

Anchored to the Gene: Sketching topographHi-C maps of the Genome

A Dissertation Presented by

Jocelyn Tourtellotte, MPH, MAT

This work was undertaken at Worcester Polytechnic Institute and the Morningside Graduate
School of Biomedical Sciences

Bioinformatics and Computational Biological
&
WPI/UMass Chan Joint Program

Under the mentorship of

David Grunwald, PhD, Thesis Advisor

Dmitry Korkin, PhD, Member of Committee

Sam Walcott, PhD, Member of Committee

Nick Rhind, PhD, Member of the Committee

20-April-2024

Abstract

The identity of each cell is defined by the genes it expresses, which is predicated by the three- and four-dimensional organization of chromatin in the nucleus. The nuclear pore complex (NPC) guides such gene expression by physically interacting with cell-type specific genes combinations. These genes are identifiable by DamID-fused nucleoporin, making the NPC a physical reference point. Guided by these nuclear waypoints, Hi-C data can be integrated to further models for organization of the genome at the nuclear periphery.

For biological and computational reasons, analyzing Hi-C data is most effective for examining interactions between regions on the same chromosome – cis-interactions. Accordingly, visualization methods primarily focused on what is happening “along the diagonal”.

We propose using kernel density estimation, a method frequently applied in ecology and epidemiology, as a chromosome-agnostic method for quantifying the probability of observing interactions at the “local” (gene) level. Subsequently, these densities are used to create contour plots that depict the regions most likely interaction along the respective lengths of the pair. Although such a targeted approach is not unique in the realm of chromatin capture as a whole – there exist targeted applications of such techniques – by its nature Hi-C data lends itself to a “global” perspective. Therefore, a framework for a “local” approach to Hi-C data analysis was constructed in conjunction with expanding our visualization toolbox. This framework will be used to integrate additional data-types and guide microscopy-based research of genome organization at the nuclear periphery.

Table of Contents

1.	Introduction: The Art of Science.....	1
2.	Overview.....	4
1.1	Don't Take This Out of Context.....	4
1.1.1	Biological Models	5
1.1.1.1	Cell lines.....	5
1.1.2	Experimental Methods	6
1.1.3	What is your alignment?.....	7
1.1.4	Analytical Methods	8
1.2	Speaking of Context... The Nuclear Pore Complex	8
1.2.1	Generalities	9
1.2.2	And who? ...are? ...you?	9
1.3	Considerations for Multiomic data Integration	9
1.3.1	The Need for Metadata	10
1.3.2	Application of Emerging Technologies: Metadata Strikes Again	11
1.3.3	Finding a Common Language in Probabilities	12
1.4	This is the Way.	13
3.	Part I. Multiomic Integration for Exploring Therapeutic Gene Targets	14
1.5	Background	15
1.5.1	Rheumatoid Arthritis.....	15
1.5.1.1	Biological Models	17
1.5.1.2	Genome-Wide Association Studies for Rheumatoid Arthritis	18
1.5.2	The Open Targets Platform > Systematic Aggregation of Potential Therapeutic Targets 19	
1.5.3	The Nucleoporins of Interest: Nup93 & Nup153	19
1.5.4	DamID	20
1.5.5	siRNA	21
1.5.6	Background Summary.....	21
1.6	Methods.....	22
1.6.1	Omic Data Acquisition and Exploration	22
1.6.1.1	Rheumatoid Arthritis GWAS	23
1.6.1.2	NPC-associated DamID & siRNA Identified Genes.....	23
1.6.1.3	RA, OA, OS, BC Gene from Open Targets Platform ⁹⁹	24
1.6.2	Multiomic Data Aggregation	24
1.6.2.1	Statistical Methods	24
1.6.2.1.1	Assessing Independence	25
1.6.2.1.2	Generalized Linear Models.....	25
1.6.2.1.3	Log-Linear Analysis: Modeling associations without defining treatment and response variables	26
1.6.2.1.4	Is this [data] Normal?.....	28
1.6.2.1.5	Comparing Group Means	29

1.6.2.2	Filtering of Genes by Experimental Results	29
1.6.2.3	Integration of the RA GWAS Dataset	29
1.6.2.4	Comparison with Potential Drug Targets	29
1.7	Results.....	30
1.7.1	Exploratory Analysis of the DamID and siRNA Datasets	30
1.7.1.1	Testing for subset independence using χ^2 -statistics	31
1.7.1.2	Log-Linear Analysis: Modeling Relationships between Subsets	38
1.7.2	Integration of the RA GWAS Gene List	41
1.7.3	Integration with Open Targets Datasets	42
1.8	Discussion	46
1.9	Conclusion.....	48
1.10	Supplemental Materials.....	49
4.	Part II. Exploratory Analysis of Hi-C Data for Gene-Gene Interactions	52
1.11	Background	52
1.11.1	Cell lines.....	52
1.11.2	Gene Targets.....	53
1.11.3	Chromatin Capture (Generally Speaking)	54
1.11.4	All-against-all Chromatin Capture: Hi-C	55
1.11.4.1	Overview of Experimental Design & Data Considerations	55
1.11.4.2	Data structure.....	57
1.11.4.3	Bias & Normalization Methods	58
1.11.5	The Exploration of Hi-C Data	59
1.11.5.1	2-Dimensional Visualization	60
1.11.5.2	3-Dimensional Visualization	61
1.11.6	Implications of Current Methods.....	62
1.12	Methods.....	64
1.12.1	Hi-C Dataset.....	64
1.12.2	Preliminary Case Study.....	64
1.12.2.1	Heatmaps	65
	Application of Geospatial Methods.....	66
1.12.2.2	Kernel Density Estimation	66
1.12.3	Pipeline Development	67
1.12.3.1	Data Processing	67
1.13	Results	69
1.13.1	Preliminary Case Study.....	69
1.13.1.1	Heatmaps	69
1.13.1.2	Kernel Density Estimation	70
1.14	Pipeline Development	71
1.15	Discussion.....	72
1.16	Conclusion.....	74
5.	Overarching Discussion.....	75
6.	References	78

Table of Figures

Figure 1.1 Accidental Gerrymandering	1
Figure 1.2 Motivating factors for moving the anchor point for local binning strategies for Hi-C data	3
Figure 2.1 Correlation, Causation and Confounding.....	4
Figure 2.2 Chromosomes included in GRCh38	7
Figure 2.3 Increased inclusion of Hi-C data in multiomic analysis.....	10
Figure 2.4 Combining imaging and sequencing to explore genomic organization	13
Figure 3.1 Multiple systems are implicated in the progression of Rheumatoid Arthritis.....	16
Figure 3.2 DamID is used for chromatin-protein interaction identification	20
Figure 3.3 Overview of siRNA knockdown.....	21
Figure 3.4 Manhattan plot of RA risk factors identified by a trans-ethnic meta-analysis.....	22
Figure 3.5 Log-Normal Distributions	25
Figure 3.6 Intersections of sets of genes as classified by siRNA, DamID, and a RA GWAS meta-analysis	31
Figure 3.7 Intersections of siRNA and/or DamID for Nup93 or Nup153 with Potential Target Genes grouped by disease as identified by the Open Target Platform.....	42
Figure 3.8 Boxplots depicting the distribution of overall association scores for genes in the intersection of the NPC assay datasets and Open Target data, stratified by disease.	43
Figure 4.1 A general overview of Hi-C data collection, including details specific to the Hi-C dataset utilized in this analysis.....	55
Figure 4.2 The relationship between bin size, coverage, and sequencing depth	56
Figure 4.3 Hi-C data is often organized into contact matrix with genomic coordinates represented by axis.....	57
Figure 4.4 Three standard representations of Hi-C contact matrices.	60
Figure 4.5 Hi-C heatmaps visualized alongside tracks.	61
Figure 4.6 Visualizing Heatmaps \equiv 2-D Histograms.....	63
Figure 4.7 Sampling Schema for Hi-C data acquisition	64
Figure 4.8 A different perspective on contact matrices: “One-versus-Any”	65
Figure 4.9 Revision of search queries to reduce search space.....	68
Figure 4.10 Heatmaps of PARD3 interacting with ZNF438 with hierarchical faceting by biological replicate and sampling time.	69
Figure 4.11 The use of traditional heatmaps to depict an alternative perspective of contact matrices: “One versus Any”	70
Figure 4.12 Two-Dimensional Kernel Density Estimation of PARD3 versus ZNF438, faceted by biological replicate and time.....	70
Figure 4.13 Kernel density estimate plots of the interactions between ZNF438 and PARD3 with time “flattened”	71
Figure 4.14 Three-Dimensional Kernel Density Estimation Plots, faceted by biological replicate	71
Figure 4.15 Two-Dimensional Kernel Density Estimation using raw Hi-C data for one biological replicate of G1 of the cell cycle in U2OS cells, stratified by time.	72

That may be the most important thing to understand about humans. It is the unknown that defines our existence. We are constantly searching, not just for answers to our questions, but for new questions. We are explorers. We explore our lives day by day, and we explore the galaxy, trying to expand the boundaries of our knowledge. And that is why I am here. Not to conquer you with weapons, or with ideas. But to coexist... and learn.

-DeepSpace9 s1e1

Ready?

Why do your people always ask if someone is ready right before you're going to do something massively unwise?

Tradition.

-Babylon5:s3e17

1 Introduction

If we could look directly at the contents of the nucleus, the genome, the blueprint of life itself, it would more closely resemble mom's spaghettiⁱ than an organized system of networks coordinating the underpinnings of life. Despite the appearance of disorganization, (most) cells can fulfill their needs either as individuals – such as yeast – or fulfill their roles within multi-cellular organisms — such as us. Given the need for functionality, it follows the organizational structure may be more akin to a refrigerator – the stuff you need fast up front and maybe some food near expiration in the back – as the stochastic drivers of evolution are unlikely to apply the “KonMari Method™ⁱⁱ” to genomic organization. This begs the question, *which component(s) of the nucleus may be “pulling to the front” the information (DNA) a cell requires regular access to?*

Underpinning the work in this thesis is the hypothesis that the nuclear pore complex (NPC) has a role beyond that of transit authority for material traveling between the cytoplasm and the nucleus. In the NPC, we observe potential for an organizational force for genomic information, ensuring the necessary components for cellular function are at the ready. We explore this potential by integrating datasets derived from distinct experimental methods – DamID-seq, siRNA-seq, Hi-C – and a curated database of potential drug target – Open Targets – to predict NPC-associated patterns of genomic organization. The aim of such predictions is to guide the selection of probes for fluorescent microscopy. Doing so would allow the cell-by-cell collection of spatial data into our modeling framework that would allow us to connect the probabilistic inferences made using sequencing data to the point-spread functions at the heart of imaging data. Current methods in bioinformatics were deconstructed to find a path appropriate to our question to go from data processing and integration through analyses. Additionally, we undertook a careful assessment data context – the source (U2OS cell line), how it was obtained, what questions the data were originally collected

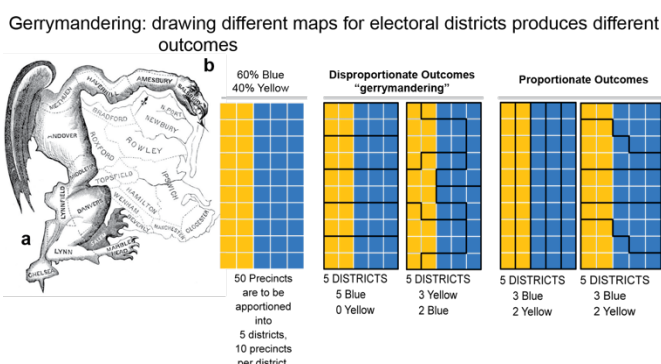


Figure 1.1 Accidental Gerrymandering

The term “Gerrymander” refers to the practice of legally altering the lines of voting districts to dramatically favor the electoral outcomes desired by one party such that said outcome is all but guaranteed. (a) A visualization of electoral districts drawn such that they favor a specific party, i.e. they are “gerrymandered” (top) versus proportionally representing the population of the district (bottom). (b) Although hardly a new practice, the term was coined after then Governor of Massachusetts, Elbridge Gerry, signed a redistricting bill on February 12, 1812. The new districts included one in his own home county of Essex that was of an especially peculiar shape. The infamous cartoon-map depicting a mythical dragon-like beast, the “Gerry-mander” appeared in the Boston Gazette after painter Gilbert Stuart added a head, wings, and claws to the new district map hanging over newspaper editor Benjamin Russel’s desk. It is said he exclaimed, “That’ll do for a salamander,” to which Russel replied, “Better say a Gerry-mander.”

Figures from (a) Klein. *The Boston Globe*. 2011;280(72). (a) M.boli et al. *Wikipedia.org*. 2022.

ⁱ The lyrics to the award winning “Lose Yourself” by Eminem include: “His palms are sweaty, knees weak, arms are heavy / There’s vomit on his sweater already: *Mom’s spaghetti* / He’s nervous, but on the surface he looks calm and ready”. This author was introduced to the nucleus-as-a-spaghetti-bowl imagery during a multiomics course given by EMBL.

ⁱⁱ The KonMari method was established by Marie Kondo and initially popularize by her first book, “The Life-Changing Magic of Tidying Up,” which has been translated into 44 languages, according to [her website](#). The method centers on only keeping such things that spark joy for an individual.

to answer, and so on – an important endeavor for any analysis but particularly so when using externally sourced datasets, such as those we used. For the analysis of “population data,” i.e. the set of transcribed genes, we went “back to basics.” Each gene was considered as an observed member of a population and the viability of statistical methods – as applied to this dataset – were assessed based on test and model assumptions. Elements that are easily (and understandably) taken for granted, such as characteristics of the reference genome, were carefully considered to identify potential confounders they may introduce. Existing tools for the analysis and visualization of Hi-C data were reviewed, both with broad strokes as well as through the lens of our specific objectives. We found that the current state of Hi-C analysis can be likened to [accidental] gerrymandering (**Fig. 1.1**). By establishing the start of binning at the start of each chromosome, the results are biased toward global features. We hence propose “redistricting” to favor local phenomena when investigating interactions between specific loci such as genes. Such redistricting is an essential step toward relating these genomic data from their “sequence” coordinate space into the “physical” space of the nanoscale distances used in microscopy. Returning to fundamentals throughout the process, particularly when faced with roadblocks, is a reoccurring theme of this research project. Knowing why something worked – or did not – was as important as the answers sought. Our method for aggregating gene assay data uniquely considered genes as a population. This approach allowed us to apply methods often employed by epidemiologists, viewing assays as one might view exposures and outcomes. Epidemiologists also have methods for finding the source or sources of an exposure, some of which correspond with those in the geoscience, such as topography. The traditional view of Hi-C is that of astronauts on the International Space Station, the global features of the world seen in their entirety in ways impossible to achieve on Earth. However, the answers we seek require more the bird’s eye view, as they can make out the undulations of the ground, the appearance of predators (or prey) – the local features of the landscape. To this end, our methodology applies topographical methods to seek local patterns of genomic interaction and organization; in this way, we posit that Hi-C data can be used to create a rough map of the gene-level perspective. Going forward, this rough map can be used to inform experimental design for quantitative fluorescent microscopy that will fill in the details. In tandem with our investigation of the NPC as an organizational force in the nucleus, we highlight the utility of publicly available data for preliminary data analyses as well as practical

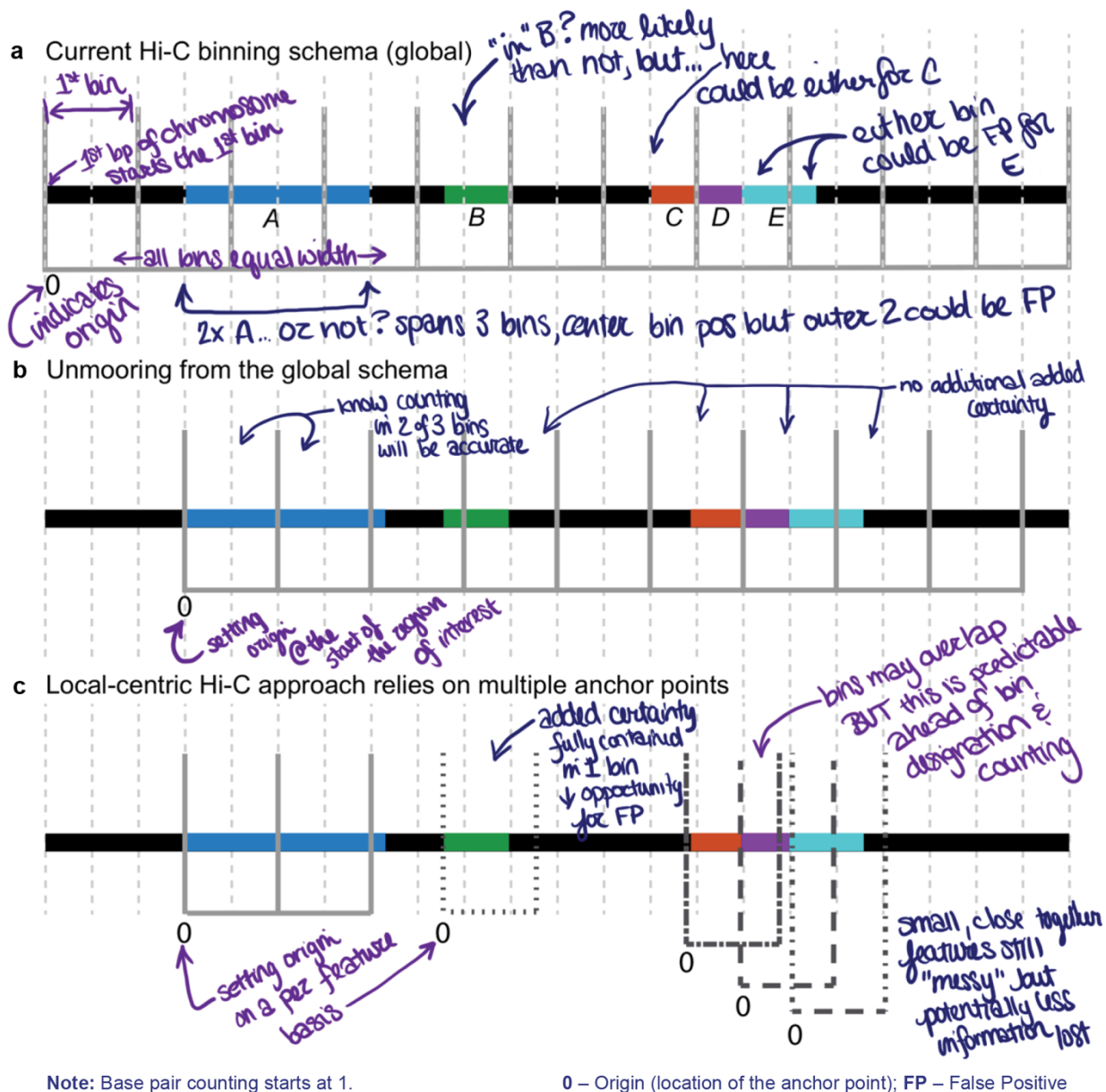


Figure 1.2 Motivating factors for moving the anchor point for local binning strategies for Hi-C data
 Global binning, the standard methodology for Hi-C data analysis, does not account for the position of feature relative to its enclosing or dividing bin(s). Often, the length of the feature relative to the bin is also overlooked – in global binning, the feature IS the bin, rather than a biologically-based feature such as a promotor or a gene. By contrast, the local approach to binning accounts for the biological-feature by setting the anchor point at the start of the feature in question, with the ability to set a slightly broader search window (say, +/- 1kb) if desired. The parameters are subject to the requirements of the question asked of the data. Similarly, the use of global versus local binning depends on the question being asked. For global patterns indicating organizational structure, anchoring the bins at the first base pair of a chromosome makes sense. For relationships between biological features on a smaller scale, feature-location based anchoring is likely preferable.

considerations surrounding its use.

2 Overview

2.1 Don't Take This Out of Context

“Doing science” requires an understanding context. Context is what a high school physics teacher sets by establishing that “magic physics land”ⁱⁱⁱ problems on an exam only respect a specified subset of physical laws. As far as this author can recall, that first taste of context consideration was never explicitly connected to a general need to consider the context in all scientific endeavors – and most of life. In science, the context is the setting in which the phenomenon in question exists – for example, if studying cellular differentiation, the microenvironment in which a cell “lives in the wild” must be accounted for, either through emulation or explained as a study limitation. The experimental context can determine the fidelity of data collection – was data for a behavioral study collected via observations or self-reporting? Historical context can clue a researcher into whether not existing research may be biased and should be excluded, or at least read with trepidation – the prevalence of undergraduate students as sample populations in sociology research or the lack of female subjects in medical research are two examples that come to mind. For those researching entirely in a “dry lab”, it is easy to lose touch with the experimental and larger biological context within which data is collected.

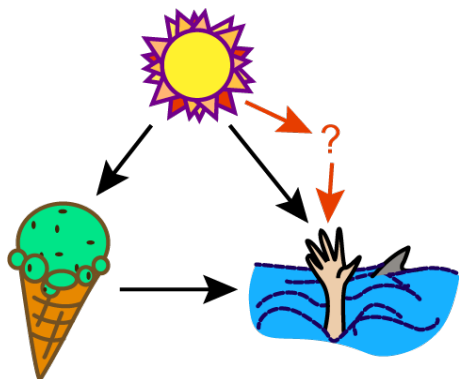


Figure 2.1 Correlation, Causation and Confounding

A classic example of confounding that is often used to illustrate “correlation is not causation.” (black arrows). The temptation is linking an increase in ice cream consumption with an increase in drowning deaths. However, hot weather is a factor that increases the incidents of both, making it a confounding variable. Say one studied the casual effect of hot weather on drowning (red arrows). If you wanted to add complexity, “access to public swimming areas” may be a “mediator”, a variable that is an intermediate step between exposure and outcome.

Considering context can aid researchers in avoiding, or at least accounting for, confounding variables. Confounding variables are those variables that are not measured by the data but affect both the exposure and the outcome. This may lead to misattribution of causality between two variables (**Fig. 2.1**). Confounding variables can result in surprising findings, such as the initial conclusion resulting from the Obesity Paradox.¹ In some studies, obesity is associated with a higher chance of survival. However, this outcome is predominantly seen in situations where patients are admitted to the ICU. Thus, the protective effect associated with survival might be ICU admittance. We say might, as even selecting for ICU patients alone, obesity seems to have a protective. This correlation may be the result of further confounding, such as water retention associated with congestive heart failure leading to a higher BMI that is not caused by a higher percentage of body fat.

Some disciplines touch upon the consideration of context as needed, it will come as no surprise that those pursuing professional degrees in public health are required to evaluate context until it becomes second nature. Such considerations include, but is not limited to, the relevant biological processes, stakeholder analysis and applicable models of health². Models are scrutinized for confounding factors, variables that may disrupt any

causal connections implied by experimental findings. Such variables may be inadvertently overlooked during study design or masked by model selection and interpretation, as may be the case in multiple studies attempting to establish what is, if any, a “safe” blood lead level^{3,4}, a debate

ⁱⁱⁱ This is the semi-sarcastic name this writer’s high school honors and Advanced Placement physics teacher, Eugene Newman, JD, used for a place without conditions such as drag and friction, unless they had been covered at that point in the course.

that continues, in various forms, to this day^{4,5}. All these examples reinforce that looking only the numbers can be very misleading: context matters. With this reflection on context in mind, a significant amount of background information is included within each section. The aim is to promote a well-informed reader^{iv} so they may follow the train of thought of the researchers, understand why decisions were made, and evaluate the results for themselves.

2.1.1 Biological Models

What does this model bring to the table? The good, the bad, the ugly...

Mice are mutated to represent certain conditions, and while we control for what we can, this system is imperfect. The imperfections of this system are discussed further in [Part I > Background](#) in the context of biological models for rheumatoid arthritis.

Overlooked experimental elements – “givens” – can impact the system under scrutiny. For instance, the impact that conditions under which laboratory animals are standardly kept can be taken for granted – until they are not. While caring for study subjects, a researcher noted the bedding used was bothering their respiratory system. Fortuitously, the researcher was studying the effects of prenatal exposure to pollution^v on behavior⁶. This coincidence inspired a new experiment, which compared the particulate matter (PM) generated by commonly used bedding substrates across common cage set-ups. The data indicated that all of the standard bedding substrates exposed rats to a substantial amount of PM_{2.5}⁷. Given PM_{2.5} inhalation⁸ is known to result in a multitude of negative health impacts, the implications of such findings are far reaching. Bedding substrates have likely contributed confounding or mediating variability in the countless prior studies that used rats as a model organism under standard laboratory conditions.

It is unclear to what extent these findings have impacted current and future standards of care for laboratory animals, the points this outcome raises are clear: 1. the biological model can impact study results in both predictable and unpredictable ways, 2. the importance of cannot be understated.

2.1.1.1 Cell lines

Cell lines are commonly used biological models and are the biological model for the data analyzed in this thesis. Therefore, it is prudent to consider how the selection of a cell line can affect the relevance of results. To highlight the importance, let us consider the following example.

To evaluate the expression of three specific non-coding mitochondrial RNA (ncmtRNA) in breast cancer, publicly available transcriptomic data from at least one breast cancer cell line was selected for analysis.⁹ The interest in these ncmtRNA was inspired by the prior findings that identified these transcripts in samples taken from cell lines derived from a multitude of tumor types. Differences in expression were described between the transcripts during G1 versus mitosis as well as when comparing tumorigenic cells to health cells of the same type. MCF-7 is a frequently used model for estrogen receptor alpha (ER α)-positive breast cancer. The association

^{iv} This author hopes this dissertation might be read by some seeking answers to questions she is able to answer, so she does not assume the reader has intimate familiarity with the specific problems or methods presented. That said, she understands and respects that some readers may have such familiarity. To those who wish to ask questions or submit critiques, the author welcome boths.

^v Pregnant mice were exposed to pollution, which was collected in Boston's Chinatown area, in a uniquely designed chamber. After they matured, the behavior of the offspring was observed during tests targeting specific aspects of animal behavior.

For those unfamiliar with the geography of Boston, Chinatown is located next to / below the exchanges between several major interstates (I-93, I-90) as well as a highly trafficked bus and train station (South Station).

between the presence of estrogen and changes in mitochondrial gene expression¹⁰ as well as the stimulating effect on cell division of 17 β -estradiol (E2) binding to E α receptors¹¹ adds additional context to data collected using the MFC-7 cell line. In the aforementioned study, the data selected was publicly available RNA-seq data including E2 treated and untreated conditions and was specifically focused ncRNA, meaning they were unlikely to be filtered out during processing.¹² An additional benefit of using data from frequently used cell line is the increased likelihood additional datasets from the same line will be available for future analysis and/or published results will be available for comparison. As we will see, similar considerations of the biological context as well as data availability were made when selecting and analyzing the datasets used in this thesis. As the amount of available data continues to grow and the construction of data-hungry deep learning models unlikely to subside, the characteristics of individual cell lines may be overwhelmed by the desire to feed the models.

2.1.2 Experimental Methods

Another potential victim of big data is an understanding of experimental methods. One point of big data driven concern is the risk of letting the limitations of research questions that can be answered with the given tools fall by the wayside. This overestimation of the worth of a dataset is a danger when analyzing survey data, for example. There is often a bias introduced by self-reporting on surveys versus collecting information through observation. Depending on the mode of collection, surveys may also suffer from “social desirability bias” – the want to provide socially acceptable answers – which is of particular concern when another person is used as the collect instrument. Luckily^{vi}, cells do not care if they make researchers happy, but the experimental methods are equally as important to consider.

Limitations are a consideration at multiple points in this thesis. For example, the detectable distance between chromatin loci is limited by the size of the cross-linking molecule and any stabilizing forces, such as proteins, to maintain the interaction (discussed in [Part II > Background](#)).

On the flip side, experimental methods are often capable of providing researchers with more information than they are aware. This thesis would do a disservice to its origin in an optics-focused lab if the unrealized potential of microscopy as a quantitative method was left unstated. While photons do have limitations as do detectors and the like, there exist removable barriers to the potential of microscopy data that are too infrequently addressed. Currently, the reported values for microscopy data are in terms of relative percentages; this arbitrary value has no intrinsic value. However, reporting in a truly quantitative way is possible, but requires lowering the barrier to accessing such information. This barrier is surmountable with an aspect of calibration focused on determining the amount of light moving through the microscope and adjusting the reported values accordingly. Though the “behind the scenes” process is more complex, the efforts of the Grunwald Lab have made it easy for individual researchers to report their data in standard units: photons.¹³⁻¹⁷ The hard part is getting everyone on board.

Another aspect of microscopy that has its own set of considerations is sequencing technology. These days, omics data relies heavily on *next generation sequencing* (NGS) in one way or another. What is often overlooked is these data are the product of imaging. Imaging and the processing of those images, the standard “is it a spot” question being asked, is the step between data collection and preparation for sequencing and the sequences output by the sequencer. This reliance on imaging makes sequencers subject to the same potential perturbations as any other optical equipment. Although the impact may be minute, it does need to be acknowledged: sequencing data is not infallible.

^{vi} This is debatable.

The decisions made during the experimental design process as well as during experiments themselves, affect the questions the resultant datasets can be used to answer. This applies not only to the questions data were originally designed to answer but to their reusability in future investigations.

2.1.3 What is your alignment?^{vii}

Alignment sits at the cross-road between data processing and analysis. Unless constructing a *de novo* alignment, this step hinges on a reference genome. The human reference genome has undergone many revisions over the years. The major revisions result in new versions, while the minor revisions are often implemented as patches. The 2009 version of human reference genome, *hg19*, is purported to exclude the mitochondrial genome.^{viii} It was then updated to include an older version of the mitochondrial genome. To make matters more complicated, versions of the human genome are stored in multiple locations, under accession numbers that differ by database, and with different properties. The current, generally accepted reference genome is *Genome Reference Consortium Human Build 38 (GRCh38)* published by the Genome Reference Consortium. GRCh38 is also referred to as *hg38* in the UCSC Genome Browser^{ix} and as version 111 by Ensembl. The most recent patched version is *GRCh38p14*. This ontology is slightly less confusing than the last build – which was named *GRCh37*, but *hg19* by UCSC – and less confusing still than the release in 2009 – which was *NCBI36/GRCh37* while referred to as *3c* by GENCODE and UCSC but as version *56* by Ensembl.

To add another level of complexity, there are subtle differences in the same reference genome between sources, even for “equivalent” genomes, which must be accounted for during analysis. These range from simple encoding issues – UCSC and GENCODE use “chr1” for chromosome 1, while Ensembl uses “1” – to how they are annotated.¹⁹ Furthermore, what is included and the associated identification numbers may differ; for example, RefSeq more stringently defines genes and includes transcript sequences that are independent of the reference genome.¹⁹

To add another layer of uncertainty, making the omics field that much less accessible, the information posted regarding a reference genome is often misleading. For example, the “ideogram” posted by the Genome Reference Consortium does not indicate GRC38.p14 contains

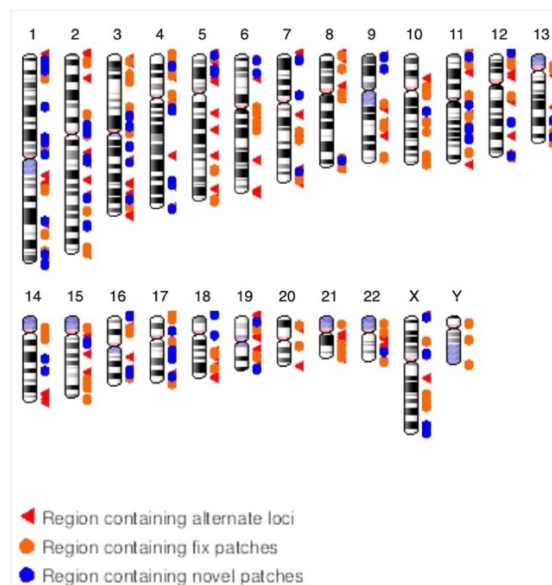


Figure 2.2 Chromosomes included in GRCh38
Ideogram of the current human genome reference assembly, GRCh38.p14, as published by the Genome Reference Consortium.

Image from the Genome Reference Consortium > Data > [Human Overview](#)

^{vii} This is a reference to the moral alignment system used in many tabletop and digital games, perhaps most widely associated with the Dungeons & Dragons, which has made notable changes to its alignment matrix when new editions of the main rulebook are published.¹⁸

^{viii} Validating this assertion has proven difficult. The original release notes for hg19 are nowhere to be found, which seems to be a result of NCBI’s data restructuring efforts. This author did have to add back the mitochondrial genome to the hg19 version of the transcriptome for prior work on the mitochondrial genome (through 2021) and it has been widely discussed on bioinformatics forums such as [BioStars](#), but a formal indication thereof was not to be found.

^{ix} This is not an “official” name for the build.

the mitochondrial genome, never mind which version it (might) include (**Fig. 2.2**). Determining what a reference contains requires an understanding of its origins as well as accepting that the information available at first glance is likely not the full story. As one goes beyond recognized, all-inclusive reference genomes into the world of such things as transcriptomes, the complexity increases. At this point, considerations such as the sequencing source (direct RNA sequencing versus creation of a cDNA library) as well as what “ome” used to align (or pseudoalign²⁰), genome or transcriptome, is dictated by the tool used.

In terms of what is included in a pre-processed, pre-aligned dataset, even when it is known exactly which genome it was aligned to, that is not the full picture of what may be found within the dataset, experiment aside. Often rRNA reads are removed either during the experimental protocol or by filtering during sequence alignment. This is a reasonable procedure, they are so prevalent that their presence may dilute the pool so much so some transcripts may elude detection, this solution can be problematic if unknown to someone repurposing data to answer their own research questions. Though a well-seasoned bioinformatician may be aware of this data limitation and know what to look for when selecting a dataset, such considerations are not always obvious. Even the specific way and order of filtering can affect results.²¹ All that to be said, yet again, context matters. We will see another example of the importance of reference genome in [Part II > Methods](#).

2.1.4 Analytical Methods

During the data analysis phase of any experimental investigation, there lies a minefield of conventions that are often accepted without examination. Frequently used statistical methods rely on underlying assumptions about properties of the data such as the distribution. For example, the *t*-test is often applied in biology to look at the difference of means between samples; among its assumptions is samples follow a normal distribution and have homogeneity of variance.²² Statistical tests are often incorrectly applied, poorly reported, and the results are frequently misinterpreted or misrepresented^{23–25}; for example, odds ratios and relative risk^x are often interchanged.²⁶ Measures of statistical significance may make or break the publication of an article though they may not carry the weight attributed to them; the *p*-value is a well-understood yet maintained instance of an overvalued make-or-break measure of significance.^{27–29} For the methods presented in this thesis, we strove to explain our decision-making process with respect to the methods employed. This view into our process includes the deluge of contingency tables in [Part I > Results](#) as well as addressing the anchor point problem in [Part II](#).

2.2 Speaking of Context... The Nuclear Pore Complex

There are multiple theories for how eukaryotic cells evolved to possess nuclear envelope – such as the membrane curving and invagination theory³⁰ – some with the “why” – such as the slow process of mRNA splicing³¹ – muddled in along the way. Either way, it did. And embedded within it, so too did the nuclear pore complex (NPC) evolve.

Though the NPC differs in composition between organisms, the individual nucleoporin (Nup) – protein subunits of the NPC – are conserved to varying degrees between species. The Nups under scrutiny in this work, Nup93 and Nup153, are conserved, Nup153 being highly conserved, potentially due to the evolutionary pressure of nucleic acids interactions.^{32–34}

^x An *odds ratio* compares the number of individuals who experience an event to the number who do not. The *relative risk* compares the number of individuals who experience an event to the total number of people who were at risk of experience the same event, including those who did.

2.2.1 Generalities

The nuclear pore complex is known for its role as transport authority between the cytoplasm and the nucleus. Contrary to what rudimentary biology textbooks would have one believe, the NPC is not a gaping hole through which molecules may travel freely³⁵; it has been suggested that individual NPCs preferentially transport types of cargo.³⁶ The density of NPCs in a given cell type is consistent as long metabolic state is consistent between cells.³⁷⁻³⁹ However, despite the contributions of surface area, volume, and DNA content to a cell's total pore number (assuming a consistent metabolic state) these three parameters are not sufficient for predicting the number of nuclear pores a cell may have.⁴⁰ What these relationships do indicate, however, is that there exists a link between cell identity and the NPC.

2.2.2 And who? ...are? ...you?

Given that metabolic state can change the NPC number in a cell as well as the factors established above, it is reasonable to hypothesize that the number of NPC is driven by cell type and need. The need is established based on external conditions as well as the ability of the cell to meet those conditions. Looking at the logic of this argument from the other direction, the laminar structure underpinning the nuclear envelop provides structural stability whereas certain chromosomes preferentially position themselves near the nuclear periphery in a cell-type dependent manner.⁴¹ These findings together suggest that NPCs are linked to cell identity. The (general) predictability of their number alongside their location along what is a stable structure in the cell indicate the NPC as a good reference point for spatial organization of the nucleus. To add to this, there is a growing pool of evidence supporting the NPC is a driver of cell identity – the binding of NPC components to cell identity linked super enhancers⁴² – as well as the influence specific nucleoporin have on development and disease.⁴³⁻⁵¹ Physical association with components of the NPC driving cell identity, development, and being linked to disease further supports the use of the NPC as a spatial reference point in the cell as the location of genes in proximity to the NPC in a cell-type dependent manner can be verified through fluorescent microscopy. Once this link is established, a coordinate system on which to modeling the 4D architecture of the nucleus can be established. Deep below, foundational to perhaps, our underlying aim of predicting and designing probes for genes interacting at the nuclear pore complex, is the hypothesis that the nuclear pore complex is an ideal reference point for a coordinate system that links 1-d sequencing to 3- and 4- physical models of genomic organization.

2.3 Considerations for Multiomic data Integration

Multiomic data integration has more than one purpose and those purposes remains specific to the research questions that are being asked. To fully model an entire biological system alongside the environmental pressures required to adequately simulate its function would require resources and methods that we currently do not have.^{xi} An addendum to the saying, “all models are bad but some are useful,” might be that no models can encompassed every aspect of a system and useful models are able to give insight into some questions.

^{xi} It also starts creeping us as a society a bit closer to testing some suppositions of arguments for and against simulation theory, which is most certainly out of scope for this thesis and far above the paygrade of this author.⁵²

The most familiar way to model data is often taught – likely by “eyeballing” a line of best fit instead of using the statistically accepted “least square method” – in Algebra I^{xii}. In the jumble of confusion caused by finding such a line’s y -intercept and calculating its slope, the notion that these values are connecting a “most likely” outcome (y) with a given condition (x) is lost. This concept of a model does not stray far from the results of studying the relationship between enzyme activity and the concentration of its substrate, the dose-response relationship is modeled and visualized in 2D with a logistic curve.⁵³ However, some types of modeling seem to stray further from the familiar, generally due to the number variables involved.

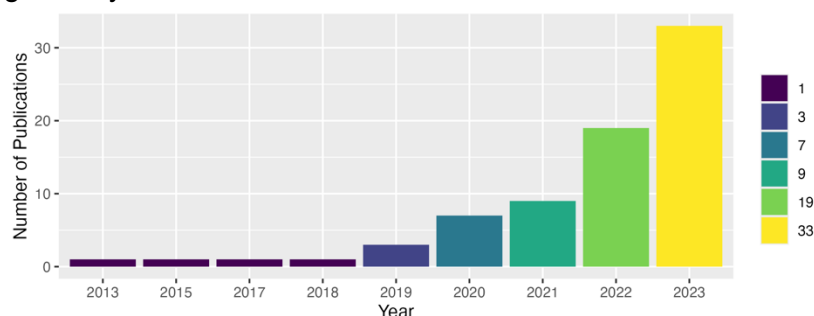


Figure 2.3 Increased inclusion of Hi-C data in multiomic analysis
The results of a PubMed search for “Multiomic” + “Hi-C” publications by year from 2013 through 2023 show a rise in publication every year after 2018.

Multiomic data has an excess of variables, frequently more than it has samples. Still, some applications of these data seek the same input/output model; for example, the association between the status of various measures (inputs) such as gene expression (RNA-seq) and chromatin accessibility (ChIP-seq) may be used to model outcomes such as cell identity.

The number of publications containing “Hi-C” and “Multiomics” has increased by an order of magnitude between 2019 and 2023 (3 to 33) (**Fig. 2.3**). This count is likely an underestimate, as the search did not include any umbrella terms that encapsulate Hi-C such as “chromatin capture,” nor names for any Hi-C adjacent technologies. Looking at these numbers, it is clear that understanding methods for data integration that include Hi-C and Hi-C-like datasets will be a necessary tool for bioinformatics research in the future. The sheer size of Hi-C data, along with its inherent sparsity, requires it be handled with an awareness of its limitations as well as the limitations imposed upon it by current analysis methods. Establishing our own awareness of these limitations as well as developing a new method for exploring Hi-C data is the major undertaking described in [Part II](#).

2.3.1 The Need for Metadata

The “*in silico lab*” researchers are not often part of the data collection processes; often, they are only called in after the data exists to be analyzed. This post hoc involvement is inherently true when analyzing pre-existing, publicly accessible data. In this situation especially, metadata is necessary for success. Metadata places data into context, from sample origin to processing methods to instrumentation used. Contemplating the many reasons why the inclusion of metadata is advantageous brings to mind a frequently uttered piece of advice regarding coding, “comment.” Anyone who needs to use your code, including future-you, needs to comprehend the flow without reinventing the wheel. Similarly, metadata has the potential to convey what was done, how it was

^{xii} Secondary school math curriculum in the United States is a topic that can be either a topic of heated debate or swept under the rug by the same folks that scoff and blow off any truthful response this author gives to “what do you do for work,” with “well, I was never a math person anyway.” It is the opinion of said author that the latter would be a less frequently hear response if issues in the former, which are really issues that start in primary school mathematics education, were ever addressed in a sensible, systematic way that fostered an understanding of numbers, their relationships, and perhaps presented them through guided discovery.

done and under what conditions, both for internal and external use and re-use of a dataset when it is properly implemented.^{54–58}

Beyond use and re-use, having metadata accompanying a dataset makes comparisons between studies more plausible; it is more clear what points of comparison exist to make. To go a step further, as multiomic approaches are being widely adopted, metadata informs the decisions regarding which datasets can be integrated and what steps, if any, must be taken to resolve any discrepancies. An example of this will be discussed in [Part II > Methods](#).

Alongside use and re-use by external researchers, lies *reproducibility* – defined as “strictly computational reproducibility” by the National Academy of Sciences, Engineering and Medicine.⁵⁹ Assessing reproducibility (as well as replicability) is crucial for “doing science”, which is addressed (in theory^{xiii}) through peer review, a process currently subject to scrutiny and redesign.^{60,61} Unfortunately, reported methods in omic^{xiv} datasets, such as RNA-seq⁶², tend to lack detailed enough methods, which hinders reproducibility. To address this issue, the various subdisciplines of the biological sciences are making strides the adoption of metadata standards, often through community-based initiatives.^{13,54,55,63–67}

Though it may seem burdensome at first, recording parameters may inspire better data collection methods, leading to a higher degree of accountability. To alleviate a fair portion of the cognitive load and reduce the strain of adopting such practices, tools such as the Micro-Meta App¹³ have been created. Alongside metadata is the need for tracking data provenance. While it may initially be another burden on the researcher, a shift in societal norms would help the process move along. Can you imagine what would happen if researchers were as revered for well-produced datasets with high-fidelity metadata as they are for publishing in “the” journals with a high impact factor?

The need for continued improvement of metadata practices is evidenced by the amount of data sourced externally. The demand for publicly available data for re-use/re-purposing is evident; as of writing this sentence (14-February-2024) 31878 citations were listed by NCBI GEO as third-party usage citations.⁶⁸ When posting processed data alongside raw data, the recommended practice is to post the processed data in the form used for analysis in the corresponding journal article.⁶⁹ Several experiences described in the sections that follow highlight the sorts of issues that arise when such best practices are not followed.

2.3.2 Application of Emerging Technologies: Metadata Strikes Again

The pull of deep learning (DL) for research in biology is its ability to capture linear and non-linear relationships, which is idea for the layered nature of biological data. DL uses hierarchical feature extraction, which frees itself from reliance on a chosen kernel function for the “kernel trick” that many machine learning algorithms use to simplify solving the complex equations.^{xv} Let us assume that creating a viable model is possible in the first place, which is not a proposition without its own issues in this space.^{xvi} There is a concern surrounding the “black box” nature of DL models. These models may be able to successfully classify biological data, but the connections between

^{xiii} As a child, I thought that experiments were done by one group of scientists, written up, and then redone by another group to confirm the results. This belief was enforced by the “whiteboarding” of methods and results for discussion between groups in my physics courses (I was lucky enough to have modeling-based instruction). Finding out otherwise was an unpleasant paradigm shift, worse even than realizing my health records weren’t something that could easily be accessed by doctors, no matter where I may get injured and need care.

^{xiv} For the purposes of this dissertation, omic(s) and multiomic(s) includes but is not limited to sequencing and imaging.

^{xv} Machine learning algorithms solving complex equations need to raise the dimensionality of a model to accommodate nonlinear relationships between variables, which leads using the kernel trick.⁷⁰

^{xvi} Biological data, particularly omics data, tends to have high dimensionality – a high number of variables, which is not necessarily bad for DL – and low sample size – this is where the problem lies.

biological processes may not be discernable by humans given the complexity of DL-derived models. This mismatch between tool and objective raises the question, is it possible to extract a simpler model from these complex models—a “good enough” model that can function practically? The want for AI model transparency necessitates metadata, as explained by Pscal Heus of Postman Open Technologies in a recent editorial; he used HuggingFace’s [Model Cards](#) – which include datasets trained on and metrics evaluated on – and [Data Cards](#) – which include information about data content and context – as exemplars of potential solutions.⁷¹ What this means for us is that we, as researchers in the biologically adjacent sciences, must take heed of the precedent set by responsible^{xvii} purveyors of AI models and follow a standard of metadata for reporting both our data and our results.

2.3.3 Finding a Common Language in Probabilities

One of the (many) barriers to multiomic data integration is the lack of common language between datasets. This absence is not in reference to the ontological hell that the multitude of databases and disciplines put everyone through, though several instances of this word salad are discussed within this manuscript, particularly within [Part II](#). The common language that is lacking is a way of quantifying the exposures and outcomes that we want to model using compatible units. We put forward probabilities as the common language that will allow the integration of data types. Conveniently, probability already exists as the backbone of machine learning.

Probabilities are similar to proportions; it is helpful to consider that probabilities are to time as proportions are to space. We as humans are not as inherently good at thinking in probabilities – our brains have evolved to be risk sensitivity, ascribing a higher risk to things with lower certainty and *vis versa*⁷² – but the world we live in is well represented by probabilities.^{xviii}

Let us first consider modeling with proportions, which is reasonable to consider in terms of baking. In baking, proportions allow scaling up and down – cooking a meal for a family versus catering a 100- or 1000-person event – or different products by using different proportions.^{xix} Moving into probabilities: it may be that the probability of it raining (outcome) is predictable given the probability of being in a tropical climate (predictor) and the probability of it being monsoon season (predictor). In this imaginary scenario, a model could be constructed to see how the change in the probability of each predictor altered the probability of it raining. The rain would not ever be guaranteed, but its potential could be demonstrably higher depending on the predictors.

Is this an oversimplified example? Perhaps.

It does illustrate what we hope to achieve by moving the results of sequencing experiments, starting with Hi-C, from the realm of counts, frequencies, and even proportions, to the language of probabilities. We then seek to use such transformed data, now translated into the language of probability densities, to predict the position of point spread functions. As one of few labs that expresses imaging data using probabilities, it is fitting that we travel further down the path, [linking sequencing data and imaging data within the same model](#).

^{xvii} This word is written with a notable amount of trepidation.

^{xviii} The author is resisting the urge to revisit her undergraduate thesis on free-will versus the existence of initial conditions, within which she found some haunting philosophical issues.

^{xix} This is entirely unlike Sims4, where the same cooking animation using the same ingredients results in many different cuisines.

2.4 This is the Way.^{xx}

Achieving our long-term goal of creating a multi-omic model of genome organization through the combination of imaging and sequencing data – enabled by translating both modalities into probabilistic language as well as using the NPC as a spatial reference point – will require quite a bit of mise-en-place^{xxi}. The research presented within this thesis is part of that preparation.

Fluorescent microscopy requires fluorescently labeled targets for them to “show up” in imaging data. Therefore, targets for labelling must be identified and a means of labeling them established. With respect to labeling the NPC, the Grunwald Lab has labeled the NPC by infecting cells with lentivirus carrying a tandem Tomato linked POM121 (POM121-tdTomato) in previous experiments.⁷⁴ As our goal includes using the NPC as a spatial reference point, we will target genes that have come into close proximity with the NPC. For this purpose, we use DamID data as an indicator of a gene having come into physical contact with either Nup93 and/or Nup153, with genes that have positive DamID for both nucleoporin having the strongest evidence supporting physical association with the NPC.

In [Part I](#), we compare sets of genes based on their classification using pre-existing DamID and siRNA data for Nup93 and Nup153. This data was collected for a study that connected the NPC via these two nucleoporin to super enhancers⁴² identified as driving cell identity related gene expression.⁷⁵ This was done to explore the connection between cell identifying genes, which are based on cell type, and the diseases that a specific cell line is used to model via potential drug targets as identified by the Open Targets Platform. We hypothesized that if genes indicated as changing expression after knockdown of the two nucleoporin and/or physically interacting with the same nucleoporin were associated with cell identity, these indicators are also linked to disease state.

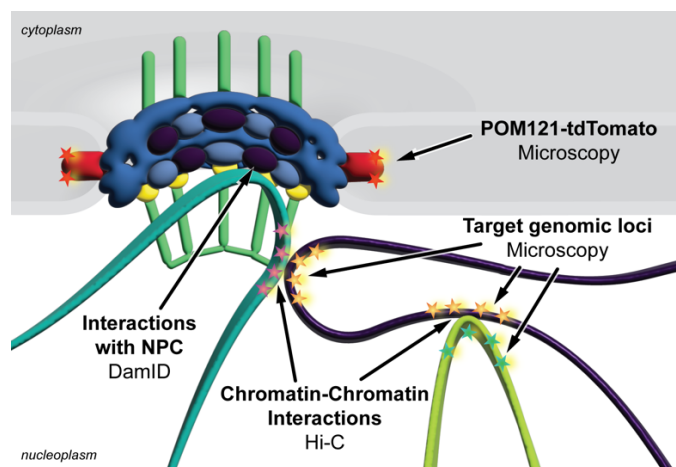


Figure 2.4 Combining imaging and sequencing to explore genomic organization

Sequencing technologies can be used to guide microscopy experiments by narrowing down *where to look* and *for whom to look*. Regions of chromatin that interact with the nuclear pore complex (NPC) can be identified using DamID, giving a

^{xx} “This is the Way,” refers to the Way of the Mandalorian and is stated when orthodox members of their society follow their ideals.

^{xxi} “Mise-en-place” is a culinary term meaning everything is prepared and “in its place”. This skill is undervalued by culinary students but found of notable worth by graduates in the workforce.⁷³

combination of where (NPC) and for whom (identified region). Chromatin-chromatin interactions can be identified by chromatin-capture technologies such as Hi-C. This indicates for whom (the interaction regions) and the where can be established as local to the NPC if one of the interacting regions has been identified as interacting with or near a region that has been identified as interacting with the NPC via DamID. Regions are then targeted for fluorescent labeling (stars), whether by FISH in fixed cells or CRISPRainbow in live cells, and imaging elevates linear sequencing information into three – spatial for fixed cells – or four – spatiotemporal for live cells – dimensions.

Having established a set of genes that have been in proximity to the NPC, we will then determine which genes have physically interacted with those genes interacting with the NPC (**Fig. 2.4**). Our initial work in identifying these genes was done using previously processed Hi-C data that has been binned at a resolution of 5kb. The details of this portion is not included within this manuscript. However, it is mentioned here to draw a connection

between the results of our research that utilizes DamID and the method of exploratory analysis of Hi-C data presented in this thesis.

The need to select regions for fluorescent labeling motivates our development an exploratory analysis method for Hi-C data. After a preliminary analysis of pre-processed Hi-C data, we sought a quote for the probes necessary to label the length of two genes. The total cost was in the hundreds of thousands of dollars for the full custom probe set.⁷⁶ As such a financial investment was not practical, we sought an alternative. We considered the same data from an alternative perspective: the regions that we wanted to create probes for should be those most likely to interact as the interaction is what we want to image. Assuming that the areas with the greatest interaction frequency in existing data were indicative of the location the greatest chance of interaction in the future, we looked for a way to identify these areas of higher probability density. Additionally, we sought to circumvent the anchor point problem presented by chromosome-based binning as well as normalizations methods that hinge on global statistics for an approach that supports exploratory analysis of Hi-C data at the gene level. In [Part II](#), we hypothesize that kernel density estimation (KDE) can be used for exploratory analyses of gene-level features. We demonstrate that transforming count data into probability densities using KDE favors local conditions over global considerations when performing exploratory analyses.

3 Multiomic Integration for Exploring Therapeutic Gene Targets

Next generation sequencing dramatically expanded the territory available for exploration in many fields, including potential therapeutics. Researchers are no longer constrained to observing a select few variables upon perturbing a biological system with treatment or comparing drug response between cell types. Anyone's who's stared at a [Roche pathway poster](#) for long enough can appreciate the potential downstream effects of altering any one of the processes tasked with maintaining homeostasis, never mind the potential for a cascade off-target effects.

Regardless of the end product, RNA or protein, the manifestation of DNA encoded information is at the heart of this biological network. The mechanisms for this data transmission do not come with a manual, yet a healthy cell promotes and suppresses expression of genes according to its needs. To do so, the cell not only must have means of information conversion, but ways of denoting what to convert.

Genes are only available for transcription if their promoters are accessible to binding proteins such as transcription factors. Subsequently, chromatin condensation and other means packing the genome within the nucleus are a means of controlling gene expression. Genome organization

does not end with configuring everything such that it not only fits in the nucleus, but the “right” genes are accessible for transcription. Consider this: what good is organizing a kitchen if every time it is done, everything is put back in a different place? It may look good from the outside to have all things neatly stowed in cabinets, but it makes cooking inefficient to have no regularity to where the pans can be found. Likewise, the nucleus benefits from having some regularity in the spatial location of genes that are expressed – particularly for those genes with transcription factors in a predictable locations. Given gene expression is linked to cell identity, it follows that cell identity would drive at least some aspects of genome organization.

Nuclear pores are posited to have a key role in cell identity, implicating them as principal organizers of the nucleus as evidenced by patterns of physical contact with chromatin that differ by cell type. Given diseased cells often have different patterns of gene expression than their healthy counterparts, we propose that these patterns will alter the interaction of chromatin at the nuclear pore complex in ways that support such a change in gene expression. If the genes at these loci are identified, they may indicate novel therapeutic targets or increase the priority of investigating suspected therapeutic targets by providing supporting evidence of their relevance.

3.1 Background

The work herein began as an (essentially) back-of-the-napkin comparison between a list of RA associated genes and a list of genes that are thought to associate with the nuclear pore complex as well as cell identity in U2OS cells. From there, the study unfolded as contextual knowledge expanded and a more thorough interrogation of data began. In this section, the reader is presented with contextual information both regarding the disease, the molecular key players, and datasets in question.

3.1.1 Rheumatoid Arthritis

Rheumatoid arthritis (RA) is an autoimmune disease most commonly known for the chronic pain suffered by those afflicted, a group to which 0.6% of the US population belong.⁷⁷ Of the affected, the prevalence is higher among women, with men experiencing later onset, leading to a closure of the incidence gender gap as age increases.⁷⁸

RA causes chronic inflammation of the synovia joints, leading to joint destruction and systemic complications^{79,80}. The progression of RA is orchestrated by cell types from multiple body systems such as B-cells, T-cells, and osteoblasts (**Fig. 3.1**). This complexity makes the disease particularly hard to model in the laboratory. Moreover, it also makes it a computationally complex system to model *in silico*, although omic data for an increasing number of cell types and evolving computational tools make developing such a model more realistic.

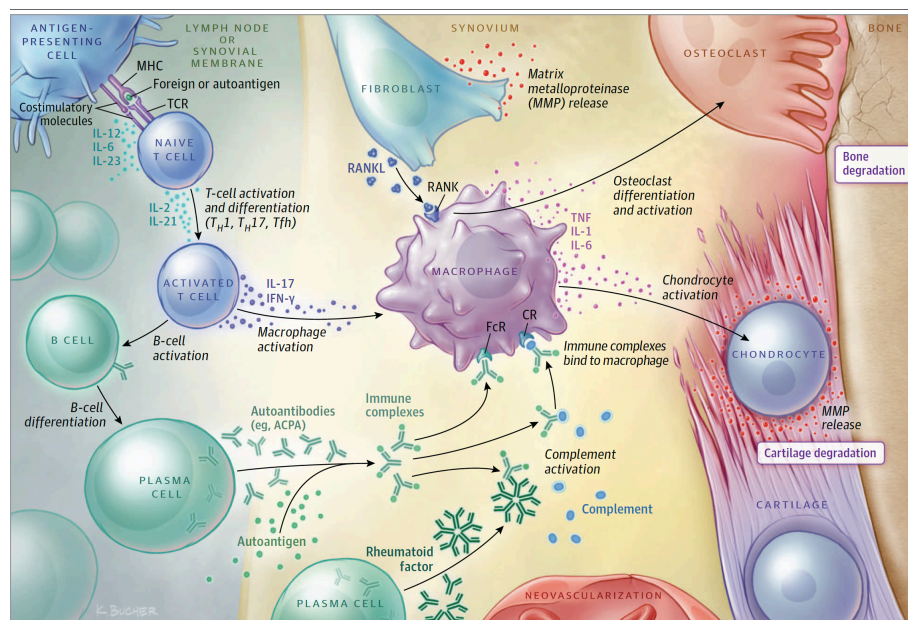


Figure 3.1 Multiple systems are implicated in the progression of Rheumatoid Arthritis. Rheumatoid arthritis is a complex disease involving multiple organ systems and a plethora of cell types. This makes it a difficult disease to model. The complexity of the intercellular signaling cascades shown is “only” the manifestation of complex interaction networks within each cell.

From Aletaha & Smolen. *JAMA*. 2018;320(13). doi:10.1001/jama.2018.13103

citrullinated proteins become epitopes recognized by autoantibodies, known as anticitrullinated protein antibodies (ACPA)^{xxiii84}, in the pre-clinical stages of RA. Subsequently, the recruitment of immune cells is triggered leading to an increase in inflammation in the area. Such RA-specific autoantibodies can be detected in laboratory tests, though percent of RA patients expressing a given autoantibody depends on the population and the specificity and sensitive, as well as the feasibility, depends on the biomarker indicated by the test.⁸⁵

Looking at the PADI association alone, the multitude of players in the pathway imply there are many entry points for the disruption of this system, leading to the development of RA. Yet, this is not the only pathway through which RA can manifest, it is ever more complicated than that. The components of the human leukocyte antigen (HLA) system, more broadly known as the of the major histocompatibility complex (MCH)^{xxiv}, are encoded in the short arm of chromosome 6

^{xxii} PADI is also referred to as PAD in RA literature. Investigation into “PADI” versus “PAD” uncovered the latter’s use as an abbreviation for peripheral artery disease, potentially instigating the adoption of PADI. The prevalence of PAD in patients with RA⁸¹ may have contributed to a want to decrease confusion, particularly as the use of search engines for literature review grew more prevalent.

^{xxiii} Also referred to as Anti-cyclic citrullinated peptide antibodies (CCPA) in literature describing the diagnostic use of the level of these antibodies⁸³ as a better biomarker for RA than Rheumatoid Factor (RF). However, in studies that use this measure, ACPA appears to be standard.

^{xxiv} Broadly in a phylogenetic sense. Because the RA risk associated genes are indicated with gene symbols starting with HLA, both terms are introduced here. The association of the acronym “HLA” was assigned *post hoc*. The acronym was initially a compromise reached at a WHO conference regarding the naming conventions proposed by two different labs (“Hu” vs “LA”), both having studied leukocyte antigens.⁸⁶

One of the sources of complexity for RA is the propagation of changes that are incited by an alteration in a single protein. For example, one of pathways leading to RA involves the citrullination of proteins, a process instigated by tissue specific isoforms of peptidyl arginine deiminase (PADI^{xxii}); PADI2 and PADI4 are two isoforms expressed in cells associated with RA joint inflammation⁸². Citrullination plays a key role in development with PADI levels increased in cells undergoing terminal differentiation or apoptosis. These

(6p21).⁸⁷ This region is thought to contain 30-50% of the genetic risk factors for RA^{xxv}. However, this figure has conflicted with results from a twin study, which cites potential prior misattribution to potentially protective genetic factors in the MCH⁸⁸ for the discrepancy.²⁶ The MCH also includes a class II MHC gene expressed only on B-cells, activated T-cells, and antigen-presenting cells, HLA-DRB1 which has *shared epitope* (SE) alleles. These alleles have a 5 amino acid sequence motif found in the HLA-DR β chain region⁸⁹. Individuals that are SE-positive are observed to be ACPA-positive more frequently than those who are SE-negative. The ACPA-positive phenotype has not only been associated with higher rates of joint destruction, it is seen in two-thirds of patients diagnosed with RA.⁸⁴ RA symptoms arise through the interaction of a variety of genetic and environmental factors⁸⁰, with environmental risk factors differing between ACPA-positive and negative RA.⁸⁴ In the words of the epidemiology Erin Welsh, “It’s complicated.”^{xxvi}

3.1.1.1 Biological Models

Complicated multi-system processes result in trade-offs when it comes to selecting experimental models for research. RA is not at all immune to this issue. The polygenicity of the condition is a significant hurdle to modeling. It is a condition found in few animals used in research.

Transgenic mice models provide options for studying specific aspects of disease progression or drug candidates but the scope of use for each mutation is a limited. RA-like conditions can be induced in mice and rat models. However, their use provides an incomplete picture. For example, induction via methylated bovine serum albumin (mBSA) can induce delayed hypersensitivity (DHS) arthritis in C57BL/6J strain mice, which shares features with RA.⁹⁰ However, models such as this one do not allow researchers to follow the chronic disease progression as it occurs in humans nor do such models lend themselves to studying the full “omic” underpinnings of the disease. Non-human primates (NHP) are also used to model RA through induction, which is slightly less limited given phenotypic similarities. One NHP, the macaque, may have an additional advantage as a model; not only does the macaque share 93% genomic similarity with humans but has been known to develop RA naturally with age.⁹¹

Tissue models of RA are possible, and are utilized, they are not ideal. The complexity of RA goes beyond cell type into anatomic structures such as synovial capsules, joint cartilage, vascularization, and so on. These factors are beyond what (co)culturing tissues can provide while *ex vivo* cultures come with decreased longevity, small sample size, and the quagmire of donor specific conditions. There is, however, hope that state-of-the-art technologies such as lab-constructed 3D multi-component joints⁹² and microfluidic “organs on a chip”⁹³ may be further developed and applied to studying RA.

Studying omics requires a source of cells, whether that be patient samples, tissue cultures, or cell cultures. Patient samples come with a degree of variation that may not be easily balanced by number of specimens, at least not in the short term. The potential for gathering omic data from biopsies and aggregating the data over the long term is a possibility. As mentioned above, tissue

^{xxv} This statement was made in a paper citing another source, and the citation was made without reading what that source paper actually found: this percent was generally accepted, but the twin study presented did not agree with said generally accepted numbers, for which there was a citation. The author gave several possible sources for the discrepancy, ranging from a difference in analytical methods to potential protective factors that had been found in the same region.

^{xxvi} *This Podcast Will Kill You* hosts Erin Welsh, PhD and Erin Allmann Updyke, MD, PhD frequently acknowledge the complexity of disease processes and, because Leishmaniasis ex the complexity of host, pathogen(s) and unknown factors in predicting disease outcomes, used the phrase for the title of the s4e62 episode on Leishmaniasis. <https://thispodcastwillkillyou.com/2020/12/15/episode-62-leishmaniasis-relationship-status-its-complicated/>

cultures also come with their own set of downsides. This leaves cell cultures, which are both at an advantage and disadvantage with their simplicity. There is an ease-of-use factor with cell cultures that is not present with tissue cultures, and the genetic variability is (in theory) decreased. However, commonly used cell lines for modeling RA are limited in the information they present and are derived from cells that are often derived from tumorigenic samples, which may bias that limited information.

For example, the U2OS cell line originates from an interosseous biopsy of an osteosarcoma of the tibia removed from a 15-year old female patient^{xxvii} in 1964.⁹⁴ Straight away, there is one obvious source of bias: any Y-chromosome associated genetic variants or any limitations of having only one allele for X-chromosome genes will not be found utilizing this cell line.^{xxviii} U2OS cells are tumorigenic when injected into immunodeficient mice, producing cells that are suited for modeling immune attraction based on follow-up xenografts.⁹⁶ Discrepancies between the molecular profile of U2OS cells and that of mature osteoblasts have been observed – this includes a lack of markers such as osteocalcin (OC) or decorin while collagen types II, IV, IX and X, are present – potentially meriting classification as fibroblastic as well as osteoblastic.⁹⁷ In a more recent study, U2OS cells were induced to differentiate into adipocytes but failed to differentiate in the osteoblasts or chondrocytes; however, tumors induced through subdermal injection of these cells in mice were shown to produce “abundant” osteoid (Os).⁹⁶ Furthermore, the U2OS cell line has been notorious for polyploidy, which can vary between passages, since its inception.⁹⁴ Despite its imperfection as a model for RA, this is still a cell line utilized to this end.

Although there are many hurdles to modeling RA, various omics methods and the integration thereof are driving current RA research. The focus of governmental funding from the National Institute of Allergy and Infectious Disease (NIAID) as well as the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) has recently included efforts to leverage omics data to model RA and other autoimmune disease.⁹⁸ Results of this have included advances in multimodal single-cell techniques to model the dynamics of HLA gene regulation, including cell-type-specific expression quantitative trait loci (cis-eQTLs) in multiple cell types.⁹⁹ It has also led to the identification of dynamic regulatory elements linked to cell-state, specifically states associated with autoimmune-driven inflammation, that may be indicative of heritability.¹⁰⁰

3.1.1.2 Genome-Wide Association Studies for Rheumatoid Arthritis

As models catch up, efforts to determine the genetic variants responsible for RA continue. Lead by Kazuhiko Yamamoto, M.D., Ph.D., The Laboratory for Autoimmune Diseases at the [RIKEN Center for Integrative Medical Sciences](#) (IMS) has conducted genome-wide association studies (GWAS) for RA risk factors, each building on the lab’s prior work. In their meta-analyses, the group aggregated the results of multiple GWAS including studies sampling from ethnic and trans-Asian populations.¹⁰¹ One of their findings speaks to the difficulty of finding a comprehensive animal model for RA: approximately 80% of RA risk variants occur in non-coding regions.¹⁰² That finding in combination with the observation that associated SNPs (single nucleotide polymorphisms) are often highly correlated with a larger number of variants¹⁰¹, the GWAS was unable to conclusively identify a causal gene or causal variant. This further substantiation of RA’s standing as complex

^{xxvii} The procedure that made the sample available was an amputation of the young woman’s left leg. Unfortunately, the cancer had already metastasized, and she passed away eight months later. It is important we acknowledge she was an individual as well as her contribution to science.

^{xxviii} There exists a non-negligible potential for genetic sex to influence disease states, including cancer. As it turns out, this might negatively impact the understanding of disease progression in men.⁹⁵ This leads to the question, how many cell lines are derived from genetically female cells and what impact does that have when translating findings from cell to male murine model to then the human male in clinical trials?

condition stemming from a polygenic, or multigenic – either or both? – is compounded by the impact of environmental factors such as smoking or infection.⁷⁸

3.1.2 The Open Targets Platform > Systematic Aggregation of Potential Therapeutic Targets

Finding the genetic underpinnings of a condition such as Rheumatoid Arthritis is only the first step in translational research. One of the practical applications is to seek viable therapeutics that target these diseases to either ameliorate symptoms and/or reverse disease processes. The [Open Targets Platform](#)¹⁰³ (OTP) is a curated database of therapeutic targets by disease association. These associations are substantiated and prioritized through the systematic weighting of evidence supporting the association of a target with a given disease. While targets go beyond protein-encoding genes to include RNA and pseudogenes, they do not include protein-complexes or other targets comprised of more than one element. The aggregation of evidence extends beyond journal articles to include sources such as clinical signs and symptoms as well as various ontologies and text mining techniques. Associations are evaluated by multiple criteria, such as the strength and type of evidence connecting the target to the disease, scored, and ranked accordingly. At the time of writing this (January 2024), the OTP is updated bi-weekly. While the platform continues to refine its data aggregation methods, curation of results, part of the initial draw of the platform for our lab, remains a part of the method refinement process.

3.1.3 The Nucleoporins of Interest: Nup93 & Nup153

Some chromosomes maintain a fairly consistent location relative to the nuclear periphery regardless of the cell type¹⁰⁴, whereas this relative positioning varies between cell types for others¹⁰⁵. Both scenarios serve as evidence of the non-random nature of chromatin organization relative to the nuclear membrane.⁴¹ One of the organizational instigators at the nuclear periphery is the nuclear pore complex (NPC). The position of an NPC correlates, for instance, with β -actin mRNA occupancy proximal to the nuclear envelope¹⁰⁶ or spatial organization¹⁰⁷. Growing evidence supports the role of NPCs as organizational hubs during development and differentiation^{108–112}, the nuclear basket component Nup153 being a key player^{47,48}. The number of NPCs installed in the nuclear member is cell-type specific¹¹³ and are evenly distributed across the throughout the nuclear envelope.^{38,40} This non-random distribution may be further indicative of their nature as an coordinator of chromatin organization.

The large, multi-protein NPC connects the nucleoplasm to the cytoplasm, creating a tunnel through the double membrane of the nuclear envelope that primarily serves as a transport mediator. Each NPC is comprised of a membrane spanning pore – itself made up of a luminal ring, two outer rings, two inner rings and lined with FG Nups, which line the channel with Phe-Gly (FG) repeat motifs – as well as elements on the cytoplasmic side, and a basket that extends from the inner nuclear membrane into the nucleoplasm.^{35,36,114,115} These elements are formed by protein subcomplexes arranged in an eight-fold rotational symmetry.

The proteins making up the subcomplexes are diverse in structure and function. The research presented herein utilizes data collected regarding two nucleoporins in particular, Nup93 and Nup153. These two nucleoporins were the selected by the Hetzer lab as indicators of NPCs physically interacting with super enhancer regions (SEs) on chromatin as corresponds with cell identity.⁴² As a component of the pore's channel, Nup93 is more proximal to the NPC core than Nup153³⁶; therefore, indicators of Nup93 interaction alongside those of Nup153 with the same region of chromatin may be used to strengthen the argument a chromosome locus for physical interaction with the NPC.⁴²

It is also noteworthy that both Nup93 and Nup153 show a high degree of conservation across vertebrates, with Nup93 distributed across all five supergroups,¹¹⁶ which raises the question, did

predecessors of Nup93 and other Nups play a role in chromatic organization at the early stages of NPC evolution in eukaryotes? However, answering such a question is outside the scope of this study. Nup93 plays roles in the cell over and above its structural one. It is pivotal to regulation of actin cytoskeleton remodeling, while overexpression seems to lend to permissible conditions for metastatic cell invasion via the extracellular matrix (ECM).⁴³ In addition to the impact on the AC, Nup93 overexpression may also accelerate movement of transcription factors into the nucleus.¹¹⁷ A potential HOXA repressor, Nup93 may exert its influence through physical contact as depletion of Nup93 uncouples HOXA gene cluster from the nuclear envelope.⁴⁶ Nup153 also serves a purpose beyond transportation facilitator with studies suggesting Nup153 influences multiple aspects of cellular structure. Depletion of Nup153 affects not only the 3D structure of chromatin, patterns of differential expression are observed, particularly in those genes implicated in development.¹¹⁸ Furthermore, Nup153 has a key role in the regulation of NPC number.¹¹⁹

3.1.4 DamID

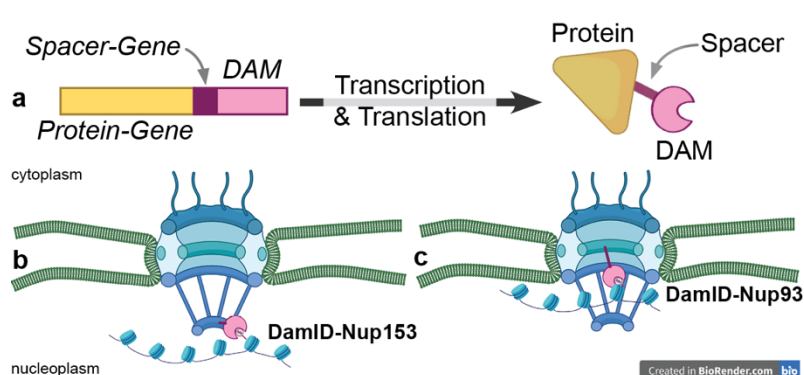


Figure 3.2 DamID is used for chromatin-protein interaction identification (a) The DAM gene is connected to a gene copy of the protein target. This protein is transcribed and translated into a DamID-protein complex including a spacer. (b) The Nup153-DamID protein methylating chromatin at the basket of the NPC. (c) The Nup93-DamID protein methylating chromatin at the inner ring of the NPC.

DNA adenine methylase identification (*DamID*)¹²⁰ is a method that indicates if and where a protein physically interacts with chromatin. By fusing an *E. coli* adenine methyltransferase (Dam) to the protein of interest, contact between chromatin and Dam-linked protein are marked by adenine methylation at GATC sites;¹²⁰ the constraints of local chromatin accessibility give this method a resolution of 1–5kb.^{121,122} To link Dam to the protein of interest, modified cell lines are created with a Dam-fusion transgene for the protein

(Fig. 3.2).

In this study, publicly available data is used that includes a DamID dataset. To produce this dataset, researchers created individual, U2OS-based cell lines with DamID fused proteins for Nup93 and Nup153 (exclusive) as well as Dam-GFP and Dam-LBR as references.⁴² A cell line expressing unlinked Dam was used as a control for calculating the relative local accessibility of chromatin.

The regions of chromatin that interacted with Nup93 and Nup153 were identified using NGS. This is achieved via production of a transgenic cell line. Though DamID was originally developed for DNA binding proteins, it has been successfully applied to Nup153 and Nup93.⁴² In the study described as well as others¹¹³ the role of NPCs in maintaining chromatin structure and gene expression as it pertains to cell identity was further elucidated.

3.1.5 siRNA

Observing the effect of turning a gene “off” can reveal information regarding the biological networks that gene influences, either directly or indirectly. RNA interference (RNAi) is a method that leverages conserved biological mechanisms of cellular defense against dsRNA^{xxix} to post-transcriptionally silence genes (see **Fig. 3.3**).^{123,125} This method can be used to systematically

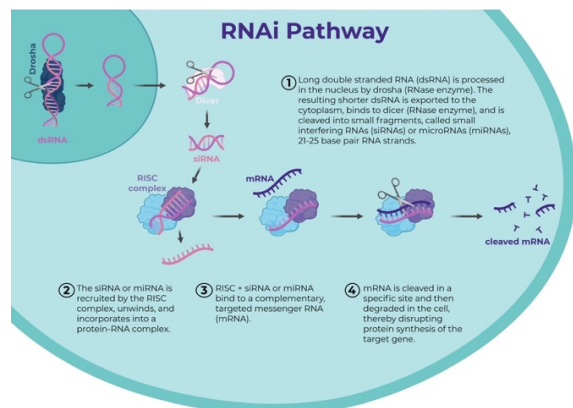


Figure 3.3 Overview of siRNA knockdown
This technology leverages endogenous RNA processing and degradation systems to bind and cleave mRNA that would otherwise be transcribed into the target protein.,.

From the RNA Therapeutics Institute @UMassChan > [What is RNAi?](#).¹²³

examine cellular pathways through the effects of gene knockdown.¹²⁶ The data utilized in this study was collected after siRNA was used to silence Nup93 and Nup153. To do so, lipid-based transfection of siRNA oligos (control, Nup93-specific, and Nup153-specific) was performed on U2OS cell cultures.⁴² These oligos were recognized by the RNA-induced Silencing Complex (RISC) and integrated into the protein for utilization in sequence recognition. The RISC-siRNA complex then recognizes and binds mRNA with the complimentary sequence to the siRNA. After binding, the mRNA is cut by RISC and degraded by the cell.¹²⁷ In this way, translation of the targeted protein is blocked. Some evidence suggests that using CRISPR/Cas9 produces more “useful” results.¹²⁸ Although this assertion is plausible given that CRISPR/Cas9 introduces a genetic mutation (knockout of the target protein) while siRNA hinges on degrading all mRNA before it can be transcribed (knockdown of the target

protein), the “usefulness” of the method is contingent on the research question being addressed. For instance, a complete knockout of a protein necessary to cell survival may have such catastrophic effects that the more subtle understanding of WHY it is necessary for survival may be lost.

3.1.6 Background Summary

The integration of various data sources requires an understanding of their origin with respect to both the biological processes involved and experimental methods utilized. With the background information in mind, the current study investigates the correlation between gene interaction with the nuclear pore complex and genes implicated in disease as well as those viewed as potential therapeutic targets. We propose that physical association with the nuclear pore complex correlates with the “diseased” facet of cellular identity.

To test this, rheumatoid arthritis (RA) will be our disease of interest and the U2OS cell line will provide the biological model. Our multiomic approach will integrate data from publicly available sources: DNA adenine methyltransferase identification (DamID) and small interfering (si) RNA experiments, the results of an intercontinental meta-analysis of RA GWAS, and a curated database, Open Targets. This assemblage of data will allow us to compare, respectively, indicators of interaction with NPCs, a “verified” list of RA gene risk factors, as well as potential therapeutic targets for select diseases, RA as well as three additional conditions for contrast.

^{xxix} For example, many viruses store their genome using dsRNA. dsRNA virus genera that act as human pathogens include: *Rotavirus*, infamous for Rotoviral enteritis, *Orbivirus*, which are transmitted via arthropods, and *Orthoreovirus*, which may cause upper respiratory infections and enteritis.¹²⁴

Furthermore, identifying those genes that contribute to cellular identity, which we defined inclusive of disease state, with the NPC, which is experimentally represented by Nup93 and Nup153, we will reduce the number of genes from which we will select gene targets for microscopy experiments.

3.2 Methods

3.2.1 Omic Data Acquisition and Exploration

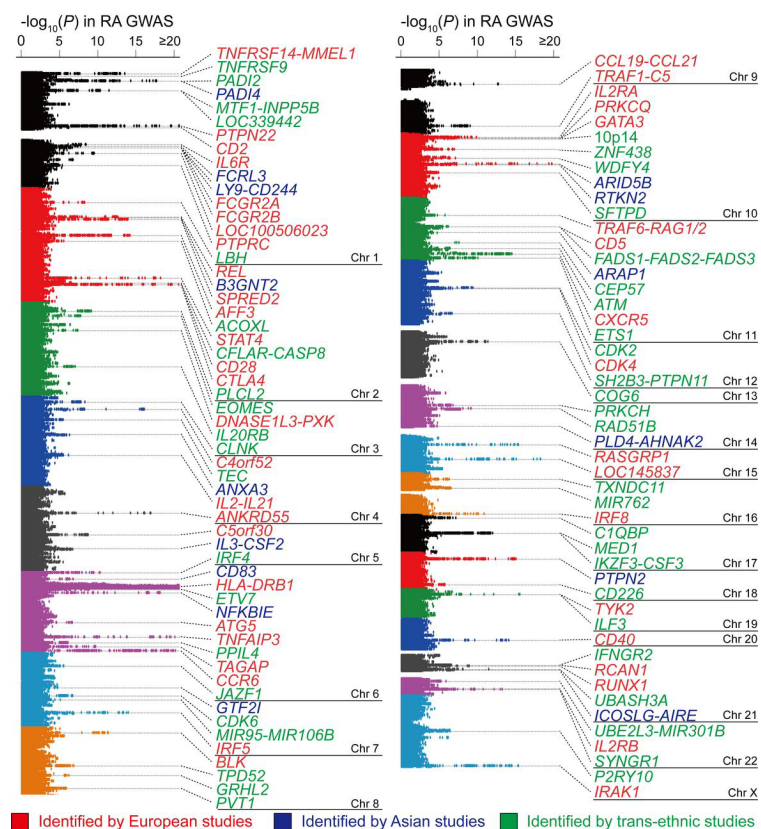


Figure 3.4 Manhattan plot of RA risk factors identified by a trans-ethnic meta-analysis

This plot includes the significance of each single nucleotide polymorphism (SNP) associated with RA in a trans-ethnic meta-analysis. From a data consumer standpoint, this figure has a pitfall. The way the lines demarcating chromosomes are inconsistently misleading. For example, “Chr 1” is placed underneath a gene that is on chromosome 2, while “Chr 7” is placed on a line underneath a gene that is on chromosome 7.

From Okada et al. *Annals of the Rheumatic Diseases*. 2019; 78(4). 10.1136/annrheumdis-2018-213678.

necessary, both of which were aligned to *hg19*, there is no need to consider if translating coordinates to align with any other processed data is required for this line of inquiry. For the sake of simplicity and because there is no need to do otherwise, a gene’s status on the GWAS gene

Before aggregating or integrating the data, the datasets are considered individually. Sometimes, this evaluation includes looking at summary statistics^{xxx}, other times it includes comparing a manually transcribed list one was provided alongside the original source. In the case of the RA GWAS data, digging through the supplementals and comparing it against the data visualization printed within the associated journal article proved invaluable to the integrity of the data. Extracted from the Manhattan plot included in the journal article (**Fig. 3.4**), the gene list used in preliminary analysis did not match the list published as part of the supplemental. One might consider the error that of an exhausted researcher (it likely was) the authors were also responsible via some poor decisions regarding the visual their visualization of the gene list. Not only does this highlight the importance of confirming one’s data, it provides an excellent example of what the position of a few lines can do to the interpretation of a plot.

As the DamID and siRNA datasets are the only sets for which genomic coordinates are

^{xxx} Unless otherwise noted, it can be assumed that any given calculation and/or statistical analysis was done using the R programming language¹²⁹ (specifically, all code was updated to R version 4.3.1 (2023-06-16) at the time of writing this manuscript).

list will be considered a gene attribute rather than untangling the alignment status of each dataset that contributed to the meta-analysis that resulted in the list.

3.2.1.1 Rheumatoid Arthritis GWAS

A list of 105^{xxxi} RA risk factors was extracted from the 2018 update of the multi-ethnic meta-analysis of RA GWAS data enumerated by the Laboratory for Autoimmune Diseases at the RIKEN center for IMS.¹⁰¹ This list was expanded on a 2014 meta-analysis by the same group, which identified 98 RA risk genes outside of the MHC region.¹⁰² The biological network of each of the 98 genes overlapped, either through direct interaction with their gene product (protein) or a protein within the product's direct protein-protein interaction (PPI) network, with that of 27 genes targeted by therapeutic drugs approved for the treatment of RA. The list was expanded in their 2018 follow-up study, which included the previously excluded MHC region using an imputation method¹³⁰ to look at the complex web of genetic interactions found in this region of chromosome 6.

3.2.1.2 NPC-associated DamID & siRNA Identified Genes

For data regarding gene-NPC interactions, we utilized a subset of the multi-omic dataset produced by the Hetzer laboratory's examination of physical contact between NPCs and super enhancers (SEs) as it pertains to cell identity.⁴² The full superset for the study included, but was not limited to: ChIP-seq (H3K4me3 and H3K27ac), DamID (Nup93 and Nup153), and RNA-seq (including the results of siRNA treatment for Nup93 and Nup153) datasets collected using the U2OS cell line. As mentioned in the [Background](#), Nup93 was used to “back-up” evidence of chromatin- Nup153 association given their relative positions within the NPC structure: channel versus basket. For the purposes of the research presented here, data was imported from the supplemental files associated with the cited study.⁴² The data contained in the supplemental were aligned to *hg19*, *NCBI37* (equivalent to *GRCh37*) as a reference genome. The results presented here are based on processed rather than raw data. The use of the raw datasets for future endeavors, as well as the drawbacks of using pre-processed data, will be addressed in the [Overarching Discussion](#).

Table 3.1 Data encoding and what it indicates in terms of regulatory effect on differential gene expression after treatment with siRNA.

Encoding	Effect on gene expression of siRNA targeting the indicated Nup	If knockdown of a Nup is associated with...	Regulatory effect
1	Increase	an <i>increase</i> in a gene's expression, its presence will be associated with a <i>decrease</i> in expression of the same gene.	Down
0	No change	<i>no change</i> in a gene's expression, its presence will be associated with <i>no change</i> in gene expression of the same gene.	None

^{xxxi} The figure caption indicated there were 106 genes; however, only 105 were listed in the figure. Clarification was sought from the author and has yet to be obtained.

-1	Decrease	a decrease in a gene's expression, its presence will be associated with an increase in expression of the same gene.	Up
----	----------	---	----

The schema used for encoding the effect of siRNA on gene expression alongside how it translates biologically is shown in **Table 3.1**. This dataset included the direction of the change, but not the magnitude (think unit vector). While data was collected with a control siRNA, siRNA targeting Nup93 and siRNA targeting Nup153, the case where both nucleoporins were targets of siRNA was not tested. The experimental conditions are of note as so that appropriate statistical tests may be selected based on an alignment between assumptions and the nature of the collected data; in this case, mutually exclusive groups can be assumed, meaning an assumption of the Chi-squared test, as will be discussed in [Combinatorial analysis of sub-setting results](#).

3.2.1.3 RA, OA, OS, BC Gene from Open Targets Platform¹⁰³

Data from the OTP was pulled^{xxxii} for four diseases, each serving a specific purpose. Rheumatoid Arthritis (RA) was the disease of interest.^{xxxiii} Osteoarthritis (OA) was used as a control with respect to bone degeneration. Osteosarcoma (OS) is the type of tumor from which the U2OS cell line was derived. Breast Cancer (BC) was selected as a control for cancer related genes, which is along the same lines of using OA to control for non-RA specific bone degradation.

3.2.2 Multiomic Data Aggregation

Rooting a multiomic analysis in biological reality provides the boundary conditions necessary to inspire pragmatic methods and contextually relevant results. For example, this study requires a general understanding of risk factors for RA, the experimental methods employed for experimentation and data collection, and the structure of data on the OTP. Additionally, such analyses must be approached systematically.

The process presented here begins with an understanding of the datasets employed, starting with the “limiting reagent” as it were. In this analysis, those genes with DamID and/or change in gene expression after siRNA treatment are the limiting reagent, as it is in these data the biological mechanism in question finds its foothold. From there, the RA GWAS gene list is explored and the overlap with the DamID and siRNA datasets examined. Finally, the DamID and siRNA datasets viewed through the lens of the OTP RA drug targets dataset, which will be juxtaposed with the RA GWAS list, alongside similar datasets for osteosarcoma (OS), osteoarthritis (OA), and breast cancer (BC).

3.2.2.1 Statistical Methods

Here we establish what statistical methods will be employed in the exploration and integration of datasets that follows. These methods include determining if sets are independent as well as modeling the relationships present in the datasets, individually and/or when integrated.

^{xxxii} These datasets were pulled in January 2021. The contents of the Open Targets Platform change over time, so should these analyses be done with current Open Targets data, the results are not expected to be the same.

^{xxxiii} The disease of interest was selected based on the appeal of the GWAS study findings, particularly the validity provided by associations with existing therapeutics.

3.2.2.1.1 Assessing Independence

For the data that is categorical in nature, namely the siRNA and DamID assay results, testing for the null hypothesis^{xxxiv} (H_0) – that membership (or not) in one group is independent of membership (or not) in another group – employs Pearson’s Chi-squared (χ^2) test¹³¹ and, if indicated by the sample size, Fisher’s exact test rather than Yate’s continuity correction will be applied, though both tend to be more conservative.¹³² These are tests are often preceded by the construction of a contingency table^{xxxv} and assume the groups being compared (each represented by a cell in

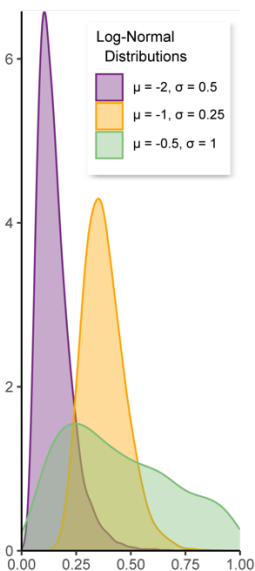


Figure 3.5 Log-Normal Distributions
Examples of log-normal distributions with different values set for the mean (μ) and standard deviation (σ) parameters.

the contingency table) are mutually exclusive. The χ^2 -test^{xxxvi} is applicable to contingency tables greater than 2×2 ; herein it was employed to look at both a 4×4 as well as $2 \times 2 \times 2 \times 2$ tables to consider potential dependency between subsets of genes as classified by assay results.

The contingency table construct can be parlayed into odds ratios that compare the likelihood of one condition given another. In the case of a $2 \times 2 \times 2 \times 2$ table, the χ^2 -test is applied by holding two conditions constant while calculating the odds ratios for each of the resulting four 2×2 contingency tables. Although interpreting values such as the χ^2 -statistic with respect to the null hypotheses is an element of understanding data, it does not paint the most tangible, “human-friendly” picture. Odds ratios do provide (slightly) more interpretable results. However, explaining such results must be done with a grain of context, as they are highly dependent on what is held constant.^{xxxvii}

For a more nuanced picture of the relationships between the frequencies and the binary variables that characterize groups, [log-linear analysis](#) is employed.¹³⁵

3.2.2.1.2 Generalized Linear Models

Before diving into the use of log-linear analysis in this study, we first outline the use of generalized linear models (GLMs). The reason for this is twofold: 1. log-linear analysis hinges on the use of a specific GLM, and 2. another form of GLM is utilized when modeling the Open Targets data. A GLM frees the dependent (response) variable (\hat{y}) in linear regression^{xxxviii} from the constraints of the normal (Gaussian) distribution.

^{xxxiv} An H_0 testable using the χ^2 -test: the set of genes that change expression after siRNA treatment for Nup153 is independent of the set of genes with positive DamID for Nup153; that is, being a member of the first set of genes makes a gene no more or less likely to be a member of the other set.

^{xxxv} Those familiar with machine learning are likely familiar with a special type of contingency table, the confusion matrix. This table compares actual and predicted to evaluate a model rather than conditional differences between classes (like exposure to a pathogen). Such *confusion matrices* look an awful lot like contingency tables used to determine sensitivity and specificity in applications like COVID test development – and not coincidentally. Unraveling inter (and intra) disciplinary jargon has been non-negligible part of this author’s research journey.

^{xxxvi} The Cochran-Mantel-Haenszel Statistic^{xxxvi} was also briefly considered; however, its use applies strictly to holding a third variable constant for a 2×2 table.^{133,134}

^{xxxvii} To compare the difference between odds ratios when an alternative set of variables are held constant, compare **Table 3.11** with **Table S3.24** for an example of different ranges in odds as well as the ease with which one becomes tongue tangled when trying to verbalize the interpretation.

^{xxxviii} Yes, THAT linear regression, standardly taught in Algebra 1 as $y = mx + b$ and progressively features more Greek letters as one gets older.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

This is done by transformation via a link function.

Assume that the observed values, y , come from a log-normal^{xxxix} distribution (**Fig. 3.5**). Then we might use log as a transformation to better fit the data. This is done via a link function, $g(z)$. We also define the family of the GLM as Gaussian, given the observed values were sampled from a log-normal distribution.

For the GLMs used to model the Open Targets data in this analysis, the data is thought to have a log-normal distribution, based on a comparison of the empirical (observed) cumulative density and

$$g(\hat{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

$$\text{where } g(z) = \log(z), \text{ therefore} \quad (2)$$

$$\hat{y} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}$$

the theoretical cumulative density function (CDF) of continuous distributions, as the dependent variable is continuous, as well as a comparison between q-q plot of empirical versus theoretical distributions. The overall association scores for each disease were modeled as dependent on the binary variables representing the assay results with GLMs constructed using `stats::glm` with parameters for Gaussian distribution, logarithmic link function, and sum contrasts. Sum contrasts recode categorical variables such that their sum is 0, similar to centering a continuous variable on the mean. In the case of binary variables, instead of the levels 0 and 1, the model uses -0.5 and 0.5. In a model with no interactions, this means the two levels will have equal magnitude but opposite direction. As a result, interpretations of the main effect are relative to the intercept (β_0), even if interaction terms are added (though they are not in this portion of the analysis).

3.2.2.1.3 Log-Linear Analysis: Modeling associations without defining treatment and response variables

Rather than looking at cell means, the log-linear model looks at cell frequencies. The focus of analysis is constructing a model for categorical outcomes that minimizes the difference between the expected – model predicted – and observed outcomes. An “ideal” log-linear model has a ratio of observed to expected (predicted) as close to 1 as possible – i.e. approaching no difference between observed and expected – while maintaining as much information as possible. This is done by step-wise optimization on the Akaike Information Criterion (AIC), a value which weights information loss against maintaining^{xi} a ratio of observed to expect close to 1. The AIC is given by:

$$AIC = -2 \log(\mathcal{L}(\hat{\theta}|y)) + 2K$$

where $\mathcal{L}(\hat{\theta}|y)$ is the maximized likelihood of a predicted unknown ($\hat{\theta}$), given the estimator y , which is generated by the model under scrutiny, and K is the number of estimable parameters in the “approximating” model.^{136,137} (3)

Log-linear analysis begins with a comparison of the *simple model*, which contains no interaction terms, and the *saturated model*, which contains all possible interaction terms. The *log-linear model*, as the name implies, is a linearization of a multiplicative independence model through the application of logarithms.

^{xxxix} In some fields, it is customary to say “log” when one means “natural log” with an utter disregard for the use of “log” with a base of 10 (or 2, for that matter) in other fields. In this manuscript, we will carry on this confusing tradition like some sort of right of ivory tower passage.

^{xi} The word maintained is used here because, as will be discussed, the algorithm starts with the fully saturated model, which has O/E = 1 BUT is, essentially, overfit.

In other words, a multiplicative model such as, where $\hat{\tau}$ is the “grand” geometric mean of the expected cell frequencies of the model, and $\hat{\tau}_i^A$ and $\hat{\tau}_j^B$ are ratios of conditional expected frequencies and $\hat{\tau}$:

$$F_{ij} = \hat{\tau} \hat{\tau}_i^A \hat{\tau}_j^B \quad (4)$$

can be transformed:

$$\log(F_{ij}) = \log(\hat{\tau}) + \log(\hat{\tau}_i^A) + \log(\hat{\tau}_j^B) \quad (5)$$

and is often written as:

$$\log(F_{ij}) = \lambda + \lambda_i^A + \lambda_j^B \quad (6)$$

which looks an awful lot like an additive linear model – because, in this form, it is – hence, it is a “log-linear” model, which is a generalized linear model (GLM) This particular log-linear model, also known as a *mutual independence model*, contains no interaction terms and is considered a simple log-linear model. This simple model is one of two models with which one starts a log-linear analysis. The other model is the saturated model, which contains all possible interaction terms. For the scenario modeled in Eq.6, which has conditions A and B, the saturated model would be written:

$$\log(F_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_i^A \lambda_j^B \quad (7)$$

The saturated model always fits the data, but it may not be the *best* model for the data. To find the best fit (with respect to the inquiry at hand), the effect individual terms have on the overall model is considered. This evaluation is requires examining all possible log-linear models. Because log-linear models are (generally) hierarchical models, they must adhere to the following condition: for all variables in the highest-level interaction term, all possible lower-level interaction terms that involve that variable must be included. The inclusion of interaction terms ensures the model encompasses all effects of a variable for the maximum “way” at which it is included. For example, the following model is not acceptable:

$$\log(F_{ijk}) = \lambda + \lambda_i^{height} + \lambda_j^{weight} + \lambda_k^{age} + \lambda_i^{height} \lambda_j^{weight} + \lambda_i^{height} \lambda_j^{weight} \lambda_k^{age} \quad (8)$$

The variable *age* is present in a three-way interaction term but the two-way interaction terms that include *age* are not present. To fit the criteria, either the three-way interaction term would need to be removed,

$$\log(F_{ijk}) = \lambda + \lambda_i^{height} + \lambda_j^{weight} + \lambda_k^{age} + \lambda_i^{height} \lambda_j^{weight} \quad (9)$$

or the two two-way interaction terms that include *age* (interaction terms with *weight* and with *height*, respectively)

$$\begin{aligned} \log(F_{ijk}) = & \lambda + \lambda_i^{height} + \lambda_j^{weight} + \lambda_k^{age} + \lambda_i^{height} \lambda_j^{weight} + \lambda_i^{height} \lambda_k^{age} \\ & + \lambda_k^{age} \lambda_j^{weight} + \lambda_i^{height} \lambda_j^{weight} \lambda_k^{age} \end{aligned} \quad (10)$$

Significance for generalized linear models can be calculated using the χ^2 -likelihood ratio test (G^2 -test)^{138,139}, given by

$$G^2 = 2 \sum O_{ij} \log_e \frac{O_{ij}}{E_{ij}} \quad (11)$$

Furthermore, models can be compared by calculating the G^2 value and determining if the overall model quality is improved based on an increase in the score, which itself comes with a corresponding p-value. Individual estimated coefficients for each model have a calculated p-value, via the z-score, which has been shown to be robust to skewed distributions, making it acceptable for Poisson distributions.¹⁴⁰

Log-linear models were constructed using `stats::glm` with parameters for Poisson distribution because the data is frequency-based, not continuous, a logarithmic link function, and (again) sum contrasts. As a result, interpretations of the main effect are relative to the intercept (λ) even as interaction terms are added, which is relevant in this portion of the analysis.

To algorithmically assess the range of possible log-linear models for a given set of data requires weighing overfitting against information lost. Given the number of possible models to assess for systems involving more than two variables, algorithmic methods of converging on a suitable model have been developed. One such method employs the AIC to iteratively assess models while adding or removing terms (depending on whether the search starts from the simple or saturate model).¹³⁷ To this end, `stats::step` is employed to obtain the best (or least-bad) log-linear model to represent the data.

In many gene-centric and genome-centric studies, the genes are each treated as an individual variable. This introduces a statistical hurdle known as the “Multiple Comparisons Problem”, which rests on the idea that the chance of encountering a spurious correlation via coincidence increases as the number of comparisons increases (i.e. a greater opportunity for Type I error: rejecting the null hypothesis when it is true). Such can be the case in transcriptome-wide RNA-seq analyses unless correction methods¹⁴¹ such as the Bonferroni correction¹⁴² are applied. This study avoids this problem by looking at genes as sets based on criteria rather than as individual variables, which lends to a smaller chance (comparatively) of uncovering meaningless, though statistically significant results.

The possibility of encountering spurious correlations due to unseen factors is also at play, owed in part to what is referred to as the *independence assumption* – genes within a set are independent and can be assumed so for the purposes of statistical testing.¹⁴³ This particular issue arises when sets of genes are pre-determined, presumably due to criteria related to the research question at hand, such may be the case when analyzing microarray data. Given the subsetting of the data is based on experimental results rather than driving the experimental results, this obstacle is avoided by this study.

3.2.2.1.4 Is this [data] Normal?

Normality – sampled from a normal distribution – is an assumption made by many statistical models. Before applying such parametric models, the normality of the data must be assessed.

To test for normality, we visually inspect the data by plotting density plots of the scores as well as apply the Shapiro-Wilk Normality Test. The Shapiro-Wilk Normality Test has an upper limit to the

number of observations it can test, $3 \leq n \leq 5000$.^{144,145} As a secondary method of testing normality, the Asymptotic One-Sample Kolmogorov-Smirnov test¹⁴⁶ was applied.

3.2.2.1.5 Comparing Group Means

Non-parametric methods are used to compare group means to avoid the assumption data were normally distributed (they were not normal, as will be indicated in the [Results](#)). To make an overall comparison of the groups, i.e. is there any difference among this collection of groups, the Kruskal-Wallis Rank Sum Test¹⁴⁷ was used. When venturing into which groups among the collection significantly differ, Pairwise Wilcoxon Rank Sum Tests^{148,149} were run.

3.2.2.2 Filtering of Genes by Experimental Results

After importing and processing all datasets, an exploratory analysis focused on DamID and siRNA treatment assay results with respect Nup93 and/or Nup153 was performed. Contingency tables were used to compared subsets based on assay results. Conditions (in terms of assay results) that had greater contributions to test statistics were further interrogated using the statistical methods previously described.

Additionally, odds ratios were used to compare groups when test-statistics indicated the sets compared were highly unlikely to be independent – in other words, the standard null hypothesis of χ^2 -testing was rejected. Odds ratios provide a view harmonious with the statistical method that follows next, which has can be interpreted using odds ratios.

At this point, the analysis shifts from tests of independence to modeling association via log-linear analysis. Because log-linear regression does not expect variables to have roles [as in independent versus dependent variables in the model] it is well suited to considering the relationships between binary variables. The purpose of such a shift is to consider the effects of interactions between variables rather than “just” the main effects. Additionally, because the models that result from log-linear analysis can be interpreted in terms of odds ratios, they are unaffected by sample size or unequal distribution between margins (distribution between either row values or column values, if one pictures an R x C table, all else being held constant).¹³⁵

3.2.2.3 Integration of the RA GWAS Dataset

The junction between each disease-based dataset as well as the NPC dataset and the RA GWAS gene list were found. These junctions were used to look at the overlap with the Nup dataset as subset by disease as well as the RA GWAS dataset. This analysis was done by visualizing the overlap using the `eulerr`¹⁵⁰ and `ComplexUpset` libraries.

3.2.2.4 Comparison with Potential Drug Targets

Initially, the datasets were downloaded^{xli} directly from the database via their browser-based GUI after searching for drug targets by disease. Subsequently, as a “proof-of-concept”, data regarding RA was pulled using the Open Targets API via `sparkR`¹⁵¹, an R library frontend for interfacing with databases built on Apache Spark.¹⁵² However, given the relative frequency with which the Open Targets database is updated and the want for a stable set of data for analysis, the data have not been refreshed regularly. For the purpose of this study, we used the original, manually pulled datasets for each disease of interest, which were downloaded in January of 2021.

^{xli} Initial downloads occurred on 12-Jan-2021 and 14-Jan-2021. The dates of download are significant due to updates Open Targets has undergone (and continues to undergo) on a regular basis.

The datasets were compared to the Nup93 and Nup153 assay data as well as the RA GWAS dataset using an UpSet plot to find the intersections. For the analysis that followed, only data that were in the intersection of each OT dataset with the assay dataset were included. The proportion of genes in the Nup subsets to the overall drug targets was calculated for each pairing for the overall NUP dataset as well as the subset.

The distribution of association scores stratified by disease were visualized using box plots and the significance between the average disease association score was determined both overall as well as pairwise by disease. For each disease, a GLM with a Gaussian distribution and the log link function was used to model the overall association score in terms of binary assay result variables. These models were limited to observing the main effects between these variables, which was done by reflecting on the p-values of the individual correlation coefficients as well as that of the overall model.

3.3 Results

Using the framework established in the [Methods](#) section, we began our analysis by characterizing the DamID and siRNA datasets. After modeling the relationships between these assay results by looking at how these results subset a list of gene names (the genes were not considered individually), the associations between these assays, the RA GWAS risk gene list and the Open Targets datasets pulled for four different diseases were considered.

3.3.1 Exploratory Analysis of the DamID and siRNA Datasets

Of the 26392^{xlii} genes assessed, roughly 38% (9938 genes) were associated with either Nup93 or Nup153. Association was established by physical contact at the gene loci via DamID¹³ with at least one Nup93 and Nup153 and/or an observed change in gene expression after cells were treated with siRNA gene suppression targeting Nup93 or Nup153 (but not both simultaneously).^{xliii} Chromatin has more often been methylated by DamID-Nup153 than DamID-Nup93; a randomly selected gene is 74.9% more likely to be associated with Nup153 than Nup93 in U2OS cells. This result is within the realm of biological plausibility given Nup153 can be found free within the nucleus as well as a component of the “basket” of the NPC, which extends into the nucleoplasm. In terms of siRNA, a randomly selected gene has only 26.4% a chance as likely of changing expression after treatment with siRNA against Nup153 (siRNA-Nup153) compared to changing expression after treatment with siRNA against Nup93 (siRNA-Nup93).

The original study identified 1021 chromatin interaction sites for Nup93, which were in close proximity to the 1851 chromatin interaction sites indicated for Nup153. These data are available in a supplemental to the associated journal article. However, they are presented as binary variables in terms of genes, rather than counts ascribed to loci, leading one to conclude multiple interaction sites were ascribe to one gene if they occurred between its start and end coordinates. Using this gene-centric data, we found that 1966 (11.3%) genes were DamID-Nup93 and/or DamID-Nup153 positive. Of those genes 851 (3.2%) were DamID-Nup93 positive (478, or 1.8%, exclusively) and 1488 (5.6%) were DamID-Nup153 positive (1115, or 4.2%, exclusively). The

^{xlii} The study that provided the DamID and siRNA datasets, 29391 genes were assayed. Upon aggregation with the Open Targets datasets, two entries for *PINX1* appeared. Initially, the temptation was to merge them; however, the osteoarthritis and breast cancer datasets had two entries for PINX1 with different Ensembl ID numbers.

^{xliii} Also known as *exclusive disjunction*

remaining 373 genes (1.4%) were positive for both DamID. The intersections of the gene subsets formed by the positive assays were then visualized alongside their representation on the RA GWAS gene list (**Fig. 3.6**), the detailed results of which are discussed [later](#). By subsetting on the assays, we moved the view of the data from gene-centric to gene-set-centric with sets categorically defined by assay results.

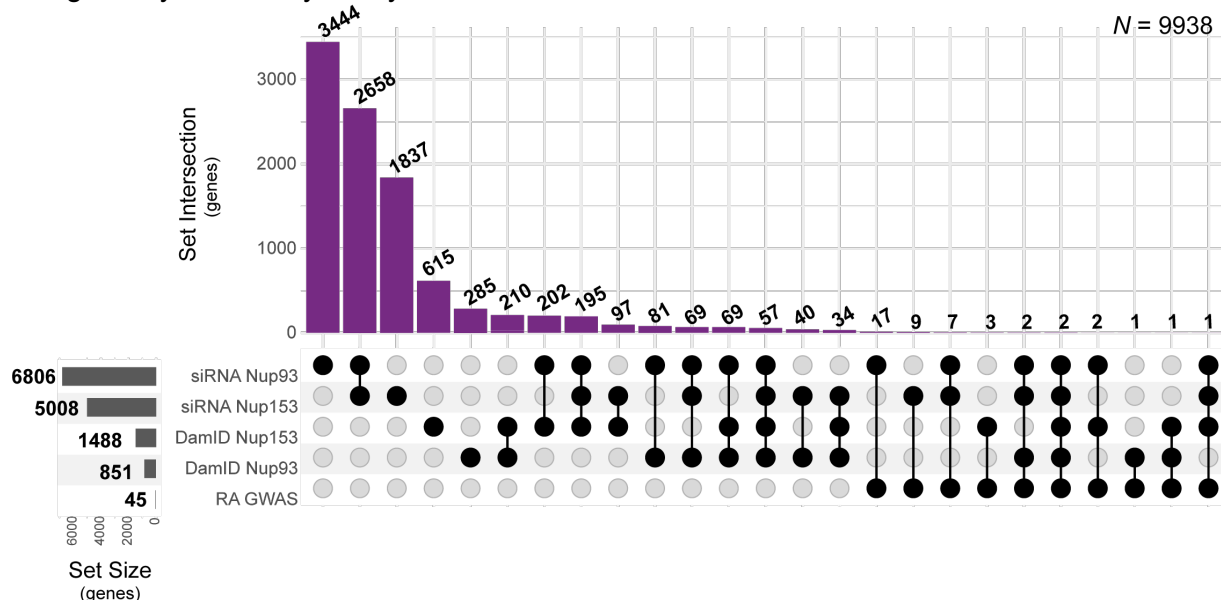


Figure 3.6 Intersections of sets of genes as classified by siRNA, DamID, and a RA GWAS meta-analysis. An UpSet plot¹³⁸ depicting the set unions of genes categorized by positive assays for DamID and siRNA with respect to Nup93 and Nup153 as well as the gene list presented in “Genetics of rheumatoid arthritis: 2018 status”.⁶⁵ The number of genes in each set are represented by bars in the lower, left corner and the size of the intersections of each set are indicated by the bar graph along the top of what is called the “interaction matrix”. The interaction matrix indicates which sets are part of an intersection with black dots vertically aligned below the bar representing the intersection. Membership within a set is indicated by a black dot’s horizontal alignment with a set name. All intersections are exclusive, meaning no genes contained in one intersection are counted within another intersection. The basis for this method of presenting set intersection is less complex (and less error prone) than Euler diagrams, while retaining the encoding of relative set and intersection size.

Graph created in R with ComplexUpset.¹³⁹ By implementing UpSet plots within the ComplexHeatmap R package, the flexibility of the visualization method was increased beyond the scope of the original UpSetR package. This flexibility includes genomic intervals as a viable set type via GRanges.

3.4 Testing for subset independence using χ^2 -statistics

To test the independence of the gene subsets, contingency tables were constructed to perform χ^2 -tests, which are provided with the results. To ease the burden on the reader as they follow this analysis, purple and green outlines surround the same subset of genes in each table. The purple border indicates genes with neither a positive result for Nup93 nor Nup153 for either assay. The green border indicates genes with at least one positive assay for Nup93 and at least one positive assay for Nup153, whether it be DamID or change in expression after siRNA treatment. It is also worth noting that the direction of change in gene expression is not considered in this analysis; a change of expression in either direction indicated it was affected. In other words, up and down regulation were assigned as one value of a binary variable with no expression as its counterpart. As we progressed through looking at the independence of siRNA and DamID assay results, contingency table dimensions begin with the familiar 2×2 arrangement (**Table 3.2**), then progress to 4×4 (**Table 3.3**), and finally a 2×2×2×2 (**Table 3.4**), which can be considered as a whole as well as a set of 2×2 tables created by stratification on two dimensions (**Table 3.5**).

Table 3.2 2x2 Contingency Table of siRNA treatment response versus DamID results with respect to either Nup93 or Nup153^a

		Change in Expression when treated with siRNA against either Nup			
		No	Yes		
Physical Association with either Nup	No	16454	7972	24426	
		16260	8166		
		2.309	4.597		92.55%
	Yes	1115	851	1966	
		1309	657		
		28.684	57.119		7.45%
		17569	8823		26392
		66.57%	33.43%		

$$\chi^2(1, N = 26392) = 92.709, p = 6.057546 \times 10^{-22}$$

For each cell:	Observed	
	Expected (rounded to nearest integer)	
	χ^2 contribution	

^a Neither DamID nor siRNA treatment were applied to both Nups simultaneously. Count and χ^2 contribution reported for each cell.

The null hypothesis that physical association via DamID with Nup93 and/or Nup153 is independent of change in expression after treatment with siRNA for Nup93 xor Nup153 was tested using χ^2 -test (**Table 3.2**) and is not supported ($\chi^2(1, N = 26392) = 92.709, p \ll 0.01$).

Table 3.3 4x4 Contingency table of siRNA treatment response^a versus DamID results with respect to either Nup93 xor as well as and Nup153^b.

		Change in Gene Expression after treatment with siRNA against...					
		Neither	Nup93	Nup153	Nups 93 & 153		
Physical association via DamID with...	Neither	16454	3461	1846	2665	24426	
		16260	3531	1867	2768		
		2.309	1.38	0.231	3.847		92.55%
	Nup93	286	81	40	71	478	
		318	69	37	54		
		3.259	2.051	0.329	5.228		1.81%
	Nup153	618	204	97	196	1115	
		742	161	85	126		
		20.799	11.379	1.63	38.376		4.22%
	Nups 93 & 153	211	69	34	59	373	
		248	54	29	42		
		5.604	4.219	1.059	6.62		1.41%
			17569	3815	2017	2991	26392
			66.57%	14.46%	7.64%	11.33%	

$$\chi^2(9, N = 26392) = 108.320, p = 3.231 \times 10^{-19}$$

For each cell:	Observed	
	Expected (rounded to nearest integer)	

χ ² contribution	The greatest contributor to the χ ² -statistic was the set of genes that were physically associated with Nup153 and showed a change in expression after both siRNA treatments (orange box).
-----------------------------	--

^a Neither DamID nor siRNA treatment were applied to both Nups simultaneously.

^b Change in gene expression was classified as a response without consideration of the direction (up or down).

Assay results were used to assign genes to mutually exclusive subsets as encoded by two categorical variables, one per assay type.

The null hypothesis remains rejected after the response to siRNA treatment and the DamID results were broken down into exclusive categories with respect to either assay: neither Nup93 nor Nup153, only Nup93, only Nup153, and Nup93 as well as Nup153 (**Table 3.3**). The null hypothesis remains rejected after the transition from binary to categorical variables ($\chi^2(9, N = 26392) = 108.320, p \ll 0.01$). The greatest contributor to the χ^2 -statistic when conditioned on DamID and siRNA class is the subset of genes positive for DamID Nup153 that change expression when treated with siRNA against Nup93 and Nup153 (χ^2 -contribution = 38.376, 35.43% of $\chi^2(9, N = 26392)$; highlighted by a yellow border in **Table 3.3**).

Given the top three χ^2 -contributors were three out of four of the subsets conditioned on DamID Nup153 being true, the potential for physical association with Nup153 being “protective” was considered. While the question, “What direction does the physical association with Nup153 go in terms of change in expression when either Nup is knocked down?” cannot be asked, given the information we have, what can be asked is, “are genes that are physically associated with Nup153 more or less likely to change expression when treated with siRNA-Nup93 or siRNA-Nup153?” To answer this, we first ran the χ^2 -test DamID-Nup153 × siRNA Nup Class ($\chi^2(3, N = 26932) = 94.101, p = 2.881 \times 10^{-20}$), rejecting the null hypothesis that DamID-Nup153 is independent of siRNA-Nup response classification. Based on the same contingency table (not shown), a gene is 36.8% more likely to show a change of expression after treatment with siRNA-Nup93 than siRNA-Nup153.

We then considered the effects of the separate treatments without considering the interaction between the two treatments, i.e. genes with a change in expression after siRNA-Nup93 may or may not also change siRNA-Nup153, and so on (tables also not shown). The χ^2 -statistics were calculated for DamID-Nup153 × siRNA-Nup93 ($\chi^2(1, N = 26932) = 76.924, p = 1.777 \times 10^{-18}$) and DamID-Nup153 × siRNA-Nup153 ($\chi^2(1, N = 26932) = 49.761, p = 1.736 \times 10^{-12}$), respectively. The odds a gene with a positive Nup153-DamID having a change in expression after siRNA-Nup93 is slightly higher (6.18%) than after siRNA-Nup153 (OR = 1.632, 95%CI:1.462–1.821 and OR = 1.537, 95%CI: 1.351–1.793, respectively).

To consider the interaction effects using χ^2 -tests, DamID-Nup153 × siRNA-Nup93 was calculated while the condition for siRNA-Nup153 was held constant (siRNA-Nup153(0): $\chi^2(1, N = 21384) = 37.601, p = 6.728362 \times 10^{-10}$; siRNA-Nup153(1): $\chi^2(1, N = 5008) = 6.984, p = 8.225 \times 10^{-3}$). DamID-Nup153 positive genes that change expression with siRNA-Nup153 are less likely to change expression with siRNA-Nup93 than those who aren't DamID-Nup153 and do not change expression with siRNA-Nup153 (86.2% or -0.215x). As an alternative approach to untangling potential interactions between variables, we will apply log-linear analysis [later in this section](#).

For a closer look at their dependency, the categorical variables in **Table 3.3** were exchanged for Nup and assay specific binary variables. To illustrate how both tables display the same data as represented by different variable types – categorical versus binary – a purple border outlines the cells where all conditions are “No” (i.e. a negative assay result) while a green border outlines the cells where at least one “Yes” exists for an siRNA as well as a DamID assay. The result was a

four-way table (2×2×2×2) that tests for independence between four binary variables simultaneously (**Table 3.4**). This construction provides little pairwise information specific to the independence of any two variables, as is evident by the singular χ^2 -statistic calculated.

Table 3.4 2×2×2×2 Contingency table of response to siRNA treatment against Nup153 xor Nup93^a versus DamID for Nup153 xor Nup93.

					Change in expression after treatment with siRNA against...					
					Nup93					
					No	Yes				
					Nup153					
					No	Yes	No	Yes		
Physical Association via DamID by...	Nup93	No	No	16454	1846	3461	2665	24426		
				16260	1867	3531	2768			
				2.309	1641.318	31040.618	23521.256		92.55%	
				618	97	204	196		1115	
				742	85	161	126			
				835.342	1.630	767.728	984.133		4.22%	
	Nup153	Yes	No	286	40	81	71		478	
				318	37	69	54			
				2981.978	91.102	2.051	5.412		1.81%	
				211	34	69	59		373	
				248	29	54	42			
				2362.277	67.511	4.059	6.620		1.41%	
					17569	2017	3815	2991		26392
					66.57%	14.46%	7.64%	11.33%		

$$\chi^2(11, N = 26392) = 6888.299, p = 0$$

^a Neither DamID nor siRNA treatment were applied to both Nups simultaneously.

^b Change in gene expression was classified as a response without consideration of the direction (up or down).

Assay results were used to assign genes to mutually exclusive subsets as encoded by four binary variables, one per Nup per assay type.

Changing from categorical to binary variables changed the order of the values due to the hierarchical nature of displaying this way versus the flattened-by-categorization nature of the other display.

The results of the χ^2 -test remain significant ($\chi^2(11, N = 26392) = 6888.299, p = 0^{\text{xliv}}$) with the higher bar created by the χ^2 -distribution for 11 (versus 9) degrees of freedom. To test for dependence between the subsets comprising **Table 3.4**, χ^2 -tests were performed on a contingency table, **Table 3.5**. This table (**Table 3.5**) represents the same subsets of data as (**Table 3.4**), now broken into four 2×2 contingency tables, one for each combination of the two binary variables being held constant. For example, the 1st quadrant contains a 2×2 contingency table representing the subset

^{xliv} Of course, the p-value is not truly equal to zero. That is how very low p-values are reported when the computer gives up on the calculation. A quick venture into the internet turns up numerous calculators for p-values, and as [one from STAT200 at University of Illinois at Champagne-Urbana](#) indicated, the statistic is out of bounds with respect to the distribution graph.

of genes that demonstrated no change in gene expression after siRNA-Nup93 as well as no physical association with Nup93-DamID.

Of note is the inability to reject the null hypothesis for two out of the four “sub”- χ^2 -tests performed after stratifying by Nup93 associated assays. The null hypothesis – change in expression after siRNA-Nup153 is independent of DamID-Nup153 – was not rejected for the set of genes that demonstrated no change in gene expression after siRNA-Nup93 and had positive DamID-Nup93 ($\chi^2(1, N = 571) = 0.320, p = 0.571$). Similarly, the null hypothesis – change in expression after siRNA-Nup153 is independent of DamID-Nup153 – is not rejected for the set of genes that changed gene expression after siRNA-Nup93 and show negative DamID-Nup93, ($\chi^2(1, N = 280) = 1.063 \times 10^{-2}, p = 0.918$).

Table 3.5 2×2×2×2 Contingency table of response to siRNA treatment against Nup153 stratified by response to siRNA treatment against Nup93^{a,b} versus DamID for Nup153 stratified by DamID for Nup93.

				Change in expression after treatment with siRNA against...							
				Nup93							
				No			Yes				
				Nup153							
				No	Yes			No	Yes		
Physical Association via DamID by ...	Nup93	No	No	16454	1846	18300	3461	2665	6126		
			Yes	618	97	715	204	196	400		
					17072	1943	19015	3665	2861	6526	
					$\chi^2(1, N = 19015) = 9.078,$ $p = 2.586 \times 10^{-3}$			$\chi^2(1, N = 6526) = 4.608,$ $p = 3.182 \times 10^{-2}$			
	Yes	No	No	286	40	326	81	71	152		
			Yes	211	34	245	69	59	128		
					497	74	571	150	130	280	
					$\chi^2(1, N = 571) = 3.200 \times 10^{-1},$ $p = 0.571$			$\chi^2(1, N = 280) = 1.063 \times 10^{-2},$ $p = 0.918$			

^a Neither DamID nor siRNA treatment were applied to both Nups simultaneously.

^b Change in gene expression was classified as a response without consideration of the direction (up or down).

Assay results were used to assign genes to mutually exclusive subsets as encoded by four binary variables, one per Nup per assay type.

What might be somewhat surprising about these results is the stark contrast in p-value between the 2×2×2×2 table (**Table 3.5**) tested for independence as a 4×4 table (**Table 3.4**) and the individual p-values when the same table is analyzed for what it is, a 2×2×2×2 table, and the results stratified by two of four of the variables. However, the difference in magnitude of the p-values does not indicate whether groups are more or less likely to be independent. The cut-off p-value for the null hypothesis, that the two groups are independent, is set prior for testing the hypothesis.

To look further at the question of how selecting variables to stratify on changes the results of testing for independence, a different set of variables were selected and more χ^2 -tests were performed. The original parameters selected were picked because their place on the “outside” of the $2 \times 2 \times 2 \times 2$ made the resulting tables visually consistent with prior contingency tables; they could be nested within the structure of the existing 4×4 table. Though other options would not produce visual consistency, additional stratification schema was tested (**Tables 3.6a–e**).

Table 3.6a $2 \times 2 \times 2 \times 2$ Contingency table of response to siRNA treatment against Nup153 versus response to siRNA treatment against Nup93^{a,b} stratified by DamID for Nup153 and Nup93 with conditional subsets for χ^2 -testing indicated by cell background color.

				Change in expression after treatment with siRNA against...			
				Nup93			
				No		Yes	
				Nup153			
Physical Association via DamID by...	Nup93	No	No	16454	1846	3461	2665
			Yes	618	97	204	196
	Yes	Nup153	No	286	40	81	71
				Yes	211	34	69

The background colors above correspond with the following contingency “sub-tables”.

Table 3.7b 2×2 Contingency Table of siRNA treatment response to Nup93 versus Nup153^a without positive DamID for Nup93 or Nup153

			Change in expression after treatment with siRNA against...		
			Nup93		
			No	Yes	
Change in expression after treatment with siRNA against...	Nup153	No	16454	3461	19915
		Yes	14920	4995	19915
	Nup93	No	157.642	470.920	
		Yes	1846	2665	4511
		No	3380	1131	
		Yes	695.954	2078.999	
			18300	6126	24426

$$\chi^2(1, N = 124426) = 3403.515, p = 4.907 \times 10^{-38}$$

Table 3.9d 2×2 Contingency Table of siRNA treatment response to Nup93 versus Nup153^a with positive DamID for Nup93 and without positive DamID for Nup153.

Table 3.8c 2×2 Contingency Table of siRNA treatment response to Nup93 versus Nup153^a without positive DamID for Nup93 and with positive DamID for Nup153.

			Change in expression after treatment with siRNA against...		
			Nup93		
			No	Yes	
Physical Association via DamID by...	Nup153	No	618	204	822
		Yes	527	295	822
	Nup93	No	15.671	28.013	
		Yes	97	196	293
		No	188	105	
		Yes	43.966	78.589	
			715	400	1115

$$\chi^2(1, N = 1115) = 166.238, p = 4.907 \times 10^{-36}$$

Table 3.10e 2×2 Contingency Table of siRNA treatment response to Nup93 versus Nup153^a with positive DamID for Nup93 and Nup153.

Change in expression after treatment with siRNA against...		Change in expression after treatment with siRNA against...		
		Nup93		
		No	Yes	
Nup153	No	286	81	367
	Yes	40	71	
Nup153	No	250	117	111
	Yes	76	35	
		5.093	10.924	
		16.838	36.113	
		326	152	478

$\chi^2(1, N = 478) = 68.967, p = 1.101 \times 10^{-16}$

Change in expression after treatment with siRNA against...		Change in expression after treatment with siRNA against...		
		Nup93		
		No	Yes	
Nup153	No	211	69	280
	Yes	34	59	
Nup153	No	184	96	93
	Yes	61	32	
		3.989	7.635	
		12.010	22.988	
		245	59	373

$\chi^2(1, N = 373) = 46.622, p = 8.608 \times 10^{-12}$

^a Neither DamID nor siRNA treatment were applied to both Nups simultaneously.

^b Change in gene expression was classified as a response without consideration of the direction (up or down).

Assay results were used to assign genes to mutually exclusive subsets as encoded by four binary variables, one per Nup per assay type.

Looking at the data from this perspective, the null hypothesis that genes subset based on results of treatment siRNA-Nup93 versus siRNA-Nup153 are independent was rejected regardless of the subset membership conditional on DamID ($0 \leq p \leq 1 \times 10^{-11}$, see **Tables 3.6a–e** for detailed results).

To maintain consistency with the χ^2 -tests above, odds ratios were calculated while holding the DamID variables for Nup93 and Nup153 constant (**Table 3.11**). The genes most least likely to change gene expression after treatment against Nup153 were those who did not demonstrate a change after siRNA treatment against Nup93 and had no physical association with Nup93 or Nup153 (OR = 6.9, 95%CI = 6.4–7.36).

Table 3.11 Odds ratios indicating the likelihood of a gene showing no change in gene expression after siRNA treatment against Nup153 after demonstrating no change in gene expression after siRNA treatment against Nup93 while holding the conditions for physical association with Nup93 and Nup153 (as indicated by DamID) constant.

DamID for Nup93	DamID for Nup153	Genes with <i>NO</i> change after siRNA-Nup93 are _____-fold <i>LESS</i> likely to change expression after siRNA-Nup153.	
		OR	95%CI
None	None	6.9	6.4 – 7.36
None	Yes	6.1	4.6 – 8.2
Yes	None	6.3	4.0 – 9.9
Yes	Yes	5.3	3.2 – 8.8

The odds were also calculated for holding siRNA treatments constant and a similar trend was found, with the likelihood a gene will be physically associated with Nup153 being greatest when a gene shows change in expression after treatment with both siRNA against Nup93 as well as against Nup153 (**Table S3.24**).

From odds and ratios, we shifted to modeling cell frequencies to develop a more holistic model of relationships between assay results, including any interactions that may exist among the conditions.

3.4.1.1 Log-Linear Analysis: Modeling Relationships between Subsets

Log-linear analysis allows us to shift the focus from testing independence and cell means to modeling building based on cell frequencies as a means of further elucidating the relationships between assay results. The GLMs used in log-linear analysis were fit to the data using `glm` function of `stat` using the Poisson family of GLMs and a log link function. The Poisson distribution is generally used for count data; thus it is appropriate for modeling relationships that hinge on cell frequencies. The log link function is useful as it allows non-additive relationships to be modeled as a linear function, which will be explained in more detail as the method is expounded upon. To start, the null, or *equiprobability*, model is established. This model represents the case when no combinations of assay results is any more likely to occur than any other.

$$\ln(F_{ij}) = \lambda \quad (12)$$

The model is fit to the data, fitting the data under the assumption none of the four variables have any effect on the classification in any given subgroup, which are mutually exclusive and defined by the assay results encoded by four binary variables.

Table 3.12 Estimated intercept of the null model for the frequency of classification into gene subsets defined by DamID siRNA treatment assays for Nup93 and Nup153

	Coefficient	SE	z-value	p-value
Intercept	7.408	0.006	1203.591	0

$G^2(15, N = 26392) = 76145, p = 0$
AIC = 76267

Subsequently, the simplest model (Eq. 13) was constructed and fit to the data (**Table 3.13**).

$$\ln(F_{ij}) = \lambda + \lambda_i^{DamID\ Nup93} + \lambda_j^{DamID\ Nup153} + \lambda_k^{siRNA\ Nup93} + \lambda_l^{siRNA\ Nup153} \quad (13)$$

As expected, has an overall p-value of 0, meaning there is a significant difference between the observed and expected frequencies for each cell and the ratio between the two approaches 0.

Table 3.13 Estimated coefficients of the simple log-linear model for the relative frequencies of gene subsets as classified by DamID siRNA treatment assays for Nup93 and Nup153

	Coefficient	SE	z-value	p-value
λ	5.217	0.02190	238.38	0
$\lambda_i^{DamID\ Nup93}$	1.701	0.01740	97.62	0
$\lambda_j^{DamID\ Nup153}$	1.409	0.01330	105.58	0
$\lambda_k^{siRNA\ Nup93}$	0.726	0.00785	92.47	0
$\lambda_l^{siRNA\ Nup153}$	0.529	0.00704	75.12	0

$G^2(11, N = 26392) = 4541.708, p = 0$
AIC = 4671.2

As a contrast to the simple model, the fully saturated model was also constructed (Eq. 14) and fit to the data (**Table 3.14**).

$$\begin{aligned}
& \ln(F_{ij}) \\
& = \lambda + \lambda_i^{DamID\ Nup93} + \lambda_j^{DamID\ Nup153} + \lambda_k^{siRNA\ Nup93} + \lambda_l^{siRNA\ Nup153} + \lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \\
& + \lambda_i^{DamID\ Nup93} \lambda_k^{siRNA\ Nup93} \\
& + \lambda_i^{DamID\ Nup93} \lambda_l^{siRNA\ Nup153} + \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} + \lambda_j^{DamID\ Nup153} \lambda_l^{siRNA\ Nup153} \\
& + \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153} + \lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} \\
& + \lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_l^{siRNA\ Nup153} + \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153} \\
& + \lambda_i^{DamID\ Nup93} \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153} + \lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153}
\end{aligned} \tag{14}$$

Not surprisingly, the model has an overall p-value of 1, meaning there is no significant difference between the observed and expected frequencies for each cell – i.e. the ratio between the two approaches is 1. This means that this model, the saturated model, is the best representation of the specific set of data used to construct it. However, it risks (almost guarantees) overfitting the data, making it less useful as a predictive model.

Table 3.14 Estimated coefficients of the saturated log-linear model for the relative frequencies of gene subsets as classified by DamID siRNA treatment assays for Nup93 and Nup153

	Coefficient	SE	z-value	p-value
λ	5.628	0.024	237.55	0
$\lambda_i^{DamID\ Nup93}$	1.229	0.024	51.891	0
$\lambda_j^{DamID\ Nup153}$	0.780	0.024	32.923	0
$\lambda_k^{siRNA\ Nup93}$	0.526	0.024	22.215	0
$\lambda_l^{siRNA\ Nup93}$	0.178	0.024	7.521	0
$\lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153}$	0.679	0.024	28.638	0
$\lambda_i^{DamID\ Nup93} \lambda_k^{siRNA\ Nup93}$	0.016	0.024	0.685	0.493
$\lambda_i^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93}$	0.042	0.024	1.778	0.075
$\lambda_i^{DamID\ Nup93} \lambda_k^{siRNA\ Nup153}$	0.021	0.024	0.903	0.366
$\lambda_i^{DamID\ Nup153} \lambda_k^{siRNA\ Nup153}$	0.057	0.024	2.395	0.017
$\lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153}$	0.453	0.024	19.104	0
$\lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93}$	0.028	0.024	1.161	0.245
$\lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup153}$	0.042	0.024	1.756	0.079
$\lambda_i^{DamID\ Nup93} \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup93}$	0.015	0.024	0.616	0.538
$\lambda_i^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup93}$	0.018	0.024	0.741	0.459
$\lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup93}$	-0.003	0.024	-0.137	0.891

$$G^2(0, N = 26392) = -1.520 \times 10^{-10}, p = 1$$

$$AIC = 151.49$$

Comparing models via their goodness-of-fit by manually evaluating the benefit of adding or leaving out a term is possible; but it quickly becomes computationally prohibitive. In the case of four binary variables, there are more than one hundred possible models to compare. Thus, an algorithm was employed to find the balance between goodness-of-fit (without overfitting) while retaining as much information as possible by minimizing the Akaike Information Criterion (AIC). The resulting model from running `stats::step` starting with the saturated model (Eq. 14) is shown in Eq. 15 while the calculated values for the coefficients and their affiliated measures of significance can be found in Table 3.15.

$$\begin{aligned}
& \ln(F_{ij}) \\
& = \lambda + \lambda_i^{DamID\ Nup93} + \lambda_j^{DamID\ Nup153} + \lambda_k^{siRNA\ Nup93} + \lambda_l^{siRNA\ Nup153} \\
& + \lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} + \lambda_j^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93} + \lambda_j^{DamID\ Nup93} \lambda_k^{siRNA\ Nup153} \\
& + \lambda_j^{DamID\ Nup153} \lambda_l^{siRNA\ Nup153} + \lambda_j^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153} + \lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_l^{siRNA\ Nup153}
\end{aligned} \tag{15}$$

Table 3.15 Estimated coefficients of a log-linear model for the relative frequencies of gene subsets as classified by DamID siRNA treatment assays for Nup93 and Nup153 as algorithmically optimized using the Akaike Information Criterion (AIC).

	Coefficient	SE	z-value	p-value
λ	5.611	0.021	264.246	0
$\lambda_i^{DamID\ Nup93}$	1.242	0.02	62.022	0
$\lambda_j^{DamID\ Nup153}$	0.777	0.021	36.949	0
$\lambda_k^{siRNA\ Nup93}$	0.546	0.017	32.582	0
$\lambda_l^{siRNA\ Nup93}$	0.161	0.021	7.585	0
$\lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153}$	0.689	0.02	34.422	0
$\lambda_i^{DamID\ Nup153} \lambda_k^{siRNA\ Nup93}$	0.065	0.017	3.849	0
$\lambda_i^{DamID\ Nup93} \lambda_k^{siRNA\ Nup153}$	0.033	0.02	1.642	0.101
$\lambda_i^{DamID\ Nup153} \lambda_k^{siRNA\ Nup153}$	0.056	0.021	2.658	0.008
$\lambda_k^{siRNA\ Nup93} \lambda_l^{siRNA\ Nup153}$	0.479	0.008	56.423	0
$\lambda_i^{DamID\ Nup93} \lambda_j^{DamID\ Nup153} \lambda_l^{siRNA\ Nup153}$	0.050	0.02	2.496	0.013

$$G^2(5, N = 26392) = 3.779, p = 0.582$$

$$AIC = 145.26$$

When looking at the results of log-linear analysis, keeping in mind the null hypothesis being tested is helpful. The null hypothesis is that the parameters used to build the model and estimated response are independent, which mean the model had no predictive value. It is one of those rare occasions where $p < 0.05$ is not sought after. Therefore, the model could be worse (and likely could be better – the value of the model is in the eye of the inquirer).

Both the saturated and optimized models are an improvement over the simple model (AIC = 151.49, $p = 1$ and AIC = 145.26, $p = 0.582$, versus AIC = 4671.2, $p = 0$). It is counterintuitive to think taking things away would improve a model; it goes against human nature as we currently understand it.^{153–155} This method, however, demonstrates that the removal of extraneous interaction terms increases the chance that the contribution of those that will describe the relationships between the data. The contributions of each individual assay results is harder to interpret with this particular type of model, both because of the log-link function as well as the number of interaction terms. What it one can do easily is note the what has been removed: two interaction terms involving siRNA-Nup93. What this might suggest is that the effect siRNA-Nup93 is less influential over the results of other assay results and/or that other assay results are less influential over siRNA-Nup93 outcomes.

Table 3.16 Comparison of estimated coefficients of the saturated and “optimized” models of relative frequencies of gene subsets as classified by DamID siRNA treatment assays for Nup93 and Nup153

	Coefficient		SE		z-value		p-value	
	Sat. ^a	Opt.	Sat.	Opt.	Sat.	Opt.	Sat.	Opt.
λ	5.628	5.611	0.024	0.021	237.55	264.246	0	0
$\lambda_i^{DamID\ Nup93}$	1.229	1.242	0.024	0.020	51.891	62.022	0	0
$\lambda_j^{DamID\ Nup153}$	0.780	0.777	0.024	0.021	32.923	36.949	0	0
$\lambda_k^{siRNA\ Nup93}$	0.526	0.546	0.024	0.017	22.215	32.582	0	0

λ_k siRNA Nup93	0.178	0.161	0.024	0.021	7.521	7.585	0	0
λ_j DamID Nup93 λ_j DamID Nup153	0.679	0.689	0.024	0.020	28.638	34.422	0	0
λ_j DamID Nup93 λ_k siRNA Nup93	0.016	–	0.024	–	0.685	–	0.493	–
λ_j DamID Nup153 λ_k siRNA Nup93	0.042	0.065	0.024	0.017	1.778	3.849	0.075	0
λ_j DamID Nup93 λ_k siRNA Nup153	0.021	0.033	0.024	0.020	0.903	1.642	0.366	0.101
λ_j DamID Nup153 λ_k siRNA Nup153	0.057	0.056	0.024	0.021	2.395	2.658	0.017	0.008
λ_k siRNA Nup93 λ_k siRNA Nup153	0.453	0.479	0.024	0.008	19.104	56.423	0	0
λ_j DamID Nup93 λ_j DamID Nup153 λ_k siRNA Nup93	0.028	–	0.024	–	1.161	–	0.245	–
λ_j DamID Nup93 λ_j DamID Nup153 λ_k siRNA Nup153	0.042	0.050	0.024	0.020	1.756	2.496	0.079	0.013
λ_j DamID Nup93 λ_k siRNA Nup93 λ_k siRNA Nup93	0.015	–	0.024	–	0.616	–	0.538	–
λ_j DamID Nup153 λ_k siRNA Nup93 λ_k siRNA Nup93	0.018	–	0.024	–	0.741	–	0.459	–
λ_j DamID Nup93 λ_j DamID Nup153 λ_k siRNA Nup93	-0.003	–	0.024	–	-0.137	–	0.891	–
λ_k siRNA Nup93								
	Saturated Model				Optimized Model			
	$G^2(0, N = 26392) = -1.520 \times 10^{-10}$				$G^2(5, N = 26392) = 3.779$			
^a Saturated Model (Sat.)	$p = 1$				$p = 0.582$			
^b Optimized Model (Opt.)	AIC = 151.49				AIC = 145.26			

3.4.2 Integration of the RA GWAS Gene List

Of the 105 genes on the RA GWAS gene list, 25 were not initially connected to results in the assay dataset. Of those 25, 16 were variants; this was determined by searching for a variation of the gene symbol, one which was truncated at the hyphen, among those genes for which there were assays results ($n = 17$) and confirming the result was on the same chromosome as that on the RA GWAS list ($n = 16$, as one was not). The total genes connected to genes in the Nup93/Nup153 dataset rose back to 17 when an alias for *C4orf52* (*SMIM20*) was found determined to be included in the assay dataset, as determined by HUGO's Multi-Symbol Checker.¹⁵⁶ In total, 8 genes were not linked to the assay dataset after searching for aliases via HUGO and/or did not have a truncated symbol that located on the same chromosome as the GWAS list indicated.

Even with this truncated list (down from 105 to 97), there is a greater than 1.44-fold chance of drawing an RA GWAS gene from the set of 1966 genes that are positive via DamID for at least one of the Nup93 or Nup153 than drawing an RA risk factor gene from the set all genes (.56% versus 0.39%). However, the null hypotheses of standard χ^2 -tests, that membership on the RA GWAS was independent of DamID classification ($\chi^2(3, N = 26392) = 3.967, p = 0.2651$) or siRNA classification ($\chi^2(3, N = 26392) = 3.099, p = 0.3766$), could not be rejected.

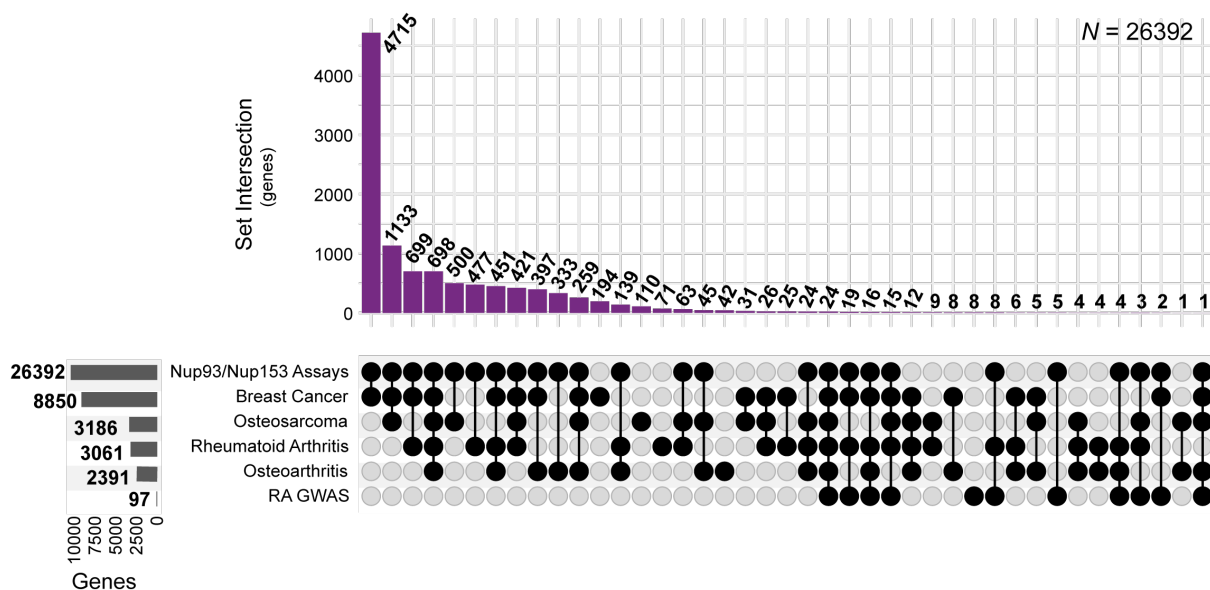


Figure 3.7 Intersections of siRNA and/or DamID for Nup93 or Nup153 with Potential Target Genes grouped by disease as identified by the Open Target Platform

Genes with at least one positive siRNA or DamID assay for Nup93 and/or Nup153 were compared with the gene target lists identified by the Open Targets platform. The lists for each disease were not mutually exclusive. Looking at set overlap was the first step in looking for possible associations between the positive assay results and association with disease.

The data used was retrieved in January 2021.

3.4.3 Integration with Open Targets Datasets

The Open Targets datasets were examined for overlap alongside the RA GWAS gene list and Nup93/Nup153 siRNA and DamID assay dataset. These results were visualized using an UpSet plot (**Fig. 3.7**); the genes in the assay dataset that did not appear in any other dataset were not included in the graph as the number of such genes (15385) was approximately 3-fold greater than the next largest intersection, that of the assay dataset with the OT BC dataset (4715). Dividing the data by disease association, a χ^2 -test was performed per disease to test for independence between the assay results. In the case of all four diseases, the null hypothesis – DamID and siRNA assay results were independent – (remained) rejected (RA: $\chi^2(9, 3061) = 31.80$, $p = 1.6 \times 10^{-4}$; OS: $\chi^2(9, 3186) = 25.11$, $p = 2.8 \times 10^{-3}$; OA: $\chi^2(9, 2391) = 29.90$, $p = 4.6 \times 10^{-4}$; BC: $\chi^2(9, 8850) = 58.56$, $p = 2.5 \times 10^{-9}$). In all four tests, the top contributor to the χ^2 -statistic was the case when DamID-Nup153 was positive.

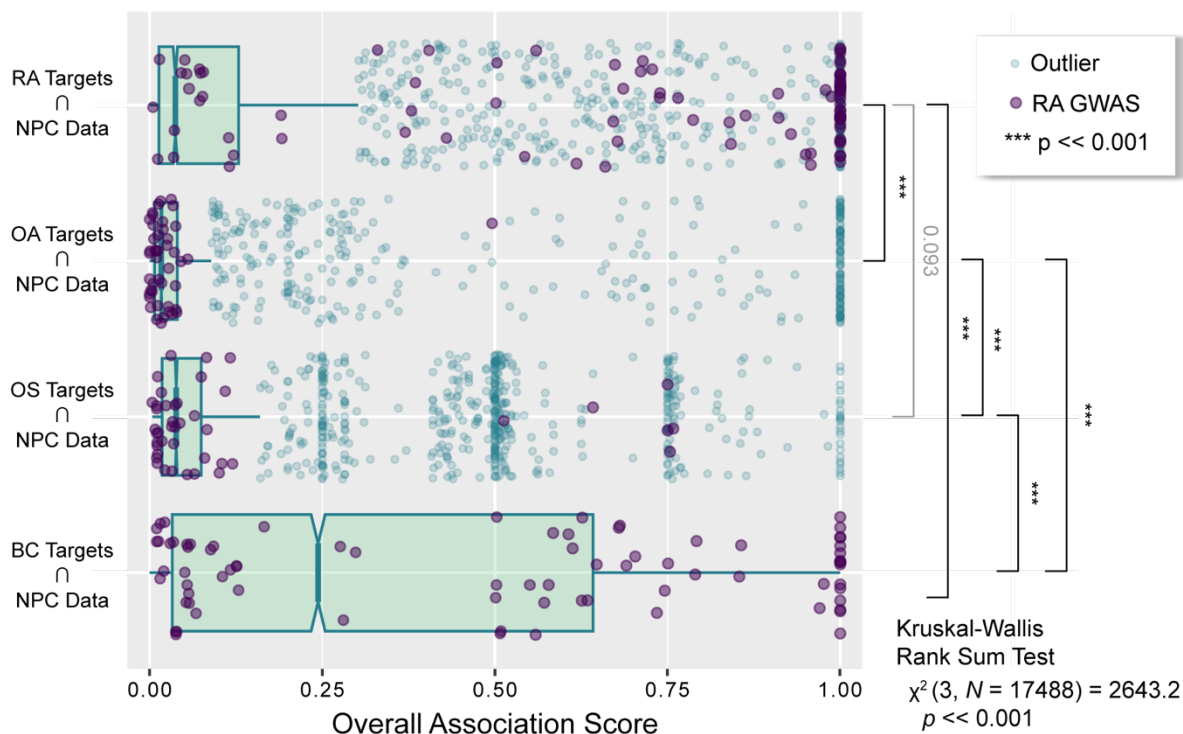


Figure 3.8 Boxplots depicting the distribution of overall association scores for genes in the intersection of the NPC assay datasets and Open Target data, stratified by disease.

The distribution of overall association scores for the given disease (if it exists) for those genes RA GWAS gene list is also included, as indicated by the purple dots. The results of the Pairwise Wilcoxon Rank Sum Test are next to the bracket indicating the pair of diseases compared. A statistically significant result translates to a rejection of the null hypothesis that there is no difference in the means compared.

Presented with continuous data (overall association scores) we visualized distributions of association scores by disease were depicted as box plots with genes in the RA GWAS dataset visualized as points on the boxplots (**Fig. 3.8**). The results of this were visually striking, with genes on the RA GWAS list showing strong association scores with both RA and BC. The median association score for BC was also greater than that of any of the other three diseases as was the interquartile range (**Table 3.17**).

Table 3.17 Summary statistics for the distribution of overall association scores for each disease

	median	IQR	mean
RA	0.037	0.014 – 0.129	0.146
OA	0.017	0.007 – 0.040	0.067
OS	0.038	0.018 – 0.074	0.114
BC	0.244	0.033 – 0.642	0.346

As was expected, a greater percentage of the genes were shared between the Open Targets data and the RA GWAS gene list than any other group (91.75%), see **Table 3.18** for all results.

Table 3.18 Comparing the percent of RA GWAS genes included in each set of disease-specific genes

	Genes in assay data	Percent of NPC Genes ($N = 26392$)	In RA GWAS	% of RA GWAS in assay data ($n = 97$)	% of gene targets shared with RA GWAS
RA	3061	11.60	89	91.75	2.91

OA	2391	9.06	45	46.39	1.88
OS	3186	12.07	43	44.33	1.35
BC	8850	33.53	77	79.38	0.87

The distribution of the association scores were checked for normality by visually inspected by plotting density plots of the scores as well as by the Shapiro-Wilk Normality Test, where possible (the Open Targets Breast Cancer dataset had too many members). The distribution of the scores was not found to be normal ($p < 2.2 \times 10^{-16}$). The Kruskal-Wallis rank sum test was performed to determine if, when stratified by gene membership on the RA GWAS list, the association scores of at least one Open Targets dataset was different from the others (Kruskal-Wallis $\chi^2(3, N = 17488) = 2643.2, p < 2.2 \times 10^{-16}$). To determine which groups significantly differed, pairwise comparisons using Pairwise Wilcoxon Rank Sum Test with continuity correction were made and Bonferroni correction for multiple comparisons applied (**Fig. 3.8**). The only groups that did not have a significant difference in means were RA and OS ($p = 0.093$).

To look for associations between the overall association scores for a disease and the assay results was sought to fit a GLM. The distribution and log link function were selected after looking at the distribution of scores using tools in the `fitdistrplus` package.¹⁵⁷ plotting empirical versus theoretical cumulative sum functions, a qq-plot, and a Cullen and Frey plot of the square of the skew versus kurtosis.¹⁵⁸ The overall picture was one of a log-normal distribution. As a result, a GLM with a Gaussian distribution and a log link function were used to model the disease association scores as a function of binary variables with values determined by assay results.

$$\log(\hat{y}) = \beta_0 + \beta_{\text{DamID_Nup93}}x_{\text{DamID_Nup93}} + \beta_{\text{DamID_Nup153}}x_{\text{DamID_Nup153}} + \beta_{\text{siRNA_Nup93}}x_{\text{siRNA_Nup93}} + \beta_{\text{siRNA_Nup153}}x_{\text{siRNA_Nup153}} + \quad (16)$$

Significance of the overall models was calculated using the likelihood ratio test with the null model, $\log(\hat{y}) = \beta_0$, for comparison.

Table 3.19 Generalized Linear Model of RA association scores as a function of siRNA and DamID assay results

	Coefficient	SE	t-value	p-value
β_0	-1.825	0.071	-25.700	0***
$\beta_{\text{DamID_Nup93}}$	-0.055	0.070	-0.777	0.437
$\beta_{\text{DamID_Nup153}}$	-0.072	0.054	-1.343	0.179
$\beta_{\text{siRNA_Nup93}}$	0.015	0.035	0.433	0.665
$\beta_{\text{siRNA_Nup153}}$	0.019	0.038	0.491	0.623

$\rho_{\text{model}} = 0.454$ *** $p < 0.001$

Reduction of the model down to only include β_0 and $\beta_{\text{siRNA_Nup153}}$ did not improve the model as determined by the likelihood ratio test.

Table 3.20 Generalized Linear Model of OA association scores as a function of siRNA and DamID assay results

	Coefficient	SE	t-value	p-value
--	-------------	----	---------	---------

β_0	-2.543	0.174	-14.637	0***
$\beta_{\text{DamID_Nup93}}$	0.290	0.174	1.662	0.097
$\beta_{\text{DamID_Nup153}}$	-0.174	0.075	-2.309	0.021*
$\beta_{\text{siRNA_Nup93}}$	0.019	0.055	0.341	0.733
$\beta_{\text{siRNA_Nup153}}$	0.062	0.061	1.013	0.311

$\rho_{\text{model}} = 0.149$ * $p < 0.05$, *** $p < 0.001$

Reduction of the model down to only include β_0 and $\beta_{\text{siRNA_Nup153}}$ did not improve the model as determined by the likelihood ratio test.

Table 3.21 Generalized Linear Model of OS association scores as a function of siRNA and DamID assay results

	Coefficient	SE	t-value	p-value
β_0	-2.093	0.067	-31.430	0***
$\beta_{\text{DamID_Nup93}}$	-0.008	0.067	-0.119	0.905
$\beta_{\text{DamID_Nup153}}$	-0.126	0.05	-2.509	0.012*
$\beta_{\text{siRNA_Nup93}}$	0.088	0.036	2.476	0.013*
$\beta_{\text{siRNA_Nup153}}$	0.015	0.037	0.398	0.691

$\rho_{\text{model}} = 0.008 < 0.01$ * $p < 0.05$, *** $p < 0.001$

Reduction of the model down to only include β_0 and $\beta_{\text{siRNA_Nup153}}$ did improve the model, making its comparison to the null significant as determined by the likelihood ratio test.

Table 3.22 Generalized Linear Model of OS association scores as a function of DamID-Nup153

	Coefficient	SE	t-value	p-value
β_0	-2.073	0.047	-44.431	0***
$\beta_{\text{siRNA_Nup153}}$	-0.115	0.047	-2.467	0.014*

$\rho_{\text{model}} = 0.024 < 0.05$ * $p < 0.05$, *** $p < 0.001$

Reduction of the model down to only include β_0 and $\beta_{\text{siRNA_Nup93}}$ ($p = 0.549$) did not improve the model nor did adding $\beta_{\text{siRNA_Nup153}}$ to the model $\log(\hat{y}) = \beta_0 + \beta_{\text{DamID_Nup153}}$ ($p = 0.945 > 0.05$).

Table 3.23 Generalized Linear Model of BC association scores as a function of siRNA and DamID assay results

	Coefficient	SE	t-value	p-value
β_0	-1.043	0.028	-37.145	0***
$\beta_{\text{DamID_Nup93}}$	0.026	0.027	0.955	0.340
$\beta_{\text{DamID_Nup153}}$	-0.056	0.020	-2.830	0.0047**
$\beta_{\text{siRNA_Nup93}}$	-0.011	0.012	-0.890	0.374

$\beta_{\text{siRNA_Nup153}}$	0.012	0.013	1.400	0.162
--------------------------------	-------	-------	-------	-------

** $p < 0.01$, *** $p < 0.001$

Reduction of the model down to only include β_0 and $\beta_{\text{siRNA_Nup153}}$ did improve the model, making its comparison to the null significant as determined by the likelihood ratio test.

The coefficients for the resulting models are given in **Tables 3.19–3.23**. In three out of four models, the only binary variable with a significant correlation with the response variable (overall association score) was DamID-Nup153. Subsetting all four disease datasets to just include RA GWAS gene list observations came up with no significant correlation with any of the predictor variables. This lack of correlation was not surprising given the null hypothesis that membership on the RA GWAS was independent of DamID classification ($\chi^2(3, N = 26392) = 3.967, p = 0.2651$) or siRNA classification ($\chi^2(3, N = 26392) = 3.099, p = 0.3766$) via the standard χ^2 -test.^{xlv} All four of the models, when compared to the null, did not stand up to scrutiny. When reduced to only the intercept and DamID-Nup153 term, the only model of note was that for the overall association score of osteosarcoma.

3.5 Discussion

There are many flavors of multiomic data integration, from the integration of single cell (sc) omics data¹⁵⁹ to the exploration of metabolic networks¹⁶⁰, multi-omic plant data¹⁶¹ ... the list goes on. As the amount of available data continues to grow, for those who have big questions that can only be answered by dissecting a few datasets and merging the carefully crafted results, it is like being a kid in the proverbial candy shop but, unlike a kid, researchers can afford the candy.^{xlvi}

So much of “doing science” hinges on the ability to compare and contrast different conditions, which makes the growing availability of public data all the more exciting. The ability to develop research questions through preliminary data analysis not only saves resources, but it broadens possibilities. These possibilities include crinkling up that hypothesis you wrote on the back of a napkin, not because it sounded stupid the next time you read it, but because you found some data, crunched some numbers, and realized it was likely a dead end.

Funding sources are understandably risk adverse, so unless you’re someone like <insert the name of a tech billionaire here>, even if that now-smoothed-out napkin idea is that start of an innovative solution to a stubborn problem, the merit of the idea must be established before dedicating significant resources to it. Within this paper, we pursued just such an idea using publicly available datasets. As the data were integrated and analyzed, considerations for such exploratory analyses have been collected. It is through this lens that the results are discussed.

We set out with the following train of thought guiding the research:

If association with the nuclear pore complex (NPC) is linked to cell identity – both by their average density of insertion across the nuclear envelope as well as their physical association with cell-type specific areas of active transcription – and disease-types are often (to some degree) cell-

^{xlv} These results occurred despite being unable to reject the null hypothesis that all three were independent via the Cochran-Mantel-Haenszel Chi-Squared Test for Count Data ($M^2(9, N = 26392) = 108, p \ll 0.001$). This seemingly odd result is not surprising, given the null hypothesis was rejected regarding the relationships between DamID and siRNA classification (see Table 3.3), but this author has been waiting for a chance to apply that test after being unable to earlier.

^{xlvi} Please do not look too hard at this metaphor, it will fall apart.

type specific, then the expression of genes associated with a disease process may be linked to the NPC.

Rheumatoid arthritis was selected as our disease of interest because a list RA risk factor genes had been compiled, updated, and verified by considering the connections between genes on the list with existing drug targets through biological networks.¹⁰¹ With this list in hand, we sought data that connected cellular identity with physical contact with the NPC within a cell line used to model RA, which we found in the Nup93/Nup153 datasets.⁴² To connect disease identity to cell identity and these assays, we sought data with predictive association scores between genes and diseases, which is what the Open Targets Platform aggregates.^{103,162–164} Thus, the first step toward flushing out our research question was successful: we found data that connected the different parts of our (fairly broad) research question.

The summary statistics gathered at the outset were promising and were copesetic with what is known about the structure of the NPC. For example, the higher incidence of methylation by DamID-Nup153 than DamID-Nup93 may be a result of: 1. The more accessible location of Nup153 as part of the nucleoplasm facing basket as compared to the channel constituent, Nup93¹⁶⁵ and, 2. the reputation of Nup153 existing within the nucleus independent of the NPC.¹⁶⁶ It is possible that the fraction of DamID-Nup153 genes that are not associated with DamID-Nup93 as well are those that exist freely within in the nucleus.

Given that a higher number of genes are physically associated with Nup153 than Nup93, one might think that a higher percentage of genes would change expression with Nup153 knockdown. Instead, a higher percent of genes are knocked down with siRNA-Nup93 (25.8% versus 19.0%), with a greater odds of siRNA-Nup153 responsive if responsive to siRNA-Nup93 (OR = 6.829, 95%CI=6.390095–7.298249). However, these results are not off-brand for Nup93, as it has been frequently indited as having a crucial role in gene regulation.^{43,43,44}

Juxtaposing the p-values associated with χ^2 -testing of genes subset through stratification by both DamID (**Tables 3.6b–e**) and those stratifying by response to siRNA-Nup93 and DamID-Nup93 (**Table 3.5**), the role stratifying variables play in the results of testing for independence between groups is readily apparent ($0 \leq p \leq 1 \times 10^{-11}$ versus $10^{-4} \leq p \leq 1$). This trend in p-values is noteworthy specifically because one stratification schema rejects the null in all cases while the other only rejects the null in some cases. While the comparative magnitudes of p-values are not indicative of the relative strength of any association, the influence of the assay type versus the Nup assayed is highlighted. The complexities of the relationships the relationships between DamID and siRNA conditional subsets are detailed in the [Results](#); the main takeaway is the relationship between the two is far more complicated than comparisons between groups can describe.

To model these intricacies, we applied [log-linear analysis](#) to develop a GLM that includes interaction terms. Although [the model](#) optimized on the AIC consisted primarily of terms making a significant contribution, the two- and three- way interaction terms make articulating the effect any one perturbation may have on frequency counts cumbersome. What can be said based on what was included in that specific model is that it appears siRNA-Nup93 may have fewer connections with the other three assay results. To flush out these relationships, a more quantitative analysis of the results may prove useful; that is, including the fold-change in expression or, at minimum, the direction of that change. A further direction to that end would likely superimpose the results thereof onto known biological networks to see what unconsidered connections may exist that influence the magnitude of these interactions (if any) or where they stem from: *are the interaction*

terms of the model representing a directly mechanistic relationship between Nup153 and Nup93 or the result of perturbing this corner of the cellular network?

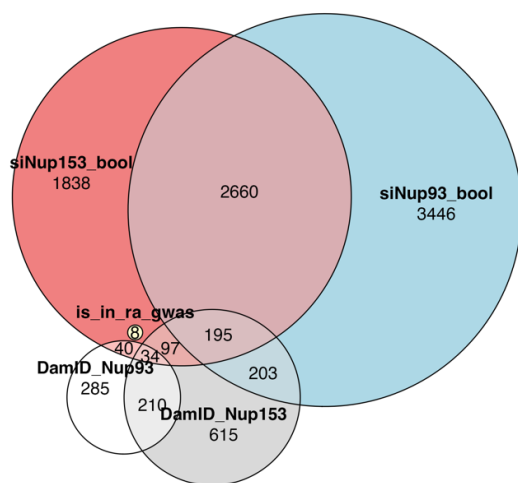
Connecting Open Targets data as well as the RA GWAS data was less successful than we would have liked, though their analyses was not without merit. For all four diseases, the dependence between gene subsets as defined by DamID and siRNA remained. This persisting dependence does not say anything particular about the diseases, as the results were in U2OS cells. What it does add to is the strength of the argument that there is a link between gene expression and association with the NPC. Trying to model the overall disease association score as a function of DamID and siRNA assay results was almost entirely unremarkable. Within the context of our analysis up to this point, the correlation coefficient for DamID-Nup153 found in the models for OS, OA and BC (not in RA) was a significant contributor to the association score for each of these three diseases, respectively. However, none of these models held up when compared to the null via the likelihood ratio test; three out of four did not hold up when reduced to the intercept and the DamID-Nup153 term. Of note, and perhaps significant note, is that such model for the overall association score of osteosarcoma did past the muster of the likelihood ratio test. Given the U2OS cell line, from which these data were derived, is an osteosarcoma-based line, these results do support the efficacy of integrating data in this way. This methodology could be extended to enable tailoring of gene targets in similar experiments to a cell line of interest, specifically those derived from cells in a disease state. To improve upon this method, additional work that considers the raw data, such as the impact the degree of change in cell expression has on disease association (if any), as dependency on pre-processed results with an unknown thresholds is a current (surmountable) limitation of the methodology.

3.6 Conclusion

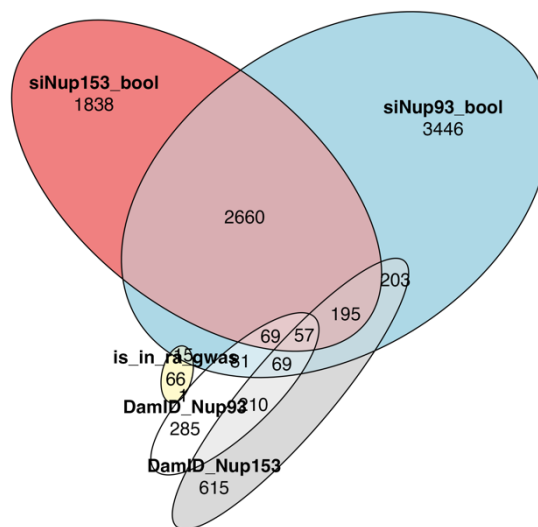
As researchers, it is frustrating to analyze data and find what, for the most part, seems to amount to nothing. However, what we found here was not nothing, it just does not definitively support our original hypothesis. Instead, the analysis present here suggests further analysis is merited, at the level of the raw data, and perhaps with the inclusion of alternative “omic” views of the problem space (e.g. ChIP-seq and Hi-C, given the focus on physical association and regions of active transcription). It also is a head nod to the understated importance of an appropriate biological model for the system being studied; the strong the need for models that reflex the complexity of diseases such as RA is also highlighted in these findings. This comes as no surprise to that community but perhaps underscoring this point will help we the over-eager data crunchers keep in mind the difference between statistically and practically significant.

3.7 Supplemental Materials

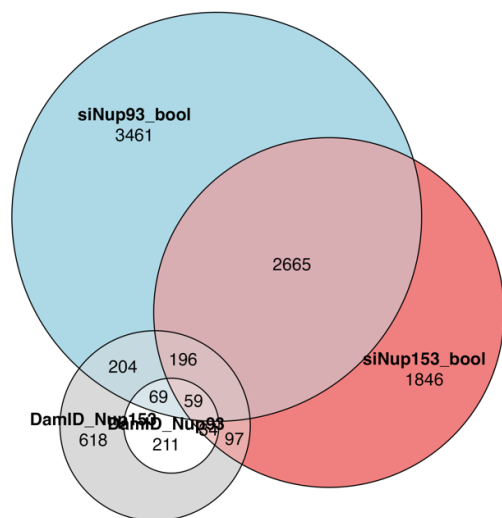
a


 $diagError \approx 0.0158$

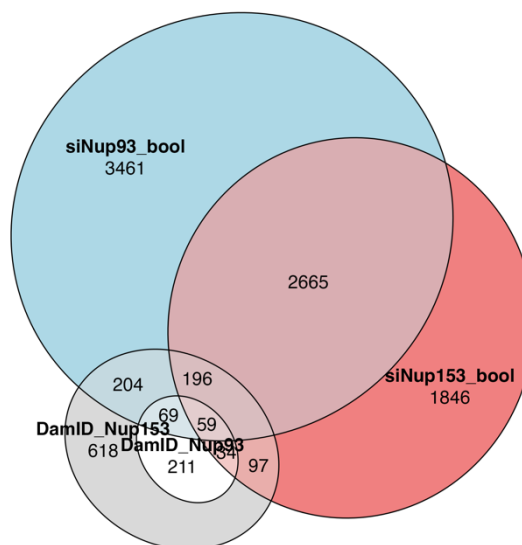
b


 $diagError \approx 0.00970$

c


 $diagError \approx 0.0288$

d


 $diagError \approx 0.0288$

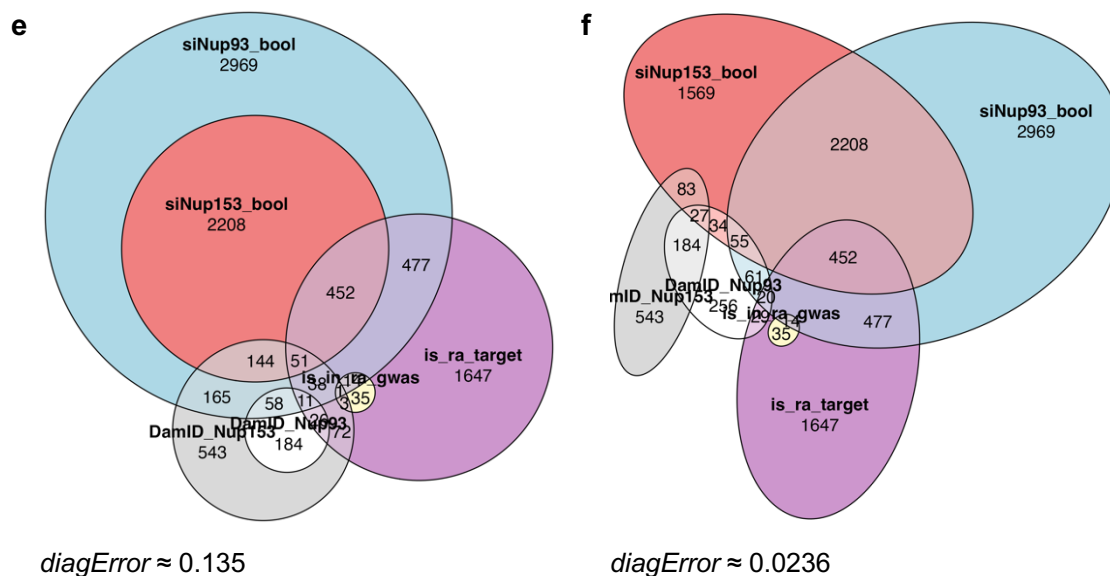


Figure S3.1 Euler diagrams of gene sets

Genes were categorized by positive assays for DamID and siRNA with respect to Nup93 and Nup153 as well as gene list presented in "Genetics of rheumatoid arthritis: 2018 status" (Fig. 3.4). The graphs were created using *eulerr*, an R package that has its underpinnings based on *eulerAPE*, which draws area-proportional circular and elliptical 3-Venn diagrams. Specifically, the optimization of a cost function is driven by the convergence of the diagonal error to ϵ , which is defined as 10^{-6} for computational purposes. This method endeavors to calculate the size and arrangement of overlapping circles (or ellipses) representing sets such that set sizes and union sizes are represented by area. *eulerr* applies extends this method beyond 3-Venns, though the results are not always ideal.

The application of this package to the data presented here was not successful. The *diagError* of all of these diagrams were $\gg \epsilon$. There were multiple discrepancies in visualization that resulted. (a) A Euler diagram depicting the same sets and intersections as shown in the Upset plot ultimately included in the manuscript. The union of genes positive for change after siRNA treatment for Nup153 and those on the RA GWAS list is indicated to have 8 members by the Euler diagram when said intersection contained 26 genes.

Table S3.24 Odds ratios indicating the likelihood of a gene being physically associated with Nup153 if it is not physically associated with Nup93 (as indicated by DamID) while holding the conditions treatment with siRNA against Nup93 xor Nup153 constant.

Change siRNA Nup93	after against	Change after siRNA against Nup153	If a gene is not physically associated with Nup93, that gene is _____-fold less likely to be physically associated with Nup153.
None		None	19.6
None		Yes	14.4
Yes		None	16.2
Yes		Yes	11.3

Table S3.25 Results of pairwise comparisons of all four Open Targets Datasets stratified by gene membership on the RA GWAS list using Wilcoxon rank sum test with continuity correction and Bonferroni correction for multiple comparisons.

	RA GWAS?	BC		OS		OA		RA
		Yes	No	Yes	No	Yes	No	Yes
BC	No	1.545 x 10 ⁻³	-	-	-	-	-	-
OS	Yes	2.892 x 10 ⁻⁶	2.937 x 10 ⁻³	-	-	-	-	-
	No	2.275 x 10 ⁻¹⁹	1.778 x 10 ⁻²⁰³	1	-	-	-	-
OA	Yes	1.894 x 10 ⁻¹⁴	5.487 x 10 ⁻¹⁵	1.601 x 10 ⁻³	4.871 x 10 ⁻⁹	-	-	-
	No	1.946 x 10 ⁻²⁶	0	1.155 x 10 ⁻²	1.448 x 10 ⁻¹¹⁰	1	-	-
RA	Yes	2.332 x 10 ⁻²	1.530 x 10 ⁻¹⁶	1.025 x 10 ⁻¹¹	1.902 x 10 ⁻³⁶	9.237x 10 ⁻¹⁸	2.750 x 10 ⁻³⁹	-
	No	2.524 x 10 ⁻¹⁸	7.494x 10 ⁻²¹⁴	1	2.283 x 10 ⁻²	8.550 x 10 ⁻⁶	6.858 x 10 ⁻⁶¹	2.516 x 10 ⁻³⁴

4 Exploratory Analysis of Hi-C Data for Gene-Gene Interactions

Gene expression is dictated by many factors including physical accessibility for transcription. Patterns of gene expression differ between cell types as do observable components of genome organization. It follows that genome organization drives cell identity.⁷⁵ Evidence points to the nuclear pore complex (NPC) as a director of nuclear organization, with chromatin associated with them being under active transcription.⁴² This activity is a contrast to the adjacent dead zones of transcriptomic activity around the nuclear periphery that are considered “on brand” for the region. This reputation comes courtesy of nuclear lamina-associated domains (LADs) that are regions of inactive chromatin mainly located at nuclear periphery^{xlvii}.¹⁶⁷ Therefore, it is reasonable to consider the nuclear pore as a regulatory mechanism influencing cell identity.¹⁶⁸

In our [prior work](#), we examined the connection between the physical association of chromatin with NPCs and cell identity, which was inclusive of type and disease state. To represent the NPC, two nucleoporins, Nup93 and Nup153, were selected. Nup153 is more accessible as a constituent of the basket extending into the nucleoplasm, while Nup93 is in the channel. Chromatin contact with Nup93 in conjunction with Nup153 provides “2-Factor Authentication” for NPC association via DamID.

Our objective is to capture information that may be lost by practices such as normalization rooted in global statistics and anchor points established at the “origin” (first base pair) of a given chromosome. Our proposed methodology combines an alternative way to visualization count data with a probabilistic framework that reflects our primary research question:

How likely is it that these regions will touch?

In principle, what is proposed here to address this question is extensible to any Hi-C dataset.

By selectively filtering Hi-C data by DamID for analysis at the local level we anticipate adding another dimension to our understanding of genomic organization at the nuclear periphery. We will be adjusting the global-centric standard methods for Hi-C analysis to accommodate our local-centric approach. This additional dimension allows us to refactor our primary research question, which becomes:

How likely is it these regions will touch... here?

The inclusion of NPC associated DamID data adds depth by establishing physical reference points (the NPCs) within the cell.

4.1 Background

Before addressing these questions, we set the stage by looking at the context within which the data was collected. After establishing the “how” of data acquisition, we will look at the way that data is structured as well as the existing means of visualizing Hi-C data.

4.1.1 Cell lines

If genomic organization influences cellular identity, then the results of any study thereof will yield cell type specific results, making the origin of a cell line key to interpreting the data collected. The origin of the U2OS cell line was [discussed in Part I](#) as it pertained to use as a model for

^{xlvii} This definition of LADs is not nearly as nuanced as the one presented in the cited paper but it does reflect prevailing thoughts of non-specialists with respect to transcriptional activity at the nuclear periphery.

Rheumatoid Arthritis (RA). The data analyzed in this study was collected from U2OS cells, which is not a coincidence. After initial discussions regarding the first study, we sought an orthogonal dataset that could be used for follow-up. To this end, we found a Hi-C dataset from the Hetzer lab, the same lab that originated the siRNA/DamID dataset, also collected from U2OS cells. Though the laboratory originating the data was not initially a consideration, it is fortuitous that these data were collected in the same lab, which means it is likely they originate from the same U2OS culture. As an immortalized cell line derived from tumorigenic cells, the potential for genetic mutations should not be overlooked^{xlviii}.

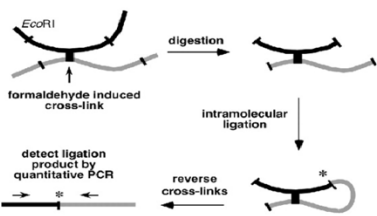
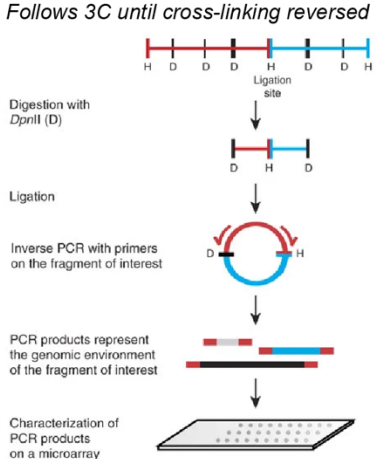
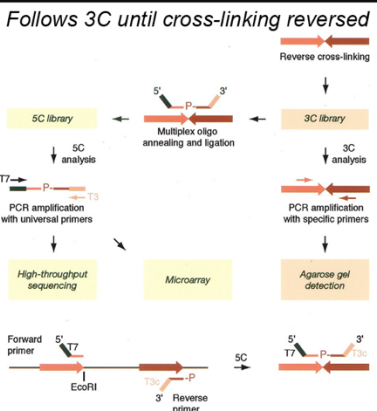
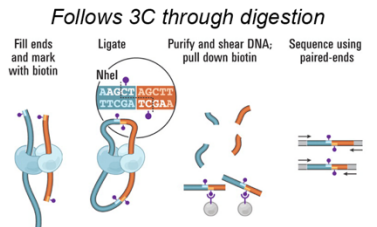
In our RA study, we observed that the genes that were associated with osteosarcoma had the most statistical significance when it came to modeling overall association scores as a function of siRNA and DamID assay results. Given the origin of the U2OS dataset, an osteosarcoma these results are not surprising. However, such results cannot be taken for granted; not all cell lines derived from (what appears to be) the same tumor-type behave in the same way.¹⁶⁹ Another study compared the expression of osteoblastic markers between osteosarcoma cell lines and “normal” human osteoblasts, finding that U2OS cells do not present the standard osteoblastic cell markers such as osteocalcin (OC) and decorin.⁹⁷ For example, some – not U2OS – have lost the ability to induce tumors in *in vivo* murine models.⁹⁶ The microenvironment for cells in culture may select for more stable cells within the laboratory setting, leading to a loss in tumorigenicity. Additionally, there is likely an impact of tissue architecture on the “behavior” of cells in culture. Osteoblasts demonstrate density-dependent alkaline phosphatase (ALP) activity and change in labelling profile; U2OS cells show no ALP activity regardless of cell density nor was their labeling profile density dependent.⁹⁷ Cells were also characterized by the cell types into which they could be enticed into differentiating. U2OS cells were induced to differentiate into adipocytes while not differentiating into osteoblasts and chondrocytes.⁹⁶ Cell line characterization experiments are valuable, and their results must be considered as part of the “baggage” a cell line brings to a study.

4.1.2 Gene Targets

For our analysis, we aggregated the information for several sets of genes that fit different criteria by downloading their information from Ensembl via biomaRt for hg38 AKA GRCh38p14. These sets were not designed to be mutually exclusive. Based on our broader goal of establishing the NPC as an organizational reference point in the cell, we will be looking at those genes that are DamID positive for both Nup93 and Nup153 based on data from a prior study.⁴² In addition, we will be looking at data posted on dbSUPER¹⁷⁰ of super enhancers (SEs) and their associated genes for three cell types: osteoblasts, adipocytes, and MCF-7 cells – an estrogen receptor (ER)-positive breast cancer cell line.¹⁷¹ The osteoblasts are included as it was the sole representative of bone tissue in the database and U2OS likely has a profile similar to osteoblasts based on their origin. Adipocytes were included because they are the cell type U2OS cells were differentiated into in a study characterizing osteosarcoma derived cell lines.⁹⁷ MCF-7 cells were included as a comparison with respect to cancer related genomic organization. The *HOXA* genes were included based on known association with Nup93.^{45,46} *MYC* and its associated enhancer/lncRNA, *CCAT1* as well as *KITLG* and its associated enhancer were also included based on evidence of their association with Nup93.¹⁷²

^{xlviii} In practice, it is unclear whether or not this is considered; if it is, it's one [of many] steps taken to be a given but not discussed, so debatable whether its assumption is merited – much like model assumptions in statistics.

Table 4.1 Overview of Chromosome Conformation Capture methods through the development of Hi-C

2002	"one-vs-one"	PCR amplification Primers for specific targets	
2006	"one-vs-all"	Circularization of fragments before PCR amplification Primer for a specific target used to amplify it and any ligated fragments Data collected using microarrays	<p><i>Follows 3C until cross-linking reversed</i></p> 
2006	"many-vs-many"	Ligation-mediated amplification (LMA): PCR amplification of strands with T7 and T3 primers Captures Interactions between regions of interest	<p><i>Follows 3C until cross-linking reversed</i></p> 
2009	"all-vs-all"	Biotinylated nucleotides mark ligated fragments for pull down Cross-links retained through ligation Next Generation Sequencing used for high-throughput	<p><i>Follows 3C through digestion</i></p> 

Figures within the table adapted from the papers commonly associated with the introduction of each method. (3C) Dekker et al. Figure 1A, *Science*. 2002; 295(5558). doi: 10.1126/science.1067799 (4C) Simonis, et al. Figure 1A, *Nature Genetics*. 2006; 38(11). doi: 10.1038/ng1896 (5C) Dostie, et al. Figure 1, *Genome Research*. 2006; 16(10). doi: 10.1101/gr.5571506 (Hi-C) Lieberman-Aiden et al. Figure 1A, *Science*. 2009; 326(289). doi: 10.1126/science.1181369

4.1.3 Chromatin Capture (Generally Speaking)

Chromatin conformation capture methods are sequencing protocols used to extricate clues regarding the 3D and/or 4D organization of chromatin within the nucleus. The fourth dimension, time, is not always measured; when it is, it is often relative to the cell cycle, which varies between cell types and conditions in standard units of time. Like evolution, while some outcomes (methods)

may be considered “better”, in reality some improvements are “better” for responding (answering) to a different situation (question). A bird’s eye view of the four major methods, up to and including Hi-C, are outlined in **Table 4.1**. Although the concept behind 5C made genome-wide studies seem possible, it was not until Hi-C drew in NGS that it became reality.

While there have been subsequent improvements in Hi-C technique as well as single-celled adaptations of Hi-C, these are outside the scope of what is necessary to provide context for this study. The data used in this study was collected using a Hi-C approach that does not stray far from what was established in 2009, which will be described in the next section.

4.1.4 All-against-all Chromatin Capture: Hi-C

Hi-C is distinguished from other methods by its application of next generation sequencing (NGS) to an “all-against-all” approach targeting the whole genome.

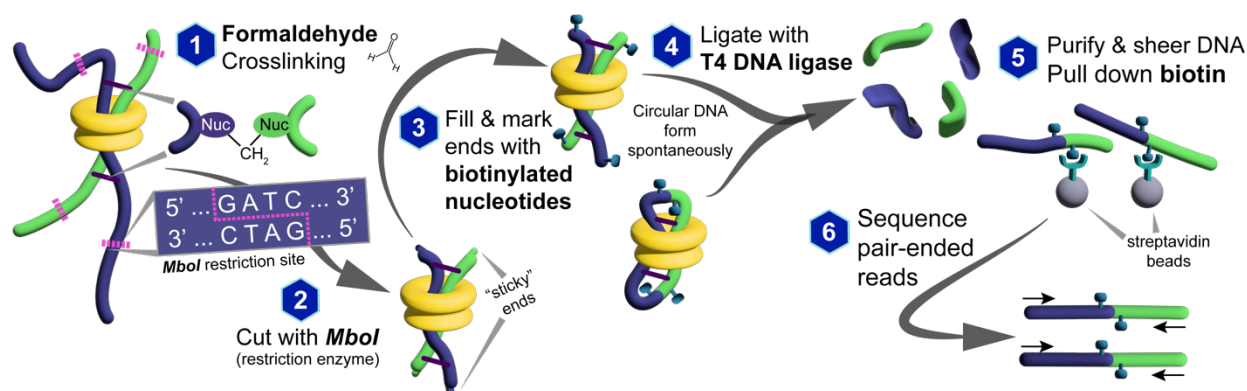


Figure 4.1 A general overview of Hi-C data collection, including details specific to the Hi-C dataset utilized in this analysis.

To capture chromatin-chromatin interactions, formaldehyde is used to crosslink chromatin. First, formaldehyde reacts with a nucleophile (labeled Nuc) at one chromatin locus (purple) to form a Schiff base.¹⁷³ Then, the base reacts with another nucleophile (green) on the interacting chromatin, crosslinking the two chromatin loci. This process is reversible. Proteins such as cohesins (yellow), which are involved in loop formation, may be present and mediate the reaction by stabilizing the chromatin position. Chromatin is cut into fragments by a restriction enzyme – *MboI* in this experiment¹⁷⁴ – which leaves “sticky” ends. These sticky ends are then filled in with biotinylated bases. Spontaneous formation of circular DNA occurs at this step, as may spurious chimeric and self-ligated DNA fragments. The process often occurs at a high dilution to avoid such unwanted fragments (they produce noise that hides the signal, i.e. the strands that ultimately become pair-ended reads). The DNA is then ligated, in this case by *T4 DNA ligase*. After the DNA sample is cleared of other matter, the crosslinkers are reversed and the pair-ended reads are pulled down by capturing the biotin markers on streptavidin beads. Loose DNA is removed, and the trapped reads are sequenced, resulting in the pair-ended data found in a raw dataset.

Adapted from Lieberman-Aiden et al. Figure 1A, *Science*. 2009; 326(289). doi: 10.1126/science.1181369

4.1.4.1 Overview of Experimental Design & Data Considerations

The choices made when designing a Hi-C experiment impact the data in non-negligible ways. Here, we introduce a few such factors and how they may impact the data.

Restriction enzymes are exogenous enzymes derived from bacteria that recognize a given sequence and cleave at that site. These enzymes recognize and cleave at a restriction site, which consists of a short sequence of base pairs. The restriction enzyme *HindIII* recognizes the sequence 5'-A A G C T T-3' and *Mbol* recognizes the sequence 5'-G A T C-3'. For a k -mer – a DNA sequence of length k – there exist 4^k possible combinations of base pairs that could make up that sequence. For example, a 4-mer can have any one of 4^4 (256) possible base pair combinations and there 4^6 (4096) possible 6-mers. In other words, a specific k -mer may be found roughly once in every of 4^k stretch of random base pairs, assuming no additional factors are at

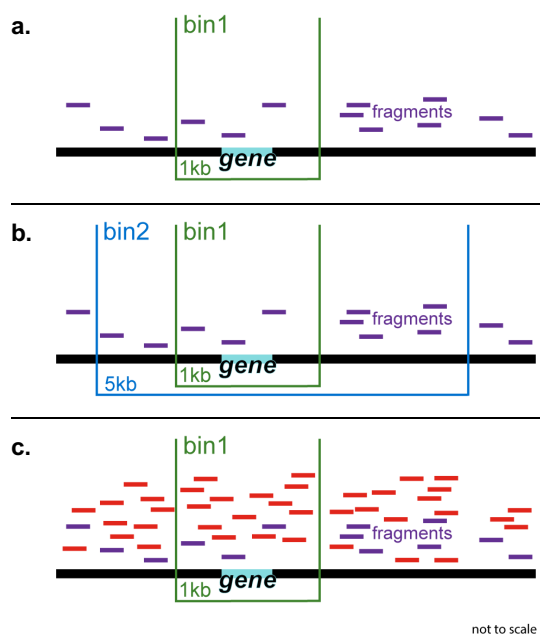


Figure 4.2 The relationship between bin size, coverage, and sequencing depth (a) Fragments (purple) have been aligned to a stretch of chromatin (black) that contains a gene of interest (teal). With a 1kb bin, only three counts are “assigned” to the gene. At this sequencing depth, the coverage is fairly low, but the sparsity of data can be improved by (b) increasing the bin size to 5kb, which increases the number of reads “assigned” to this bin. (c) To increase the resolution to 1kb, a linear improvement, the read depth needs to be increased quadratically (red).

play. Thus, the length of the restriction site an enzyme recognizes is a major determinant of how much it can “chop up” chromatin. Of course, considerations such as G/C content and chromatin accessibility also come into play.¹⁷⁵ “Small” modifications of a sequence such as methylation render some restriction enzymes less likely or unable to fragment DNA at that location.¹⁷⁶

Resolution is a broadly used but vaguely-defined term that holds (somewhat) distinctive meanings in different spaces, not so much in the overall concept but in the underlying parameters. In microscopy, resolution is defined loosely as “what level of detail can be reliably discerned^{xlix} in a given image” as characterized by labeling density and localization uncertainty.¹⁷⁷ With respect to Hi-C data, the resolution is the size of the bin^l used. The size of the bin is used to increase coverage by spreading out the value of the fragments.¹⁷⁸ (Fig. 4.2a–b). To increase the resolution while maintaining coverage, the read depth must be increased quadratically while the resolution increases linearly (Fig. 4.2c). What determines the bin size? Great question. We were unable to find a tried-and-true heuristic, though some were suggested within methods sections, such as constructing the bins with a size that results in 80% of loci have at least 1000 contacts¹⁷⁹; “The map resolution is meant to reflect the finest scale at which one can reliably discern local features when visually examining the data.” On the whole, however, it seems as though this is – as so many things are – context dependent.¹⁸⁰

Although binning itself seems to be a statistical exercise, the contacts that result in the binned fragments have spatial limitations. There are many physical constraints to chromatin interacting particularly for any duration of time. Two such factors are: 1. the nature of chromatin as a physical

^{xlix} If possible, this author would dive headfirst in the rabbit hole of what qualifies as “reliable” in terms of discerning and who is the observer doing the discerning, what are their technical limitations, and what improvements in computing may impact this. But, as it stands, to many rabbits have already been chase in the writing of this paper.

^l We will later observe that the interaction between two bins is visually represented by a square in a heatmap, which appropriately makes a heatmap look like a pixelated image.

entity, which is often modeled as a polymer and included in normalization calculations as part of the expected interaction frequency count, and 2. the limit imposed by the size of the crosslinking moleculeⁱⁱ, which is about 2 Å (0.2 nm) in the case of formaldehyde.¹⁸¹ The physical distance required for cross-linking is likely more of a limitation than the ability of DamID-fused proteins to interact with chromatin, which is something to keep in mind. However, further investigation is warranted before asserting this with any certainty.

4.1.4.2 Data structure

The raw sequencing data produced comes in strings of sequenced pair-ended reads, which require quality assessment, alignment, (re)pairing, and sorting before questions of counting, binning and normalization can be addressed. Between sorting and what is thought of as “processed” exists a sort of limbo: the point at which all fragments are arranged in an orderly fashion but have yet to be counted. This is akin to students before being assigned groups for a project. They are all there, (hopefully) orderly, and waiting for assignment. It is only after assigning fragments to bins that we arrive processed data. At this point, the data could be left as raw counts, though it is generally normalized.

Processed Hi-C data comes in the form of a *contact matrix* containing a pair of interaction regions – *bins* – of the same base pair (bp) length – *bin size* – and the number of interactions counted for that combination, arrange in a sparse matrix.

This dataset can also be structured into a dense matrix, which a row represents the i^{th} bin, a column represents the j^{th} bin, and the counts, c_{ij} , represent the number of interactions between the two genomic locations (**Fig. 4.3**). If the fourth dimension, time, is introduced, such data is collected by taking another sample of the same cell population and another dataset results. In the case of the dataset used in the work presented here, time was measured in minutes based on the expected trajectory of the U2OS cell population through the cell cycle, producing eight datasets per biological replicate (**Fig. 4.7**).

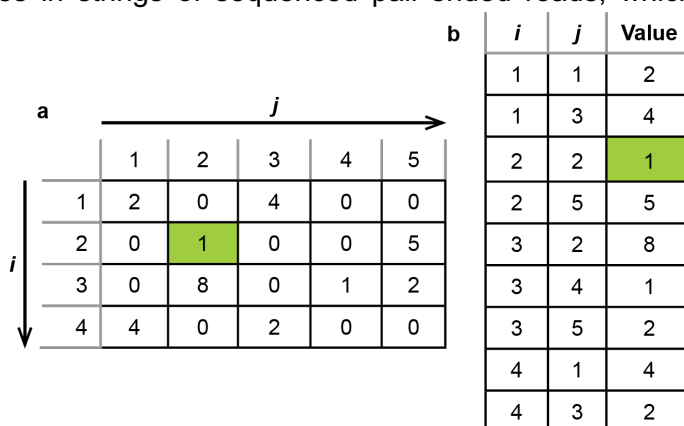


Figure 4.3 Hi-C data is often organized into contact matrix with genomic coordinates represented by axis.

(a) A dense matrix includes all possible combinations of loci/bins between two lengths of chromatin. The diagonal of the matrix indicates interactions between a location and itself. In each cell, denoted C_{ij} , the number of contact points between the two loci represented by the row (i) and column (j) is stored. (b) Interactions can also be stored in a sparse matrix, which retains only information about those pairs of loci with non-zero contactⁱⁱⁱ.¹⁸⁵

ⁱⁱ I did not find evidence of studies on any limitations imposed by cross-linking agents on establishing contacts between chromatin strands, nor was a researcher with the Dekker lab aware of any such research.

ⁱⁱⁱ Here is an example that epitomizes a jargon problem: opposing definitions for the same terms. The nomenclature used in this caption originates from *Hi-C Data Formats*, Chapter 6 of *Hi-C Data Analysis* (2022, Springer Nature), the dense matrix is an $n \times m$ matrix with $n \times m$ entries, a significant number of which are 0, and the sparse matrix only consists of non-zero values. At first glance, it seemed like a typo, until it was discovered others define these two terms this way.. as well as in the exact opposite manner. This conflict does not seem to have discipline specific borders.^{182–184}

Binning is all well and good, but it cannot be done before defining an anchor point. Setting the anchor point comes down to asking: where to start counting?

Counting, something we learn to do as children, is a deceptively difficult task.^{liii} The “number” zero is not even considered natural. It should come as no surprise then that defining a point at which to start counting is non-trivial.^{liv} Previously, the starting point was an emergent property of the technology used. The “one-versus-one”, “one-versus-many”, and so on that encapsulate the nature of each chromatin capture iteration, including the *anchor point* for counting interactions between genomic loci. Leaving no ambiguity, the “one-vs-one” loci comparison of 3C places the anchor point is the start of the gene. However, this starting point became more ambiguous as the scope of the technology broadened, leading to our current question: where do we start counting in the “all-vs-all” information provided by a Hi-C dataset?

The answer to this question depends on what research question you are addressing.

Currently, Hi-C data are processed with an anchor point set at the first base pair of each chromosome. From this point, bins of equal width are enumerated to which chromatin fragments will be assigned upon alignment. For looking at global organizational patterns such as A/B compartments, topological domains (TADs) and loops, this choice keeps the information scaled to a level appropriate to the inquiry. It also increases the effective coverage by decreasing the search space.¹⁷⁸ However, this approach may lose information when applied to local features of the genome such as genes (**Fig. 1.2**). This is one of the issues that we address in the [Methods](#) section.

Once aligned to a reference genome^{lv}, fragments are assigned to these bins and a contact matrix is constructed. The signal, the “true” interaction counts, is contained in the contact matrix. The construct often undergoes adjustments to compensate for noise created by bias¹⁸⁷ and is generally normalized with respect to the whole genome.¹⁸⁸ Processed data is often log transformation prior to visualization. This transformation amplifies strong signals, such as intra-chromosomal regions proximal to each other, while dampening weaker signals, such as those between inter-chromosomal regions, making patterns visually easier to distinguish.

4.1.4.3 Bias & Normalization Methods

Hi-C experiments come with many sources of potential bias, some which are known and accounted for during data processing. Often, methods for bias correction are baked into normalization methods.¹⁸⁸ Such methods can be explicit, applying *a priori* information such as fragment length, GC content and mappability.¹⁸⁷ They assume that biases can be known and taken into account. Others are implicit, using only the data collected during the experiment. Behind this approach is the notion that biases are equally applicable to all data collected in a

^{liii} “Counting is hard” was first read by this author in her sophomore year of college, haunting her ever since.¹⁸⁶ Seven out of the thirteen students withdrew from the combinatorics course that semester. She was one of them. A majority of them had already defected from the physics department, seeking shelter in mathematics from the handwaving, making this but another rockface in their uphill climb.

^{liv} Programming languages don’t agree on this. For example, Python starts indexing at 0 while R begins at 1. This makes sense when you consider R was designed for use in statistics, which is based in observations, of which having a zeroth observation makes very little sense outside of a patient-zero scenario. However, it remains a tripping hazard when jumping between both, especially when you realize not all bioinformatics software starts counting in the same place either.

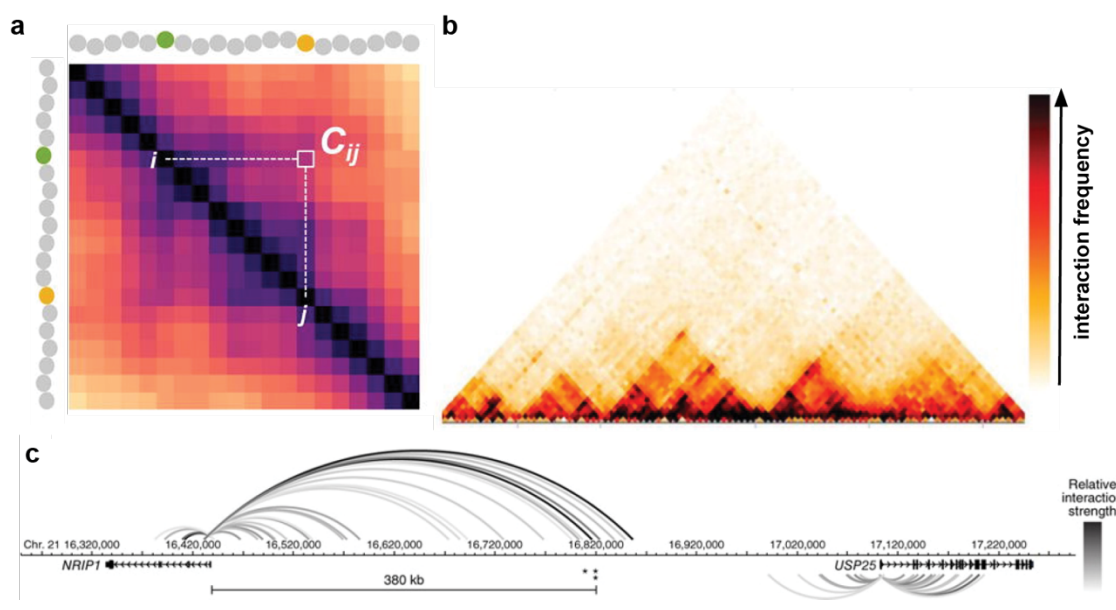
^{lv} This author has not encountered the use of *de novo* alignment in the Hi-C space, the sparsity of the data does not inspire its use for such a task.

given experiment.^{lvi} Explicit methods include HiCNorm and *hicpipe*. HiCNorm uses a Poisson regression model to account for things such as fragment length.¹⁸⁹ *hicpipe* establishes that sources of bias for *cis*- and *trans*-interactions may differ and also accounts for bias introduced by random cleavage events (versus those induced by restriction enzymes).¹⁸⁷ Sequential component normalization (SCN)¹⁹⁰ and Iterative correction and eigenvector decomposition (ICE)¹⁹¹ are implicit methods. They draw upon scaling and iterative correction methods, respectively. It is worth noting that ICE does use external information for validation, though it *a priori* information is not included in calculations when applying the method.

At this point, it is important to note that most, if not all, methods are focused on normalization on the global scale. Given the available resolution and the questions Hi-C datasets are generally used to address, this global scale normalization is appropriate. However, scaling in this way is not necessarily appropriate for looking at interactions at the local level. What is meant by local is not so much a gene and its nearest neighboring gene, although it could, but gene/gene interactions on the whole, a shift back toward the “one-versus-one” or “one-versus-many” perspectives gained by earlier chromatin capture technologies while using existing Hi-C datasets.

4.1.5 The Exploration of Hi-C Data

The purpose of modeling Hi-C data is to characterize features of genome organization. Typical Hi-C data analysis is not inclusive of all features of genomic organization; the focus tends to be on A/B compartments, topologically associated domains (TADs), lamina-associated domains (LADs), and chromatin loops. The conceptual model that is tethered to chromatin-chromatin interactions is Activity-by-Contact, which posits that physically associated chromatin regions may influence the expression of their respective genes.¹⁹² This model does not limit or require genes to touch directly, influence may also be exerted through shared enhancers, enhancer-promoters contacts and so on. The major mode of gaining insights as to the nature of such contacts and the structures they inhabit is through visualization, which is both the inspiration for and the standard against which analyses of Hi-C data is compared.^{193,194}



^{lvi} While this broke stroke assumption might be reasonable to make regarding data from a single run, this author is dubious that it is reasonable to assume such a thing between runs.

Figure 4.4 Three standard representations of Hi-C contact matrices.

(a) Square matrices represent two sets of genomic coordinates by bin, one on each axis, and the interaction counts or relative strength between the binned locations is using the color channel. (b) Triangular matrices are used to highlight patterns self-interaction between stretches of chromatin. Such a heatmap is derived from the square representation by cutting the plot along the diagonal, upper left to lower right, then retaining and rotating only the lower portion of the map such that the diagonal becomes the horizontal axis. (c) Arc diagrams visualize interactions between genomic locations by connecting them with an arc. The arc may be used to indicate the “strength” of interactions, based on the number of counts, through color or linewidth channels. An alternative form of this includes multiple chromosomes in a circular representation (not shown).

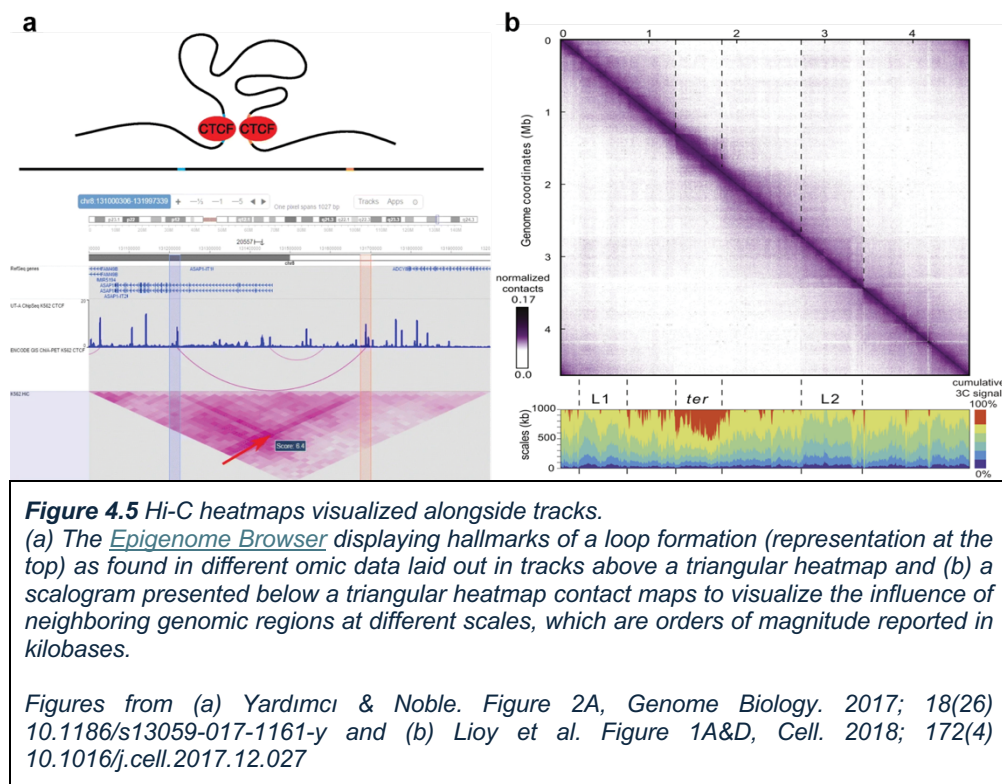
Figures from (a) Shinkai, Onami, and Nakato. Figure 1A, *Computational and Structural Biotechnology Journal*. 2020; 18, doi: 10.1016/j.csbj.2020.08.014 (b) Lajoie, Dekker, and Kaplan. Figure 9, *Methods*. 2014; doi: 10.1016/j.ymeth.2014.10.031 (c) Mifsud et al. Supplementary Figure 5, *Nature Genetics*. 2015; 47(6), doi: 10.1038/ng.3286

Hi-C visualizations often depend on transformation of sequencing pipeline output into a format that can be fed into and visualized using available tools (i.e. a contact matrix). The R/Bioconductor¹ package *GenomicInteractions*¹⁹⁵ crosses the bridge between standalone software (e.g. HOMER¹⁹⁶) as well as packages developed for singular chromatin capture methods (e.g. *diffHic*¹⁹⁷) to provide a standard data structure that can be used for visualization and further analysis Hi-C data. Further improvements to processing, manipulation, and visualization were necessitated by the increase in computation demands brought about by increasing Hi-C resolution and are addressed by *HiCBricks*, an R/Bioconductor framework¹⁹⁸.

4.1.5.1 2-Dimensional Visualization

The standard for visualizing the contact matrix is the heatmap, in either a square or triangular form, and the arc diagram (**Fig. 4.4a–c**). Other representations include those built upon the initial plot with “tracks” – aligned by location relative to genomic sequence – informed by other omics data, and interpolated 3D models of chromatin structure. The latter is the least common, likely due to the computational expense of creating such models.

As a result of how contact matrices are transformed during normalization – often preferentially amplifying intra-chromosomal interactions – most standard Hi-C visualization methods are better suited for exploring *cis*-interaction patterns. For example, TADs are easily identified through visual inspection of heatmaps as dark regions at the peaks of triangles where *cis*-interactions are plotted – along the diagonal of a square heatmap or along the base of a triangular heatmap (**Fig. 4.4c**).¹⁹⁹ *Trans*-interactions are not easily distinguished in traditional heatmaps, where they are washed out by the prevalence of intrachromosomal contacts.



Visualizations that build on linear genomic coordinates can be annotated with “tracks”, with additional information displayed on a parallel axis using the same genomic coordinates and scale. JuiceBox^{200,201} is an Hi-C data visualizer that incorporates the track-based Integrative Genome Viewer (IGV)^{202,203}, which allows the aggregation of multiple genomic data-types for display with the corresponding Hi-C heatmap. Similarly, 2D data can be overlaid on the heatmap, allowing the identification of chromatin structures. For example, the directionality of a binding protein with respect to chromatin loops was determined by juxtaposing CTCF-binding motifs on Hi-C data (**Fig. 4.5a**).¹⁷⁹ The scalogram can be used as a track to close the gap between scales by linking resolutions (**Fig. 4.5b**). Scalograms illustrate the “tightness” of a region by considering what percent of the signal in a bin comes from interacting with flanking region. As for the availability of multi-scale encoding, HiGlass and JuiceBox allow zooming and include gene symbol searches. Although visual encoding does not change based on the scale, the refinement level of the data shown increases as resolution permits.

The 2D visualizations reviewed above, even those that are available to scale, largely focus on global rather than local effects, even when used to visualize features that trend toward the small scale such as loops. One potential exception to this is the arc diagram. However, the data itself is scaled globally, which has the potential to alter the perspective. Before further consideration of issues of scale, we consider an additional dimension as the genomic organization does not occur within the confines of 2D.

4.1.5.2 3-Dimensional Visualization

Visualization in three dimensions is inherently messy because of limitations in how we perceive information as three-dimensional. The information available along the z-axis diminishes as a result, making encoding in three-dimensions render more like two-and-a-half dimensions of information²⁰⁴, even when perception (linear as well as texture and size gradients), occlusion and

the kinetic depth effect are leveraged. As shapes become less familiar, the benefit of using these indicators decreases, making them less useful for the unfamiliar shapes organized chromatin take when visualized. These complications are compounded by the lack of a common coordinate system for the nucleus, which would otherwise allow integration of imaging and sequencing data. Such a common coordinate system is under development as one of the stated objectives of the 4D Nucleome Data Coordination and Integration Center.²⁰⁵ As a result, visualizing Hi-C data layers its own set of challenges on top of those associated with 3D visualization.

Although GrapHi-C²⁰⁶ workflow does not display Hi-C data in 3D, the representation of multi-way chromosomal interactions implies dimensionality beyond the 2D matrix. Representing a contact matrix with nodes and edges emerges naturally as a form of visualizing dense contact matrices (**Fig. 4.3b**). In its simplest form, coordinates may be used as nodes with each row indicating a connection between two nodes with the weight the connecting edge encoded by the corresponding value. The prediction of chromatin organization with multi-way contacts within the 4D context of the nucleus requires a statistical means of extrapolating such interactions. This criterion contributes to the computational complexity of modeling genomic structure. Though statistical methods for predicting such interactions exist, they, along with other methods, are more applicable to *intra*-chromosomal interactions vis-à-vis the use of polymer models.²⁰⁷

Where prior tools either visualized curated data only or were developed primarily for modeling protein structures (e.g. Chimera²⁰⁸), HiC-3DViewer²⁰⁹ allows researchers to import their own data into a graphical user interface (GUI) developed specifically for chromatin-centric data. The ability to overlay additional “omic” tracks such as ChIP-seq data is also included, strengthening HiC-3DViewer’s efficacy for viewing *trans*-interactions.

4.1.6 Implications of Current Methods

More often than not, interaction counts between genome loci are normalized before any exploratory analysis or visualization take place. Generally, the normalization methods selected for Hi-C data are rooted in global, rather than local, characteristics of the dataset. Additionally, a polymer model is the foundation of calculating expected interaction counts, against which the observed are compared to establish significance. However, the polymer model isn’t applicable when considering *trans*-chromosomal interactions or *cis*-interactions over a certain distance.

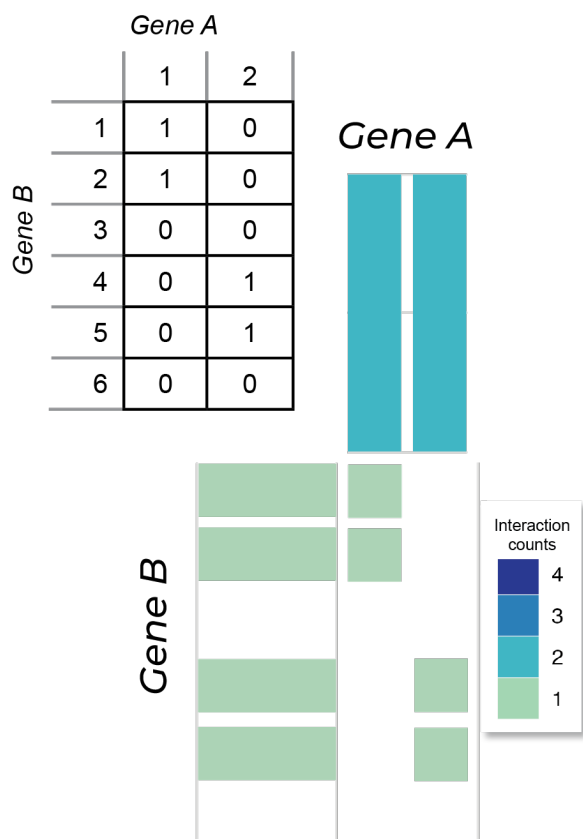


Figure 4.6 Visualizing Heatmaps \equiv 2-D Histograms Comparing a contact matrix for Genes A and B with 1D histograms of counts for Gene A (x-axis) and Gene B (y-axis) flanking a heatmap of the same count data demonstrates the congruence of heatmaps and histograms.

As we've seen, normalized interaction data are most often visualized as a heatmap – essentially a two-dimensional histogram (**Fig. 4.6**) – which is not well suited to exploratory analysis at the gene-scale. While more target-specific data collection methods such as 3C may be employed, those are also not appropriate^{lvii} for the sort of exploratory analysis one may endeavor with a list of a few hundred potential target gene pairs. So, what is it about heatmaps for Hi-C that isn't suitable for looking at interactions at a local scale? The nature of the underlying data and the normalization methods generally applied during processing decrease the usefulness of standard heatmaps for gene level data. This depreciation is in no small part due to the resolution available versus the length of a gene. To address this problem, we looked to other disciplines that ask questions of data structured around position and time.

Heatmaps, probabilities, and tracking dynamic systems overtime are fundamental to geospatial data analysis. Beyond using heatmaps to explore raw data, geospatial data are often transformed and displayed as probability densities. These are frequently visualized using contour plots. Much like the epidemiologists who model areas of high-to-low risk, computational biologists model chromatin contacts. It follows that methods drawn from geospatial analysis may be applicable to visually represent gene-to-gene interaction likelihood. As the protocols for Hi-C improve²¹⁰,

improving the overall resolution, and other chromatin capture techniques are developed the utility of exploratory analysis at the local level will become ever more necessary.

We propose using kernel density estimation (KDE) as a means of visual exploratory data analysis. As a smoothing method, KDE is inherently a normalization method. Unlike the “usual” Hi-C normalization methods, which hinge on the summary statistics of the whole genome, KDE uses information regarding neighboring data to smooth – normalize – the data in a local-centric manner. Additionally, KDE circumvents the anchor point problem, which refers to the influence the starting point of binning has on the information conveyed by a histogram and heatmap.²¹¹ Furthermore, the dynamic nature of chromatin warrants a probabilistic approach to account for variation between individual cells; this approach is of particular importance for bulk-sequencing data such as that used in this research. This approach can be adapted to visualizing the persistence of signal over time, as chromosome organization is dynamic²¹², changing over the course of the cell cycle.

^{lvii} Furthermore, such methods assume the existence of specific data. In the case of a preliminary data analysis, it may be more pragmatic to use existing datasets, particularly given the high cost of such experiments.

4.2 Methods

Our methods were divided into two parts: 1) the [initial case study](#), which focused on a specific gene pair to devise a method for applying locally appropriate statistical methods and visualization for exploratory data analysis using pre-processed Hi-C data; and 2) a pipeline for applying the developed methods to a set of gene pairs utilizing raw Hi-C data processed using an adaptation of the 4DN pipeline.

4.2.1 Hi-C Dataset

The Hi-C dataset employed in this endeavor is a SubSeries (GSE141067) that exists as part of SuperSeries (GSE141139) hosted by the NCBI in the GEO database, where it is publicly available.²¹³ The Hi-C experiments as well as

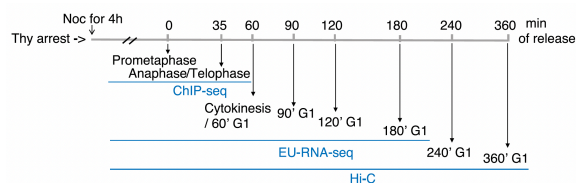


Figure 4.7 Sampling Schema for Hi-C data acquisition

Points during the cell cycle at which samples were collected from U2OS cell culture as indicated in minutes as well as relative to the state of the cells.¹⁷⁴

From Kang et al. Figure 1A, *Genes & Development*. 2020;34(13-14). doi:10.1101/gad.335794.119

other omic experiments used the U2OS cell line²¹⁴, the same cell line underlying the data in [Part I](#). The Hi-C dataset consists of 16 “runs”: time was measured in minutes based on the expected trajectory of the U2OS cell population through the cell cycle, producing 8 sets of pair-ended reads for each of 2 *biological replicates*^{lviii} (**Fig. 4.7**). The restriction enzyme *Mbol* was used to cut the chromatin into fragments after cross-linking with formaldehyde (**Fig. 4.1**).

During the original analysis, reads were aligned to hg38 via bwa-mem and prepared using

Hypergeometric Optimization of Motif EnRichment (HOMER)¹⁹⁶ to establish interaction counts and Juicer²⁰⁰ to create, balance, normalize the contact matrices (posted in *hic* format). In the original study, coordinates were lifted from hg38 to hg19 for juxtaposition with the other datasets in the [SuperSeries](#), as those were aligned to hg19. For this initial case study, these processed data were retrieved from the NCBI GEO²¹³ using the browser-based GUI. For the subsequent portion of this venture, the raw (unprocessed/unaligned) data were downloaded from the NCBI GEO database straight to the SCI in *fastq* format using *curl*²¹⁵ after obtaining the urls from the metadata via *ffq*²¹⁶, both via the command-line interface (CLI).

4.2.2 Preliminary Case Study

A case study was the initial approach to develop the methods presented here. Our original objective was to select probes for the fluorescent tagging of two genes.

The criteria for selection were:

1. DamID for Nup93 and Nup153 suggest both genes interact with the NPC complex at the nuclear periphery, and
2. Hi-C data suggests the two genes interact with each other.

^{lviii} Although from the same cell line, the replicates were two different “populations” that underwent separate synchronizing arrests, releases, and data collection. This differs from *technical replicates*, which are samples from the same population taken to assess variability among measurements. The classification was confirmed by the lab that original performed the experiments as this author found it difficult to determine as originally described.

The *cis*-chromosomal gene pair used as an exemplar, *ZNF438* and *PARD3*, was selected during preliminary data exploration of the pre-processed data from this dataset.^{lix} PaintSHOP²¹⁷ was used to design probes for chromosome walking²¹⁸ along both genes.^{lx} Labeling the entire length of each gene was not practical⁷⁶ – the quoted price was greater than that of a modest home in the area at the time of writing. To reduce the number of oligos necessary, and thus reduce the cost, we sought a method of selecting the region(s) with the highest potential for interaction.

We determined sustained interaction would be an indicator of a high potential for interaction in future experiments as well as a characteristic beneficial to fluorescent microscopy experiments.

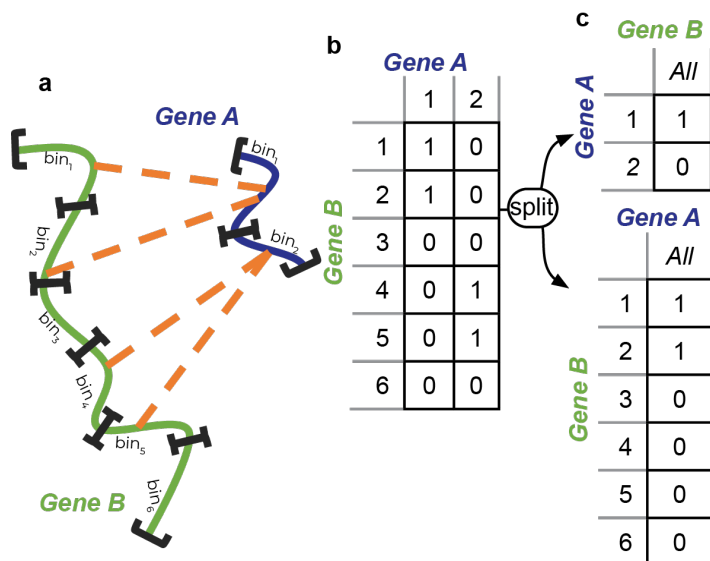


Figure 4.8 A different perspective on contact matrices: “One-versus-Any”

(a) Visualization of interactions between Gene A and B, as seen between a hypothetical gene pair. The resulting gene matrix (an example of a dense matrix) is then split such that: (b) the 1×2 matrix carries information regarding the two Gene A bins interacting with any location on Gene B, and (c) the 6×1 matrix holds information regarding the six Gene B bins interacting with any location on Gene A.

The binning in this figure uses an anchor point situated at the start position of each gene. While not significant to the purpose of this diagram, in this hypothetical, the contact assignment ambiguity is minimized to zero. If the genes did not (conveniently) have a length that is a multiple of the bin length, the only potential ambiguity would exist in the last bin of the gene(s).

We posit that bins with interaction counts that existed at multiple time points and/or between replicates would indicate the region(s) of the genes most likely to interact in the same cell line.^{lix} In other words, the most persistent signal over time and/or between replicates is likely just that, signal, not noise.

4.2.2.1 Heatmaps

Initially, contact matrices for the selected target genes, *ZNF438* and *PARD3*, were visualized as traditional heatmaps with the addition of hierarchical faceting by replicate and time. As far as we know, there is not a standard way of visualizing Hi-C data over time. The counts for each cell (C_{ij}) are encoded using color, depicting all possible i^{th} and j^{th} bin combinations along the length of both genes, an intuitive approach given the “all-against-all” nature of Hi-C data.

To reduce clutter and gain a clearer picture of any underlying patterns of contact such as signal persisting across

time, we considered both genes individually (**Fig. 4.8**). Heatmaps were generated that plotted the sum of interactions of each binned loci of one gene with any part of the opposing gene. Such plots were constructed for both genes and faceted by replicate.

^{lix} As it turns out, they likely aren’t a pair suitable for our long-term objectives, but the methods developed here are and the results with respect to this pair is consistent within its biological context.

^{lx} Chromosome walking is a bit of a misnomer in this situation; the “walk” would have been limited to the length of each gene rather than the whole chromosome.

^{lix} The merit of this prediction has been left for a future paper.

For selecting genomic regions for labeling with probes, as is our goal for imaging experiments, narrowing down the target area for each probe set is beneficial. That target area needs to be the areas that have the greatest probability of coming into contact in the future based on the distribution of signal between the two regions at rest. To do this, we need to add back the previously condensed element of one-to-one specificity (as far as Hi-C resolution allowed). To do so, we look to geospatial analysis for a potentially applicable method.

Application of Geospatial Methods

Two methods often considered for creating surfaces from discrete data points in the geospatial analysis space are inverse distance weighted interpolation (IDW) and kernel density estimation (KDE). IDW creates a surface between sampled points using the distance between the points as weights.²¹⁹ KDE uses a kernel, such as Gaussian distribution, to estimated probability densities such that the total probability around any given point is 1, which results in a probability density surface.²¹⁹ Because Hi-C collects all-against-all data, rather than sampling, KDE was the more appropriate choice.

4.2.2.2 Kernel Density Estimation

Given n is the number of datapoints, K is a kernel function and \mathbf{H} is the bandwidth, a symmetric, positive definite, $d \times d$ matrix, the kernel density estimator is defined as:

$$\hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

The bandwidth \mathbf{H} controls the magnitude and direction of the smoothing applied by the kernel function:

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-\frac{1}{2}} K(\mathbf{H}^{-\frac{1}{2}}\mathbf{x})$$

Two-dimensional KDE was implemented in R using the `ggplot2` library, which imports the `kde2d` function from the `MASS` library²²⁰ to perform the underlying calculations.²²¹ This function uses a bivariate normal kernel, which is aligned with the x- and y-axes, on a grid. The default bandwidth is selected via *normal reference distribution*, the accepted “rule-of-thumb” method for use with Gaussian kernels.²²² The surfaces were plotted, faceted by time and with replicates presented in separate plots.

Evidence supports the (relative) stability of global chromosome architecture during G1, though changes in epigenetic markers as well as movement on the megabase scale.²²³ Therefore, we aggregated at the same data three different ways: all samples aggregated, regardless of timepoint; only samples during taken during G1, excluding cytokinesis; and only samples taken during mitosis, including cytokinesis.

For an alternative look at the change in interaction probability throughout the cell cycle, the `ks` library was used, which has the capacity to perform KDE in up to six dimensions. The result was plotted calling the `p1ot` function, a base R function extended by `ks` using the `p1ot3D` library (**Fig. Error! Reference source not found.**). Interactivity was achieved by opting for `rg1` extension of the base plot function rather than `p1ot3D`, which allows the user to rotate the graph in three dimensions. This makes examining the surfaces easier.

4.2.3 Pipeline Development

4.2.3.1 Data Processing

“... space is generally cheaper than time.”

Whoever wrote this^{lxii} has not worked with Hi-C data.

The processing pipeline used was adapted from the [4DNucleome Hi-C Processing Pipeline](#).²²⁴ The raw Hi-C data was transferred from NCBI GEO to the [Scientific Computing for Innovation Cluster](#) (SCI),¹ a high-performance computing (HPC) cluster, in the *fastq* format using *curl*.^{lxiii} The accession-based URLs for transferring the files, expected file sizes and other metadata were retrieved using the *ffq* python library.²¹⁶

The raw data was aligned to hg38²²⁶ via *bwa mem*.²²⁷ The aligned fragments were parsed and sorted into read-pairs using *pairtools*.²²⁸ It is at this point we began to diverge from the 4DN pipeline. Pairs were not deduplicated using *pairtools* as we plan to look more closely at the characteristics of the duplicates in a future study. The two replicates were not merged, as would this is not appropriate to do with biological replicates. It is at this point we fully deviated from the 4DN-recommended pipeline; we did not bin, matrix balance, or normalization the read-pairs.

When contemplating the allocation of resources, we began by determining the set of genes from which we would draw gene pairs, which hinged in part on how many gene pairs would arise from a given set. *Multi-choose k* – combinations with replacement and without repetition – was used for this calculation:

$$\binom{n}{k} = \frac{(n+k-1)!}{r!(n-1)!} \quad (17)$$

For example, if all possible genes ($n = 26392$) in the siRNA/DamID dataset were paired ($k = 2$), the result would be 691,610,048 combinations of genes. Given that many of these combinations will be “duds” with respect to the phenomenon under investigation – interaction in the immediate vicinity of NPCs^{lxiv} – and that this was a pilot study of applying KDE to minimally processed sequencing data, we opt to narrow this list down to only those genes double positive for interaction with Nup93 and Nup153 via DamID.

At this point, we began by using *pairtools select* to filter out reads by gene. This turned out to be exceedingly inefficient, both in terms of time as well as hard drive space. Thus, we sought a better approach and considered alternative tools for extracting reads from the read files. We began by examining the relationships between the start and end of the genes we to get a better handle on their distribution between chromosomes (**Fig. 4.9**). In this data, we saw clusters of genes, so k-means clustering was applied. This application greatly reduced the number of queries by clustering genes by relative position on each chromosome.

^{lxii} The read [documentation for write_rds\(\)](#) states, “write_rds() does not compress by default as space is generally cheaper than time.”

^{lxiii} The recommended *fastrq* failed to transfer these files, even if proceeded with *pre-fetch*.²²⁵

^{lxiv} It was VERY tempting to calculate the volume of the nucleus and compare it to the shell formed between the envelope and the distance away from NPCs within which a given section of chromatin is likely to interact with an NPC considering density of NPCs within the membrane; but this rabbit hole seemed a bit more of a distraction than a benefit, particularly given the number of cell-type dependent variables.

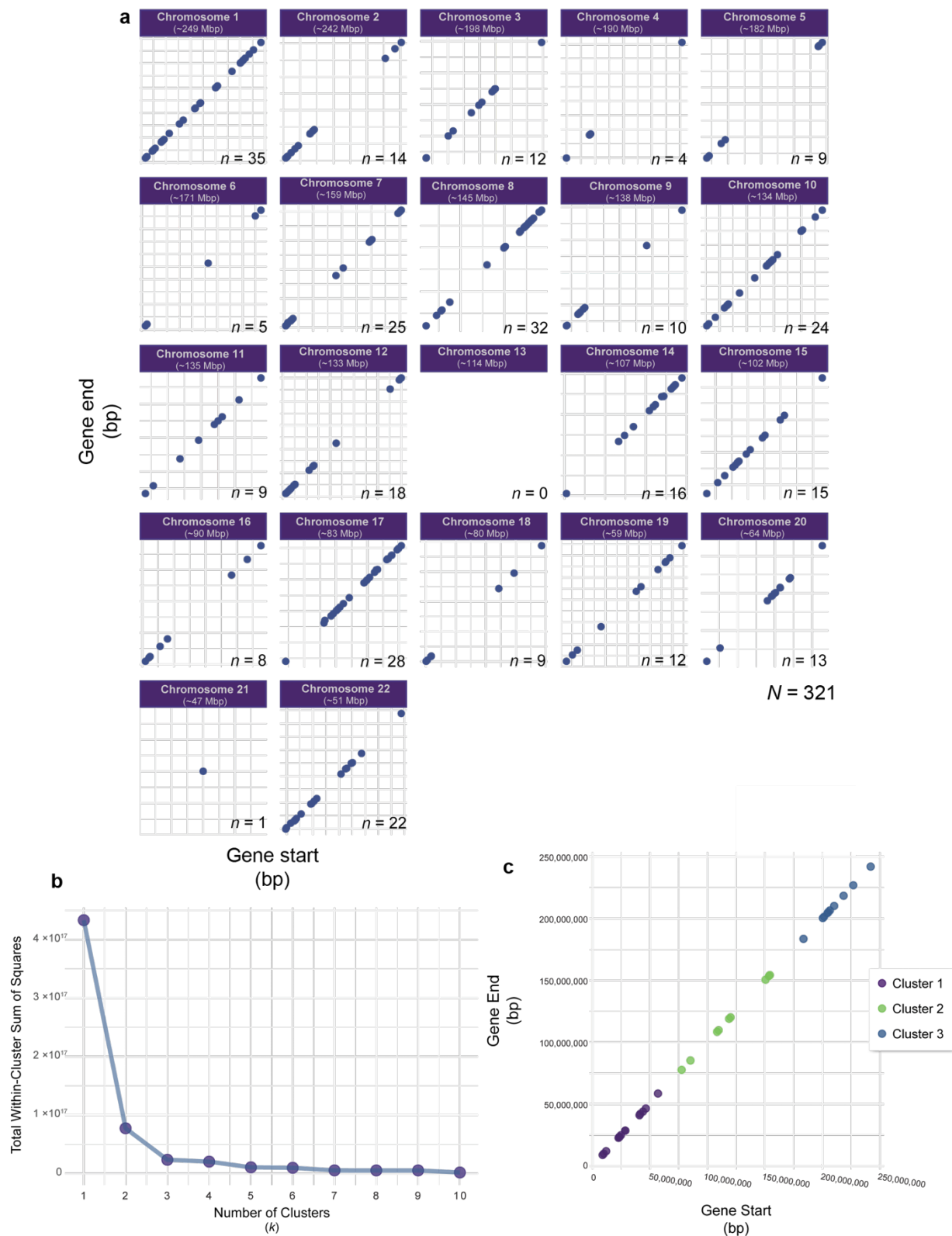


Figure 4.9 Revision of search queries to reduce search space
 (a) The start and end points of the set of DamID-Nup93 DamID-Nup153 double positive genes were plotted after stratification by chromosome and clustering was observed. k -means²²⁹ clustering was used to balance reducing the

number of searches with decreasing the search space when extracting gene-gene interactions from Hi-C data after alignment and pairing. (a) The Elbow Curve for genes on chromosome 1 used to select the number of clusters (k) based on the total within-cluster sum of squares calculated for all values of k , where $\{k \in \mathbb{Z} \mid 1 \leq k \leq 10\}$. The number of clusters is then selected visually by choosing the value at the point where the “elbow” bends, $k = 3$ in the graph shown. (b) The results of k -means clustering encoded with color applied to the scatterplot of the start versus end position of DamID-Nup93 and DamID-Nup153 positive genes on chromosome 1 (from a).

With the number of queries reduced, we sought help from Open2C regarding using `pairtools` for extraction; they suggested using `pairix` instead. This application indexes each pairs file before any queries are made, greatly reducing the amount of time needed to extract matches as well as the resources used to store the results.

These intermediate `.pairs` files were imported into RStudio run in a singularity instance using an interactive job on the SCI. As these files contained information regarding clusters of genes rather than individual genes, they were run through a rudimentary search algorithm that used rounded coordinates as well as a margin on either side of the expected start and end of a gene to attribute reads to genes. This method can be applied to any pair of sequence-based features if provided with the chromosomes and base pair range within which to search for reads.

After creating data frames connecting gene information to read pairs, we first tested the visualization method developed in the case study to the raw reads extracted from replicate 1 for PARD3 and ZNF438 with respect to G1 time points for comparison.

4.3 Results

4.3.1 Preliminary Case Study

4.3.1.1 Heatmaps

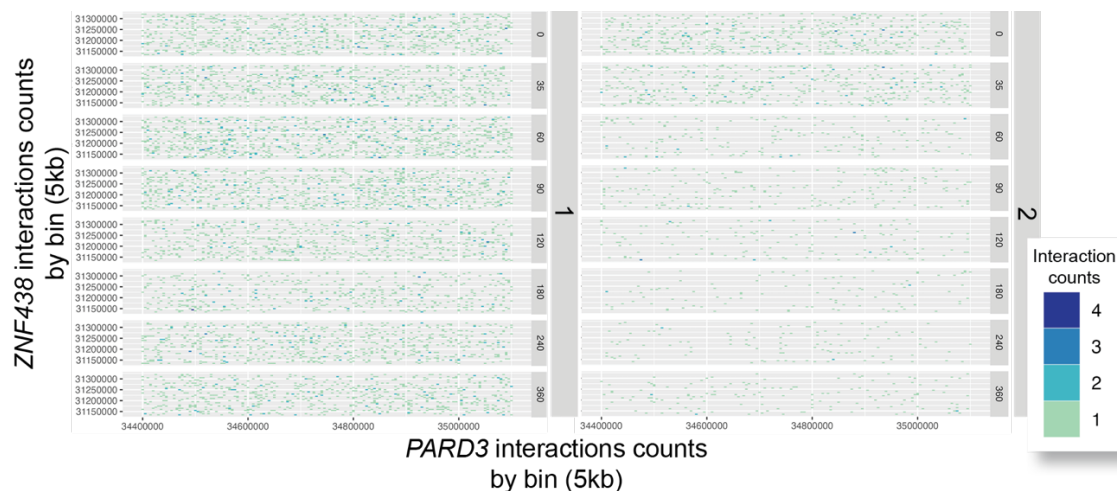


Figure 4.10 Heatmaps of PARD3 interacting with ZNF438 with hierarchical faceting by biological replicate and sampling time. It is difficult to distinguish any patterns at the “local” (gene) level using a traditional heatmap.

Distinguishing interaction “hotspots” from these plots was limited (**Fig. 4.10**); the resultant gene-level heatmaps were not rich with easily distinguishable visual patterns such as the traditionally observed triangles as compared to the global-scoped heatmaps traditionally used to see patterns such as loops and TADs.

Compared to the interactions for replicate 1, the interactions for replicate 2 were sparse. Both replicates had relatively persistent signals over time in regions of each gene (for example, *PARD3* in the region near 34,800kb) (**Fig. 4.11**).

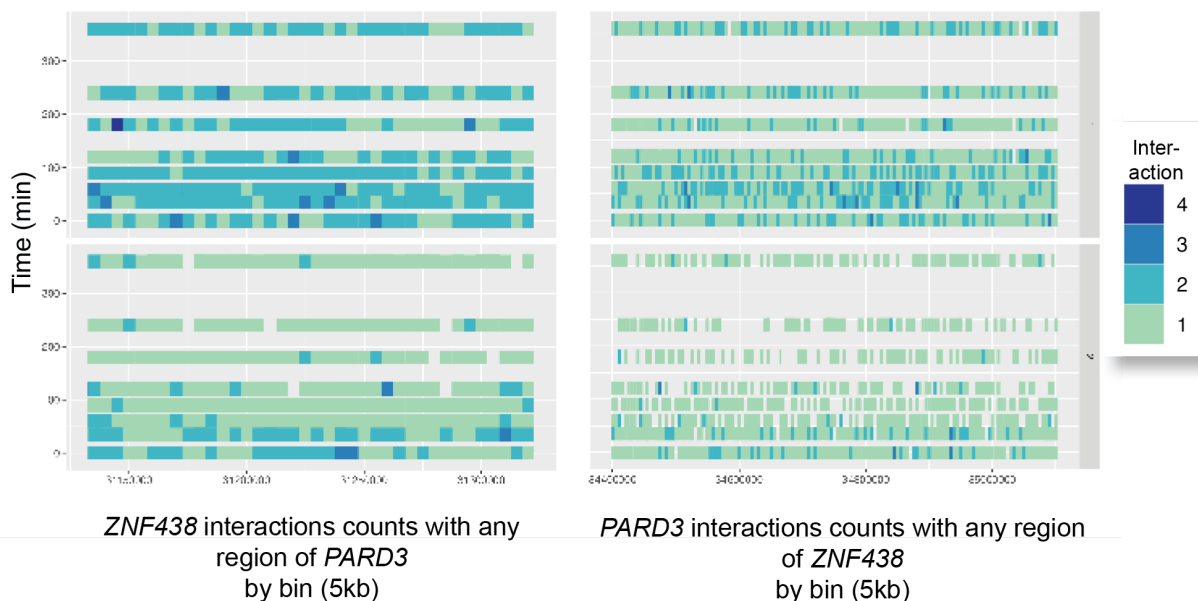


Figure 4.11 The use of traditional heatmaps to depict an alternative perspective of contact matrices: “One-versus-Any”
Heatmap visualization of interactions over time between binned locations on individual genes (*ZNF438* and *PARD3*) with any region of a second gene (*PARD3* and *ZNF438*, respectively), faceted by replicate.

4.3.1.2 Kernel Density Estimation

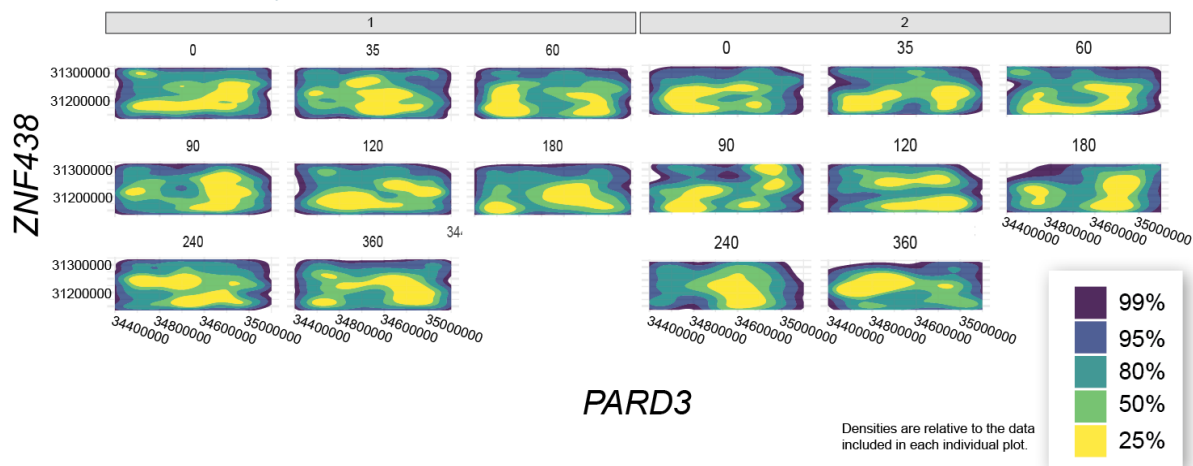


Figure 4.12 Two-Dimensional Kernel Density Estimation of *PARD3* versus *ZNF438*, faceted by biological replicate and time
Kernel density estimation gives the probability density of an event occurring within a given space and/or time, in this case a chromatin-chromatin interaction occurring within a certain span of two gene loci.

KDE plots of *PARD3* versus *ZNF438* were made after faceting data by replicate and time. These plots show consistent “hotspots” in some regions over time, such as that seen in the lower left quadrant, which persists across most time points for both replicates.

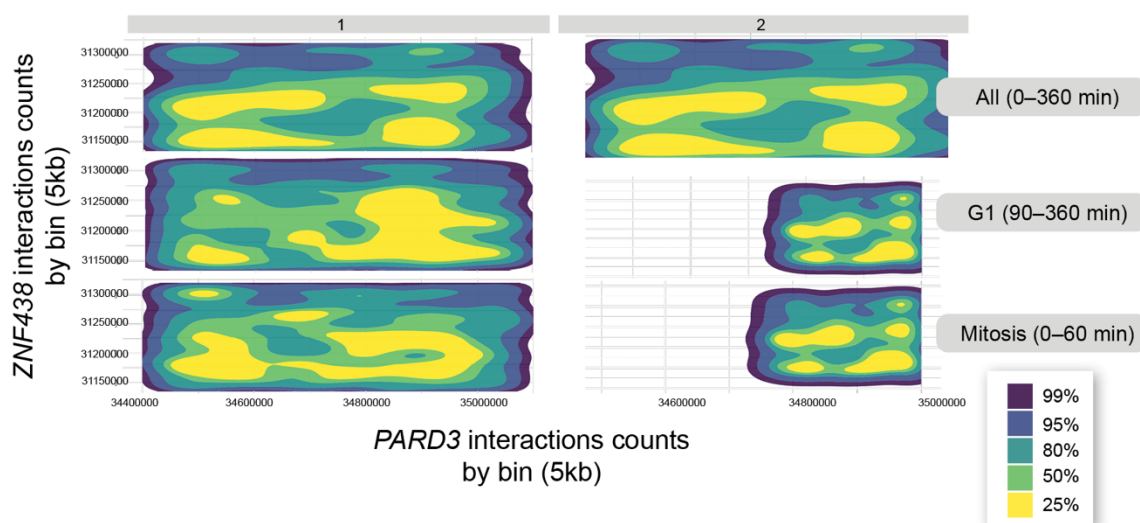


Figure 4.13 Kernel density estimate plots of the interactions between ZNF438 and PARD3 with time “flattened” i.e. multiple timepoints aggregated rather than plotted separately.: (a) All time points aggregated; (b) aggregation of only timepoints where chromatin condensation is not expected (i.e. predominantly taken during G1, excluding Cytokinesis/G1 (samples taken at 90 minutes and beyond); and (c) aggregation of only timepoints where chromatin condensation is expected (up to and including the sample at 60 minutes). See Fig. 4.7 for the sampling timeline. The KDE plot represents the probability density independent of the number of fragments contributing to each level. The concept of fragments contributing to each level differing by plot is expounded upon in using raw fragments in Fig. 4.13.

As expected, the plots displaying data collected during mitosis and cytokinesis has more area covered in the top 20% of the probability density, meaning the area where an interaction is likely to happen is less specific – the points of interaction underlying the kernel density estimation are more evenly distributed over a larger area, they are more spread out. The contrast between the mitosis and G1 is more evident in replicate 1 than in replicate 2.

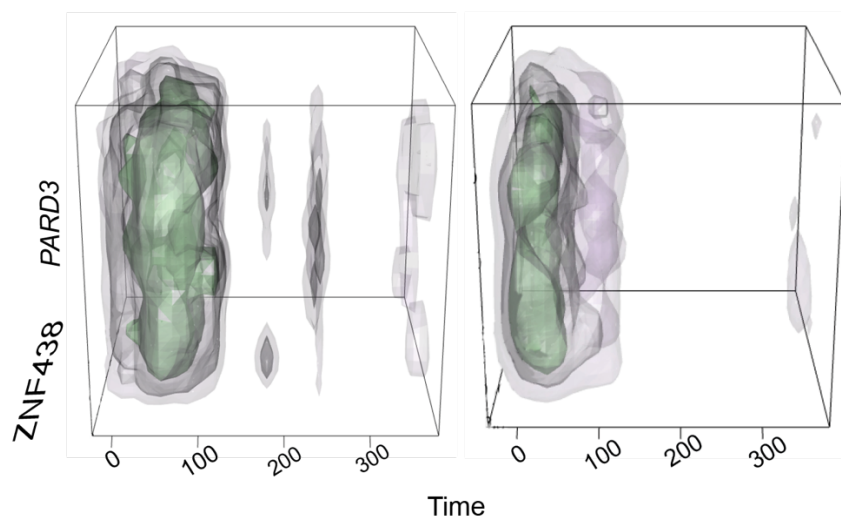


Figure 4.14 Three-Dimensional Kernel Density Estimation Plots, faceted by biological replicate (replicate 1, left; replicate 2, right). Most of the interactions between these two genes were seen during the prometaphase through cytokinesis, which was consistent between replicates.

As expected, and consistent with prior plots, the probability of contact was higher during mitosis. The plot for replicate 1 also indicates the potential for interaction at 240 minutes that is not present for replicate 2.

4.4 Pipeline Development

The initial results of constructing KDE plots from lightly processed sequencing data provide a series of plots with consistent regions of

interaction across the five visualized time points. These plots are overlaid with datapoints representing where the first base pair of a fragment aligns with the reference genome. Future iterations of this visualization, which are already under development, will be constructed as a shiny app and will include the ability to extract fragment information through interaction with the plot.

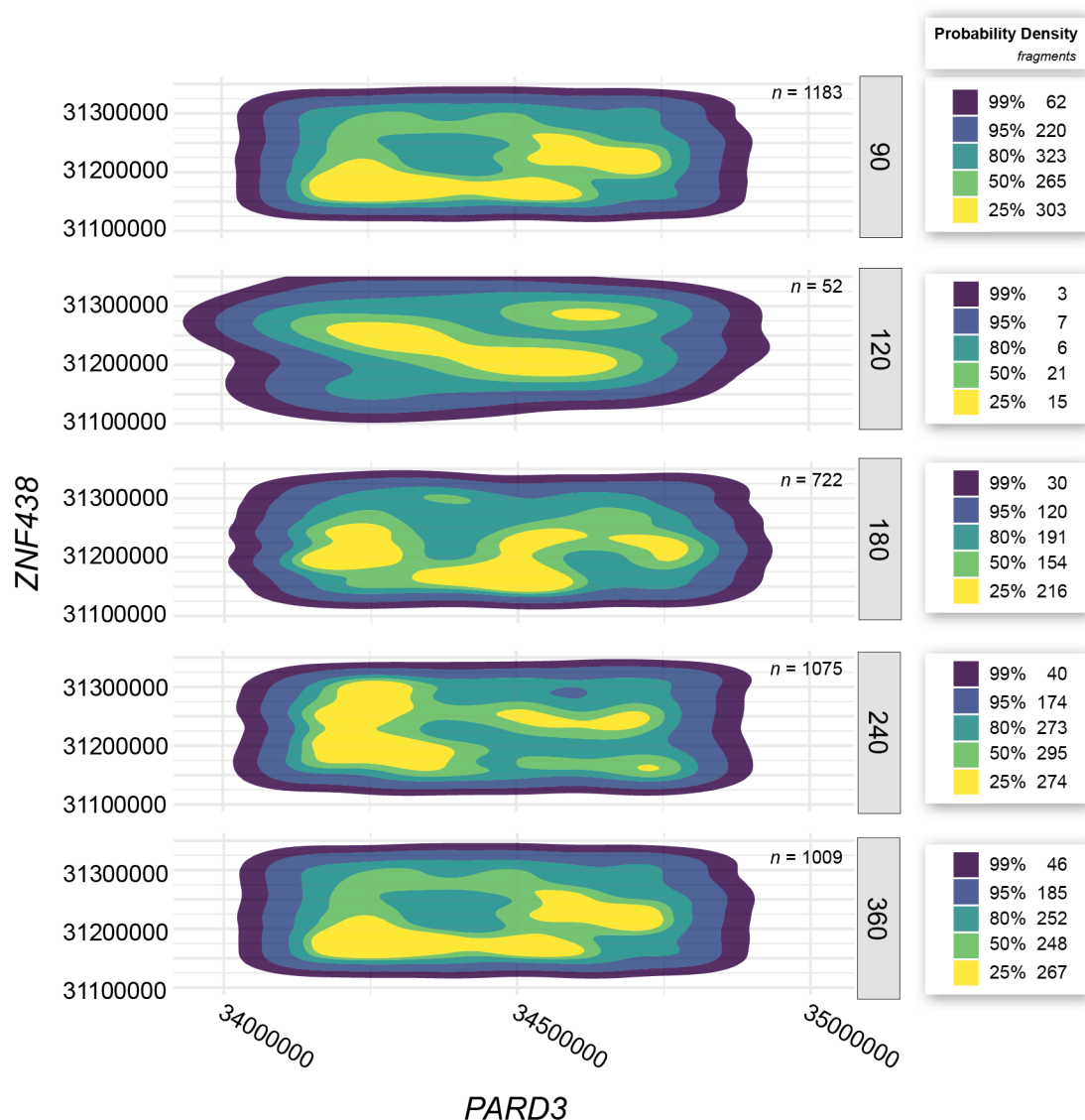


Figure 4.15 Two-Dimensional Kernel Density Estimation using raw Hi-C data for one biological replicate of G1 of the cell cycle in U2OS cells, stratified by time. The data used in these plots was aligned, paired, by not binned or normalized, nor were duplicates removed. Although the colors are the same for each level, they represent the percent relative to the sample size (n , located in the upper right corner of each plot). Listed next to the probability density are the number of fragments from the raw data contributing to each level.

4.5 Discussion

Viewing the same Hi-C dataset from different perspectives benefits researchers. Each visualization method can be used to highlight different features of chromatin organization with

degrees of success hinging, in part, on the scale. For example, an arc diagram suffers from visual clutter when taking a macrolevel view of the entire chromosome. However, it can be useful for looking at genes within the same biological network and/or the same region of a chromosome. The different visualization methods can also be used to complement each other, both when juxtaposed (ex. IGV tracks laid alongside an arc diagram and heatmap) and when used in new configurations (ex. nodes and edges used to present a 3D representation of inter-chromosomal interactions).

Traditional use of heatmaps to visualize Hi-C data does not account for the interactions between neighboring bins. Furthermore, the bins are arbitrarily assigned, with the first bin set relative either to the overall coordinate system of the chromosome or the beginning of a region of the chromosome. The interactions assigned to a bin not only depend on the resolution of the data, which influences bin size selection, there is a dependence on where the bins start – the anchor point. Depending on the location of the bin, the interactions attributed to that bin may affect neighboring bins more or less. To circumvent this issue, we extracted raw data and circumvented the use of the bin entirely. However, this leaves a massive gap between not binned and globally binned data. To address this gap, one of our immediate next steps include the implementation of a local binning schema to connecting of not-binned to binned data at differing resolutions. In this local schema, the first base pair of a feature will be set as the anchor point. To bridge the gap further, the fragments associated with the bin as well as appropriate summary statistics for those fragments within the bin will be available to users. Allowing access to information such as C/G content at this level of granularity facilitates the assessment of the quality of the reads specific to the features of interest. Normalization abstracts the identity of the fragments within a bin away from the count attributed to the bin. Therefore, this access is only possible when interactions are looked at without normalization.

Initially we sought to step away from traditional normalization techniques to reduce the influence of the counts between disparate portions of the genome. Running this train of thought in reverse, we realized that including the influence of interactions between contiguous regions might be beneficial. This potential impact brought to mind infectious disease epidemiology, which looks for point sources by back tracing through interactions.

Epidemiology applies spatiotemporal analysis to examine the probability of an event, such as exposure from a point-source, over time. The effect of neighboring regions is accounted for by such methods, as the proximity in both space and time are considered non-negligible parameters to such models.

Now that these methods are “on the radar”, the next step would be to apply this methodology to exploring the relationships outlined in the overarching introduction (**Fig. 2.4**). Furthermore, a comparison between local and global binning, specifically looking at information loss and read quality as it varies by gene would guide method selection as it depends on the research question. These methods can also be compared to current methods with respect interactions that are not well suited to the most commonly used methods of analysis, namely long range and *trans*-interactions. Additionally, this method of exploration can be applied to probe selection, as originally intended. This includes the ability to look at fragment level or probability density level information when determining the section of a gene to target. These techniques can be used to begin unlocking the potential of NPCs as spatial reference points for the study of genome organization within the nucleus.

The limitations of this study may be better understood through a narrative describing one of the many issues that arose during the research process that hardly seem unique, and yet also seem to remain unaddressed.

For the sake of posterity, we went back to the pre-processed data and tried to repeat the procedure we had repeated previously, but with new information in hand. Over the course of writing, a few details regarding the reference genome originally used to align the Hi-C data came to light, as did the (still a bit unclear) normalization status of the data. To resolve the confusion, the lab group that originated the data was contacted. They graciously answered our inquiries, including those regarding inconsistencies in their methods that made it difficult to discern what was done to arrive at the pre-processed data, as well as providing code that was referenced simply as “in-house code”. What still does not “sit quite right” is with regards to the reference genome originally used to align the data, the implementation of `liftOver` to “downgrade” the results from hg38 to hg19, and to which version of the genome were the pre-processed hic files aligned to at the time they were posted to NCBI GEO. While the answers I was given were delivered with certainty, the fact remains: when I erroneously used hg19 coordinates to search for interactions between PARD3 and ZNF438 in the original data, the result of which can be found in Figs. 4.10–4.13, there were interactions between the two genes. Given they are in close proximity on chromosome 10, the interactions between the two being consistent during mitosis and all-but-non-existent otherwise, holds up to scrutiny. However, having realized the erroneous assumption made previously, the pre-processed hic files was revisited with hg38 coordinates to reproduce the results with the correct coordinates. No interactions came up, though self-with-self interactions did come up for both genes. Even more curious is the pair did not lack read-pairs when the unbinned fragments extracted after aligning the raw reads (replicate 1 contains over 200k read-pairs for the gene pair).

Thus far, the investigation into this peculiar occurrence has not arrived at any answers, but it has come to a conclusion: documentation is exceedingly important both for reproducibility as well as reuse of data.

4.6 Conclusion

Visualizations based on such methods can be applied to Hi-C datasets to depict regions of probability for gene interactions, signal, as well as the persistence of signal over time. We demonstrated this use case by exploring cis-chromosomal interactions between two genes, ZNF438 and PARD3, inspired by a need to reduce the number of oligo probes designed for fluorescent imaging. With the potential interaction space now translatable into probabilities rather than counts, the applications for imaging of this methodology go beyond informing probe design. It is our intent that this research be the first step in bringing together sequencing and imaging data.

5 Overarching Discussion

“Trust the Process.”

This phrase was the first thought this author had when sitting down to write these final words. Except “when sitting down” is a bit of a lie, as she has been at this for over twenty-four hours. That duration was not intentional, but it does speak to the somewhat obsessive nature that this process, the process of squeezing potentially useful information out of imperfect data, requires. Something prolific that ties together all of the research within these pages, and perhaps some of what did not make it onto these pages, is what traditionally goes at the end of a dissertation. However, with no certainty can grandiose proclamations of having discovered anything specifically noteworthy go here. Then again, who is to say what is of note.

The pages of this dissertation include a hypothesis that did not^{lxv} find support, per say. However, the lack of support of the hypothesis can support that the method used to arrive at that conclusion—that is, the methodology resulted in a conclusion that was biologically substantiated. Such was the case in Part I, *Multiomic Integration for Exploring Therapeutic Gene Targets*. The results of this part came with (at least) one clear message: Sometimes the only result available is the most obvious one. If nothing else, at least coming to an obvious conclusion through a myriad of calculations means the methodology itself is worth applying in the future. As trite as it sounds, we found that osteoblasts, regardless of their tumorigenicity, will identify as osteoblasts. Furthermore, the assertion that disease state may be considered a part of cell identity was well enough indicated that it may warrant pursuit using the same methodology in additional cell lines. The result of refining this multiomic analysis may result in process for choosing gene or chromatin locus targets for methodology research.

In terms of selecting probes, it is not yet clear if the results of Part II, *Exploratory Analysis of Hi-C Data for Gene-Gene Interactions* are successful, as this will require *in vitro* verification. The behind-the-scenes process involved in this portion of our research, highlighted that exploring new methods of exploration is incredibly messy, which was in no small part due to side effects of bigger issues within the scientific community at large. We are going to be funding the storage of a lot of data that will not produce robust models if metadata standards are not agreed upon, and compliance verified. But this author will step off that soap box for the time being, as she stood on it long enough in the introduction.

From the messy data and the months (years) of learning to extract information from Hi-C datasets, came an idea that looks promising as an alternative method for analyzing Hi-C data. It may extend the usefulness of these large datasets, as they can be used as preliminary analysis for answering questions regarding specific interactions between genes before applying more targeted approaches. This will facilitate a reduction in expenses by reducing the genomic search space.

In terms of microscopy, once the gene set is narrowed down by other means, such as those in Part I, the methods of Part II can be applied to the subset of a Hi-C dataset inclusive of the genes of interest. The interactions between those genes can then be looked at in the local context, which can guide probe selection for fluorescent imaging.

On the whole, it is the process itself that we come away with, which has laid the groundwork for our future work of integrating sequencing and imaging data. It is no small feat to undertake, which is more apparent having done the work herein, but it is ours to carry forward.

^{lxv} This sentence almost stopped about here with the words “come true.” But really, isn’t that part of doing research? Really hoping your hypothesis comes true, while the rest of academia hopes you’re honest enough to admit when it doesn’t.

6 Additional Notes

The code specifically associated with the analyses within this dissertation are available by request. Additional code related to the methods will be available as it relates to future publications. The author does not recommend downloading the raw data associated with these analyses unless the reader has multiple terabytes available – more if you want to run the Hi-C analysis as the intermediate files are also rather large.

8 Acknowledgements

The order of these acknowledgements does not, in any way, indicate the order of importance. Also, if you have contributed and you are not listed here, it does not reflect my feelings toward you, it reflects my scattered brain – something I would like to claim is caused by the dissertation process but anyone who knows me would see right through that excuse.

Mother... thank you. For everything. For being the mother bunny in the story that you used to read to me. For dragging book reports out of me in middle school – and finding my inclusion of an eyeball on the cover of a book report for *The Congo* amusing rather than horrifying. You've always encouraged my creativity and allowed me a childhood free of any notion there were things that girls just didn't do – including play baseball, even though I wasn't particularly good. I have the urge to pour out a million more hints at the things you've done for me but that would be another dissertation length work and still not cover it all. You are a badass, an inspiration – I'm blessed to have been raised by such a strong, independent woman.

Ken, you've really set the step-parent bar high. I didn't understand how I could love people so much until I met my own step-children... and I have been lucky to have you in my life.

There are many Kevins, but none are like you. Congratulations, you made it with me from one side to the other of a graduate degree... which honestly seems like one of the lesser challenges we have been through together (already?!). If we can get all this done in three years, imagine what we can do together over our lifetimes. OOS!

Doc Brown, thank you so much for encouraging me down this path. I hope I can encourage others to seek places of acceptances, places where they belong and are appreciated for their talents and their quirks, the way you helped me do just that. I might not be a DVM, but you have seen me through becoming a Doctor.

Thank you to my Academic Committee... You have been incredibly patient with me, and it is much appreciated. You gave me a chance to overcome quite a few challenges and here we are, I'm graduating! Seriously, thank you. I took none of your support for granted and I hope I can do the same for others in the future.

David, you get an extra thank you because you saw in me what I am pretty sure few others did, regardless of how dark my hour seemed. You are the change we need to see in academia. And, of course, you're right because by giving me the tools, I can help spread that change. One person at a time.

... and this is where I know that I'm leaving a ton of people out, but hopefully that ton of people know that I am writing this 10 minutes before its due so...

Thank you, and good night!

9 References

1. Danziger, J. & Zimolzak, A. J. Residual Confounding Lurking in Big Data: A Source of Error. in *Secondary Analysis of Electronic Health Records* (ed. MIT Critical Data) (Springer, Cham (CH), 2016).
2. Rejeski, W. J. & Fanning, J. Models and theories of health behavior and clinical interventions in aging: a contemporary, integrative approach. *Clin Interv Aging* **14**, 1007–1019 (2019).
3. Rossi, E. Low Level Environmental Lead Exposure – A Continuing Challenge. *Clin Biochem Rev* **29**, 63–70 (2008).
4. Roy, S., Dietrich, K. N., Gomez, H. F. & Edwards, M. A. Considering Some Negative Implications of an Ever-Decreasing U.S. Centers for Disease Control and Prevention (CDC) Blood Lead Threshold and “No Safe Level” Health Messaging. *Environ. Sci. Technol.* **57**, 12935–12939 (2023).
5. Coulton, C. *et al.* Making the case for lead safe housing: Downstream effects of lead exposure on outcomes for children and youth. *Health Place* **84**, 103118 (2023).
6. Nephew, B. C. *et al.* Traffic-related particulate matter affects behavior, inflammation, and neural integrity in a developmental rodent model. *Environ Res* **183**, 109242 (2020).
7. Hudda, N. *et al.* Bedding-generated particulate matter: implications for rodent studies. *Inhalation Toxicology* **31**, 368–375 (2019).
8. Thangavel, P., Park, D. & Lee, Y.-C. Recent Insights into Particulate Matter (PM_{2.5})-Mediated Toxicity in Humans: An Overview. *Int J Environ Res Public Health* **19**, 7511 (2022).
9. Petitto, J. Mito-Nuclear Retrograde Communication via Mitochondrially Derived Long Non-coding RNA. (2020).
10. Sanchez, M. I. G. L. *et al.* Estrogen-Mediated Regulation of Mitochondrial Gene Expression. *Molecular Endocrinology* **29**, 14–27 (2015).
11. The origins of estrogen receptor alpha-positive and estrogen receptor alpha-negative human breast cancer. *Breast Cancer Research* **6**, 240–245 , pmid = 15535853 (2004).
12. Sun, M. and G. S. S. and K. D. S. and K. W. L. Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Molecular Cell* **59**, 698–711 , pmid = 26236012 (2015).
13. Rigano, A. *et al.* Micro-Meta App: an interactive tool for collecting microscopy metadata based on community specifications. *Nature Methods* **18**, 1489–1495 (2021).
14. Grunwald, D., Huisman, M., Smith, C. & United States Patent and Trademark Office. Systems and Methods of Fluorescence Microscope Calibration. 31 (2022).
15. Huisman, M. *et al.* *Minimum Information Guidelines for Fluorescence Microscopy: Increasing the Value, Quality, and Fidelity of Image Data.*
16. Smith, C. S., Stallinga, S., Lidke, K. A., Rieger, B. & Grunwald, D. Probability-based particle detection that enables threshold-free and robust in vivo single-molecule tracking. *Molecular Biology of the Cell* **26**, 4057–4062 (2015).
17. Smith, C. S. *et al.* An automated Bayesian pipeline for rapid analysis of single-molecule binding data. *Nature Communications* **10**, (2019).

18. Geremia, H. Roll for Alignment: The Application of Moral and Ethical Systems in Tabletop and Digital Dungeons & Dragons. (University of Wollongong, 2022).
19. Genome Browser FAQ > Frequently Asked Questions: Gene tracks. *UCSC Genome Browser* <https://genome.ucsc.edu/FAQ/FAQgenes.html#ens>.
20. Bray, N. L. and P. H. and M. P. and P. L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016).
21. Zhang, P. and S. D. C. and L. B. and S. T. and P. J. and S. Y. and G. Y. Mitochondria sequence mapping strategies and practicability of mitochondria variant detection from exome and RNA sequencing data. *Briefings in Bioinformatics* **17**, 224–232 (2016).
22. Kim, T. K. & Park, J. H. More about the basic assumptions of t-test: normality and sample size. *Korean J Anesthesiol* **72**, 331–335 (2019).
23. Diong, J., Butler, A. A., Gandevia, S. C. & Héroux, M. E. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. *PLoS One* **13**, e0202121 (2018).
24. Brown, A. W., Kaiser, K. A. & Allison, D. B. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences* **115**, 2563–2570 (2018).
25. Thiese, M. S., Arnold, Z. C. & Walker, S. D. The misuse and abuse of statistics in biomedical research. *Biochem Med (Zagreb)* **25**, 5–11 (2015).
26. Davies, H. T. O., Crombie, I. K. & Tavakoli, M. When can odds ratios mislead? *BMJ* **316**, 989–991 (1998).
27. Dahiru, T. P-Value, a true test of statistical significance? a cautionary note. *Annals of Ibadan Postgraduate Medicine* **6**, 21, pmid = 25161440 (2011).
28. Gelman, A. The Problems With P-Values are not Just With P-Values.
29. Ge, X. *et al.* Clipper: p-value-free FDR control on high-throughput data from two conditions. *Genome Biology* **22**, (2021).
30. Devos, D. *et al.* Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol* **2**, e380 (2004).
31. Martin, W. & Koonin, E. V. Introns and the origin of nucleus–cytosol compartmentalization. *Nature* **440**, 41–45 (2006).
32. Makarov, A. A., Padilla-Mejia, N. E. & Field, M. C. Evolution and diversification of the nuclear pore complex. *Biochemical Society Transactions* **49**, 1601–1619 (2021).
33. Baptiste, E., Charlebois, R. L., Macleod, D. & Brochier, C. The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure. *Genome Biology* **6**, R85 (2005).
34. Field, M. C. & Rout, M. P. Pore timing: the evolutionary origins of the nucleus and nuclear pore complex. *F1000Research* **8**, 369 (2019).
35. Kim, S. J. *et al.* Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).
36. Wenthe, S. R. & Rout, M. P. The Nuclear Pore Complex and Nuclear Transport. *Cold Spring Harbor Perspectives in Biology* **2**, (2010).
37. Maul, G. G. *et al.* Time sequence of nuclear pore formation in phytohemagglutinin-stimulated lymphocytes and in HeLa cells during the cell cycle. *Journal of Cell Biology* **55**, 433–447 (1972).

38. Maul, G. G., Price, J. W. & Lieberman, M. W. Formation and distribution of nuclear pore complexes in interphase. *Journal of Cell Biology* **51**, 405–418 (1971).
39. Wunderlich, F. and F. W. W. Structure of macronuclear envelopes of *Tetrahymena pyriformis* in the stationary phase of growth. *The Journal of cell biology* **38**, 458–462 (1968).
40. Maul, G. & Deaven, L. Quantitative determination of nuclear pore complexes in cycling cells with differing DNA content. *Journal of Cell Biology* **73**, 748–760 (1977).
41. Zuleger, N., Robson, M. I. & Schirmer, E. C. The nuclear envelope as a chromatin organizer. *Nucleus* **2**, 339–49 (2011).
42. Ibarra, A., Benner, C., Tyagi, S., Cool, J. & Hetzer, M. W. Nucleoporin-mediated regulation of cell identity genes. *Genes Dev* **30**, 2253–2258 (2016).
43. Bersini, S. *et al.* Nup93 regulates breast tumor growth by modulating cell proliferation and actin cytoskeleton remodeling. *Life Science Alliance* **3**, e201900623 (2020).
44. Nguyen, T. D. *et al.* Nucleoporin93 (Nup93) Limits Yap Activity to Prevent Endothelial Cell Senescence. *bioRxiv* 2023.11.10.566598 (2023) doi:10.1101/2023.11.10.566598.
45. Labade, A. S., Salvi, A., Karmodiya, K. & Sengupta, K. *Nup93 and CTCF Co-Modulate Spatiotemporal Dynamics and Function of the HOXA Gene Cluster during Differentiation*. <http://biorxiv.org/lookup/doi/10.1101/646224> (2019) doi:10.1101/646224.
46. Labade, A. S., Karmodiya, K. & Sengupta, K. HOXA repression is mediated by nucleoporin Nup93 assisted by its interactors Nup188 and Nup205. *Epigenetics & Chromatin* **9**, (2016).
47. Toda, T. *et al.* Nup153 Interacts with Sox2 to Enable Bimodal Gene Regulation and Maintenance of Neural Progenitor Cells. *Cell Stem Cell* **21**, 618–634.e7 (2017).
48. Jacinto, F. V., Benner, C. & Hetzer, M. W. The nucleoporin Nup153 regulates embryonic stem cell pluripotency through gene silencing. *Genes Dev* **29**, 1224–38 (2015).
49. Zhou, L. & Panté, N. The nucleoporin Nup153 maintains nuclear envelope architecture and is required for cell migration in tumor cells. *FEBS Letters* **584**, 3013–3020 (2010).
50. Ng, S. C. *et al.* Barrier properties of Nup98 FG phases ruled by FG motif identity and inter-FG spacer length. *Nature Communications* **14**, (2023).
51. Singer, S. *et al.* Nuclear Pore Component Nup98 Is a Potential Tumor Suppressor and Regulates Posttranscriptional Expression of Select p53 Target Genes. *Molecular Cell* **48**, 799–810 (2012).
52. Bostrom, B. N. Are we living in a computer simulation? *Philosophical Quarterly* **53**, 243–255 (2003).
53. Garfinkel, A., Shevtsov, J. & Guo, Y. *Modeling Life*. (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-59731-7.
54. Mcquilton, P. *et al.* BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database* **2016**, baw075 (2016).
55. Hammer, M. *et al.* Towards community-driven metadata standards for light microscopy: tiered specifications extending the OME model. *Nature Methods* **18**, 1427–1440 (2021).
56. Gillman, D. Achieving Transparency: A Metadata Perspective. *Data Intelligence* **5**, 261–274 (2023).
57. Linkert, M. *et al.* Metadata matters: access to image data in the real world. *J Cell Biol* **189**, 777–782 (2010).
58. Rajesh, A. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biology* **22**, 106 (2021).

59. Engineering National Academies of Sciences. *Reproducibility and Replicability in Science*. (The National Academies Press, Washington, DC, 2019).
60. Reviewing peer review. *Nat Immunol* **4**, 297–297 (2003).
61. Nature will publish peer review reports as a trial. *Nature* **578**, 8–8 (2020).
62. Simoneau, J. and D. S. and G. R. and S. M. S. Current RNA-seq methodology reporting limits reproducibility. *Briefings in Bioinformatics* (2019) doi:10.1093/bib/bbz124 , file = ::
63. Boehm, U. *et al.* QUAREP-LiMi: a community endeavor to advance quality assessment and reproducibility in light microscopy. *Nature Methods* **18**, 1423–1426 (2021).
64. Schmied, C. *et al.* Community-developed checklists for publishing images and image analyses. *Nature Methods* (2023) doi:10.1038/s41592-023-01987-9.
65. Edgar, R. & Barrett, T. NCBI GEO standards and services for microarray data. *Nat Biotechnol* **24**, 1471–1472 (2006).
66. Marti-Renom, M. A. *et al.* Challenges and guidelines toward 4D nucleome data and model standards. *Nat Genet* **50**, 1352–1358 (2018).
67. Heil, B. J. *et al.* Reproducibility standards for machine learning in the life sciences. *Nature Methods* **18**, 1132–1135 (2021).
68. Citation listings - GEO - NCBI. *NCBI* <https://www.ncbi.nlm.nih.gov/geo/info/citations.html>.
69. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat Rev Genet* **14**, 89–99 (2013).
70. Kang, M., Ko, E. & Mersha, T. B. A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics* **23**, bbab454 (2022).
71. Heus, P. AI has a metadata problem.... *Medium* <https://plgah.medium.com/ai-has-a-metadata-problem-78b30ca1936b> (2023).
72. Hintze, A., Olson, R. S., Adami, C. & Hertwig, R. Risk sensitivity as an evolutionary adaptation. *Sci Rep* **5**, 8242 (2015).
73. Fisher, H. & Louw, I. Teaching mise-en-place: Student perceptions of the cooking pro forma process. *International Journal of Gastronomy and Food Science* **22**, 100245 (2020).
74. Grunwald, D. & Singer, R. H. In vivo imaging of labelled endogenous beta-actin mRNA during nucleocytoplasmic transport. *Nature* **467**, 604–7 (2010).
75. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**, 934–947 (2013).
76. Abby Sadler. Pricing Proposal. (2021).
77. Gabriel, S. E. & Michaud, K. Epidemiological studies in incidence, prevalence, mortality, and comorbidity of the rheumatic diseases. *Arthritis Res Ther* **11**, 229 (2009).
78. Jawaheer, D., Lum, R. F., Gregersen, P. K. & Criswell, L. A. Influence of male sex on disease phenotype in familial rheumatoid arthritis. *Arthritis Rheum* **54**, 3087–3094 (2006).
79. Burmester, G. R. & Pope, J. E. Novel treatment strategies in rheumatoid arthritis. *The Lancet* **389**, 2338–2348 (2017).
80. Aletaha, D. & Smolen, J. S. Diagnosis and Management of Rheumatoid Arthritis: A Review. *JAMA* **320**, 1360–1372 (2018).
81. Brevetti, G., Giugliano, G., Brevetti, L. & Hiatt, W. R. Inflammation in Peripheral Artery Disease. *Circulation* **122**, 1862–1875 (2010).

82. Nijenhuis, S., Zendman, A. J. W., Vossenaar, E. R., Pruijn, G. J. M. & vanVenrooij, W. J. Autoantibodies to citrullinated proteins in rheumatoid arthritis: clinical performance and biochemical aspects of an RA-specific marker. *Clinica Chimica Acta* **350**, 17–34 (2004).
83. Abdul Wahab, A., Mohammad, M., Rahman, M. M. & Mohamed Said, Mohd. S. Anti-cyclic citrullinated peptide antibody is a good indicator for the diagnosis of rheumatoid arthritis. *Pak J Med Sci* **29**, 773–777 (2013).
84. van der Woude, D. *et al.* Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis & Rheumatism* **60**, 916–923 (2009).
85. van Boekel, M. A., Vossenaar, E. R., van den Hoogen, F. H. & van Venrooij, W. J. Autoantibody systems in rheumatoid arthritis: specificity, sensitivity and diagnostic value. *Arthritis Res Ther* **4**, 87 (2001).
86. Paul I. Terasaki. Transplant Antigens: A Brief History of HLA. in *Textbook of Organ Transplantation* 30–35 (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118873434.ch3.
87. Choo, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Med J* **48**, 11–23 (2007).
88. van der Helm-van Mil, A. H. M. *et al.* An independent role of protective HLA class II alleles in rheumatoid arthritis severity and susceptibility. *Arthritis Rheum* **52**, 2637–2644 (2005).
89. Holoshitz, J. The Rheumatoid Arthritis HLA-DRB1 Shared Epitope. *Curr Opin Rheumatol* **22**, 293–298 (2010).
90. Atkinson, S. M. *et al.* Establishment and characterization of a sustained delayed-type hypersensitivity model with arthritic manifestations in C57BL/6J mice. *Arthritis Res Ther* **14**, R134 (2012).
91. Zhao, T. *et al.* How to Model Rheumatoid Arthritis in Animals: From Rodents to Non-Human Primates. *Frontiers in Immunology* **13**, (2022).
92. Damerou, A., Lang, A., Pfeiffeberger, M., Buttgereti, F. & Gaber, T. FRI0002 Development of an in vitro multi-component 3d joint model to simulate the pathogenesis of arthritis | Annals of the Rheumatic Diseases. *Annals of Rheumatic Diseases* **76**, 420.2-420 (2017).
93. Lin, Z. *et al.* Osteochondral Tissue Chip Derived From iPSCs: Modeling OA Pathologies and Testing Drugs. *Front Bioeng Biotechnol* **7**, 411 (2019).
94. Ponten, J. & Saksela, E. Two established in vitro cell lines from human mesenchymal tumours. *Intl Journal of Cancer* **2**, 434–447 (1967).
95. Kido, T. & Lau, Y.-F. C. Roles of the Y chromosome genes in human cancers. *Asian J Androl* **17**, 373–380 (2015).
96. Mohseny, A. B. *et al.* Functional characterization of osteosarcoma cell lines provides representative models to study the human disease. *Lab Invest* **91**, 1195–1205 (2011).
97. Pautke, C. *et al.* Characterization of Osteosarcoma Cell Lines MG-63, Saos-2 and U-2 OS in Comparison to Human Osteoblasts. *ANTICANCER RESEARCH* (2004).
98. RFA-AR-21-016: Accelerating Medicines Partnership Autoimmune and Immune-Mediated Diseases: Technology and Analytic Cores (TACs) and Research Management Unit (RMU) (UC2 Clinical Trial Not Allowed). <https://grants.nih.gov/grants/guide/rfa-files/RFA-AR-21-016.html>.
99. Kang, J. B. *et al.* Mapping the dynamic genetic regulatory architecture of HLA genes at single-cell resolution. *Nat Genet* **55**, 2255–2268 (2023).

100. Gupta, A. *et al.* Dynamic regulatory elements in single-cell multimodal data implicate key immune cell states enriched for autoimmune disease heritability. *Nat Genet* **55**, 2200–2210 (2023).
101. Okada, Y., Eyre, S., Suzuki, A., Kochi, Y. & Yamamoto, K. Genetics of rheumatoid arthritis: 2018 status. *Ann Rheum Dis* **78**, 446–453 (2019).
102. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–81 (2014).
103. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* **49**, D1302–D1310 (2021).
104. Croft, J. A. *et al.* Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. *Journal of Cell Biology* **145**, 1119–1131 (1999).
105. Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8**, 104–115 (2007).
106. Smith, C. S. *et al.* Nuclear accessibility of β -actin mRNA is measured by 3D single-molecule real-time tracking. *Journal of Cell Biology* **209**, 609–619 (2015).
107. Schermelleh, L. *et al.* Subdiffraction Multicolor Imaging of the Nuclear Periphery with 3D Structured Illumination Microscopy. *Science* **320**, 1332–1336 (2008).
108. Pinzaru, A. M. *et al.* Telomere relocalization to the nuclear pore complex in response to replication stress. (2020).
109. Khadaroo, B. *et al.* The DNA damage response at eroded telomeres and tethering to the nuclear pore complex. *Nat Cell Biol* **11**, 980–7 (2009).
110. Oza, P. & Peterson, C. L. Opening the DNA repair toolbox: localization of DNA double strand breaks to the nuclear periphery. *Cell Cycle* **9**, 43–9 (2010).
111. Oza, P., Jaspersen, S. L., Miele, A., Dekker, J. & Peterson, C. L. Mechanisms that regulate localization of a DNA double-strand break to the nuclear periphery. *Genes Dev* **23**, 912–27 (2009).
112. Boumendil, C., Hari, P., Olsen, K. C. F., Acosta, J. C. & Bickmore, W. A. Nuclear pore density controls heterochromatin reorganization during senescence. *Genes Dev* **33**, 144–149 (2019).
113. Sun, J., Shi, Y. & Yildirim, E. The Nuclear Pore Complex in Cell Type-Specific Chromatin Structure and Gene Regulation. *Trends in Genetics* **35**, 579–588 (2019).
114. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
115. Ibarra, A. & Hetzer, M. W. Nuclear pore proteins and the control of genome functions. *Genes Dev* **29**, 337–49 (2015).
116. Neumann, N., Lundin, D. & Poole, A. M. Comparative Genomic Evidence for a Complete Nuclear Pore Complex in the Last Eukaryotic Common Ancestor. *PLOS ONE* **5**, e13241 (2010).
117. Nataraj, N. B. *et al.* Nucleoporin-93 reveals a common feature of aggressive breast cancers: robust nucleocytoplasmic transport of transcription factors. *Cell Rep* **38**, 110418 (2022).
118. Kadota, S. *et al.* Nucleoporin 153 links nuclear pore complex to chromatin architecture by mediating CTCF and cohesin binding. *Nature Communications* **11**, (2020).

119. McCloskey, A., Ibarra, A. & Hetzer, M. W. Tpr regulates the total number of nuclear pore complexes per cell nucleus. *Genes Dev* **32**, 1321–1331 (2018).
120. Aughey, G. N., Cheetham, S. W. & Southall, T. D. DamID as a versatile tool for understanding gene regulation. *Development* **146**, dev173666 (2019).
121. Vogel, M. J., Peric-Hupkes, D. & Van Steensel, B. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nat Protoc* **2**, 1467–1478 (2007).
122. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat Biotechnol* **18**, 424–428 (2000).
123. What is RNAi - RNAi Biology. *UMass Chan Medical School* <https://www.umassmed.edu/rti/biology/rna/how-rnai-works/> (2013).
124. Michael Loeffelholz, Richard Hodinka, Stephen Young, & Benjamin Pinsky. *Clinical Virology Manual*. (Wiley-Blackwell, 2016).
125. Ashley J Pratt & Ian J MacRae. The RNA-induced Silencing Complex: A Versatile Gene-silencing Machine. *The Journal of Biological Chemistry* **284**, 17897–17901 (2009).
126. Semizarov, D., Kroeger, P. & Fesik, S. siRNA-mediated gene silencing: a global genome view. *Nucleic Acids Res* **32**, 3836–3845 (2004).
127. RNA Interference (RNAi). <https://www.ncbi.nlm.nih.gov/probe/docs/technai/>.
128. Wang, F. *et al.* A comparison of CRISPR/Cas9 and siRNA-mediated ALDH2 gene silencing in human cell lines. *Mol Genet Genomics* **293**, 769–783 (2018).
129. R Core Team. *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/> (2023).
130. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).
131. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**, 157–175 (1900).
132. Haviland, M. G. Yates’s correction for continuity and the analysis of 2 × 2 contingency tables. *Statistics in Medicine* **9**, 363–367 (1990).
133. Mantel, N. & Haenszel, W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute* **22**, 719–748 (1959).
134. Cochran, W. G. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics* **10**, 417–451 (1954).
135. Howell, D. C. *Statistical Methods for Psychology*. (Wadsworth, Cengage Learning, 2010).
136. Burnham, K. P. & Anderson, D. R. Information Theory and Log-Likelihood Models: A Basis for Model Selection and Inference. in *Model Selection and Inference* 32–74 (Springer New York, New York, NY, 1998). doi:10.1007/978-1-4757-2917-7_2.
137. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974).
138. Cressie, N. & Read, T. R. C. Pearson’s X^2 and the Loglikelihood Ratio Statistic G^2 : A Comparative Review. *International Statistical Review / Revue Internationale de Statistique* **57**, 19–43 (1989).

139. Wilks, S. S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* **9**, 60–62 (1938).
140. Chapman, R. A. Applicability of the z Test to a Poisson Distribution. *Biometrika* **30**, 188 (1938).
141. Lee, S. & Lee, D. K. What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology* **71**, 353–360 (2018).
142. Bonferroni, C. E. Teoria statistica delle classi e calcolo delle probabilità. (1936).
143. Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I. & Wright, F. A. Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets. *BMC Genomics* **11**, 574 (2010).
144. Royston, P. Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44**, 547–551 (1995).
145. Royston, P. Approximating the Shapiro-Wilk W-test for non-normality. *Stat Comput* **2**, 117–119 (1992).
146. Quade, D. On the Asymptotic Power of the One-Sample Kolmogorov-Smirnov Tests. *The Annals of Mathematical Statistics* **36**, 1000–1018 (1965).
147. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* **47**, 583–621 (1952).
148. Blair, R. C. & Higgins, J. J. Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin* **97**, 119–128 (1985).
149. Neuhausser, M. & Bretz, F. Nonparametric all-pairs multiple comparisons. *Biometrical Journal* **43**, 571–580 (2001).
150. Johan Larsson. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. (2024).
151. The Apache Software Foundation. *SparkR: R Front End for 'Apache Spark'*. (2023).
152. Zaharia, M. *et al.* Apache Spark: a unified engine for big data processing. *Commun. ACM* **59**, 56–65 (2016).
153. Adams, G. S., Converse, B. A., Hales, A. H. & Klotz, L. E. People systematically overlook subtractive changes. *Nature* **592**, 258–261 (2021).
154. Meyvis, T. & Yoon, H. Adding is favoured over subtracting in problem solving. *Nature* **592**, 189–190 (2021).
155. *Less Is More: Why Our Brains Struggle to Subtract*. (2021).
156. Home | HUGO Gene Nomenclature Committee. *HUGO Gene Nomenclature Committee at the University of Cambridge* <https://www.genenames.org/>.
157. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software* **64**, 1–34 (2015).
158. Frey, H. C. & Burmaster, D. E. Methods for Characterizing Variability and Uncertainty: Comparison of Bootstrap Simulation and Likelihood-Based Approaches. *Risk Analysis* **19**, 109–130 (1999).
159. Lee, J., Hyeon, Y. D. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine* **52**, 1428–1442 (2020).
160. Droste, Miebach, S., Niedenfhr, S., Wiechert, W. & Nh, K. Visualizing multi-omics data in metabolic networks with the software Omix-A case study. *BioSystems* **105**, 154–161 (2011).

161. Rajasundaram, D. and S. J. More effort - more results: Recent advances in integrative 'omics' data analysis. *Current Opinion in Plant Biology* **30**, 57--61 (2016).
162. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51**, D1353–D1359 (2023).
163. Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research* **47**, D1056–D1065 (2019).
164. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research* **45**, D985–D994 (2017).
165. Ullman, K. S., Shah, S., Powers, M. A. & Forbes, D. J. The Nucleoporin Nup153 Plays a Critical Role in Multiple Types of Nuclear Export. *Molecular Biology of the Cell* **10**, 649–664 (1999).
166. Rabut, G., Doye, V. & Ellenberg, J. Mapping the dynamic organization of the nuclear pore complex inside single living cells. *Nat Cell Biol* **6**, 1114–1121 (2004).
167. Briand, N. & Collas, P. Lamina-associated domains: peripheral matters and internal affairs. *Genome Biology* **21**, (2020).
168. Pascual-Garcia, P. & Capelson, M. The nuclear pore complex and the genome: organizing and regulatory principles. *Curr Opin Genet Dev* **67**, 142–150 (2021).
169. Sharma, S. V., Haber, D. A. & Settleman, J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer* **10**, 241–253 (2010).
170. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Research* **44**, D164–D171 (2016).
171. Comşa, Ş., Cîmpean, A. M. & Raica, M. The story of MCF-7 breast cancer cell line: 40 Years of experience in research. *Anticancer Research* **35**, 3147--3154 (2015).
172. Zhu, X. *et al.* Acute depletion of human core nucleoporin reveals direct roles in transcription control but dispensability for 3D genome organization. *Cell Rep* **41**, 111576 (2022).
173. Hoffman, E. A., Frey, B. L., Smith, L. M. & Auble, D. T. Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes. *Journal of Biological Chemistry* **290**, 26404–26411 (2015).
174. Kang, H. *et al.* Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. *Genes & Development* **34**, 913–930 (2020).
175. Gregory J Phillips, Jonathan Arnold, & Rober Ivarie. Mono- through hexanucleotide composition of the Escherichia coli genome: a Markov chian analysis. *Nucleic Acids Research* **15**, 2611–2626 (1987).
176. Bohlin, J., Skjerve, E. & Ussery, D. W. Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* **9**, 104 (2008).
177. Nieuwenhuizen, R. P. J. *et al.* Measuring image resolution in optical nanoscopy. *Nat Methods* **10**, 557–562 (2013).
178. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* **72**, 65–75 (2014).
179. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665--1680 (2014).

180. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* **17**, 743–755 (2016).
181. Jocelyn Tourtellotte & Bastiaan Dekker. HiC Cross linking distance. (2023).
182. sparse matrix. <https://xlinux.nist.gov/dads/HTML/sparsematrix.html>.
183. Yan, D., Wu, T., Liu, Y. & Gao, Y. An efficient sparse-dense matrix multiplication on a multicore system. in *2017 IEEE 17th International Conference on Communication Technology (ICCT)* 1880–1883 (2017). doi:10.1109/ICCT.2017.8359956.
184. Sparse Matrices - MATLAB & Simulink. <https://www.mathworks.com/help/matlab/sparse-matrices.html>.
185. Lee, S. Hi-C Data Formats. in *Hi-C Data Analysis: Methods and Protocols* (eds. Silvio Bicciato & Francesco Ferrari) vol. 2301 133–141 (Springer Nature, 2022).
186. Martin, G. E. *Counting: The Art of Enumerative Combinatorics*. (Springer New York, New York, NY, 2001). doi:10.1007/978-1-4757-4878-9.
187. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* **43**, (2011).
188. Lyu, H., Liu, E. & Wu, Z. Comparison of normalization methods for Hi-C data. *BioTechniques* **68**, 56–64 (2020).
189. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
190. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).
191. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* **9**, 999–1003 (2012).
192. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* **51**, 1664–1669 (2019).
193. Yardimci, G. G. & Noble, W. S. Software tools for visualizing Hi-C data. (2017).
194. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophys Rev* **11**, 67–78 (2019).
195. Harmston, N., Ing-Simmons, E., Perry, M., Baresic, A. & Lenhard, B. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics* **16**, 963 (2015).
196. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
197. Lun, A. T. L. & Smyth, G. K. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**, (2015).
198. Pal, K., Tagliaferri, I., Livi, C. M. & Ferrari, F. HiCBricks: building blocks for efficient handling of large Hi-C datasets. *Bioinformatics* **36**, 1917–9 (2019).
199. Shinkai, S., Itoga, H., Kyoda, K. & Onami, S. PHi-C2: interpreting Hi-C data as the dynamic 3D genome state. *Bioinformatics* **38**, 4984–4986 (2022).
200. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101 (2016).
201. Robinson, J. T. *et al.* Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Systems* **6**, 256-258.e1 (2018).

202. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–92 (2013).
203. Robinson, J. T., Thorvaldsdottir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**, (2023).
204. Colin Ware. Designing with a 2 1/2D Attitude. *Information Design Journal* **10**, 255–262 (2001).
205. Peter J Park. *4D Nucleome Network Data Coordination and Integration Center*. (2020).
206. Mackay, K., Kusalik, A. & Eskiw, C. H. GrapHi-C: graph-based visualization of Hi-C datasets. *BMC Research Notes* **11**, (2018).
207. Liu, L., Zhang, B. & Hyeon, C. Extracting multi-way chromatin contacts from Hi-C data. *PLoS Computational Biology* **17**, e1009669 (2021).
208. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–12 (2004).
209. Djekidel, M. N., Wang, M., Zhang, M. Q. & Gao, J. HiC-3DViewer: a new tool to visualize Hi-C data in 3D space. *Quantitative Biology* **5**, 183–190 (2017).
210. Lafontaine, D. L., Yang, L., Dekker, J. & Gibcus, J. H. Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture. *Current Protocols* **1**, e198 (2021).
211. José E Chacón & Tarn Duong. Multivariate Kernel Smoothing and Its Applications.
212. Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**, 20–32 (2018).
213. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2012).
214. Akan, P. *et al.* Comprehensive analysis of the genome transcriptome and proteome landscapes of three tumor cell lines. *Genome Medicine* **4**, 86 (2012).
215. Hostetter, M., Kranz, D. A., Seed, C., Terman, C. & Ward, S. Curl: a gentle slope language for the Web. *World wide web journal* **2**, 121–134 (1997).
216. Gálvez-Merchán, Á., Min, K. H. (Joseph), Pachter, L. & Boeshaghi, A. S. Metadata retrieval from sequence databases with ffq. *Bioinformatics* **39**, btac667 (2023).
217. Hershberg, E. A. *et al.* PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. *Nat Methods* **18**, 937–944 (2021).
218. Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet* **14**, e1007872 (2018).
219. Shi, X. *et al.* Estimation of environmental exposure: interpolation, kernel density estimation or snapshotting. *Annals of GIS* **25**, 1–8 (2019).
220. Brian D. Ripley *et al.* Support Functions and Datasets for Venables and Ripley's MASS. CRAN (2023).
221. Deng, H. & Wickham, H. Density estimation in R. (2011).
222. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, New York, NY, 2002).
223. Golloshi, R., Sanders, J. T. & McCord, R. P. Genome organization during the cell cycle: unity in division. *Wiley Interdiscip Rev Syst Biol Med* **9**, (2017).

224. Hi-C Processing Pipeline – 4DN Data Portal.
https://data.4dnucleome.org/resources/data-analysis/hi_c-processing-pipeline.
225. SRA Toolkit Development Team. The SRA Toolkit. NCBI.
226. Pan, B. and K. R. and X. W. and Z. Y. and L. Z. and X. C. and S. S. and G. W. and G. P. and Z. C. and G. W. and S. L. and T. W. and H. H. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* 2019 20:2 **20**, 17--29 (2019).
227. Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
228. Open2C *et al.* Pairtools: from sequencing data to chromosome contacts. 2023.02.13.528389 Preprint at <https://doi.org/10.1101/2023.02.13.528389> (2023).
229. Macqueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1**, 281–297 (1967).