair). This is due to the increased variability between replicates in the Bru-seq data compared to the RNA-seq data. As a result, for the purpose of comparing Bru-seq to RNA-seq or Hi-C within a cell cycle stage, another method of identifying genes with allele specific expression was performed on the Bru-seq data. For this purpose we subtracted the expression values of each gene in the maternal sample from those in the paternal sample and created a histogram of the amount of change for each gene. We then selected a threshold above which the

genes were considered to have large differences in allelic expression. The threshold was selected by fitting the histogram of genes with differences less than 4 (cutting off the extreme values in the tail) to a exponential. The exponential was then used to calculate the theoretical 95% cutoff which was used as the threshold. This method identified 565 genes with allele specific expression in G1, 594 with MAE in S, and 610 with MAE in G2.

6.4.3 Location based consistency in MAE

Previous work has shown that there is no preference for genes on a single chromosome to have the same allele expressed except for the well studied chromosome X inactivation [42]. To test if there was any preference for genes on the same chromosome to have the same parental expressed allele in our dataset, a binomial test was run to see if the percent of paternally expressed genes was significantly different than 50%. Figure 6.4A shows the number of maternal, paternal, and biallelic genes on each chromosome. Chromosomes 5 and 22, were significant at a nominal p-value ≤ 0.05 indicating a slight allele imbalance. The three stars above chromosome X mean that the result was highly significant with $p \leq 10^{-6}$ as is expected for chromosome X based on its inactivation.

Next, we tested if there is any statistical evidence that genes in the same TAD, tend to express the same allele. There are a total 1056 TADs with more than a single gene (out of 2347). Figure 6.4B shows the number of genes with maternal, paternal, or biallelic expression in each TAD along chromosome 22 (those in which MAE could not be evaluated are not included).

To test for grouping genes with the same expressed allele, random sets of genes were grouped into TADs and the number of TADs in which more than 90% of the genes expressed a single allele was counted. The histogram in Figure 6.4 show the number of TADs with strong paternal, maternal, or biallelic expression in red lines

as well as the random distribution. TADs on chromosome X were excluded due to the strong allelic preference for all of the genes on the chromosome. There is no evidence of a tendency for genes with MAE to be clustered in TADs with other similarly MAE genes. Interestingly, TADs containing genes with biallelic expression are statistically significantly more likely to contain almost purely genes with biallelic expression ($p \leq 0.0005$).

The same analysis was repeated using cell cycle specific definitions of monoallelic expression. As described above, this was determined through a negative binomial test comparing maternal and paternal expression of each gene. This test is more conservative, and identifies less genes as MAE, thus smaller number of TADs experience purely paternal or maternal expression. A single TAD had strongly maternal expression in S and G2/M while no other TADs had strong monoallelic expression in any cell cycle stage. During each cell cycle stage between 915 and 925 TADs had strong biallelic expression all of which are strongly within the expected range ($p \geq 0.40$). This shows that there is no evidence that MAE clusters within TADs.There are only 162 and 73 genes that show paternal and maternal expression, respectively, making it very unlikely that they are clustered in the same TAD.

6.5 Discussion

Studying allele specific structure and function within the nucleus will be critical to forming a deep understanding of the principles guiding genomic organization. This work will be combined analysis of allele specific genomic interactions to help understand the differences between the two copies of each chromosome.

$1,762$ genes with MAE during the cell cycle were identified as well as $1,789$ genes that change expression through the cell cycle using RNA-seq data. Bru-seq identified 393 genes whose nascent expression was stronger for a single allele and 247 genes whose nascent transcription changed through the cell cycle stages. Consistent with

previous work exploring allele specific gene expression, this indicates that there are widespread differences in allelic expression beyond the well studied chromosome X inactivation.

Further work will identify genomic regions with different structures between the two copies of each chromosome. Once these regions are identified we can compare the structural changes to the functional changes to identify regions of overlapping or diverging difference. We will also further explore how these allele specific structures and functions change through the cell cycle.

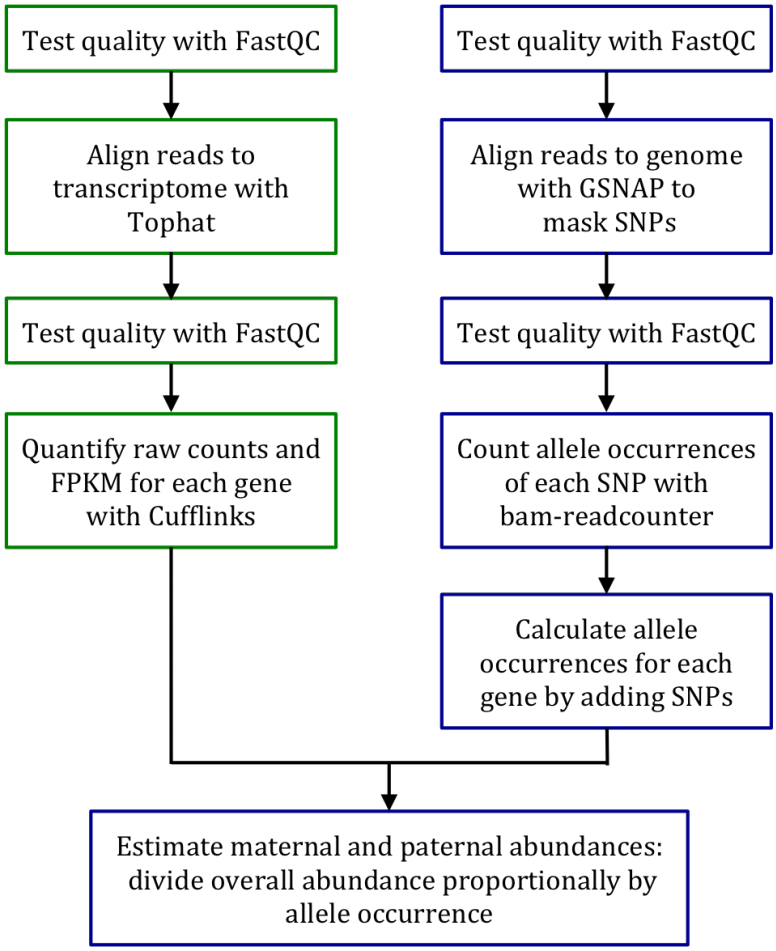

Figure 6.1: Bioinformatic processing pipeline for analyzing MAE in RNA-seq and Bru-seq datasets. The green boxes indicate the steps involved in standard non-allele specific analysis of RNA-seq or Bru-seq data including alignment and gene expression quantification. Blue boxes indicate the steps that are unique to allele specific analysis and estimation of maternal and paternal specific expression levels of each gene.

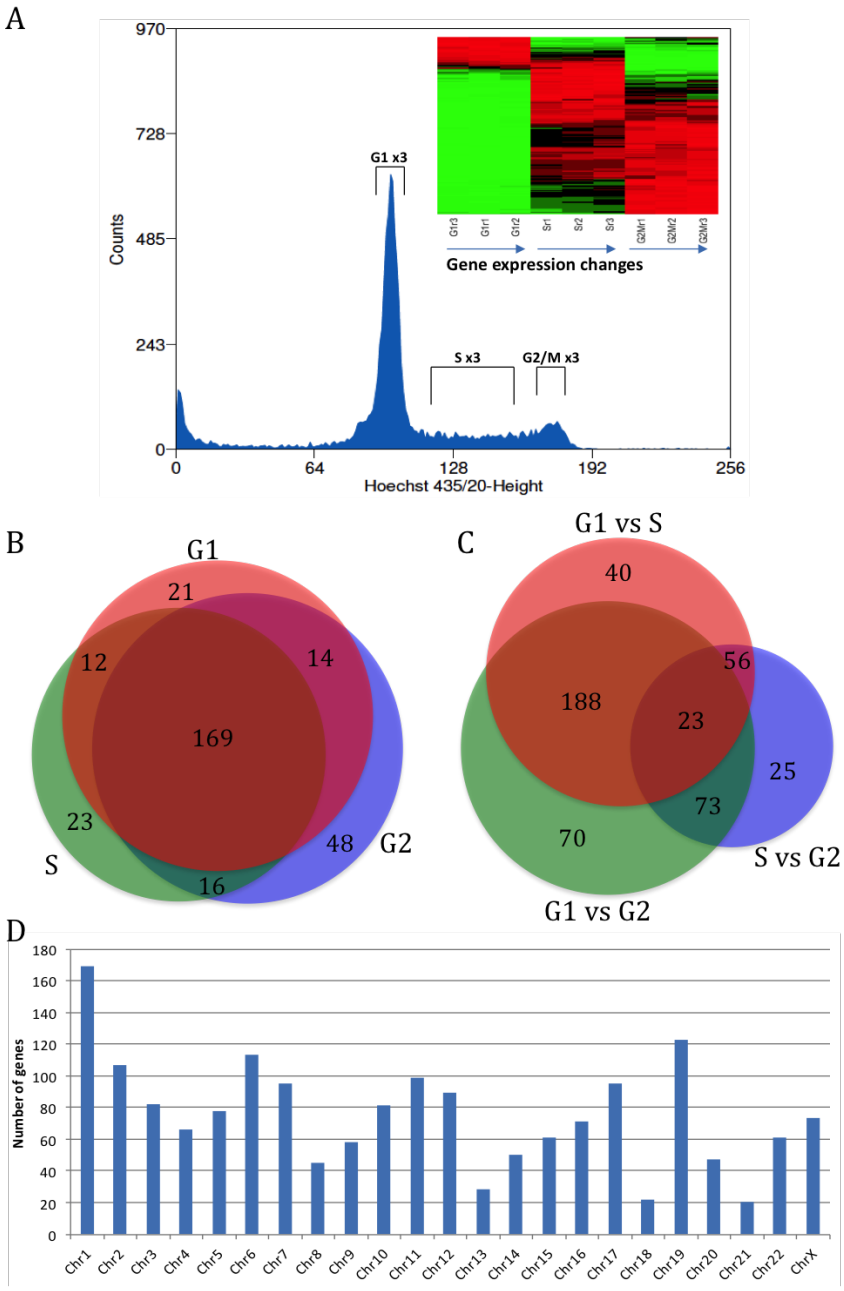

Figure 6.2: MAE of RNA through the cell cycle. A) Flow cytometry enrichment of G1, S, and G2/M cells (blue regions) and correspondent gene expression changes (insert). B) Venn diagram indicating the number of maternally and paternally expressed genes for each sample. C) Venn diagram indicating the number of genes differentially expressed between each pair of cell cycle stages. D) The number of genes on each chromosome with MAE.

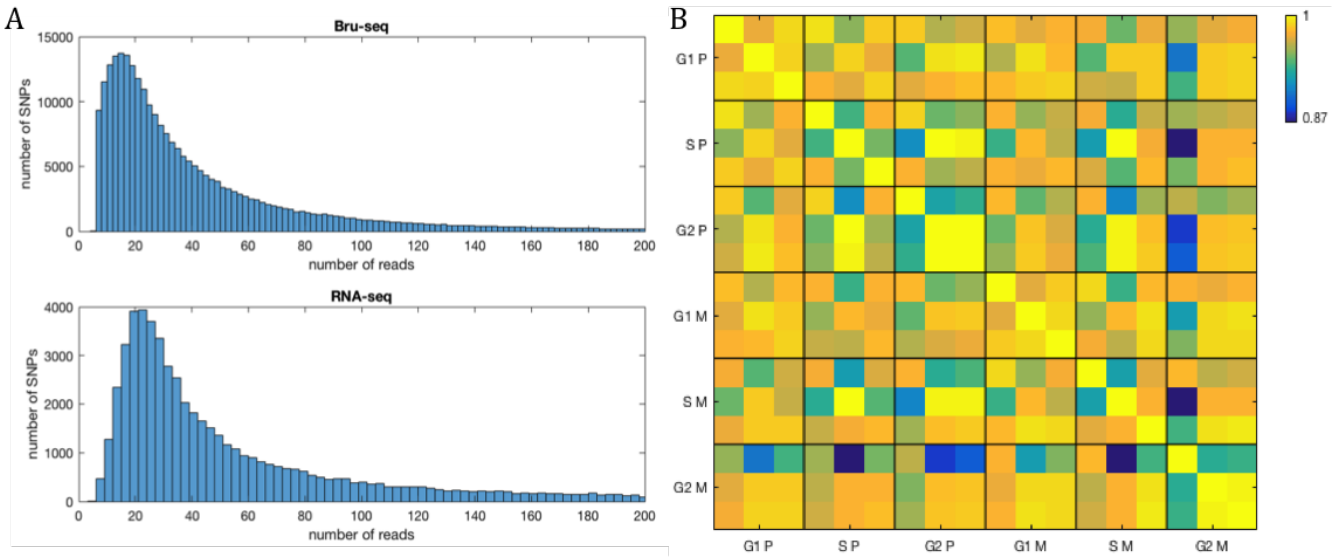

Figure 6.3: Monoallelic nascent expression through the cell cycle measured with Bru-seq. A) Histograms of the number of reads spanning a SNP for all of the SNPs with at least 5 reads for Bru-seq (top) and RNA-seq (bottom) shows that Bru-seq has many more SNPs with small numbers of reads. B) The correlation between gene expression replicates and samples shows high overall correlation between the samples with some replicates that are less similar than others.

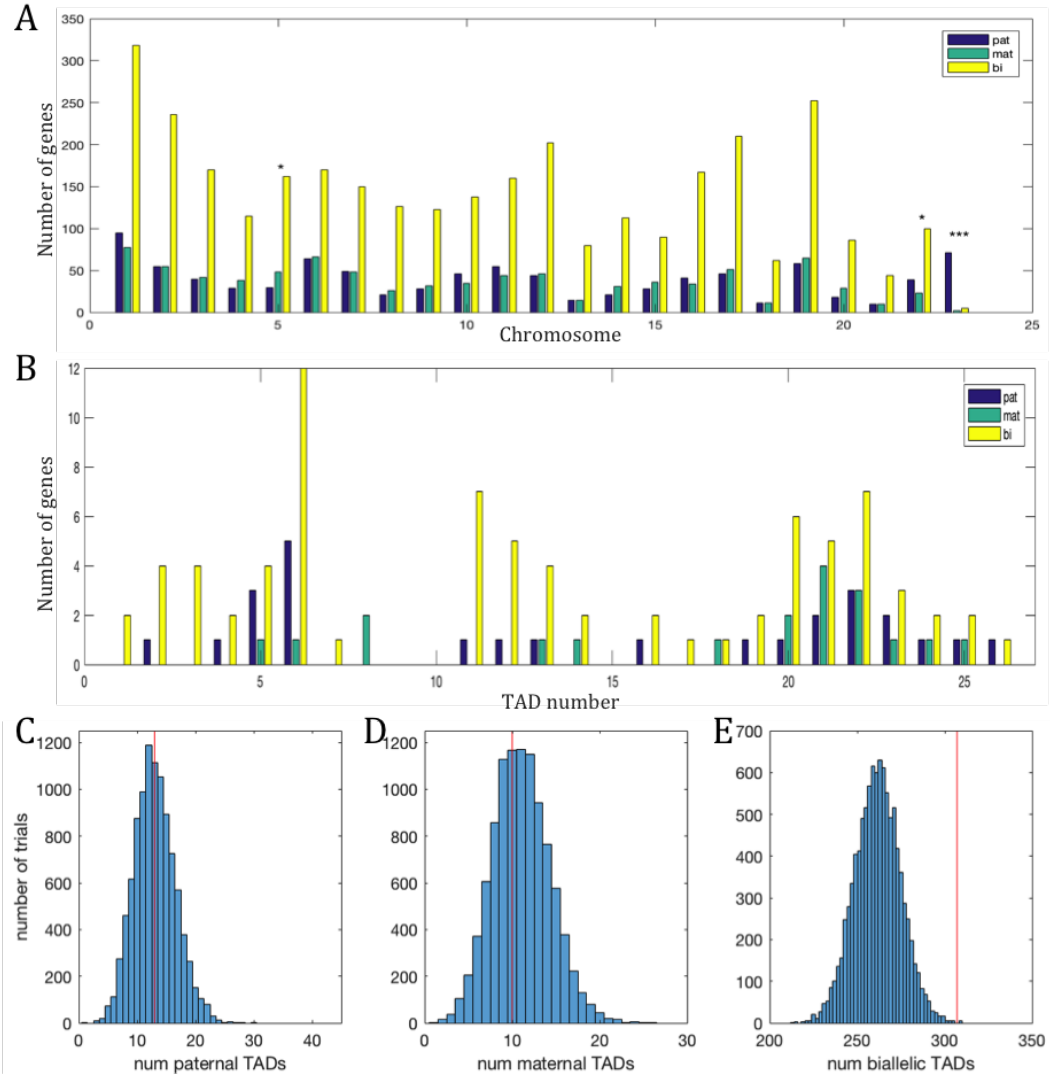

Figure 6.4: Allelic consistency. A) A histogram showing that most chromosomes have similar numbers of maternally expressed genes (blue) and paternally expressed genes (green) as well as far more genes with biallelic expression (yellow). *** indicates that chromosome X has a highly significant tendency towards paternally expressed genes ($p \leq 10^{-6}$). * indicates that chromosomes 5 and 21 have a slight bias towards a single allele ($p \leq 0.05$). B) A histogram of the number of genes with paternal (blue), maternal (green), and biallelic (yellow) expression in each TAD along chromosome 22. The observed number of TADs in which at least 90% of genes showed C) paternal, D) maternal, or E) biallelic expression compared to the expected number. The number of TADs with many genes with biallelic expression is more than expected ($p \leq 0.0005$).

CHAPTER VII

Concluding Remarks

As the second most common cause of death in the United States [1], understanding and identifying treatments for cancer is a high priority with the biomedical research community. This dissertation aims to further that goal by applying new technologies for understanding genome organization to cancer cells. A number of challenges face the field some of which were addressed here. The goal is to help understand how genetic changes in cancer lead to the cancer phenotype through changes in the 4D nucleome.

One such challenge is that quantitative rigor generally was not used when studying nuclear morphology as a diagnostic for cancer or other genetic diseases [59, 21, 27]. In Chapter II, nuclear shape was measured for cell-cycle synchronized primary human fibroblasts from six different individuals at 32 time points over a 75 hour period. An algorithm was developed to calculate the dimensions of an approximating ellipsoid for each nucleus and used to determine what periodicities were present in the dataset. Two prominent frequencies were found: a 17 hour period consistent with the cell cycle of these cells and a 26 hour period that might be related to the natural circadian rhythms of the cells. The work shows that the shape of the nucleus changes naturally over time and thus any time-invariant shape property may provide a misleading characterization of cellular populations. The algorithm developed provides a statistical

framework for analyzing populations of fixed cell and shows that a single sample in time provides an incomplete picture especially if the cells are not all in the same cell cycle phase. Previous work showed that changes in cellular shape might lead to local chemical gradients and thus to amplification of signals including transcriptional regulation at a cellular level [97]. At the nuclear level, a similar mechanism might be at work where changes in nuclear shape influence the distribution of chemicals at different times in the cell cycle, leading to transcriptional changes. Future work should further explore changes in nuclear shape and expand the body of knowledge on the mechanisms and importance of these changes.

Another challenge in the biomedical community is the significant differences between model systems including various types of cell culture [135, 56]. Growing mammalian cells *in vitro* is an indispensable technique for cell biology and biomedical research. However, 2D cell cultures do not resemble the natural 3D structures of body tissues, and as a result cells grown in 2D have considerable discordances in cellular morphology, physiology, pathology, cell-cell interaction and communication compared with natural tissues [46, 3, 20, 108]. In Chapter III the effect of 2D and 3D cell culture was explored with chromosome conformation capture and gene expression in fibroblasts derived from human foreskin. The analysis of RNA-seq data identified large numbers of differentially expressed genes between the culture conditions. By combining this analysis with analysis of Hi-C data it was shown that many of the changes in gene expression are localized to genomic regions that display structural changes. Additionally, the nuclear structure and function of 3D cultured cells was more similar to native skin tissue indicating that 3D culture might be a better model system for understanding how real tissues respond. This leads more support to the trend of working with better research models including using 3D culture to help with drug discovery and translational work [56, 125, 37].

Chromosomal translocations and aneuploidy are hallmarks of cancer genomes;

however, the impact of these aberrations on the nucleome are not yet understood [47, 41]. Previous studies of cancer genomes using Hi-C showed long range interactions between known risk loci for the development of CRC and regulatory regions [58], demonstrated proto-oncogene activation by disruption of chromosome neighborhoods [51] and showed that changes in genomic copy number subdivide the domain structure of chromosomes [119]. In Chapter IV, the nucleome of colorectal cancer cell line HT-29 was analyzed through collection of Hi-C to measure genome structure and RNA-seq to measure consequent changes in function. A new normalization method that can correct for non-constant copy number was developed and the sites of translocation and copy number changes were determined at high resolution from the Hi-C data. The data show that the relationship between structure and function that is well studied in normal cells is maintained in cancer samples. Additionally, for a small region around the translocation, the correlation between structure and function increases indicating that chromatin accessibility more directly reflects transcription. By analyzing a high copy number region on chromosome 8 that contains the oncogene *MYC*, we show that the region interacts with many regions across the genome in different cells and strongly interacts with an amplified region on chromosome 2 that contains the oncogene *STARD7*.

The methods developed in IV chapter can be used to identify chromosomal abnormalities at high resolution and allow analysis regardless of karyotype. These tools have been released as part of a MATLAB package that contains functions for loading data, normalization, defining topologically associating domains, exploring translocations, and analyzing time series datasets. It is important for tools to be publicly available for a number of reasons including to increase reproducibility of data, increasing the rate of progress by sharing work, and allowing comparison of the tools to determine which work best. A first attempt at comparing many of the tools developed for defining TADs was recently published [39] and more studies of this kind focusing

on various steps in the analysis pipeline are needed.

Cancer stem cells are a subpopulation of cancer cells with distinct properties that are thought to be particularly important to therapeutic resistance and metastasis [83, 136, 132]. Chapter V covers an ongoing project that extends the analysis of nuclear structure and function to cancer stem cells. By gathering Hi-C and RNA-seq datasets for genetically identical CSC and non-CSC populations, we explore structural and functional signatures of CSCs. Regions with CSC specific interactions are identified including a region that contains *MYC* and *POUF51B*, two transcription factors associated with reprogramming healthy cells into stem cells. These results show that cellular subpopulations have unique structures that might be important for the understanding how cancer cells evade therapeutics and go on to cause often deadly metastasis.

As more Hi-C experiments are performed on a variety of cancer subpopulations, cancer types, and cell lines, the Analyses used in this dissertation can be used to extend the results to new areas and draw new conclusions on how nuclear structure and function interplay in cancer. Analysis of larger sets of translocations and copy number alterations will help verify the conclusions drawn from this exploration of HT-29 and K562. This work can be extended to explore mechanisms by collecting datasets from a cancer development model system at multiple stages [49] and by studying more subpopulations and the distinct properties they display [130, 84]. One technology that will undoubtedly extend the ability to analyze nuclear structure that is particularly interesting for cancer, is single cell Hi-C [96, 118]. By studying individual cells, the heterogeneity of nuclear structure in cancer can be studied and previously uncharacterized subpopulations will likely be identified.

One of the biggest difficulties with Hi-C arises from the fact that most Hi-C cannot distinguish between copies of the same region. The genome is composed of two copies of each chromosome folded inside each cell yet almost all studies of genomic

folding ignore this instead analyzing the combination of the chromosomes. Chapter VI explores how SNPs can be used to distinguish the two copies of each chromosome to study allele specific structures and functions. Preliminary results on overall and nascent gene expression identify over one thousand genes with monoallelic expression in B-lymphocytes including some that change through the cell cycle. Work in this area [98, 30] is also important for cancer studies. If this work on distinguishing the two copies in each cell can be extended to work on high copy number regions or translocations then many of the simplifying assumptions required for current analysis of chromosomal aberrations could be removed or reduced. Instead of studying the general properties of a high copy number region, it would be possible to look at the copies individually, study how they interact, and what differences there are between copies.

Projects with the goal of characterizing the full dynamics of the cancer nucleome to help identify the best paths for reprogramming them are beginning. Current projects have explored how to reprogram a fibroblast into a myotube [74] and have developed an algorithm to predict what transcription factors will reprogram a fibroblast into any other cell type [105]. With more study of cancerous systems and their dynamics, these algorithms will be extended to the diseased state. The goal is to use ongoing experiments along with the information gathered here to reprogram cancer cells to make them die, make them more like normal cells, or make them more susceptible to traditional therapeutics.

APPENDIX

APPENDIX A

Supplemental Materials

A.1 Nuclear morphology: supplemental proof

Using the method of Lagrange multipliers, the appropriate objective function is given by the following sum of two terms

$$F \triangleq F_D + F_\lambda. \tag{S1}$$

F_D represents the weighted sum of squared distances from equation 2.4 in the paper:

$$F_D \triangleq \sum_{p_i \in P} w_i \, D^2(p_i, \mathbf{Q}_*), \tag{S2}$$

and F_λ represents the constraint from equation 2.3 in the paper weighted by the Lagrange multiplier λ:

$$F_\lambda \triangleq \lambda \, (a_*^\gamma + b_*^\gamma + c_*^\gamma - 3). \tag{S3}$$

The first term can be expanded via

$$\begin{aligned} F_D &= \sum_{p_i \in P} w_i \left(p_i - o_*\right)^T Q_*^{-1} \left(p_i - o_*\right) \\ &= \sum_{p_i \in P} w_i \left(p_i - o_*\right)^T R_* \begin{bmatrix} a_*^{-2} & 0 & 0 \\ 0 & b_*^{-2} & 0 \\ 0 & 0 & c_*^{-2} \end{bmatrix} R_*^T \left(p_i - o_*\right) \\ &= \begin{pmatrix} \sum_{p_i \in P} w_i \, x_i'^2 \\ \sum_{p_i \in P} w_i \, y_i'^2 \\ \sum_{p_i \in P} w_i \, z_i'^2 \end{pmatrix} \cdot \begin{pmatrix} a_*^{-2} \\ b_*^{-2} \\ c_*^{-2} \end{pmatrix} = p'^2 \cdot \begin{pmatrix} a_*^{-2} \\ b_*^{-2} \\ c_*^{-2} \end{pmatrix} \end{aligned} \tag{S4}$$

where

$$p_i' \triangleq \begin{pmatrix} x_i' \\ y_i' \\ z_i' \end{pmatrix} \triangleq R_*^T \left(p_i - o_*\right) \tag{S5}$$

and

$$p'^2 \triangleq \begin{pmatrix} \sum_{p_i \in P} w_i \, x_i'^2 \\ \sum_{p_i \in P} w_i \, y_i'^2 \\ \sum_{p_i \in P} w_i \, z_i'^2 \end{pmatrix}. \tag{S6}$$

Taking derivatives of F with respect to the components of o_* and setting them to 0, it is easy to verify that $\sum_{p_i \in P} w_i \left(p_i - o_*\right)$ must yield the null vector. But the matrix Q_*^{-1} is non-singular, so

$$o_* = \bar{p} \triangleq \frac{\sum_{p_i \in P} w_i \, p_i}{\sum_{p_i \in P} w_i}, \tag{S7}$$

verifying equation 2.5 in the paper.

Furthermore, the particular rotation R_C diagonalizes all the squared energy in the sum of outer products of transformed points p_i' from equation S5. In other words, by

choosing

$$R_* = R_C \tag{S8}$$

where R_C is derived from an eigenanalysis of the dataset's covariance matrix C (equations 2.6 and 2.7 in the paper), we extremize each coordinate of the vector p'^2 which forms the diagonal of this outer product. More concretely, defining C' as the weighted sum of outer products of transformed points p', we have

$$\begin{aligned}
w\,C' &\triangleq \sum_{p_i \in P} w_i\, p'_i \otimes p'_i \\
&= \sum_{p_i \in P} w_i\, R_C^T\, (p_i - \bar{p})\, (p_i - \bar{p})^T\, R_C \\
&= R_C^T \left(\sum_{p_i \in P} w_i\, (p_i - \bar{p}) \otimes (p_i - \bar{p}) \right) R_C \\
&= R_C^T\, C\, R_C \\
&= R_C^T \left(R_C \begin{bmatrix} a_C^2 & 0 & 0 \\ 0 & b_C^2 & 0 \\ 0 & 0 & c_C^2 \end{bmatrix} R_C^T \right) R_C \\
&= \begin{bmatrix} a_C^2 & 0 & 0 \\ 0 & b_C^2 & 0 \\ 0 & 0 & c_C^2 \end{bmatrix}
\end{aligned} \tag{S9}$$

where the covariance matrix C is given by equation 2.7 in the paper:

$$w\,C \triangleq \sum_{p_i \in P} w_i\, (p_i - \bar{p}) \otimes (p_i - \bar{p}) \tag{S10}$$

and $w \triangleq \sum_{p_i \in P} w_i$. Thus

$$p^{'2} = w \, \mathrm{diag}(C') = w \begin{pmatrix} a_C^2 \\ b_C^2 \\ c_C^2 \end{pmatrix} \tag{S11}$$

and the first term of the objective reduces to

$$F_D = w \begin{pmatrix} a_C^2 \\ b_C^2 \\ c_C^2 \end{pmatrix} \cdot \begin{pmatrix} a_*^{-2} \\ b_*^{-2} \\ c_*^{-2} \end{pmatrix} \tag{S12}$$

by equation S4. Taking derivatives,

$$\frac{\partial F}{\partial a_*} = -2 \, w \, a_C^2 \, a_*^{-3} + \lambda \, \gamma \, a_*^{\gamma - 1}, \tag{S13}$$

and equating to 0 we obtain

$$a_C^2 = \frac{\lambda \, \gamma}{2 \, w} a_*^{\gamma + 2} \tag{S14}$$

so that

$$a_*^2 \propto a_C^{\frac{4}{\gamma + 2}}. \tag{S15}$$

The derivation is similar for the other optimal scale factors b_* and c_*, with the same constant of proportionality, yielding equation 2.8 in the paper. Finally, applying the constraint (equation 2.3) yields equation 2.9 in the paper.

A.2 Nuclear morphology: supplemental tables

Table S1: Best fit frequency and phase for all individuals. The phase θ, nMSE ($MSE/\sum fi^2$), $PSNR$, α, and β for each individual for each measurements' best two frequencies, ω, based on the average of the periodograms shown in the left panel of Figure 3A. Ellip, thresh, eccen stand for ellipsoid, threshold, and eccentricity, respectively.

		Best Frequency					
		ω	α	β	θ	$nMSE$	$PSNR$
Ellip	S1	0.0446	−55.3	−63.52	0.7163	0.0003	4.44
Volume	S2	0.1166	51.8	58.44	0.7255	0.0006	0.172
	S3	0.0566	105	−54.6	−1.092	0.0087	−13.5
	S4	0.1506	76.4	−38.37	−1.105	0.0004	2.87
	S5	0.0566	−43.1	56.75	−0.6494	0.0049	−9.92
	S6	0.0546	−12.7	192.9	−0.0658	0.0083	−14.5
Thresh	S1	0.0866	146	−0.9363	−1.564	0.0074	−12.1
Volume	S2	0.1106	−0.428	93.19	−0.0046	0.0012	−3.87
	S3	0.0566	91.1	−73.05	−0.895	0.0014	−5.23
	S4	0.0446	105	8.879	1.487	0.0013	−3.93
	S5	0.0566	15.2	134.5	0.1125	0.0005	−0.876
	S6	0.0866	262	−31.78	−1.45	0.0079	−15.6
Eccen	S1	0.0726	−0.0029	0.0052	−0.515	0.0009	36
	S2	0.1166	−0.0007	−0.0037	0.188	0.0011	37.3
	S3	0.0586	−0.0051	−0.0057	0.7337	0.0066	27.8
	S4	0.0306	−0.0082	−0.0037	1.152	0.0045	29.2
	S5	0.0566	−0.0084	−0.0072	0.8565	0.0070	24.7
	S6	0.0426	−0.0057	−0.0057	0.7887	0.0018	31.8
Short	S1	0.0446	−0.115	−0.1486	0.6598	0.0061	17.5
Axis	S2	0.1166	0.0766	0.1483	0.4766	0.0005	25.8
	S3	0.0566	0.134	0.0589	1.157	0.0046	20.4
	S4	0.0386	−0.16	−0.0235	1.425	0.0053	18.4
	S5	0.0566	0.0616	0.1516	0.3859	0.0001	26.3
	S6	0.0406	−0.062	0.2327	−0.2604	0.005	16.5
Middle	S1	0.0426	−0.0512	−0.2856	0.1772	0.0030	15.4

Axis	S2	0.0326	−0.331	0.1015	−1.274	0.0029	16.5
	S3	0.0566	0.26	−0.2617	−0.7828	0.0079	12.7
	S4	0.0646	−0.208	0.1868	−0.8401	0.0010	22.8
	S5	0.0606	−0.12	0.3135	−0.3655	0.0103	11.6
	S6	0.0546	−0.0255	0.4535	−0.0562	0.0065	11.9
Long	S1	0.0726	−0.195	0.2225	−0.7197	0.0017	20.4
Axis	S2	0.1386	0.219	0.2171	0.7898	0.0007	22
	S3	0.0606	−0.171	−0.462	0.3551	0.0111	10.4
	S4	0.0646	−0.308	0.1774	−1.048	0.0062	15.1
	S5	0.0526	−0.221	−0.2828	0.6643	0.0069	12.3
	S6	0.0546	−0.0597	0.4008	−0.1479	0.0040	15.4

			Second	Frequency			
		ω	α	β	θ	$nMSE$	$PSNR$
Ellip	S1	0.1006	−51.6	15.97	−1.27	0.0042	−7.74
Volume	S2	0.0326	−71.8	25.99	−1.224	0.0028	−6.17
	S3	0.0346	−61.2	57.28	−0.8185	0.0029	−8.8
	S4	0.0666	−24.8	64.39	−0.367	0.0017	−3.66
	S5	0.0406	38.3	55.96	0.5998	0.0042	−9.22
	S6	0.0386	−119	62.13	−1.091	0.0043	−11.6
Thresh	S1	0.0466	−87.8	57.38	−0.9921	0.0015	−5.31
Volume	S2	0.0726	−71.6	25.14	−1.233	0.0003	2.32
	S3	0.0346	−56.3	95.04	−0.5349	0.0063	−11.6
	S4	0.1606	−73.7	11.08	−1.422	0.0002	3.59
	S5	0.1226	78.7	−1.591	−1.551	0.005	−10.9
	S6	0.0366	−230	−75.2	1.255	0.0063	−14.6
Eccen	S1	0.0866	−0.0042	−0.0019	1.154	0.0049	28.8
	S2	0.0686	0.0031	0.0004	1.452	0.0002	44.1
	S3	0.1386	0.0043	0.0058	0.6367	0.0022	32.6
	S4	0.1406	0.0063	0.002	1.265	0.0058	28.1
	S5	0.0926	−0.0006	0.0072	−0.0811	0.0031	28.2
	S6	0.0766	−0.0005	0.008	−0.0653	0.0041	28.1
Short	S1	0.0866	0.093	0.03664	1.196	0.0009	25.7
Axis	S2	0.0706	−0.122	0.07048	−1.047	0.0006	25.8
	S3	0.0386	0.01	0.1302	0.0770	0.0068	18.7
	S4	0.1426	−0.116	0.0597	−1.095	0.0016	23.7

	S5	0.1646	0.0153	0.1506	0.1011	0.0054	18.9
	S6	0.0546	−0.0468	0.2182	−0.2112	0.0039	17.2
Middle	S1	0.0746	−0.166	0.1931	−0.7101	0.0011	19.9
Axis	S2	0.0666	−0.299	0.0808	−1.307	0.0045	14.7
	S3	0.0326	−0.264	0.1109	−1.173	0.0050	14.7
	S4	0.1486	0.263	−0.0271	−1.468	0.0007	24.7
	S5	0.0386	−0.0682	0.2203	−0.3002	0.0041	15.5
	S6	0.0846	0.322	0.1044	1.258	0.0037	14.5
Long	S1	0.0306	−0.186	−0.1315	0.9555	0.0010	22.6
Axis	S2	0.0346	−0.298	0.0806	−1.306	0.0034	15.3
	S3	0.1426	0.321	−0.0486	−1.42	0.0020	17.8
	S4	0.1106	0.249	0.0639	1.319	0.0026	18.8
	S5	0.0386	−0.0122	0.3259	−0.0373	0.0058	13.1
	S6	0.0826	0.216	0.2564	0.701	0.0036	15.9

Table S2: Best fit frequency and phase for each individual. The phase θ, squared residual energy, $F(\omega)$, α, and β for each sample for each measurements' best two frequencies, ω. Ellip, thresh, eccen stand for ellipsoid, threshold, and eccentricity, respectively.

		Best Frequency					
		ω	α	β	θ	$nMSE$	$PSNR$
Ellip	S1	0.0546	12.8	9.075	0.9543	0.0003	4.44
Volume	S2		22.9	−15.17	−0.9866	0.0007	0.172
	S3		118	14.39	1.449	0.0087	−13.5
	S4		−0.149	24.09	−0.0062	0.0004	2.87
	S5		−65.2	26.97	−1.179	0.0049	−9.92
	S6		−12.7	192.9	−0.066	0.0083	−14.5
Thresh	S1	0.0886	116	−77.42	−0.9842	0.0074	−12.1
Volume	S2		−28.4	28.09	−0.7909	0.0012	−3.87
	S3		−43.8	−33.4	0.919	0.0015	−5.23
	S4		28.5	−39.76	−0.6211	0.0013	−3.93
	S5		9.33	37.39	0.2446	0.0005	−0.876
	S6		184	−187.5	−0.7761	0.0079	−15.6
Eccen	S1	0.0586	−0.002	−0.0022	0.7365	0.0009	36

	S2		−0.0021	−0.0003	1.452	0.0011	37.3
	S3		−0.0051	−0.0057	0.7337	0.0066	27.8
	S4		−0.0042	−0.0051	0.686	0.0045	29.2
	S5		−0.0105	−0.0028	1.307	0.0070	24.7
	S6		−0.0048	0.0019	−1.189	0.0018	31.8
Short	S1	0.0406	0.0626	−0.1551	−0.3837	0.0061	17.5
Axis	S2		0.013	0.0531	0.2403	0.0005	25.8
	S3		0.0688	0.1031	0.5881	0.0046	20.4
	S4		−0.128	0.0988	−0.9145	0.0053	18.4
	S5		0.0575	−0.0370	−0.9988	0.0001	26.3
	S6		−0.062	0.2327	−0.2604	0.0046	16.5
Middle	S1	0.0586	−0.156	−0.1725	0.7347	0.0031	15.4
Axis	S2		−0.114	−0.2592	0.4141	0.0029	16.5
	S3		0.0529	−0.3635	−0.1446	0.0079	12.7
	S4		−0.0774	−0.1056	0.6328	0.0010	22.8
	S5		−0.243	0.2255	−0.8232	0.0103	11.6
	S6		0.333	0.2163	0.9943	0.0066	11.9
Long	S1	0.06057	−0.0769	−0.1109	0.606	0.0017	20.4
Axis	S2		−0.0912	−0.0951	0.7641	0.0007	22
	S3		−0.171	−0.462	0.3551	0.0111	10.4
	S4		−0.266	−0.1971	0.9338	0.0069	15.1
	S5		−0.205	0.2559	−0.6761	0.0069	12.3
	S6		0.309	0.0244	1.492	0.0040	15.4

			Second	Frequency			
		ω	α	β	θ	$nMSE$	$PSNR$
Ellip	S1	0.0386	53.2	−32.39	−1.024	0.0042	−7.74
Volume	S2		−5.03	58.56	−0.0856	0.0028	−6.17
	S3		7.62	70.72	0.1074	0.0029	−8.8
	S4		−39.2	31.36	−0.8966	0.0017	−3.66
	S5		2.98	67.08	0.0444	0.0042	−9.22
	S6		−119	62.13	−1.091	0.0043	−11.6
Thresh	S1	0.0546	61.6	15.1	1.33	0.0015	−5.31
Volume	S2		18.3	6.91	1.209	0.0003	2.32
	S3		115	−9.334	−1.49	0.0063	−11.6
	S4		19.3	6.909	1.228	0.0002	3.59

	S5		−53.2	110.3	−0.4493	0.005	−10.9
	S6		−50.2	227.4	−0.2173	0.0063	−14.6
Eccen	S1	0.0426	0.0008	0.0067	0.1175	0.0049	28.8
	S2		−0.0009	0.0004	−1.107	0.0002	44.1
	S3		−0.0041	−0.0010	1.328	0.0022	32.6
	S4		0.0030	−0.0068	−0.4134	0.0058	28.1
	S5		0.0033	0.0064	0.4798	0.0031	28.2
	S6		−0.0057	−0.0057	0.7887	0.0041	28.1
Short	S1	0.0566	0.0534	0.0383	0.9483	0.0009	25.7
Axis	S2		0.0523	0.0096	1.39	0.0005	25.8
	S3		0.134	0.0589	1.157	0.0068	18.7
	S4		0.0312	0.0856	0.35	0.0016	23.7
	S5		0.0616	0.1516	0.3859	0.0054	18.9
	S6		0.116	0.1924	0.542	0.0039	17.2
Middle	S1	0.0346	0.0134	0.134	0.0997	0.0011	19.9
Axis	S2		−0.275	0.2085	−0.9215	0.0045	14.7
	S3		−0.195	0.2175	−0.7308	0.0050	14.7
	S4		−0.0295	0.099	−0.29	0.0006	24.7
	S5		−0.184	0.1135	−1.019	0.0041	15.5
	S6		−0.271	−0.1068	1.195	0.0037	14.5
Long	S1	0.0366	−0.0298	0.1026	−0.2825	0.0010	22.6
Axis	S2		−0.236	0.1598	−0.976	0.0034	15.3
	S3		−0.0512	0.2102	−0.2389	0.0020	17.8
	S4		−0.0726	0.2069	−0.3375	0.0026	18.8
	S5		−0.173	0.2626	−0.5834	0.0058	13.1
	S6		−0.266	−0.0909	1.242	0.0036	15.9

Table S3: All nuclei axis lengths for all the nuclei. Available as SuppTable3.csv, http://www.tandfonline.com /doi/full/10.1080/ 19491034.2015.1095432

A.3 Nuclear morphology: supplemental figures

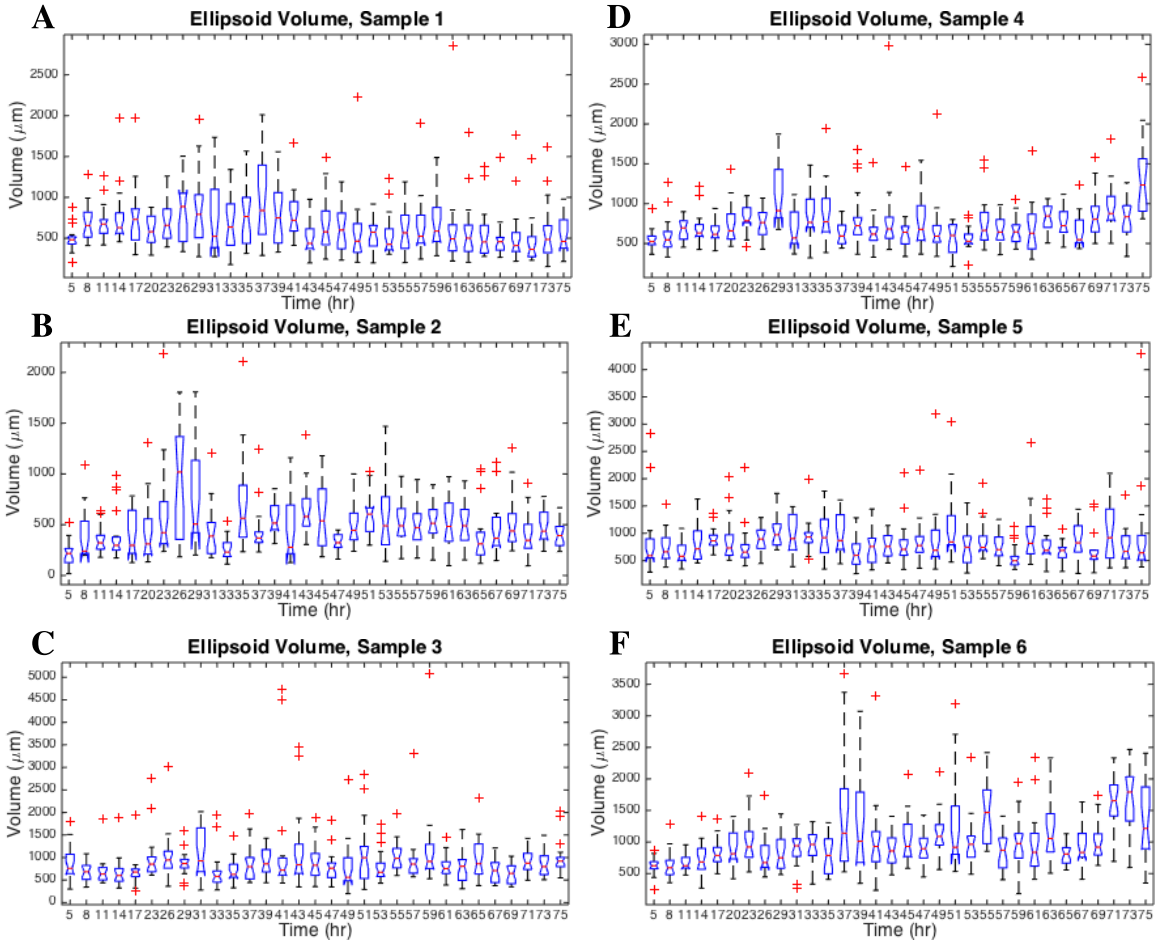

Figure S1: Ellipsoid volume box plots by sample. The box plots showing the distribution of ellipsoid volumes for each sample over time. The fact that the boxes do not all line up shows that the distribution changes significantly over time for all 6 individuals.

A.4 Spheroids versus monolayers: supplementary materials

All supplementary materials for this paper including Tables S1-5 are available at http://www.tandfonline.com/doi/ full/10.1080/19491034.2017.1280209

A.5 Nucleome analysis: supplemental figures

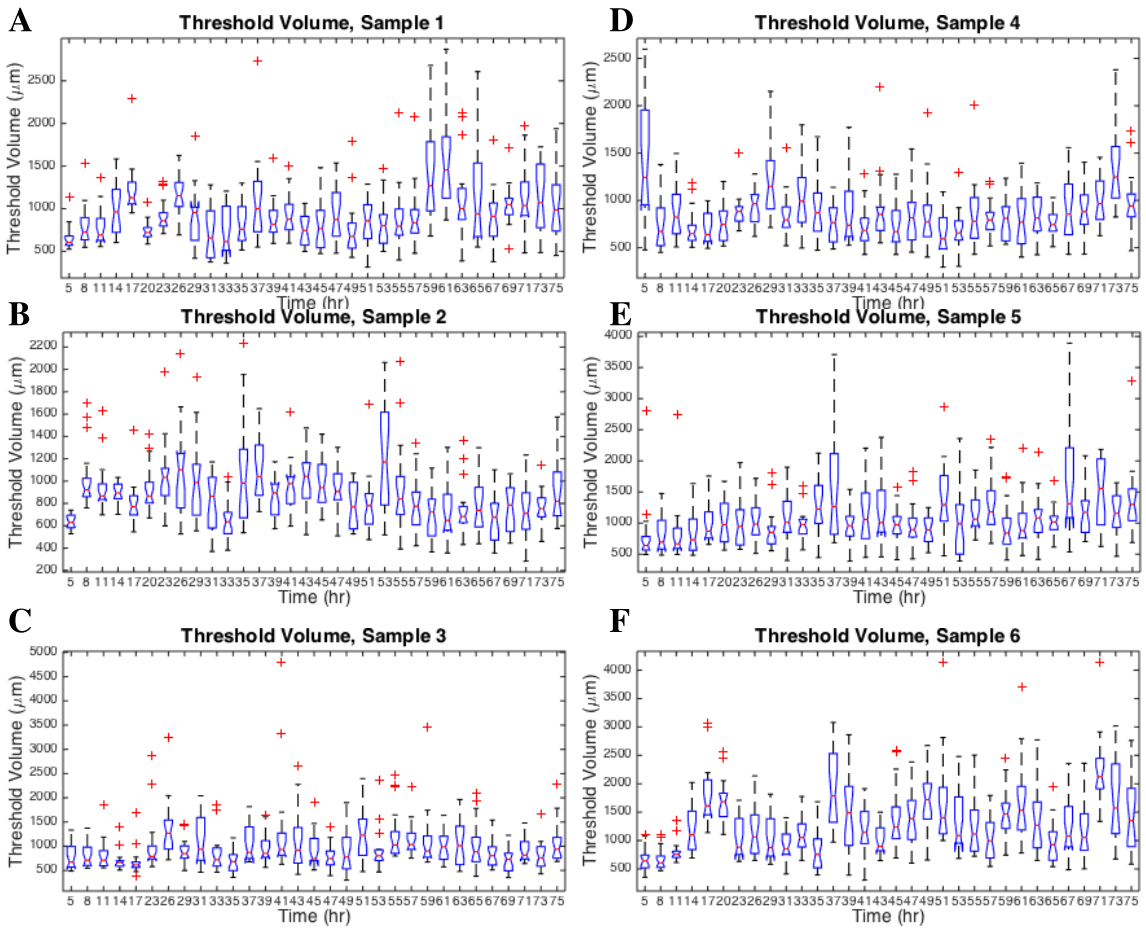

Figure S2: Threshold volume box plots by sample. The box plots showing the distribution of threshold volumes for each sample over time. The fact that the boxes do not all line up shows that the distribution changes significantly over time for all 6 individuals.

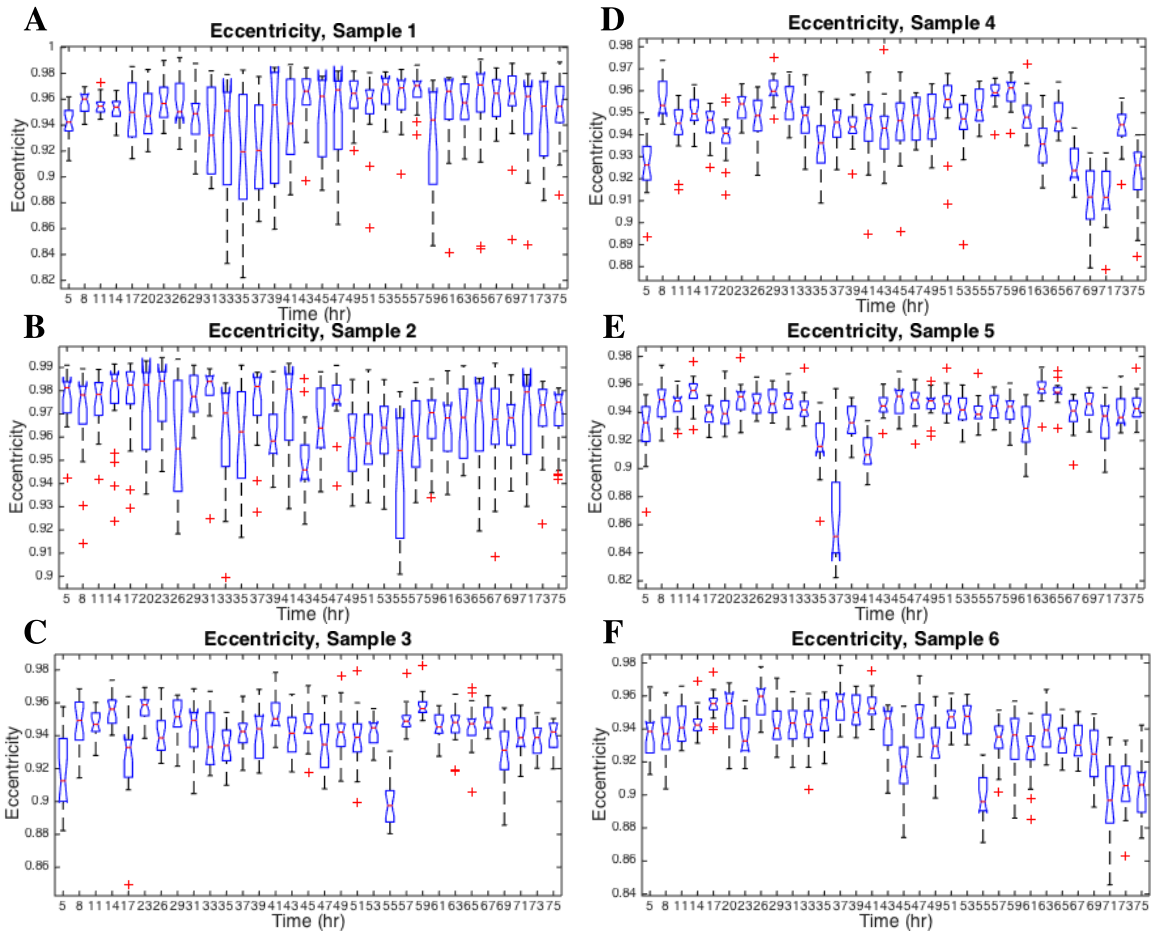

Figure S3: Eccentricity box plots by sample. The boxplots showing the distribution of eccentricities for each sample over time. The fact that the boxes do not all line up shows that the distribution changes significantly over time for all 6 individuals.

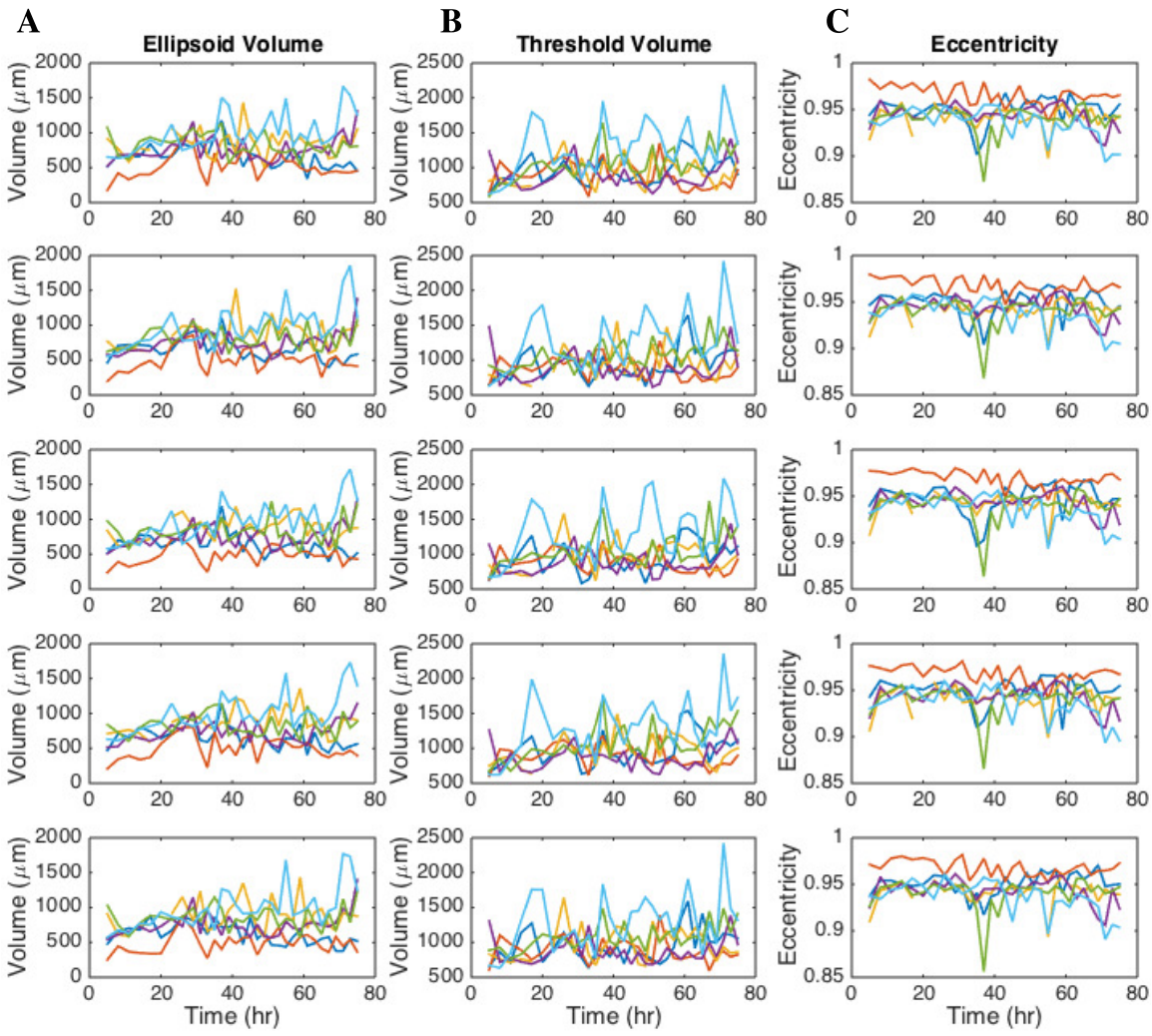

Figure S4: Random sample nuclear shape dynamics. Each column holds either the ellipsoid volume, threshold volume, or eccentricity over time. Each row is a different random sample of half the data. These data were used to validate the consistency of the periodicities seen.

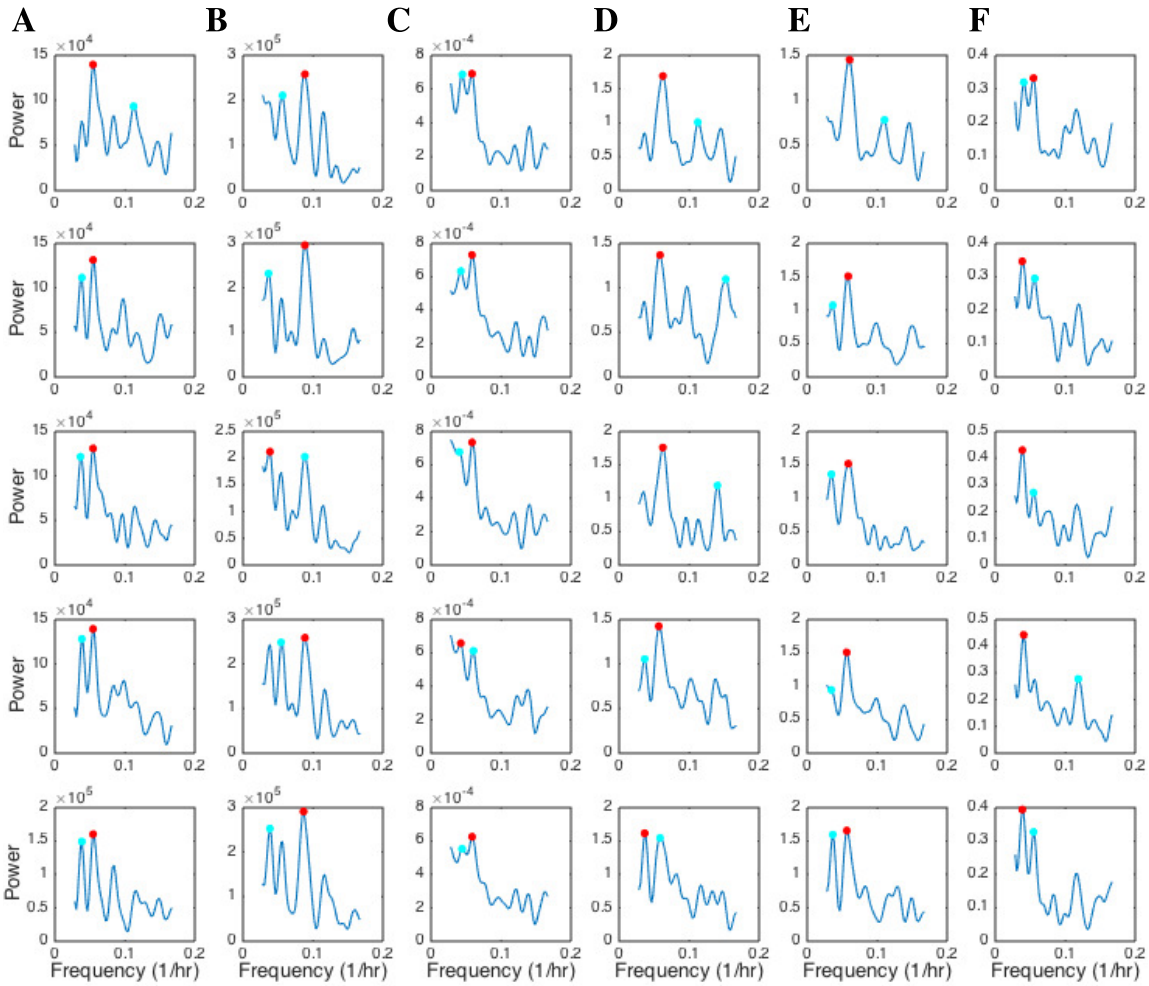

Figure S5: Random sample frequency spectrums. the columns are the average spectrum for the A) ellipsoid volume, B) threshold volume, C) eccentricity, D) shortest axis, E) middle axis, and F) longest axis. Each row's spectrums were calculated from corresponding the random sample of half the data whose dynamics are shown in figure S4.

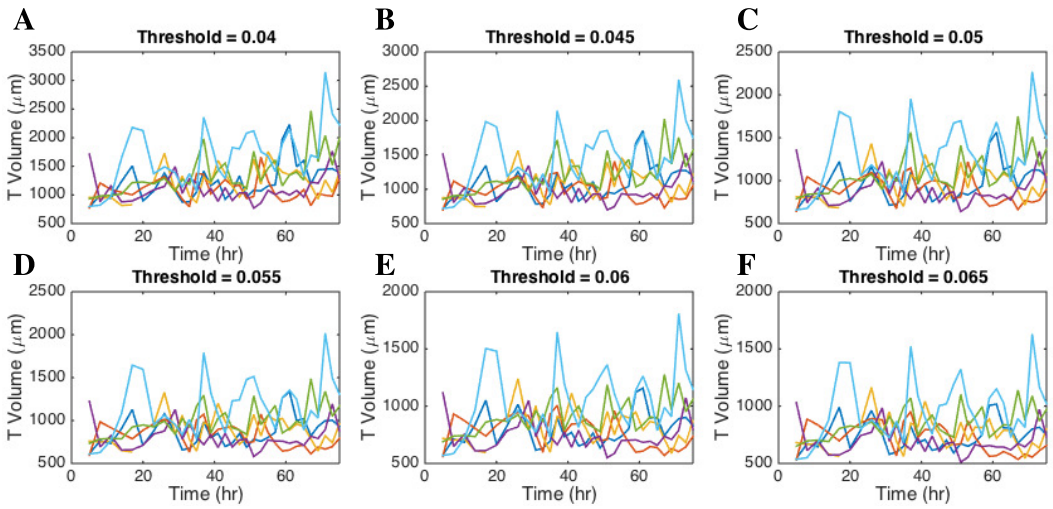

Figure S6: Threshold volume over time separated by individual for thresholds of A) 0.04, B) 0.045, C) 0.05, D) 0.55, E) 0.06, and F) 0.065. C is a repeat of Figure 2C since 0.05 was the threshold used in the rest of the paper.

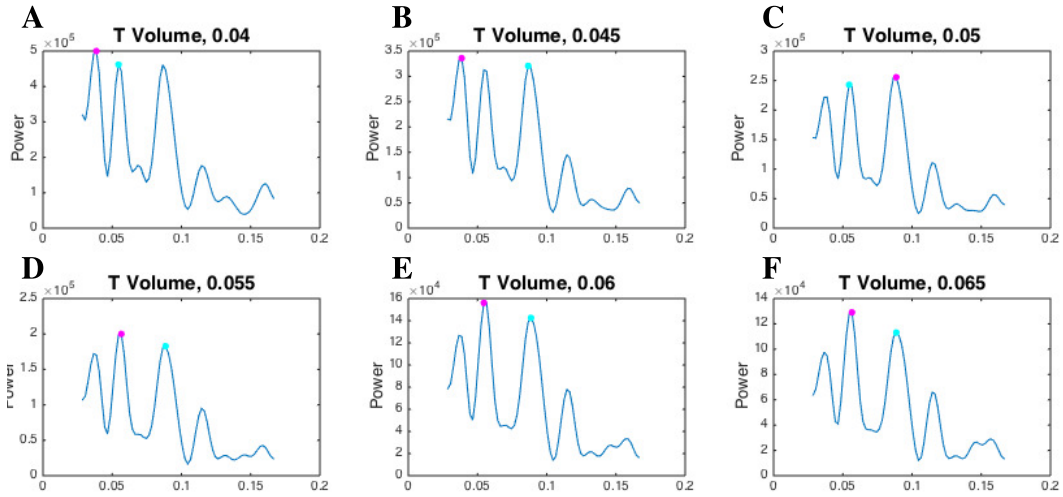

Figure S7: Spectrums from threshold volume for thresholds of A) 0.04, B) 0.045, C) 0.05, D) 0.55, E) 0.06, and F) 0.065. C is repeated from Figure 3A since 0.05 was the threshold used in the rest of the paper. The highest peak is marked in magenta and the second highest is marked in cyan.

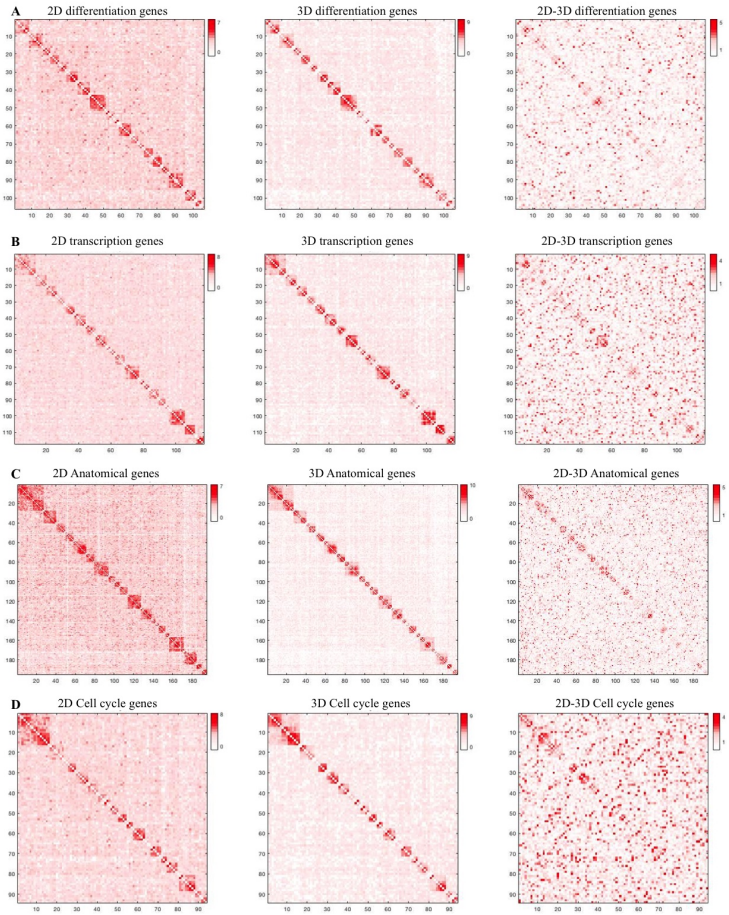

Figure S8: Chromatin interactions for differentially expressed gene sets. A) The chromatin interactions at 1 Mb resolution for four sets of differentially expressed genes clustered under GO terms A) "transcription", B) "cell differentiation", C) "anatomical structure development", and D) "cell cycle". The first two showed decreased expression while the last two showed increased expression in 3D culture relative to 2D culture. The left panel is the 2D cell matrix, the center one is the 3D cell matrix, and the right panel shows the difference between the two.

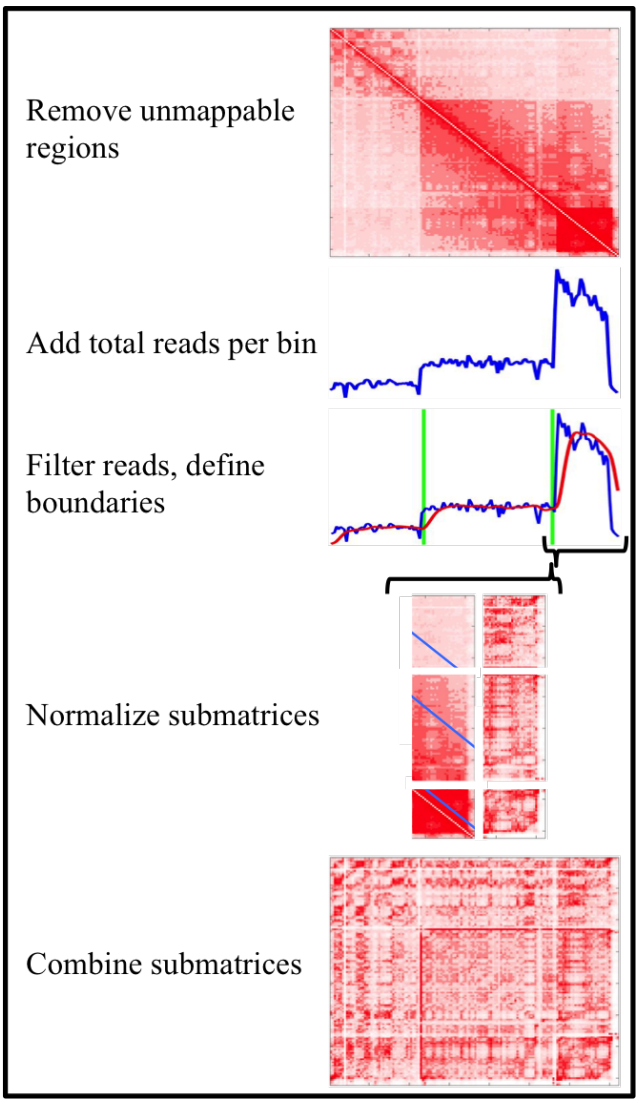

Figure S9: Copy number based normalization method. Chromosome 8 from the HT-29 2D12hr sample is shown as the example.

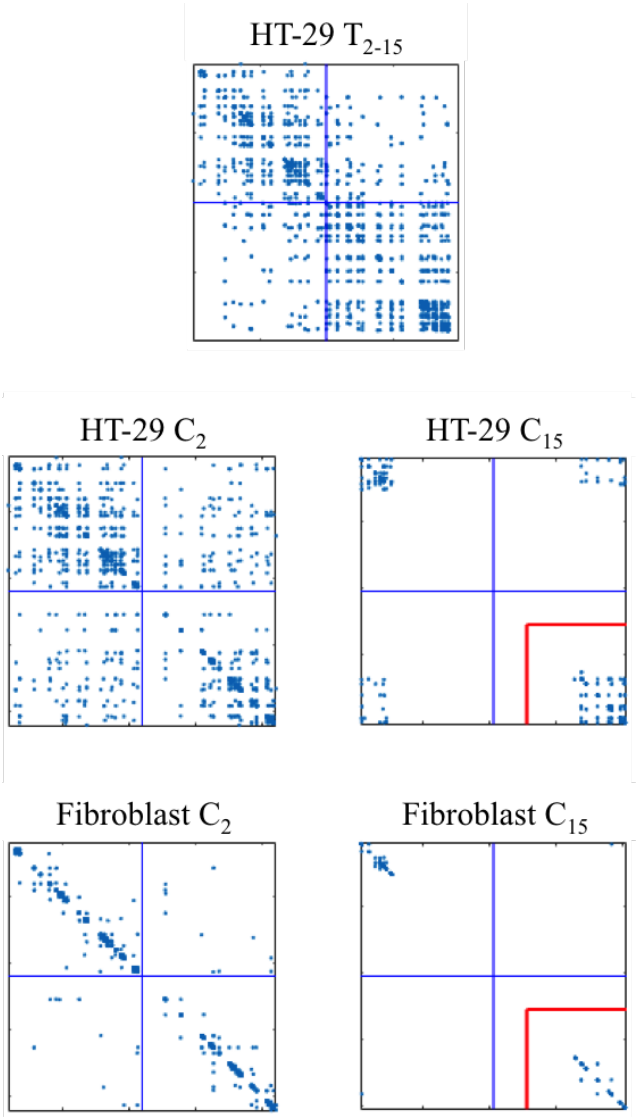

Figure S10: Translocation $2 - 15$ at read level. The 200 kb surrounding the two breakpoints in the t(2; 15) translocation at read level. The top row shows the reconstructed chromosome in HT-29 2D12hr. The middle row shows the non-translocated chromosomes in HT-29 2D12hr while the bottom shows the same chromosomes in fibroblasts. Boxes indicate the locations of genes colored by their expression in the appropriate cell line (red = off, orange = low expression, green = high expression) and the magenta lines across the top indicates CTCF binding sites.

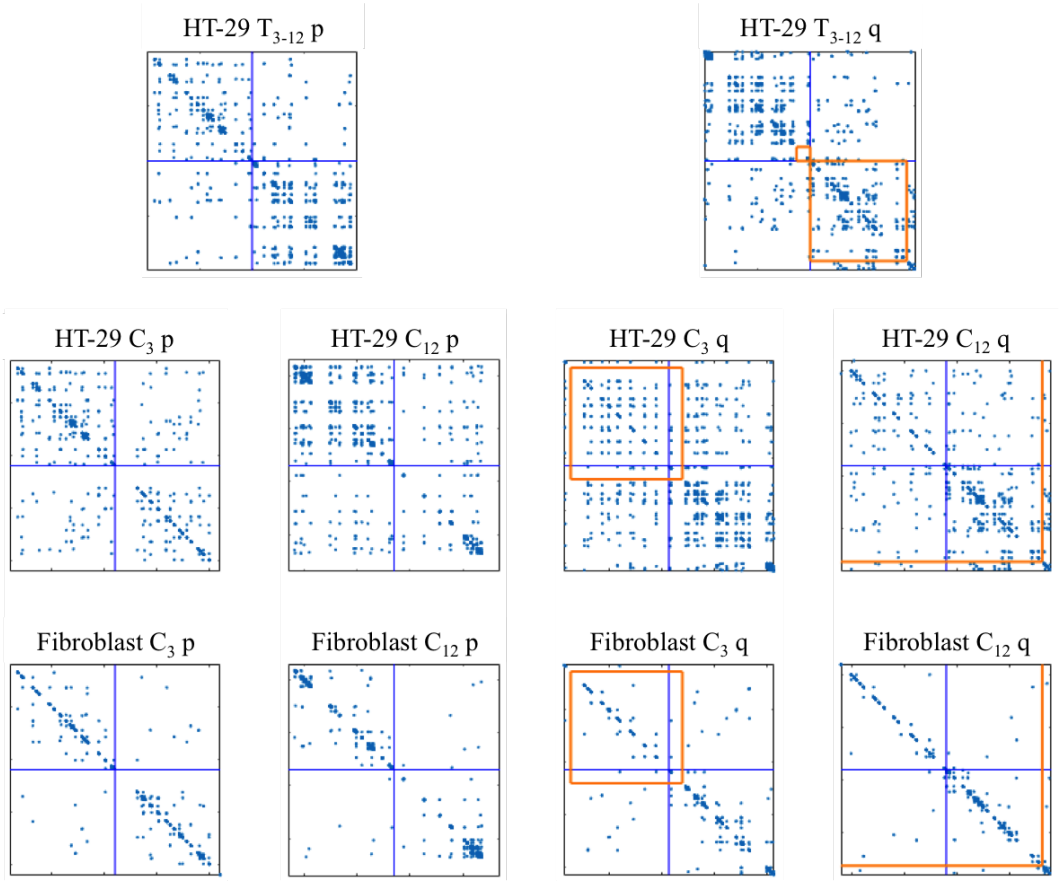

Figure S11: Translocation $3-12$ at read level. The 200 kb surrounding the two breakpoints in the $t(3;12)$ translocation at read level. The top row shows the reconstructed chromosome in HT-29 2D12hr. The p indicates the translocation closer to the p-end (lower genomic coordinate) while the q indicates the translocation closer the q terminal (higher genomic coordinate). The middle row shows the non-translocated chromosomes in HT-29 2D12hr while the bottom shows the same chromosomes in fibroblasts. Boxes indicate the locations of genes colored by their expression in the appropriate cell line (red = off, orange = low expression, green = high expression.

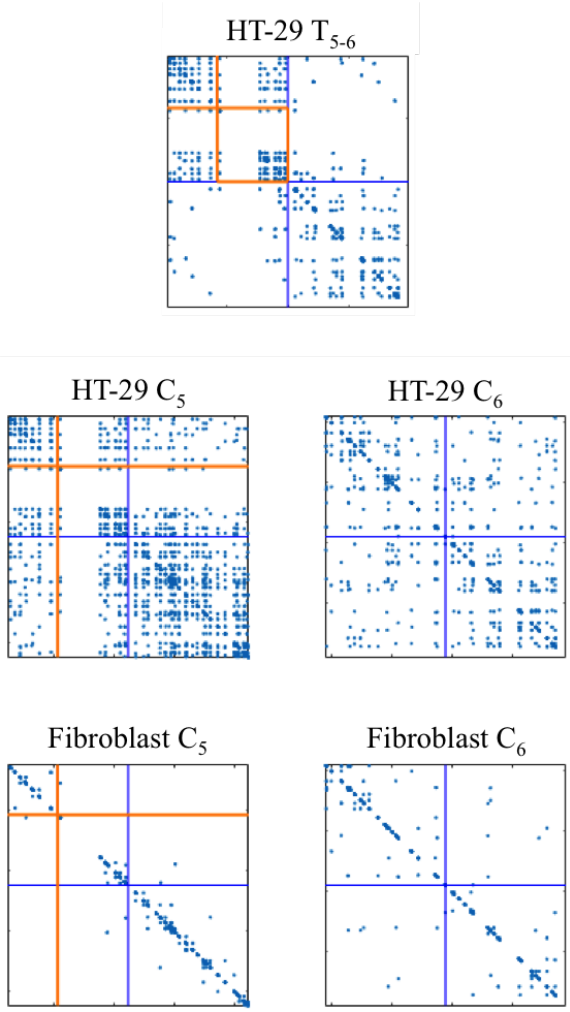

Figure S12: Translocation $5-6$ at read level. The 200 kb surrounding the breakpoint in the t(5; 6) translocation at read level. The top row shows the reconstructed chromosome in HT-29 2D12hr. The middle row shows the non-translocated chromosomes in HT-29 2D12hr while the bottom shows the same chromosomes in fibroblasts. Boxes indicate the locations of genes colored by their expression in the appropriate cell line (red = off, orange = low expression, green = high expression).

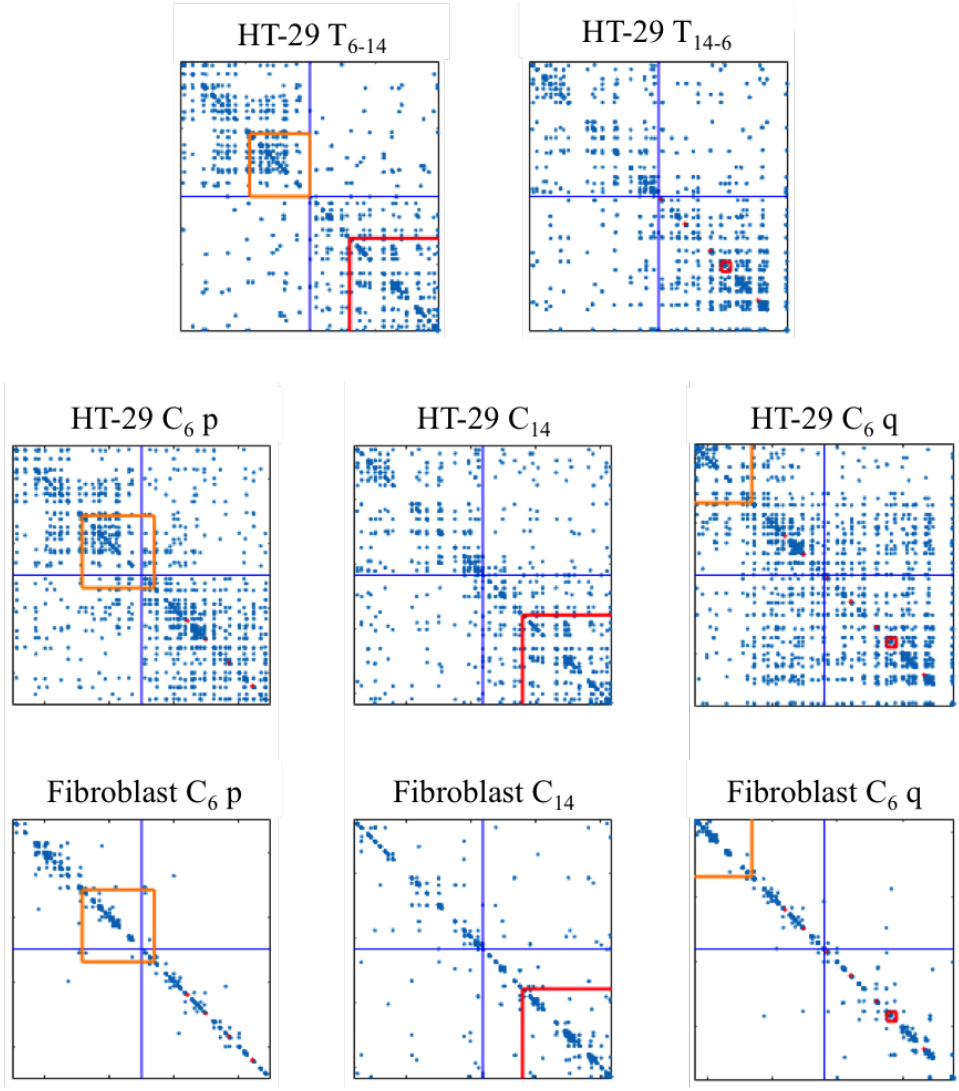

Figure S13: Translocation $6 - 14$ at read level. The 200 kb surrounding the $t(6;14)$ and $t(14;6)$ breakpoints at read level. The top row shows the reconstructed chromosome in HT-29 2D12hr. The p indicates the translocation closer to the p-end (lower genomic coordinate) while the q indicates the translocation closer the q terminal (higher genomic coordinate). The middle row shows the non-translocated chromosomes in HT-29 2D12hr while the bottom shows the same chromosomes in fibroblasts. Boxes indicate the locations of genes colored by their expression in the appropriate cell line (red = off, orange = low expression, green = high expression).

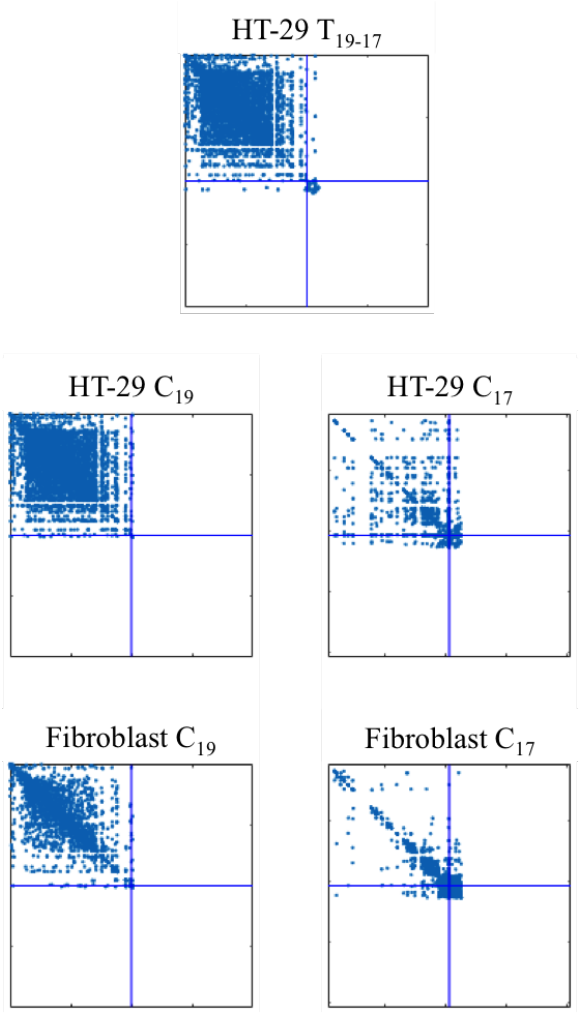

Figure S14: Translocation $19 - 17$ at read level. The 200 kb surrounding the two breakpoints in the $t(19;17)$ translocation at read level. The top row shows the reconstructed chromosome in HT-29 2D12hr. The middle row shows the non-translocated chromosomes in HT-29 2D12hr while the bottom shows the same chromosomes in fibroblasts. The lack of reads on the right and bottom of each region is due to the presence of the centromeres to which reads cannot be aligned. Boxes indicate the locations of genes colored by their expression in the appropriate cell line (red = off, orange = low expression, green = high expression) and the magenta lines across the top indicates CTCF binding sites.

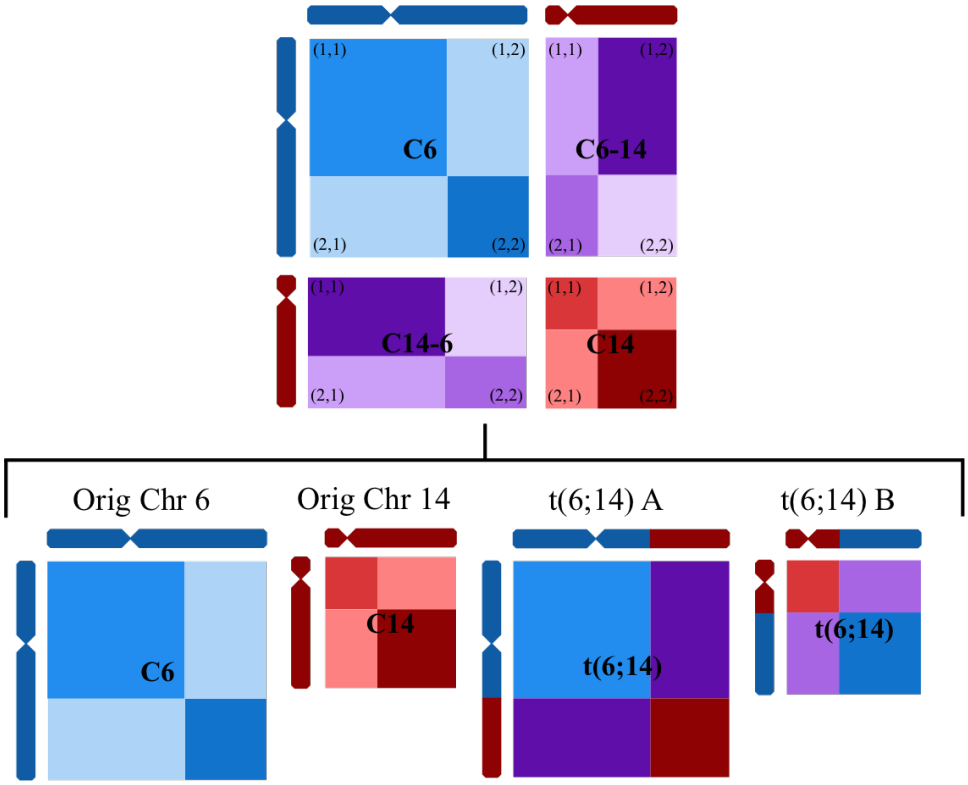

Figure S15: Translocated chromosome analysis from Hi-C data. A) A diagram of chromosomes 6 and 14 divided into sections based on the translocation's location. In addition to the two original chromosomes, two translocated chromosomes were formed by combining the relevant sections of the original chromosomes with parts of the inter-chromosomal matrix. We labeled the two t(6; 14) chromosomes with A and B. A denotes the larger one (with the beginning of chromosome 6) while B denotes the smaller chromosome (with the beginning of chromosome 14).

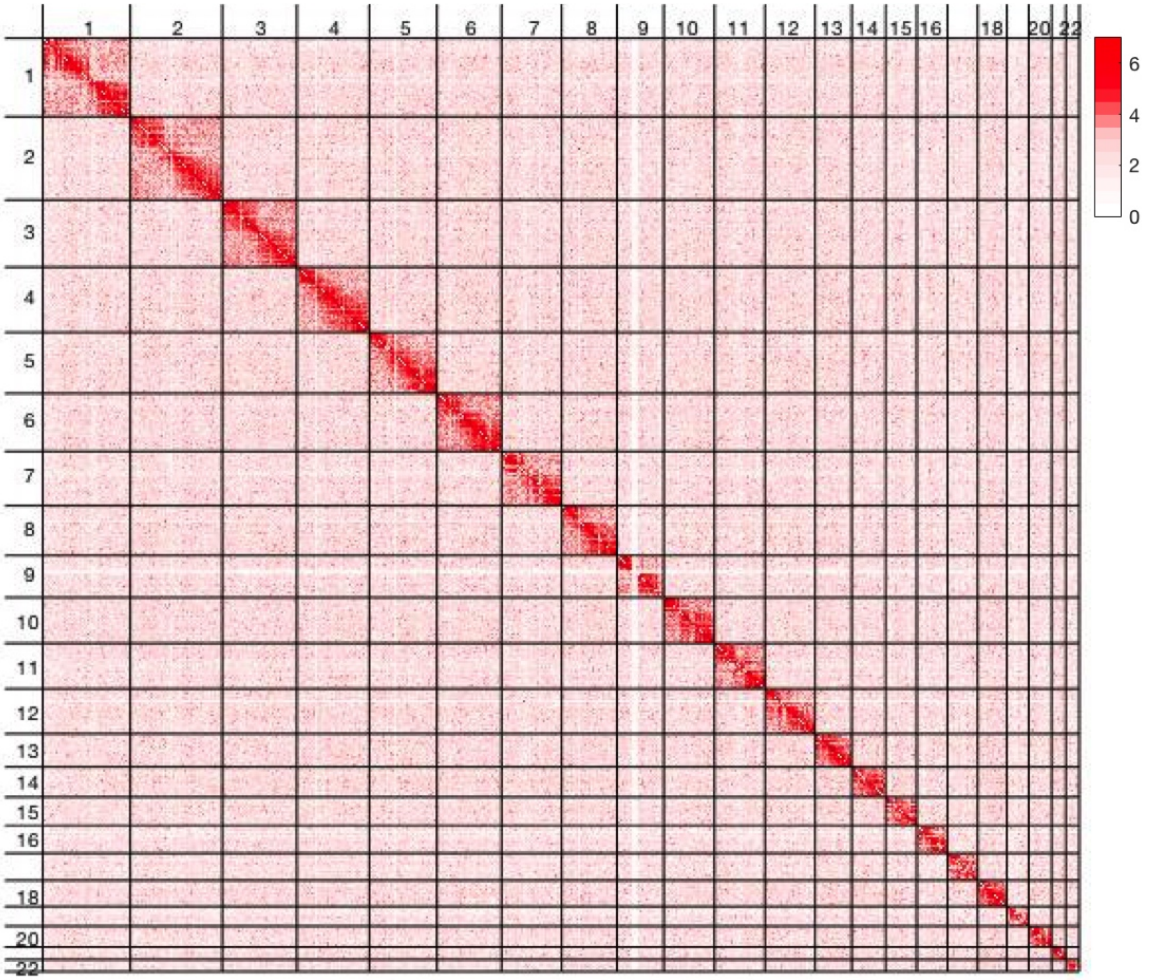

Figure S16: Genome-wide Hi-C matrix for 2D-cultured human fibroblast cells.

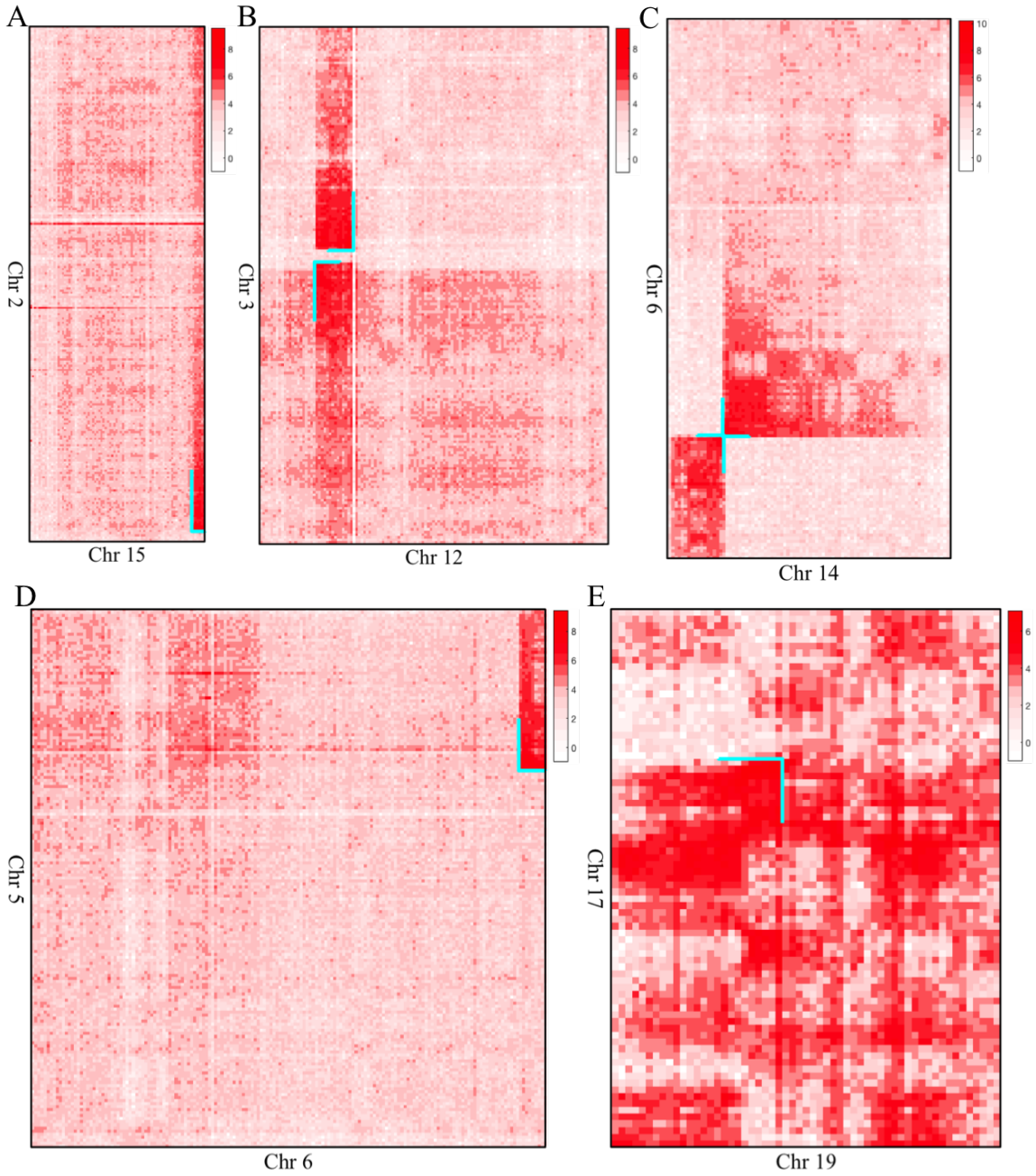

Figure S17: Interchromosomal matrices for translocations. A) the matrix showing the interactions between chromosomes 2 and 15, B) chromosomes 3 and 12 C) chromosomes 6 and 14, D) chromosomes 5 and 6, and E) chromosomes 17 and 19 at 100 kb resolution in the HT-29 2D12hr sample. Blue Ls indicate the sites of translocations.

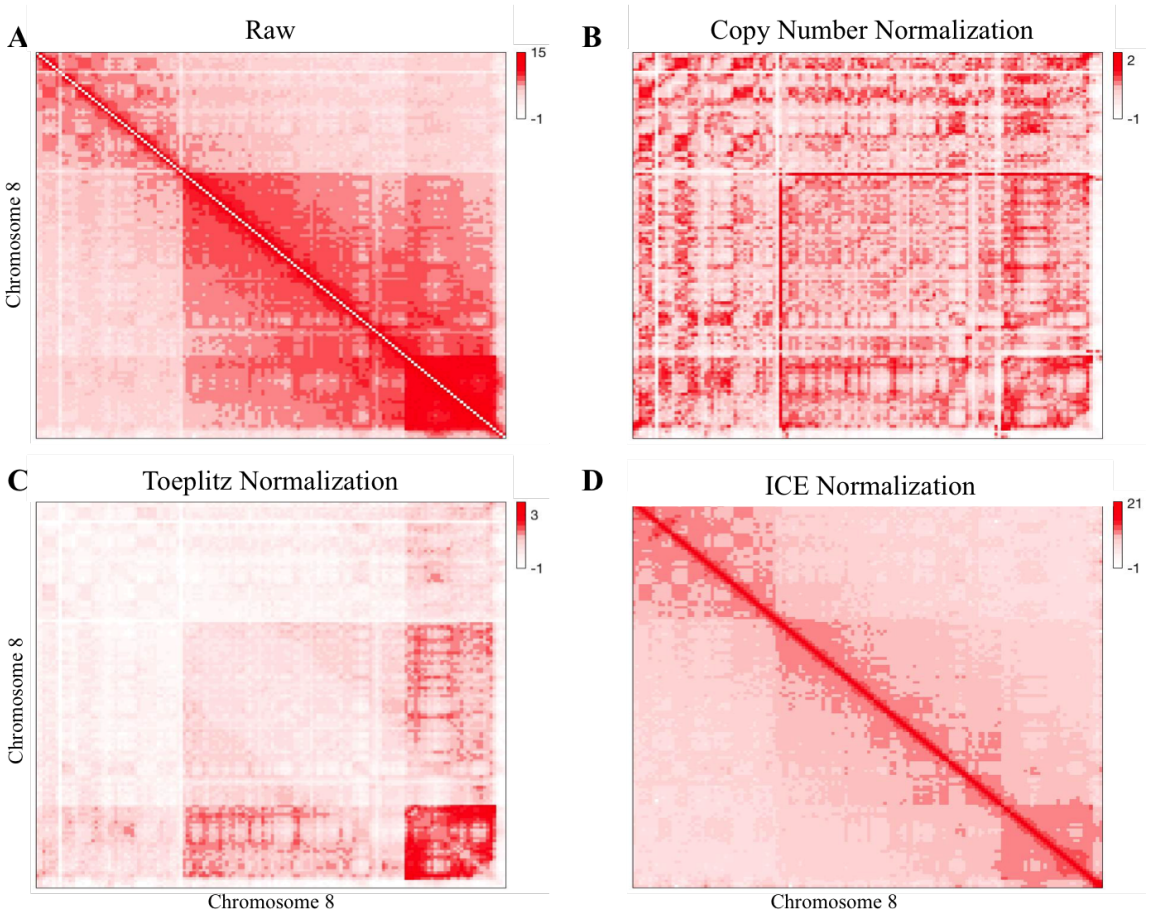

Figure S18: Comparison of normalization methods. A) The Hi-C matrix for chromosome 8 of the 2D12 hour sample of HT-29 in which three distinct copy number regions can been seen. B) The matrix after new block based normalization described in Figure 1E in which underlying patterns can be seen over the copy number blocks. C) The matrix after Toeplitz normalization [18]. D) The matrix after ICE [55]. The structure-function correlation as measured using block, Toeplitz and ICE normalization is 0.60, 0.57, and 0.16 respectively.

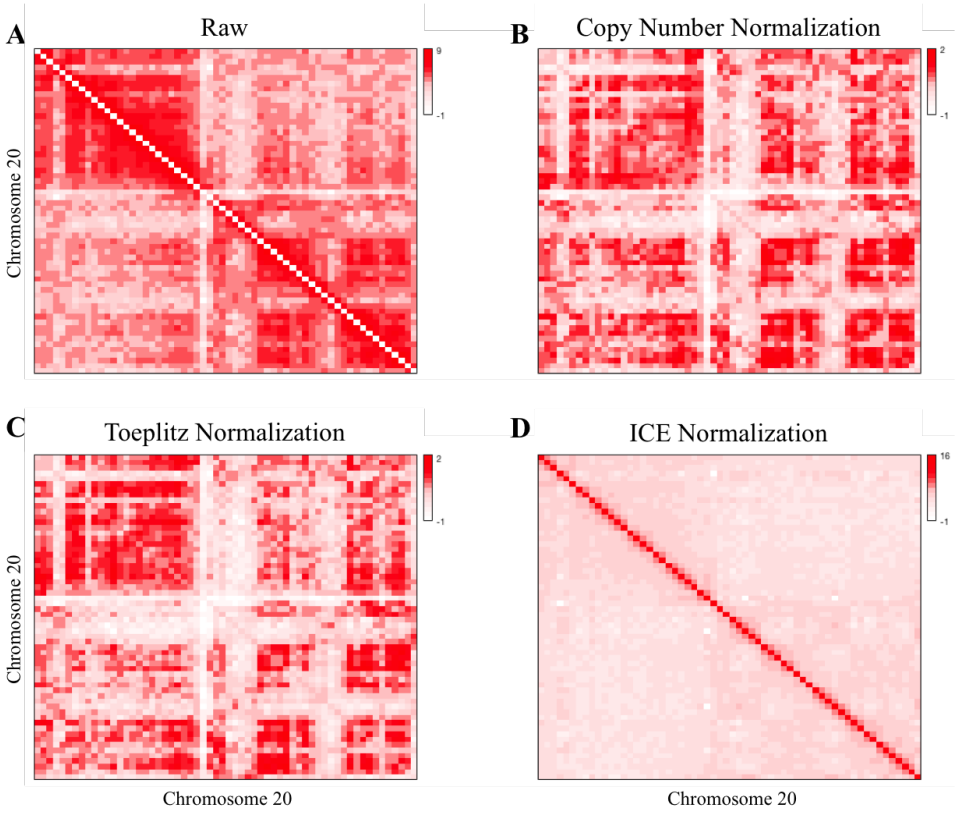

Figure S19: Normalization methods on K562 data. A) The Hi-C matrix for chromosome 20 of the K562 sample in which the first have is present at two copies while the second only has a single copy. B) The matrix after new block based normalization described in Figure 1E. C) The matrix after Toeplitz normalization [16]. D) The matrix after ICE [55]. The structure-function correlation as measured using block, Toeplitz and ICE normalization is 0.63, 0.59, and 0.20. respectively.

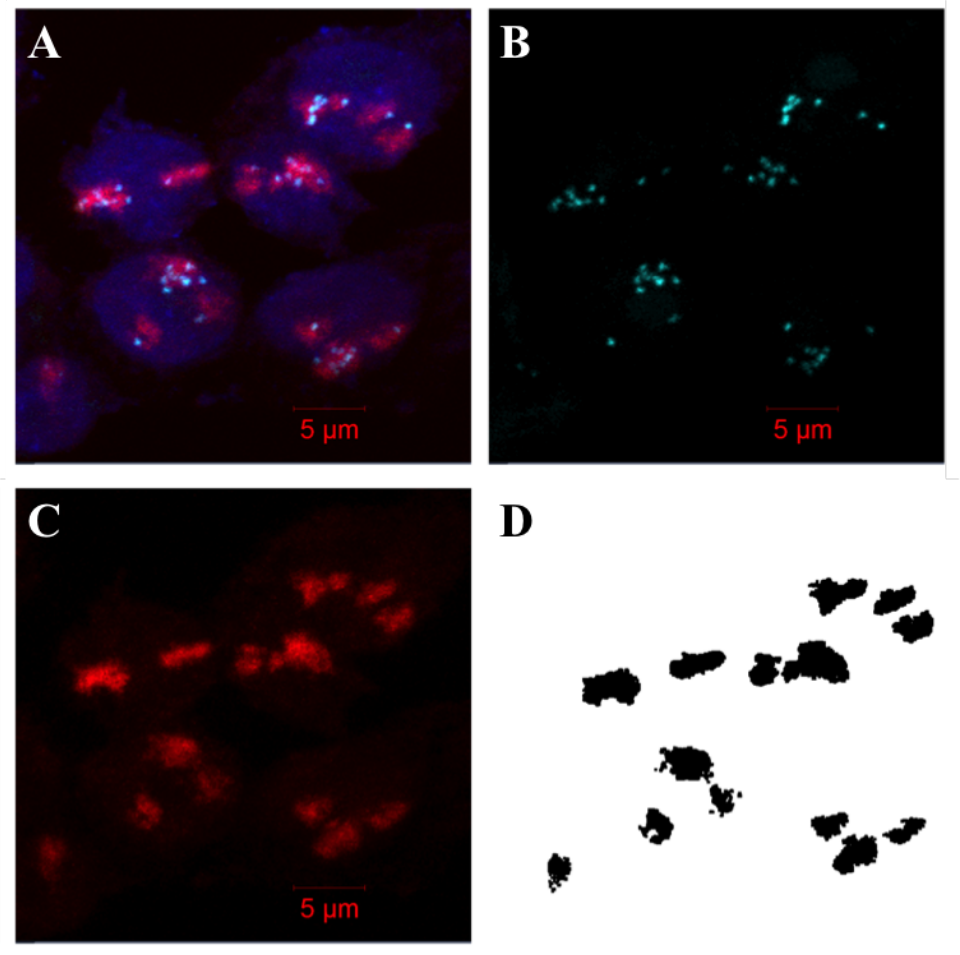

Figure S20: Measuring size of chromosome 8 territories. A) The overlay showing stained nuclei in blue, chromosome 8 territories in red, and *MYC* in bright blue/white. B) The locations of the *MYC* gene alone. C) The chromosome territories alone. D) The binarized image from which territory areas were calculated.

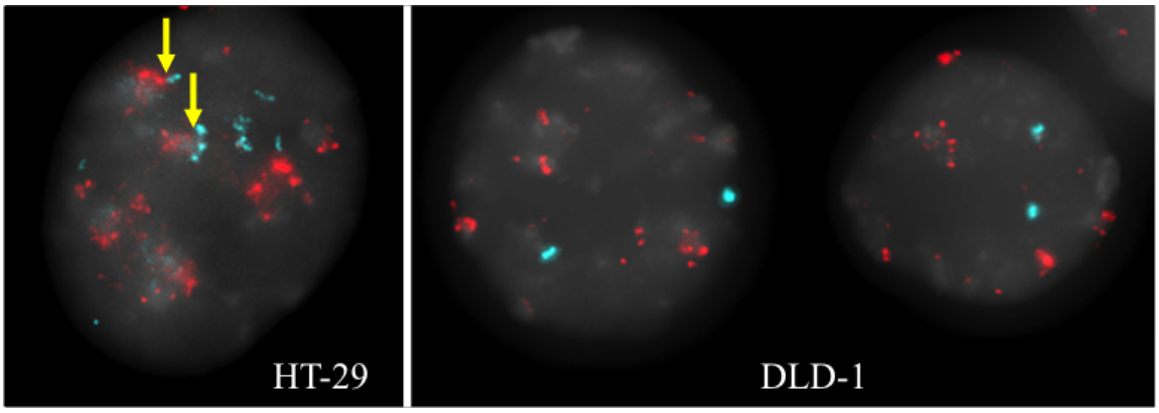

Figure S21: Interactions with the HSR for all samples. Graphs of the total genomic interactions for each interchromosomal bin against their interactions with just the HSR for the A) 3D12, B) 3D5day C) 2D12hr and D) 2D5day samples. The red line shows the best-fit line for a region's interactions with the HSR. The red point shows the point with the largest residual in each sample which in all cases is the amplified region on chromosome 2.

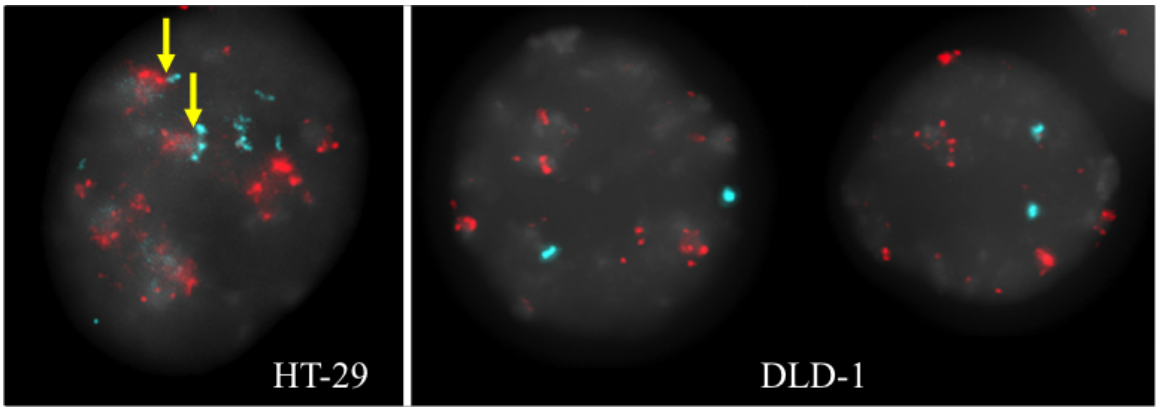

Figure S22: Interactions between the HSR and chromosome 2.The *MYC* locus within the HSR on chromosome 8 is shown in blue and a locus on chromosome 2 is shown in red. The chromosome 2 region was found to interact strongly with the HSR which is supported by these images that show that the region is amplified giving it more opportunities to interact with the HSR. The same loci are shown in DLD-1 where they are not near each other as a negative control.

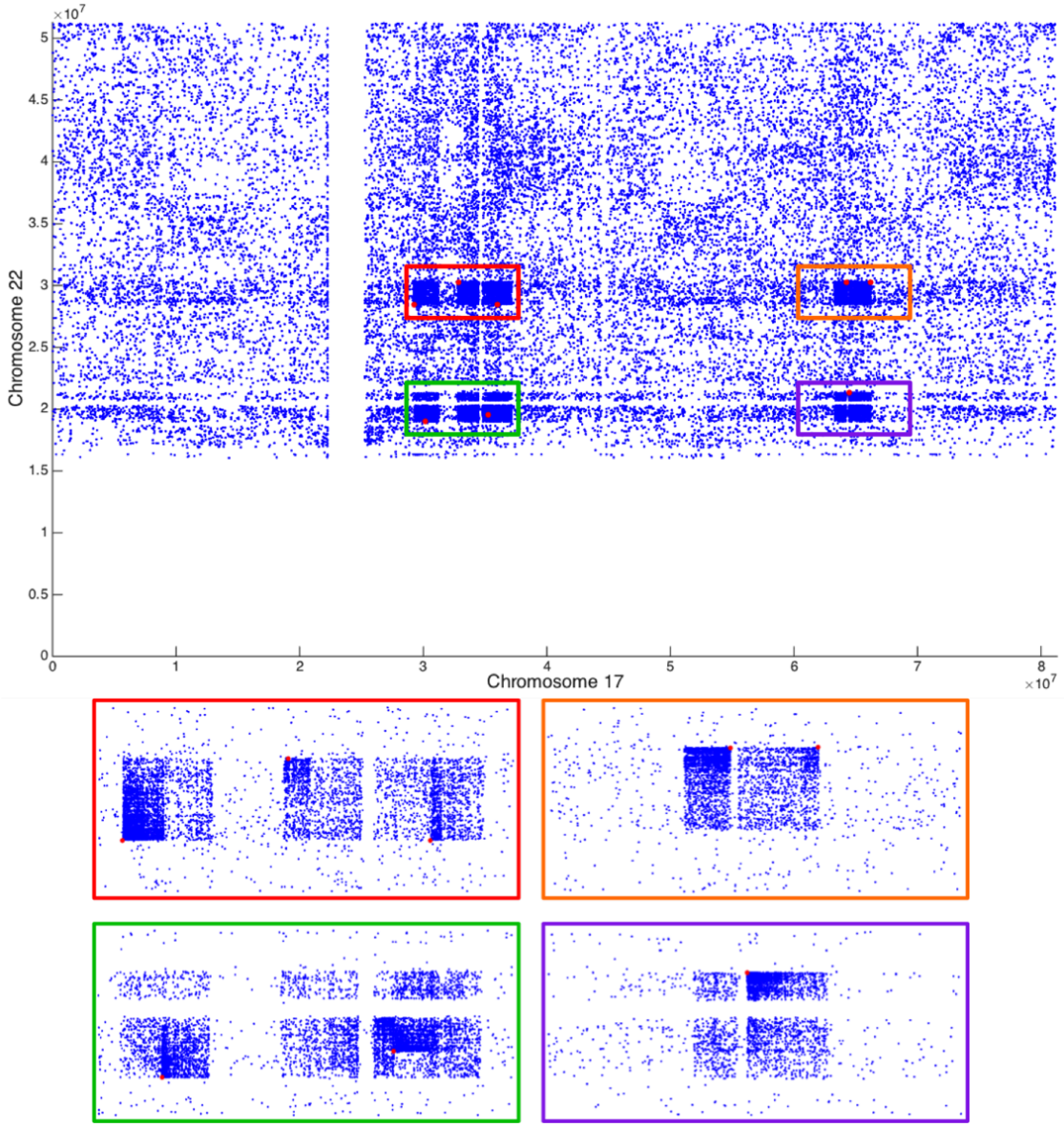

Figure S23: Read level interactions between chromosomes 17 and 22 indicate there several breakpoints between the chromosomes. The top shows all interactions between chromosomes 17 and 22. Each dot represents a single paired end read. There are at least 8 different translocations that occur between the two chromosomes each of which is marked with a red dot. The bottom images zoom in on the regions where the translocations occur. Due to the difficulty of knowing which ones combine together to create chromosomes which combination occur in the cells, this translocation was not analyzed elsewhere in the paper.

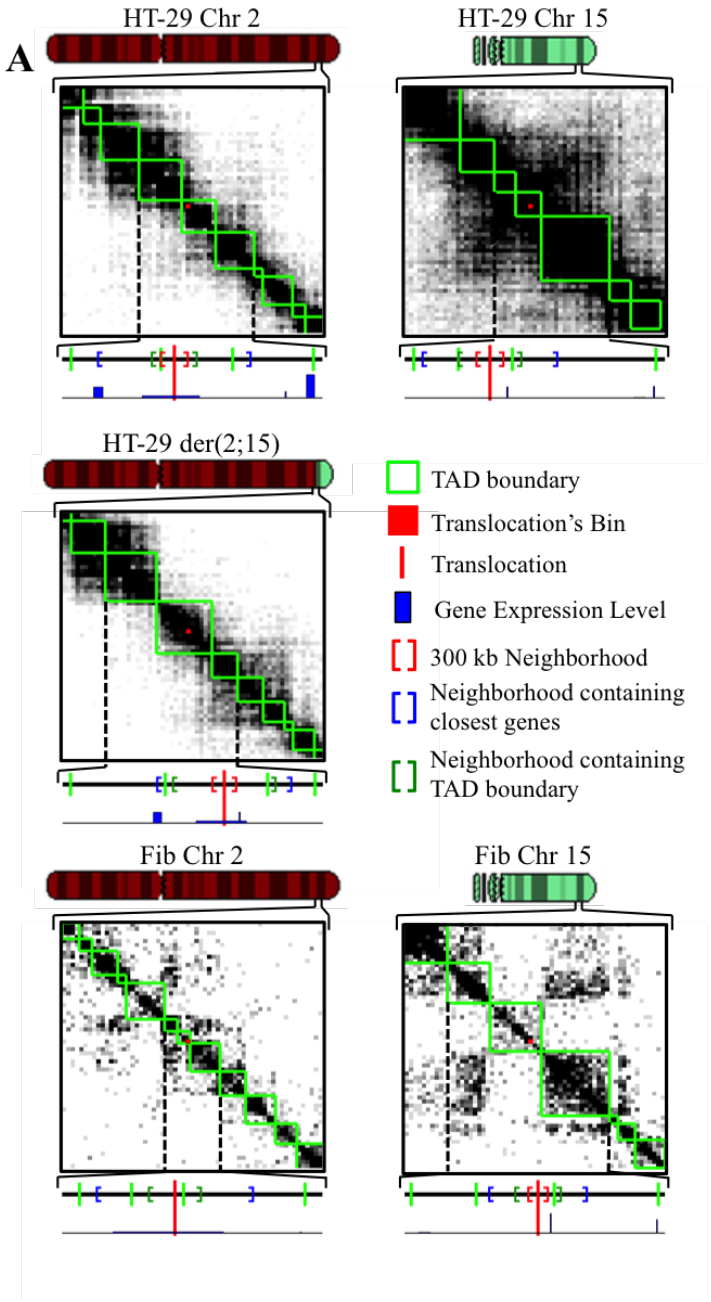

Figure S24: Structural stability and gene expression of der(2; 15) in HT-29. der(2; 15) and the normal copies of the chromosomes in CRC (HT-29 cell line), fibroblasts (BJ cell line). The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

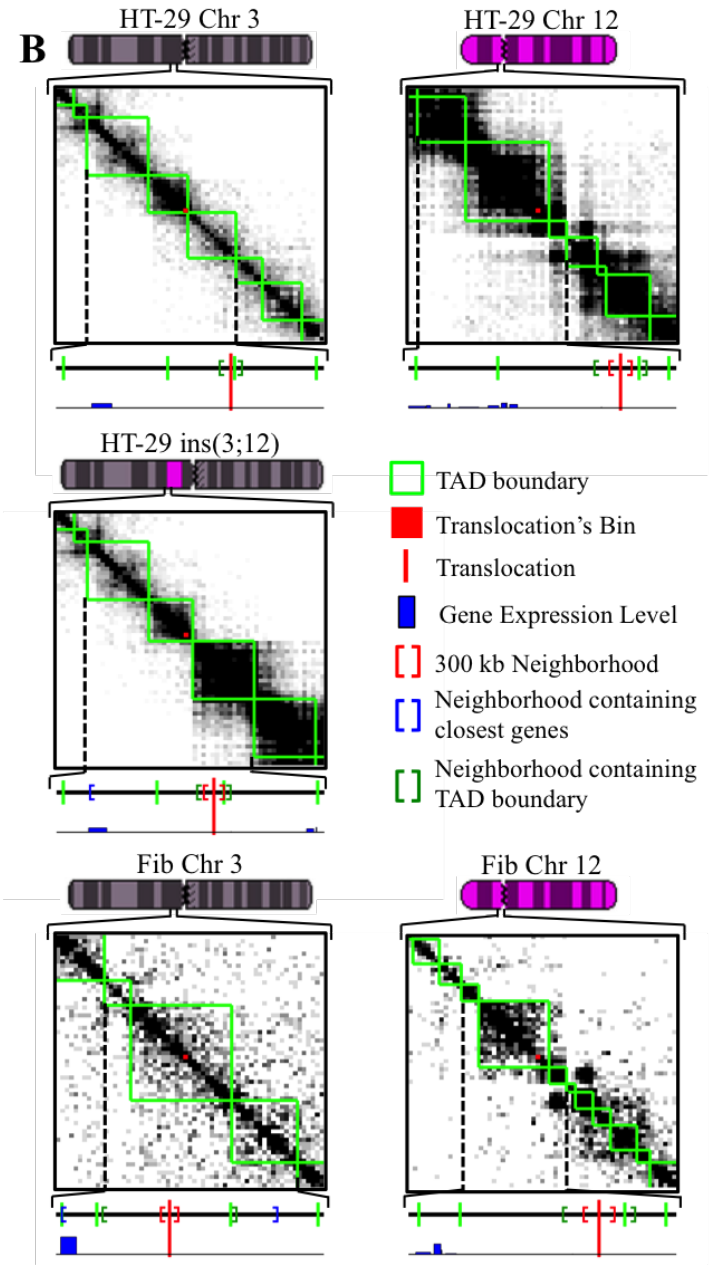

Figure S25: Structural stability and gene expression of ins(3; 12) in HT-29. ins(3; 12) and the normal copies of the chromosomes in HT-29 and fibroblasts for the first (more p-terminal) break in ins(3; 12). The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

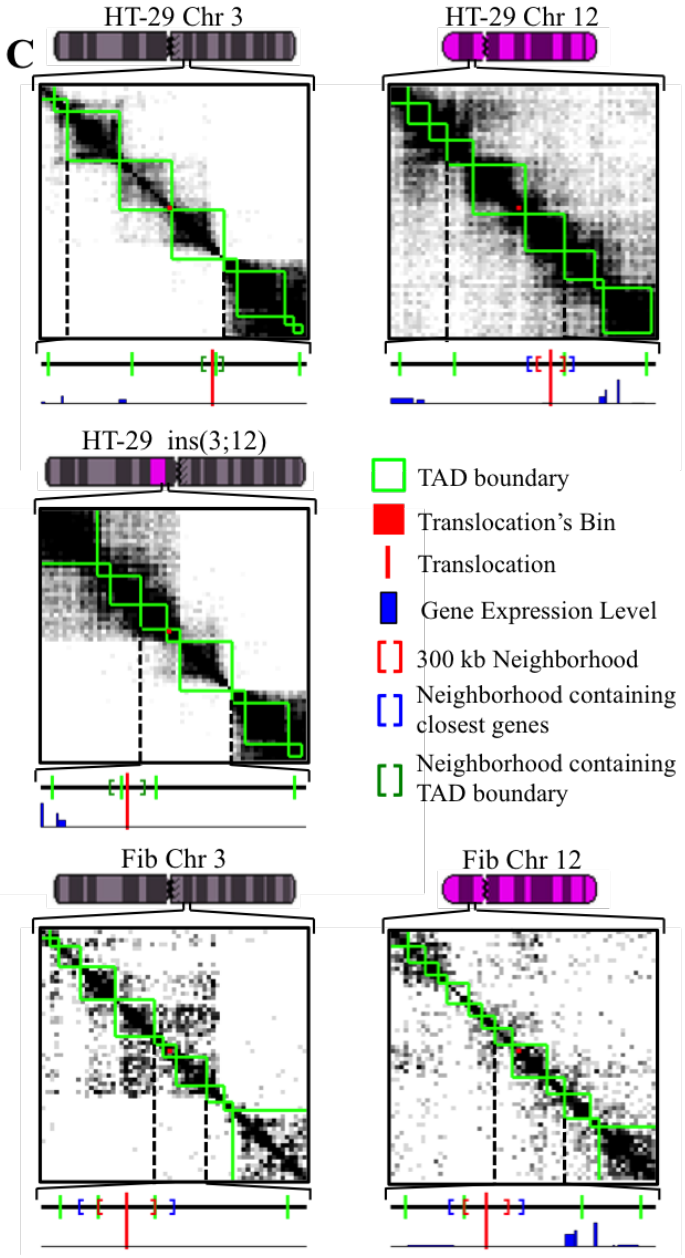

Figure S26: Structural stability and gene expression of ins(3; 1) in HT-29. ins(3; 12) and the normal copies of the chromosomes in HT-29 and fibroblasts for the second (closer to the centromere) break in ins(3; 12). The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

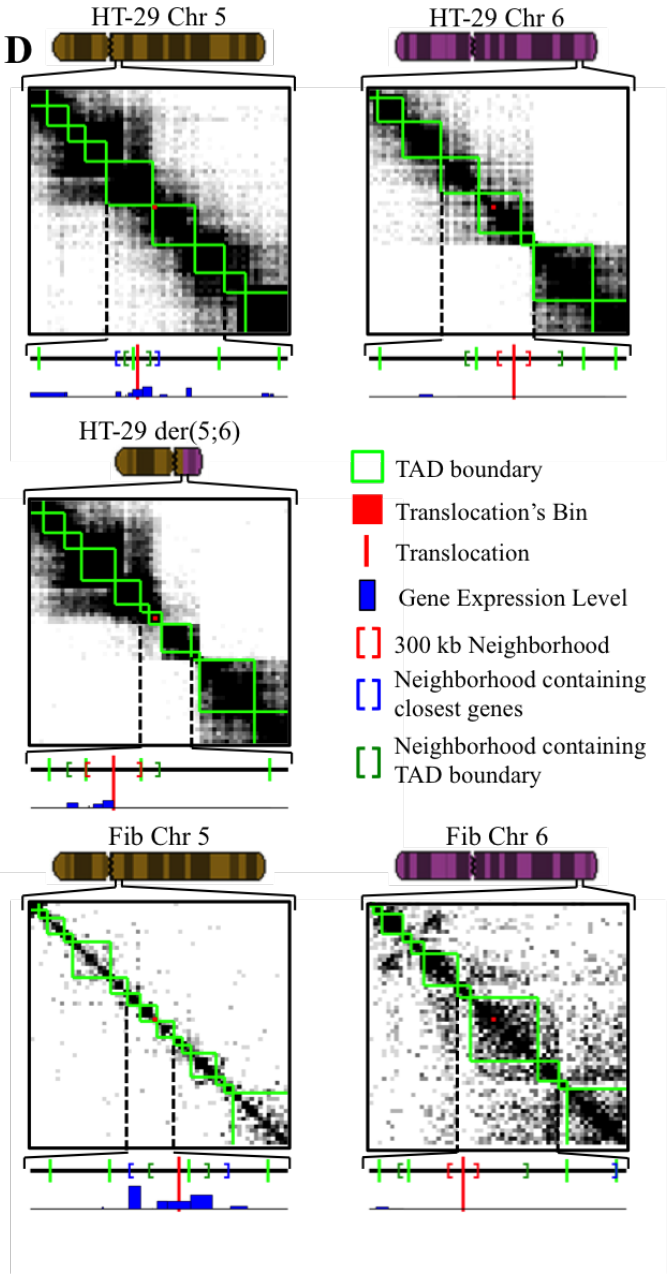

Figure S27: Structural stability and gene expression of der(5; 6) in HT-29. der(5; 6) and the normal copies of the chromosomes in HT-29 and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

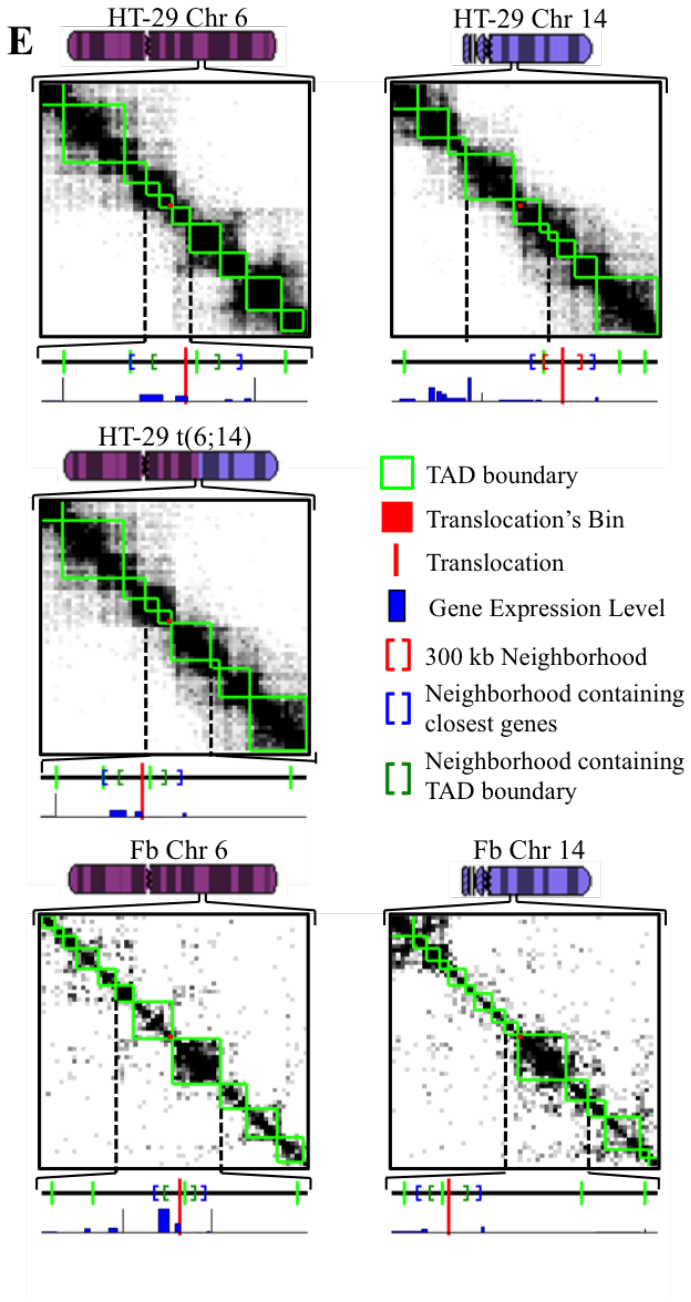

Figure S28: Structural stability and gene expression of t(6; 14) in HT-29. The $6-14$ chromosome of the seemingly balanced t(6; 14) and the normal copies of the chromosomes in HT-29 and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

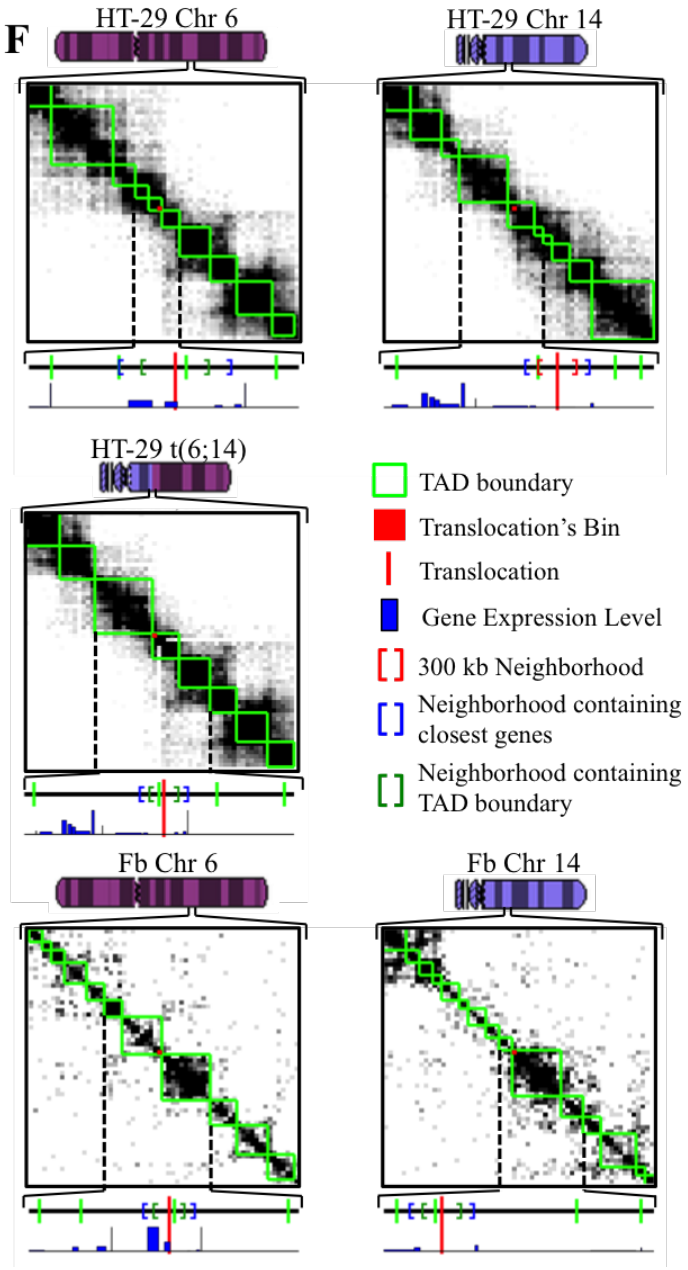

Figure S29: Structural stability and gene expression of t(6; 14) in HT-29. The $14-6$ chromosome of the seemingly balanced t(6; 14) and the normal copies of the chromosomes in HT-29 and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 aMb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

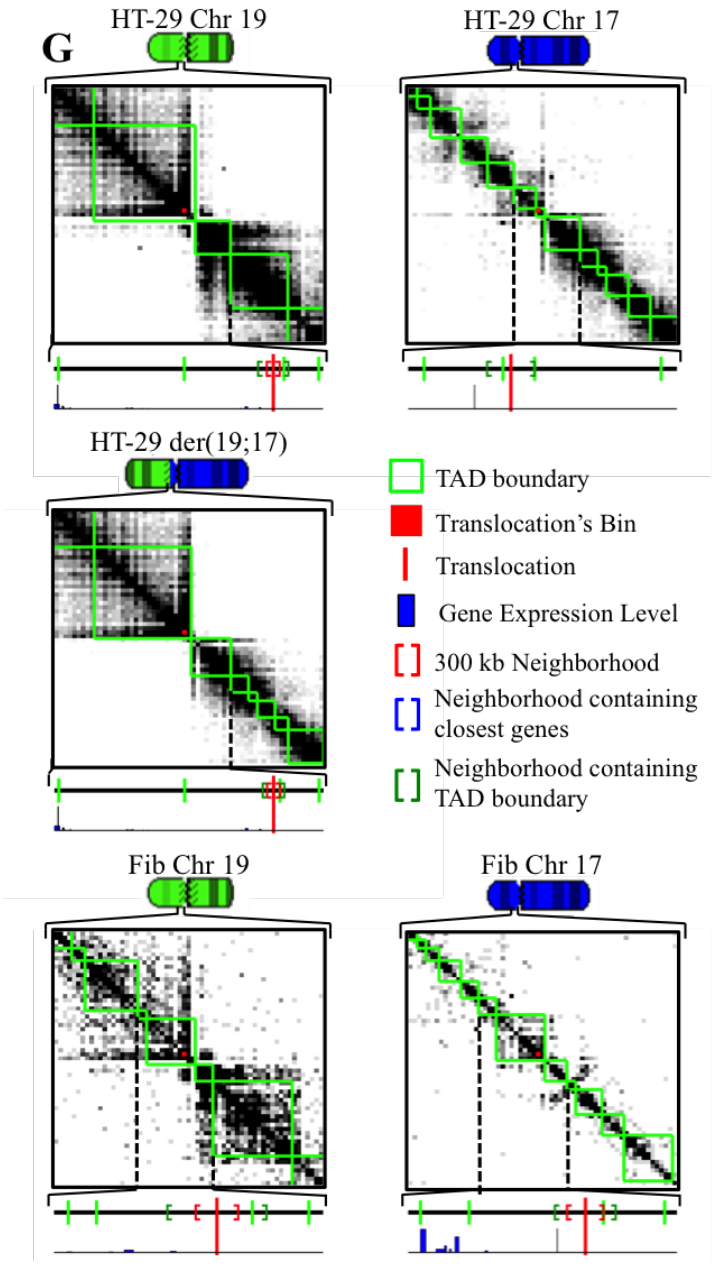

Figure S30: Structural stability and gene expression of der(19; 17) in HT-29. der(19; 17) and the normal copies of the chromosomes in HT-29 and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

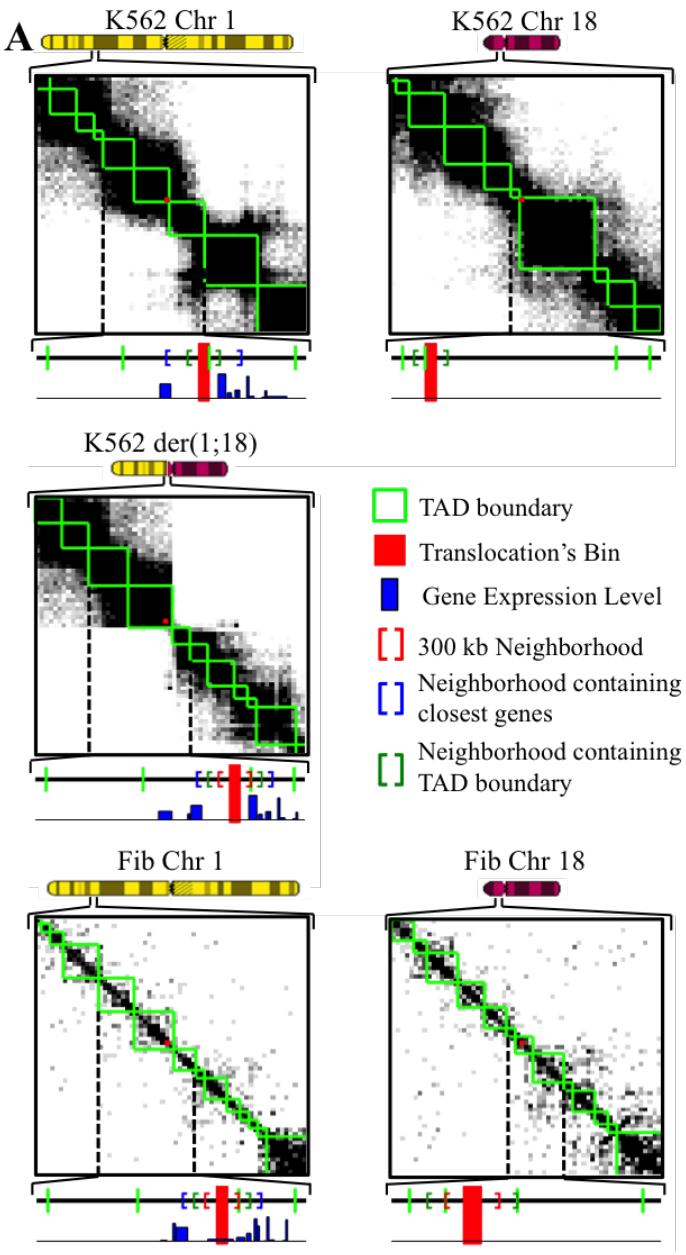

Figure S31: Structural stability and gene expression of der(1; 18) in K562. A) der(1; 18) and the normal copies of the chromosomes in chronic myelogenous leukemia (K562 cell line), and fibroblasts (BJ cell line). The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

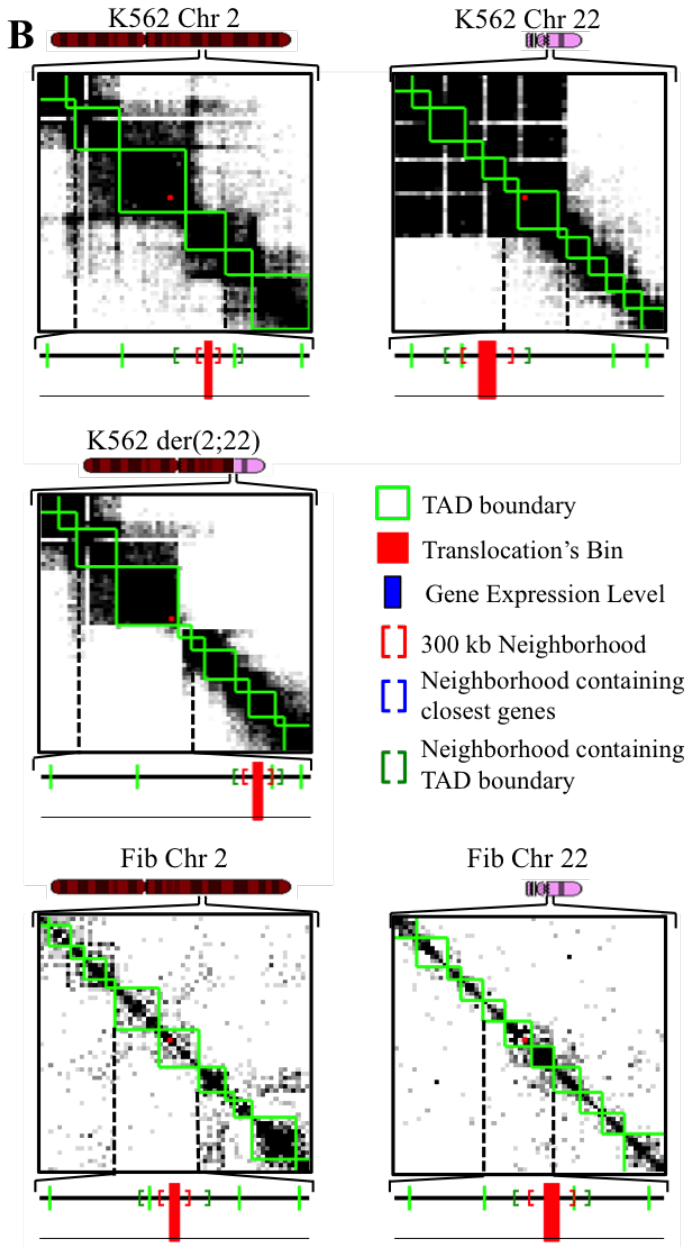

Figure S32: Structural stability and gene expression of der(2; 22) in K562. A) der(2; 22) and the normal copies of the chromosomes in K562 cell line, and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

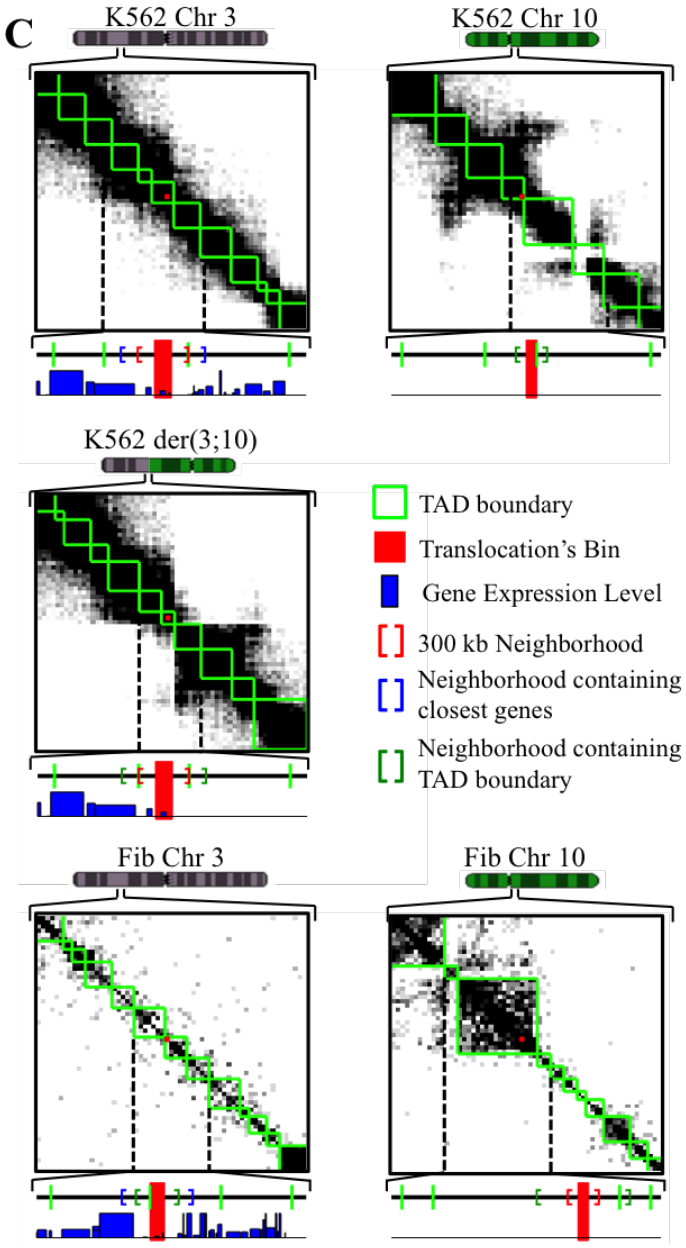

Figure S33: Structural stability and gene expression of der(3; 10) in K562. A) der(3; 10) and the normal copies of the chromosomes in K562 cell line, and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

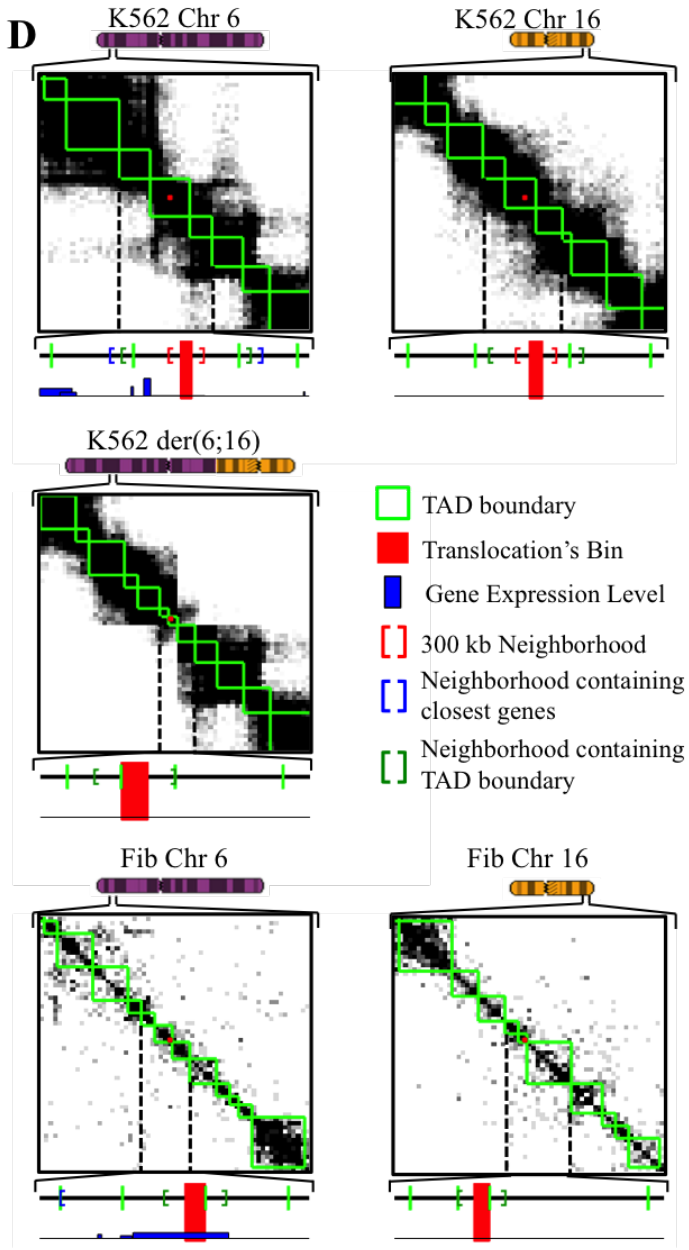

Figure S34: Structural stability and gene expression of der(6; 16) in K562. A) The first of two der(6; 16) and the normal copies of the chromosomes in K562 and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

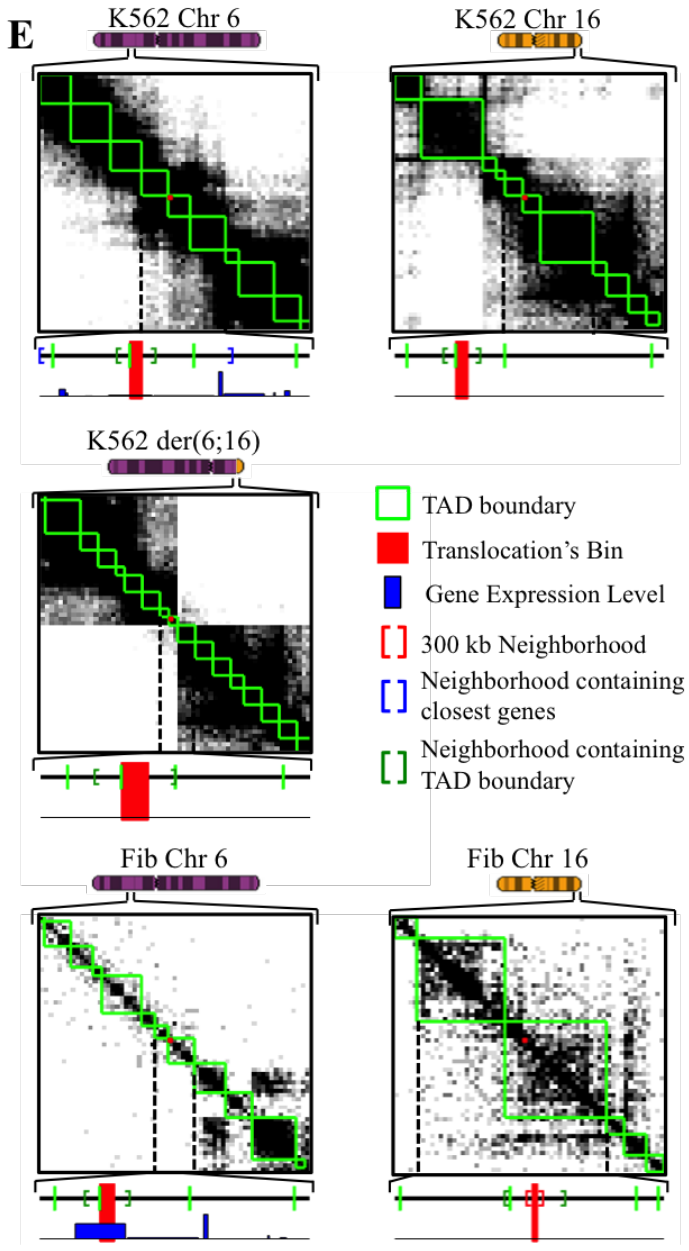

Figure S35: Structural stability and gene expression of der(6; 16) in K562. A) The second of two der(6; 16) and the normal copies of the chromosomes in K562 and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

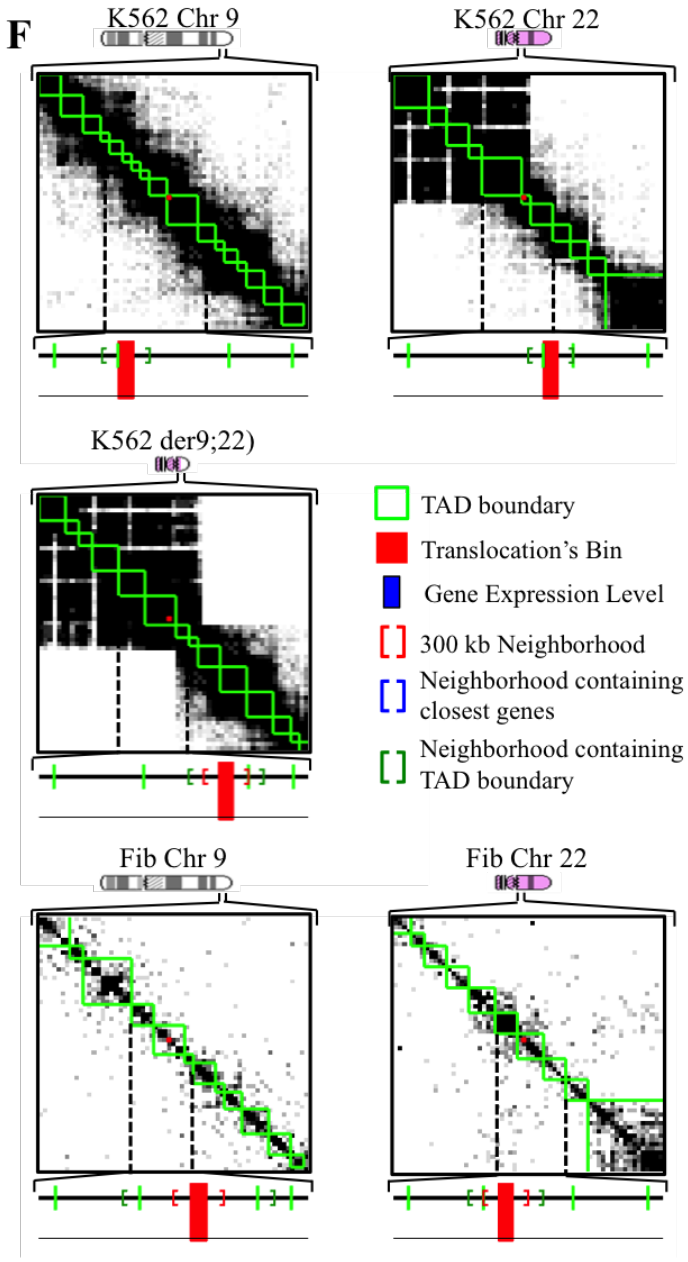

Figure S36: Structural stability and gene expression of der(9; 22) in K562. A) der(9; 22) and the normal copies of the chromosomes in K562 cell line, and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

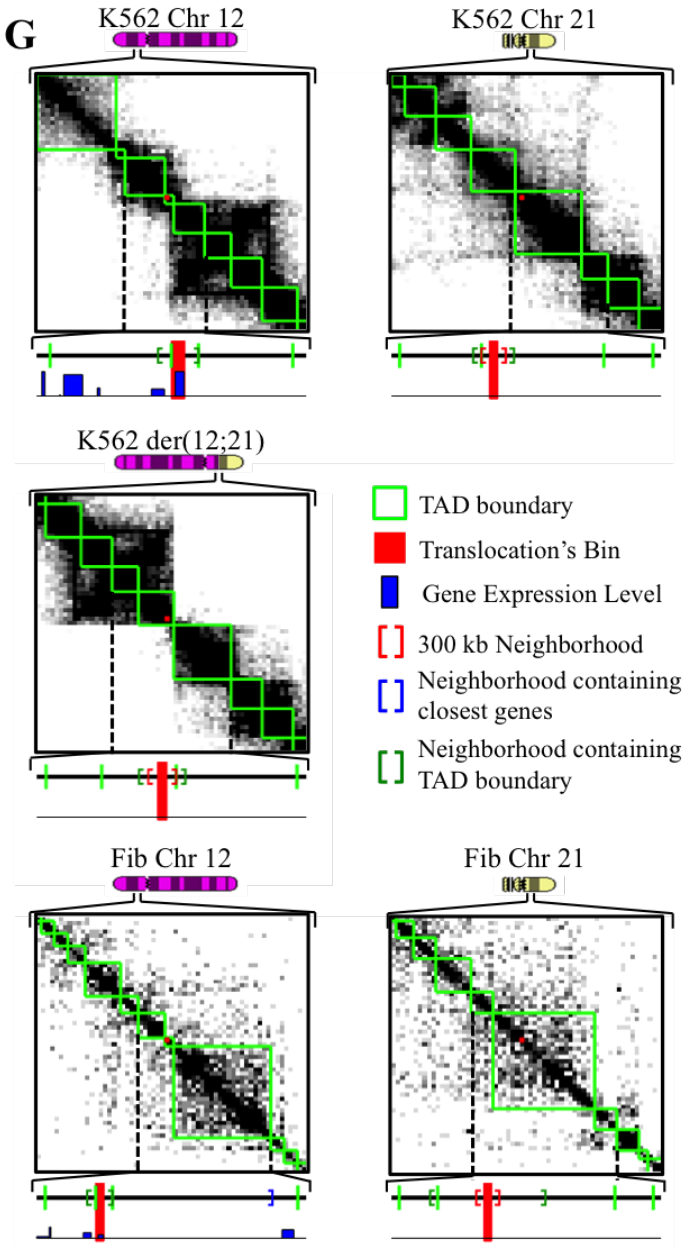

Figure S37: Structural stability and gene expression of der(12; 21) in K562. A) der(12; 21) and the normal copies of the chromosomes in K562 cell line, and fibroblasts. The heat maps show the raw Hi-C matrix at 100 kb resolution for a 3 Mb region centered on the translocation. The number line shows the site of translocation, TAD boundaries, and neighborhoods encompassing 300 kb, the closest TAD boundary, and a gene on each side for a region including 3 TADs. The blue bar plot shows the gene expression for each gene in the 3 TAD region with the boundaries of the blocks representing the edges of the genes.

A.6 Nucleome analysis: supplemental tables

Table S4: The threshold and number of segments for each chromosome during normalization.

Chr	Threshold	Num Segments
1	50	1
2	70	1
3	42	2
4	40	1
5	38	2
6	20	2
7	95	1
8	24	3
9	110	1
10	60	1
11	65	1
12	50	1
13	40	2
14	25	1
15	70	1
16	68	1
17	65	1
18	26	2
19	40	2
20	60	1
21	20	2
22	22	3

Table S5: Copy number, gene expression, and Fiedler number by chromosome arm for 2D12hr sample as well as the correlation between each pair (bottom).

Chr Arm	RNA-seq	Hi-C	CGH
1p	−0.15	5.3×10^{17}	1.06
1q	−0.22	1.2×10^{18}	1.00
2p	−0.31	1.2×10^{18}	1.02
2q	−0.31	9.8×10^{17}	0.99
3p	−0.44	2.3×10^{17}	0.77
3q	−0.28	1.2×10^{17}	1.12
4p	−0.24	8.5×10^{16}	0.96
4q	−0.52	1.9×10^{17}	0.88
5p	−0.32	4.3×10^{17}	1.10
5q	−0.46	1.0×10^{18}	0.96
6p	−0.24	5.1×10^{17}	0.97
6q	−0.64	6.7×10^{17}	0.89
7p	−0.14	1.4×10^{18}	1.22
7q	0.01	1.8×10^{18}	1.30
8p	−0.51	7.8×10^{17}	0.58
8q	0.11	7.4×10^{17}	0.96
9p	−0.15	8.6×10^{17}	1.61
9q	−0.25	9.3×10^{17}	0.89
10p	−0.32	1.1×10^{18}	0.97
10q	−0.30	7.0×10^{17}	0.96
11p	−0.05	6.9×10^{17}	1.26
11q	−0.22	2.9×10^{18}	1.21
12p	−0.37	1.0×10^{18}	1.03
12q	−0.14	7.0×10^{17}	0.99
13q	−0.30	3.0×10^{16}	1.21
14q	−0.43	1.2×10^{18}	0.72
15q	−0.16	1.2×10^{18}	0.31
16p	0.09	6.4×10^{17}	1.01
16q	−0.12	1.2×10^{18}	1.00
17p	−0.26	9.5×10^{17}	0.80

17q	0.03	1.1×10^{18}	1.16
18p	−0.54	3.7×10^{16}	1.13
18q	−0.46	4.1×10^{17}	0.72
19p	−0.31	1.6×10^{18}	0.87
19q	−0.11	1.5×10^{18}	1.33
20p	0.01	4.8×10^{17}	1.07
20q	0.26	1.2×10^{18}	1.45
21q	−0.24	7.3×10^{16}	0.67
22q	−0.25	1.0×10^{18}	0.91
RNA & Hi-C	0.269		
Hi-C& CGH	0.275		
RNA & CGH	0.648		

Table S6: Chromosome territory quantification. The size of chromosome territories in μm^2. Two sets of single territories were unable to be separated and treated as a small territory in analysis.

Nucleus	Area μm^2	Description
1	3.8	HSR
	1.4	
	1.5	
2	5.0	HSR
	0.3	
	1.4	
3	5.3	HSR
	3.2	
	3.6	
4	8.5	HSR
	3.0	2 single
5	4.8	HSR
	2.2	
	2.7	
6	6.9	HSR
	4.5	2 single
7	5.3	HSR
	1.7	
	2.8	
8	4.0	HSR
	2.2	2 single
9	4.5	HSR
	0.9	
	1.4	
10	6.8	HSR
	1.3	
	2.5	

Table S7: Translocations in K562.

Chr	Loc	Entropy Fib	Entropy K562	S-F Fib	S-F K562
$9-22$	237		1.80		0.16
9	1137	2.49	1.65	0.03	0.40
22	237	2.29	1.51	0.77	0.33
$6-16$	857		2.07		0.67
6	168	2.92	1.78	0.04	0.85
16	857	2.86	1.95	0.10	0.00
$6-16$	1130		2.02		0.55
6	382	2.78	1.72	0.37	0.16
16	783	2.62	1.96	0.55	0.56
$12-21$	1112		2.80		0.42
12	228	3.10	2.39	0.63	0.00
21	256	3.10	2.17	0.32	0.00
$3-10$	483		1.89		0.69
3	483	2.50	1.84	0.26	0.47
10	879	2.21	1.56	0.83	0.69
$2-22$	518		1.86		0.36
2	1518	2.60	1.40	0.59	0.94
22	229	2.44	1.50	0.70	0.13
$1-18$	553		1.05		0.15
1	553	1.84	1.06	0.70	0.08
18	221	1.78	0.95	0.06	0.28
No Trans Avg		2.54	1.68	0.43	0.32
Trans Avg			1.93		0.43

Table S8: ANOVA genes that changed between 2D and 3D samples.

TRABD2A	GPR128	DONSON	RDH11	SLC6A8
STRIP2	CENPK	CENPL	TRMT11	LIN54
LOC654433	RNF144B	RAD54L	CCNE2	GINS4
KRT18	HEXA-AS1	KRT8	WDHD1	ARNTL2
MYT1	UBQLN1	ADAMTSL2	NKD1	IL1R2
FGFR2	RRM2	PTPRO	E2F7	NCAPG
ARHGAP11A	CEP78	HAT1	GINS1	RRM1
CANX	RBMS2	ANKRD32	BIRC5	KLK12
CYFIP2	KIAA1211	RAD18	MT1B	SHCBP1
LINC00857	FAM105A	CERK	ADH6	NEIL3
MTHFD2	DCLRE1B	MTFR2	SKP2	WDR76
TPPP3	EPC1	RFC3	SRP54	CRLF3
ESCO2	CFTR	ARL4C	SCAMP5	DHFR
GAS2L3	KIF11	SKA1	HNF4A	CDK1
PKD2L1	SLC11A2	S100A16	TYMS	SKA2
DNAJC21	SNRPD1	MAST4	CCNB1	IBTK
PLEKHG4	PLK4	ATF4	SYP	ACTL7A
ANLN	TTK	HFE	ZMIZ2	KIF18B
KIAA1524	HPRT1	ECT2	STARD8	MAD2L1
MLK7-AS1	TCF19	TUBB	SASS6	PDIA6
MIS18BP1	GPR126	KDM1B	PSMD10	RANBP9
ATP6V0A1	KIAA1875	PCNXL2	UBR7	RAD51AP1
FGFBP1	FBXO5	ASPH	SPATS2L	TSPAN32
TYRO3	COQ3	ADCY5	PDE2A	HPS3
SLC39A10	DCBLD2	SLC10A2	SLC31A1	RARRES1
NUSAP1	SBF2-AS1	TMEM178B	TMEM198	CTTNBP2
TMEM106C	IPO8	HSD11B2	IDUA	DEPDC1B
CENPI	SLC25A43	PLEKHB1	C15orf41	AADAC
LINC00669	KIF18A	KIF14	KPNA2	GTSE1
EPC2	CDC7	CCNA2	BRSK2	DNA2
FLJ12825	SPC25	NCAPG2	DTL	KANSL1
C1orf112	C5orf34	AHCYL2	NUP43	HMGB1
LOC93432	ADRBK2	MT1H	SEMA3E	XKRX

RAD54B	CENPQ	CHEK1	ZDHHC1	EPOR
SLC23A3	IYD	ASPM	NUP37	DEK
PBK	MNT	IAH1	EXO1	PPM1H
NXF4	UBE2V2	KLF15	BTBD11	CHST3
KIAA0101	UGT2B7	DEPDC1	WNK4	TXNDC17
MASTL	PPM1L	GPRIN3	NUPL1	CLK1
PUM1	ALDH7A1	NOSTRIN	CNIH4	DUOX1
LOC286467	MELK	NOX1	MIS18A	GPN3
HAUS6	NOTCH1	MFSD6	BTBD9	ANP32E
FLJ13197	AJUBA	ZNF367	GPSM2	CBFA2T3
RNASE4	MECOM	RNF103	CLSTN3	SMC4
NAALADL2	FOXL1	PRPS2	SKIV2L2	STT3B
SEMA3A	KAP2	BRIP1	TNFAIP8	H2AFV
RNASEH2A	PIGR	R3HDML	STMN1	MAD2L2
IHH	CENPH	PCLO	AP5Z1	SGOL1
HNRNPH3	TUBA1B	CENPA	SLC26A1	PER3
PSMA3	RBL1	TNFRSF11A	NPM2	PRIM2
SPATA25	SENP1	HMGB3	AURKA	FGD6
FBXL19-AS1	MUC1	MAP2K6	LMNB1	BUB1
NUP205	PROM1	NR5A2	RTTN	C9orf64
ATAD2	BMP7	IDNK	FFAR2	ACSL1
LOC100287314	ATP1A3	BARD1	TFRC	PGAM1
TTC3P1	RHNO1	TLE1	POLE2	SUV39H2
LOC253039	POC1A	MAP3K6	CDK2	RASD2
GALNT8	NUCKS1			

Table S9: ANOVA genes that changed between samples with 12 hr or 5 day growth.

RNASE4	PRPS2	IL13RA1	STIL	CSPG5
TRABD2A	GPD2	PRIM2	MCM6	USP14
ESCO2	NCAPG2	CENPO	KLHDC2	DEPDC1
WDHD1	TPPP3	CCNE1	CENPP	CSE1L
E2F7	TTK	IL18	LMNB1	MTFR2
KIAA1524	FGFBP1	CEP78	TIMM8B	INTS7
LINC00857	RBL1	HMGN1	RDH11	PCNA-AS1

PHKB	PTPRB	HMGB1	S100A2	RGS7
DTL	SLC39A10	TYMS	SMC4	LMNB2
DNA2	CHEK1	ZDHHC6	CHST3	TRA2A
C1orf112	KIAA0101	GEN1	CASC5	WNK4
IL1R2	BARD1	SKA3	DHRS2	NAMPT
KRT18	EXO1	DDX5	TRMT11	NUP43
RRM2	SKA1	NXF4	SLC10A2	DNMT3B
DHFR	MELK	LRP8	APAF1	ABCB8
HPRT1	NUP205	PLK1S1	PSMA1	GMFG
CCNE2	ARNTL2	TUBA1B	ZWILCH	HMGB3
CENPI	RELT	LIN54	GINS4	C6orf48
RAD54B	ATF4	TK1	SUV39H2	SKP2
PGAM1	CENPK	TNFAIP8	BRCA2	FAM192A
ATAD2	CPOX	C4orf21	CKAP2	UBR7
ANLN	MCM4	RRM1	NUP155	RPP30
TCF19	CENPQ	E2F1	STK3	DLD
SKIV2L2	DCLRE1B	SLC39A8	UHRF1	DEK
CDC7	SGOL1	KIF18B	PGM2	CDC6
CDK2	E2F8	RTKN2	VPS29	HNRNPU
ZNF367	SNRPD1	NEIL3	SH3KBP1	CDK1
RAD54L	LDHA	NUPL1	ASPH	PPME1
FBXO5	HERPUD1	UGT1A6	PBX2	SHCBP1
MASTL	DONSON	GTSE1	SF3B3	GINS1
CCNA2	LIPG	ADPRM	TUBB	ISPD
PKD2L1	EIF1B	TXNDC17	LGALSL	MAD2L1
ARRDC4	DCBLD2	RNASEH2A	MILR1	ITM2C
NCAPG	FAM111B	TMOD3	ANP32E	ASPM
SPC25	HFE	C9orf64	ESPL1	CDC5L
DERA	HAT1	LYRM1	SPATS2L	LUC7L2
XPO1	KIF18A	IL36RN	RAD18	BLM
PLK4	HELLS	SMC2	USP28	ZDHHC1
KIF11	DARS	ENO1	POLR2G	YWHAB
WDR76	NEO1	MGST1	SNRPF	PSMC2
SPRY4	TCN1	RMI1	PSMA2	CPSF3
STRIP2	E2F2	SASS6	BBX	IARS2
PBK	C15orf41	GPR110	MIS18BP1	MSH6

BRIP1	CEP76	STMN1	MCM5	ANKRD32
RFC3	DSCC1	FAM115C	UNC13D	SKAP2
PKD1P1	KLHL24	CYB5R4	KIF14	CENPM
NAGPA	FANCL	NCBP1	SQLE	UBALD1
KCNMB4	PANK1	MT1B	HADH	RMI2
KIF20B	VIMP	CCNB1	CLOCK	PTAFR
GSTCD	APOL1	C19orf66	SYNGAP1	SDCBP
NCAPD3	PM20D2	TPP1	AURKA	LIN9
DHRS9	COX20	CENPH	API5	SRGAP3
TRPV1	NUP107	NHLRC4	FBXL5	ACER3
IL1RAP	AJUBA	CACYBP	RASSF4	MTHFD2
PPIH	HMGB2	FBXO32	SBF2-AS1	CDCA2
CENPL	TRA2B	BRCA1	SMAP2	BORA
BIRC5	BANF1	HMGN2	IDUA	COQ3
PCNA	POLQ	ACTL6A	EI24	FOXM1
PGD	TPM3	GINS3	XYLB	ERGIC2
VBP1	NUP50	HMGCR	SENP1	ERCC6L
CIRBP	UBE2V2	MCM2	CCDC109B	IPO11
C7orf41	SMC1A	MTMR10	SLC11A2	INCENP
DVL3	SNRPG	XRCC2	UNKL	NUCKS1
PGPEP1	KLHL3	ZFP91	KIAA1875	SLC25A42
SLC9A2	POC1A	MRAP2	IPO9	RRAGB
TEX30	PARPBP	GATS	RABAC1	MMS22L
ATP5G1	CLCC1	DAZAP1	C10orf91	NAP1L4
GGH	RPE	C12orf36	RPA1	RHOB
S100A16	NIT2	TRIP13	WHSC1	CXCL11
RNASEH2B	ZRANB1	HMGCS1	PSMD11	DQX1
CLSTN3	CDC45	ALDH9A1	CELF6	CCDC126
CAPNS1	BRI3BP	PIR	DLGAP5	FMR1
CENPW	PPP2R1B	FANCA	CFL1	RALB
HAUS6	FAM120A	HAS3	SMARCA5	RPS6KA4
PKNOX1	TP53INP1	GPHN	SOSTDC1	HNRNPM
NDUFA9	GPR115	FAM83D	TUG1	VWA5B2
IDS	SKA2	TPT1-AS1	RQCD1	TMEM194B
PTGES3	EIF5A	PPIA	LOC729966	SRD5A3
PKHD1	NUDCD1	RACGAP1	ACTL7A	MYBL2

CTNS	TFDP2	GGT7	C14orf80	NMU
COQ2	ZNF860	NUP37	LDHB	HPSE
HMMR	CDCA4	MALL	PSMD14	CDCA5
WDR81	RASA4	HNRNPC	GAS2L3	WDR54
CPSF1	TFDP1	GOT2	CHAF1B	THNSL1
KDM6B	C17orf103	HNRNPL	PDPR	PTGES
PROCR	TICRR	ABCG2	FBXO4	GPR153
LRRCC1	BTBD11	ASRGL1	SENP7	SLC4A11
SGOL1-AS1	KLF15	TMEM5	GPR126	NICN1
UBQLN1	RNF41	CEBPD	AADAC	KLHL2
STT3B	SAE1	NRGN	NRM	SYTL5
DHX15	ACOX3	C5orf34	AP4B1	SRGAP2
APOE	EED	WDR67	FHL3	CST6
NPAT	CDC25B	CNIH4	CYP1A1	KNTC1
EXOSC3	FFAR2	MAP2K1	RPAIN	ENOPH1
DNAJC21	SYTL4	ECT2	NDUFA6	XPO7
ASCC1	KIF23	SLC16A14	SNORA40	METTL21D
SPATA20	RBM8A	FIGNL1	SAAL1	DDX46
ACAT2	FAM111A	CAPRIN1	HJURP	CLCN7
WDR62	COX7B	PI3	CKS2	RB1
BUB1B	AGPAT5	ZNF655	EIF1	LINC00669
ADK	TTC9	H2AFV	PHLPP1	CEACAM19
BMP7	PRTFDC1	ACOT7	RASA1	LOC154761
NCAPH	LRR1	MIS18A	CDKN3	HIST1H2BK
HSPE1	IHH	PRDX1	CANX	LOC100287314
MRPL51	HSD17B7	POLA1	CDCA3	APOBEC3B
RNF123	PACS2	BUB3	C6orf15	KIAA1841
SCEL	VRK1	RNF169	CAMLG	MAPKAPK2
XRCC5	UBA6	RPA3	EXOC6	LOC100505666
FAM105B	MATR3	WDR45	IL1RN	LOC286467
ZNF215	LMO2	GLO1	WDR6	TMEM106C
DPH1	GPSM2	MT1F	ATP5J2	KRT8
MTBP	GAS5	DEPDC1B	YWHAQ	MAPRE1
USP1	JAK3	MTFP1	MEF2D	SPRY2
MCMBP	PAICS	YBX1	ME2	RAD51
NDC80	BUB1	POLE2	ZW10	NUSAP1

MT1H	ADORA2B	KLK12	ANKRD22	KPNA2
PFKFB4	SEMA3A	PIK3CG	POLA2	PLK1
DPP3	ENOSF1	ZNF525	CENPA	TPMT
F8	EZH2	RFC4	CTSE	FRRS1
OLA1	RAD51AP1	CKLF	PSMA3	NUP188
TFR2	TUBGCP3	RASD2	PDK1	POLD2
COTL1	MLF1IP	TMEM48	RBBP7	FPR3
ARHGAP11A	TMEM167B	WNT2B	EPT1	MMP7
LOC100130744	DYNLT1	RFC5	CACNA2D2	PAIP2
LOC100507118	C11orf82	FANCB	HNRNPD	FEN1
ZNF283	PKM	YWHAE	SFXN2	ARL4C
TIMELESS	HINT1	ANP32A	GINS2	ALDH3B2
TMPO	PIK3R3	UBE2T	EGR2	IDI1
BTBD10	VPS4B	SFR1	EPOR	PGP
G3BP1	FDPS	CSTF3	FIBP	UBASH3B
THSD4	DTYMK	DHCR24	XRCC3	NDUFB2
HNRNPA2B1	PRPS1	TTC26	FH	

Table S10: Gene ontology enrichment terms for genes differentially expressed between 2D and 3D growth.

GO term	No. genes	Bonferroni P
Cell cycle process	27	7.91×10^{-7}
M phase	19	1.31×10^{-6}
Cell cycle	29	2.81×10^{-6}
Mitotic cell cycle	19	8.09×10^{-6}
Cell cycle phase	20	1.08×10^{-5}
Cell cycle checkpoint	10	4.92×10^{-5}
Mitosis	15	1.06×10^{-4}
M phase of mitotic cell cycle	15	1.28×10^{-4}
Regulation of progression through cell cycle	18	3.22×10^{-4}
Regulation of cell cycle	18	3.43×10^{-4}
Regulation of mitosis	9	9.44×10^{-4}

Table S11: Gene ontology enrichment terms for genes differentially expressed between 12 hr and 5 day growth.

GO term	No. genes	Bonferroni P
Cell cycle	66	5.45E-16
Cell cycle phase	44	2.99E-14
M phase	39	3.80E-14
Cell cycle process	56	1.06E-13
Mitotic cell cycle	39	2.01E-12
DNA metabolic process	59	5.29E-11
DNA replication	31	8.19E-11
Cell division	31	4.86E-10
Mitosis	29	3.11E-09
Response to DNA damage stimulus	35	3.40E-09
M phase of mitotic cell cycle	29	4.53E-09
Cell cycle checkpoint	17	1.33E-08
DNA repair	31	1.35E-08
Response to endogenous stimulus	37	7.10E-07
DNA-dependent DNA replication	17	8.60E-06
Regulation of progression through cell cycle	31	2.58E-05
Regulation of cell cycle	31	2.86E-05
Interphase	16	2.58E-04
Interphase of mitotic cell cycle	15	9.61E-04
Regulation of mitosis	12	3.09E-03
Mitotic cell cycle checkpoint	9	5.97E-03
DNA recombination	14	1.04E-02
Regulation of DNA metabolic process	10	1.78E-02
Cellular metabolic process	221	3.34E-02
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	120	4.59E-02

A.7 Nucleome analysis: supplemental methods

A.7.1 K562 analysis

K562 data used were publicly available from GSE63525 (Hi-C) and long RNA from ENCODE (GSM765405). In order to make the analysis more comparable with our previous analysis of HT-29, the Hi-C data were down sampled by randomly extracting one tenth of the reads from each chromosome pair. The same analysis methods were used as described for HT-29.

A.7.2 Cell culture and cross-linking of chromatin

HT-29 human colorectal adenocarcinoma cells (Cat # HTB-38, ATCC, Manassas, VA) were propagated in growth medium, composed of McCoy's 5A medium (Life Technologies), 10% fetal bovine serum, and 1% penicillin/streptomycin solution. Cells grown in T-75 flasks were dissociated with 0.25% trypsin to single cell suspension for replating into culture plates for 2D or 3D growth. For 2D cultures, 4×106 cells were plated in each 150mm dish. For 3D spheroids, 2×105 cells were plated in each well of a 96-well PERFECTA3D ® Hanging Drop Plate (3D Biomatrix, Ann Arbor, MI). 2D and 3D cultures were incubated for 12 hours or 5 days in growth medium at 37° C with 5% CO2. Cells grown in 2D were cross-linked with 1% formaldehyde in serum free-medium for 15 minutes at room temperature, and then quenched with glycine to a final concentration of 0.128 M. Spheroids from one hanging drop plate were harvested, cross-linked and quenched as described above. Cross-linked cells were flash frozen in liquid nitrogen and then stored at $-80°$ C until the construction of Hi-C libraries. Fibroblast data is the merged 2D 48 hr samples from Chen H 2015 [17].

A.7.3 RNA isolation and RNA-seq

We used RNeasy product (Qiagen) to extract RNA from cells grown in 6-well plates or 96-well PERFECTA3D® Hanging Drop Plate. Total RNA was treated with RNAse-free DNaseI, then submitted to the University of Michigan sequencing Core lab for library construction and RNA-seq on the Illumina Hiseq-2000 platform. Single-end 50-base sequence reads were generated at a multiplex of 4 per sequencing lane.

In RNA-seq data process, the raw reads were checked with FastQC (version 0.10.1) to identify potential quality problems in the reads data. Tophat (version 2.0.11) and Bowtie (version 2.1.0.0) were used to align the reads to the reference transcriptome (HG19). Default parameter settings were used for alignment, with the exception of: "–b2-very-sensitive" as well as "–no-coverage-search" and "–no-novel-juncs" to limit the search to known transcripts. A second round of quality assessment was performed using FastQC on the aligned reads. The data is of overall excellent quality. Cufflinks/CuffDiff (version 2.2.1) were used for expression quantification and differential expression analysis, using UCSC hg19.fa and hg19.gtf as the reference genome and transcriptome. For the CuffDiff analysis, the following parameter settings were used: "–multi-read-correct" to adjust expression (FPKM) calculations for reads that map in more than one locus, and "–upper-quartile-norm" for normalization across samples. A locally developed R script using CummeRbund was used to format the Cufflinks output.

Gene level analysis was performed using FPKM values outputted by Cufflinks and $\log_2$ FC with pseudocounts, $\log_2$ FC = $\log_2$(HT-29 + 10^{-20}) − $\log_2$(fib+10^{-20}), for comparisons of samples and properties. Bin level gene expression vectors were calculated using raw counts outputted by Cufflinks and adding up the counts for all the genes in each bin then normalizing by million reads to convert them to FPM. Gene length normalization was not performed for bin level since each bin is the same

size.

A.7.4 Generation of Hi-C libraries for sequencing

For each Hi-C library generation [8], approximately 20×106 cells were cross-linked and resuspended in 1mL lysis buffer, consisting of 10mM Tris-HCl, 10mM NaCl, 0.2% Igpel (Cat $\#8896 - 50$mL, Sigma-Aldrich), and 10 mL protease inhibitor cocktail (Cat # $P8340 - 1$ml, Sigma-Aldrich), incubated on ice for 15 minutes. Cells were homogenized in a Dounce homogenizer on ice with pestle A, and the lysate was transferred to a 1.7mL tube. Cells were collected by spinning for 5 minutes at $2,000$xg, then washed twice in 500μL of ice cold $1\times$ NEB buffer 2. Cells were distributed between 41.7 ml centrifuge tubes (50μL per tube). Chromatin was digested with 400u of restriction enzyme HindIII (Cat # R0104M, New England BioLabs, Ipswich, MA) in $1\times$ NEB buffer 2 at 37°C overnight on a spin wheel.

After HindIII digestion, restriction site overhanging ends were filled and labeled with biotin with DNA polymerase I large (Klenow) fragment (Cat # M0210L, New England BioLabs) in a reaction containing dATP, dGTP, dTTP, and biotin-14-dCTP (Cat $\#19518 - 018$, Life Technologies) in each of the 4 HindIII digestion tube. DNA fragments labeled biotin-14-dCTP from each of the 4 tubes were ligated at 16° C for 4 hours in an 8.23 mL reaction containing $1\times$ ligation buffer, 1% Triton X-100 (Cat # $T8787 - 250$ML, Sigma-Aldrich), 1 mg/ml Bovine serum albumin (Cat # $BP9706 - 100$, Fisher Scientific), 10 mM ATP (Cat # $A9187 - 1$g, Sigma-Aldrich), and 50u T4 DNA ligase (Cat $\#15224 - 025$, Life Technologies).

Reverse cross-linking was performed in two steps. First, 50 μ of 10 mg/ml proteinase K (Cat $\#25530 - 015$, Life Technologies) was added to each ligation reaction tube and incubated at 65°C for 4 hours. Then, another 50 μl of proteinase K were added and continued incubating at 65°C overnight. Next, DNA was extracted with saturated phenol:chloroform ($1 : 1$) (Cat #1100631, Fisher Scientific), and desalted

by using AMICONA® Ultra Centrifugal Filter Unit (Cat # UFC503024, Millipore, Billerica, MA) with 1× TE buffer. The final volume of desalted DNA was adjusted to 100μL in 1× TE buffer.

Removal of Biotin from un-ligated ends was carried out in 8 individual reactions each of 50 μL containing 5 μg of Hi-C DNA, 1 mg/ml Bovine serum albumin,1× NEB buffer 2, 25 nM dATP, 25 nM dGTP, and 15u T4 DNA polymerase at 20°C for 4 hours. The Hi-C DNA was then pooled in a single tube, purified with single phenol extraction, and precipitated by ethanol. The DNA was re-dissolved in 105 μl of water, and transferred to a microTUBE AFA tube (Cat #520045, Covaris, Woburn, Massachusetts). DNA fragmentation was performed in a Sonicator (Covaris S2, Covaris). The DNA fragments in a size of $200 - 400$ bp were recovered with Agencourt AMPure® XP mixture (Cat # A63880, Beckman Coulter, Indianapolis IN) following the manufacturer's protocols.

DNA fragment ends were repaired in a 70 μL reaction containing 1× ligation buffer (Cat # B0202, New England BioLabs), 0.25 mM of dNTP mixture, 7.5u of T4 DNA polymerase (Cat # M0203L, New England BioLabs), 25u of T4 polynucleotide kinase (Cat # M0201S, New England BioLabs), 2.5u of DNA polymerase I large fragment at 20° C for 30 minutes. The reaction is purified with a MinElute column (Cat 28204, Qiagen, Valencia, CA). The DNA is eluted in 32 μL of elusion buffer for A-tailing, which was performed in a 50 μL reaction containing purified DNA (5 μg), 1× NEB buffer 2, 0.2 mM dATP, 15u Klenow fragment ($3' \rightarrow 5'$ exo-) (Cat # M0212L, New England BioLabs). The reaction was incubated at 37°C for 30 minutes, then at 65°C for 20 minutes to inactivate Klenow (exo-).

For Streptavidin pull-down, the biotinylated Hi-C ligation products are mixed with MyOne C1 streptavidin bead solution (Cat #65001, Life Technologies). Non-specifically binding DNA was removed by washing with 1× binding buffer (5 mM Tris-HCl (pH8.0), 0.5 mM EDTA, and 1 M NaCl), then with with 1× T4 Ligation buffer

(Cat #46300 − 018, Life Technologies). The DNA-bound beads were resuspended in 38.75μl of 1× ligation buffer for adapter ligation.

Illumina adapter ligation was performed in a 50 μL reaction by adding to the DNA-bound beads suspension of 1× T4 ligation buffer, 90 pM of Illumina paired end adapter, 3u of T4 DNA ligase (Cat #15224 − 025, Life Technologies). The reaction was incubated at room temperature for 2 hours. The beads were reclaimed, and the supernatant discarded. The beads were washed twice in Tween Wash Buffer (5 mM Tris-HCl pH8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20), and once in 1× binding buffer (5 mM Tris-HCl (pH8.0), 0.5 mM EDTA, and 1 M NaCl), and twice in 1× NEB buffer 2. After the last wash, the beads were resuspended in 20μl of 1X NEB buffer 2.

The Hi-C DNA sample was amplified by 16 PCR cycles (optimized in the log amplification phase) for Illumina HiSeq sequencing. Each PCR reaction in 25 l, 1.5μl of Bead-bound Hi-C DNA, 0.35μl of PE primer 1.0, 0.35 μl of PE primer 2.0, 0.2 μl of 25mM dNTP, 2.5 μl of 10X PfuUltra buffer, 19.6 μl of H_2O, and 0.5 μl of PfuUltra Fusion DNA polymerase. The PCR cycling parameters were 98° for 30 seconds, followed by 15 cycles at 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, and a final extension at, 72°C for 7 minutes. PCR products pooled from the supernatant of multiple reactions were subjected to AMpure XP beads purification to remove primer dimers. A standard quality control procedure was performed on the purified PCR products (Hi-C library). Each Hi-C library passed the QC procedure was then sequenced in a single lane of a flow cell on a HiSeq 2000 sequencer to generate paired-end sequence reads at 100 bases per end read.

A.7.5 Fluorescence in situ hybridization

Metaphase chromosomes were prepared by incubating the cells for 1 − 2 hours to 0.02 mg/ml Colcemid (Invitrogen; Grand Island, NY). The cells were then incubated

in a 0.075M hypotonic solution for twenty minutes before fixation in methanol/acetic acid (3:1). The cells were washed with fixative three times and dropped in a controlled humidity chamber. Interphase cells were prepared the same way except for the addition of Colcemid.

BACs were purchased from BAC/PAC Resources (Oakland, CA) for *MYC* (8q24) and to confirm the translocation of chromosome 6 and 14. The BAC clone contig for *MYC* consisted of 3 overlapping BAC clones anchored to the gene of interest, labeled with Spectrum Orange-dUTP (Abbott Molecular Inc, IL) using nick translation. To confirm the $t(6;14)$ translocation, BACs were ordered on 6q23.2 before 132890000 ($RP11-951D6$) and after 132825000 ($RP11-295F4$). On chromosome 14q13.2, BACs were ordered before and after 36508800 (CTD-2326N4 and $RP11-266A10$, respectively). The extracted DNA was labeled via nick-translation with Dy590 (Dyomics; Jena, Germany) for $RP11-951D6$; Dy505 (Dyomics; Jena, Germany) for $RP11-295F4$; Dy415 (Dyomics; Jena, Germany) for CTD-2326N4; and Dy547 (Dyomics; Jena, Germany) for $RP11-266A10$. Whole chromosome paint probes were generated in-house using PCR labeling techniques for the following chromosomes: 2 (Spectrum Orange dUTP), 8 (Dy505) and 15 (Dy505).

We followed our standard FISH protocol for hybridization and detection [127]. Approximately 20 metaphase nuclei were imaged using the Leica DM-RXA fluorescence microscope (Leica; Wetzlar, Germany) equipped with custom optical filters and a $40\times$ objective. All slides were counterstained with $4',6-$diamidino$-2-$phenylindole. Area measurements for chromosome territories were completed using a simple threshold.

A.8 CSC nucleome: introduction to centrality

Centrality measures how central or important a node is within a graph. This can be measured a number of different ways but one of the simplest is degree centrality. The degree centrality of a node is the number of edges surrounding the node (its

degree). Graphs can be written as an adjacency matrix ($\mathbf{A}$) in which each row and column represents a node and each entry is 1 if an edge connects the two nodes and 0 otherwise. An example graph is shown in Figure S38.

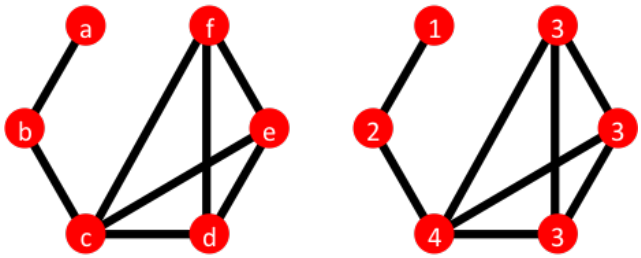

Figure S38: The same graph is shown with node labels and the degree of each node labeled on the left and right copies of the graph, respectively.

$$\mathbf{A} = \begin{array}{c} \begin{matrix} a & b & c & d & e & f \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array} \begin{matrix} a \\ b \\ c \\ d \\ e \\ f \end{matrix} \tag{S16}$$

The letters around adjacency matrix are not part of the matrix; they indicate which node that row or column represents. The degree centrality of a node is the number of edges surrounding the node. Graphically this is simply the number of black lines surrounding the node and algebraically this is the column sum of the adjacency matrix, both of which are shown below.

$$\deg(A) = \text{sum}(A) = [1 \quad 2 \quad 4 \; 3 \; 3 \; 3] \tag{S17}$$

For this particular graph, it shows that node c is the most central since it has a degree of 4, the highest in the graph, while node a is the least central since its degree

of 1 is the smallest. The example used here was an unweighted graph but the concept can be easily extended to weighted graphs (edges have weights w_i instead of just 0 or 1) with the degree still being the sum of the adjacency matrix.

Another measure of centrality is betweenness which measure the number of shortest paths between all pairs of nodes cross the node of interest. Additionally, closeness is the sum of the shortest paths from the selected node to all other nodes. The goal of all of these measures is to determine how central the node is to the graph.

A.9 CSC nucleome: types of centrality

8 different types of centrality were used for analysis of the genomic network as measured by Hi-C. They are:

- degree centrality - the nodal degree is defined as the sum of the edge weights, i.e. Hi-C, contacts surrounding each node,

$$\text{degree}(i) = \sum_{j=1}^{n} [\mathbf{A}]_{i,j} \tag{S18}$$

  , This indicates the spatial proximity between node i and all of the other nodes.

- eigenvector centrality - the eigenvector centrality is defined as the principal eigenvector of the adjacency matrix corresponding to its largest eigenvalues, i.e.

$$\text{eig}(i) = [v]_i = \frac{1}{\lambda_1(\mathbf{A})} \sum_{j=1}^{n} [\mathbf{A}]_{i,j} [v]_j \tag{S19}$$

  , where $\lambda_1(\mathbf{A})$ is the maximum eigenvalue of $\mathbf{A}$ in magnitude, and v is the associated eigenvector. Eigenvector centrality relies on the principle that a node is important it if it connected to many other important nodes, thus it accounts for the full network centrality more than degree centrality.

- local Fiedler vector centrality (LFVC) - LFVC evaluates a node's importance by evaluating the network centrality through the Fiedler vector,

$$\mathrm{LFVC}(i) = \sum_{j \in edges} ([y]_i - [y]_j)^2 \tag{S20}$$

, where y is the Fiedler vector. Since the Fiedler vector partitions the network into well separated clusters, i.e. A/B compartments, LFVC characterizes network partitioning and the nodal significance.

- closeness - closeness is defined by the shortest-path distance between a node and all other nodes,

$$\mathrm{close}(i) = \frac{1}{\sum_{j=1}^{n} \rho(i,j)} \tag{S21}$$

where $\rho(i,j)$ is the shortest-path distance between nodes i and j. The closeness reflect how far a node is from the center of the network.

- betweenness - betweenness is defined as the fraction of the number of shortest paths passing through a node relative to the total number of shortest paths,

$$\mathrm{betw}(i) = \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\sigma_{k,j}(i)}{\sigma_{k,j}} \tag{S22}$$

where $\sigma_{k,j}$ is the total number of shortest paths from node k to j and $\sigma_{k,j}(i)$ is the number of such shortest paths crossing through node i. Betweenness characterizes nodes that might make the network disconnect if removed because they are hub nodes.

- local clustering coefficient - local clustering coefficient measures how interconnected a nodes neighbors are,

$$\mathrm{LCC} = \frac{2|(j,k)|(j,k) \in \mathrm{edges}, \forall j,k \in N_i|}{|N_i|(|N_i|-1)}, \tag{S23}$$

where N_i is the direct neighbors of node i, and *numerator* denotes the number of edges in the neighborhood of node i. The local clustering coefficient characterized the local connectedness.

- h-hop walks. The h-hop walk is defined as the number of the edge weights associated with the paths departing from the node and traversing through h edges, which can be computed iteratively,

$$\mathrm{w}^{(h+1)} = \mathbf{A}\mathrm{d}^{(h)} + \mathbf{B}\mathrm{w}^{(h)}, \mathrm{d}^{(h)} = \mathbf{A}\mathrm{d}^{(h-1)}, \quad \mathrm{d}^{(0)} = 1, \mathrm{w}^{(1)} = \mathbf{A}1, h = 1, 2, ..., \tag{S24}$$

where $w^{(h)}$ is the vector of h-hop walk weights, and $\mathbf{B}$ is the binary Hi-C matrix with $[\mathbf{B}]_{\mathbf{i,j}} = 1$ if $[\mathbf{A}]_{\mathbf{i,j}} > 0$, and $[\mathbf{B}]_{\mathbf{i,j}} = 0$ otherwise. This accounts for indirect interactions among nodes. hops between 1 and 5 in length were used in this analysis.

- distance to a reference node - given a set of nodes of interest, the distance to each node can be measured and used as a structural feature.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Health, united states, 2015. Technical report, U.S.Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2015.

[2] Cancer statistics. Technical report, National Cancer Institute, 2017.

[3] Delphine Antoni, Hélène Burckel, Elodie Josset, and Georges Noel. Three-dimensional cell culture: a breakthrough in vivo. *International journal of molecular sciences*, 16(3):5517–5527, 2015.

[4] Sepideh Babaei, Waseem Akhtar, Johann De Jong, Marcel Reinders, and Jeroen De Ridder. 3d hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. *Nature communications*, 6, 2015.

[5] Tomas Babak, Brian DeVeale, Emily K Tsang, Yiqi Zhou, Xin Li, Kevin S Smith, Kim R Kukurba, Rui Zhang, Jin Billy Li, Derek van der Kooy, et al. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature genetics*, 47(5):544–549, 2015.

[6] Yael Baran, Meena Subramaniam, Anne Biton, Taru Tukiainen, Emily K Tsang, Manuel A Rivas, Matti Pirinen, Maria Gutierrez-Arcelus, Kevin S Smith, Kim R Kukurba, et al. The landscape of genomic imprinting across diverse adult human tissues. *Genome research*, 25(7):927–936, 2015.

[7] A Rasim Barutcu, Bryan R Lajoie, Rachel P McCord, Coralee E Tye, Deli Hong, Terri L Messier, Gillian Browne, Andre J van Wijnen, Jane B Lian, Janet L Stein, et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome biology*, 16(1):1, 2015.

[8] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. Hi–c: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276, 2012.

[9] Ralf Bender and Stefan Lange. Adjusting for multiple testing: when and how? *Journal of Clinical Epidemiology*, 54(4):343–349, 2001.

[10] Christelle Borel, Pedro G Ferreira, Federico Santoni, Olivier Delaneau, Alexandre Fort, Konstantin Y Popadin, Marco Garieri, Emilie Falconnet, Pascale

Ribaux, Michel Guipponi, et al. Biased allelic expression in human primary fibroblast single cells. *The American Journal of Human Genetics*, 96(1):70–80, 2015.

[11] Carolyn J Brown, Andrea Ballabio, James L Rupert, Ronald G Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F Willard. A gene from the region of the human x inactivation centre is expressed exclusively from the inactive x chromosome. *Nature*, 349(6304):38–44, 1991.

[12] Carolyn J Brown, Brian D Hendrich, Jim L Rupert, Ronald G Lafrenière, Yigong Xing, Jeanne Lawrence, and Huntington F Willard. The human xist gene: analysis of a 17 kb inactive x-specific rna that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–542, 1992.

[13] Clotilde Cadart, Ewa Zlotek-Zlotkiewicz, Maël Le Berre, Matthieu Piel, and Helen K Matthews. Exploring the function of cell shape and size during mitosis. *Developmental Cell*, 29(2):159–169, 2014.

[14] Jordi Camps, Quang Tri Nguyen, Hesed M Padilla-Nash, Turid Knutsen, Nicole E McNeil, Danny Wangsa, Amanda B Hummon, Marian Grade, Thomas Ried, and Michael J Difilippantonio. Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. *Genes, Chromosomes and Cancer*, 48(11):1002–1017, 2009.

[15] Emmanuelle Charafe-Jauffret, Christophe Ginestier, Flora Iovino, Julien Wicinski, Nathalie Cervera, Pascal Finetti, Min-Hee Hur, Mark E Diebel, Florence Monville, Julie Dutcher, et al. Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer research*, 69(4):1302–1313, 2009.

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

[17] Haiming Chen, Nicholas Comment, Jie Chen, Scott Ronquist, Alfred Hero, Thomas Ried, and Indika Rajapakse. Chromosome conformation of human fibroblasts grown in 3-dimensional spheroids. *Nucleus*, 6(1):55–65, 2015.

[18] Jie Chen, Alfred O Hero, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, page btw221, 2016.

[19] Derek Y Chiang, Gad Getz, David B Jaffe, Michael JT O'kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, 2009.

[20] Se Hoon Choi, Young Hye Kim, Matthias Hebisch, Christopher Sliwinski, Seungkyu Lee, Carla DAvanzo, Hechao Chen, Basavaraj Hooli, Caroline Asselin, Julien Muffat, et al. A three-dimensional human neural cell culture model of alzheimer/'s disease. *Nature*, 515(7526):274–278, 2014.

[21] Kin-Hoe Chow, Rachael E. Factor, and Katharine S. Ullman. The nuclear envelope enviornment and its cancer connections. *Nature Reviews Cancer*, 12:196–209, 2012.

[22] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[23] Hans Clevers. Modeling development and disease with organoids. *Cell*, 165(7):1586–1597, 2016.

[24] ENCODE Project Consortium et al. A user's guide to the encyclopedia of dna elements (encode). *PLoS biology*, 9(4):e1001046, 2011.

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

[26] Felipe Cucker and Steve Smale. On the mathematics of emergence. *Japanese Journal of Mathematics*, 2(1):197–227, 2007.

[27] Kris Noel Dahl, Alexandre J.S. Ribeiro, and Jan Lammerding. Nuclear shape, mechanics, and mechanotransduction. *Circulation Research*, 102:1307–1318, 2008.

[28] Robert L Davis, Harold Weintraub, and Andrew B Lassar. Expression of a single transfected cdna converts fibroblasts to myoblasts. *Cell*, 51(6):987–1000, 1987.

[29] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[30] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331, 2015.

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

[32] Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Systems*, 3(1):99–101, 2016.

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

[34] Per-Henrik D Edqvist, Linn Fagerberg, Björn M Hallström, Angelika Danielsson, Karolina Edlund, Mathias Uhlén, and Fredrik Pontén. Expression of human skin-specific genes defined by transcriptomics and antibody-based profiling. *Journal of Histochemistry & Cytochemistry*, 63(2):129–141, 2015.

[35] Jonathan I. Esptein, Stephen J. Berry, and Joseph C Eggleston. Nuclear roundness factor. a predictor of progression in untreated stage a2 prostate cancer. *Cancer*, 54:1666–1671, 1984.

[36] Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, 9(1):14, 2014.

[37] Claudia Fischbach, Ruth Chen, Takuya Matsumoto, Tobias Schmelzle, Joan S Brugge, Peter J Polverini, and David J Mooney. Engineering tumors with 3d scaffolds. *Nature methods*, 4(10):855, 2007.

[38] Jésica Flores-Martin, Viviana Rena, Sofía Angeletti, Graciela M Panzetta-Dutari, and Susana Genti-Raimondi. The lipid transfer protein stard7: structure, function, and regulation. *International journal of molecular sciences*, 14(3):6170–6186, 2013.

[39] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for hi-c data analysis. *Nature Methods*, 14(7):679–685, 2017.

[40] Geoff Fudenberg, Gad Getz, Matthew Meyerson, and Leonid A Mirny. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature biotechnology*, 29(12):1109–1113, 2011.

[41] B Michael Ghadimi and Thomas Ried. Chromosomal instability in cancer cells, 2015.

[42] Alexander Gimelbrant, John N Hutchinson, Benjamin R Thompson, and Andrew Chess. Widespread monoallelic expression on human autosomes. *Science*, 318(5853):1136–1140, 2007.

[43] L Goff, C Trapnell, and D Kelley. cummerbund: Analysis, exploration, manipulation, and visualization of cufflinks high-throughput sequencing data. *R package version*, 2(0), 2013.

[44] Robert D. Goldman, Dale K. Shumaker, Michael R. Erdos, Maria Eriksson, Anne E. Goldman, Leslie B. Gordon, Yosef Gruenbaum, Satya Khuon, Melissa Mendez, Renée Varga, and Francis S. Collins. Accumulation of mutant lamin a causes progressive changes in nuclear architecture in hutchinson–gilford progeria syndrome. *Proceedings of the National Academy of Sciences*, 101:8963–8968, 2004.

[45] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[46] Linda G Griffith and Melody A Swartz. Capturing complex 3d tissue physiology in vitro. *Nature reviews Molecular cell biology*, 7(3):211–224, 2006.

[47] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[48] Louise Harewood, Frédéric Schütz, Shelagh Boyle, Paul Perry, Mauro Delorenzi, Wendy A Bickmore, and Alexandre Reymond. The effect of translocation-induced nuclear reorganization on gene expression. *Genome research*, 20(5):554–564, 2010.

[49] Aaron N Hata, Matthew J Niederst, Hannah L Archibald, Maria Gomez-Caraballo, Faria M Siddiqui, Hillary E Mulvey, Yosef E Maruvka, Fei Ji, Hyoeun C Bhang, Viveksagar Krishnamurthy Radhakrishna, et al. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nature medicine*, 22(3):262, 2016.

[50] H Hayashi, T Arao, Y Togashi, H Kato, Y Fujita, MA De Velasco, H Kimura, K Matsumoto, K Tanaka, I Okamoto, et al. The oct4 pseudogene pou5f1b is amplified and promotes an aggressive phenotype in gastric cancer. *Oncogene*, 34(2):199–208, 2015.

[51] Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-Laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie, Zi Peng Fan, Alla A Sigova, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458, 2016.

[52] Pamela M Holland, Richard D Abramson, Robert Watson, and David H Gelfand. Detection of specific polymerase chain reaction product by utilizing the 5'—-3'exonuclease activity of thermus aquaticus dna polymerase. *Proceedings of the National Academy of Sciences*, 88(16):7276–7280, 1991.

[53] Douglas A Hosack, Glynn Dennis, Brad T Sherman, H Clifford Lane, and Richard A Lempicki. Identifying biological themes within lists of genes with ease. *Genome biology*, 4(10):R70, 2003.

[54] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

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

[56] Yoshinori Imamura, Toru Mukohara, Yohei Shimono, Yohei Funakoshi, Naoko Chayahara, Masanori Toyoda, Naomi Kiyota, Shintaro Takao, Seishi Kono, Tetsuya Nakatsura, et al. Comparison of 2d-and 3d-culture models as drug-testing platforms in breast cancer. *Oncology reports*, 33(4):1837–1843, 2015.

[57] Salk Institute. Homer hi-c background models, 04 2015.

[58] Roland Jäger, Gabriele Migliorini, Marc Henrion, Radhika Kandaswamy, Helen E Speedy, Andreas Heindl, Nicola Whiffin, Maria J Carnicer, Laura Broome, Nicola Dryden, et al. Capture hi-c identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications*, 6, 2015.

[59] Predrag Javetić, Lisa J Edens, Lidija D Vuković, and Daniel L Levy. Sizing and shaping the nucleus: mechanisms and significance. *Current Opinion Cell Biology*, 28:16–27, 2014.

[60] Junghyun Jo, Yixin Xiao, Alfred Xuyang Sun, Engin Cukuroglu, Hoang-Dai Tran, Jonathan Göke, Zi Ying Tan, Tzuen Yih Saw, Cheng-Peow Tan, Hidayat Lokman, et al. Midbrain-like organoids from human pluripotent stem cells contain functional dopaminergic and neuromelanin-producing neurons. *Cell Stem Cell*, 19(2):248–257, 2016.

[61] Peter A Jones and Stephen B Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, 2007.

[62] Eun Yong Kang, Lisa J Martin, Serghei Mangul, Warin Isvilanonda, Jennifer Zou, Eyal Ben-David, Buhm Han, Aldons J Lusis, Sagiv Shifman, and Eleazar Eskin. Discovering single nucleotide polymorphisms regulating human gene expression using allele specific expression from rna-seq data. *Genetics*, 204(3):1057–1064, 2016.

[63] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[64] Turid Knutsen, Vasuki Gobu, Rodger Knaus, Hesed Padilla-Nash, Meena Augustus, Robert L Strausberg, Ilan R Kirsch, Karl Sirotkin, and Thomas Ried. The interactive online sky/m-fish & cgh database and the entrez cancer chromosomes search database: Linkage of chromosomal aberrations with the genome sequence. *Genes, Chromosomes and Cancer*, 44(1):52–64, 2005.

[65] Kai Kretzschmar and Hans Clevers. Organoids: modeling development and the stem cell niche in a dish. *Developmental Cell*, 38(6):590–600, 2016.

[66] Oscar M Lancaster, Maël Le Berre, Andrea Dimitracopoulos, Daria Bonazzi, Ewa Zlotek-Zlotkiewicz, Remigio Picone, Thomas Duke, Matthieu Piel, and Buzz Baum. Mitotic rounding alters cell geometry to ensure efficient bipolar spindle formation. *Developmental cell*, 25(3):270–283, 2013.

[67] Thomas J Langan and Richard C Chou. Synchronization of mammalian cell cultures by serum deprivation. *Cell Cycle Synchronization*, 761:75–83, 2011.

[68] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[69] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[70] Charalampos Lazaris, Stephen Kelly, Panagiotis Ntziachristos, Iannis Aifantis, and Aristotelis Tsirigos. Hic-bench: comprehensive and reproducible hi-c data analysis designed for parameter exploration and benchmarking. *BMC genomics*, 18(1):22, 2017.

[71] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

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

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

[74] Sijia Liu, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Lindsey A Muir, Pin-Yu Chen, Gerald Higgins, Pierre Baldi, et al. Genome architecture leads a bifurcation in cell identity. *bioRxiv*, page 151555, 2017.

[75] Kenneth J Livak and Thomas D Schmittgen. Analysis of relative gene expression data using real-time quantitative pcr and the 2- $\delta\delta$ct method. *methods*, 25(4):402–408, 2001.

[76] Nick R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462, 1976.

[77] Paul Martin. Wound healing–aiming for perfect skin regeneration. *Science*, 276(5309):75–81, 1997.

[78] Adolf Mathias, Florian Grond, Ramon Guardans, Detlef Seese, Miguel Canela, Hans H Diebner, and Giovanni Baiocchi. Algorithms for spectral analysis of irregularly sampled time series. *Journal of Statistical Software*, 11(2):1–30, 2004.

[79] Vea Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E Kel, Olga V Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, 2003.

[80] Karen J Meaburn, Tom Misteli, and Evi Soutoglou. Spatial genome organization in the formation of chromosomal translocations. In *Seminars in cancer biology*, volume 17, pages 80–90. Elsevier, 2007.

[81] Patrick Mehlen and Alain Puisieux. Metastasis: a question of life or death. *Nature Reviews Cancer*, 6(6):449–458, 2006.

[82] Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10):685–696, 2010.

[83] Joana Monteiro and Riccardo Fodde. Cancer stemness and metastasis: therapeutic consequences and perspectives. *European journal of cancer*, 46(7):1198–1203, 2010.

[84] RA Mora-Rodriguez and JA Molina-Mora. Characterization of heterogeneous response to chemotherapy by perturbation-based modeling of fluorescent sphingolipid metabolism in cancer cell subpopulations. In *Central American and Panama Convention (CONCAPAN XXXVI), 2016 IEEE 36th*, pages 1–7. IEEE, 2016.

[85] Anwesha Nag, Virginia Savova, Ho-Lim Fung, Alexander Miron, Guo-Cheng Yuan, Kun Zhang, and Alexander A Gimelbrant. Chromatin signature of widespread monoallelic expression. *Elife*, 2:e01256, 2013.

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

[87] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[88] Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[89] Tomi Pastinen and Thomas J Hudson. Cis-acting regulatory variation in the human genome. *Science*, 306(5696):647–650, 2004.

[90] Michelle T Paulsen, Artur Veloso, Jayendra Prasad, Karan Bedi, Emily A Ljungman, Brian Magnuson, Thomas E Wilson, and Mats Ljungman. Use of bru-seq and bruchase-seq for genome-wide assessment of the synthesis and stability of rna. *Methods*, 67(1):45–54, 2014.

[91] Dénes Petz. Entropy, von neumann and the von neumann entropy. In *John von Neumann and the foundations of quantum physics*, pages 83–96. Springer, 2001.

[92] Daniel Pinkel and Donna G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37:S11–S17, 2005.

[93] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes in C*, volume 2. Citeseer, 1996.

[94] ME Prince, R Sivanandan, A Kaczorowski, GT Wolf, MJ Kaplan, P Dalerba, IL Weissman, MF Clarke, and LE Ailles. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. *Proceedings of the National Academy of Sciences*, 104(3):973–978, 2007.

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

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

[97] Padmini Rangamani, Azi Lipshtat, Evren U. Azeloglu, Rhodora Cristina Calizo, Mufeng Hu, Saba Ghassemi, James Hone, Suzanne Scarlata, Susana R. Neves, and Ravi Iyengar. Decoding information in cell shape. *Cell*, 154:1356–1369, 2013.

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

[99] S. V. Razin, A. A. Gavrilov, A. Pichugin, M. Lipinski, O. V. Iarovaia, and Yegor S. Vassetzky. Transcription factories in the context of the nuclear and genome organization. *Nucleic Acids Research*, 39:9085–9092, 2011.

[100] Wolf Reik and Jörn Walter. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32, 2001.

[101] Thomas Ried, Yue Hu, Michael J Difilippantonio, B Michael Ghadimi, Marian Grade, and Jordi Camps. The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(7):784–793, 2012.

[102] Markus Riester, Hua-Jun Wu, Ahmet Zehir, Mithat Gönen, Andre L Moreira, Robert J Downey, and Franziska Michor. Distance in cancer gene expression from stem cells predicts patient survival. *PloS one*, 12(3):e0173589, 2017.

[103] Kimberly Robasky and Martha L Bulyk. Uniprobe, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–dna interactions. *Nucleic acids research*, 39(suppl_1):D124–D128, 2010.

[104] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. Edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[105] Scott Ronquist, Geoff Patterson, Markus Brown, Haiming Chen, Anthony Bloch, Lindsey Muir, Roger Brockett, and Indika Rajapakse. An algorithm for cellular reprogramming. *arXiv preprint arXiv:1703.03441*, 2017.

[106] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl 1):D91–D94, 2004.

[107] Federico A Santoni, Georgios Stamoulis, Marco Garieri, Emilie Falconnet, Pascale Ribaux, Christelle Borel, and Stylianos E Antonarakis. Detection of imprinted genes by single-cell allele-specific gene expression. *The American Journal of Human Genetics*, 100(3):444–453, 2017.

[108] Yoshiki Sasai. Next-generation regenerative medicine: organogenesis from stem cells in 3d culture. *Cell stem cell*, 12(5):520–530, 2013.

[109] Virginia Savova, Jon Patsenker, Sébastien Vigneau, and Alexander A Gimelbrant. dbmae: the database of autosomal monoallelic expression. *Nucleic acids research*, 44(D1):D753–D756, 2016.

[110] Ueli Schibler and Felix Naef. Cellular oscillators: rhythmic gene expression and metabolism. *Current opinion in cell biology*, 17(2):223–229, 2005.

[111] Marc W Schmid, Stefan Grob, and Ueli Grossniklaus. Hicdat: a fast and easy-to-use hi-c data analysis tool. *BMC bioinformatics*, 16(1):277, 2015.

[112] EDMS Schrock, S Du Manoir, T Veldman, B Schoell, et al. Multicolor spectral karotyping of human chromosomes. *Science*, 273(5274):494, 1996.

[113] Laura Seaman, Haiming Chen, Markus Brown, Darawalee Wangsa, Geoff Patterson, Jordi Camps, Gilbert S Omenn, Thomas Ried, and Indika Rajapakse. Nucleome analysis reveals structure-function relationships for colon cancer. *Molecular Cancer Research*, pages molcanres–0374, 2017.

[114] Alvaro Sebastian and Bruno Contreras-Moreira. footprintdb: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, page btt663, 2013.

[115] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):1–11, 2015.

[116] Michel Siffre. Six months alone in a cave. *National Geographic*, 147:426–435, 1975.

[117] A Sivaraman, JK Leach, S Townsend, T Iida, BJ Hogan, D Beer Stolz, R Fry, LD Samson, SR Tannenbaum, and LG Griffith. A microscale in vitro physiological model of the liver: predictive screens for drug metabolism and enzyme induction. *Current drug metabolism*, 6(6):569–591, 2005.

[118] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife OShaughnessy-Kirwan, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59–64, 2017.

[119] Phillippa C Taberlay, Joanna Achinger-Kawecka, Aaron TL Lun, Fabian A Buske, Kenneth Sabir, Cathryn M Gould, Elena Zotenko, Saul A Bert, Katherine A Giles, Denis C Bauer, et al. Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations. *Genome Research*, pages gr–201517, 2016.

[120] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

[121] Igor Tamm, Toyoko Kikuchi, Eugenia Wang, and Lawrence M Pfeffer. Growth rate of control and β-interferon-treated human fibroblast populations over the course of their in vitro life span. *Cancer research*, 44(6):2291–2296, 1984.

[122] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[123] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012.

[124] Barbara J Trask. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Reviews Genetics*, 3(10):769–778, 2002.

[125] Yi-Chung Tung, Amy Y Hsiao, Steven G Allen, Yu-suke Torisawa, Mitchell Ho, and Shuichi Takayama. High-throughput 3d spheroid culture and drug testing using a 384 hanging drop array. *Analyst*, 136(3):473–478, 2011.

[126] Jeff Vierstra, Andreas Reik, Kai-Hsin Chang, Sandra Stehling-Sun, Yuanyue Zhou, Sarah J Hinkley, David E Paschon, Lei Zhang, Nikoletta Psatha, Yuri R Bendana, et al. Functional footprinting of regulatory dna. *Nature methods*, 2015.

[127] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[128] Britta Weigelt, Cyrus M Ghajar, and Mina J Bissell. The need for complex 3d culture models to unravel novel pathways and identify accurate biomarkers in breast cancer. *Advanced drug delivery reviews*, 69:42–51, 2014.

[129] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, 2014.

[130] Yanhua Wen, Yanjun Wei, Shumei Zhang, Song Li, Hongbo Liu, Fang Wang, Yue Zhao, Dongwei Zhang, and Yan Zhang. Cell subpopulation deconvolution reveals breast cancer heterogeneity based on dna methylation signature. *Briefings in bioinformatics*, 18(3):426–440, 2016.

[131] Marius Wernig, Alexander Meissner, Ruth Foreman, Tobias Brambrink, Manching Ku, Konrad Hochedlinger, Bradley E Bernstein, and Rudolf Jaenisch. In vitro reprogramming of fibroblasts into a pluripotent es-cell-like state. *Nature*, 448(7151):318–324, 2007.

[132] Max S Wicha, Suling Liu, and Gabriela Dontu. Cancer stem cells: an old ideaa paradigm shift. *Cancer research*, 66(4):1883–1890, 2006.

[133] Huntington F Willard. X chromosome inactivation, xist, and pursuit of the x-inactivation center. *Cell*, 86(1):5–7, 1996.

[134] Hua-Jun Wu and Franziska Michor. A computational strategy to adjust for copy number in tumor hi-c data. *Bioinformatics*, page btw540, 2016.

[135] Kenneth M Yamada and Edna Cukierman. Modeling tissue morphogenesis and cancer in 3d. *Cell*, 130(4):601–610, 2007.

[136] Zuoren Yu, Timothy G Pestell, Michael P Lisanti, and Richard G Pestell. Cancer stem cells. *The international journal of biochemistry & cell biology*, 44(12):2144–2151, 2012.