Disentangling the 4D Nucleome

by

Stephen Lindsly

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2021

Doctoral Committee:

Associate Professor Indika Rajapakse, Chair
Professor Gilbert S. Omenn
Professor Anthony Bloch
Professor Daniel Burns
Research Assistant Professor Lindsey Muir

Stephen Lindsly

Lindsly@umich.edu

ORCID iD: 0000-0001-8787-1746

ACKNOWLEDGMENTS

I would like to thank my advisor, Indika Rajapakse, for his support over the past eight years. From our very first conversation, just a few weeks after starting as an undergraduate at the University of Michigan, your passion for science has been a constant inspiration. I would also like to thank my dissertation committee members, Gilbert S. Omenn, Anthony Bloch, Daniel Burns, and Lindsey Muir for their scientific guidance and expertise during my undergraduate and graduate work.

Thank you to all of the Rajapakse lab members, both past and present, for your advice and helpful discussions: Scott Ronquist, Walter Meixner, Haiming Chen, Sijia Liu, Wenlong Jia, Can Chen, Sivakumar Jeyarajan, Geoff Patterson, Laura Seaman, Emily Crossette, Gabrielle Dotson, Christopher York, Charles Ryan, Sam Dilworth, Drew McKearney, Anthony Cicalo, and Cooper Stansbury. I would also like to thank my collaborators and research associates: Maya Gupta, Alnawaz Rehemtulla, Alfred Hero, Ren Wang, Tianqi Chen, Xingzhao Wen, Shuai Cheng Li, Michael Perlman, and Max Wicha. Additionally, I would like to thank the Bioinformatics graduate program coordinators and staff for their support and counsel.

Finally, I would like to thank my family for their love and unwavering support. Thank you to my brothers, Chris Humphreys and Chad Lindsly, for helping maintain my sanity and encouraging me to be the best that I can be. Thank you to my wonderful parents, Jean and Joe Lindsly, for the incredible amount of support you have given me. I cannot express in words how much you mean to me, and I dedicate this work to you.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

1D one-dimensional

2D two-dimensional

3C chromatin conformation capture

3D three-dimensional

4DN 4D Nucleome

5C chromosome conformation capture carbon copy

ABE allele-biased expression

AD allelic divergence

AID activation-induced cytidine deaminase

AM affinity maturation

ATAC-seq assay for transposase-accessible chromatin using sequencing

BLOSUM blocks substitution matrix

Bru-seq bromouridine sequencing

CBE cell cycle-biased expression

ChIP-seq chromatin immunoprecipitation sequencing

CTCF CCCTC-binding factor

DNase-seq DNase I hypersensitive sites sequencing

FISH fluorescent in situ hybridization

FPKM fragments per kilobase of transcript per million mapped reads

GC germinal center

Hi-C genome wide chromosome conformation capture

InDel insertions or deletions

kb kilobase

LE laplacian eigenmaps

LP Larntz and Perlman

MAE monoallelic expression

Mb megabase

MR master regulator

PCA principal component analysis

PE paired-end

Pol II RNA polymerase II

RNA-seq RNA sequencing

SNV single nucleotide variation

SHM somatic hypermutation

RPKM reads per kilobase of transcript per million mapped reads

TAD topologically associating domain

TPM transcripts per million

TF transcription factor

t-SNE t-distributed stochastic neighbor embedding

UMAP uniform manifold approximation and projection

XCI X-chromosome inactivation

ABSTRACT

The dynamical relationship between 3D genome structure, genome function, and cellular phenotype is referred to as the 4D Nucleome (4DN). 4DN analysis remains difficult, since multiple data modalities must be integrated and comprehensively studied in order to obtain new insights. In my dissertation work, I present a computational toolbox which offers both novel and established methods to integrate and analyze time series genome structure and function data. I also provide an extension of the 4DN that captures the contributions of the maternal and paternal genomes. I uncover differences between the two genomes' structural and functional features across the cell cycle, and reveal an allele-specific relationship between local genome structures and gene expression. In addition, I present a computational framework for analyzing multi-way genomic interactions which allow us to identify transcription clusters in the human genome. Finally, I introduce a computational method to characterize the differences between memory and plasma B cells in the adaptive immune system, which guide us to develop an immune system inspired learning system.

CHAPTER I

Introduction

1.1 Research overview

The 4D Nucleome (4DN) is defined by the dynamical relationship between 3D genome structure, genome function, and cellular phenotype. In recent years, 4DN analysis has gained attention for its ability to uncover cellular regulation principles and advance the field of cellular reprogramming [40, 111]. 4DN analysis remains difficult, since multiple data modalities must be integrated and studied comprehensively in order to obtain new insights. These data are often large and represented in disparate formats. In addition, methods for 4DN analysis remain underdeveloped since most established analysis techniques focus on a single data type. Successful analysis of the 4DN requires an in depth knowledge of multiple fields, including biology, mathematics, and computer science.

In this dissertation, I present my contribution towards the advancement of knowledge on the 4DN. This includes the development of 4DN analysis tools and methods, the introduction of novel data sets, and insights into cell dynamics within the immune system. The remainder of the research overview will outline the methods and concepts presented in later chapters.

Chapter II, "4DNvestigator: Time Series Genomic Data Analysis Toolbox," describes the 4DNvestigator, a toolbox designed for the analysis of 4DN data. The

4DNvestigator is a MATLAB toolbox that processes time series RNA-seq and Hi-C data, and applies both novel and established analysis methods to these data. The 4DNvestigator can automatically apply standard methods like differential expression analysis and topologically associated domain (TAD) identification to time series data. In addition, the 4DNvestigator includes recently developed analysis techniques, including a statistical test for comparing Hi-C matrices and a formalized method to integrate gene expression with genome structure features over time and compare their dynamics. We provide a publicly available code repository for others to use for their own research, with pre-written examples to directly recapitulate the analysis presented in this work. This allows for both experienced researchers and novices to explore and analyze the 4DN in a mathematically rigorous manner.

Chapter III, "Functional Organization of the Maternal and Paternal Human 4D Nucleome," examines the relationship between genome structure and genome function in the maternal and paternal genomes across the cell cycle. Steady state gene expression (RNA-seq), nascent gene expression (Bru-seq), and genome structure (Hi-C) were separated into their maternal and paternal contributions to determine how the two genomes work together to give rise to cellular phenotype. We investigated differences between maternal and paternal genome structure and function from cell cycle sorted B-lymphocytes, and discovered significant relationships between changes in local genome structure and differentially expressed genes. In addition, we found that maternal and paternal alleles of many genes have different dynamics through the cell cycle. In contrast, we also discovered that genes which regulate the cell cycle and other critical biological pathways have extraordinarily similar gene expression patterns between their alleles. We introduce an extension of the 4DN, which accounts for the maternal and paternal genomes independently, allowing for more in-depth analysis of 4DN data.

Chapter IV, "Deciphering Multi-way Interactions in the Human Genome," describes a recently developed genome conformation capture technique, Pore-C, and methods for its analysis. We introduce a novel dataset for Pore-C in human fibroblasts and use publicly available Pore-C data from B-lymphocytes for our investigation. We outline how the long range multi-way contacts between genomic loci from Pore-C can be represented computationally through incidence matrices and hypergraphs. We also develop methods for analysis of Pore-C data. This includes measurements of entropy, which can be used to describe the amount of order within a genome and to compare different cell types or cell states. In addition, we created a pipeline that integrates Pore-C with other data modalities to identify biologically important genomic features called transcription factories (transcription clusters). This ongoing work represents a major advancement for the 4DN, since the incorporation of multi-way contacts will be crucial to elucidate more complex and elusive genomic structural features going forward.

Chapter V, "Understanding Memory B Cell Selection," represents a slight change in focus from the genome within each cell to the dynamics of cells themselves. Specifically, we examine the dynamics of B cells during the adaptive immune response. We hypothesize what mathematical objective functions the immune system is optimizing for during a process called affinity maturation, in which naive B cells are mutated in order to gain affinity to a foreign antigen. We consider the differences between optimizing for plasma B cells, which have extremely high affinity and eradicate the current antigen, and memory B cells, which have a lower affinity but are stored within the body for repeated attacks from the same (or similar) antigen in the future. We provide a detailed algorithm that outlines the steps that the immune system performs during affinity maturation, and offer a simulated immune response to quantitatively assess our hypotheses.

Together, the methods and experiments within this dissertation represent an advancement in 4DN data analysis, and insights into cell dynamics within the immune system. This work will guide future research on genome structure and function, with implications for the studies of genetic diseases, immunology, and personalized medicine.

1.2 Hi-C and RNA-seq

The human genome contains over 20,000 genes which work together to give rise to cellular phenotype. Within individual cells, not all genes are active. Depending on the cell type and cell state, certain genes express while others are transcriptionally repressed. Over the past two decades, it has been found that different cell types and cell states can be characterized by their gene expression profiles. For example, a skin cell has a different set of genes that are turned "on" compared to a neuron, and a cancerous cell may have genes which are mistakenly over- or under-expressed. Because gene expression can be used to differentiate cell types and cell states, it can also serve as a valuable tool for analyzing the differences and identifying the causes for these types and states.

The most commonly used technique for extracting gene expression is RNA-sequencing (RNA-seq) [137]. RNA-seq is a high-throughput DNA sequencing method, which quantifies RNA transcripts after conversion of RNA to cDNA (reverse transcription) [94, 137]. RNA-seq offers many benefits over previous technologies such as genomic tiling microarrays and cDNA Sanger sequencing. Microarrays offer high resolution, but rely on previous knowledge of the genome sequence and have high background noise. Sanger sequencing of cDNA provides accurate reads but is expensive and has low throughput. On the other hand, RNA-seq is able to quantify gene expression that does not correspond to known genomic sequences, enabling analysis on unmapped regions of the human genome or on other organisms whose genomes have yet to be sequenced. Additionally, RNA-seq is a relatively inexpensive, high throughput, and accurate technique, and is able to detect a wide range of expression levels (i.e. extremely low or high expression) [137]. Because of these advantages, RNA-seq has been critical in studying and discovering the boundaries of genes, splicing isoforms of genes, and novel transcribed regions [137]. RNA-seq has become the standard for biological experiments since its conception. It has proved to be an invaluable tool for understanding gene expression in numerous disease states, cell types, and biological time (i.e. the cell cycle).

Computationally, RNA-seq data are represented as a vector of expression levels for all genes. These vectors of gene expression can then be compared between settings to determine changes in gene expression. Comparisons of expression can be as simple as subtracting one sample's expression vector from the other and determining the genes with the largest differences. Differential expression analysis is another common way of comparing samples, where differences are determined by finding a ratio, or fold-change, of the genes between settings [80]. Comparisons of gene expression profiles in diseased versus healthy cells has been a major focus within the past two decades. These analyses have to led to countless discoveries of genes related to disease states, and the development of therapeutic technologies which have the ability to directly target malfunctioning genes.

The 3D organization of the genome enables approximately three billion nucleotide base pairs to be contained within a cell's nucleus. Genome organization and its importance for proper cellular function has been a long standing question, but this field has made dramatic advances in recent years through the development of chromatin conformation capture (3C) technology and its related extensions, namely genome-wide chromatin conformation capture (Hi-C) [42]. To extract Hi-C data, chromatin is fixed in place and neighboring DNA sequences are crosslinked. The DNA is then digested, or cut with restriction enzymes, which isolates potentially distant (in terms of 1D sequence) regions of the genome that are in close 3D proximity. These small sections of DNA are then ligated, sequenced, and aligned with a reference genome.

This data provides information on DNA-DNA ‘contacts’ which is stored within a matrix for analysis [42].

Hi-C data has provided unprecedented levels of information on the structure of the genome, both in terms of data resolution and throughput [42]. Analysis of data derived from Hi-C (and its predecessors) has confirmed the existence of euchromatin and heterochromatin, or loosely packed and tightly packed DNA, as well as chromosome territories which had previously been observed through microscopy (Figure 1.1). In addition, Hi-C has enabled the discovery of finer details of genome organization such as topologically associated domains (TADs) and chromatin loops between enhancers, insulators, and genes genome-wide [42, 74, 107].

Chromatin contacts are represented as matrices computationally, where rows and columns correspond to genomic loci at a given resolution. Hi-C data are commonly referred to as contact frequency matrices, as bulk Hi-C data has the largest values in the indices that represent the most frequent contacts between genomic loci over a population of cells. Hi-C matrices are characterized by a block-like structure. These represent areas of the genome which have a preference to interact with one another. The most significant block structures are formed by euchromatin and heterochromatin. Euchromatic regions tend to interact with other euchromatic regions but not heterochromatic regions (and vice-versa). The more subtle block structures within euchromatin and heterochromatin correspond to TADs, in which chromatin is highly self-interacting but is insulated from neighboring chromatin. Similar to RNA-seq data, Hi-C can be used to differentiate between cell types and identify disease states. For example, Hi-C has been used to identify chromosomal aberrations, such as translocations or aneuploidy, which are hallmarks of cancers [114, 115].

1.3 The 4D Nucleome

Gene expression and the 3D organization of the genome are intrinsically linked to one another. When unfolded, the human genome is approximately two meters long, yet it is able to fit within a cell nucleus with a diameter of only six micrometers [19]. This is only possible due to the incredible organization of the genome, which consists of a highly complex folding of DNA at multiple scales. DNA is wrapped around histone octamers and is then further organized into loops, topologically domains, and chromatin compartments [92]. This organization is non-random, and the functional output of genes is connected to the specific configuration of the genome. Genome organization and gene expression are not static, but instead are constantly changing over time due to cell replication, transcriptional bursting, and other factors like the circadian rhythm (Figure 1.1) [23, 33]. The dynamical relationship of genome structure, gene expression, and cell phenotype is known as the 4D Nucleome.

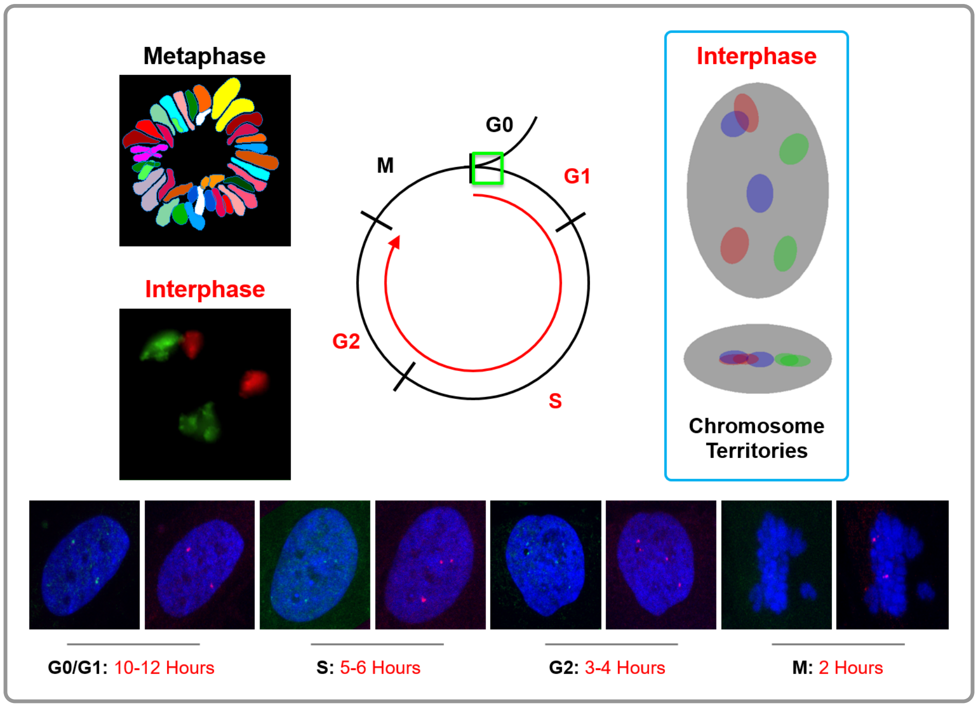


Figure 1.1: Imaging and simplified schematic of the cell cycle. Chromosome territories occupy distinct regions within the interphase nucleus. The maternal and paternal copies of each chromosome are often far apart.

The integration of gene expression and genome architecture, RNA-seq and Hi-C data respectively, offers novel insights into their relationship and how they give rise to cellular phenotype. This integration remains difficult, as the two data modalities are inherently different. First, RNA-seq offers a measurement of gene expression which only covers RNA coding regions of the genome. That is, the areas between transcribed regions do not provide information. Hi-C data can be derived from all non-repetitive areas of the genome but Hi-C is not able to provide a "gene-scale" contact matrix without being prohibitively expensive. This discrepancy can be mostly resolved through binning, or grouping of genomic loci into lower resolution subdivisions. Second, RNA-seq gives a vector of values (one value per gene) while Hi-C provides a matrix of contacts between genomic loci. The conflict of data representation can be addressed through the creation of a correlation matrix of gene expression, or the reduction of a Hi-C matrix to a vector of values. Multiple methods have been proposed for the latter, such as principle component analysis (PCA), but these methods innately lose important spatial information which is critical for more nuanced analyses.

Other promising methods for the integration of gene expression and genome organization come from network theory. A network is a simplified representation of a system, which captures patterns of interactions between its components. One can consider the genome as a network, where genomic loci (e.g. genes) are the nodes and contacts between loci are the edges. From this network perspective, both gene expression and genome architecture can be used to study the genome between cell states or through time. One can use concepts that are well known in network theory such as network centrality, clustering coefficients, and entropy to determine changes in the network architecture and how these changes are related to phenotypical differences [79]. Genomic networks have been used to study genes essential to the circadian rhythm and cell cycle, and elucidate structural and functional changes during cellular reprogramming [23, 79].

1.4 Pore-C/Multi-Way Interactions in the Genome

A recent advance in genome sequencing by Oxford Nanopore Technologies led to the development of Pore-C. Pore-C offers some distinct advantages over Hi-C. Pore-C data are characterized by their long read lengths and ability to capture long range interactions between multiple loci at once (multi-way interactions). Since this technology was only made commercially available recently, few tools or methods have been created to process and analyze these data.

Pore-C can capture multi-way interactions between genomic loci, since the reads contain multiple sequences concatenated together. We process these data and construct hypergraphs from the Pore-C reads. Normally, graphs contain edges between pairs of nodes (which can describe two-way interactions from other forms of data, like Hi-C), while hypergraphs can contain any number of nodes within a single hyperedge. This allows us to represent multi-way contacts (3+ loci) from Pore-C data in a more succinct and efficient manner. To work with Pore-C data computationally, we use incidence matrices. In incidence matrices, rows represent nodes (loci) and columns represent hyperedges (multi-way contacts). Pore-C data also include the pair-wise contacts offered by Hi-C, both by directly capturing pairs of interacting loci and by decomposing multi-way contacts into all pair-wise combinations of loci.

One application for Pore-C data is for the identification of biologically relevant interactions between genomic loci. We focus on the identification of features called transcription clusters (also called transcription factories), where multiple loci come together for more efficient gene transcription. They are characterized by high concentrations of proteins called transcription factors and RNA Polymerase II, which facilitate gene transcription. We integrate the multi-way contact data from Pore-C with other data modalities, such as chromatin accessibility (ATAC-seq), RNA Polymerase II binding (ChIP-seq), transcription factor binding motifs, and gene expression (RNA-seq) to identify these transcription clusters. The identification of transcription clusters can help explain the relationship between genome structure and gene function, and offer insights into gene expression regulation which can be leveraged for therapeutic applications in the future.

1.5 Adaptive Immune System

The immune system is critical in maintaining the health of humans. It identifies and defends against the great variety of pathogens one encounters in daily life, even defending against threats not previously encountered and perhaps never to be seen again. Considering the range and severity of the problems the immune system confronts, its properties are truly amazing, well beyond current technology's capabilities.

When confronting a threat, the immune system usually does not attempt to create an entirely new solution. Throughout a person's life, the immune system deals with many antigens and uses these experiences to help fight the current threat. The first line of defense is called the innate immune system. It includes innate immune cells such as macrophages and dendritic cells, and is able to detect and destroy invaders that show clear non-self patterns. The innate immune system responds very quickly, but generically, to threats.

The adaptive immune system is a second line of defense that more specifically detects and destroys foreign pathogens, and generates antibodies that are critical for resolving infections. Generating highly specific antibodies is a multistage process that starts with certain signals from the innate immune system. When those signals (antigens) reach regions of lymphoid tissue, such as lymph nodes, a specialized structure forms called the germinal center. Within the germinal center, B cells undergo an optimization process called affinity maturation, where interactions between B cells and T cell promote survival of B cells whose receptors bind an antigen very well. These interactions eventually lead to the production of plasma B cells which then secrete a highly specific antibody against the antigen. Further, this process produces memory B cells that can quickly be activated if the same antigen, or something similar, reappears in the future. Thus key elements of a successful solution are stored for future use.

CHAPTER II

4DNvestigator: Time Series Genomic Data Analysis Toolbox

This chapter is based on a paper by Stephen Lindsly, Can Chen, Sijia Liu, Scott Ronquist, Samuel Dilworth, Michael Perlman, and Indika Rajapakse [76].

2.1 Abstract

Data on genome organization and output over time, or the 4D Nucleome (4DN), require synthesis for meaningful interpretation. Development of tools for the efficient integration of these data is needed, especially for the time dimension. We present the "4DNvestigator", a user-friendly network based toolbox for the analysis of time series genome-wide genome structure (Hi-C) and gene expression (RNA-seq) data. Additionally, we provide methods to quantify network entropy, tensor entropy, and statistically significant changes in time series Hi-C data at different genomic scales.

2.2 Introduction

4D nuclear organization (4D Nucleome, 4DN) is defined by the dynamical interaction between 3D genome structure and function [23, 40, 109]. To analyze the 4DN, genome-wide chromosome conformation capture (Hi-C) and RNA sequencing

(RNA-seq) are often used to observe genome structure and function, respectively (Figure 2.1A). The availability and volume of Hi-C and RNA-seq data are expected to increase as high throughput sequencing costs decline, thus the development of methods to analyze these data is imperative. The relationship of genome structure and function has been studied previously [23, 45, 46, 74, 79], yet comprehensive and accessible tools for 4DN analysis are underdeveloped. The 4DNvestigator is a unified toolbox that loads time series Hi-C and RNA-seq data, extracts important structural and functional features (Figure 2.1B), and conducts both established and novel 4DN data analysis methods. We show that network centrality can be integrated with gene expression to elucidate structural and functional changes through time, and provide relevant links to the NCBI and GeneCards databases for biological interpretation of these changes [122, 140]. Furthermore, we utilize entropy to quantify the uncertainty of genome structure, and present a simple statistical method for comparing two or more Hi-C matrices.

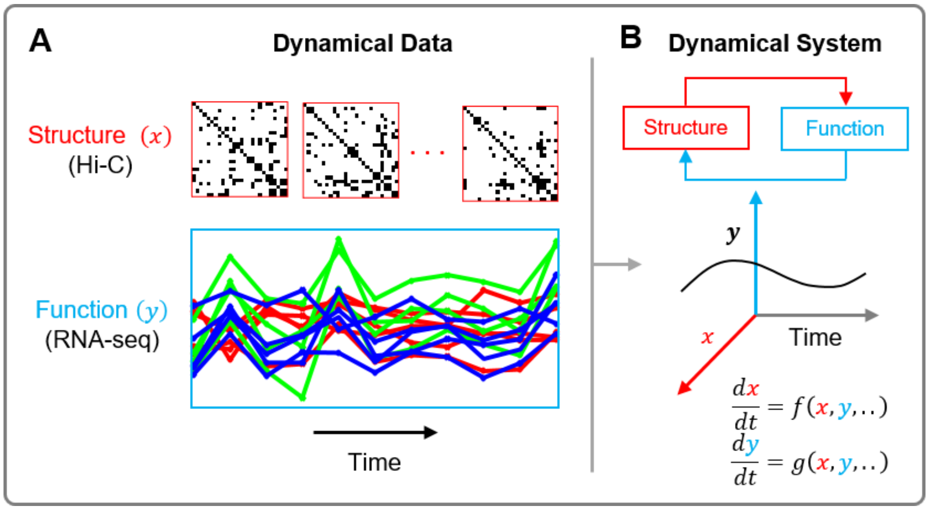


Figure 2.1: The 4D Nucleome. (A) Representative time series Hi-C and RNA-seq data correspond to genome structure and function, respectively. (B) Genome structure and function are intimately related. The 4DNvestigator integrates and visualizes time series data to study their dynamical relationship. This figure was taken from Lindsly *et al.* [76].

2.3 Materials and Methods

An overview of the 4DNvestigator workflow is depicted in Figure 2.2, and a Getting Started document is provided to guide the user through the main functionalities of the 4DNvestigator. The 4DNvestigator takes processed Hi-C and RNA-seq data as input, along with a metadata file which describes the sample and time point for each input Hi-C and RNA-seq file (See Supplementary Materials "Data Preparation" in Lindsly *et al.* [76]). A number of novel methods for analyzing 4DN data are included within the 4DNvestigator and are described below.

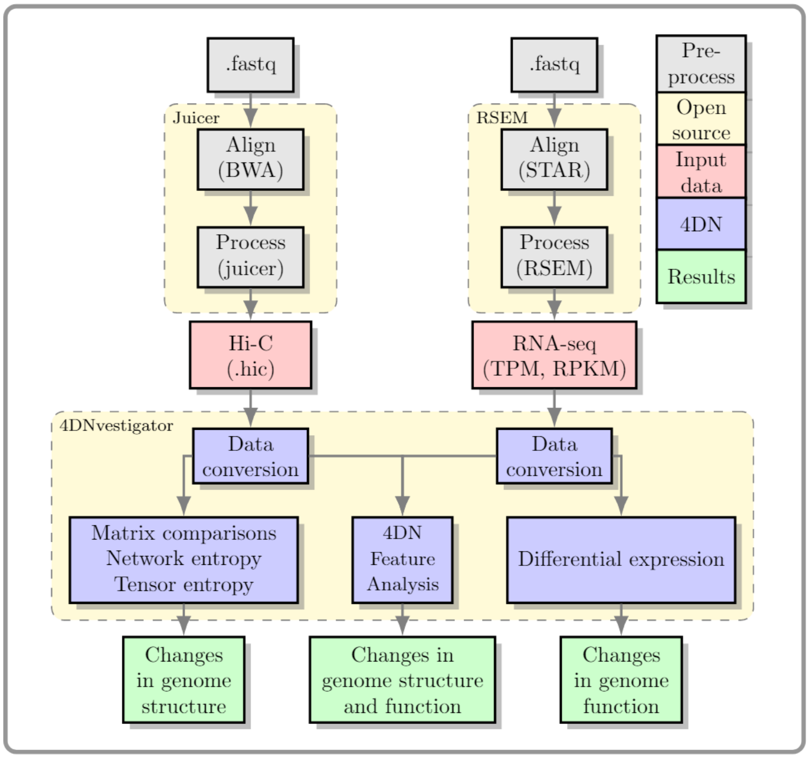


Figure 2.2: Overview of the 4DNvestigator data processing pipeline. Within this diagram, 4DN refers to the 4DNvestigator. This figure was taken from Lindsly *et al.* [76].

2.3.1 4DN Feature Analyzer

The "4DN feature analyzer" quantifies and visualizes how much a genomic region changes in structure and function over time. To analyze both structural and functional data, we consider the genome as a network. Nodes within this network are genomic loci, where a locus can be a gene or a genomic region at a particular resolution (i.e. 100 kb or 1 Mb bins). Edges in the genomic network are the relationships or interactions between genomic loci.

Algorithm 1 4DN Feature Analyzer

```
1: Input: Hi-C matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{n\times n}$, and RNA-seq vectors $\mathbf{r}^{(m)} \in \mathbb{R}^{n\times 1}$, $m = 1, \ldots, T$
2: Compute degree, eigenvector, betweenness, and closeness centrality of $\mathbf{A}^{(m)}$, and define as $\mathbf{b}_{deg}^{(m)}$, $\mathbf{b}_{eig}^{(m)}$, $\mathbf{b}_{bet}^{(m)}$, $\mathbf{b}_{close}^{(m)}$, respectively, where each $\mathbf{b}^{(m)} \in \mathbb{R}^{n\times 1}$
3: Compute the first principal component (PC1) of $\mathbf{A}^{(m)}$
4: Form the feature matrices $\mathbf{X}^{(m)} = [\mathbf{b}_{deg}^{(m)}, \mathbf{b}_{eig}^{(m)}, \mathbf{b}_{bet}^{(m)}, \mathbf{b}_{close}^{(m)}, \mathbf{r}^{(m)}]$, where $\mathbf{X}^{(m)} \in \mathbb{R}^{n\times 5}$
5: Normalize the columns of $\mathbf{X}^{(m)}$
6: Compute the common low dimensional space $\mathbf{Y}^{(m)}$
7: Visualize the low dimensional projection $\mathbf{Y}^{(m)}$ or 4DN phase plane
8: Return: Low dimensional space $\mathbf{Y}^{(m)}$ and genes in loci with the largest structure-function changes
```

2.3.1.1 Structural Data

Structure in the 4DN feature analyzer is derived from Hi-C data. Hi-C determines the edge weights in our genomic network through the frequency of contacts between genomic loci. To analyze genomic networks, we adopt an important concept from network theory called centrality. Network centrality is motivated by the identification of nodes which are the most "central" or "important" within a network [96]. The 4DN feature analyzer uses *degree*, *eigenvector*, *betweenness*, and *closeness* centrality (step 1 of Algorithm 1), which have been shown to be biologically relevant [79]. For example, eigenvector centrality can identify structurally defined regions of active/inactive gene expression, since it encodes clustering information of a network [79, 97]. Additionally, betweenness centrality measures the importance of nodes in regard to the flow of information between pairs of nodes. Boundaries between euchromatin and heterochromatin, which often change in reprogramming experiments, can be identified in a genomic network through betweenness centrality [79].

2.3.1.2 Functional Data

Function in the 4DN feature analyzer is derived from gene expression through RNA-seq. Function is defined as the $\log_2$ transformation of Transcripts Per Million (TPM) or Reads Per Kilobase Million (RPKM). For regions containing more than one gene, the mean expression of all genes within the region is used. The 4DN feature analyzer can also use other one-dimensional features (e.g. ChIP-seq, DNase-seq). The interpretation of the results and visualizations would change accordingly.

2.3.1.3 Integration of Data

Hi-C data are naturally represented as a matrix of contacts between genomic loci. Network centrality measures are one-dimensional vectors that describe important structural features of the genomic network. We combine network centrality with RNA-seq expression to form a structure-function "feature" matrix that defines the state of each genomic region at each time point (Figure 2.3A, step 3 of Algorithm 1). Within this matrix, rows represent genomic loci and columns are the centrality measures (structure) and gene expression (function) of each locus. The z-score for each column is computed to normalize the data (step 4 of Algorithm 1).

2.3.1.4 4DN Analysis

The 4DN feature analyzer reduces the dimension of the structure-function feature matrix for visualization and further analysis (steps 5 and 6 of Algorithm 1). We include the main linear dimension reduction method, Principal Component Analysis (PCA), and multiple nonlinear dimension reduction methods: Laplacian Eigenmaps (LE) [10], t-distributed Stochastic Neighbor Embedding (t-SNE) [134], and Uniform Manifold Approximation and Projection (UMAP) [85] (Figure 2.3C). These methods are described in more detail in Supplementary Materials "Dimension Reduction" in Lindsly *et al.* [76]. The 4DN feature analyzer can also visualize the dynamics of genome structure and function using the 4DN phase plane (step 6 of Algorithm 1) [23, 77]. We designate one axis of the 4DN phase plane as a measure of genome structure (e.g. eigenvector centrality) and the other as a measure of genome function (gene expression). Each point on the phase plane represents the structure and function of a genomic locus at a specific point in time (Figure 2.3B). The 4DN feature analyzer identifies genomic regions and genes with large changes in structure and function over time, and provides relevant links to the NCBI and GeneCard databases [122, 140].

2.3.2 Additional 4DNvestigator Tools

2.3.2.1 General Structure and Function Analysis

The 4DNvestigator also includes a suite of previously developed Hi-C and RNA-seq analysis methods. Euchromatin and heterochromatin compartments can be identified from Hi-C [25, 74], and regions that change compartments between samples are automatically identified. Significant changes in gene expression between RNA-seq samples can be determined through differential expression analysis using established methods [4].

2.3.2.2 Network Entropy

Entropy measures the amount of uncertainty within a system [34]. We use entropy to quantify the organization of chromatin structure from Hi-C data, where higher entropy corresponds to less structural organization. Since Hi-C is a multivariate analysis measurement (each contact coincidence involves two variables, the two genomic loci), we use multivariate entropy as follows:

$$\mathbf{Entropy} = -\sum_i \lambda_i \ln \lambda_i, \tag{2.1}$$

where λ_i represents the dominant features of the Hi-C contact matrix. In mathematics, these dominant features are called eigenvalues [123]. Biologically, genomic regions with high entropy likely correlate with high proportions of euchromatin, as euchromatin is more structurally permissive than heterochromatin [82, 105]. Furthermore, entropy can be used to quantify stemness, since cells with high pluripotency are less defined in their chromatin structure [86]. We provide the full algorithm for network entropy and calculate the entropy of Hi-C data from multiple cell types in Supplementary Materials "Network Entropy" in Lindsly *et al.* [76].

2.3.2.3 Tensor Entropy

The notion of transcription factories supports the existence of simultaneous interactions involving three or more genomic loci [33]. This implies that the configuration of the human genome can be more accurately represented by k-uniform hypergraphs, a generalization of networks in which each edge can join exactly k nodes (e.g. a standard network is a 2-uniform hypergraph). We can construct k-uniform hypergraphs from Hi-C contact matrices by computing the multi-correlations of genomic loci. Tensor entropy, an extension of network entropy, measures the uncertainty or disorganization of uniform hypergraphs [22]. Tensor entropy can be computed from the same entropy formula (2.1) with generalized singular values λ_j from tensor theory [22, 36]. We provide the definitions for multi-correlation and generalized singular values, the algorithm to compute tensor entropy, and an application of tensor entropy on Hi-C data in Supplementary Materials "Tensor Entropy" in Lindsly *et al.* [76].

2.3.2.4 Larntz-Perlman Procedure

The 4DNvestigator includes a statistical test, proposed by Larntz and Perlman (the LP procedure), that compares correlation matrices [67, 71]. The LP procedure is applied to correlation matrices from Hi-C data, and is able to determine whether multiple Hi-C samples are significantly different from one another. Suppose that

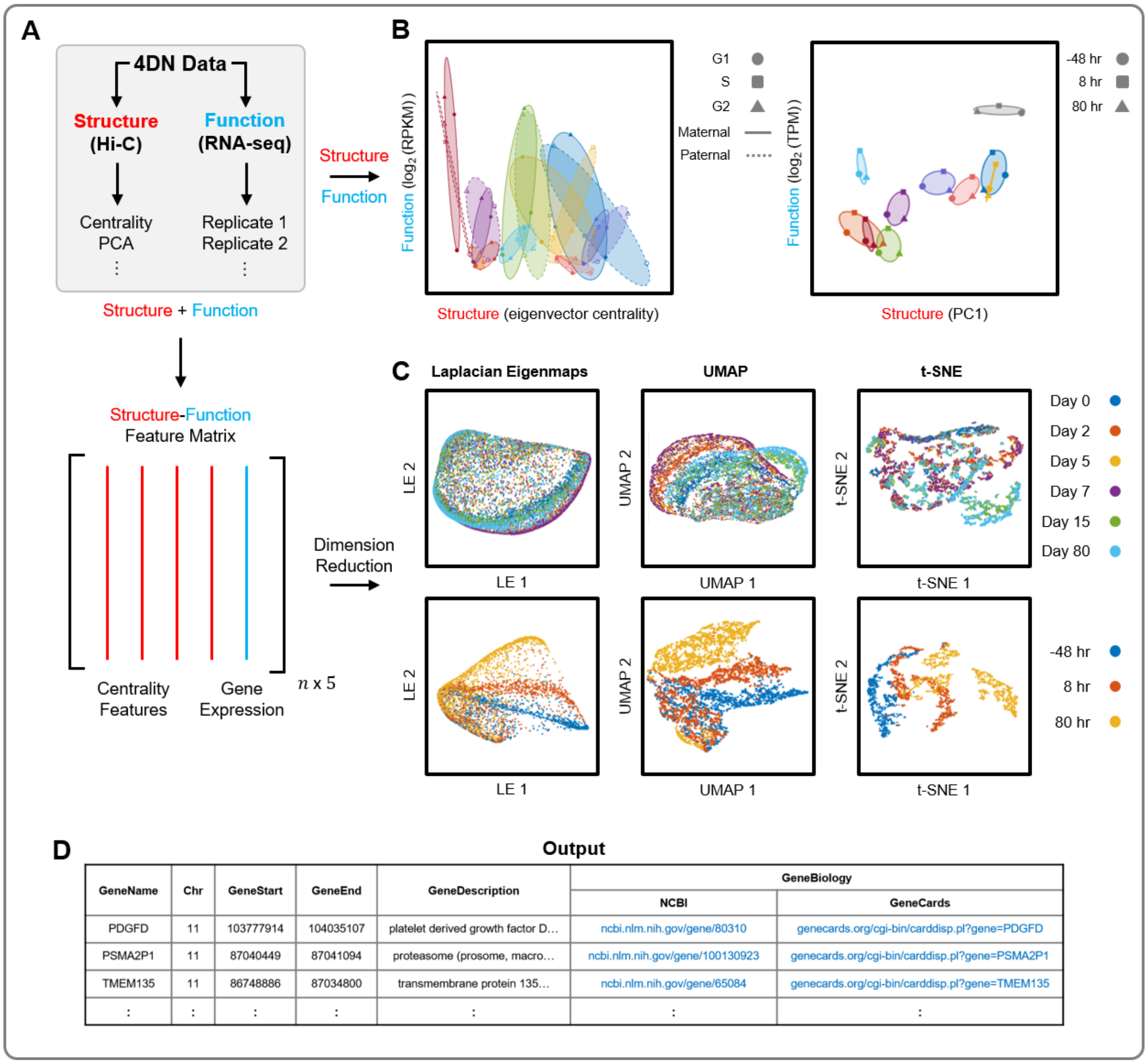
$\mathbf{C}^{(m)} \in \mathbb{R}^{n\times n}$ are the sample correlation matrices of Hi-C contacts with corresponding population correlation matrices $\mathbf{P}^{(m)} \in \mathbb{R}^{n\times n}$ for $m = 1, 2, \ldots, k$. The null hypothesis is $H_0 : \mathbf{P}^{(1)} = \cdots = \mathbf{P}^{(k)}$. First, compute the Fisher z-transformation $\mathbf{Z}^{(m)}$ by

$$\mathbf{Z}_{ij}^{(m)} = \frac{1}{2} \ln \frac{1 + \mathbf{C}_{ij}^{(m)}}{1 - \mathbf{C}_{ij}^{(m)}}. \tag{2.2}$$

Then form the matrices $\mathbf{S}^{(m)}$ such that

$$\mathbf{S}_{ij}^{(m)} = (n-3) \sum_{m=1}^{k} (\mathbf{Z}_{ij}^{(m)} - \bar{\mathbf{Z}}_{ij})^2, \tag{2.3}$$

where, $\bar{\mathbf{Z}}_{ij} = \frac{1}{k} \sum_{m=1}^{k} \mathbf{Z}_{ij}^{(m)}$. The test statistic is given by $T = \max_{ij} \mathbf{S}_{ij}$, and H_0 is rejected at level α if $T > \chi^2_{k-1,\epsilon(\alpha)}$ where $\chi^2_{k-1,\epsilon(\alpha)}$ is the chi-square distribution with $k-1$ degree of freedom, and $\epsilon(\alpha) = (1-\alpha)^{2/(n(n-1))}$ is the Šidák correction. Finally, calculate the p-value at which $T > \chi^2_{k-1,\epsilon(\alpha)}$. We note that this p-value is conservative, and that the the actual p-value may be smaller depending upon the amount of correlation among the variables. The LP procedure determines the statistical significance of any differences between multiple Hi-C samples for a genomic region of interest. We provide benchmark results of the LP procedure with other Hi-C comparison methods in Supplementary Materials “LP Procedure for Comparing Hi-C Matrices” in Lindsly *et al.* [76].



GeneName	Chr	GeneStart	GeneEnd	GeneDescription	GeneBiology	
					NCBI	GeneCards
PDGFD	11	103777914	104035107	platelet derived growth factor D…	ncbi.nlm.nih.gov/gene/80310	genecards.org/cgi-bin/carddisp.pl?gene=PDGFD
PSMA2P1	11	87040449	87041094	proteasome (prosome, macro…	ncbi.nlm.nih.gov/gene/100130923	genecards.org/cgi-bin/carddisp.pl?gene=PSMA2P1
TMEM135	11	86748886	87034800	transmembrane protein 135…	ncbi.nlm.nih.gov/gene/65084	genecards.org/cgi-bin/carddisp.pl?gene=TMEM135
:	:	:	:	:	:	:

Figure 2.3: 4DN feature analyzer. (A) 4DN data is input to the 4DN feature analyzer. Top: Structure data (Hi-C) is described using one dimensional features for compatibility with function data (RNA-seq). Bottom: Multiple structural features and function data are integrated into the structure-function feature matrix. (B) The 4DN feature analyzer can use structure and function data directly to visualize a system's dynamics using the 4DN phase plane [23, 77]. Structure defines the x-axis (left: eigenvector centrality, right: PC1) and function defines the y-axis (left: $\log_2$(RPKM), right: $\log_2$(TPM)), and points show structure-function coordinates through time. Left: Maternal and paternal alleles of nine cell cycle genes through G1, S, and G2/M phases of the cell cycle (adapted from [77]). Right: Top ten genomic regions (100 kb) with the largest changes in structure and function during cellular reprogramming [79]. (C) Multiple dimension reduction techniques can be used to visualize the 4DN feature analyzer's structure-function feature matrix (from left to right: LE, UMAP, and t-SNE). Top: 100 kb regions of Chromosome 4 across six time points during cellular differentiation [147]. Bottom: 100 kb regions of Chromosome 11 across three time points during cellular reprogramming [79]. (D) Example output of the 4DN feature analyzer. The output includes genes contained in loci with the largest changes, and links to their NCBI and GeneCards database entries [122, 140]. This figure was taken from Lindsly *et al.* [76].

2.4 Results

We demonstrate how the 4DN feature analyzer can process time series structure and function data (Figure 2.3A) with three examples (Figure 2.3B-D).

Example 1: Cellular Proliferation. Hi-C and RNA-seq data from B-lymphocytes (NA12878) capture the G1, S, and G2/M phases of the cell cycle for the maternal and paternal genomes [77]. We visualize the structure-function dynamics of the maternal and paternal alleles for nine cell cycle regulating genes using the 4DN phase plane (Figure 2.3B, left). We are interested in the importance of these genes within the genomic network through the cell cycle, so we use eigenvector centrality as the structural measure. This analysis highlights the coordination between the maternal and paternal alleles of these genes through the cell cycle.

Example 2: Cellular Differentiation. We constructed a structure-function feature matrix from time series Hi-C and RNA-seq data obtained from differentiating human stem cells [147]. These data consist of six time points which include human embryonic stem cells, mesodermal cells, cardiac mesodermal cells, cardiac progenitors, primitive cardiomyocytes, and ventricular cardiomyocytes [147]. We analyze Chromosome 4 across the six time points in 100 kb resolution by applying three dimension reduction techniques to the structure-function feature matrix: LE, UMAP, and t-SNE (Figure 2.3C, top). There is a better separation of the cell types during differentiation using UMAP and t-SNE than from LE. The optimal methods for visualization and analysis are data dependent, so the 4DNvestigator offers multiple tools for the user's own exploration of their data.

Example 3: Cellular Reprogramming. Time series Hi-C and RNA-seq data were obtained from an experiment that reprogrammed human dermal fibroblasts to the skeletal muscle lineage [79]. We analyze samples collected 48 hr prior to, 8 hr after, and 80 hr after the addition of the transcription factor MYOD1. The ten

100 kb regions from Chromosome 11 that varied most in structure and function are visualized using the 4DN phase plane in Figure 2.3B (right). We also construct a structure-function feature matrix of Chromosome 11 in 100 kb resolution. Similar to the differentiation data analysis, we use LE, UMAP, and t-SNE to visualize the structure-function dynamics. These low dimensional projections show the separation of the three time points corresponding to before, during, and after cellular reprogramming (Figure 2.3C, bottom). We show an example output of the 4DN feature analyzer, which highlights genes contained in the genomic loci that have the largest structure-function changes through time and provides links to the NCBI and GeneCards database entries for these genes (Figure 2.3D) [122, 140].

2.5 Discussion

The 4DNvestigator provides rigorous and automated analysis of Hi-C and RNA-seq time series data by drawing on network theory, information theory, and multivariate statistics. It also introduces a simple statistical method for comparing Hi-C matrices, the LP procedure. The LP procedure is distinct from established Hi-C matrix comparison methods, as it takes a statistical approach to test for matrix equality, and allows for the comparison of many matrices simultaneously. Thus, the 4DNvestigator provides a comprehensive toolbox that can be applied to time series Hi-C and RNA-seq data simultaneously or independently. These methods are important for producing rigorous quantitative results in 4DN research.

CHAPTER III

Functional Organization of the Maternal and Paternal Human 4D Nucleome

This chapter is based on a paper by Stephen Lindsly, Wenlong Jia, Haiming Chen, Sijia Liu, Scott Ronquist, Can Chen, Xingzhao Wen, Cooper Stansbury, Gabrielle A. Dotson, Charles Ryan, Alnawaz Rehemtulla, Gilbert S. Omenn, Max Wicha, Shuai Cheng Li, Lindsey Muir, and Indika Rajapakse [77] (under review).

3.1 Abstract

Every human somatic cell inherits a maternal and a paternal genome, which work together to give rise to cellular phenotypes. However, the allele-specific relationship between gene expression and genome structure through the cell cycle is largely unknown. By integrating haplotype-resolved genome-wide chromosome conformation capture, mature and nascent mRNA, and protein binding data, we investigate this relationship both globally and locally. We introduce the maternal and paternal 4D Nucleome, enabling detailed analysis of the mechanisms and dynamics of genome structure and gene function for diploid organisms. Our analyses find significant coordination between allelic expression biases and local genome conformation, and notably absent expression bias in universally essential cell cycle and glycolysis genes. We pro-

pose a model in which coordinated biallelic expression reflects prioritized preservation of essential gene sets.

3.2 Introduction

Biallelic gene expression in diploid genomes inherently protects against potentially harmful mutations. Disrupted biallelic expression of certain genes increases vulnerability to disease in humans, such as in familial cancer syndromes that have loss of function in one allele [66]. BRCA1 and BRCA2 are quintessential examples, for which missense, nonsense, or frameshift mutations affecting function of one allele significantly increase the risk of breast cancer in women [57, 84]. Imprinted genes are also associated with multiple disease phenotypes such as Angelman and Prader-Willi syndromes [18, 146]. Other genes with monoallelic or allele-biased expression (MAE, ABE) may be associated with disease, but the contribution of allelic bias to disease phenotypes remains poorly understood.

ABE can occur with single nucleotide variants (SNVs), insertions or deletions (InDels), and chromatin modifications [28, 69, 107, 112, 128]. Analyses of allelic bias suggest high variance across tissues and individuals, with estimates ranging from 4% to 26% of genes in a given setting [72, 112]. In addition, higher order chromatin conformation and spatial positioning in the nucleus shape gene expression [31, 91, 92, 106]. As the maternal and paternal alleles can be distant in nucleus, their spatial positions may promote ABE [9, 35].

A major step towards understanding the contribution of allelic bias to disease is to identify ABE genes, recognizing that important biases may be transient and challenging to detect. Allele-specific expression and 3D structures are not inherently accounted for in genomics methods such as RNA-sequencing and genome-wide chromosome conformation capture (Hi-C). These limitations complicate interpretations of structure-function relationships, and complete phasing of the two genomes remains a significant challenge.

To improve understanding of ABE in genomic structure-function relationships, we developed a novel phasing algorithm for Hi-C data, which we integrate with allele-specific RNA-seq and Bru-seq data across three phases of the cell cycle in human B-lymphocytes (Figure 3.1). RNA-seq and Bru-seq data were separated into their maternal and paternal components through SNVs/InDels [104]. Our algorithm, HaploHiC, uses phased SNVs/InDels to impute Hi-C reads of unknown parental origin. Publicly available allele-specific protein binding data (ChIP-seq) were also included to better understand potential regulatory elements involved in allelic bias [26, 112]. In addition to identifying known ABE genes silenced by X-Chromosome inactivation (XCI) or imprinting, our analyses find novel expression biases between alleles and cell cycle phases in several hundred genes, many of which had corresponding bias in allele-specific protein binding. Furthermore, the alleles of ABE genes were significantly more likely to differ in local structure compared to randomly selected alleles. In contrast, we observed a pronounced lack of ABE in crucial biological pathways and essential genes. Our findings highlight advantages of integrating genomics analyses in a cell cycle and allele-specific manner and represent an allele-specific extension of the 4D Nucleome (4DN) [23, 40, 76, 109]. This approach will be beneficial to investigation of human phenotypic traits and their penetrance, genetic diseases, vulnerability to complex disorders, and tumorigenesis.

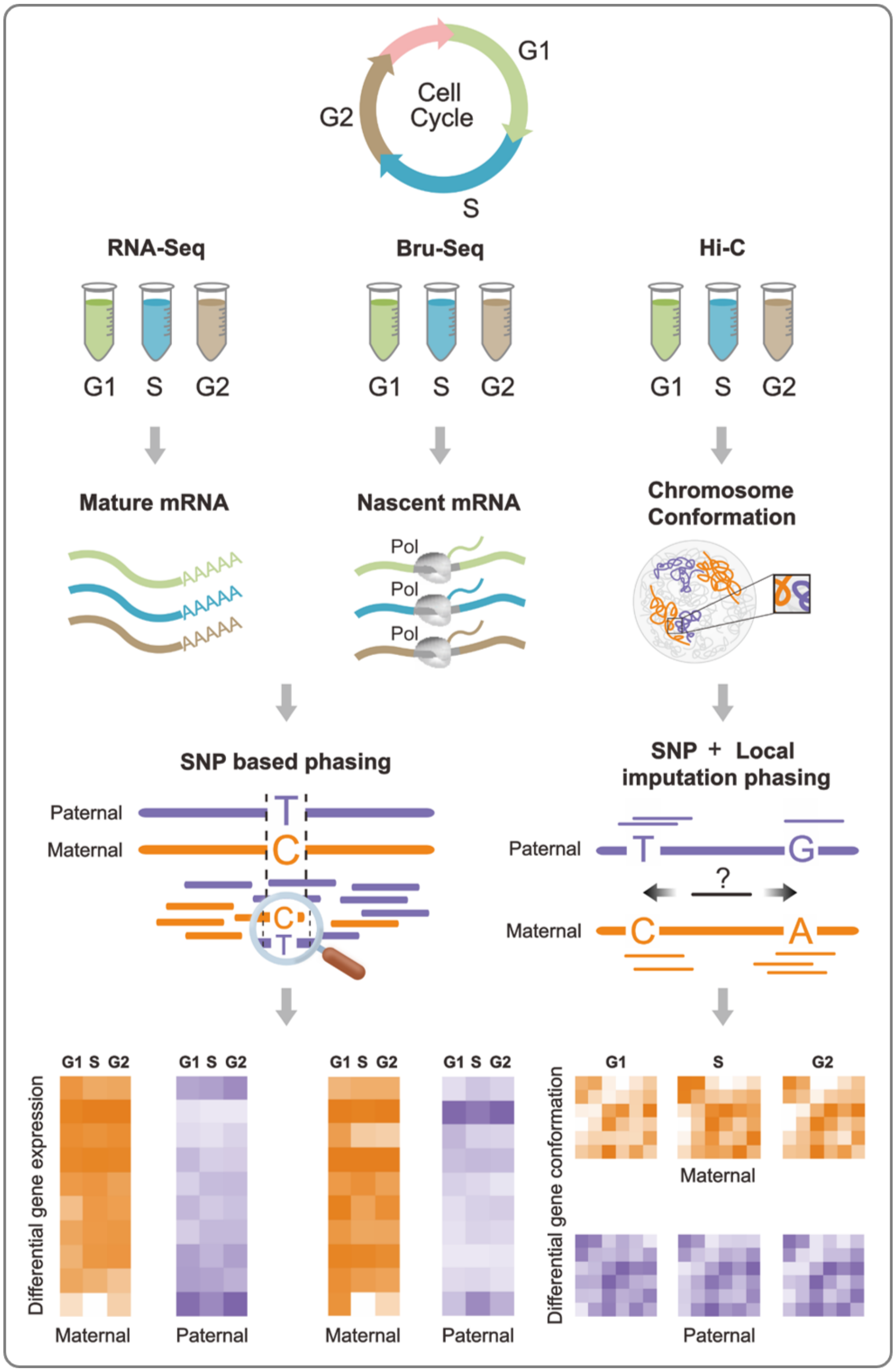


Figure 3.1: Experimental and allelic separation workflow. Cell cycle sorted cells were extracted for RNA-seq, Bru-seq, and Hi-C (left to right, respectively). RNA-seq and Bru-seq data were allelicly phased via SNVs/InDels (left). SNV/InDel based imputation and haplotype phasing of Hi-C data using HaploHiC (right). These data provide quantitative measures of structure and function of the maternal and paternal genomes through the cell cycle. This figure was taken from Lindsly *et al.* [77].

3.3 Materials and Methods

3.3.1 Cell Culture and Cell Cycle Sorting

Human GM12878 cells were cultivated in RPMI1640 medium supplemented with 10% fetal bovine serum (FBS). Live cells were stained with Hoechst 33342 (Cat #B2261, Sigma-Aldrich), and then sorted by fluorescence-activated cell sorting (FACS) to obtain cell fractions at the corresponding cell cycle phases G1, S, and G2/M (Figure S2 in Lindsly *et al.* [77]).

3.3.2 RNA-seq and Bru-seq Sequencing

Total RNA was extracted from sorted live cells for both RNA-seq and Bru-seq. We performed 5'-bromouridine (Bru) incorporation in live cells for 30 minutes, and the Bru-labeled cells were then stained on ice with Hoechst 33342 for 30 minutes before sorting at 4°C to isolate G1, S, and G2/M phase cells. The sorted cells were immediately lysed in TRizol (Cat # 15596026, ThermoFisher) and frozen. To isolate Bru-labeled RNA, DNAse-treated total RNA was incubated with anti-BrdU antibodies conjugated to magnetic beads [74]. We converted the transcripts from the RNA-seq and Bru-seq experiments for all samples into cDNA libraries and deep-sequenced at 50-base length on an Illumina HiSeq2500 platform. The RNA-seq and Bru-seq data each consist of three biological replicates. From our RNA-seq replicates, we obtained a total of 193.4, 197.2, and 202.0 million raw reads for G1, S, and G2/M, respectively. From our Bru-seq replicates, we obtained a total of 162.5, 149.9, and 138.0 million raw reads for G1, S, and G2/M, respectively.

3.3.3 Hi-C Sequencing

For cells used in construction of Hi-C libraries, cells were crosslinked with 1% formaldehyde, the reaction was neutralized with 0.125 M glycine, then cells were stained with Hoechst 33342 and sorted into G1, S, and G2/M fractions. Cross-linked chromatin was digested with the restriction enzyme MboI for 12 hours. The restriction enzyme fragment ends were tagged with biotin-dATP and ligated in situ. After ligation, the chromatins were de-cross-linked, and DNA was isolated for fragmentation. DNA fragments tagged by biotin-dATP, in the size range of 300-500 bp, were pulled down for sequencing adaptor ligation and polymerase chain reaction (PCR) products. The PCR products were sequenced on an Illumina HiSeq2500 platform. Respective to G1, S, and G2/M, we obtained 512.7, 550.3, and 615.2 million raw Hi-C sequence reads.

3.3.4 RNA-seq and Bru-seq Data Processing

RNA-seq and Bru-seq analysis were performed as previously described [102, 115]. Briefly, Bru-seq used Tophat (v1.3.2) to align reads without de novo splice junction calling after checking quality with FastQC (version 0.10.1). 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 [102] for full details). Similarly for RNA-seq data processing, the raw reads were checked with FastQC. Tophat (version 2.0.11) and Bowtie (version 2.1.0.0) were used to align the reads to the reference transcriptome (HG19). Cufflinks (version 2.2.1) was used for expression quantification, using UCSC hg19.fa and hg19.gtf as the reference genome and transcriptome, respectively. A locally developed R script using CummeRbund was used to format the Cufflinks output.

3.3.5 Separation of Maternal and Paternal RNA-seq and Bru-seq Data

To determine allele-specific transcription and gene expression through Bru-seq and RNA-seq, all reads were aligned using GSNAP, a SNV aware aligner [68, 144]. HG19 and UCSC gene annotations were used for the reference genome and gene annotation, respectively. The gene annotations were used to create the files for mapping to splice sites (used with –s option). Optional inputs to perform SNV aware alignment were also included. Specifically, –v was used to include the list of heterozygous SNVs and –use-sarray=0 was used to prevent bias against non-reference alleles [38].

After alignment, the output SAM files were converted to BAM files, sorted and indexed using SAMtools [73]. SNV alleles were quantified using bam-readcounter to count the number of each base that was observed at each of the heterozygous SNV locations. Allele-specificity of each gene was then assessed by combining all of the SNVs in each gene. For RNA-seq, only exonic SNVs were used. Bru-seq detects nascent transcripts containing both exons and introns, so both exonic and intronic SNVs were used. Maternal and paternal gene expression were calculated by multiplying the genes' overall read counts by the fraction of the SNV-covering reads that were maternal and paternal, respectively. We identified 266,899 SNVs from the Bru-seq data, compared with only 65,676 SNVs from RNA-seq data. However in the Bru-seq data, many SNVs have low read coverage depth. We required at least 5 SNV-covering reads for a SNV to be used to separate the maternal and paternal contributions to gene expression. This criterion found that there were similar numbers of informative SNVs (19,394 and 19,998) in the RNA-seq and Bru-seq data, respectively. Genes which did not contain informative SNVs were divided equally into their maternal and paternal contributions.

3.3.6 Allele-specific Differential Expression

For a gene's expression to be considered for differential expression analysis, we require each of the three replicates to have an average of at least 10 SNV-covering reads mapped to at least one of the alleles in all three cell cycle phases. This threshold was introduced to reduce the influence of technical noise on our differential expression results. From the 23,277 Refseq genes interrogated, there were 4,193 genes with at least 10 read counts mapped to either the maternal or paternal allele (or both) in the RNA-seq data. From Bru-seq, there were 5,294 genes using the same criterion. We refer to these genes as "allele-specific" genes for their respective data sources. We observed that there were larger variances between samples and lower read counts in the Bru-seq data set than in RNA-seq. We identified differentially expressed genes between alleles and between cell cycle phases for both RNA-seq and Bru-seq using a MATLAB implementation of DESeq [4]. To reduce the possibility of false positives when determining differential expression, we imposed a minimum FPKM level of 0.1, a false discovery rate adjusted p-value threshold of 0.05, and a fold change cutoff of $\mathrm{FC} > 2$ for both RNA-seq and Bru-seq [12].

3.3.7 Separation of Maternal and Paternal Hi-C Data by HaploHiC

Hi-C library construction and Illumina sequencing were performed using established methods [107]. In this study, we separate the maternal and paternal genomes' contributions to the Hi-C contact matrices to analyze their similarities and differences in genome structure. In order to determine which Hi-C reads come from which parental origin, we utilize differences in genomic sequence at phased SNVs/InDels. As these variations are unique to the maternal and paternal genomes, they can be used to distinguish reads. When attempting to separate the maternal and paternal genomes, complications arise when there are sections of DNA that are identical. There are a relatively small number of allele-specific variations, and the resulting segregated maternal and paternal contact matrices are sparse. In order to combat this problem, we seek to infer contacts of unknown parental origin.

We propose a novel technique, HaploHiC, for phasing reads of unknown parental origin using local imputation from known reads. HaploHiC uses a data-derived ratio based on the following hypothesis: if the maternal and paternal genomes have different 3D structures, we can use the reads with known origin (at SNV/InDel loci) to predict the origin of neighboring unknown reads (Figure 3.4A, Supplemental Methods A in Lindsly *et al.* [77]) [116]. For example, if we observe that many contacts between two loci can be directly mapped to the paternal genome but few to the maternal genome, then unphased contacts between those loci are more likely to be from the paternal genome as well, and vice versa. This process of imputing Hi-C reads of unknown origin based on nearby known reads is similar to the methods developed by Tan *et al.* [128].

HaploHiC marks paired-end reads as haplotype-known or -unknown depending on their coverage of heterozygous phased SNVs/InDels. Haplotype-known reads are directly assigned to their corresponding haplotype, maternal or paternal. HaploHiC uses a local contacts-based algorithm to impute the haplotype of haplotype-unknown reads using nearby SNVs/InDels. If the minimum threshold (ten paired-ends) of haplotype-known reads for local imputation is not reached, HaploHiC randomly assigns the haplotype-unknown reads to be maternal or paternal (less than 5% of all haplotype-unknown reads). Detailed materials and methods for haplotype phasing and Hi-C construction are provided in Supplemental Methods A in Lindsly *et al.* [77].

Our validation shows that HaploHiC performs well, with an average accuracy of 96.9%, 97.2%, and 97.3% for G1, S, and G2, respectively, over 10 trials each (Table S12 in Lindsly *et al.* [77]). Each validation trial randomly removed 10% of the heterozygous phased SNVs/InDels, and calculated imputation accuracy by the fraction of correctly imputed reads from the haplotype-known Hi-C reads covering these removed heterozygous mutations (Supplemental Methods A.8 in Lindsly *et al.* [77]). Our validation of imputation accuracy is similar to the method presented in Tan *et al.* [128]. We also perform multiple simulations for further validation (Supplemental Methods A.8 in Lindsly *et al.* [77]). HaploHiC is available through a GitHub repository.

After haplotype assignment through HaploHiC, Hi-C paired-end reads (PE-reads) were distributed to intra-haplotype (P-P and M-M) and inter-haplotype (P-M and M-P) (Supplemental Methods A.1-A.8 in Lindsly *et al.* [77]). Juicer was applied on intra-haplotype PE-reads, and outputs maternal and paternal contact matrices which were normalized through the Knight-Ruiz method of matrix balancing [50, 65]. Inter-haplotype contact matrices were generated by HaploHiC (Supplemental Methods A.7 in Lindsly *et al.* [77]). Intra- and inter-haplotype contacts are shown in Figure 3.4B and Figure S8 in Lindsly *et al.* [77]. Both base pair level and fragment level matrices were constructed. The resolution of base pair level matrices are 1 Mb and 100 kb. Gene-level contacts were converted from fragment level matrices by HaploHiC.

3.4 Results

3.4.1 Chromosome-Scale Maternal and Paternal Differences

Spatial positioning of genes within the nucleus is known to be associated with transcriptional status [31, 91, 92, 106]. One might expect that the maternal and paternal copies of each chromosome would stay close together to ensure that their respective alleles have equal opportunities for transcription. Imaging of chromosome territories has shown that this is often not the case (Figure 3.2A, Supplemental Methods B in Lindsly *et al.* [77]) [17, 110, 119]. This inspired us to investigate whether the two genomes operate in a symmetric fashion, or if allele-specific differences exist between the genomes regarding their respective chromatin organization patterns (structure) and gene expression profiles (function). We analyzed parentally phased whole-chromosome Hi-C and RNA-seq data at 1 Mb resolution to identify allele-specific differences in structure and function, respectively. We subtracted each chromosome's paternal Hi-C matrix from the maternal matrix and found the Frobenius norm of the resulting difference matrix. The Frobenius norm is a measure of distance between matrices, where equivalent maternal and paternal genome structures would result in a value of zero. Similarly, we subtracted the phased RNA-seq vectors in log_2 scale and found the Frobenius norm of each difference vector. The Frobenius norms were adjusted for chromosome size and normalized for both Hi-C and RNA-seq.

We found that all chromosomes have allelic differences in both structure (Hi-C, blue) and function (RNA-seq, red) (Figure 3.2B). Chromosome X had the largest structural difference, as expected, followed by Chromosomes 9, 21, and 14. Chromosome X had the most extreme functional differences as well, followed by Chromosomes 13, 7, and 9. A threshold was assigned at the median Frobenius norm for Hi-C and RNA-seq (Figure 3.2B green dashed lines). The majority of chromosomes with larger structural differences than the median in Hi-C also have larger functional differences than the median in RNA-seq. There is a positive correlation between chromosome level differences in structure and function, which is statistically significant only when Chromosome X is included (R = 0.66 and $p < 0.05$ with Chromosome X; R = 0.30 and $p = 0.17$ without Chromosome X).

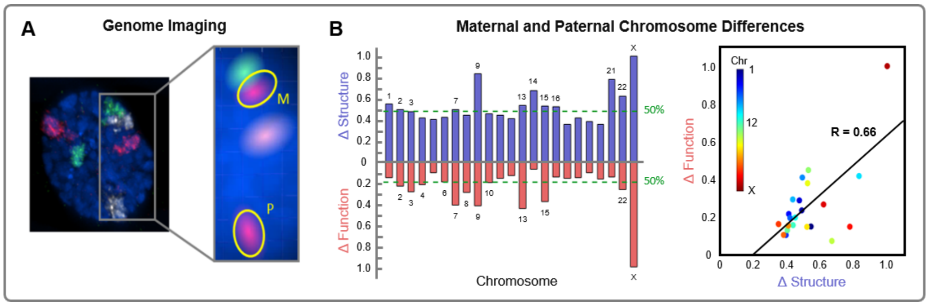


Figure 3.2: Genome imaging and chromosome differences. **(A)** Nucleus of a primary human fibroblast imaged using 3D FISH with the maternal and paternal copies of Chromosome 6, 8, and 11 painted red, green, and white, respectively (left). Subsection highlighting the separation between the maternal and paternal copies of Chromosome 11, now colored red (right). **(B)** Normalized chromosome level structural and functional parental differences of GM12878 cells. Structural differences (Δ Structure, blue) represent the aggregate changes between maternal and paternal Hi-C over all 1 Mb loci for each chromosome, adjusted for chromosome size in G1. Functional differences (Δ Function, red) represent the aggregate changes between maternal and paternal RNA-seq over all 1 Mb loci for each chromosome, adjusted for chromosome size in G1. Green dashed lines correspond to the median structural (0.48, chromosome 3) and functional (0.20, chromosome 6) differences, in the top and bottom respectively, and all chromosomes equal to or greater than the threshold are labeled. Scatter plot of maternal and paternal differences in structure and function with best-fit line (R = 0.66 and p < .05). This figure was taken from Lindsly *et al.* [77].

3.4.2 Allele-Specific RNA Expression

After confirming allelic differences in RNA expression at the chromosomal scale, we examined allele-specific expression of individual genes through RNA-seq and Bru-seq. We hypothesized that the chromosome scale expression differences were not only caused by known cases of ABE such as XCI and imprinting, but also by widespread ABE over many genes [41, 56]. Therefore, we evaluated all allele-specific genes (genes with sufficient reads covering SNVs/InDels) for differential expression across the six settings: maternal and paternal in G1, S, and G2/M (hereafter, G2). These settings give rise to seven comparisons which consist of maternal versus paternal within each of the cell cycle phases (three comparisons), as well as G1 versus S and S versus G2 for the maternal and paternal genomes, respectively (two comparisons for each genome).

First, we identified genes with ABE and cell cycle-biased expression (CBE) from RNA-seq. While ABE refers to differential expression between alleles in each cell cycle phase, CBE refers to significant changes in expression from one cell cycle phase to another in each allele. From 23,277 RefSeq genes interrogated, there were 4,193 genes with sufficient coverage on SNVs/InDels to reliably determine allele-specific expression [98]. We performed differential expression analysis for the seven comparisons to identify which of the 4,193 genes had ABE or CBE [4]. We identified 615 differentially expressed genes from RNA-seq: 467 ABE genes, 229 CBE genes, and 81 genes with both ABE and CBE (Tables S2, S4 in Lindsly *et al.* [77]). Both exons and introns containing informative SNVs/InDels were used for our Bru-seq data, from which 5,294 genes had sufficient coverage. We identified 505 differentially expressed genes from Bru-seq: 380 ABE genes, 164 CBE genes, and 39 genes with both ABE and CBE (Tables S3, S5 in Lindsly *et al.* [77]). We also identified 130 genes that had ABE in both RNA-seq and Bru-seq. While this is substantially smaller than total number of ABE genes for RNA-seq and Bru-seq (467 and 380, respectively), the number of genes that are allele-specific in both data modalities is also smaller. That is, only 285 of the ABE genes from RNA-seq are allele-specific in Bru-seq and 192 of the ABE genes from Bru-seq are allele-specific in RNA-seq. The remaining genes did not have sufficient expression or SNV coverage to be included in the downstream analysis. We then separated the differentially expressed genes into their respective chromosomes to observe their distribution throughout the genome. From RNA-seq (Bru-seq), we found that autosomes had 3-14% ABE (1-11%) in their allele-specific genes which is comparable to previous findings [72]. As expected, Chromosome X had a particularly high percentage of ABE genes at 90% (91%).

We identified 288 genes that have ABE in all three cell cycle phases from RNA-seq (160 paternally biased, 128 maternally biased) and 173 from Bru-seq (129 paternally biased, 44 maternally biased). This is the most common differential expression pattern among ABE genes and these genes form the largest clusters in Figure 3.3A. These clusters include, but are not limited to, XCI, imprinted, and other MAE genes. Known examples within these clusters are highlighted in the ‘X-Linked’ and ‘Imprinted’ sections of Figure 3.3B. We also identified hundreds of genes that are not currently appreciated in literature to have ABE, with examples shown in the ‘Autosomal Genes’ sections of Figure 3.3B for both mature and nascent RNA. Approximately half of all ABE genes were only differentially expressed in one or two cell cycle phases, which we refer to as transient allelic biases. These genes form the smaller clusters seen in Figure 3.3A. Examples of genes with transient allelic biases are also presented in the ‘Autosomal Genes’ section of Figure 3.3B. Transient expression biases like these may be due to coordinated expression of the two alleles in only certain cell cycle phases, though the mechanism behind this behavior is unclear.

Among the ABE genes from RNA-seq analysis, we found 117 MAE genes. In addition to the requirements for differential expression, we impose the thresholds of a FC ≥ 10 and for the inactive allele to have < 0.1 Fragments Per Kilobase of transcript per Million (FPKM), or FC ≥ 50 across all three cell cycle phases. Our analysis confirmed MAE for imprinted and XCI genes, with examples shown in Figure 3.3B. Imprinted and XCI genes are silenced via transcriptional regulation, which was verified by monoallelic nascent RNA expression (Bru-seq). The *XIST* gene, which is responsible for XCI, was expressed in the maternal allele reflecting the deactivation of the maternal Chromosome X. XCI was also observed from Hi-C through large heterochromatic domains in the maternal Chromosome X, and the absence of these domains in the paternal Chromosome X (Figure 3.4C). The inactive Chromosome X in our cells is opposite of what is commonly seen for the GM12878 cell line in literature (likely due to our specific GM12878 sub-clone), but is consistent between our data modalities [26, 107, 128]. The MAE genes also include six known imprinted genes, four expressed from the paternal allele (*KCNQ1OT1*, *SNRPN*, *SNURF*, and *PEG10*) and two from the maternal allele (*NLRP2* and *HOXB2*). Some of the known imprinted genes that were confirmed in our data are associated with imprinting diseases, such as Beckwith-Wiedemann syndrome (*KCNQ1OT1* and *NLRP2*), Angelman syndrome (*SNRPN* and *SNURF*), and Prader-Willi syndrome (*SNRPN* and *SNURF*) [1, 20]. These genes and their related diseases offer further support for allele-specific analysis, as their monoallelic expression could not be detected otherwise.

After observing that approximately half of all ABE genes had transient expression biases, we hypothesized that alleles may have unique dynamics through the cell cycle. We then focused our investigation on allele-specific gene expression through the cell cycle to determine if alleles had CBE, and whether alleles were coordinated in their cell-cycle dependent expression (Figure S1B in Lindsly *et al.* [77]). We compared the expression of each allele between G1 and S as well as between S and G2, which provides insight into the differences between maternal and paternal alleles' dynamics across the cell cycle. In the G1 to S comparison, there are 88 (55) genes in RNA-seq (Bru-seq) which have similar expression dynamics in both alleles. These genes' maternal and paternal alleles are similarly upregulated or downregulated from G1 to

S. In contrast, 87 (97) genes in RNA-seq (Bru-seq) have different expression dynamics between alleles. That is, only one allele is up or downregulated in the transition from G1 to S. In the S to G2 comparison, there are 26 (3) genes in RNA-seq (Bru-seq) that have similar expression dynamics in both alleles and 56 (12) genes with different expression dynamics between alleles. From these data, we see a coordination of expression between many, but certainly not all, alleles through the cell cycle.

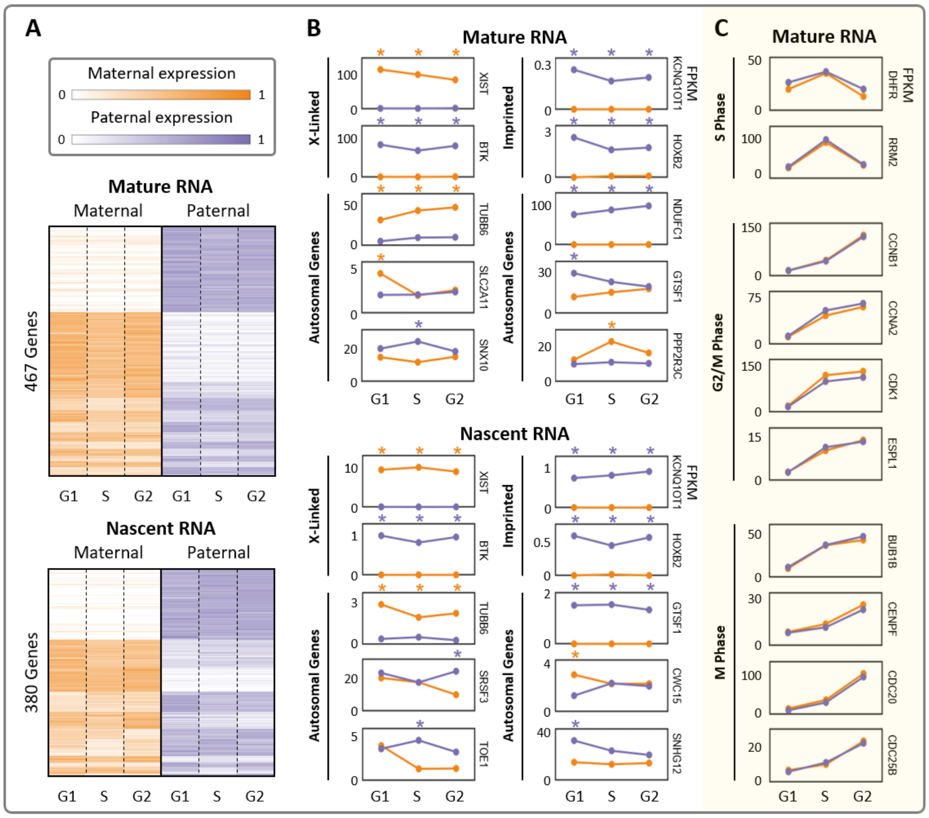


Figure 3.3: Allele-specific mature and nascent RNA expression. **(A)** Differentially expressed genes' maternal and paternal RNA expression through the cell cycle. Expression heatmaps are average FPKM values over three replicates after row normalization. Genes are grouped by their differential expression patterns (Figure S1 in Lindsly *et al.* [77]). **(B)** Representative examples of X-linked, imprinted, and other autosomal genes with allelic bias. Top shows mature RNA levels (RNA-seq) and bottom shows nascent RNA expression (Bru-seq) in both (A) and (B). **(C)** Examples of cell cycle regulatory genes' mature RNA levels through the cell cycle. These genes are grouped by their function in relation to the cell cycle and all exhibit CBE but none have ABE. All example genes in (B) and (C) reflect average FPKM values over three replicates, and ABE in a particular cell cycle phase is marked with an orange or purple asterisk for maternal or paternal bias, respectively. G2 includes both G2 and M phase. This figure was taken from Lindsly *et al.* [77].

3.4.3 Biallelic Expression and Cellular Function

We observed from our analysis of CBE that multiple cell cycle regulatory genes had no instances of ABE (Figure 3.3C). We expanded this set of genes to include all allele-specific genes contained in the KEGG cell cycle pathway [61]. Again, we found zero instances of ABE. This may suggest that genes with certain crucial cellular functions, like cell cycle regulation, may have coordinated biallelic expression to ensure their sufficient presence as a means of robustness. This is supported by previous findings which showed restricted genetic variation of enzymes in the essential glycolytic pathway [27].

We hypothesized that genes implicated in critical cell cycle processes would be less likely to have ABEs. We tested additional modules derived from KEGG pathways containing at least five allele-specific genes, with the circadian rhythm module supplemented by a known core circadian gene set [23]. Examples of modules with varying proportions of ABE are shown in Table 3.1 (Table S7 in Lindsly *et al.* [77]), where "Percent ABE" refers to the proportion of genes with ABE to the total number of allele-specific genes in that module. We found that there are multiple crucial modules, including the glycolysis/gluconeogenesis and pentose phosphate pathways, which also had zero ABE genes. To explore the possibility of a global phenomenon by which genes essential to cellular fitness are significantly less likely to have biased expression, we analyzed the frequency of ABE in 1,734 genes experimentally determined to be essential in human cells [16]. Using the 662 allele-specific genes in this set, we found that these essential genes were significantly less likely to have ABE than a random selection of allele-specific genes (5.8% versus 11.1%, $p < 0.001$, Supplemental Methods D in Lindsly *et al.* [77]), consistent with our hypothesis that critical genes are likely to be expressed by both alleles. In total, we offer a model in which coordinated biallelic expression reflects prioritized preservation of essential gene sets.

Table 3.1: Allelic bias in biological modules.

Module	**Percent ABE**
Cell Cycle	0%
Glycolysis/Gluconeogenesis	0%
Pentose Phosphate	0%
BCR Signaling	8%
Circadian Rhythm	9%
p53 Signaling	11%
Wnt Signaling	16%
Hippo Signaling	21%
Whole Genome	**11%**
Pathways with 0% ABE are highlighted	

3.4.4 Allele-Specific Genome Structure

Motivated by our observations of chromosome level structural differences between the maternal and paternal genomes, we examined the HaploHiC separated data in more detail to determine where these differences reside. The genome is often categorized into two compartments: transcriptionally active euchromatin and repressed heterochromatin. In studies comparing multiple types of cells or cells undergoing differentiation, areas of euchromatin and heterochromatin often switch corresponding to genes that are activated/deactivated for the specific cell type [43]. We explored this phenomenon in the context of the maternal and paternal Hi-C matrices to determine if the two genomes had differing chromatin compartments. Chromatin compartments can be identified from Hi-C data using methods such as principal component analysis or spectral clustering (Supplemental Methods C in Lindsly *et al.* [77]) [25]. We applied spectral clustering to every chromosome across all three cell cycle phases at 1 Mb resolution. We found that there were slight changes in chromatin compartments for all chromosomes, but the vast majority of these changes took place on the borders between compartments rather than an entire region switching compartments. These border differences were not enriched for ABE genes. This implies that, although the structures may not be identical, the maternal and paternal genomes have similar overall compartmentalization (aside from Chromosome X).

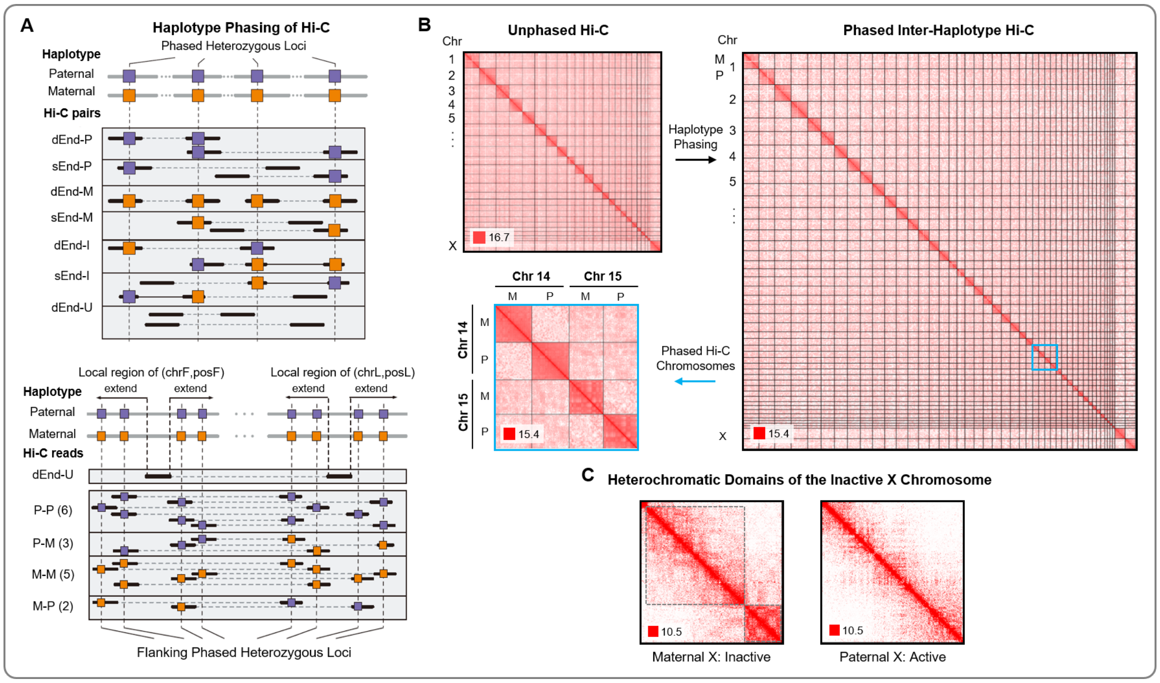


Figure 3.4: Haplotype phasing of Hi-C data. **(A)** HaploHiC separates paired-end reads into groups based on parental origin determined through SNVs/InDels (left, Supplemental Methods A.3 in Lindsly *et al.* [77]). Reads are grouped by: (i) reads with one (sEnd-P/M) or both ends (dEnd-P/M) mapped to a single parent, (ii) reads are inter-haplotype, with ends mapped to both parents (d/sEnd-I), and (iii) reads with neither end mapped to a specific parent (dEnd-U) . An example of a paired-end read (dEnd-U) with no SNVs/InDels has its origin imputed using nearby reads (right, Supplemental Methods A.6 in Lindsly *et al.* [77]). A ratio of paternally and maternally mapped reads is found in a dynamically sized flanking region around the haplotype-unknown read's location (Supplemental Methods A.4 in Lindsly *et al.* [77]). This ratio then determines the likelihood of the haplotype-unknown read's origin. **(B)** Whole-genome Hi-C of GM12878 cells (top left). Inter- and intra-haplotype chromatin contacts after phasing Hi-C data using HaploHiC (right). Chromosomes 14 and 15 highlight inter- and intra-chromosome contacts within and between genomes (bottom left). Visualized in log_2 scale 1 Mb resolution in G1. **(C)** Haplotype phasing illustrates that the inactive maternal X Chromosome is partitioned into large heterochromatic domains, outlined in dotted black boxes. Visualized in log_2 scale 100 kb resolution in G1. This figure was taken from Lindsly *et al.* [77].

We next applied spectral clustering recursively to the Hi-C data at 100 kb resolution to determine whether there were differences in TADs between the two genomes throughout the cell cycle [25]. While the current understanding of genomic structure dictates that TAD boundaries are invariant (between alleles, cell types, etc), it is also known that “intra-TAD” structures are highly variable [43, 44, 53]. The spectral identification method has an increased ability to discern these subtle structural changes. We found that TAD boundaries were variable between the maternal and paternal genomes and across cell cycle phases in all chromosomes. This supports previous findings of allelic differences in TADs for single cells, and we predict that they are even more variable across cell types [25, 53]. Differences in TAD boundaries were observed surrounding MAE genes, ABE genes, and genes with coordinated biallelic expression (Figure S6 in Lindsly *et al.* [77]). This indicated that changes in TAD boundaries were not directly related to allelic expression differences.

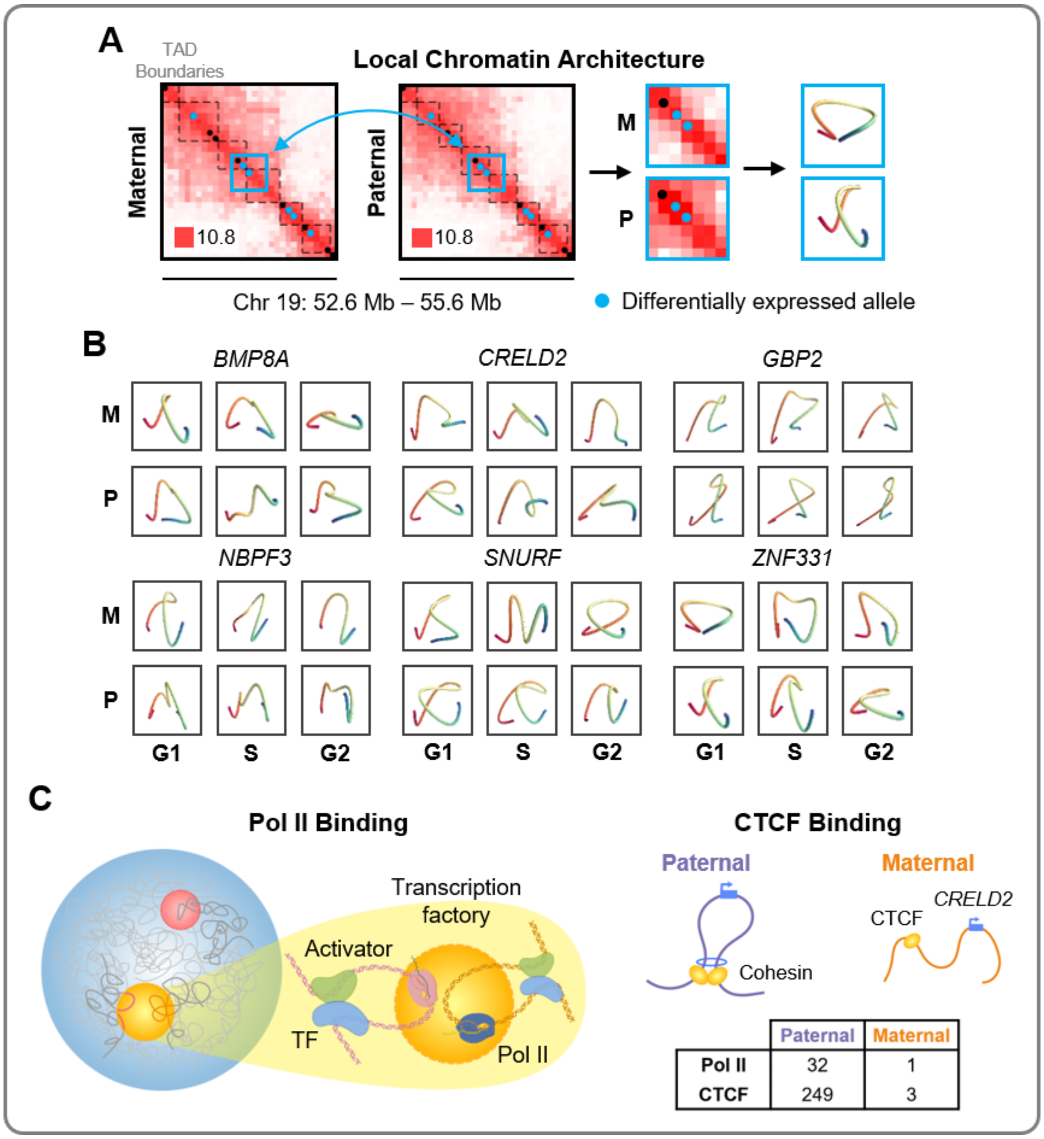
Although we did not find a direct relationship between TAD boundary differences and ABE genes, we observed during this analysis that the local genome structure around the six imprinted genes had noticeable differences. We then sought to analyze all genes with ABE or CBE to find out if they had corresponding structural differences at a local level. We analyzed the local Hi-C matrices for each of the 615 RNA-seq and 505 Bru-seq differentially expressed genes. Using a 300 kb flanking region centered at the 100 kb bin containing the transcription start site, we isolated a 7x7 matrix (700 kb) for each differentially expressed gene (Figure 3.5A, B). These matrices represent the local genomic structure of the differentially expressed genes, and are slightly smaller than average TAD size ($\sim$1 Mb). We then compared the correlation matrices of the log_2-transformed local Hi-C data and determined whether or not the matrices have statistically significant differences ($p < 0.05$) [67, 76]. We applied this comparison to all genes that were differentially expressed in RNA-seq (Bru-seq) and found that 515 (403) genes had at least one comparison in which both the expression and local structure had significantly changed. While changes in local genome structure and changes in gene expression do not have a one-to-one relationship, we found that both ABE and CBE genes are more likely to have significant architectural differences than randomly sampled allele-specific genes ($p < 0.01$) (Figure S7 and Supplemental Methods D in Lindsly *et al.* [77]). This lends further support to the idea that there is a relationship between allele-specific differences in gene expression and genome structure.

3.4.5 Allele-Specific Protein Binding

To uncover the mechanisms behind the relationship between allele-specific gene expression and genome structure, we looked to DNA binding proteins such as RNA polymerase II (Pol II), CCCTC-binding factor (CTCF), and 35 other transcription factors. We used publicly available protein binding data from AlleleDB in tandem with RNA-seq and found 114 genes that have an allelic bias in both gene expression and binding of at least one such protein [26]. We identified 13 genes which have ABE and biased binding of Pol II, with bias agreement in 11 cases (85%). That is, the allelic expression and Pol II binding were biased toward the same allele. For CTCF, 33 of 72 cases have bias agreement (46%), and for all other transcription factors analyzed, 20 of 29 cases have bias agreement (69%) (Table S6 in Lindsly *et al.* [77]). The CTCF binding bias agreement of around 50% is expected, based on previous studies [112]. This is likely due to CTCF's role as an insulator, since an allele could be expressed or suppressed by CTCF's presence depending on the context. To avoid potential inconsistencies between our data and the protein data from AlleleDB, we excluded Chromosome X when testing for ABE and protein binding biases.

We evaluated the relationship between TAD boundary differences between the maternal and paternal genomes and allele-specific CTCF binding sites. We found multiple instances of biased binding of CTCF and corresponding changes to the boundaries of TADs containing ABE genes. Examples of this phenomenon are shown in the center of Figure S6A in Lindsly *et al.* [77], where TAD boundaries from the maternal (paternal) Hi-C data are closer to a maternally (paternally) biased CTCF binding site in some cell cycle phases near the ABE genes *ANKRD19P*, *C9orf89*, and *FAM120A*. Despite observing individual instances of biased CTCF binding corresponding to TAD boundary differences and ABE genes, there were insufficient data to evaluate this relationship genome-wide. We hypothesize that differences in TAD boundaries would correspond to allele-biased CTCF binding provided that there were enough data, as it has been repeatedly shown that TAD boundaries are enriched with CTCF binding [44, 145].

We analyzed the 11 genes with allelic expression and Pol II binding bias agreement further to determine if they also had significant changes in local genome structure. Through local Hi-C comparisons, we found that all 11 of these genes had significant changes in structure in at least one cell cycle phase. 3D models for six of these genes are shown in Figure 3.5B, which highlight differences in local genome structure (Supplemental Methods F in Lindsly *et al.* [77]) [135]. The genes with bias agreement and changes in local genome structure include known imprinted genes such as *SNURF* and *SNRPN*, as well as genes with known allele-specific expression (and suggested imprinting in other cell types) like *ZNF331* [11]. Additionally, there are multiple genes with known associations with diseases or disorders such as *BMP8A*, *CRELD2*, and *NBPF3* [64, 103, 143]. These findings suggest that changes in local structure often coincide with changes in expression due to the increased or decreased ability of a gene to access the necessary transcriptional machinery within transcription factories [31, 99]. We visualize this relationship for the gene *CRELD2* as an example (Figure 3.5C).



	Paternal	Maternal
Pol II	32	1
CTCF	249	3

Figure 3.5: Local chromatin structure and transcription factor binding. **(A)** Local regions around differentially expressed genes are tested for significant conformation changes. These regions are modeled to visualize the conformations around each allele through G1, S, and G2 (Supplemental Methods E in Lindsly *et al.* [77]). Example of local chromatin structure extraction is shown for *ZNF331* in G1 phase (center of blue box). Hi-C matrices are shown in log_2 scale 100 kb resolution. **(B)** 3D models of the local genome structure around 6 ABE genes with bias agreement of Pol II and significant changes in local genome structure. **(C)** Schematic representation of allele-specific Pol II and CTCF binding, with highlighted gene *CRELD2*, which had binding biases in both. Table shows extreme binding biases of Pol II and CTCF on *CRELD2* as an example. This figure was taken from Lindsly *et al.* [77].

3.4.6 The Maternal and Paternal 4D Nucleome

We define the maternal and paternal 4DN as the integration of allele-specific genome structure with gene expression data through time, adapted from Chen *et al.* [23]. Many complex dynamical systems are investigated using a network perspective, which offers a simplified representation of a system [96, 125]. Networks capture patterns of interactions between their components and how those interactions change over time [23]. We can consider genome structure as a network, since Hi-C data captures interactions between genomic loci [92, 106]. In network science, the relative importance of a node in a network is commonly determined using network centrality [96]. For Hi-C data, we consider genomic loci as nodes and use network centrality to measure the importance of each locus at each cell cycle phase [76, 79]. We initially performed a global analysis of the maternal and paternal 4DN by combining RNA-seq with multiple network centrality measures (Supplemental Methods D in Lindsly *et al.* [77]) [76]. We found differences between the maternal and paternal genomes and across cell cycle phases, but only Chromosome X had clear maternal and paternal separation (Figures S9, S10 in Lindsly *et al.* [77]).

In our earlier analysis, we found a significant relationship between ABE and changes in local genome structure. We also observed that genes in multiple critical biological modules had coordinated biallelic expression. Motivated by these results, we performed an integrated analysis of structure and function to determine allele-specific dynamics of targeted gene sets. We constructed a sub-network for each gene set (analogous to an in silico 5C matrix), by extracting rows and columns of the Hi-C matrix containing genes of interest for each cell cycle phase [49]. We used eigenvector centrality (similar to Google's PageRank) to quantify structure, and used the average expression from the three RNA-seq replicates to quantify function, for each allele in the sub-network [100]. We utilized the concept of a phase plane to plot the maternal and paternal 4DN (4DN phase plane, adapted from Chen *et al.*) (Figure 3.6) [23]. We designated one axis as a measure of structure and the other as a measure of function. Coordinates of each point in the 4DN phase plane were determined from normalized structure data (x-axis, sub-network eigenvector centrality) and function data (y-axis, FPKM). The 4DN phase plane contains three points for each allele, which represent G1, S, and G2. We define allelic divergence (AD) as the average Euclidean distance between the maternal and paternal alleles across all cell cycle phases in the 4DN phase plane (Supplemental Methods D in Lindsly *et al.* [77]).

We show four example 4DN phase planes of gene sub-networks with various ADs in Figure 3.6. Genes which are known to be crucial for cell cycle regulation have a mean AD of 0.0245 (Figure 3.6B, middle-left). Given that GM12878 is a B-lymphocyte cell line, we were interested in the AD of genes which are related to B cell receptor functionality. We found that these genes had a mean AD of 0.0225 (Figure 3.6B, left). The ADs of cell cycle regulating genes and B cell specific genes are smaller than the mean AD of randomly selected allele-specific genes (AD = 0.0301 over 10,000 samples). This may be indicative of a robust coordination between the alleles to maintain proper cellular function and progression through the cell cycle, and therefore a lack of ABE genes or large structural differences. We show a random set of allele-specific genes with a mean AD of 0.0249 as an example (Figure 3.6B, middle-right). MAE genes had a mean AD of 0.1748 (Figure 3.6B, right), significantly higher than randomly selected allele-specific genes ($p < 0.01$, Supplemental Methods E in Lindsly *et al.* [77]). This approach is useful for quantifying differences between maternal and paternal genomes throughout the cell cycle, highlighting gene sets with large structural or expression differences over time. In previous work, we have also shown that this method may be broadly applicable to time series analysis of different cell types [76].

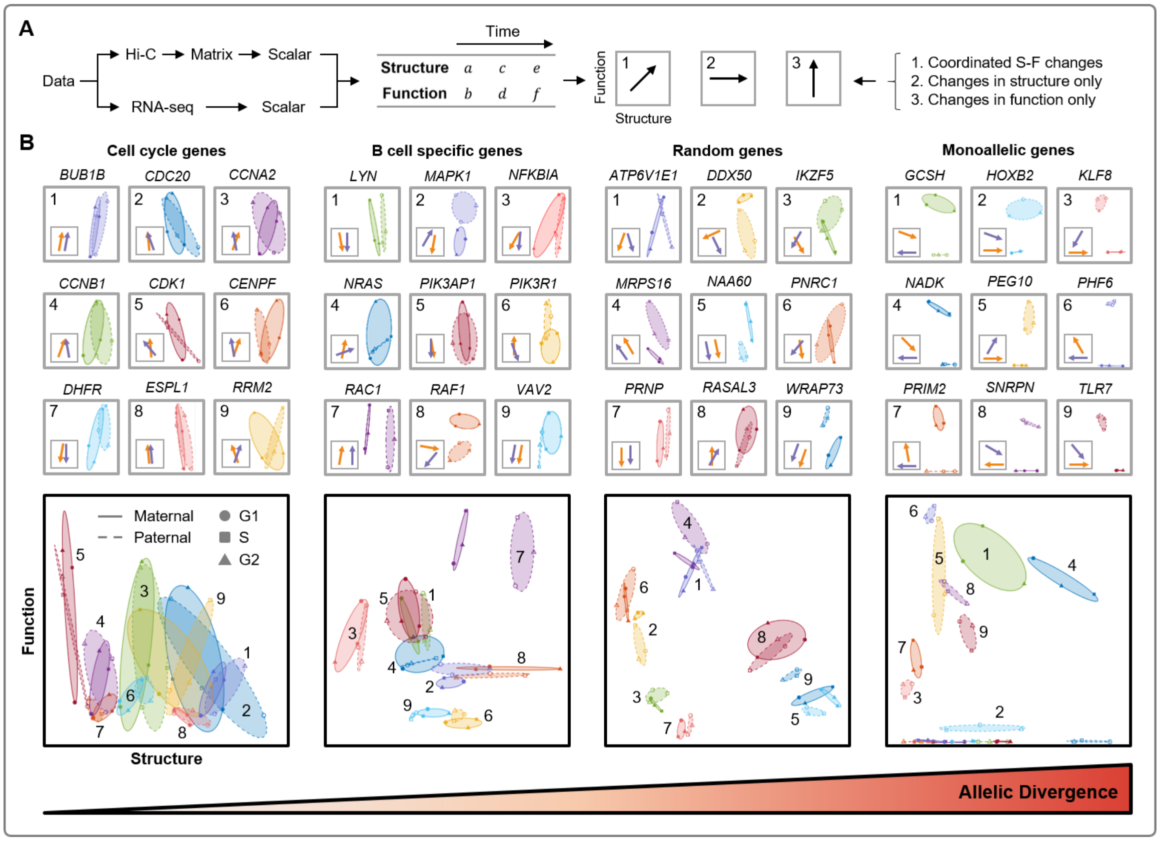


Figure 3.6: 4DN phase planes reveal a wide range of allelic divergences in gene sub-networks. **(A)** Workflow to obtain structure and function measures. Simplified phase planes are shown with linear relationships between changes in structure and function, changes in structure with no changes in function, and changes in function with no changes in structure. **(B)** 4DN phase planes of cell cycle genes, genes specific to B cell function, random allele-specific genes, and MAE genes, highlighting the similarities and differences between their alleles. Genes such as *BUB1B* and *PIK3AP1* have similar phase planes between alleles, while *RAC1* differs in structure and *WRAP73* differs in function. The bottom plot for each column combines the phase planes of the nine example genes. This figure was taken from Lindsly *et al.* [77].

3.5 Discussion

In this study, we present evidence for the intimate relationship among allele-specific gene expression, genome structure, and protein binding across the cell cycle. We validated our data and methods using known allele-specific properties such as the monoallelic expression of imprinted and X-linked genes, broad similarities of chromatin compartments between the maternal and paternal genomes, and large heterochromatic domains of Chromosome X [6, 8, 107, 108, 113]. Unique to this study, we established a coordination of allele-biased expression and changes in local genome structure, which included hundreds of genes not commonly associated with allele-biased expression. We observed further evidence of this coordination through corresponding protein binding biases.

Through our analysis of mature (nascent) RNA, we found 467 (380) genes to be differentially expressed between the two alleles and 229 (164) genes with differential expression through the cell cycle. Approximately half of the genes with allele-biased expression are only differentially expressed in certain cell cycle phases, and over half of the genes with CBE are only differentially expressed in one allele. Further research is needed to explore why certain genes have coordinated cell cycle dynamics across both alleles, while other genes have disparate expression in some cell cycle phases. We predict that these transient allelic biases may be associated with developmental pathologies and tumorigenesis, similar to imprinted and other MAE genes. Conversely, we found no allele-biased expression from genes in multiple biological modules, such as the cell cycle and glycolysis pathways (Table 3.1). We were not able to establish a statistical significance here due to the limited number of allele-specific genes in these modules, so we surveyed a set of 662 essential genes and found that they are significantly less likely to have allele-biased expression [16]. This supports our hypothesis of highly coordinated biallelic expression in universally essential genes.

We developed a novel phasing algorithm, HaploHiC, which uses Hi-C reads mapped to phased SNVs/InDels to predict nearby reads of unknown parental origin. This allowed us to decrease the sparsity of our allele-specific contact matrices and increase confidence in our analysis of the parental differences in genome structure. While we found that the overall compartmentalization (euchromatin and heterochromatin) of the two genomes was broadly similar, there were many differences in TAD boundaries and local genome structure between the two genomes and between cell cycle phases. We focused our search for allele-specific differences in genome structure by calculating the similarity of local contacts surrounding differentially expressed genes [67, 76]. We found that differentially expressed genes were significantly more likely to have corresponding changes in local genome structure than random allele-specific genes.

We incorporated publicly available allele-specific protein binding data for Pol II and CTCF to explore the mechanisms behind the gene expression and local genome structure relationship [112]. In genes that had both allele-biased expression and Pol II binding biases, we found that 85% of these genes had allelic bias agreement. Additionally, we found that all of the genes with expression and Pol II binding bias agreement had significant changes in local genome structure. Analysis of the relationships among allele-specific gene expression, genome structure, and protein binding is currently hindered by the amount of information available and our limited understanding of the dynamics of cell-specific genome structure and gene expression variability [52]. The ability to separate maternal and paternal gene expression and protein binding is dependent on the presence of a SNV/InDel within the gene body and nearby protein binding motifs. As SNVs/InDels are relatively rare in the human genome, the number of genes available to study is severely limited. Once we are able to separate the maternal and paternal genomes through advances in experimental techniques, we will be able to fully study these relationships.

Overall, these data support an intimate allele-specific relationship between genome structure and function, coupled through allele-specific protein binding. Changes in genome structure, influenced by the binding of proteins such as CTCF, can affect the ability of transcription factors and transcription machinery to access DNA. This results in changes in the rate of transcription of RNA, captured by Bru-seq. The rate of transcription leads to differential steady state gene expression, captured by RNA-seq. Integration of these data into a comprehensive computational framework led to the development of a maternal and paternal 4DN, which can be visualized using 4DN phase planes and quantified using allelic divergence. Allele-specific analysis across the cell cycle will be imperative to discern the underlying mechanisms behind many diseases by uncovering potential associations between deleterious mutations and allelic bias, and may have broad translational impact spanning cancer cell biology, complex disorders of growth and development, and precision medicine.

CHAPTER IV

Deciphering Multi-way Interactions in the Human Genome

This chapter is based on ongoing work by Stephen Lindsly, Can Chen, Sam Dilworth, Sivakumar Jeyarajan, Walter Meixner, Anthony Cicalo, Nicholas Beckloff, Charles Ryan, Gilbert S. Omenn, Amit Surana, Lindsey Muir, and Indika Rajapakse.

4.1 Abstract

The organization of the genome is non-random and has a high degree of order. The current standard for experimentally capturing the genome's organization is through genome wide chromosome conformation capture (Hi-C). A recently developed sequencing technology called Pore-C contains the information of Hi-C, but also includes multi-way interactions which cannot be directly derived from Hi-C. Hi-C is often used to observe structural features through the aggregation of pair-wise contacts genome-wide, but these features cannot be captured directly. Multi-way contacts from Pore-C can be used to unambiguously observe higher order structural features, where instances of nearby multiple genomic loci are captured together as single reads. In this work, we use Pore-C data in the form of hypergraphs to develop methods for quantifying entropy of genome structure and to compare the genomes of different cell

types. In addition, we integrate Pore-C data with multiple data modalities to find biologically important multi-way interactions.

4.2 Introduction

The intricate folding of the genome allows for approximately two meters of DNA to fit within a cell nucleus while remaining accessible for transcription. The folding patterns of the genome, or genome structure, is a rapidly advancing study. With the advent of experimental techniques based on chromosome conformation capture (3C), we have uncovered an immense amount of knowledge about how the genome is organized and how it affects genome function. Many of the advancements in this field have focused on expanding the amount of interactions between genomic loci that can be captured. Originally, 3C could only capture an interaction between two genomic loci. This was later extended to interactions from one locus to all others (4C), many loci's interactions with many others (5C), and eventually all loci to all loci (Hi-C) [39, 49, 74, 118]. While extraordinarily useful in their own right, all of these technologies are only able to capture interactions between pairs of loci.

Recent advances in sequencing technologies has brought forth the ability to capture multiple loci at once genome-wide. Pore-C reads contain fragments from multiple interacting loci at once, allowing for new methods of analysis on genome structure. We extract the multi-way contacts from Pore-C reads to construct hypergraphs. Hypergraphs are an extension of graphs, where hyperedges can contain any number of nodes while an edge can only contain two nodes. We consider genomic loci as nodes in our hypergraph, and multi-way contacts as hyperedges. We use incidence matrices to represent hypergraphs, where rows in the incidence matrices represent genomic loci and columns contain individual hyperedges. From this representation, we are able to make quantitative measurements of the genome's organization through hypergraph entropy, compare different cell types through hypergraph distance, and identify functionally important multi-way contacts in multiple cell types.

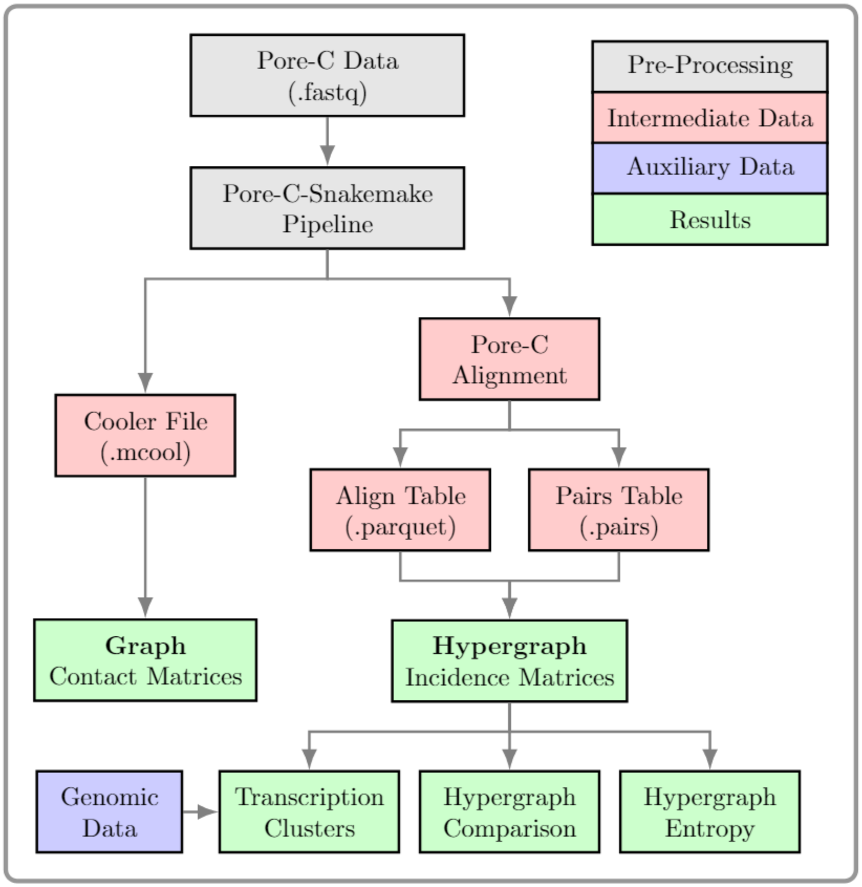


Figure 4.1: Flowchart overview of Pore-C analysis.

4.3 Results

To obtain Pore-C reads, DNA is cross-linked to histones, digested by a restriction enzyme, ligated together, and then sequenced (Figure 4.2A). Once these sequences are aligned to the genome, we can determine the locations where each fragment originated and construct a multi-way contact (Figure 4.2B). We use hypergraphs to represent multi-way contacts, where individual hyperedges contain at least two loci (Figure 4.2C, left). Hypergraphs provide a simple and concise way to depict multi-way contacts, and allow for abstract representations of genome structure. Computationally, we represent multi-way contacts as incidence matrices (Figure 4.2C, right). Using more standard experimental techniques, such as Hi-C, adjacency matrices are often used to capture the pair-wise genomic contacts. Multi-way contacts are not able to be represented in this manner, since the rows and columns of adjacency matrices only account for individual loci. In contrast, incidence matrices allow us to include more than two loci per contact and provide a clear visualization of multi-way contacts. Multi-way contacts can also be decomposed into pair-wise contacts, similar to Hi-C, by extracting all combinations of loci (Figure 4.2D).

4.3.1 Decomposing Multi-way Contacts

We conducted Pore-C experiments using adult human dermal fibroblasts and obtained additional publicly available Pore-C data from B lymphocytes [132].) From these data, we constructed hypergraphs as multiple resolutions (read level, 100 kb, 1 Mb, and 25 Mb). We first analyzed individual chromosomes at 100 kb resolution, and decomposed the multi-way contacts into their pair-wise counterparts to identify topologically associated domains (TADs, *Materials and Methods*). Examples of TADs from Chromosome 22 for fibroblasts and B lymphocytes can be seen in Figures 4.3 and 4.4, respectively.

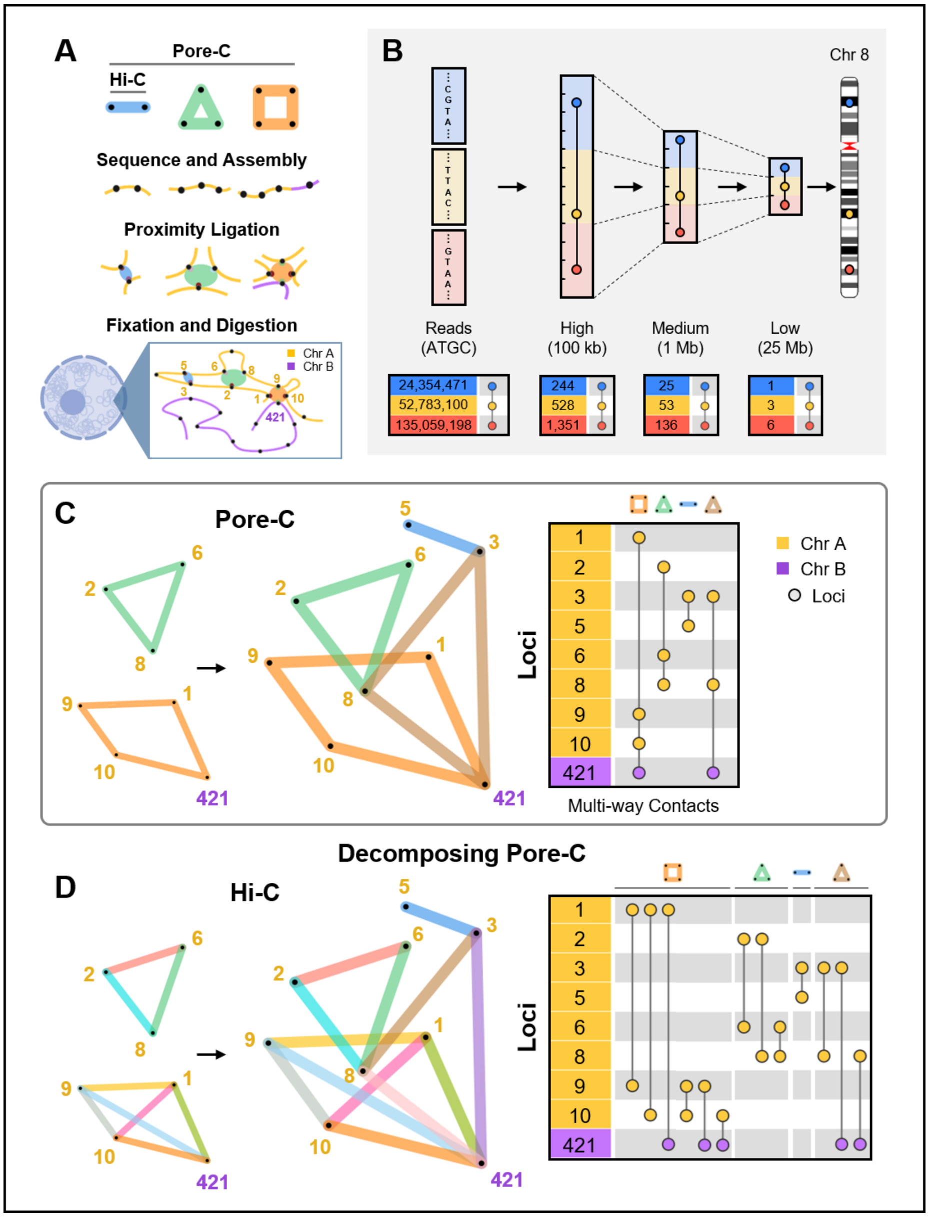


Figure 4.2: Pore-C experimental and data workflow. (A) The Pore-C experimental protocol, which captures pairwise and multi-way contacts (*Materials and Methods*). (B) Representation of multi-way contacts at different resolutions (top, not to scale). Incidence matrix visualizations of a representative example from Chromosome 8 in human fibroblasts at each resolution (bottom). The numbers in the left columns represent the location of each genomic locus present in a multi-way contact, where values are either the chromosome base-pair position (read-level) or the bin into which the locus was placed (binning at 100 kb, 1 Mb, or 25 Mb). (C) Hypergraph representation of Pore-C contacts (left) and an incidence matrix (right) of four sets of multi-way contacts within (yellow-to-yellow) and between (yellow-to-purple) chromosomes. Contacts correspond to examples from part A. The numbers in the left column represent a bin in which a locus resides. Each vertical line represents a multi-way contact, with nodes at participating genomic loci. (D) Multi-way contacts can be decomposed into pairwise contacts. Decomposed multi-way contacts can be represented using graphs (left) or incidence matrices (right). Contacts correspond to examples from parts A and C.

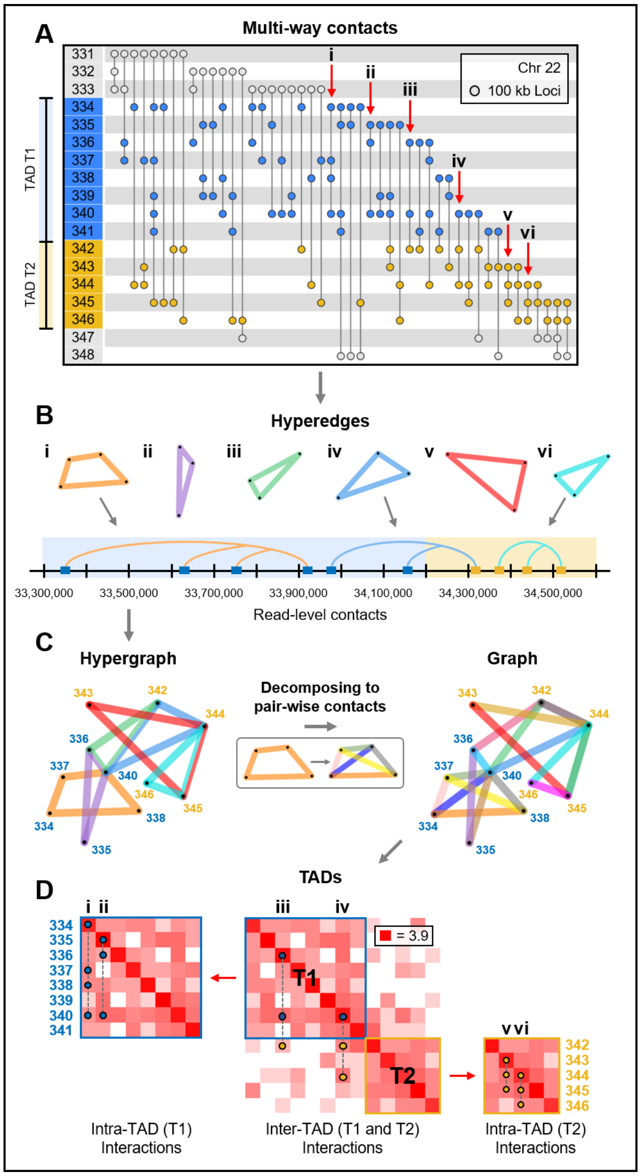


Figure 4.3: Local organization of the genome. (A) Incidence matrix visualization of a region in Chromosome 22 from fibroblasts (V1-V4). The numbers in the left column represent genomic loci in 100 kb resolution, vertical lines represent multi-way contacts, where nodes indicate the corresponding locus' participation in this contact. The blue and yellow regions represent two TADs, T1 and T2. The six contacts, denoted by the labels i-vi, are used as examples to show intra- and inter-TAD contacts in B, C, and D. (B) Hyperedge and read-level visualizations of the multi-way contacts i-vi from A. Blue and yellow rectangles (bottom) indicate which TAD each loci corresponds to. (C) A hypergraph is constructed using the hyperedges from B (multiway contacts i-vi from A). The hypergraph is decomposed into its pair-wise contacts in order to be represented as a graph. (D) Contact frequency matrices were constructed by separating all multi-way contacts within this region of Chromosome 22 into their pairwise combinations. TADs were computed from the pair-wise contacts using the methods from [25]. Example multi-way contacts i-vi are superimposed onto the contact frequency matrices. Multi-way contacts in this figure were determined in 100 kb resolution after noise reduction, originally derived from read-level multi-way contacts (*Materials and Methods*).

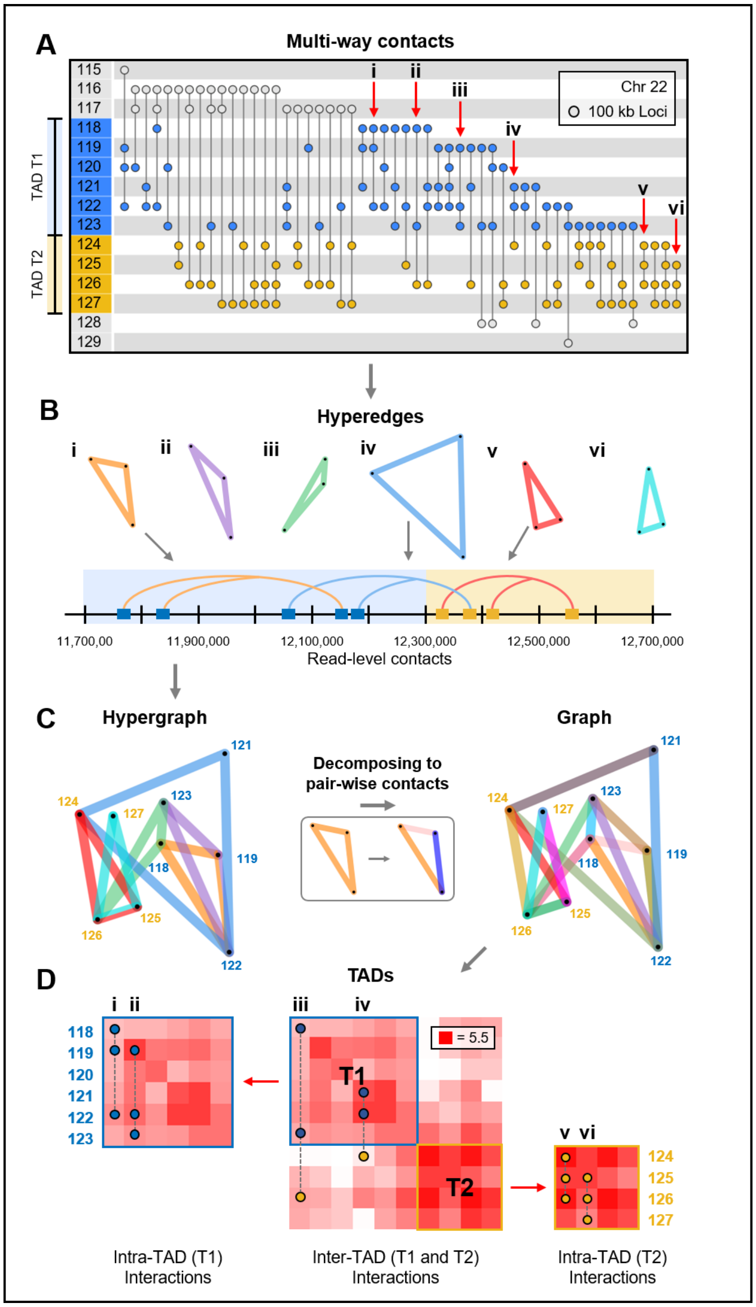


Figure 4.4: Local organization of the genome. (A) Incidence matrix visualization of a region in Chromosome 22 from B lymphocytes. The numbers in the left column represent genomic loci, vertical lines represent multi-way contacts, where nodes indicate the corresponding locus' participation in this contact. The blue and yellow regions represent two TADs, T1 and T2. The six contacts, denoted by the labels i-vi, are used as examples for hypergraph and genomic folding pattern visualizations. (B) Hypergraph visualization of the multi-way contacts i-vi from A. Blue and yellow labels indicate which TADs these loci participate in. (C) Contact frequency matrices were constructed by separating all multi-way contacts within and between the two TADs into their pairwise combinations. Example multi-way contacts are superimposed onto contact frequency matrices. All multi-way contacts in this figure were determined in 100 kb resolution after noise reduction (*Materials and Methods*).

4.3.2 Representing Chromosomes as Hypergraphs

To gain a better understanding of genome structure with multi-way contacts, we constructed hypergraphs for entire chromosomes in 1 Mb resolution. We show the incidence matrix of Chromosome 22 as an example in Figure 4.5A. In Figure 4.5B, we show Chromosomes 22's distribution of 1 Mb contacts at multiple orders (2-way contacts, 3-way contacts, etc). Figure 4.5C highlights the most common intra-chromosome contacts within Chromosome 22 using multi-way contact "motifs", which we use as a simplified way to show hyperedges. Figure 4.5D shows a zoom-in of a 3-way contact to highlight how a low resolution multi-way contact can contain many contacts at higher resolutions. Figure 4.5E visualizes the multi-way contacts contained in Figure 4.5D as a hypergraph.

We also identified multi-way contacts that contain loci from multiple chromosomes. These inter-chromosomal multi-way contacts can be seen in 1 Mb resolution in Figure 4.5F and in 25 Mb resolution for both fibroblasts and B lymphocytes in Figure 4.6A and 4.6B, respectively. Figure 4.6 gives a summary of the entire genome's multi-way contacts, by showing the most common intra- and inter-chromosomal multi-way contacts across all chromosomes. We highlight examples of multi-way contacts with loci that are contained within a single chromosome ("intra only"), spread across unique chromosomes ("inter only"), and a mix of both within and between chromosomes ("intra and inter"). Finally, we found the most common inter-chromosomal multi-way contacts across all chromosomes, which we summarize with five example chromosomes in Figure 4.7 using multi-way contact motifs.

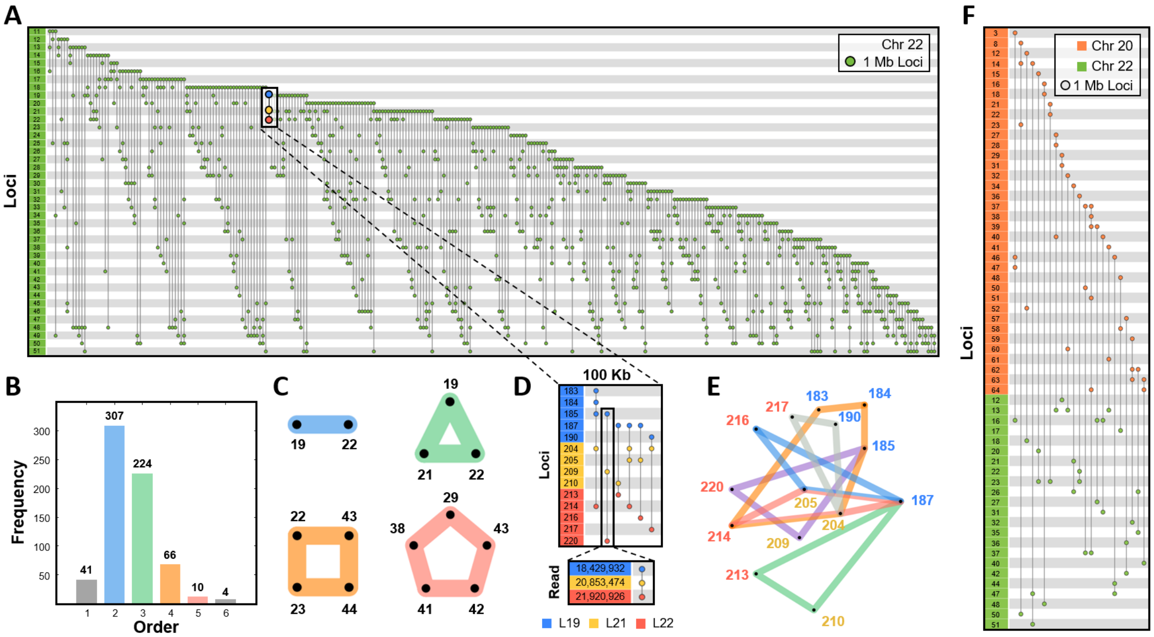


Figure 4.5: Patterning of intra- and inter-chromosomal contacts. (A) Incidence matrix visualization of Chromosome 22 in fibroblasts. The numbers in the left column represent genomic loci in 1 Mb resolution. Each vertical line represents a multi-way contact, in which the nodes indicate the corresponding locus' participation in this contact. (B) Frequencies of Pore-C contacts in Chromosome 22. Bars are colored according to the order of contact. Blue, green, orange, and red correspond to 2-way, 3-way, 4-way, and 5-way contacts. (C) The most common 2-way, 3-way, 4-way, and 5-way intra-chromosome contacts within Chromosome 22 are represented as motifs, color-coded similarly to B. (D) Zoomed in incidence matrix visualization in 100 kb resolution shows the multi-way contacts between three 1 Mb loci L19 (blue), L21 (yellow), and L22 (red). An example 100 kb resolution multi-way contact is zoomed to read-level resolution. (E) Hypergraph representation of the 100 kb multi-way contacts from D. Blue, yellow, and red labels correspond to loci L19, L21, and L22, respectively. (F) Incidence matrix visualization of the inter-chromosomal multi-way contacts between Chromosome 20 (orange) and Chromosome 22 (green) in 1 Mb resolution. Within this figure, all data are derived from one fibroblast sequencing run (V2) and multi-way contacts were determined after noise reduction at 1 Mb or 100 kb resolution accordingly (*Materials and Methods*).

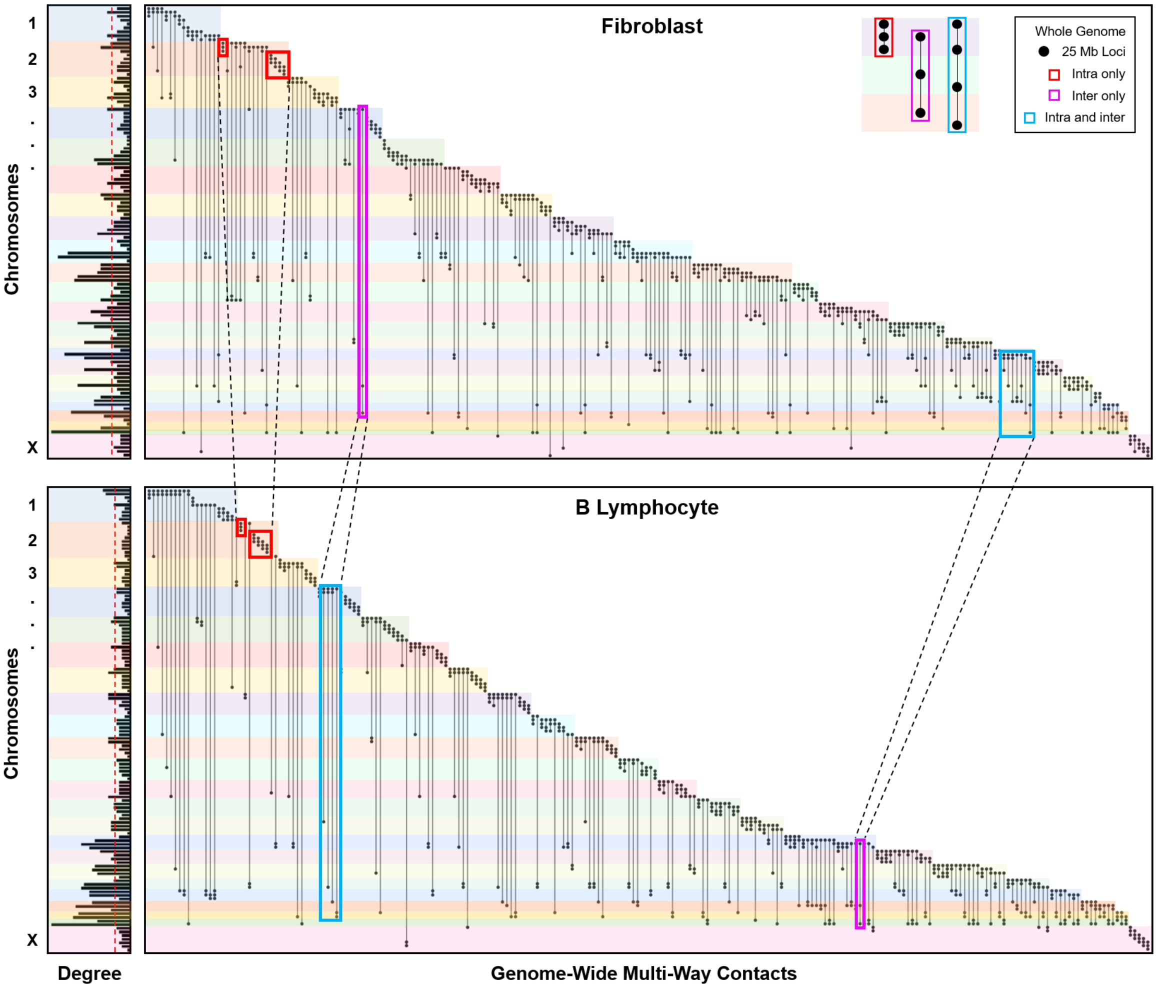


Figure 4.6: Genome-wide patterning of multi-way contacts. Incidence matrix visualization of the top 10 most common multi-way contacts per chromosome. Matrices are constructed at 25 Mb resolution for both fibroblasts (top, V1-V4) and B lymphocytes (bottom). Specifically, 5 intra-chromosomal and 5 inter-chromosomal multi-way contacts were identified for each chromosome with no repeated contacts. If 5 unique intra-chromosomal multi-way contacts are not possible in a chromosome, they are supplemented with additional inter-chromosomal contacts. Vertical lines represent multi-way contacts, nodes indicate the corresponding locus' participation in a multi-way contact, and color-coded rows delineate chromosomes. Highlighted boxes indicate example intra-chromosomal contacts (red), inter-chromosomal contacts (magenta), and combinations of intra- and inter-chromosomal contacts (blue). Examples for each type of contact are shown in the top right corner. Multi-way contacts of specific regions are compared between cell types by connecting highlighted boxes with black dashed lines, emphasizing similarities and differences between fibroblasts and B lymphocytes. Normalized degree of loci participating in the top 10 most common multi-way contacts for each chromosome in fibroblast and B-lymphocytes are shown on the left. Red dashed lines indicate the mean degree for fibroblasts and B lymphocytes (top and bottom, respectively). Genomic loci that do not participate in the top 10 most common multi-way contacts for fibroblasts or B lymphocytes were removed from their respective incidence and degree plots. Multi-way contacts were determined in 25 Mb resolution after noise reduction (*Materials and Methods*).

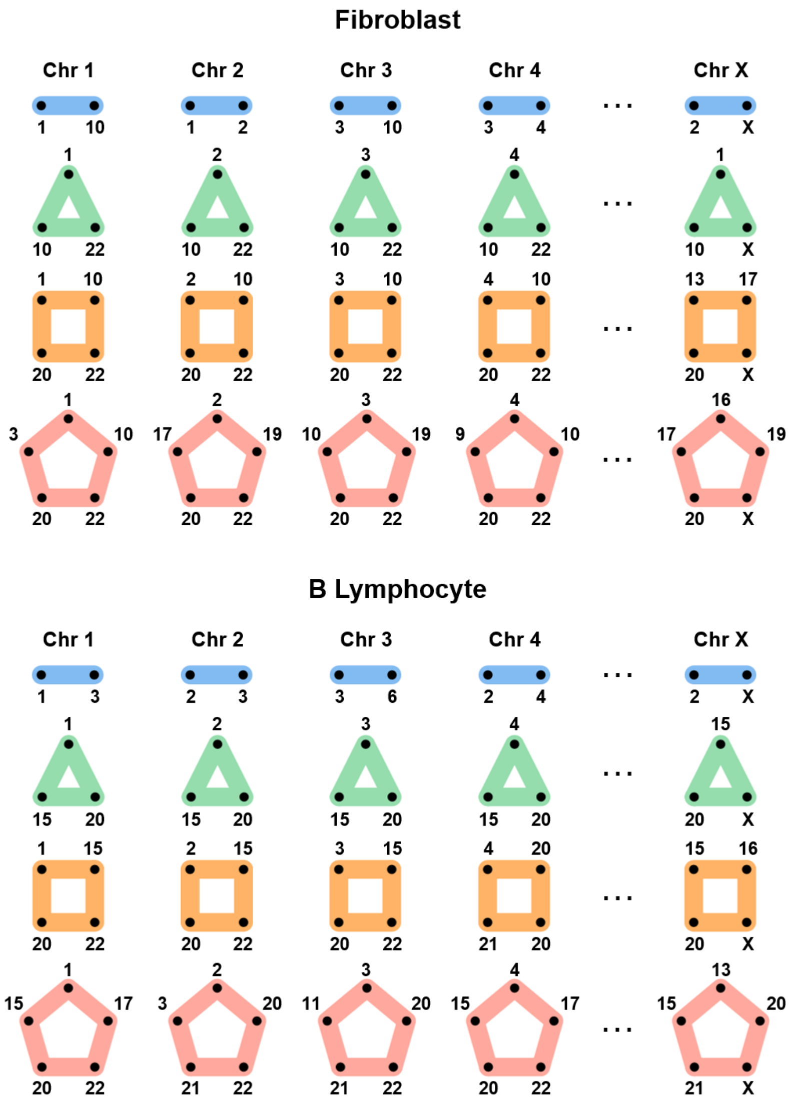


Figure 4.7: Inter-chromosomal interactions. The most common 2-way, 3-way, 4-way, and 5-way inter-chromosome combinations for each chromosome are represented using motifs from fibroblasts (top, V1-V4) and B lymphocytes (bottom). Rows represent the combinations of 2-way, 3-way, 4-way, and 5-way inter-chromosome interactions, and columns are the chromosomes. Inter-chromosomal combinations are determined using 25 Mb resolution multi-way contacts after noise reduction (*Materials and Methods*) and are normalized by chromosome length. Here we only consider unique chromosome instances (i.e. multiple loci in a single chromosome are ignored).

4.3.3 Transcription Clusters

Genes are transcribed in short sporadic bursts and transcription occurs in localized areas with high concentrations of transcriptional machinery [32, 33, 127]. This includes transcriptionally engaged polymerase and the accumulation of necessary proteins, called transcription factors. Multiple genomic loci can colocalize at these areas for more efficient transcription. In fact, it has been shown using fluorescence in situ hybridization (FISH) that genes frequently colocalize during transcription [99]. Simulations have also provided evidence that genomic loci which are bound by common transcription factors can self-assemble into clusters, forming structural patterns commonly observed in Hi-C data [33]. We refer to these instances of highly concentrated areas of transcription machinery and genomic loci as *transcription clusters*. The colocalization of multiple genomic loci in transcription clusters naturally leads to multi-way contacts, but these interactions can not be fully captured from the pair-wise contacts of Hi-C. Multi-way contacts derived from Pore-C reads can detect interactions between many genomic loci, and are well suited for identifying potential transcription clusters (Figure 4.8).

Using the initial criteria of chromatin accessibility and RNA Pol II binding, we identified 16,080 and 16,527 potential transcription clusters from fibroblasts and B lymphocytes, respectively (Table 4.1, *Materials and Methods*). The majority of these clusters involved at least one expressed gene (72.2% in fibroblasts, 90.5% in B lymphocytes) and many involved at least two expressed genes (31.2% in fibroblasts, 58.7% in B lymphocytes). While investigating the colocalization of expressed genes in transcription clusters, we found that over 30% of clusters containing multiple expressed genes had common transcription factors based on binding motifs (31.0% in fibroblasts, 33.1% in B lymphocytes) and that over half of these common transcription factors were master regulators (56.6% in fibroblasts, 74.7% in B lymphocytes). Two example transcription clusters derived from 3-way, 4-way, and 5-way contacts from both fibroblasts and B lymphocytes are shown in Figure 4.9. These example clusters contain at least two genes and have at least one common transcription factor.

We tested the criteria for potential transcription clusters for statistical significance (*Materials and Methods)*. That is, we tested whether the identified transcription clusters are more likely to include genes, and if these genes are more likely to share common transcription factors, than random multi-way contacts in both fibroblasts and B lymphocytes. We found that the transcription clusters were significantly more likely to include ≥ 1 gene and ≥ 2 genes than random multi-way contacts ($p < 0.01$). In addition, transcription clusters containing ≥ 2 genes were significantly more likely to have common transcription factors and common master regulators ($p < 0.01$). After testing all order multi-way transcription clusters, we also tested the 3-way, 4-way, 5-way, and 6-way (or more) cases individually. We found that all cases were statistically significant ($p < 0.01$) except for clusters for common transcription factors or master regulators in the 6-way (or more) case for both fibroblasts and B lymphocytes. We hypothesize that these cases were not statistically significant due to the fact that the large number of loci involved in these multi-way contacts will naturally lead to an increase of overlap with genes. This increases the likelihood that at least two genes will have common transcription factors or master regulators. Approximately half of transcription clusters with at least two genes with common transcription factors also contained at least one enhancer locus (~51% and ~44% in fibroblasts and B lymphocytes, respectively) [54]. This offers even further support that these multi-way contacts represented real transcription clusters.

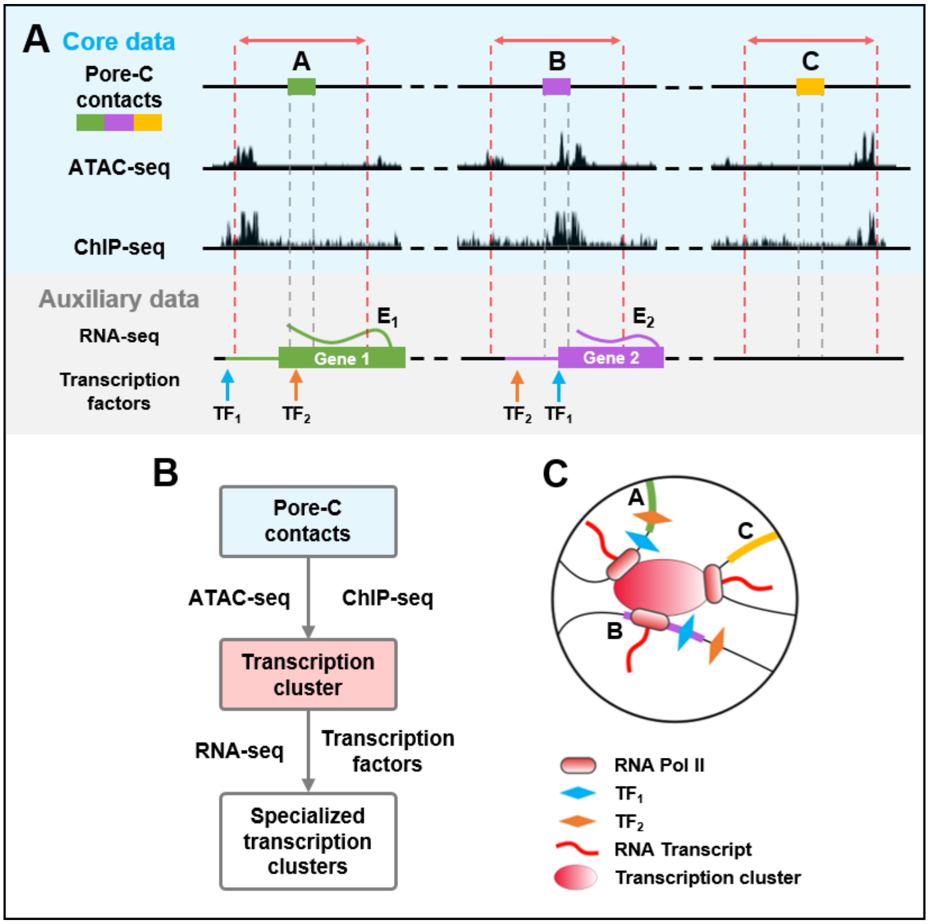


Figure 4.8: Data-driven identification of transcription clusters. (A) A 5 kb region before and after each locus in a Pore-C read (between red dashed lines) is queried for chromatin accessibility and RNA Pol II binding (ATAC-seq and ChIP-seq, respectively). Multi-way contacts between accessible loci that have ≥ 1 instance of Pol II binding are indicative of potential transcription clusters. Gene expression (RNA-seq, E1 for gene 1 and E2 for gene 2, respectively) and transcription factor binding sites (TF1 and TF2) are integrated to determine potential coexpression and coregulation within multi-way contacts with multiple genes. Transcription factor binding sites are queried ± 5 kb from the gene's transcription start site (Materials and Methods). (B) Simplified pipeline for extracting transcription clusters (Algorithm 3). Genes are colored based on the overlapping Pore-C locus, and the extended line from each gene represents the 5 kb flanking region used to query transcription factor binding sites. (C) Schematic representation of a transcription cluster.

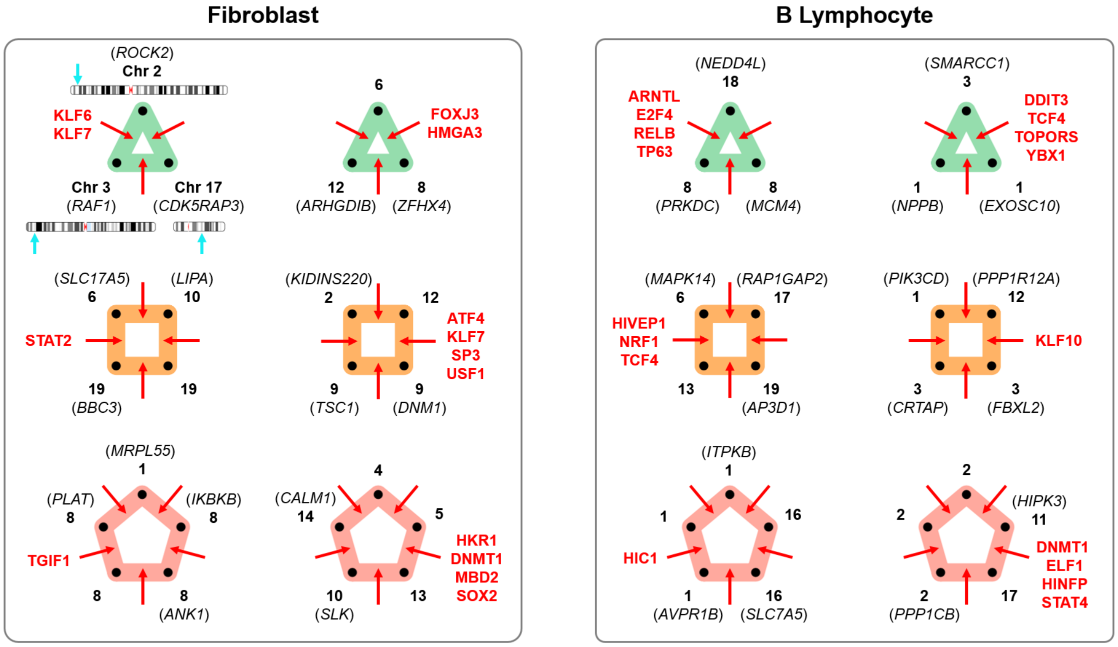


Figure 4.9: Example transcription clusters. Six examples of potential transcription clusters are shown for fibroblasts (left) and B lymphocytes (right) as multi-way contact motifs. Black labels indicate genes and chromosomes (bold). Red labels correspond to transcription factors shared between ≥ 2 genes within a transcription cluster. Blue arrows indicate a gene's position on it's respective chromosome. Multi-way contacts used for fibroblasts include all experiments (V1-V4). Examples were selected from the set of multi-way contacts summarized in the "Clusters with Common TFs" column of Table 4.1.

Order	Multi-way Contacts	Transcription Clusters	Clusters with $\geq$ 1 Gene	Clusters with $\geq$ 2 Genes	Clusters with Common TFs	Clusters with Common MRs
3	379,165	11,261	7,782	2,986	1,191	679
	240,477	8,384	7,384	4,157	2,006	1,536
4	181,554	3,254	2,519	1,214	276	153
	227,352	4,345	3,972	2,686	822	606
5	98,272	1,021	831	473	63	35
	196,423	1,996	1,881	1,434	277	193
6+	142,575	544	477	341	24	13
	1,000,231	1,802	1,727	1,419	109	67

Table 4.1: Summary of multi-way contacts. Multi-way contacts from fibroblasts (gray rows, V1-V4) and B lymphocytes (white rows) are listed after different filtering criteria. Multi-way contacts are considered to be potential transcription clusters if all loci within the multi-way contact are accessible and at least one locus has binding from RNA Pol II. These multi-way contacts are then queried for nearby expressed genes. If a transcription cluster candidate has at least two expressed genes, we determine whether the genes have common transcription factors (TFs) through binding motifs. From the set of transcription clusters with common transcription factors, we calculate how many clusters are regulated by at least one master regulator (MR).

4.4 Discussion

Through advancements in sequencing technology, we are now able to capture multi-way contacts within the genome. Multi-way contacts will become increasingly important within biological studies, as the relationship higher-order chromatin structures and genome function are intrinsically linked. An incredible amount of knowledge has already been gained within the last two decades using pair-wise contacts from prior chromosome conformation capture techniques. We expect that rate of new discoveries will only accelerate with the addition of unambiguous multi-way contacts.

4.5 Materials and Methods

Cell culture. Human adult dermal fibroblasts were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1X Glutamax (Thermo Fisher Scientific cat no. 35050061) and 1X non-essential amino acid (Thermo Fisher Scientific cat no. 11140050).

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

Restriction enzyme digest. The cell pellet was resuspended in 500 µL of cold permeabilization buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630, 100 µL of protease inhibitor cock-tail Roche cat no. 11836170001) and placed on ice for 15 min. One tablet of protease inhibitor cocktail was dissolved in 1 ml nuclease free water and 100 µL from that was added to a 500 µL permeabilization buffer. Cells were centrifuged at 500 x g for 10 min at 4°C after which the supernatant was aspirated and replaced with 200 µL of chilled 1.5X New England Biolabs (NEB) cutsmart buffer. Cells were centrifuged again at 500 x g for 10 min at 4°C, then aspirated and re-suspended in 300 µL of chilled 1.5X NEB cutsmart buffer. To denature the chromatin, 33.5 µL of 1% w/v sodium dodecyl sulfate (SDS, Invitrogen cat no. 15553-035) was added to the cell suspension and incubated for exactly 10 min at 65°C with gentle agitation then placed on ice immediately afterwards. To quench the SDS, 37.5 µL of 10% v/v Triton X-100 (Sigma Aldrich cat no. T8787-250) was added for a final concentration of 1%, followed by incubation for 10 min on ice. Permeabilized cells were then digested with a final concentration of 1 U/µL of NlaIII (NEB-R0125L) and brought to volume with nuclease-free water to achieve a final 1X digestion reaction buffer in 450 µL. Cells were then mixed by gentle inversion. Cell suspensions were incubated in a thermomixer at 37°C for 18 hours with periodic rotation.

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

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

Nanopore sequencing. Purified DNA was Solid Phase Reversible Immobilization (SPRI) size selected before library preparation with a bead ratio of 0.48X for fragments $>$ 1.5 kb. The $>$ 1.5 kb products were prepared for sequencing using the protocol provided by Oxford Nanopore Technologies. In brief, 1 µg of genomic DNA input was used to generate asequencing library according to the protocol provided for the SQK-LSK109 kit. (Oxford Nanopore Technologies, Oxford Science Park, UK). After the DNA repair, end prep, and adapter ligation steps, SPRI select bead suspension (Cat No. B23318, Beckman Coulter Life Sciences, Indianapolis, IN, USA) was used to remove short fragments and free adapters. A bead ratio of 1X was used for DNA repair and end prep while a bead ratio of 0.4X was used for the adapter ligation step. Qubit dsDNA assay (ThermoFisher Scientific, Waltham, MA, USA) was used to quantify DNA and ~300-400 ng of DNA library was loaded onto a GridION flow cell (version R9, Flo-MIN 106D). In total, 4 sequencing runs were conducted generating a total of 6.25 million reads (referred to as V1-V4).

Sequence processing. Reads which passed Q-score filtering (`--min_qscore 7`, 4.56 million reads) from basecalling on the Oxford Nanopore GridION were used as input for the Pore-C-Snakemake pipeline (`https://github.com/nanoporetech/Pore-C-Snakemake`, commit 6b2f762). The pipeline maps multi-way contacts to a reference genome and stores the hyperedges data in a variety of formats. The reference genome used for mapping was GRCh38.p13 (`https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/`). The pairs and parquet files output from the pipeline were converted into MATLAB tables to construct hyperedges and the cooler files were used to create the pair-wise adjacency matrices (Figures 4.3 and Figure 4.4). The individual tables from the four sequencing runs were assigned a sequencing run label and then concatenated. The combined tables were used as standard inputs for all downstream software processes.

Noise Reduction. After obtaining the pairwise read table, we first bin the chromosome contact positions (in MATLAB, we use the command `ceil(chr_pos/resolution)`). Next, we create an array containing unique combinations of read IDs and bin numbers participating in pairwise contacts. This removes experimental artifacts such as spuriously repeated reads. We then count the number of reads containing each unique pairwise contact between bins. Only those contacts between bins occurring in at least ϵ reads were further considered. In other words, if the count of a unique pairwise contact is less than the threshold ϵ, we consider it as noise and remove all instances of that pairwise contact from the pairwise read table. The threshold ϵ was determined by finding the 85th percentile of pairwise contact counts. By considering the frequency of pairwise contacts rather than the frequency of hyperedges, we can maintain hyperedges with mostly the same constituents (few differences in loci) while still eliminating noise. This also maintains the overlapping pairwise contacts from different read IDs. The 85th percentile threshold was determined empirically, and can be adjusted by making simple modifications to the provided MATLAB code.

Hypergraphs and Incidence Matrices. The Pore-C technique has inspired us to consider the human genome as a hypergraph, a generalization of a graph in which its hyperedges can join any number of nodes [14]. Hypergraphs can capture higher-order connectivity patterns and represent multidimensional relationships unambiguously [13, 142]. Therefore, modeling the human genome via hypergraphs can offer new insights into its organization. In this genomic hypergraph, nodes are genomic loci, where a locus can be a gene or a genomic region at a particular resolution (i.e. read level, 100 kb, 1 Mb, or 25 Mb bins). Hyperedges are the relationships or interactions among multiple genomic loci. Pore-C determines the hyperedges in the genomic hypergraph through the multi-way contacts among genomic loci. In much of our analysis, we consider unweighted hypergraphs (i.e. ignore the frequency of contacts) since most of higher-order contacts are unique in Pore-C data at high resolution. For low resolution (i.e. 1 Mb and 25 Mb), we consider the weights of edges to find the most common intra- and inter-chromosomal contacts.

An incidence matrix of the genomic hypergraph is an $n \times m$ matrix, where n is the total number of genomic loci, and m is the total number of unique Pore-C contacts (including self-contacts, pairwise contacts, and higher-order contacts). For each column of the incidence matrix, if the genomic locus i involved with the corresponding Pore-C contact, the ith element of the column is equal to one. If not, it is equal to zero. Thus, the nonzero elements of a column tell us their associated genomic loci that are close in space, and the number of nonzero elements gives rise to the order of the Pore-C contact. The incidence matrix of the genomic hypergraph can be nicely visualized via PAOHvis [133]. PAOHvis is an online hypergraph visualization software, in which genomic loci are parallel horizontal bars and Pore-C contacts are vertical lines that connects multiple loci (see Figures 4.2, 4.3, 4.5, and 4.6). Furthermore, incidence matrices play an significant role in the mathematical analysis of hypergraphs.

Algorithm 2 Hypergraph incidence construction

```
Input: Aligned Pore-C data
for each multi-way contact $j$ do
    if multi-way contact contains locus $i$ then
        $\mathbf{H}(i, j) = 1$
    else
        $\mathbf{H}(i, j) = 0$
    end if
end for
Return: Hypergraph incidence matrix $\mathbf{H} \in \mathbb{R}^{n \times m}$ where $n$ is the total number of loci, and $m$
    is the total number of multi-way contacts.
```

Data-driven identification of transcription clusters. We use Pore-C data in conjunction with multiple other data sources to identify potential transcription clusters (Figure 4.8). Each locus in a Pore-C read, or multi-way contact, is queried for chromatin accessibility and RNA Pol II binding (ATAC-seq and ChIP-seq peaks, respectively). Multi-way contacts are considered to be potential transcription clusters if all loci within the multi-way contact are accessible and at least one locus has binding from RNA Pol II. These multi-way contacts are then queried for nearby expressed genes. A 5 kb flanking region is added before and after each locus when querying for chromatin accessibility, RNA Pol II binding, and nearby genes [95]. Gene expression (RNA-seq) and transcription factor binding sites are integrated to determine coexpression and coregulation of genes in multi-way contacts. If a transcription cluster candidate has at least two genes present, we determine whether the genes have common transcription factors through binding motifs. From the set of transcription clusters with common transcription factors, we calculate how many clusters are regulated by at least one master regulator, a transcription factor that also regulates its own gene (Figure 4.8).

Algorithm 3 Identification of Transcription Clusters

```
Input: Hypergraph incidence matrix $\mathbf{H}$, gene expression $\mathbf{R}$ (RNA-seq), RNA Pol II $\mathbf{P}$ (ChIP-seq), chromatin accessibility $\mathbf{C}$ (ATAC-seq), transcription factor binding motifs $\mathbf{B}$
for each multi-way contact $j$ in $\mathbf{H}$ do
    if all loci are accessible from $\mathbf{C}$ and $\geq 1$ locus has Pol II binding from $\mathbf{P}$ then
        multi-way contact $j$ from $\mathbf{H}$ is added to the set of potential transcription clusters $\mathbf{T}_p$
    end if
end for
for each potential transcription cluster $k$ in $\mathbf{T}_p$ do
    if loci contain $\geq 2$ expressed genes from $\mathbf{R}$ which have $\geq 1$ common TFs from $\mathbf{B}$ then
        multi-way contact $k$ from $\mathbf{T}_p$ is added to the set of transcription clusters $\mathbf{T}_c$
    end if
    if loci contain $\geq 2$ expressed genes from $\mathbf{R}$ which have $\geq 1$ common MRs from $\mathbf{B}$ then
        multi-way contact $k$ from $\mathbf{T}_p$ is added to the set of transcription clusters $\mathbf{T}_s$
    end if
end for
Return: Potential transcription clusters $\mathbf{T}_p$, transcription clusters $\mathbf{T}_c$, and specialized transcription clusters $\mathbf{T}_s$
```

Transcription factor binding motifs: Transcription factor binding site motifs were obtained from "The Human Transcription Factors" database [70]. FIMO (https://meme-suite.org/meme/tools/fimo) was used to scan for motifs within ± 5kb of the transcription start sites for protein-coding and microRNA genes. The results were converted to a $22,083 \times 1,007$ MATLAB table, where rows are genes, columns are transcription factors, and entries are the number of binding sites for a particular transcription factor and gene. The table was then filtered to only include entries with three or more binding sites in downstream computations. This threshold was determined empirically and can be adjusted by simple changes to the MATLAB code.

Public data sources. Pore-C data for B lymphocytes were downloaded from Ulahannan *et al.* [132]. ATAC-seq and ChIP-seq data were obtained from ENCODE to assess chromatin accessibility and RNA Pol II binding, respectively. These data were compared to read-level Pore-C contacts to determine whether colocalizing loci belong to accessible regions of chromatin and had RNA Pol II binding for both fibroblasts and B lymphocytes. RNA-seq data were also obtained from ENCODE to ensure that genes within potential transcription factories were expressed in their respective cell types. A summary of these data sources can be found in Table 4.2.

Data Type	**Cell Line**	**Data Description and Source**
Pore-C	Adult Human Dermal Fibroblasts	Fibroblast Pore-C multi-way contacts generated in this study
Pore-C	GM12878	B-lymphocyte Pore-C multi-way contacts obtained from Ulahannan *et al.* [132]
ATAC-seq	IMR-90	Fibroblast chromatin accessibility (ENCFF310UDS)
ATAC-seq	GM12878	B-lymphocyte chromatin accessibility data (ENCFF410XEP)
ChIP-seq	IMR-90	Fibroblast RNA Polymerase II binding data (ENCFF676DGR)
ChIP-seq	GM12878	B-lymphocyte RNA Polymerase II binding data (ENCFF912DZY)
RNA-seq	IMR-90	Fibroblast gene expression data averaged over two replicates (ENCFF353SBP, ENCFF496RIW)
RNA-seq	GM12878	B-lymphocyte gene expression data averaged over two replicates (ENCFF306TLL, ENCFF418FIT)

Table 4.2: Data sources. Data obtained from ENCODE unless otherwise specified [29].

Hypergraph Entropy. Network entropy often is used to measure the connectivity and regularity of a network [22, 90, 101]. We use hypergraph entropy to quantify the organization of chromatin structure from Pore-C data, where higher entropy corresponds to less organized folding patterns (e.g. every genomic locus is highly connected). There are different definitions of hypergraph entropy [15, 22, 60]. In our analysis, we exploit the eigenvalues of the hypergraph Laplacian matrix and fit them into the Shannon entropy formula [15]. In mathematics, eigenvalues can quantitatively represent different features of a matrix [124]. Denote the incidence matrix of the genomic hypergraph by $\mathbf{H}$. The Laplacian matrix then is an n-by-n matrix (n is the total number of genomic loci in the hypergraph), which can be computed by $\mathbf{L} = \mathbf{H}\mathbf{H}^\top \in \mathbb{R}^{n\times n}$, where $\top$ denotes matrix transpose. Therefore, the hypergraph entropy is defined by

$$\textbf{Hypergraph Entropy} = -\sum_{i=1}^{n} \lambda_i \ln \lambda_i, \tag{4.1}$$

where λ_i are the normalized eigenvalues of $\mathbf{L}$ such that $\sum_{i=1}^{n} \lambda_i = 1$, and the convention $0 \ln 0 = 0$ is used. Biologically, genomic regions with high entropy are likely associated with high proportions of euchromatin, as euchromatin is more structurally permissive than heterochromatin [81, 105].

Algorithm 4 Hypergraph entropy [15]

```
1: Input: Hypergraph incidence matrix $\mathbf{H} \in \mathbb{R}^{n\times m}$
2: Construct the hypergraph Laplacian matrix $\mathbf{L} = \mathbf{H}\mathbf{H}^\top \in \mathbb{R}^{n\times n}$
3: Compute the eigenvalues $\lambda_i$ of $\mathbf{L}$ using eigendecomposition
4: Normalize the eigenvalues $\bar{\lambda}_j = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}$
5: Compute the hypergraph entropy
       $$S = -\sum_j \bar{\lambda}_j \ln \bar{\lambda}_j$$
6: Return: Hypergraph entropy $S$.
```

Hypergraph Distance. Comparing graphs is a ubiquitous task in data analysis and machine learning [48]. There is a rich body of literature for graph distance with examples such as Hamming distance, Jaccard distance, and other spectral-based distances [48, 51, 141]. Here we propose a spectral-based hypergraph distance measure which can be used to quantify global difference between two genomic hypergraphs G_1 and G_2 from two cell lines. Denote the incidence matrices of two genomic hypergraphs by $\mathbf{H}_1 \in \mathbb{R}^{n\times m_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{n\times m_2}$, respectively. For $i = 1, 2$, construct the normalized Lapalacian matrices:

$$\tilde{\mathbf{L}}_i = \mathbf{I} - \mathbf{D}_i^{-\frac{1}{2}}\mathbf{H}_i\mathbf{E}_i^{-1}\mathbf{H}_i^\top\mathbf{D}_i^{-\frac{1}{2}} \in \mathbb{R}^{n\times n}, \tag{4.2}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix, $\mathbf{E}_i \in \mathbb{R}^{m_i\times m_i}$ is a diagonal matrix containing the orders of hyperedges along its diagonal, and $\mathbf{D}_i \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the degrees of nodes along its diagonal [148]. The degree of a node is equal to the number of hyperedges that contain that node. Therefore, the hypergraph distance between G_1 and G_2 is defined by

$$\textbf{Hypergraph Distance}(\mathsf{G}_1, \mathsf{G}_2) = \frac{1}{n}(\sum_{i=1}^n |\lambda_{1j} - \lambda_{2j}|^p)^{\frac{1}{p}}, \tag{4.3}$$

where λ_{ij} is the jth eigenvalue of $\tilde{\mathbf{L}}_i$ for $i = 1, 2$, and $p \geq 1$. In our analysis, we choose $p = 2$. The hypergraph distance (4.3) can be used to compare two genomic hypergraphs in a global scale since the eigenvalues of the normalized Laplacian are able to capture global connectivity patterns within the hypergraph.

Algorithm 5 Comparing Hypergraphs

```
1: Input: Two hypergraph incidence matrices $\mathbf{H}_1 \in \mathbb{R}^{n\times m_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{n\times m_2}$
2: Construct the normalized hypergraph Laplacian matrices
       $$\tilde{\mathbf{L}}_i = \mathbf{I} - \mathbf{D}_i^{-\frac{1}{2}}\mathbf{H}_i\mathbf{E}_i^{-1}\mathbf{H}_i^\top\mathbf{D}_i^{-\frac{1}{2}} \in \mathbb{R}^{n\times n},$$
   where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix, $\mathbf{E}_i \in \mathbb{R}^{m_i\times m_i}$ is a diagonal matrix containing the orders of hyperedges along its diagonal, and $\mathbf{D}_i \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the degrees of nodes along its diagonal, for $i = 1, 2$ [148].
3: Compute the hypergraph distance
       $$d = \frac{1}{n}\left(\sum_{i=1}^{n} |\lambda_{1j} - \lambda_{2j}|^p\right)^{\frac{1}{p}},$$
   where $\lambda_{ij}$ is the $j$th eigenvalue of $\tilde{\mathbf{L}}_i$ for $i = 1, 2$, and $p \geq 1$.
4: Return: Hypergraph distance $d$ between $\mathbf{H}_1$ and $\mathbf{H}_2$.
```

Statistical Significance via Permutation Test. In order to assess the statistical significance of the transcription cluster candidates we determined using our criteria (Figure 4.8), we use a permutation test which builds the shape of the null hypothesis (i.e. the random background distribution) by resampling the observed data over N trials. We randomly select n 3rd, 4th, and 5th order multi-way contacts from our Pore-C data, where n is based on the number of transcription cluster candidates we determined for each order using our criteria. For example, we randomly selected $n = 11,261$ multi-way contacts from the set of 3rd order multi-way contacts in fibroblasts (Table 4.1). For each trial, we determine how many of these randomly sampled "transcription clusters" match our remaining criteria: transcription clusters with ≥ 1 gene, ≥ 2 genes, common TFs, and common MRs. The background distribution for each of the criteria can then be constructed from these values. The proportion of values in these background distributions that are greater than their counterparts from the data-derived transcription cluster candidates yields the p-value. For this analysis, we chose $N = 1,000$ trials. This analysis is based on the assumption that transcription clusters will be more likely to contain genes and that those genes are more likely to have common transcription factors than arbitrary multi-way contacts.

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

CHAPTER V

Understanding Memory B Cell Selection

This chapter is based on a paper by Stephen Lindsly, Maya Gupta, Cooper Stansbury, and Indika Rajapakse [75] (under review).

5.1 Abstract

The mammalian adaptive immune system has evolved over millions of years to become an incredibly effective defense against foreign antigens. The adaptive immune system's humoral response creates plasma B cells and memory B cells, each with their own immunological objectives. The affinity maturation process is widely viewed as a heuristic to solve the global optimization problem of finding B cells with high affinity to the antigen. However, memory B cells appear to be purposely selected earlier in the affinity maturation process and have lower affinity. We propose that this memory B cell selection process may be a heuristic solution to two expected risk optimization problems: optimizing for affinity to similar antigens in the future despite mutations or other minor differences, and optimizing to warm start the generation of plasma B cells in the future. We use simulations to provide evidence for our hypotheses, taking into account data showing that certain B cell mutations are more likely than others. Our findings are consistent with memory B cells having high-affinity to mutated antigens, but do not provide strong evidence that memory B cells will be more useful than

selected naive B cells for seeding the secondary germinal centers.

5.2 Introduction

The immune system is an effective threat mitigation system that deploys a number of learned identification algorithms. While the *innate immune system* is adept at identifying foreign invaders, or antigens, it must engage the *adaptive immune system* to create a more massive and specific response. A core aspect of the adaptive immune system's humoral response is training two types of B cells through a process called affinity maturation (AM): plasma B cells which generate antibodies to identify the current antigen, and memory B cells which are used in subsequent immune responses to identify similar antigens in the future. The AM process is highly unusual, in that a specific region of DNA within participating B cells is mutated to generate offspring which are selected to have higher affinity to the antigen in question. The preservation of DNA sequences is usually of utmost importance in most cells, but the region of the genome which defines the shape of the B cell receptor must be rapidly modified for the B cell receptor to have a chance of becoming better at recognizing the antigen of interest [89]. These mutations are responsible for the B cells' incredible ability to recognize practically any antigen that they are presented, making the mammalian adaptive immune system one of the most effective learned identification systems in the natural world. In this paper, we consider whether the plasma B cell and memory B cell generation processes can be interpreted as trying to satisfy specific objectives, and if so, can we state these objectives precisely?

We borrow standard ideas from machine learning, where it is common to first specify an ideal mathematical objective to be optimized (such as minimizing the expected error rate of a learned identification system), then propose heuristic algorithms that approximately optimize that mathematical objective. Similarly, we hypothesize that the AM processes act like heuristics that approximately optimize for idealized immunological objectives due to evolutionary pressures. We hypothesize what those evolutionarily-adaptive immunological objectives might be, then compare how well different B cells satisfy these objectives when faced with adversarially-mutated antigens via simulations. These findings lead us to propose new hypotheses about the implicit objectives of the immune system's training of naive B cells to become memory B cells.

First in Section 5.3, we review how naive B cells are recruited and trained to become plasma B cells, and present a hypothesis for the objectives of this training mathematically. In Section 5.4, we consider the mathematical optimization objectives of the more enigmatic training process that leads to the generation of memory B cells. We define *plasma B cell training* or *training plasma B cells* as the process of generating plasma B cells from naive B cells during AM. Similarly, we define *memory B cell training* or *training memory B cells* as the process of generating memory B cells from naive B cells during AM. We test our hypotheses via simulations in Section 5.5, and conclude with a discussion of open questions in Section 5.6.

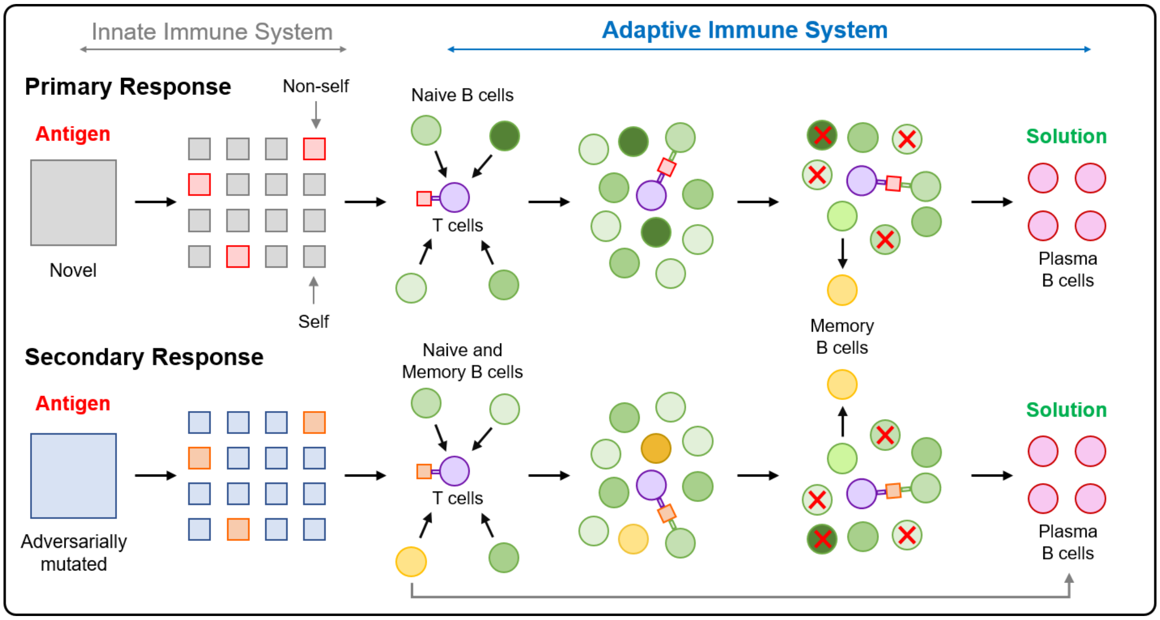


Figure 5.1: High-level illustration of the adaptive immune system. First, an antigen enters the body, then the innate immune system identifies pieces of the antigen as non-self. (Top:) The adaptive immune system's response to the identified antigen. A diverse set of random naive B cells that have some initial affinity to the antigen flock together and form a germinal center [89, 129]. These B cells proliferate and mutate when selected by T_{FH} cells for their affinity to the antigen. B cells with moderate affinity are stored for later use as memory B cells. High affinity B cells differentiate into plasma B cells, which are the solution to a particular antigen. (Bottom:) A mutated version of a previously encountered antigen, or an antigen from a related pathogen, is presented to the adaptive immune system. It responds by forming germinal centers with both random naive B cells (with some initial affinity to the antigen) and memory B cells from the first encounter. Memory B cells can also be used directly by differentiating into plasma B cells. This figure was taken from Lindsly *et al.* [75].

5.3 Plasma B cell training

We review the AM process that trains naive B cells to become plasma B cells, then consider what mathematical criteria the plasma B cell training may have evolved to optimize.

5.3.1 Affinity Maturation Of Plasma B Cells

AM begins by recruiting naive B cells with some initial affinity to the antigen to secondary lymphoid organs. These naive B cells, along with T follicular helper (T_{FH})

cells and follicular dendritic cells, concentrate into temporary structures known as germinal centers (Fig. 5.1A) [89, 129]. Germinal centers (GCs) ensure that these cells are in close proximity, thus facilitating rapid mutation and evaluation of B cells receptor sequences. During AM, B cells are evaluated by $\mathrm{T_{FH}}$ cells for their affinity to the antigen through the length of interaction between them, based on antigen presentation by B cells [87]. If the initial affinity of a B cell is high, it receives a chemical signal from the $\mathrm{T_{FH}}$ cell to move to a separate area of the GC and proliferate. While the B cell is proliferating, a specific section of the genome called the hypervariable region is exposed to an enzyme, activation-induced cytidine deaminase (AID) [7, 93]. AID is able to deaminate cytosine creating uracil, a nucleotide that is not normally found in DNA. The operation that repairs these changes is error-prone, leading to mutations in the DNA sequence [83].

The process of deamination and mutations during repair is referred to as *somatic hypermutation* (SHM). After proliferating, the B cells return to the area of the GC containing $\mathrm{T_{FH}}$ cells and are reevaluated for their affinity towards the antigen. This iterative process of proliferation, mutation, and affinity evaluation continues until the B cells have a sufficiently high affinity to the antigen. At this point, the B cells differentiate into plasma B cells and begin to produce antibodies which allow for the immune system to eradicate the antigen.

5.3.2 Affinity Maturation As An Algorithm

We model AM in Algorithm 6, which we use to simulate AM in our experiments. We simplify a few known or uncertain issues about AM, detailed in Subsection 5.3.3.

Each *naive* B cell receptor sequence is generated randomly by a combinatorial mix of its V, D, J, and C gene segments, as well as through junctional diversity between these segments [120]. Naive B cells span at least 100 million possibilities [120]. The naive B cells recruited to a germinal center are cells that already have some promising affinity s to the antigen a. SHM then mutates nucleotides in the DNA sequence region roughly 1,500 base pairs (bp) long that defines the B cell receptor structure. Mutations in the hypervariable region are on the scale of 10^6 times more likely than mutations outside of this region [83]. Mutations can be swaps, insertions, and deletions in a categorical space modeled as $\{A, T, C, G, \emptyset\}^{1500}$, where $\emptyset$ connotes a deletion. It should be noted that, although the combinatorial space of possible mutations is large, many specific mutations immediately lead to apoptosis. AM optimization is parallelized and distributed over G germinal centers, which algorithmically can be thought of as G different parallel processors. We model the different germinal centers as working independently, though there may be biochemical signaling between them. In practice, an organism trains for multiple independent antigens simultaneously, but for simplicity, we consider one antigen at a time.

5.3.3 Known Simplifications Of Algorithm 6

We note that Algorithm 6 simplifies a few known characteristics of AM. We believe these simplifications are minor and that they do not affect the major conclusions of this work.

We model the algorithm as T discrete iterations, but in practice, AM is continuous process that is partly time-limited because of antigen decay and external pressures. However, to our knowledge, there is not an exact limit to the number of divisions that can occur or an exact timeline that must be met during AM. Furthermore, there is a chance that the immune system does not find a solution fast enough, causing the host to die. While we recognize that this occurs in the natural system, it is not the focus of our simulation. Therefore, we use a fixed number of iterations as an approximation of the time limits the real immune system faces. In addition, we identify the highest affinity B cells at the end of our simulation as plasma B cells for simplicity, but plasma B cells are not selected simultaneously at the end of AM.

Before SHM begins, the initial B cell population may have undergone undirected proliferation, which means the initial random sample may be better modeled as random clusters of B cells. Algorithm 6 allows the germinal centers to grow without bounds, though the die-off rate d will tend to limit the population size in the germinal centers. In practice, the size of germinal centers is also bounded by physical volume constraints and biochemical resource constraints. We use a constant die-off rate d, but there is some evidence die-off probability decreases as affinity increases [5]. We include a simplified version of this behavior by biasing the death probabilities of B cells in the germinal center towards lower affinity cells.

$\mathrm{T_{FH}}$ cells measure *nearby* B cells for their affinity, so there is only some probability that a specific B cell will have its affinity measured, and that probability a B cell's affinity gets measured is thought to be a function of spatial organization (which is indirectly affected by affinity) and direct affinity. The affinity and spatial proximity of a particular B cell influences its likelihood to be selected and induced to proliferate. We do not consider the spatial organization between $\mathrm{T_{FH}}$ cells and B cells explicitly. As affinity of a B cell increases, so does the likelihood that the $\mathrm{T_{FH}}$ cell will send a proliferation signal. Similar to the increased likelihood of death for a low affinity B cell, we bias the selection of B cells for proliferation towards higher affinity B cells. In addition, the strength of the proliferation signal is generally proportional to the affinity of the B cell, such that a high affinity B cell is more likely to proliferate many times before returning for another iteration of affinity evaluation [37, 87].

Algorithm 6 may oversimplify AM in other ways as well that we are not aware of, or that are not yet known.

Algorithm 6 Affinity Maturation Algorithm for Training Plasma and Memory B Cells

```
Require: an antigen $a \in \mathcal{A}$
Require: an affinity score $s(b, a) \rightarrow \mathbf{R}$ for B cell $b \in \mathcal{B}$ and antigen $a \in \mathcal{A}$
Require: a low affinity threshold $\epsilon$ to enter a germinal center
Require: a high affinity threshold $\tau >> \epsilon$ to become a plasma B cell
Require: a die-off rate $d \in [0, 1]$ for germinal center B cells
Require: probability $p$ that a B cell is measured by a $\mathrm{T_{FH}}$ cell
Require: probability function $q(t, s)$ of producing a memory B cell on iteration $t$ given affinity score
    $s$ that is monotonically increasing in $s$, and might be monotonically decreasing or unimodal in $t$
Require: probability function $r(s)$ of a proliferation signal from a $\mathrm{T_{FH}}$ cell after affinity measurement
Require: a Bernoulli random number generator $Bernoulli(p)$ that outputs 1 with probability $p$
    and 0 otherwise
initialize the set of plasma B cells $B^* = \emptyset$ and the set of memory B cells $V^* = \emptyset$
for $g = 1, \ldots, G$ germinal centers do
    sample an initial set of $J_g$ naive B cells $B_g^0$ such that $s(b, a) > \epsilon$ for all $b \in B_g^0$
    for $t = 1, \ldots, T$ iterations do
        for $b \in B_g^t$ do
            if Bernoulli(p) == 1 then
                the B cell $b$ is observed by some nearby $\mathrm{T_{FH}}$ cell which measures $s(b, a)$
                if $s(b, a) \geq \tau$ then
                    insert $b$ into the set of plasma B cells $B^*$ such that $b \in B^*$
                    break
                end if
                if Bernoulli($q(t, s(b, a))$) == 1 then
                    insert $b$ into the set of memory B cells $V^*$ such that $b \in V^*$
                    break
                end if
                if Bernoulli($r(s(b, a))$) == 1 then
                    proliferate: B cell $b$ sent to divide and mutate some number of times, and its
                    mutated copies are added to the set $B_g^{t+1}$
                    break
                end if
                if Bernoulli(d) then
                    $b$ dies
                    break
                end if
            end if
            $b$ is added to the set $B_g^{t+1}$
        end for
    end for
end for
return the set of plasma cells $B^*$ and the set of memory B cells $V^*$
```

5.3.4 Plasma B Cells Are Created To Optimize Antigen Affinity Given Limited Time

AM is a process that has long been framed as the immune system acting as a global optimization algorithm trying to find a B cell that best identifies a given antigen through SHM [131]. That is, AM acts as if it were a heuristic to solve,

$$\arg\max_{b \in \mathcal{B}} s(b, a) \tag{5.1}$$

where a is a given antigen, b is a B cell from the set $\mathcal{B}$ of all possible B cells, and s is the affinity function that models the quality of the lock-and-key physical and biochemical interaction of b and a. Note that $\mathcal{B}$ is a very large categorical space defined by the variable-length DNA sequence that encodes the B cell receptor.

However, the objective (5.1) does not recognize the fact that a plasma B cell does not need to be a perfect match to the antigen. In fact, there appears to be a sufficient affinity τ such that once an affinity of τ is reached, the B cell is induced to differentiate into a plasma B cell. Further, the immune system is under time pressure to produce such sufficiently high-affinity plasma B cells as fast as possible.

Therefore, we propose that a more realistic model of what the plasma B cell generation process is optimizing should also depend on the sufficient affinity $\tau > 0$, and the given set $\mathcal{B}_0$ of initial naive B cells in the germinal center. To capture the time pressure, we model probabilistic mutations to B cells in the germinal center at each discrete time iteration. Given a naive B cell $b \in \mathcal{B}_0$, let $M(b) \in \mathcal{B}$ be a new random B cell produced by a single random mutation of b. Let $M^K(b) = M(M(\ldots(M(b))\ldots))$ denote the random B cell generated after K random mutations of b, so that the random B cell $M^K(b)$ can be any of the B cells reachable by K mutations M of the initial B cell b, with the corresponding probabilities dependent on the sum of the likelihood of the different mutations paths that could produce $M^K(b)$ starting from *b*.

Then we hypothesize the plasma B cell selection process is a heuristic evolved to minimize the number of mutations $K \in \mathcal{N}$ needed so that on average K random mutations will produce at least one B cell in the germinal center with sufficient affinity τ to the antigen:

$$\min K \text{ such that } \left(E_{M^K} \left[\max_{b \in \mathcal{B}_0} s(M^K(b), A) \right] \right) \geq \tau, \tag{5.2}$$

where $E[\cdot]$ is the standard expectation operator (average) with respect to the random variable's possible outcomes weighted by their probabilities. This objective is consistent with our Algorithm 6.

Clearly, the immune system is not a sentient entity that explicitly formulates the criteria (5.2) and subsequently identifies a heuristic to optimize it. Rather, our hypothesis is that evolutionary pressures have selected for a plasma B cell generation process that better optimizes (5.2).

5.4 Memory B Cell Training

Similar to plasma B cells, memory B cells are created within the germinal center, but there are key differences in the generation of these two cell types to achieve their respective objectives [87, 126, 138]. We first review how memory B cells are created, and then consider what criteria they are optimized for, analogous to our criteria (5.2) for plasma B cells.

5.4.1 Background On Memory B Cells

While the name *memory B cell* may invoke the idea that a memory B cell is long-term copy of a plasma B cell, the truth is more complicated. Memory B cells do not undergo the entire AM process like plasma B cells do. In fact, memory B cells are characterized by their relatively low affinity compared to plasma B cells, and low SHM load (number of mutations gathered from SHM) [126]. This implies that while memory B cells have initially high affinity to an antigen relative to the naive B cell repertoire, they do not undergo AM to the extent of plasma B cells, and on average have lower affinity to the current antigen than plasma B cells.

The gene *BACH2* plays an important role in the development of memory B cells within the germinal center, and in their eventual differentiation [117, 126]. *BACH2* has been found to be inversely correlated with the help a B cell receives from T_{FH} cells, and the resulting weak interactions with T_{FH} cells allow for *BACH2* expression to remain high. Critically, the relationship between *BACH2* expression and T_{FH} cells allows for *some* help from T_{FH} cells in order for cell survival within the germinal center, but prevents the B cell precursor from entering the area where it would proliferate and mutate via SHM. This leads to three subsets of B cells within the germinal center: (1) high affinity B cells which are selected for by T_{FH} cells to proliferate and mutate via SHM, and eventually lead to plasma B cell differentiation, (2) moderate affinity B cells (low compared to plasma B cell precursors, high compared to the average naive B cell) whose selection by T_{FH} cells is tempered by *BACH2*, leading to memory B cells, and (3) low affinity B cells which receive little or no help from T_{FH} cells leading to apoptosis [88, 130].

Memory B cells are similar to naive B cells in terms of their transcriptional profiles, which enables them to circulate freely within the organism and to monitor for future instances of antigens. Despite these similarities, they exhibit over-expression of anti-apoptotic genes which allows for the memory B cell to live for extraordinarily long periods of time and therefore the ability to recognize antigens in the future [126].

5.4.2 What Are Memory B Cells Optimized To Do?

If the immune system's objective were rote memorization of the highest affinity B cell receptors to the current antigen a, we might expect the memory B cell receptor repertoire to be nearly identical to the plasma B cells receptor repertoire, but they are not. One might alternatively expect AM to take advantage of the luxury of time it has before it needs the memory B cells to mutate more so that the memory B cells could have even higher affinity $s(b, a)$ to the antigen than the plasma B cells, further optimizing (5.1). That also does not appear to be the case. While both those options should be biologically feasible, the immune system does something radically different to create memory B cells: it selects memory B cells *earlier* in AM than plasma B cells, and thus the memory B cells on average have *lower* affinity $s(b, a)$ than plasma B cells.

To explain why the memory B cells are so poorly-fit to the current antigen a, we propose two hypotheses for the objective function that memory B cells may be heuristically trying to optimize. Our two hypotheses follow from the dual role of memory B cells [47, 138]. First, when *future* incarnations of the antigen a attack, the memory B cells are used as-is to differentiate into plasma B cells and eradicate the *mutated* antigen. Second, memory B cells are used to warm start AM's secondary training of plasma B cells. In fact, recent evidence shows that a large portion of the plasma B cells in the secondary response are memory B cells from the first response, and some memory B cells are also used to seed the new germinal centers to optimize secondary response plasma B cells [88].

5.4.3 Training For Affinity To A Mutated Antigen

We propose that the key issue for memory B cells is that the future instance of the antigen they must mitigate is almost certainly a mutation $\tilde{a}$ of the original antigen a. At the time the memory B cells are created, the future mutated antigen $\tilde{a}$ is unknown, but we can characterize it as a randomly mutated antigen $\tilde{A}$. In this paper, we use the standard probability notation that a capital letter denotes a random variable, and its corresponding lower-case letter denotes the realization of that random variable. For example, if you roll a six-sided die, the random value $X \in \{1, 2, \ldots, 6\}$ refers to the die roll before you look at it because at that point you only know the probability of its six values, but once you see the die roll, it is a deterministic value $x \in \{1, 2, \ldots, 6\}$.

Let $\tilde{A}$ be a random antigen drawn from some conditional probability distribution $P_{\tilde{A}|a}$ that depends on the current antigen a, and models the probability of possible future mutations to a and the probability that such a mutation is presented to the host organism within the lifespan of memory B cells selected during the immune response to a. If only a single memory B cell were required for a secondary response, a logical generalization of (5.1) would be to select a memory B cell that will, on average, have high affinity to the random mutated antigen $\tilde{A}$.

$$\arg\max_{b \in \mathcal{B}} \; E_{\tilde{A}}\left[s(b, \tilde{A})\right]. \tag{5.3}$$

If all mutations of the antigen a are equally likely, then the solution to (5.3) might be the same as (5.1). But if there is substantive asymmetry in the probability of different antigen mutations, then the solution to (5.3) will be different than the solution to (5.1). This is the same principle as in the famous Wayne Gretzky quote about hockey, *I skate to where the puck is going to be, not where it has been.*

However, the situation is more complex, because in each germinal center AM actually produces a *set* of N memory B cells. We hypothesize that AM is evolved to try to produce a *diverse* set of N memory B cells that maximizes the expected affinity between the best-fit of the N memory B cells and the random mutated antigen $\tilde{A}$:

$$\arg\max_{\substack{\{b_n \in \mathcal{B}\},\\ n=1,\ldots,N}} \; E_{\tilde{A}}\left[\max_{n=1,\ldots N}\left[s(b_n, \tilde{A})\right]\right]. \tag{5.4}$$

Fig. 5.2 illustrates the criterion in (5.4), showing that diversity in the memory B cells helps cover the space of probable mutations of the original antigen a. The criterion (5.4) rewards a larger number N of memory B cells, but there is downward pressure on N due to the physical resources needed to store and maintain those cells, and time pressure before the antigen decays away.

5.4.4 Optimizing For Warm Starting Training For A Mutated Antigen

A second role of memory B cells is to warm start future AM processes for $\tilde{a}$. We argue that this role calls for a different criteria as to what makes for a good set of memory B cells. Specifically, analogous to (5.2), we hypothesize that the set of N memory B cells $\{b_n \in \mathcal{B}\}$ should be chosen to minimize the number K of mutations in the secondary response needed to produce a set of N randomly mutated B cells $\{M^K(b_n)\}$ such that one of them is expected to become a secondary response plasma B cell, that is, that it meets the affinity threshold τ with respect to the randomly mutated antigen $\tilde{A}$:

$$\underset{\substack{\{b_n \in \mathcal{B}\},\\ n=1,\ldots,N}}{\arg\min} \left(K \text{ such that } \left(E_{\tilde{A}} \left[E_{M^K} \left[\max_{n=1,\ldots,N} s(M^K(b_n), \tilde{A}) \right] \right] \right) \geq \tau \right). \tag{5.5}$$

We do not mean to suggest that the memory B cells directly optimize (5.5), but rather that evolutionary pressures might have preferred memory B cell selection processes that better optimize (5.5).

Goal (5.4) and goal (5.5) will likely have different optimal solutions depending on the probability of different mutations of the antigen and the B cells, though the same heuristic memory B cell selection process might do pretty well at both objectives. It is not yet known how important memory B cells are to the secondary response plasma B cell training in germinal centers, yet some evidence shows that secondary response germinal centers are comprised of more naive B cells than one might expect [88].

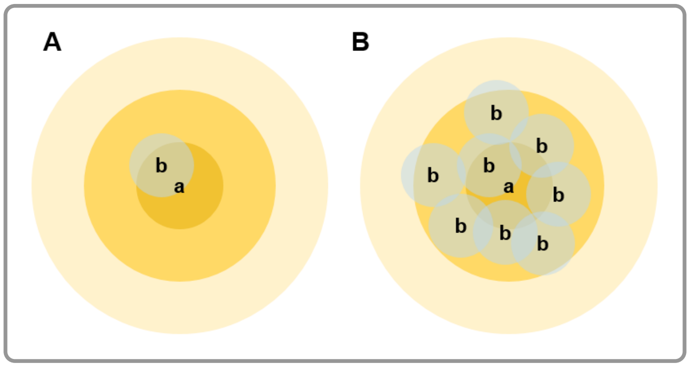


Figure 5.2: Memory B cell coverage of the antigen mutation space. Given an antigen a, we expect to see a mutated version $\tilde{a}$ in the future. Suppose we cannot predict which mutations are more likely, so the probability distribution of the future mutated antigen $\tilde{a}$ is symmetric in the mutation space, shown here as yellow rings of decreasing probability from the original antigen a. (A) If you only get to choose one memory cell, and the antigen is equally likely to mutate randomly in any way, it is optimal to be a copy of a very good plasma cell, marked by b. The blue circle shows the likely mutations of b after its K mutations in a secondary germinal center. (B) If you get to keep a set of N memory cells, spreading the memory B cells out will lead to a higher affinity to more of the possible antigen mutations, rather than keeping N copies of the plasma B cells. This figure was taken from Lindsly *et al.* [75].

5.4.5 Why Are Memory B Cells Not Copies Of Plasma B Cells?

Both (5.4) and (5.5) *appear* to require knowledge of the probability distribution $P_{\tilde{A}}$ of different mutations the antigen may undergo, and the conditional probability distribution $P_{M|b}$ of different mutations of a B cell. However, we note that for many simple choices of $P_{\tilde{A}}$ and $P_{M|b}$, the exact distributions would not matter: the immune system could cheaply achieve a good approximate solution to (5.4) and (5.5) by simply making the memory B cells copies of the plasma B cells; there does not appear to be any biochemical restriction preventing exact replication. However, the memory B cells do in fact appear to be selected for differently than the plasma B cells. We present two hypotheses as to why.

Our first hypothesis was already introduced in Section 5.4.3 and Fig. 5.2: because the immune system gets to select a *set* of memory B cells in (5.4), it pays to have more diversity in the memory B cells than one gets by copying the plasma B cells. Plasma B cells tend to be less diverse because they are trained to optimize (5.1), which even with multiple local minima in the shape of $s(b, a)$, will limit their diversity. Memory B cells are more diverse than plasma B cells because they are selected earlier in the maturation process. We believe this diversity is important to optimize affinity to the mutated antigen as per (5.5) because the true affinity function s is a highly nonlinear function of the amino acid sequences of a B cell b and antigen a that arise from complex biochemical properties and physical lock-and-key structures [2, 21, 63].

Our second hypothesis is that the warm start objective (5.5) for the secondary germinal centers is not well-optimized by a copy of the plasma B cell set because there is evidence that the probabilities $P_{M|b}$ of the random mutations of the B cells in SHM are *asymmetric*: certain mutations of B cells are much more likely than others. That makes some B cells a more flexible starting point for warm-starting than the original plasma B cells, which may have trouble mutating to match the new antigen. Evidence for asymmetric $P_{M|b}$ is that many researchers have noted AID preferential targeting of specific motifs [24, 62, 83, 121]. As mutations would, by definition, change the specific sequence that AID targets, it is reasonable to infer that the first mutation of this location is easier than future ones. Once the sequence is changed, AID is less likely to target this location. Overall, the preferential targeting of AID would make it harder for this region to mutate further or reverse back to the original sequence.

This asymmetry in the probability of moving around the space of all B cells via mutations during AM creates a disconnect between the plasma B cell objective (5.1) and the objective of being a good warm start solution to future plasma B cell training as per (5.5). Specifically, a plasma B cell might have made many difficult-to-reverse mutations to optimize (5.1) for the current antigen a. In contrast, the chosen memory B cells appear to be under-optimized for fitting the current antigen a, but we hypothesize they can more easily mutate in a secondary response AM to better fit the random future antigen $\tilde{A}$. Overall, we note that how well the objectives (5.1), (5.4) and (5.5) align depends on the symmetry of $P_{\tilde{A}}$, $P_{M|b}$, and the nonlinearity of s.

5.5 Simulations

We use the AM algorithm (Algorithm 6) to model how plasma B cells and memory B cells are trained, and show through two simulations that the simulated memory B cells are better than the simulated plasma B cells at optimizing our hypothesized objectives (5.4) and (5.5), thus providing evidence that these objectives are biologically reasonable. We first demonstrate the mechanics of affinity maturation in simulated primary immune responses, then compare different potential initial conditions for simulated secondary immune responses.

These simulations do not account for all of the real-world issues at play, such as that a viral mutation must not harm the virus's functionality, and the issues described in 5.3.3. Despite these limitations, we argue these simulations capture many of the key issues needed to illustrate that our hypothesized objectives are consistent with the difference in plasma and memory B cell training. Complete code for our simulations will be made available upon request.

5.5.1 Simulation Set-up

Our simulations follow Algorithm 6 for the AM process. We initialize a naive B cell repertoire (10,000 cells) with B cell receptors that are represented by a random sequence of 10-50 amino acids, where each amino acid is drawn uniformly over the space of 61 non-stop codons (creating a non-uniform distribution over the amino acids). We simulate the antigens as sequences derived from known antigenic sequences of chicken ovalbumin, bovine milk, and wheat [59, 78]. Each antigenic sequence is 17 amino acids long for consistency. We simulate the affinity metric s between a B cell and the antigen using the standard `localalign` MATLAB function, which finds the optimal alignment between two sequences using the `BLOSUM50` matrix and returns a score reflecting how similar two sequences are in this alignment [58]. We use this score as a measure of affinity between the B cell receptor and the antigen for simplicity, but note that it is only a rough approximation of the more complex structural compatibility between a B cell receptor and an antigen.

Mutations of the B cell during AM are modeled in the codon space, where codons of the B cell receptor are replaced with one of the 61 codon possibilities. While SHM mutates B cells on a single nucleotide level, working in the codon space prevents the added complication of filtering out nonsensical B cell receptor sequences. In addition to the replacement of codons, codons in B cell receptor sequences can be inserted or deleted. Insertions and deletions are less likely to occur than swapping for another codon, based on rates of each type of mutation observed in humans [28, 55, 149]. The codons defining each B cell receptor are chosen to mutate at random, but we simulate codons that contain C cytosines to be $C + 1$ times more likely to be mutated than codons without cytosine, reflecting biological biases to nucleotide sequence motifs [24, 62, 83, 121]. We also impose a transition bias between codons, making some swaps more likely than others based on a mutability matrix derived from the `BLOSUM50` matrix [58, 136].

For our initial simulation of a primary adaptive immune response to an antigen (primary response), we randomly select 50 naive B cells from the B cell repertoire from the top 1,000 of the 10,000 naive B cell repertoire in terms of affinity to said antigen. This reflects the recruitment of naive B cells with some partial affinity by $\mathrm{T_{FH}}$ cells to germinal centers [87, 129]. These 50 'founder' B cells are then duplicated 20 times to form a germinal center population of 1,000 cells, reflecting the growth period of germinal center formation [3]. For each iteration of the simulation, 50 B cells are selected for proliferation and 50 B cells are selected for removal. Higher affinity B cells have a higher selection probability for proliferation, while lower affinity B cells have a higher selection probability for removal. The B cells selected for proliferation are duplicated and mutated, replacing all B cells selected during this iteration. This process imitates apoptosis of low affinity B cells from lack of $\mathrm{T_{FH}}$

cell help and proliferation of B cells with high affinity after being selected by T_{FH} cells. These mutations have the possibility to increase, decrease, or have no effect on the affinity of the B cell receptors. We also impose constraints on which mutations can occur on a particular iteration, simulating a negative selection process due to damaging or potentially dangerous mutations. This entire process repeats over 100 iterations.

As a basis for both Simulations 1 and 2, we extract 50 cells during the first half of our primary response simulation as the simulated memory B cells. The cells are randomly chosen from the top 25th percentile of germinal center B cells. Another 50 cells are selected at the very end of the primary response to represent plasma B cells, where these cells exhibit the highest 50 affinity scores to the antigen. This reflects a slight deviation from Algorithm 6, as we do not know the threshold τ *a priori*. We establish τ for the secondary response based on the affinity of the plasma cells chosen at the end of the primary response. As expected, the simulated memory B cells have overall lower affinity to the antigen compared to the plasma B cells, but higher than the initial set of naive B cells.

Our simulated mutations of the antigen for the secondary response are derived from a uniform random swap of any of the codons for any other (including possibly itself, i.e. a no-op), which creates a non-uniform distribution over the amino acids as some amino acids are coded for by multiple codons. Mutations of the antigen can also include insertions or deletions of codons, with equal probability to any codon.

5.5.2 Simulation 1: Affinity To A Mutated Antigen

A secondary infection could involve an antigen that has been mutated or an antigen that is similar from a related pathogen. We simulate the changes in the antigen from the primary to secondary infection by causing adversarial mutations to our antigenic sequence. First, we generate 1,000 uniformly random mutations of antigen

a's sequence. The random mutations may be a swap of any amino acid to any other, an insertion of any amino acid, or a deletion of an amino acid. Of those candidate mutations, we keep the one that has the lowest average affinity (worst case) to the set of 50 plasma B cells from the primary response, to reflect that a potentially dangerous secondary infection would likely be from a more challenging mutation. We repeat this random process a total of K times to produce an antigen with K mutations. We take that *adversarial* antigen $\tilde{a}$ to be the worst case realization of $P_{\tilde{A}}$ from the candidate mutations.

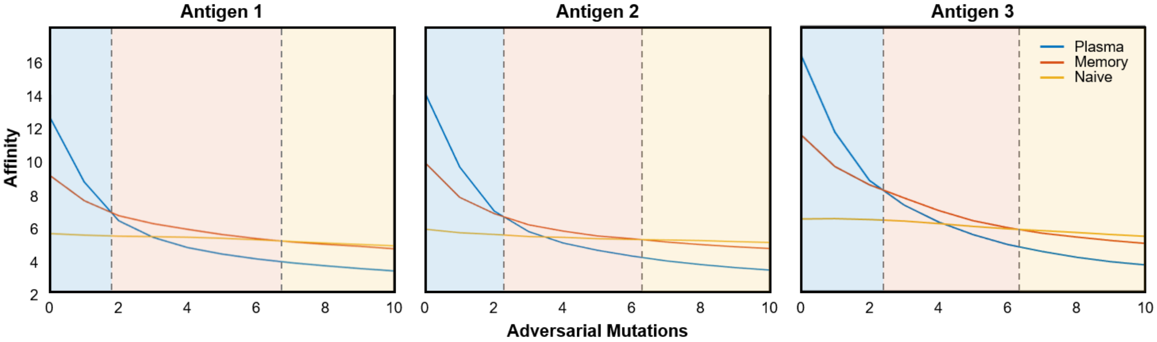


Figure 5.3: Potential initial conditions for a secondary adaptive immune response. Simulation results averaged over 100 independent runs of the primary response for three antigens (from left to right: chicken ovalbumin, bovine milk, and wheat). Plots show average affinity of naive, plasma, and memory B cells to adversarially mutated antigens with $K = 1, \ldots, 10$ mutations. Ranges of mutations where plasma, memory, and naive B cells are optimal are shaded blue, orange, and yellow respectively. This figure was taken from Lindsly *et al.* [75].

We then test the different B cell populations by the hypothesized goal of (5.4). Fig. 5.3 shows the average affinity between increasingly mutated antigens (chicken ovalbumin, bovine milk, and wheat) with mutations $K = 1, \ldots, 10$ for different cell populations at the end of the primary response, averaged over 100 independent runs [59, 78]. Fig. 5.3 shows that our simulated plasma B cells are the best choice to maximize (5.4) for a small number of adversarial mutations, our simulated memory B cells are the best choice between ∼2-6 mutations, and naive B cells are best after many mutations (7+). Our simulations are too simplified for the specific transition points to be meaningful, but we argue they do provide strong evidence that the plasma B cells are likely not optimal for identifying substantially-mutated antigens. This may further suggest that there is some region in mutation space where memory B cells are more useful than plasma B cells or naive B cells as an initial condition for a secondary response.

We hypothesize that the fact that plasma B cells do not always have the highest affinity to mutated antigens is driven by the greater diversity of the memory B cells and naive B cells. While the mutations occur in DNA space, the relevant diversity is in the resulting nonlinear physical and biochemical space that defines the affinity to the antigen. Approximately measuring the B cell diversity in each population using the pairwise BLOSUM similarities in each set shows substantial diversity differences, with the plasma B cells having average within-set BLOSUM similarity of $\sim$46, the memory B cells having much lower average within-set BLOSUM similarity of $\sim$27, and the naive B cells having even lower within-set BLOSUM similarity of $\sim$9 (from a representative primary response).

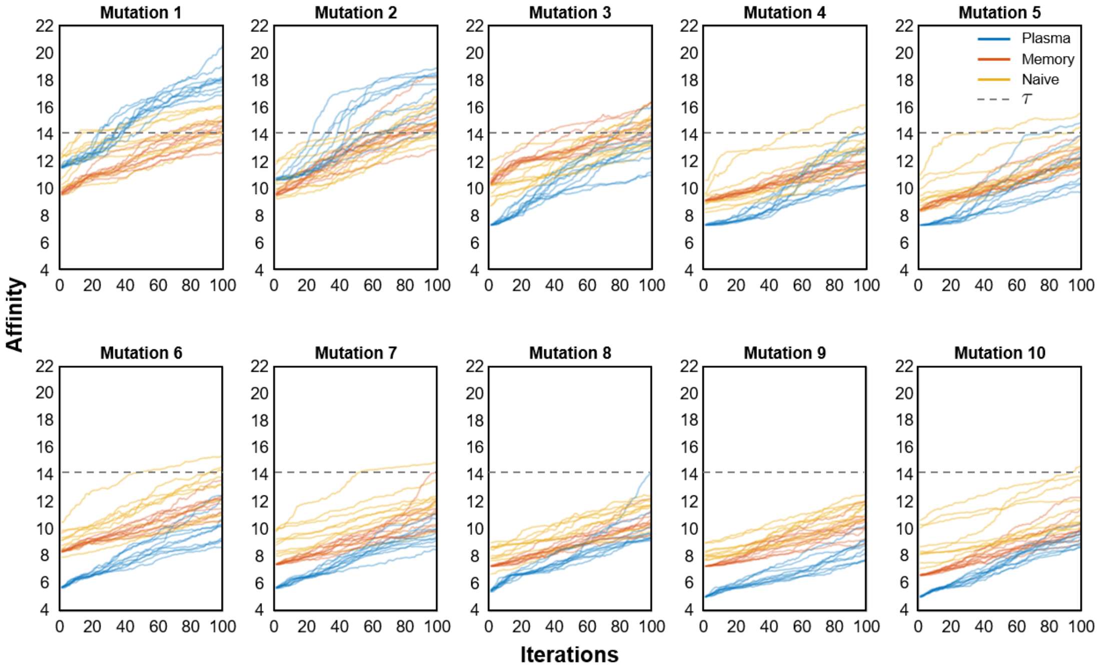


Figure 5.4: Example performance of different populations as initial conditions for secondary responses. We chose one representative run of the primary response simulation to the bovine milk antigen [78]. Each line represents one of the 10 secondary response simulations as a function of SHM iterations when initialized by a set of 50 plasma, memory, or naive B cells (as marked), for $K = 1, \ldots, 10$ sequential adversarial mutations. The y-axis marks the average affinity of the top 50 (out of 1,000) highest affinity B cells during the secondary response to the mutated antigen, representing the new potential plasma B cells from the secondary response. Dashed horizontal line reflects the average affinity, τ, of the plasma cells from the primary response. For this run of the simulation, the plasma B cells had difficulty achieving high affinity to the mutated antigens after three adversarial mutations. The memory B cells were the most useful seeds for the three-mutation case. Once the antigen was mutated four times, only the naive B cells were sometimes able to achieve the affinity threshold before 100 iterations of SHM. Surprisingly, the average affinities plotted grew fairly linearly and at roughly the same rate for most of the lines, suggesting fairly constant progress, and that none of the populations became stuck, but rather just started from a worse initial affinity. This figure was taken from Lindsly *et al.* [75].

5.5.3 Simulation 2: Mutations Needed For Secondary Response Plasma B Cell Training

In this simulation, we compare how well the three types of B cells perform as seeds for the secondary response germinal center. We mimic a secondary response training of a new set of plasma B cell's optimized for high affinity to a mutated antigen $\tilde{a}$ (described in Simulation 1). We initialize the secondary response plasma B cell optimization with one of three choices: *(i)* the $N = 50$ plasma B cells generated in the primary response for the original antigen a, *(ii)* the $N = 50$ memory B cells generated in the primary response for the original antigen a, or *(iii)* $N = 50$ naive B cells. We populate the germinal center in an identical way to the primary response, using the three sets of 50 B cells as our new founder cells. As in the primary response, the naive B cells are random, but chosen to have some initial affinity to the now-mutated antigen to simulate recruitment to the germinal center. Each of these three secondary response germinal centers undergo AM in an identical fashion to the primary response.

Fig. 5.4 are representative examples of multiple secondary response simulations, given the same primary response and the same $K = 1, \ldots, 10$ mutations on the bovine milk antigen [78]. Specifically, it shows the average affinity of the 0.05% highest affinity cells over 100 iterations in secondary responses. Fig. 5.4 highlights that for just one or two antigen mutations, plasma B cells tend have the highest initial affinity to the mutated antigen. For three or more adversarial mutations, naive and memory B cells are better initial conditions for the secondary response than the primary response plasma B cells, and reach an affinity of τ in fewer iterations (Figs. 5.4 and 5.5). Fig. 5.4 also conveys how the recruited secondary response naive B cells tend to have higher initial affinity to the mutated antigen than either the primary response plasma B cell or memory B cells once the antigen has been sufficiently mutated. This was expected, but we were surprised at how few mutations it took for the naive B cells to have the highest affinity at the first iteration of the secondary response. That is, for relatively few codon mutations in the new antigen, we find empirical evidence that seeding a secondary response with primarily B cells from the naive repertoire is an advantageous strategy. We were not able to simulate a region in mutation space where the memory B cells from the primary response consistently had the highest affinity at the first iteration of the secondary response, likely due to considerable variation between simulated secondary responses.

Similarly, Fig. 5.5 shows the average number of iterations for plasma, memory, and naive B cells to reach τ over all secondary response simulations for all three antigens. The iteration number, where the threshold τ is reached, is averaged over all secondary response simulations for 50 independent simulations of the primary response for each antigen. Again, our simulations did not show a region of mutation space where memory B cells were consistently the best warm start conditions. We found this to persist across a number of refinements of our simulation, suggesting this finding might not be an artifact of too coarse a simulation. We hypothesize that this is due to the selection of naive B cells for the secondary germinal center being biased to have some initial affinity to the mutated antigen (as described in the primary response simulation). That is, naive B cells in Fig. 5.4 are selected for initial affinity to $\tilde{a}$, while memory and plasma B cells derived from naive B cells selected for having some initial affinity to the original antigen a (Fig. 5.3). We initially suspected that our affinity bias for the selection of naive cells was too strong, but both naive and memory B cells are recruited simultaneously in real secondary germinal centers in order to have the best chance of creating new plasma B cells. In fact, recent experimental evidence suggests that secondary response germinal centers are comprised of more naive B cells than previously thought [88]. It is also possible that there is an antigen mutation regime for which the memory B cells are indeed more effective than naive B cells for warm starting, but that our simulations were not realistic enough to capture it.

We note that the secondary response simulations often take more than 100 iterations to reach the affinity threshold τ from their corresponding primary simulation. When calculating the average number of iterations to reach τ in Fig. 5.5, we set these cases to the maximum iteration number of 100. We hypothesize that this phenomenon occurs because the adversarial mutations of $\tilde{a}$ may create a more difficult problem for the simulated germinal center to solve.

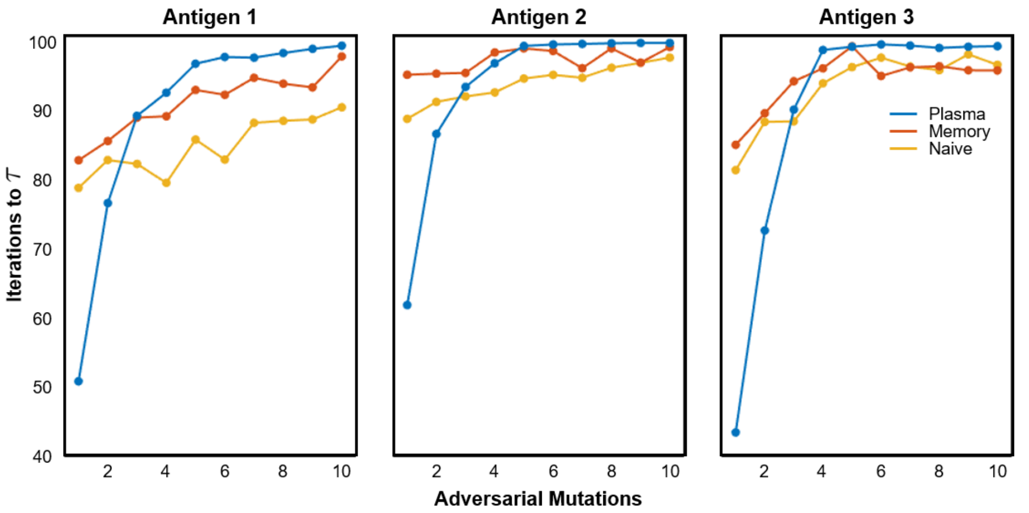


Figure 5.5: Convergence rate for potential initial conditions across all secondary responses. Time of convergence to τ for plasma, memory, and naive B cells during simulated secondary germinal centers' warm starts for $K = 1, \ldots, 10$ mutations averaged over 50 independent runs of the entire simulation, where each of the primary response simulations had 10 corresponding secondary response simulations. Lower values indicate a faster time to convergence, showing that the plasma B cells are the best warm starts for one or two of the simulated adversarial mutations (dependent on the antigen). Naive B cells are the best warm starts for almost all cases with three or more simulated adversarial mutations. Antigens 1, 2, and 3 correspond to chicken ovalbumin, bovine milk, and wheat, respectively [59, 78]. This figure was taken from Lindsly *et al.* [75].

5.6 Conclusions and Open Questions

We hypothesized that the dual role of memory B cells can be captured by two objectives (5.4) and (5.5), and that these objectives would not be as well-optimized by copying plasma B cells that are trained for (5.1), due to their over-fitting the original antigen. Our simulations, while limited, provide strong evidence that plasma B cells would not optimize (5.4) or (5.5) once the antigen underwent sufficient adversarial mutations. We believe this suboptimality of plasma B cells against mutated antigens provides a role for the different selection mechanism used for memory B cells.

Our simulations show a limited range of antigen mutations over which our simulated memory B cells may be optimal; for substantial mutations, we show random naive B cells can work even better. These findings are consistent with our knowledge of the adaptive immune system. Plasma B cells are a one-time solution and are highly overfit to the current antigen of interest. Memory B cells provide a more approximate solution to the current antigen, which is kept within the body to recognize future antigens with similar characteristics. If a future antigen is so different from what has been previously encountered by the immune system that no memory B cells are able to identify it, a new solution is formed from scratch using naive B cells.

Memory B cells play two roles, both differentiating into plasma B cells and re-initiating germinal centers, but these roles may be played by distinct subpopulations [47, 138]. These distinct subpopulations of memory B cells might have resulted from distinct AM processes, or changes in the AM process over the AM time span that we have not explicitly modeled in our Algorithm 6 [139]. Thus our two memory B cell objectives in (5.4) and (5.5) may apply to independent memory B cell populations. Here we investigated a simplified model, where memory B cells were considered a unified group. However, we were not able to show via simulations a regime in which the memory B cells were clearly better than secondary naive B cells for re-initiating the secondary germinal response. Our results suggest this re-initialization task might be a weaker or rarer role of the memory B cells. These results align with recent experimental evidence that similarly noted memory B cells were less prevalent in secondary germinal centers than previously assumed [88]. However, even if memory B cells are not always needed for the secondary response, it might be that in some cases they are very important for warm starting, which might still exert evolutionary pressure on their selection process.

The evolutionary pressures on memory B cell selection in nature are not known, but may be elucidated through the integration of computational simulations and biological experiments. Actively monitoring the affinity of B cells during affinity maturation, as well as detecting when and why GC B cells become memory B cells, may assist with the development of more accurate and complex models in the future.

CHAPTER VI

Concluding Remarks

The study of genomics has made extraordinary advancements in recent decades, and will be a major focus for the foreseeable future. The human genome was only mapped within the last 20 years, yet we have made great strides towards understanding how it operates [30]. New technologies, such as RNA-seq, Hi-C, ChIP-seq, ATAC-seq, and Pore-C have allowed us to study the genome with unprecedented levels of detail and gain access to massive quantities of data. These data have given us a better understanding of how the genome operates normally, and we have begun to learn how small mutations in the genome are associated with certain diseases. These findings have helped us predict individual's predispositions for disease and how certain medications will behave on an individual basis. In addition, the recent discovery of gene editing through CRISPR-Cas9 will allow us to manipulate the genome for therapeutic treatments. Using all of this information as a foundation, the next goal in genomic research is to understand the dynamical relationship of genome structure and function and how it affects cell phenotype. Research of the 4D Nucleome will be crucial in this pursuit.

One of the major challenges of studying the 4DN is the integration and analysis of different data types. In Chapter II, we introduce the 4DNvestigator MATLAB toolbox. The 4DNvestigator allows for both novice and experienced users to process and analyze data in a simple yet mathematically rigorous manner. We provide functions which are able to do standard gene expression and genome structure analysis for time series data. In addition, we offer methods to integrate Hi-C with multiple data modalities, find statistically significant changes between Hi-C data from different settings, and visualize low-dimensional representations of these data. We provide multiple examples of how the 4DNvestigator can be applied to different scenarios, including data from cellular proliferation, differentiation, and reprogramming experiments. The 4DNvestigator allows for 4DN data to be quickly explored and characterized, without the need for an in-depth understanding of the data types' disparate formats and computational nuances.

Research on the 4DN has been gaining traction in recent years, but many studies ignore the fact that cells contain a copy of both the maternal and paternal genomes. Chapter III describes an extension of 4DN analysis which separates the two genomes. This analysis uses many of the features included in the 4DNvestigator, such as TAD identification, differential expression of genes, and the integration of genome structure and function through time. Specifically, this work applies these techniques to genome structure and gene expression data that has been separated into the maternal and paternal components across the cell cycle. We identified hundreds of genes which were differentially expressed between their maternal and paternal alleles through RNA-seq, and that these genes had significant changes in their local genome structures through Hi-C. We also integrated publicly available protein binding data which provided further evidence that the maternal and paternal genomes can have distinct properties, outside of the known cases like X-Chromosome inactivation and imprinting. These findings represent a significant step towards comprehensive genome analysis. Further research with allele-specific data will be crucial for understanding the complex relationship between genome structure and function, and its impact on cellular phenotypes. Extensions to this work will become increasingly important as the development of personalized medicine and therapies rises in popularity.

Hi-C data are the current standard for analyzing genome structure. With the development of new technologies, such as Pore-C, we are now able to extend the pair-wise contacts of Hi-C to multi-way contacts. Multi-way contacts offer a unique advantage, as we can now observe long range interactions of multiple loci at once. From this, we can gain a more detailed understanding of the higher order structural features of the genome. In Chapter IV, we introduce methods to process and analyze multi-way contacts from Pore-C. We offer algorithms to construct hypergraphs in the form of incidence matrices, calculate entropy of those hypergraphs, and compare hypergraphs to one another. In addition, we provide a pipeline to identify transcription clusters using the multi-way contacts of Pore-C in conjunction with other data modalities. These techniques will be increasingly useful as the study of genomics progresses, since multi-way contacts can better inform us of how genes are regulated. This information can then be used to influence gene expression, and could have applications for clinical therapies.

Finally in Chapter V, we deviate from studying the internal details of the genome and instead focus on cellular dynamics within the immune system. We mathematically characterize the objectives that the adaptive immune system is optimizing for, and provide simulated results to distinguish the objectives of memory and plasma B cells. While the affinity maturation algorithm and simulations we define are vastly simplified compared to the real mammalian immune system, we believe that they are able to offer valuable insights into how the adaptive immune system operates. Additional data derived from experiments on the adaptive immune system paired with increasingly detailed simulations may help us engineer better ways to fight novel infections and improve our treatments for a multitude of diseases, such as autoimmunity or cancers.

APPENDIX

APPENDIX A

Supplemental Materials

All supplemental methods, figures, and tables for Chapter II and Chapter III can be found in Lindsly *et al.* [76] and Lindsly *et al.* [77], respectively.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Adams, J. Imprinting and genetic disease: Angelman, Prader-Willi and Beckwith–Weidemann syndromes. *Nat Educ* **1** (2008).

2. Ambrosetti, F., Jim'enez-Garc'ia, B., Roel-Touris, J. & Bonvin, A. M. J. J. Modeling Antibody-Antigen Complexes by Information-Driven Docking Author links open overlay panel. *Structure,* 119–129 (2020).

3. Amitai, A., Mesin, L., Victora, G. D., Kardar, M. & Chakraborty, A. K. A population dynamics model for clonal diversity in a germinal center. *Frontiers in microbiology* **8,** 1693 (2017).

4. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106 (2010).

5. Anderson, S. M. *et al.* Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *The Journal of Immunology* **183,** 7314–7325 (2009).

6. Babak, T. *et al.* Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature Genetics* **47,** 544 (2015).

7. Bannard, O. & Cyster, J. G. Germinal centers: programmed for affinity maturation and antibody diversification. *Current opinion in immunology* **45,** 21–30 (2017).

8. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Research* **25,** 927–936 (2015).

9. Beliveau, B. J. *et al.* Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nature Communications* **6,** 1–13 (2015).

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

11. Ben-David, E., Shohat, S. & Shifman, S. Allelic expression analysis in the brain suggests a role for heterogeneous insults affecting epigenetic processes in autism spectrum disorders. *Human Molecular Genetics* **23,** 4111–4124 (2014).

12. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289–300 (1995).

13. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353,** 163–166 (2016).

14. Berge, C. *Hypergraphs: combinatorics of finite sets* (Elsevier, 1984).

15. Bloch, I. & Bretto, A. *A new entropy for hypergraphs* in *International Conference on Discrete Geometry for Computer Imagery* (2019), 143–154.

16. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350,** 1092–1096 (2015).

17. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol* **3,** e157 (2005).

18. Buiting, K. *Prader–Willi syndrome and Angelman syndrome* in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **154** (2010), 365–376.

19. Campbell, N. A., Mitchell, L. G., Reece, J. B. & Taylor, M. R. *Biology: concepts & connections* (Benjamin/Cummings, 2000).

20. Cao, Y. *et al.* A novel deletion of SNURF/SNRPN exon 1 in a patient with Prader-Willi-like phenotype. *European Journal of Medical Genetics* **60,** 416–420 (2017).

21. Carneiro, J. & Stewart, J. Rethinking Shape Space: Evidence from Simulated Docking Suggests that Steric Shape Complmementarity Is Not Limiting For Antibody-Antigen Recognition And Idiotypic Interactions. *Journal of Theoretical Biology* **169,** 391–402 (1994).

22. Chen, C. & Rajapakse, I. Tensor Entropy for Uniform Hypergraphs. *IEEE Transactions on Network Science and Engineering* **7,** 2889–2900 (2020).

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

24. Chen, J. & MacCarthy, T. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. *PLoS computational biology* **13,** e1005471 (2017).

25. Chen, J., Hero III, A. O. & Rajapakse, I. Spectral identification of topological domains. *Bioinformatics* **32,** 2151–2158 (2016).

26. Chen, J. *et al.* A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nature Communications* **7,** 11101 (2016).

27. Cohen, P. W., Omenn, G., Motulsky, A., Chen, S.-H. & Giblett, E. Restricted variation in the glycolytic enzymes of human brain and erythrocytes. *Nature New Biology* **241,** 229–233 (1973).

28. Consortium, 1. G. P. A global reference for human genetic variation. *Nature* **526,** 68 (2015).

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

30. Consortium, I. H. G. S. *et al.* Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931 (2004).

31. Cook, P. R. A model for all genomes: the role of transcription factories. *Journal of Molecular Biology* **395,** 1–10 (2010).

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

33. Cook, P. R. & Marenduzzo, D. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic acids research* **46,** 9895–9906 (2018).

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

35. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harbor Perspectives in Biology* **2,** a003889 (2010).

36. De Lathauwer, L., De Moor, B. & Vandewalle, J. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* **21,** 1253–1278 (2000).

37. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nature reviews immunology* **15,** 137–148 (2015).

38. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–3212 (2009).

39. Dekker, J. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).

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

41. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).

42. De Wit, E. & De Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & development* **26,** 11–24 (2012).

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

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

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

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

47. Dogan, I. *et al.* Multiple layers of B cell memory with different effector functions. *Nature immunology* **10,** 1292–1299 (2009).

48. Donnat, C. & Holmes, S. Tracking network dynamics: A survey using graph distances. *The Annals of Applied Statistics* **12,** 971–1012 (2018).

49. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research* **16,** 1299–1309 (2006).

50. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3,** 95–98 (2016).

51. Faust, K. & Skvoretz, J. Comparing networks across space and time, size and species. *Sociological methodology* **32,** 267–299 (2002).

52. Finn, E. H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365** (2019).

53. Finn, E. H. *et al.* Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176,** 1502–1515 (2019).

54. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic acids research* **48,** D58–D64 (2020).

55. Gibbs, R. A. *et al.* The international HapMap project (2003).

56. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318,** 1136–1140 (2007).

57. Gudmundsson, J. *et al.* Different tumor types from BRCA2 carriers show wild-type chromosome deletions on 13q12–q13. *Cancer Research* **55,** 4830–4832 (1995).

58. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89,** 10915–10919 (1992).

59. Honma, K. *et al.* Allergenic epitopes of ovalbumin (OVA) in patients with hen's egg allergy: inhibition of basophil histamine release by haptenic ovalbumin peptide. *Clinical & Experimental Immunology* **103,** 446–453 (1996).

60. Hu, D., Li, X. L., Liu, X. G. & Zhang, S. G. Extremality of graph entropy based on degrees of uniform hypergraphs with few edges. *Acta Mathematica Sinica, English Series* **35,** 1238–1250 (2019).

61. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28,** 27–30 (2000).

62. Keim, C., Kazadi, D., Rothschild, G. & Basu, U. Regulation of AID, the B-cell genome mutator. *Genes & development* **27,** 1–17 (2013).

63. Kilambi, K. P. & Gray, J. J. Structure-based cross-docking analysis of antibody–antigen interactions. *Scientific reports* **7,** 1–15 (2017).

64. Kim, Y. *et al.* Elevated urinary CRELD2 is associated with endoplasmic reticulum stress–mediated kidney disease. *JCI Insight* **2** (2017).

65. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33,** 1029–1047 (2013).

66. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* **68,** 820–823 (1971).

67. Koziol, J. A. *et al.* A graphical technique for displaying correlation matrices. *The American Statistician* **51,** 301–304 (1997).

68. Kukurba, K. R. & Montgomery, S. B. RNA sequencing and analysis. *Cold Spring Harbor Protocols* **2015,** pdb–top084970 (2015).

69. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317 (2015).

70. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172,** 650–665 (2018).

71. Larntz, K. & Perlman, M. D. A simple test for the equality of correlation matrices. *Rapport technique, Department of Statistics, University of Washington* **141** (1985).

72. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350 (2015).

73. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

74. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

75. Lindsly, S., Gupta, M., Stansbury, C. & Rajapakse, I. Understanding Memory B Cell Selection. *arXiv preprint arXiv:2012.05817* (2021).

76. Lindsly, S. *et al.* 4DNvestigator: time series genomic data analysis toolbox. *Nucleus* **12:1** (2021).

77. Lindsly, S. *et al.* Functional Organization of the Maternal and Paternal Human 4D Nucleome. *bioRxiv* (2021).

78. Liu, C. & Sathe, S. K. Food allergen epitope mapping. *Journal of agricultural and food chemistry* **66,** 7238–7248 (2018).

79. Liu, S. *et al.* Genome architecture mediates transcriptional control of human myogenic reprogramming. *iScience* **6,** 232–246 (2018).

80. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 1–21 (2014).

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

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

83. Martin, A., Chahwan, R., Parsa, J. Y. & Scharff, M. D. in *Molecular Biology of B cells* 363–388 (Elsevier, 2015).

84. Maxwell, K. N. *et al.* BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. *Nature Communications* **8,** 1–11 (2017).

85. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

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

87. Mesin, L., Ersching, J. & Victora, G. D. Germinal center B cell dynamics. *Immunity* **45,** 471–482 (2016).

88. Mesin, L. *et al.* Restricted clonality and limited germinal center reentry characterize memory B cell reactivation by boosting. *Cell* **180,** 92–106 (2020).

89. Meyer-Hermann, M. *et al.* A theory of germinal center B cell selection, division, and exit. *Cell reports* **2,** 162–174 (2012).

90. Minello, G., Rossi, L. & Torsello, A. On the von Neumann entropy of graphs. *Journal of Complex Networks* **7,** 491–514 (2019).

91. Misteli, T. The inner life of the genome. *Scientific American* **304,** 66 (2011).

92. Misteli, T. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* (2020).

93. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102,** 553–563 (2000).

94. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).

95. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150,** 1274–1286 (2012).

96. Newman, M. *Networks* (Oxford University Press, 2018).

97. Ng, A. Y., Jordan, M. I., Weiss, Y., *et al. On spectral clustering: Analysis and an algorithm* in *NIPS* **14** (2001), 849–856.

98. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745 (2016).

99. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics* **36,** 1065–1071 (2004).

100. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank citation ranking: Bringing order to the web.* tech. rep. (Stanford InfoLab, 1999).

101. Passerini, F. & Severini, S. The von Neumann entropy of networks. *arXiv:0812.2597* (2008).

102. Paulsen, M. T. *et al.* Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* **67,** 45–54 (2014).

103. Petroziello, J. *et al.* Suppression subtractive hybridization and expression profiling identifies a unique set of genes overexpressed in non-small-cell lung cancer. *Oncogene* **23,** 7734–7745 (2004).

104. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

105. Rajapakse, I., Groudine, M. & Mesbahi, M. What can systems theory of networks offer to biology? *PLoS computational biology* **8,** e1002543 (2012).

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

107. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).

108. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics* **2,** 21 (2001).

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

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

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

112. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology* **7,** 522 (2011).

113. Santoni, F. A. *et al.* Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. *The American Journal of Human Genetics* **100,** 444–453 (2017).

114. Seaman, L. & Rajapakse, I. 4D nucleome Analysis Toolbox: analysis of Hi-C data with abnormal karyotype and time series capabilities. *Bioinformatics* **34,** 104–106 (2017).

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

116. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology* **31,** 1111 (2013).

117. Shinnakasu, R. *et al.* Regulated selection of germinal-center cells into the memory B cell compartment. *Nature Immunology* **17,** 861 (2016).

118. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics* **38,** 1348–1354 (2006).

119. Solovei, I. *et al.* Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Experimental Cell Research* **276,** 10–23 (2002).

120. Sompayrac, L. M. *How the immune system works* (John Wiley & Sons, 2019).

121. Stavnezer, J. Complex regulation and function of activation-induced cytidine deaminase. *Trends in immunology* **32,** 194–201 (2011).

122. Stelzer, G. *et al.* The GeneCards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **54,** 1–30 (2016).

123. Strang, G. *Introduction to Linear Algebra* (Cambridge Press, 2016).

124. Strang, G. *Introduction to linear algebra* (Wellesley-Cambridge Press Wellesley, MA, 1993).

125. Strogatz, S. H. Exploring complex networks. *Nature* **410,** 268–276 (2001).

126. Suan, D., Sundling, C. & Brink, R. Plasma cell and memory B cell differentiation from the germinal center. *Current Opinion in Immunology* **45,** 97–102 (2017).

127. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332,** 472–474 (2011).

128. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361,** 924–928 (2018).

129. Tas, J. M. *et al.* Visualizing antibody affinity maturation in germinal centers. *Science* **351,** 1048–1054 (2016).

130. Taylor, J. J., Pape, K. A., Steach, H. R. & Jenkins, M. K. Apoptosis and antigen affinity limit effector cell differentiation of a single naive B cell. *Science* **347,** 784–787 (2015).

131. Theodosopoulos, P. K. & Theodosopoulos, T. V. in *Evolution as Computation* 41–66 (Springer, 2002).

132. Ulahannan, N. *et al.* Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv,* 833590 (2019).

133. Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N. & Fekete, J.-D. Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE transactions on visualization and computer graphics* **27,** 1–13 (2019).

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

135. Varoquaux, N., Ay, F., Noble, W. S. & Vert, J.-P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30,** i26–i33 (2014).

136. Veerassamy, S., Smith, A. & Tillier, E. R. A transition probability model for amino acid substitutions from blocks. *Journal of Computational Biology* **10,** 997–1010 (2003).

137. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10,** 57–63 (2009).

138. Weisel, F. & Shlomchik, M. Memory B cells of mice and humans. *Annual review of immunology* **35,** 255–284 (2017).

139. Weisel, F. J., Zuccarino-Catania, G. V., Chikina, M. & Shlomchik, M. J. A temporal switch in the germinal center determines differential output of memory B and plasma cells. *Immunity* **44,** 116–130 (2016).

140. Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **36,** D13–D21 (2007).

141. Wills, P. & Meyer, F. G. Metrics for graph comparison: a practitioner's guide. *PloS one* **15,** e0228728 (2020).

142. Wolf, M. M., Klinvex, A. M. & Dunlavy, D. M. *Advantages to modeling relational data using hypergraphs versus graphs* in *2016 IEEE High Performance Extreme Computing Conference (HPEC)* (2016), 1–7.

143. Wu, F.-J. *et al.* BMP8A sustains spermatogenesis by activating both SMAD1/5/8 and SMAD2/3 in spermatogonia. *Science Signaling* **10** (2017).

144. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).

145. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal* **36,** 3573–3599 (2017).

146. Zakharova, I. S., Shevchenko, A. I. & Zakian, S. M. Monoallelic gene expression in mammals. *Chromosoma* **118,** 279–290 (2009).

147. Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature genetics* **51,** 1380–1388 (2019).

148. Zhou, D., Huang, J. & Scholkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems* **19,** 1601–1608 (2006).

149. Zia, A. & Moses, A. M. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC bioinformatics* **12,** 1–14 (2011).