A Method for Observing Genome-Wide Chromatin Conformation Changes Caused by Premature Transcription Termination of lncRNA *HOTAIRM1*

Natalie Schalick
Biochemistry
McGill University, Montréal

October 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Science in Biochemistry

Table of Contents

I. Abstract

English

HOTAIRM1 is a long non-coding RNA located between, and antisense to, *HOXA1* and *HOXA2*. Prior work in our laboratory shows that *HOTAIRM1* regulates the local chromatin architecture at the *HOXA* locus 3' end during retinoic acid-induced differentiation in NT2-D1 cells through a mechanism involving interactions with the histone-modifying complexes PRC2 and UTX/MLL. As *HOTAIRM1* is thought to contribute to the severity of different cancer types and other health conditions- some of which mechanistically involve PRC2- a genome-wide survey of *HOTAIRM1*'s effect on chromatin conformation is desirable. To this end, I present an approach to identify chromatin conformation changes in the presence or absence of *HOTAIRM1* during cellular differentiation. This method uses the CRISPR-Cas9 technology to knock-in polyA tail downstream of *HOTAIRM1*'s transcriptional start site, terminating its transcription, and harnesses the power of Hi-C to map chromatin conformation genome-wide at different stages of cellular differentiation.

Français

HOTAIRM1 est un long ARN non codant situé entre, et antisens de, *HOXA1* et *HOXA2*. Des travaux antérieurs dans notre laboratoire démontrent qu'*HOTAIRM1* régule l'architecture locale de la chromatine en 3' du locus *HOXA* lors de la différenciation induite par l'acide rétinoïque dans les cellules NT2-D1 par un mécanisme impliquant des interactions avec les complexes modificateurs d'histones PRC2 et UTX/MLL. Puisque *HOTAIRM1* contribue à la gravité de différents types de cancer et d'autres problèmes de santé - certains d'entre eux impliquant le complexe PRC2 - une enquête de l'effet de *HOTAIRM1* sur la conformation de la chromatine à l'échelle du génome est souhaitable. À cette fin, je présente une méthode combinant la technologie CRISPR-Cas9 pour insérer un signal de polyadénosylation immédiatement en aval du site d'initiation de transcription d'*HOTAIRM1*, mettant fin à sa transcription, ainsi que la méthode Hi-C utilisée pour cartographier l'effet de l'expression de cet ARN non-codant sur l'architecture tridimensionelle du génome à différents stages de différentiation cellulaires.

II. Acknowledgements

The work detailed in this thesis was supervised by Dr. Josée Dostie. Dr. Mathieu Blanchette and Dr. Jerry Pelletier served on the Research Advisory Committee.

Dana Segal provided fixed NCCIT cell pellets used to generate Hi-C libraries and provided guidance in laboratory activities and equipment usage.

The protocols detailed in this thesis were based in part on protocols developed by Dana Segal, David Wang, Denis Paquette, and Kris Cao. Specifically, the work of Denis Paquette and Kris Cao contributed to the Hi-C protocol, and the work of Dana Segal and David Wang contributed to the bacterial cloning and tissue culture protocols.

Kaiwan Rahimian provided guidance in using Excel to compare chromosome compartment types and the script for calculating the GC percentage of genomic DNA per bin. Christopher Cameron and Dr. Mathieu Blanchette provided guidance in using the HIFI program, and Dr. Blanchette provided additional guidance in using Juicer and interpreting Eigenvector calculations. Samy Coulombe provided guidance in the usage of C3G's Genpipes pipelines.

Thank you to all current and former laboratory members, including Afrida Ahmed, Ian Zhao, Rachel Ding, Romane Monnet, Xian Jin Lian, and Theodora Georgakopoulou in addition to those mentioned above, for camaraderie and advice.

III. Contribution of Authors

All chapters were written by me. Feedback was provided by Dr. Josée Dostie and Dr. Natasha Chang.

The script used to calculate average GC content per genomic bin (calculate_GC.py, Supplementary Materials) was written by Kaiwan Rahimian and edited by me for compatibility with a Windows operating system.

IV. List of Figures and Abbreviations

Figures and Tables:

Supplementary Figures and Tables:

File S1. Supplementary Tables

Table SI. Characteristics of the bin(s) genes of interest occupy.

Table SII. Characteristics of the compartment(s) genes of interest occupy.

Supplementary Data Files:

File S2. Processed and binned data derived from Hi-C experiments in H1-hESC and definitive endoderm (Akgol Oksuz et al., 2021) including Eigenvector values, compartment calls, RNA-seq expression data, and gene density per bin.

File S3A. Processed data derived from the H1-hESC Hi-C dataset described above wherein the compartments are concatenated.

File S3B. Processed data derived from the definitive endoderm Hi-C dataset described above wherein the compartments are concatenated.

File S4-S31. Excel Workbooks containing data from the starting and ending points of genes of interest. The gene name and whether the data describes bins or concatenated compartments are indicated in the filenames.

File S32. The text contents of all Python scripts described throughout this thesis.

File S33. The sequences of all plasmids and oligonucleotides designed for this project.

Abbreviations:

HOTAIRM1: HOX antisense intergenic RNA myeloid 1

lncRNA: long non-coding RNA

TAD: topologically associated domain

sgRNA: small guide RNA

NCCIT: a pluripotent human cell line

NT2-D1: a pluripotent human cell line

RA: retinoic acid

polyA: Polyadenylation

RT-qPCR: real-time quantitative polymerase chain reaction

ChIP: Chromatin immunoprecipitation

LIHC: liver hepatocellular carcinoma

ccRCC: clear cell renal cell carcinoma

TAMR: tamoxifen resistance

3C: Chromatin conformation capture

4C: Circular chromosome conformation capture

5C: Chromatin conformation capture carbon copy

V. Introduction

Non-coding RNAs have a variety of functions in mammalian cells, from acting as decoys for microRNA or transcription factors to serving as scaffolds for protein complexes [1]. Notably, some long non-coding RNAs (lncRNAs) have demonstrable effects on chromatin conformation. Individual chromosomes organize in three-dimensional space at several levels, including transcriptionally active or inactive compartments and topologically associating domains (TADs). Compartment types and TAD boundary locations in chromatin often correspond to the transcriptional activity of those loci, as elements such as transcription start sites, transcription factors, or insulator proteins are enriched near boundaries [2]. These boundaries can fluctuate during processes such as cell differentiation, in HOX clusters for example [3]. Prior research in the Dostie laboratory has shown that the temporal activation of the *HOXA* cluster is dependent on the lncRNA *HOTAIRM1*, as it facilitates unfolding of the cluster's 3' end and interacts with histone-modifying complexes UTX/MLL and PRC2 [4]. Indeed, depleting HOTAIRM1 during HOXA gene activation prevented opening of a long-range chromatin loop at the gene cluster. Recent publications have shown that the *HOXA-HOTAIRM1* regulatory axis is relevant to human disease; increased levels of *HOTAIRM1* may be useful as a cancer marker and possible therapeutic target in prostate and liver cancers for example [5, 6]. Based on the incidence of *HOTAIRM1*'s involvement with gene dysregulation in cancers and other diseases, its recognized interaction with histone-modifying complexes, and documented role in shaping chromatin, I hypothesize that *HOTAIRM1* is required for chromatin conformational changes instrumental in cell differentiation, not only at the *HOXA* cluster but also at other locations throughout the genome. These findings may have implications for human disease research, as chromatin conformation is tied to gene expression.

To identify *HOTAIRM1*-dependent chromatin conformation changes genome-wide, I suggest using CRISPR-Cas9 to insert a polyadenylation signal downstream of *HOTAIRM1*'s transcription start site to prematurely terminate its transcription in the NCCIT cellular differentiation model. In this approach, the lentiCRISPRv2 plasmid is used as it encodes Cas9, the gRNA scaffold, as well as antibacterial resistance for selection in bacterial and human cells [7]. The sgRNA sequence and repair template were designed for Cas9 to cut Chromosome 7 slightly downstream of both documented transcriptional start sites of *HOTAIRM1* and knock-in a polyA signal through homologous repair [8, 9]. I successfully cloned and sequenced the plasmid construct and designed a repair template; the edited plasmid and repair template were transfected into NCCIT cells.

The second part of the project aims to generate Hi-C DNA libraries from undifferentiated and all-trans retinoic acid (RA)-treated polyA knock-in NCCIT cells. As a first step towards implementing an existing Hi-C library protocol, I cultured and used NT2-D1 cells, which represent a similar differentiation system to NCCIT. Both NT2-D1 and NCCIT are pluripotent embryonal carcinoma cell lines wherein differentiation can be induced by RA; upon induction, NT2-D1 follow a neuronal differentiation path [10] while NCCIT can differentiate into ectoderm, mesoderm, and endoderm derivatives [11]. Several steps of the protocol from a previous laboratory standard were updated, ensuring that enzymes function at their peak performance. Quality control steps were optimized for higher confidence, while the adapter ligation steps require further optimization. When NCCIT clonal lines are successfully generated, Hi-C libraries from undifferentiated and RA-treated polyA knock-in fixed NCCIT cells as well as control libraries using undifferentiated and RA-treated NCCIT cells can be produced.

The third part of the project aims to use computational analysis tools to analyze Hi-C sequencing data. I explored several Hi-C sequencing data processing pipelines in pursuit of this goal. I successfully used the Juicer pipeline to generate contact matrices using available data [12]. I also used HiCUP to align Hi-C sequencing data and produce reports containing information on experimental artifacts [13]. HiCUP produces a paired-end read BAM file; to create a contact matrix for visualization with Juicebox [14], I used another program named HIFI [15]. Based on my experimentation using these various programs, I propose a method for generating Juicebox-compatible contact matrices from Hi-C sequencing data, assigning A and B compartment types using RNA-seq expression values per bin (adapted from [16]), and comparing loci between two datasets.

VI. Literature review

i. Long non-coding RNAs and the HOXA cluster

Decades of rigorous characterization have led to a detailed understanding of the evolutionarily conserved *HOX* homeodomain transcription factor family (Reviewed in [17]). Their colinear expression across the embryonic anteroposterior axis was first described by Edward Lewis in 1978; generally, individual *HOX* genes are expressed in the 3' to 5' order in which they appear on each chromosome, which also links to the temporal axis of development [18, 19]. In other words, 3' *HOX* genes within a cluster will be expressed earliest in development in the somites, and 5' genes are expressed later towards the posterior of the embryo (Reviewed in [20]). Most vertebrate genomes contain four *HOX* clusters: *HOXA*, *B*, *C*, and *D*.

Not all genes contained in *HOX* clusters encode proteins, however. Non-coding RNAs (ncRNAs) are plentiful in mammalian genomes and are classified based on a variety of factors including size and polarity. Long non-coding RNAs (lncRNAs) are >200nt in length, while small ncRNAs are derived from lengthier precursors (Reviewed in [21]). LncRNAs have a variety of functions; in the nucleus, they have been shown to regulate transcription by interacting with histone-modifying complexes and DNA methyltransferases (Reviewed in [22]). They can also regulate transcription by directly interacting with target DNA sequences and transcription factors or other proteins [22]. Furthermore, lncRNAs can act as post-transcriptional regulators by influencing splicing. In the cytoplasm, lncRNAs have been observed to "sponge" miRNA preventing the miRNA from regulating their targets through translation repression and/or degradation [22]. They can also be processed themselves into small RNA or mediate mRNA degradation or stability through direct binding or interacting with protein complexes. Further

proposed roles for lncRNAs include direct translational regulation and post-translational modification of peptide products [22].

Mammalian *HOX* clusters contain an abundance of antisense lncRNAs, some being intergenic and another portion overlapping with the exons of protein-coding genes (Reviewed in [23]). For example, *HOTTIP* is proposed to be involved in controlling the expression of the 5' end of the *HOXA* cluster during differentiation and is involved in the pathogenesis of many types of cancers [23]. Other notable lncRNAs include *HOTAIR*, which is shown to affect chromatin conformation in cancers, and *HOXA-AS2*, which is proposed to stimulate the proliferation and migration of cancer cells [23]. The lincRNA (long intergenic non-coding RNA) *HOTAIRM1* (HOX antisense intergenic RNA myeloid 1), situated between the *HOXA1* and *HOXA2* genes, was first described in 2009 as having a role in myelopoiesis through regulation of *HOXA* genes, particularly the 3' end of the cluster. The gene shares the CpG island which indicates the transcriptional initiation site of *HOXA1*, and its expression can be induced with endogenous retinoic acid (RA) [24].

Further work in NB4 acute promyelocytic leukemia cells showed that *HOTAIRM1* regulates members of the integrin family to influence cell growth and maturation [25]. While initial studies indicated that the expression of *HOTAIRM1* was highly specific to myeloid lineages, more recent studies have shown its role as a regulatory RNA in other cell types [26]. For example, in human ventral spinal cord motoneurons derived from iPSCs, nuclear *HOTAIRM1* modulates *NEUROGENIN 2* (*NEUROG2*) expression in *trans* through the recruitment of Polycomb Repressive Complex 2 (PRC2) to the *NEUROG2* promoter [26]. PRC2 forms part of the Polycomb (PcG) group proteins, composed of evolutionarily conserved epigenetic modifying complexes (Reviewed in [27]). PRC2 contains core subunits SUZ12, EED,

and one of the two methyltransferases EZH2 or EZH1; the complex catalyzes the mono-, di-, and trimethylation of lysine 27 found on histone H3 (H3K27). It can assist in maintaining the compact state of inactive chromatin as compact chromatin is kinetically ideal for PRC2 activity [28]. Thus, the interaction between *HOTAIRM1* and PRC2 implicates chromatin conformation as a likely mechanism for *HOTAIRM1*'s regulatory functions during neurogenesis.

ii. HOTAIRM1 in human disease

Naturally, gene expression dysregulation is of interest for treating medical conditions. In fact, abnormal expression of *HOTAIRM1* and/or *HOXA1* is associated with many disease states. In comparing samples from normal brain tissue with glioma and glioblastoma multiforme tissues, it was observed that higher expression of *HOTAIRM1* correlated with higher grade (WHO I-IV) tumors, and knockdown of *HOTAIRM1* reduced GBM cell proliferation in vitro and in a mouse model [29]. These findings were further validated by another study, which then used real-time quantitative polymerase chain reaction (RT-qPCR) to investigate possible mechanisms [30]. This study found that miR-153-5p, which directly targets the transcription factor SNAI2, was significantly increased after knockdown of *HOTAIRM1* in GBM cells. SNAI2 transcriptionally regulates *HOTAIRM1* in GBM cells in addition to repressing CDH1, a negative regulator of *HOTAIRM1*. SNAI2 is associated with migration in certain cell types, aligning with the increased migration of GBM cells observed when *HOTAIRM1* is highly expressed [31].

It was also found by Li et al. that *HOTAIRM1* regulates *HOXA1* in *cis* in GBM cells, as silencing *HOTAIRM1* followed by RT-qPCR showed that the absence of *HOTAIRM1* reduces the amount of *HOXA1* mRNA. Chromatin immunoprecipitation (ChIP) experiments showed that

HOXA1 interacts with the PRC2 subunit EZH2, G9a, which monomethylates and dimethylates H3K9 leading to the repressive histone mark H3K9me2, and DNA methyltransferase which can silence genes by methylating CpG islands [29]. In a study specifically on isocitrate dehydrogenase (IDH)-wild type glioblastoma cells which cites the three prior papers [29, 30, 31], it was observed that higher *HOTAIRM1* expression correlated with poor survival in GBM patients, and *HOTAIRM1* is the only lncRNA on chromosome 7 which is upregulated despite the frequency of increased copy numbers of that chromosome [32]. This study found that knocking down *HOTAIRM1* led to increased sensitivity to radiation therapy in vitro and in a mouse model. The proposed mechanism, evidenced by mitochondrial dysfunction upon *HOTAIRM1* knockdown and miRNA overexpression experiments, is that *HOTAIRM1* sponges the miRNA *miR-17-5p* which typically targets the transcript of Transglutaminase 2 (TGM2), which is implicated in mitochondrial function [32]. Taken together, these studies in GBM show that *HOTAIRM1* likely has a variety of functions in brain tumor cells, making it a promising target for treatment.

HOTAIRM1 upregulation does not always correlate with tumor aggressiveness, however; a study in liver cancer cells found that the DNA methyltransferase RIZ1 was upregulated in liver hepatocellular carcinoma (LIHC) tissue compared to other cancers and normal liver samples [33]. RIZ1 methylates H3K9 and is considered a tumor suppressor in many cancers, contrary to findings in liver cancer (Reviewed in [34]). On the other hand, *HOTAIRM1* RNA was significantly depleted in LIHC samples compared to non-cancer samples. ChIP and RT-qPCR experiments showed that H3K9me1 was enriched at the *HOTAIRM1* promoter region in HEPG2 and HCC-LM3 cells, a phenotype which could be reversed by knocking down RIZ1. Further experiments showed that *HOTAIRM1* RNA directly binds to the miRNA miR-125b, thus when

HOTAIRM1 expression is suppressed the miRNA is upregulated, leading to increased liver cancer cell growth and proliferation.

A recent study in clear cell renal cell carcinoma (ccRCC) cells also indicates that downregulation of *HOTAIRM1* is a marker of malignancy [35]. It was observed that in particular, the splice variant HM1-3, which contains exons 1 and 3 of the *HOTAIRM1* gene, is downregulated in renal cell carcinomas (RCCs), along with breast and colorectal cancers, when compared to normal tissues; the HM1-3 isoform was found to be the major isoform in non-cancerous renal tissue. In a normal renal cell line, CAKI-1, the study found evidence that HM1-3 regulates the hypoxia pathway through performing RNA sequencing (RNA-seq) on control and HM1-3 knockdown cells. The differential expression of hypoxia pathway genes, particularly DDAH1 and ANGPTL4, was confirmed by a HM1-3 overexpression rescue experiment. While the implications of HM1-3 downregulation in ccRCC cells must be explored further, the study found novel evidence that HM1-3 regulates HIF1 signaling, which is an oncogenic pathway in ccRCC tumors.

HOTAIRM1 was also suggested to be a tumor suppressor in colorectal cancer (CRC) as a 2016 study noted that its expression in CRC tissues and cancer patients' peripheral blood plasma was lower compared to non-cancerous colon and rectal mucosa and circulating plasma levels [36]. This study found through a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway and gene ontology analysis of the gene expression profile on control and siRNA *HOTAIRM1* knockdown HT29 cells that *HOTAIRM1* knockdown affects genes implicated in cell proliferation, particularly focal adhesion and ECM-receptor interaction pathways. Fourteen differentially expressed genes (DEGs) were further validated by RT-qPCR. ROC (Receiver Operating Characteristic) curve analysis showed that *HOTAIRM1* level as a circulating

biomarker has a comparable sensitivity and specificity as CEA, a biomarker used in the diagnosis and observation of CRC.

HOXA1 overexpression was observed to correlate with poor prognosis in breast cancer patients [37]. As *HOTAIRM1* was shown to regulate HOX1 in GBM [29], a study was conducted on estrogen receptor-positive (ER+) breast cancer cells to test whether *HOTAIRM1* regulates *HOXA1* in breast cancer as well [38]. ER+ breast cancer patients are frequently given selective estrogen-receptor modulators to treat their illness. Tamoxifen, one such drug, may be effective at first with tumors becoming resistant over time (Reviewed in [39]). Thus, biomarkers for tamoxifen resistance (TAMR) can help medical professionals make decisions about treatment plans. Kim et al. found that *HOTAIRM1* was upregulated in TAMR cells compared to ER+ breast cancer cells and knocking down *HOTAIRM1* re-sensitized TAMR cells to Tamoxifen, indicating the lncRNA as a possible biomarker [38]. The mechanism proposed in this paper involves *HOTAIRM1* interacting with PRC2 subunit EZH2, aligning with other evidence of interaction between the lncRNA and the protein. ChIP experiments showed that the repressive mark H3K27me3 was diminished in TAMR cells compared to MCF7 cells at the putative *HOXA1* promoter region, which overlaps with *HOTAIRM1*. Further ChIP experiments combined with *HOTAIRM1* knockdown showed that depleting *HOTAIRM1* in TAMR cells leads to an increase in EZH2 binding and H3K27me3 marks at the same locus. The direct interaction between *HOTAIRM1* and EZH2 was confirmed using RNA immunoprecipitation (RIP), strengthening the evidence for *HOTAIRM1* upregulating *HOXA1* by inhibiting EZH2's ability to methylate histone H3K27 through competitive binding. These results strongly implicate *HOTAIRM1*'s effect on chromatin landscape as a factor in Tamoxifen resistance.

iii. Chromatin architecture and methods of capturing it

Within mammalian nuclei, chromatin occupies three-dimensional (3D) space in a structural hierarchy. The highest-order structures in nuclear architecture are chromosome territories, wherein chromosomes fold into forms along their length in a manner conducive to transcription, replication, splicing and DNA repair (Reviewed in [40]). Smaller-scale structures termed chromatin compartments are found within territories; compartments are defined as long-range physical interactions where the chromatin has consistent compactness and epigenetic markers denoting either open or closed chromatin. Compartments are typically denoted as falling into either groups A (open, transcriptionally active) or B (closed, transcriptionally inactive), though they can be further classified within these categories based on histone modification patterns [41]. Within compartments are Topologically Associating Domains (TADs), defined as consecutive regions which greatly interact within themselves and are bordered by short sections with minimal interactions [42]. These bordering regions, termed TAD boundaries, are enriched for insulator protein CTCF occupation, factors associated with active promoters, SINE elements, and housekeeping genes relative to the rest of the genome. TAD boundaries can be conserved across cell types and are relatively consistent between mouse and human genomes at syntenic regions; altogether, these observations implicate TADs as crucial for transcriptional management. Chromatin loops, the next order of organization, frequently occur between boundary regions and often bring promoters and their associated enhancer elements together [41]. Loops commonly contain CTCF binding sites, the majority of which are convergent; that is to say, when the two loop anchors are brought together, the CTCF binding motifs most often face in the same direction.

In addition to gene expression regulation, chromatin conformational states can affect other important cellular processes such as replication and DNA repair. Thus, chromatin conformation capture (3C) was developed to address limitations in observation of chromatin conformation through microscopy-based methods [43]. In short, intact nuclei are treated with formaldehyde to crosslink proteins to DNA and other proteins. Crosslinked DNA is digested with a restriction enzyme, then restriction fragments are ligated together; though random ligation products are possible, cross-linked DNA fragments are more likely to be ligated together in dilute conditions. After ligation, crosslinking is reversed, and specific products can be amplified by polymerase chain reaction (PCR). While useful for studies focused on a small number of loci, 3C is not fit for analyzing physical contacts in large regions of chromatin.

Circular chromosome conformation capture (4C) and 3C-Carbon Copy (5C) both seek to address this limitation. 4C takes advantage of the circular nature of 3C crosslinking ligation products; after crosslinking is reversed, primers targeting the locus of interest proximal to the ligation junction will amplify the sequences the locus of interest interacts with [44]. 5C libraries are generated by performing multiplexed ligation-mediated amplification on a 3C library. 5C primers anneal to DNA across a ligation junction in a multiplex fashion and are ligated together; the full library can be amplified by PCR using universal primers as all forward 5C primers contain the T7 promoter sequence, and all reverse 5C primers contain the complementary T3 sequence [45]. The products can then be analyzed on dedicated microarrays or by deep sequencing, allowing for higher-throughput analysis compared to 3C.

4C and 5C still require target loci, which is not suitable for whole-genome studies. Thus, Hi-C was introduced to address this constraint [46]. In Hi-C, DNA is crosslinked and digested with a restriction enzyme as it would be in 3C; however, before ligation, sticky ends left by the

restriction digest are repaired using a mixture which includes a biotinylated nucleotide. Thus, contact-proximity ligation products contain biotin which can be pulled down on streptavidin beads after purification, shearing, and size selection. After the biotin-streptavidin pulldown step, sequencing adapters are ligated onto Hi-C fragments while stabilized on the beads. PCR primers designed to anneal to the adapter sequences are then used to amplify the Hi-C library, and the library is sent for sequencing. A graphic comparison of 3C, 4C, 5C, and Hi-C is shown in Figure 1. While systematic biases are still introduced throughout the Hi-C protocol, sequencing data can be filtered and normalized to offset the varying frequencies at which fragments appear based on their GC content, distance between restriction sites, and other factors [47]. Hi-C therefore provides the ability to study the physical contacts between sites in the chromatin genome-wide in a relatively unbiased manner. In order to further reduce bias, the standard protocol in recent years performs the proximity-contact ligation step in intact nuclei; research comparing this to dilute in-solution ligation shows that in-nuclear (also called in-situ) ligation reduces background noise, reproducibility, reduced fragment length bias, and sharper features [48]. In conjunction with expression data and epigenetic marker data, in-situ Hi-C can contribute to the understanding of the relationship between chromatin conformation and gene expression.

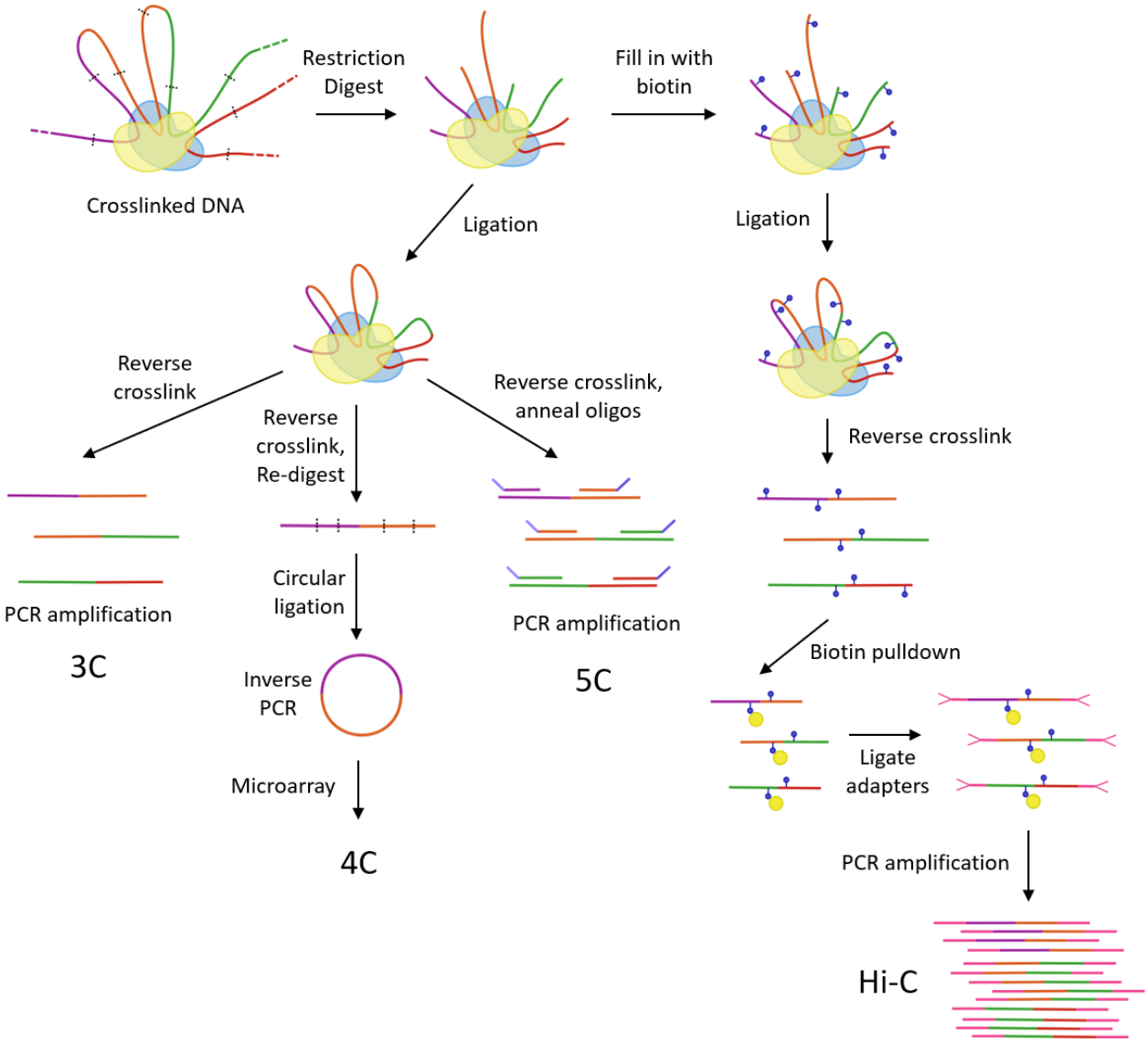


Figure 1. Overview of select Chromatin Conformation Capture methodologies. Adapted from Cao et al., 2015 [49].

iv. Processing Hi-C sequencing data

Hi-C sequencing data is returned as raw paired-end reads in the fastq format, a large text file. In order to perform downstream analysis, these reads must be aligned to the genome and filtered for quality; contact matrices can also be generated for further evaluation and visualization as a "heatmap" of contact frequencies along the genome. When Hi-C was first introduced, interaction matrices were computed on a 1 Mb scale, where each 1 Mb by 1 Mb "bin" contains the number of interactions found within that bin [46]. This method and its successors are not designed to efficiently analyze data containing Terabase-scale numbers of reads in Hi-C libraries [50-52]. Thus, the Aiden lab developed Juicer, a pipeline which maps Hi-C reads to the genome using BWA [53], calculates a contact matrix, optionally normalizes the matrix, and can be used to annotate the matrix with TAD boundary calls and loop features [12]. Normalization can be applied using the contact probability for loci at varying genomic distances [46] or using a matrix balancing algorithm [54]. Importantly, the contact matrices computed using Juicer are stored at 18 total resolutions in a compressed format: the *hic* file type. Both the normalized and non-normalized matrices are stored in the same file, the size of which is approximately 80 gigabytes per Terabyte of raw sequencing data. These files can be viewed along with annotation tracks in the accompanying program Juicebox, also from the Aiden lab [14].

Different file formats for interaction matrix data storage have been introduced, including other binary formats such as MRH and *butlr* [52, 55]. Hi-C interaction matrices can also be stored in an HDF5 container [56], for example using the *cooler* format, which comprises a sparse array allowing for fine scaling [57] Of the wide variety of options, Juicer and Juicebox were selected for this study as they are used by collaborators of the laboratory and Juicer is compatible

with the Slurm Workload Manager used on Compute Canada clusters maintained by the Digital Research Alliance of Canada.

v. HOTAIRM1 and chromatin architecture

Previous work in our laboratory has shown that *HOTAIRM1* has differences in function between NB4 and NT2-D1 cell differentiation models, indicating tissue-specific regulatory patterns [4]. As in NB4 cells, *HOTAIRM1* expression is induced upon RA treatment in NT2-D1 cells, however, different splice variants are present; while in NB4 cells a ~500 nucleotide (nt) variant was reported, a ~4kb unspliced variant and ~1.1 kb spliced variant are found in NT2-D1 [4]. The unspliced variant preferentially localizes to the nucleus, suggesting a specific role of the unspliced form in gene regulation at the transcriptional level. In NB4 cells, depletion of *HOTAIRM1* leads to decreased expression of *HOXA1*, *A4*, and *A5*, while in NT2-D1, shRNA knockdown of *HOTAIRM1* led to decreased expression of *HOXA1* and *A2* and increased expression of *A4* and *A5* [4].

These observations introduce the question of the mechanistic difference between *HOTAIRM1*'s function in pluripotent NT2-D1 cells compared to NB4 cells. Prior research in our laboratory into the three-dimensional structure of the *HOXA* locus using 3C methodology showed that the cluster has a tight looped structure when it is transcriptionally silent, while when it is transcriptionally active the contacts between the 3' and 5' ends slacken [3]. Thus, to investigate a possible mechanism, 5C-seq was employed to test whether 3D chromatin organization is affected by *HOTAIRM1* in NT2-D1. The results showed that *HOXA1*, *HOTAIRM1* and *HOXA2* comprise a subTAD, with *HOXA3* and *A4* in another subTAD beside it. *HOXA5*, *A6* and *A7* encompass an additional subTAD, which has a high degree of interaction

with the *HOXA1/2* subTAD prior to differentiation [3]. After RA treatment, this interaction frequency decreases. Depleting *HOTAIRM1* transcript with shRNA resulted in stronger contacts between the *HOXA1/2* and *HOXA5/6/7* subTADs. Repeating similar experiments with NB4 cells showed that these contacts are infrequent even prior to differentiation, suggesting that *HOTAIRM1* is necessary for chromatin conformational changes in NT2-D1 but not NB4 [3]. Chromatin immunoprecipitation sequencing (ChIP-seq) and RNA immunoprecipitation (RIP) experiments showed that in NT2-D1, *HOTAIRM1* binds to both the repressive histone modifying complex PRC2 and the activating histone-modifying complex UTX/MLL; PRC2 prefers the spliced variant, while UTX binds to the unspliced variant [3]. Taken together, these data suggest a mechanism by which *HOTAIRM1* coordinates the opening of the looped structure observed in a transcriptionally silent *HOXA* cluster upon induction with RA.

In order to further our understanding of *HOTAIRM1*'s regulatory mechanism as it pertains to chromatin conformation both locally within the *HOXA* cluster and possibly in other locations across the genome, a cell model was designed wherein CRISPR-Cas9-facilitated homologous recombination will be used to knock-in a poly-adenylation (polyA) signal slightly downstream of documented *HOTAIRM1* transcription start sites. This model is based on work in Wendy Bickmore's laboratory, where a similar method was used to study the lncRNA *Hottip*, found in the 5' end of the *HOXA* cluster, in mouse cells [8]. The polyA knock-in model is appealing as, compared with RNAi-mediated knock-down, any regulatory effect from the act of transcription itself is minimized. The polyA signal sequence in the *Hottip* model is synthetic and is preferred over the human α2 globin polyA site, indicating higher efficiency [9]. Thus, though its initial usage was in mouse, recent research has used the same polyA sequence in knock-in experiments in human cells; for example, in one study wherein the lncRNA *C1QTNF1-AS1* was

observed [58]. Following these examples, stable NCCIT cell lines will be generated where *HOTAIRM1* contains a polyA signal early in its sequence. NCCIT cells are, like NT2-D1, pluripotent and derived from embryonal carcinoma cells and their differentiation can be induced by RA [57, 58]. Thus, studies in these knock-in cells will be comparable to previous work in the laboratory. In-situ Hi-C and subsequent analysis will be used to observe chromatin conformation across the genome in control (RA-treated and untreated) cells and polyA knock-in (RA-treated and untreated) cells.

VII. Methods

Equipment

An Eppendorf Centrifuge 5424 was used to centrifuge all solutions in 1.7mL Eppendorf tubes. An Eppendorf Centrifuge 5810 R (15 amp version) was used to centrifuge all solutions in tubes larger than 1.7mL unless otherwise specified. A Covaris M220 was used for sonication. Liquid bacterial cultures were grown in a Thermo Scientific MaxQ 4450. A Bio-Rad MyCycler thermal cycler was used for PCR and all other processes requiring a thermocycler. pH was measured using a Sartorius Docu-pH meter.

Cell culture

Dulbecco's Modified Eagle Medium (DMEM) [Gibco, #11965-092] is supplemented with 10% Fetal Bovine Serum (FBS) [Gibco, #12483-020] and warmed to 37 °C in a water bath prior to starting cell work. If needed, Phosphate-buffered saline (PBS) [Gibco, #70011044] and Trypsin [Multicell, #325-045-EL] are also warmed to 37 °C and RT (room temperature), respectively. To plate fresh cells, cells were thawed from their cryogenically frozen state by gently warming them in a 37 °C water bath. Once thawed, 1 mL cells were added to a centrifuge tube containing 9 mL media and centrifuged at 100 xg for 5 minutes at RT to remove debris. Supernatant was removed and cells were resuspended in 10 mL fresh media, then the centrifuge step was repeated. After removing the supernatant, cells were resuspended in the correct volume of media for the given plate. Cells are incubated at 37 °C until they reach approximately 80-90% confluency. If cells were to be passaged, old media was removed, and cells were washed with 5mL of PBS per 10 cm plate. PBS was removed and 1 mL Trypsin solution was added to each plate, which were gently rotated such that the surface is covered. The plate(s) was placed in the 37 °C incubator for

2 minutes to allow the Trypsin to activate. 9 mL fresh media is added to each plate to bring the total volume to 10 mL. The solution was gently pipetted up and down to mix out any clumps. 1 mL of this solution was added to a fresh plate, then the volume was brought to 10 mL with fresh media. Newly passaged plates were incubated at 37 °C. This protocol is followed for both NT2-D1 cells and NCCIT cells.

Cell fixing ***(adapted from David Wang)***

1% formaldehyde [Sigma, #F8775-500ML] complete media is prepared prior to starting cell work. One plate of 80~90% confluent cells was set aside for cell counting. Media was removed from the 10 cm plate(s) with an aspiration pipette. Cells were washed once with 5 mL PBS. 5 mL 1% formaldehyde media was added to each plate. Plates were gently rocked every 2 minutes for a total of 10 minutes at room temperature (RT). 263 uL 2.5 M sterile-filtered Glycine stock was added to each plate to yield a final glycine concentration of 0.125 M to stop the cross-linking reaction. Plates were incubated for 5 minutes at room temperature, then a minimum of 15 minutes on ice. Fixed cells were scraped off the plate using a rubber scraper and the cell count estimation was used to distribute cells into 15 mL conical tubes at approximately 1e7 to 2e7 cells per tube. The tubes were centrifuged at 2,000 RPM at 4 °C for 5 minutes. The supernatant was removed and pellets were washed once with 8 mL cold PBS. Cells were centrifuged at 2,000 RPM at 4 °C for 10 minutes. The supernatant was removed and cells were snap frozen with dry ice, then stored at -80 °C. This protocol is followed for both NT2-D1 cells and NCCIT cells.

Preparation of competent bacterial cells *(adapted from Siedman et al., 2001 [60])*

All open-air steps are performed in the vicinity of an open flame. A frozen source of bacteria was streaked on an LB-agar [Multicell, #800-010-GC] plate containing no antibiotics then incubated overnight. A 50 mL liquid LB (1% w/v Casein Digest Peptone [Multicell, #800-155-LG], 0.5% w/v Yeast Extract [Fisher, #BP9727-500], 0.5% w/v NaCl [BDH, #ACS 783]) culture was prepared using a discreet colony of the bacteria, then grown at 37 °C in a shaking incubator overnight. CaCl2 solution (60 mM CaCl2 [BioShop, #CCL302.500], 15% glycerol [BioShop, #GLY001.1], 10 mM PIPES (pH 7.0) [OmniPur, #6910]) was prepared and sterile-filtered using a SteriCup Quick Release vacuum filtration system [Millipore, #S2HVU02RE] following manufacturer's instructions. The solution was chilled at 4 °C overnight. The next morning, 4 mL of the 50 mL bacterial culture was added to a 2 L flask containing 396 mL LB medium. The culture was grown at 37 °C with gentle shaking. After 1 hour of growth, 1 mL aliquots were taken every 20-30 minutes to measure the OD590 using 1 cm cuvettes and a Spectrophotometer. LB medium was used to blank the machine. Once the cell culture reached an OD590 of between 0.375 and 0.399, the culture was evenly divided into pre-chilled 250 mL sterile centrifuge tubes. The tubes rested on ice for 10 minutes, then were centrifuged using a GSA rotor for 7 minutes at 1,600 xg at 4 °C. The supernatant was decanted, then each pellet was gently resuspended in ¼ volume cold CaCl2 solution. The cells were centrifuged at 1,100 xg for 8 minutes at 4 °C. The supernatant was removed by pipet, then the pellets were gently resuspended in ≤1/25 volume $CaCl_2$ solution. Using a plastic pipet, competent cells were dispensed into pre-chilled 1.7 mL Eppendorf tubes in 100-500 uL aliquots. Competent cells were flash frozen in liquid nitrogen then stored at -80 °C.

Bacterial transformation and plasmid constructs

100 ng unedited lentiCRISPRv2, a gift from Feng Zhang (Addgene plasmid #52961; http://n2t.net/addgene:52961; RRID: Addgene_52961), was transformed into competent Stbl3 cells. Competent cells were removed from -80 °C storage and thawed on ice for 20 minutes, then 50 uL competent cells were added to a plastic tube containing 100 ng plasmid. The mixture was incubated on ice for 30 minutes, then the cells were heat shocked at 42 °C for 45 seconds. After resting on ice for 2 minutes, 5x volume of antibiotic-free SOC liquid media (2% (w/v) Tryptone [Fisher, #BP9726-500], 0.5% (w/v) Yeast extract [Fisher, #BP9727-500], 10 mM NaCl [BDH, #ACS 783], 2.5 mM KCl [BioShop, #POC306.500] 10 mM MgCl2 [BioShop, #MAG510.500], 20 mM Glucose [BioShop, #GLU501.500]) was added to the tubes. The cells were allowed to grow in a 37 °C incubator for 45 minutes with no shaking. 100 uL of the transformed cell mixture was plated on LB-agar plates containing ampicillin. Discrete colonies were selected for 4 mL LB-ampicillin culture. The cells were expanded to 100 mL cultures and a midi-prep kit [Invitrogen, #K210005] was used to purify the plasmid. The concentration of purified plasmid was estimated using a Nanodrop spectrophotometer, and the solution was diluted to approximately 1 ug/uL. Approximately 8 ug purified plasmid backbone was digested with 2 uL 10 U/uL BsmBIv2 restriction enzyme [NEB #R0739S] in a 50 uL reaction containing 5 uL buffer 3.1 [NEB, #B7203S]. The restriction digest was incubated at 37 °C for 3 hours, then the full digest reaction was run on an 0.8% agarose gel. The ~13 kb band was excised from the gel and extracted using a gel extraction kit [Qiagen, #28704]. Dry sgRNA oligos were resuspended to 80 uM in TE buffer. In a low-bind PCR tube, 1.2 uL of each resuspended forward oligo, 1 uL 10X T4 DNA ligase buffer [NEB, #B0202S], and 6.6 uL nuclease-free ddH2O were mixed

together. The tube was placed in a thermocycler and the following program was used to anneal the oligos:

30 min 37 °C
5 min 95 °C
5 min 90 °C
5 s 80 °C (ramp 0.1 °C/s)
30 s 70 °C (ramp 0.1 °C/s)
30 s 60 °C (ramp 0.1 °C/s)
1 min 50 °C (ramp 0.1°C/s)
2 min 40 °C (ramp 0.1°C/s)
2 min 30 °C (ramp 0.1°C/s)

The annealed oligos were diluted 1:200 in EB. In a fresh low-bind PCR tube, 25 ng digested plasmid, 1 uL dilute oligo duplex solution, 1 uL 10X T4 DNA ligase buffer [NEB, #B0202S], 1 uL T4 DNA ligase [NEB, #M0202S], and enough ddH2O to bring the reaction to 10 uL were mixed together. A control reaction was prepared with all of the same reagents except the oligo duplex solution. The reactions were incubated at 16 °C overnight in the thermocycler, then T4 DNA ligase was heat-inactivated at 65 °C for 10 minutes. Competent Stbl3 bacterial cells were transformed with the edited plasmid following an identical protocol to the unedited plasmid transformation. Six discrete colonies were selected at a time for 4 mL liquid LB-ampicillin culture. Glycerol stocks were made from each clone by adding 500 uL liquid culture to sterile Eppendorf tubes containing 500 uL sterile 50% glycerol, then gently mixing and storing at -80 °C. The remaining volume of each culture was used for plasmid purification using a homebrew mini-prep procedure. Per each clone, two tubes containing 1.5 mL bacterial culture were spun down at 3,000 xg for 1 minute at room temperature. Media was removed and 200 uL Solution P1 (50 mM Glucose, 25 mM Tris-HCl pH 7.5 [BioShop, #TRS002.5] 10 mM EDTA [BioShop, #EDT001.500]) was added to the pellets. Pellets were then resuspended and allowed to rest for 5 minutes. 200 uL Solution P2 (1% SDS [BioShop, #SDS001.500], 0.2 M NaOH

[BioShop, #SHY777.500]) was added to each tube. Tubes were mixed by inverting until the solution was clear. 200 uL Solution P3 (3 M KOAc [BioShop, #POA303.500], 11.5% v Glacial Acetic Acid [BioShop, #ACE333.4]) was added to each tube, which were inverted to mix. The tubes were centrifuged at 13,000 xg at RT for 5 minutes. Supernatant was transferred to a new tube; if any debris was picked up in the transfer, the centrifugation was repeated. 600 uL Isopropanol [Fisher, #BP2618-4] was added to each tube, and tubes rested on ice for 15 minutes. The tubes were spun down at 13,000 xg for 10 minutes at 4 °C to pellet out DNA. The pellets were washed with 500 uL 70% EtOH then resuspended in 45 uL TE buffer. 5 uL 10 mg/mL RNAse A [Thermo, #EN0531] was added to each tube, then tubes were incubated at 37 °C for 30 minutes. 5 uL of each miniprep was run on an 0.8% agarose gel.

Sequence confirmation

To ensure that plasmid editing was successful, PCR primers were designed to span the ligation region of lentiCRISPRv2 such that a successful ligation would generate a fragment 152 bp long, while a self-ligation event would produce a fragment 132 bp or smaller in length. Four clones which returned the expected 152 bp fragment were sent for Sanger sequencing by Genome Québec.

In-situ Hi-C library generation *(Adapted from Rao et al., 2015 [39])*

i. Cell lysis and DNA digestion

Fixed cell pellets were resuspended in 1 mL Hi-C lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal [Sigma, #I8896-100ML], 40 uL PIC [Sigma, #P8340], stored at -20 °C) per 10 million cells. After resting 15 minutes on ice, cell lysate was spun down at 2,500 xg for 5

minutes at 4 °C. Intact nuclei were washed with 500 uL Hi-C lysis buffer and spun down at 2,500 xg for 5 minutes at 4 °C. Nuclei were resuspended in 200 uL 0.5% SDS with 1x 3.1 buffer and incubated for 10 minutes at 65 °C. 22 uL 10% Triton X-100 [MP Bio, #194854] was added to each tube after cooling to RT, then incubated at 37 °C for 30 minutes. 11 uL 3.1 buffer, 37 uL ddH2O, and 40 uL (400U) NcoI [NEB, #R0193S] were added to each tube, which incubated at 37 °C overnight. The next morning, NcoI was heat inactivated as each tube was incubated at 80 °C for 20 minutes.

ii. Biotin fill-in and end ligation

To fill in the restriction enzyme overhang and mark the ends with biotin, 52 uL fill-in master mix (4.5 uL mix of 10 mM dATP, 10 mM dGTP, and 10 mM dTTP [Invitrogen, #10297-018], 37.5 uL 0.4 mM biotin-14-dCTP [Invitrogen, #19518-018], 10 uL Klenow DNA Polymerase I Large Fragment (5 U/uL) [NEB, #M0210L]) was added to each tube. Tubes were incubated at 37 °C for 45 minutes. Klenow was heat-inactivated at 75 °C for 20 minutes. A 10 uL digested and filled-in sample was collected for quality checking (QC1). 833 uL ligation master mix (601 uL ddH2O, 120 uL 10x T4 DNA ligase buffer [NEB, #B0202], 12 uL 10 mg/mL BSA [NEB, #B9001S], 100 uL 10% Triton X-100) was added to each tube, then 5 uL T4 DNA ligase 400 U/uL [NEB, #M0202] was added. Contents were mixed by inverting then incubated at RT for 6 hours. Nuclei were spun down at 3,000 xg for 5 minutes at 4 °C to remove in-solution ligation product, then washed with 200 uL 10 mM Tris-HCl pH 8.0. Each pellet was resuspended in 400 uL 10 mM Tris-HCl pH 8.0, then 4.5 uL 10% SDS was added and mixed well. 50 uL Proteinase K 10 mg/mL [NEB, #P8107S] was added to each tube, which were incubated at 64 °C overnight. The volume of QC1 was brought to 400 uL, then SDS and Proteinase K were also added to its

tube. The next morning, 50 uL Proteinase K was added to each tube, which were then incubated at 64 °C for 2 hours.

iii. DNA purification

After cooling to RT, Phenol extraction was performed by adding 500 uL Phenol [Fisher, #BP1750I-400] to each tube, vortexing for 2 minutes, and spinning down at 3,500 rpm for 15 minutes at RT. The upper phase was transferred to new tubes, then Phenol extraction was repeated. The upper phase was transferred to new tubes, then Phenol:Chloroform extraction was performed by adding 500 uL of Phenol:Chloroform to each tube, vortexing for 2 minutes, and spinning down at 13,000 rpm for 10 minutes at RT. Supernatant was transferred to new tubes, then Chloroform extraction was performed by adding 500 uL Chloroform [Fisher, #BP1145-1] to each tube, vortexing for one minute, then spinning down at maximum speed for 5 minutes at RT. Supernatant was transferred to new tubes, then 0.1 volume (50 uL) 3 M NaOAc [BioShop, #SAA 304] pH 5.2 was added to each tube and mixed well. 2.5 volumes (1.1 mL) cold 100% EtOH [Commercial Alcohols, #P016EAAN] was added to each tube, which were then incubated at -80 °C overnight. The next morning, DNA was pelleted out by spinning down at 13,000 rpm for 25 minutes at 4 °C. Ethanol was removed, then five EtOH washes were performed by resuspending each pellet in 1 mL cold 100% EtOH and spinning down at 13,000 rpm for 25 minutes at 4 °C each time. Each pellet was dissolved in 25 uL TE and library tubes were pooled together, leaving QC1 separate. QC1 along with 0.2 uL, 0.4 uL, and 0.8 uL of the purified Hi-C library were run on a 0.8% Agarose D1-LE [Multicell, #800-015-CG] gel using 250 ng of the 1 kb DNA ladder [NEB, #N3232S].

iv. PCR titration of Hi-C DNA

To test ligation efficiency, a PCR titration was performed. 2 uL of the purified Hi-C library was diluted 1:4 in H2O, which was then used to make a 2-fold dilution series leaving 4 uL Hi-C DNA in each tube with 12 dilutions in total. 4 uL H2O was used as the control. 8.2 uL PCR master mix (Volume per tube: 2.5 uL 10x HIFI buffer (600 mM Tris-SO4 pH 8.9, 180 mM Ammonium Sulfate [BioShop, #AMP301.1]), 2 uL 50 mM MgSO4 [NEB, #B1003S], 25 mM dNTPs, 2 uL 5 uM primer 1, 1.5 uL 1:100 Salmon Testes DNA [Sigma, #D7656]) and 12.8 uL Taq master mix (Volume per tube: 10.6 uL ddH2O, 2 uL 5uM primer 2, 0.2 uL Taq DNA Polymerase [NEB, #M0273L]) were added to each low-bind PCR tube. PCR was performed using the following program:

1 cycle:	94 °C	5 minutes
35 cycles:	94 °C	30 seconds
	65 °C	30 seconds
	72 °C	30 seconds
1 cycle:	94 °C	1 minute
	65 °C	45 seconds
	72 °C	8 minutes

After completion, 12.5 uL of each reaction was run on 1.5% Agarose gel along with 250 ng of the pKS+HaeIII ladder [home-brew]. 12.5 uL each of the first five reactions of the PCR titration were pooled together. Using these aliquots of the PCR products, four digestion reactions as described in the table below were set up:

reaction	**A (ctrl)**	**B**	**C**	**D**
Buffer 3.1 10 X (NEB)	5 uL	5 uL	5 uL	5 uL
PCR Product	15 uL	15 uL	15 uL	15 uL
Water	30 uL	27.5 uL	27.5 uL	25 uL
NcoI 10U/µl	0	0	2.5 uL	2.5 uL
NsiI 10U/µl	0	2.5 uL	0	2.5 uL
vol in µl	**50**	**50**	**50**	**50**

The cut site of NsiI is inverse to the cut site of NcoI, thus PCR products derived from Hi-C DNA which underwent an efficient initial NcoI digestion, fill in, and ligation are expected to be cut with NsiI. The reactions were incubated at least 1 hour at 37 °C. 25 µl of each reaction were run on a 2% agarose gel (50 mL) using 250 ng of the 100 bp DNA Ladder at 100 Volts for 60 minutes.

iv. Fragment shearing and size selection

Hi-C DNA was sheared to approximately 300 bp in volumes of 130 uL at a time by sonication using the DNA_0300_bp_130uL_Snap_Cap_Micro_TUBE Covaris Focused-Ultrasonicator protocol. All aliquots were pooled together. 10 uL of sonicated library was run on a 1.5% agarose gel along with 250 ng of the 100 bp ladder [NEB, #N3231S]. Another 10 uL aliquot was saved for quality checking (QC2). Sonicated DNA was divided into 130 uL aliquots and size selected by adding 78 uL Ampure XP [Beckman Coulter, #A63881] beads to each tube. After gently mixing using a vortex for 30 minutes, the tubes were placed on a magnetic stand for two minutes. The supernatant was transferred to new tubes, to which 32.5 uL Ampure XP beads were added. After gently mixing using a vortex for 30 minutes, the tubes were placed on a magnetic stand for two minutes. The supernatant was discarded, and the beads were washed twice with 500 uL fresh 70% EtOH. Beads were air dried for no more than 5 minutes, then DNA was eluted with 60 uL of EB (10 mM Tris-HCl pH 8.5). The supernatant was pooled together. 20 uL size-selected DNA was saved for quality checking (QC3) and 5 uL was used to quantify using the Quant-iT kit [Invitrogen, #P11496] following manufacturer instructions.

v. Biotin pull-down

Biotin pull-down was performed by washing 60 uL of Dynabeads MyOne Streptavidin C1 beads twice with 400 uL of Tween Wash Buffer (TWB) (5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween [BioShop, #TWN510.500]) and once with 300 uL 2X Binding Buffer (2XBB) (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 2 M NaCl). For all washes, beads were incubated for 3 minutes at RT with rotation, unless otherwise specified, and placed on a magnetic stand for 2 minutes. After washing the beads, they were resuspended in 300 uL 2XBB. 300 uL of Hi-C DNA was added to the beads, then incubated at RT with rotation for 15 minutes. The supernatant was removed, and the beads were washed twice with 1X Binding Buffer (1XBB) (5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 M NaCl); for the first wash, the tube was heated to 55 °C without rotation for 3 minutes. The beads were washed once with 1X T4 Ligase Buffer [NEB, #B0202S].

vi. End repair, A-tailing, and adapter ligation

Supernatant was removed and 100 uL End Repair Mix (10 uL 10X T4 Ligase Buffer, 4 uL 10 mM dNTPs, 5 uL T4 DNA Polymerase (3 U/uL) [NEB, #M0203S], 5 uL T4 Polynucleotide Kinase (10 U/uL) [NEB, #M0201S], 1 uL Klenow (5 U/uL), 75 uL ddH2O) was added to the beads to repair DNA ends and remove any overhangs. The resuspended beads were transferred to a low-bind PCR tube and incubated at 20 °C for 30 minutes, then heat-inactivated at 75 °C for 20 minutes. The mixture was transferred to a 1.7 mL tube and placed on a magnetic stand. Supernatant was removed and the beads were washed twice with 200 uL of TWB, heating at 55 °C with no rotation for 2 minutes for the first wash. The beads were washed twice with 200 uL of EB. To ligate a single Adenosine onto the 5’ end of each fragment, 50 uL of Adenosine

Mix (5 uL 10X Buffer 2 [NEB, #B7002S], 1 uL 10 mM dATP, 3 uL Klenow (3'to 5' exo$^-$) [NEB, #M0212S], 41 uL ddH2O) was added to the beads. Beads were transferred to a low-bind PCR tube and incubated at 37 °C in a thermocycler for 60 minutes. The beads were transferred to a 1.7 mL Eppendorf tube and placed on the magnetic stand for 2 minutes before removing the supernatant. The beads were washed twice with 200 uL TWB; the tube was heated to 55 °C for two minutes for the first wash. The beads were then washed with 200 uL EB twice. To ligate adapters onto each fragment, 45 uL Adapters Ligation Mix (5 uL 10X T4 DNA Ligase Buffer, ddH2O, Illumina adapter oligonucleotides) was added to the beads. The amount of Illumina adapter oligos (Index #2 and Universal; see Supplementary Materials for sequences) necessary for each reaction is calculated based on the amount of DNA present based on the Quant-iT kit analysis of QC3; for every 1 ug of Hi-C DNA, 6 pmol of adapter oligos were added to the reaction. Adapter oligos are annealed immediately prior to ligation by adding 1.5 uL each of the Index#2 and Universal oligos to a low-bind PCR tube along with 1.5 uL 10X T4 ligase buffer and 10.5 uL ddH2O. The adapters are annealed in a thermocycler using the following program:

1 cycle:	37 °C	2 minutes
	95 °C	5 minutes
	90 °C	1 minute
	85 °C	1 minute
	80 °C	1 minute
	75 °C	1 minute
	70 °C	1 minute
	65 °C	1 minute
	60 °C	1 minute
1 cycle:	55 °C	1 minute
	50 °C	1 minute
	45 °C	1 minute
	40 °C	1 minute
	35 °C	1 minute
	30 °C	1 minute
	25 °C	1 minute
1 cycle:	4 °C	hold

After Adapter Ligation Mix has been added to the Streptavidin beads, the mixture is transferred to a low-bind PCR tube and 5 uL T4 DNA Ligase (1 Weiss U/uL) [Invitrogen, #15224017] is added to the tube. The mixture is incubated at 16 °C for 60 minutes in a thermocycler, then the solution is transferred to a 1.7 mL Eppendorf tube and placed on the magnetic stand for 2 minutes. The beads are washed twice for 6 minutes with 400 uL TWB; the solution is heated to 55 °C for the first wash. The beads are washed with 400 uL EB twice for 6 minutes each. The beads are resuspended in 50 uL EB.

vii. Diagnostic and large-scale PCR

To test the efficiency of adapter ligation, a PCR diagnostic is performed. Three sets of PCR reactions are run simultaneously; each experimental low-bind PCR tube contains 1 uL Hi-C DNA on beads, 1.25 uL each 10 uM Hi-C PCR forward and reverse primers, 16.1 uL ddH2O, and 5.4 uL Phusion Master Mix (per reaction: 5 uL 5X Phusion buffer [Thermo, #F-518], 0.2 uL 25mM dNTPs, 0.2 uL Phusion DNA Polymerase [Thermo, #F-530S]. Each control reaction contains 17.1 uL water, 1.25 uL each 10 uM Hi-C PCR forward and reverse primers, and 5.4 uL Phusion Master Mix. Each set of reactions is run on a different thermocycler using the following program:

1 cycle:	98C	30 seconds
10, 8, or 6 cycles:	98C	10 seconds
	65C	30 seconds
	72C	30 seconds
1 cycle:	72C	7 minutes

15 uL of each product mixed with 3 uL 6X Blue loading dye is run on a 2.0% Agarose gel (TBE) for 60 minutes at 100 V along with 250 ng of 100 bp DNA ladder. A large-scale Phusion PCR

reaction can then be performed to amplify the library for sequencing using the minimum number of cycles found in the testing step.

End Repair Testing

In order to test End Repair efficiency, test reactions were performed using a purified PCR product. The method for End Repair, A-tailing, and ligating adapters to the PCR product was as consistent as possible with the Hi-C methodology, however, instead of using buffers to wash DNA bound to Streptavidin beads, a PCR purification kit [Qiagen, #28106] was used to purify the fragment after each enzymatic reaction. Several iterations were performed using different enzymes for the PCR reaction; Taq DNA Polymerase was used initially, as the PCR reaction setup was identical to the PCR titration step in the Hi-C protocol using *PGK1* ligation junction primers. Phusion and Q5 DNA polymerases were also tested; the Phusion reaction was carried out identically to the test PCR step of the Hi-C protocol, but *PGK1* ligation junction primers were used. Q5 PCR reactions were carried out by adding 17.35 uL Q5 PCR master mix (Volume per reaction: 1.25 uL 10 uM Primer 1, 16.1 uL ddH2O) and 6.65 uL Q5 Enzyme master mix (Volume per reaction: 1.25 uL 10 uM Primer 1, 5 uL 5X Q5 Reaction Buffer [NEB, #B9027S], 0.2 uL 25 mM dNTPs, 0.2 uL Q5 DNA Polymerase [NEB, #M0491S]) to low-bind PCR tubes containing 1 uL Hi-C DNA. The control reaction contains 1 uL ddH2O rather than DNA. The reactions were performed on a thermocycler using the following program:

1 cycle:	98 °C	30 seconds
35 cycles:	98 °C	10 seconds
	65 °C	20 seconds
	72 °C	30 seconds
1 cycle:	72 °C	2 minutes
1 cycle:	4 °C	hold

After ligating adapters to purified, end-repaired PCR fragments, test Phusion PCR was performed using Hi-C end PCR primers designed to hybridize to Illumina adapter sequences.

Hi-C Data Analysis

Hi-C contact matrices were generated using the CPU version of Juicer on Compute Canada's Graham cluster following distributor instructions [10]. HiCUP (version 0.7.3) and was used with default parameters to align and filter raw Hi-C reads in fastq format [11]. HIFI (version 1.0.0) step one, BAMtoSparseMatrix.py, was used to generate a sparse matrix for each chromosome from HiCUP-aligned BAM files following the program's documentation [13]. Step four of HIFI, SparseToFixed.py, was used to generate fixed-binning matrices from each output file from step one using the --no-score flag to create an input file compatible with Juicer Tools pre. Pre-compatible chromosome files were concatenated and Juicer Tools pre was used to construct a contact matrix in .hic format [10]. Following contact matrix (.hic file) generation, Juicer Tools (version 1.21.01) were used to calculate the principal component (Eigenvector) of each map, allowing for demarcation of compartments. As principal component values are assigned positive or negative signs arbitrarily during calculation, a modified version of a Miura et al. protocol for assigning A and B compartment types from Eigenvector values was used (Miura et al., 2018). Rather than correcting the signs by gene density, RNA-seq expression values were used. Raw RNA-seq data were downloaded from publicly available sources and processed using the Genpipes rnaseq pipeline (version 4.1.2) following distributors' instructions [61]. In short, bedtools' makewindows function was used to create a genomic bins file from the chrom.sizes file found with the sequencing data of the proper reference genome [62]. The bedtools intersect function was then used to calculate the RNA-seq gene expression per genomic bin, outputting a

bedGraph file. Using UCSC's bedGraphToBigWig tool, the bedGraph file was converted to the binary bigWig format. BedGraph files were created from the Eigenvector calculations of each chromosome using awk and grep, then bedGraphToBigWig was used to convert them to the bigWig format. Using UCSC's wigCorrelate tool in conjunction with awk, the signs of the Eigenvector calculation per bin were corrected such that positive values reflect A compartments and negative values reflect B compartments based on the correlation of the expression value per bin. Bins which have a calculated Eigenvector value of 0 are labeled "other". In earlier phases, compartment assignment was also performed using gene density as described in Miura et al. and by average GC content per bin; as euchromatin tends to have higher GC content, it is expected that A compartments will have higher average GC content than B compartments [63]. GC content per bin was calculated using a modified script originally written by Kaiwan Rahimian designed to count the ratio of GC to AT base pairs in a given reference genome per bin of any specified size (see Supplementary Information). Using the output of the script, compartments were corrected by calculating the average GC percentage of bins with positive Eigenvector values compared to negative values. If the average GC content of bins with negative Eigenvector values was higher than that of those with positive values, all Eigenvector values were multiplied by -1 such that positive values reflect A compartments and negative values reflect B compartments. While it was decided that RNA-seq gene expression data per bin was the optimal method of correcting Eigenvector values to reflect compartment types, the GC content and gene density bin calculations were retained for analysis. In order to compare entire compartments rather than individual bins, a script was written to take a .csv file containing all binned information for a given chromosome and concatenate statistics of each compartment. In short, the script counts the number of consecutive A, B, or "other" compartment assignments and, for

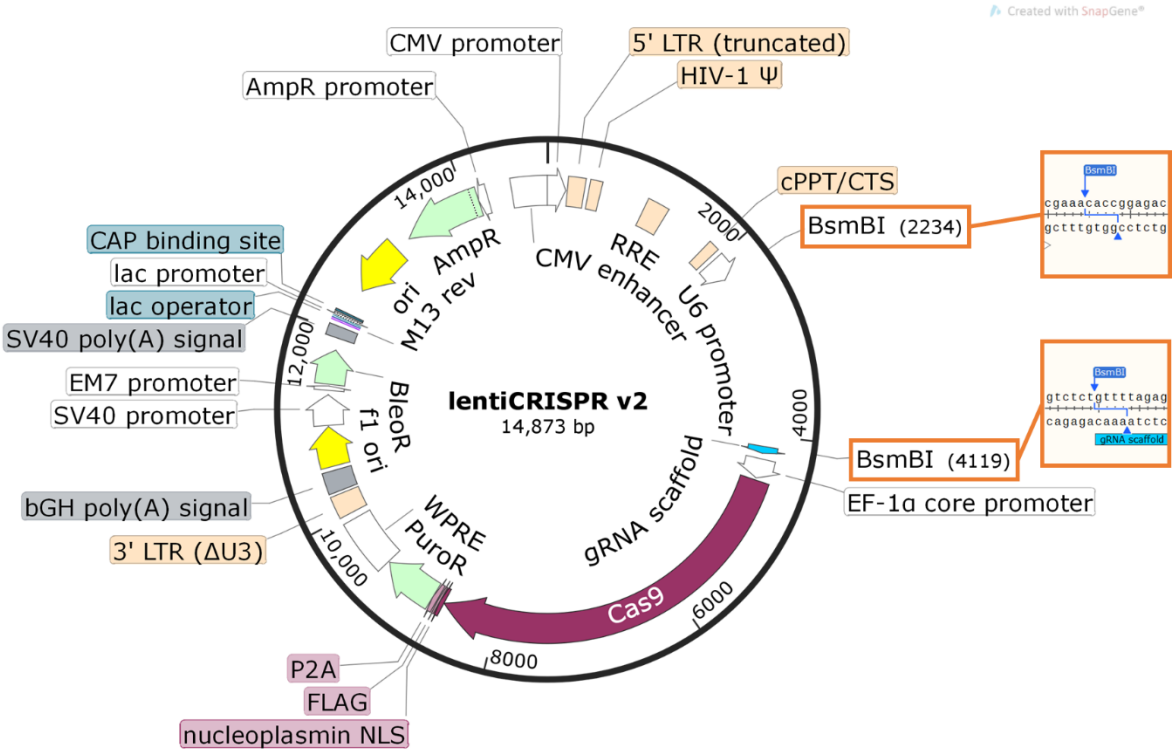
each compartment, returns the start and end bin, the compartment size (using bins as the unit), the average of the GC content across the compartment, a total and average of the RNA-seq expression data across the compartment, and a total and average of the gene density across the compartment. This script was used to concatenate compartments from control and experimental Hi-C data. To easily compare attributes of specific genes in control and experimental Hi-C maps, a script was written which takes as arguments the start and end points of the gene, the .csv containing the control Hi-C datapoints for the gene's chromosome, and the .csv containing the experimental Hi-C datapoints for the gene's chromosome. Running the script will output the compartment type and RNA-seq expression data of the start and end of the gene in each dataset for comparison, along with the consistent attributes (GC content and gene density) of the locus. Any gene of interest can be compared in this manner, though gene start and end sites must be entered individually.

VIII. Results

Successful sgRNA cloning

Per the plasmid designers' recommendations, Stbl3 were used to amplify both unedited and edited lentiCRISPRv2. Stbl3 are recombination-deficient and are thus ideal for a lentiviral plasmid containing long-terminal repeats (LTRs); the Zhang laboratory notes that of the options for recombination-deficient bacteria, Stbl3 performed the best [7, 63]. Competent Stbl3 cells were prepared in-house as described in [60]. The plasmid contains two BsmBI cut sites (Figure 2A). The small guide RNA (sgRNA) oligonucleotide sequences (shown in Supplementary Materials) were designed such that Cas9 would target the *HOTAIRM1* gene slightly downstream of both transcription start sites annotated in the UCSC genome browser (Figure 2B). As the BsmBI cut sites are not identical, the sgRNA oligos were designed with overhangs specifically matching the sticky ends left by BsmBI, allowing for directional ligation. After plasmid digestion, ligation, transformation into bacterial cells and purification, a diagnostic digest was performed using HindIII to confirm that the linker sequence is no longer present (Figure 3A, B). Following a successful digest, a diagnostic PCR was performed to determine whether sgRNA insertion was successful (Figure 3C). Sanger sequencing was used to confirm the sequence of four clones, one of which was selected for transfection into NCCIT cells along with a single stranded oligo-deoxynucleotide repair template containing the poly-adenosine sequence for knock-in into *HOTAIRM1*. The sequence is shown in Supplementary Materials File S33.

A



B

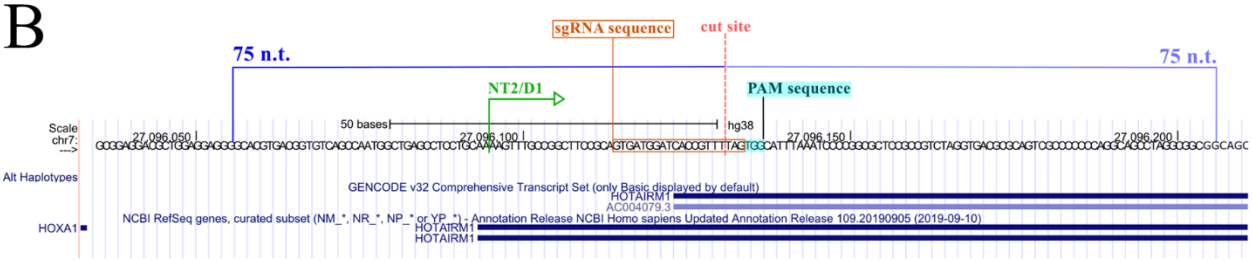


Figure 2. A) Overview of the structure of the lentiCRISPRv2 plasmid. The BsmBI sites are shown in detail; notably, the two cut sites have different sequences, thus the overhangs of the sgRNA oligos differ and sgRNA is ligated directionally. **B)** Overview of the editing region of the *HOTAIRM1* gene. The sgRNA sequence is boxed in orange. The blue 75nt segments represent the design of the homology arms of the repair template. The NT2-D1 transcription start site is marked in green (Wang and Dostie, 2017). The PAM (NGG) sequence is boxed in aqua and the cut site is shown in red.

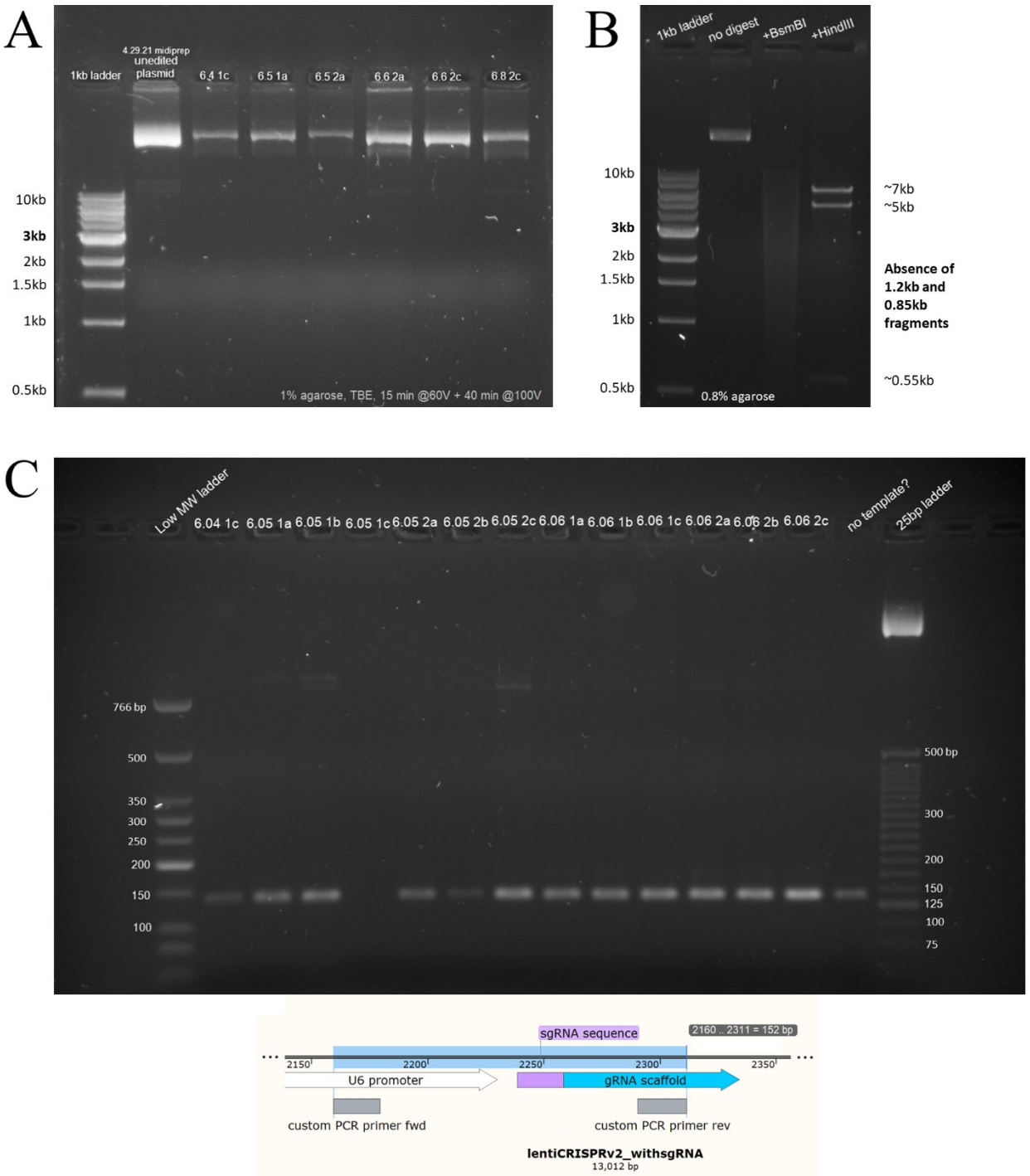


Figure 3. A) An agarose gel showing purified edited plasmid (6.4 1c, 6.5 1a, 6.5 2a, 6.6 2a, 6.6 2c, 6.8 2c) and unedited plasmid (4.29.21 midiprep). As the plasmid is ~15kb prior to editing and ~13kb after editing, the difference is not clearly shown. **B)** An edited plasmid clone digested with BsmBI and HindIII. The BsmBI should not have cut, but the HindIII reaction shows the expected result for a plasmid which does not contain the spacer region. **C)** PCR reactions of multiple clones of edited plasmid using PCR primers designed to amplify the editing region. Though the PCR product did not appear to run at 152bp as expected, four of these clones (6.01 1a, 6.05 1b, 6.06 2a, 6.08 2d) had their sequences confirmed using Sanger sequencing.

Optimization of the Hi-C library protocol

Initial attempts at implementing the lab's protocol for generating Hi-C libraries were conducted using NT2-D1 cells. As NCCIT cells are easier to manipulate– including editing with CRSPR/Cas9–, it was decided to switch from NT2-D1 to NCCIT cells. As a result, the first major quality check step, the PCR titration, needed to be altered. The PCR titration step is designed to amplify a fragment derived from a ligation junction and is meant to assess the efficiency at which filled-in digested fragments have ligated in the sample. To this end, a set of primers is designed to anneal adjacent to non-contiguous restriction sites corresponding to the enzyme selected for library preparation. The NT2-D1 protocol uses ligation junction primers located in a gene desert (GD09 and GD10, sequences shown in Supplementary Material file S33).

In NCCIT libraries however, these primers produced very weak or no bands (Figure 4). Thus, new ligation junction primers were designed. As chromatin availability could have caused difficulty with the gene desert primer set, primers were designed within the *PGK1* and *HOXA1* genes, which are both expected to be in an open and accessible chromatin state in pluripotent cells. I designed different primer sets to obtain a good one that could also be used for the second step of the PCR titration aimed at checking the initial digestion efficiency. Ideally, the PCR fragment amplified during the titration would not be digested by NcoI but would be digested by NsiI, an enzyme with an inverse cut site, indicating that the chromatin was completely digested and completely ligated in a head-to-head fashion. As the first set of *PGK1* primers happened to amplify a product containing a NcoI cut site, which is problematic, I designed a second primer pair within *PGK1* to improve the usefulness of the PCR titration as a quality check step. The *HOXA1* primers produced weak bands. All three new primer pairs are shown in Figure 5.

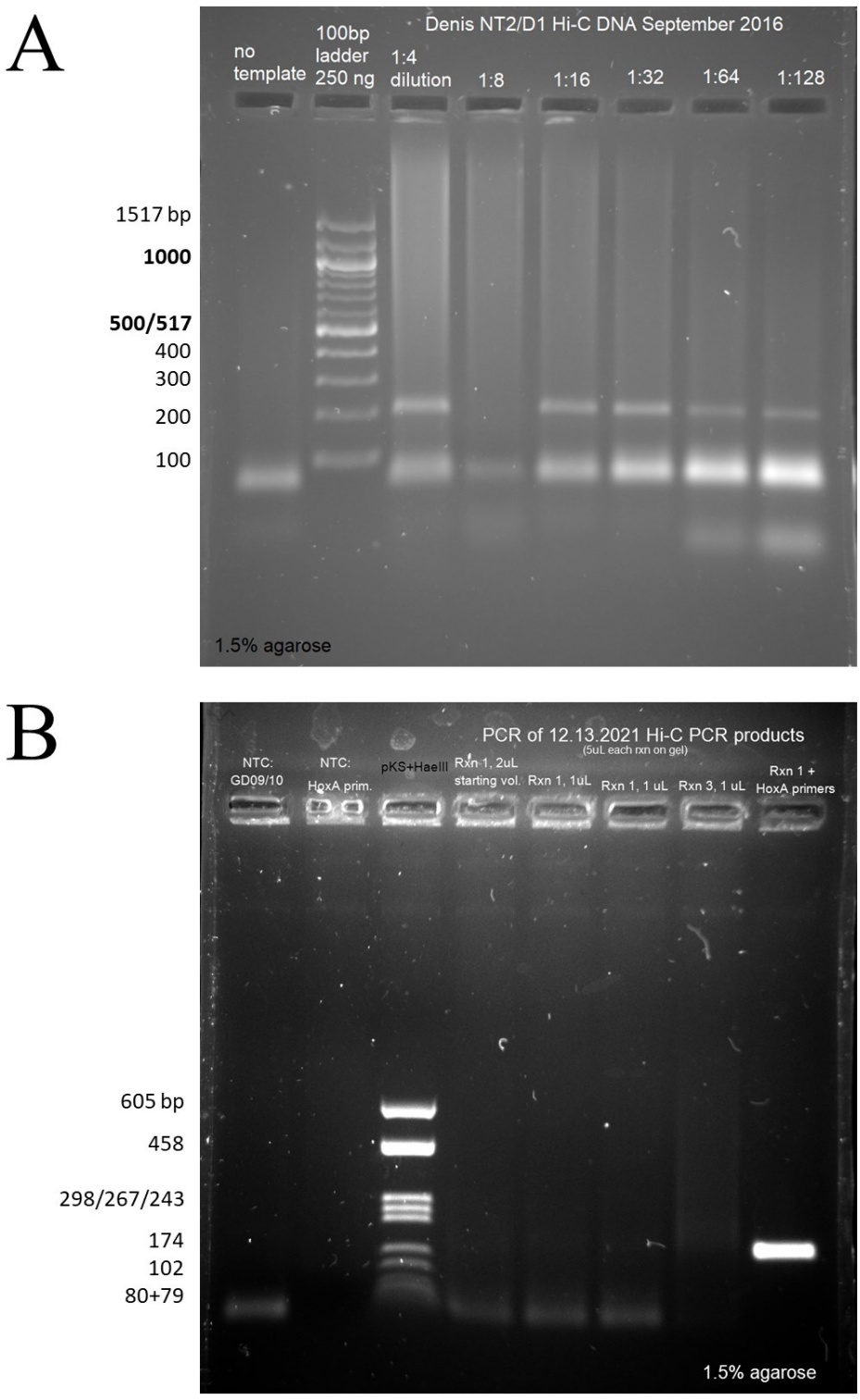


Figure 4. A) GD09/GD10 primers used on Hi-C DNA derived from NT2-D1 cells. **B)** GD09/GD10 primers used on Hi-C DNA derived from NCCIT cells. The rightmost lane shows a positive control to test the efficiency of the PCR itself; non-ligation junction primers (designed for RT-qPCR) located within the *HOXA* gene were used.

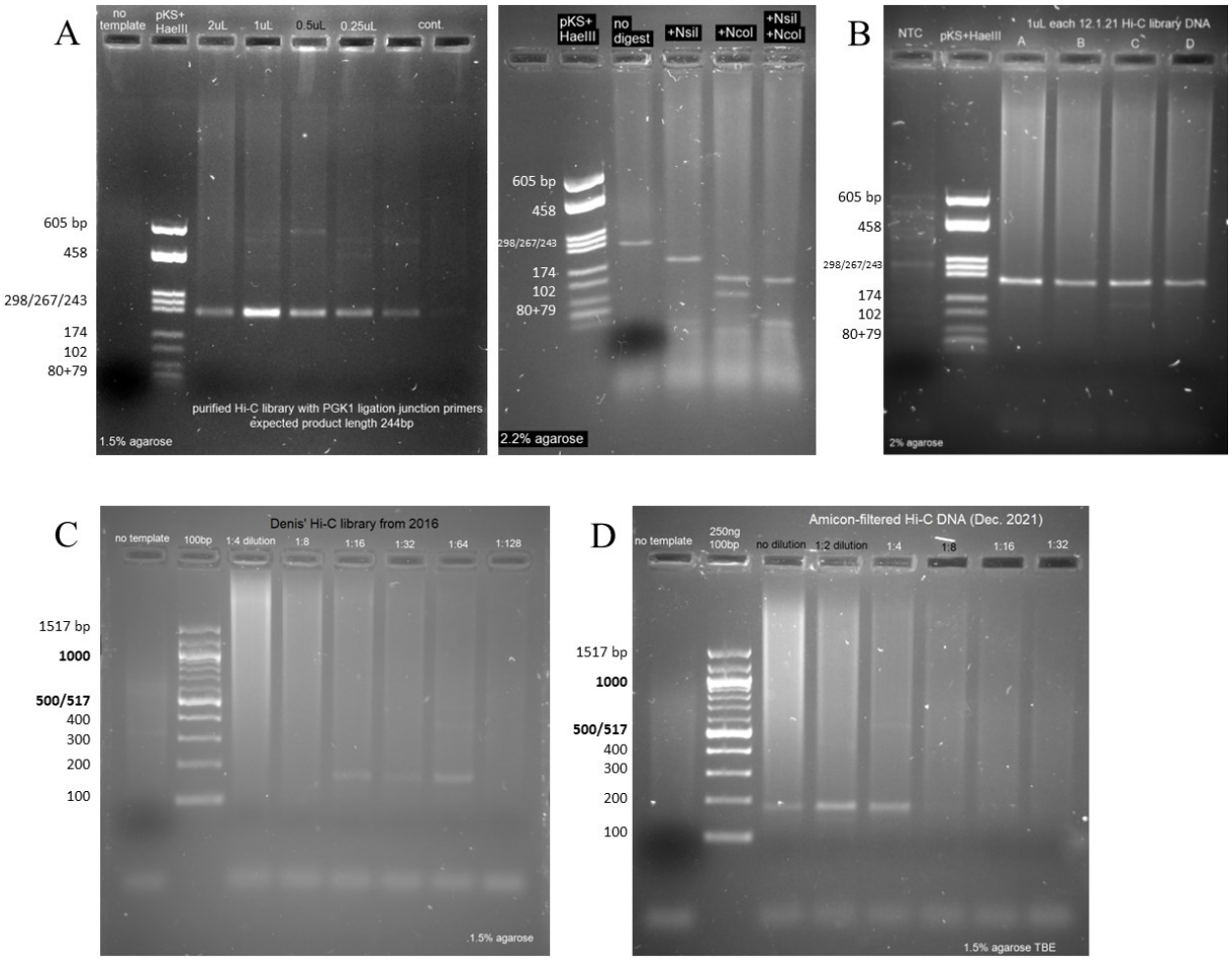


Figure 5. A) ***Left panel:*** PCR titration of NCCIT Hi-C DNA using PGK1v1 ligation junction primers. ***Right panel:*** diagnostic digest of PCR product obtained from titration; NcoI is not expected to cut the product but the product contains a pseudo-NcoI cut site which may be digested here. **B)** Test PCR using PGK1v2 ligation junction primers on NCCIT Hi-C DNA. **C)** Test PCR titration using *HOXA1* ligation junction primers on NT2-D1 Hi-C DNA. **D)** Test PCR titration using *HOXA1* ligation junction primers on NCCIT Hi-C DNA.

Adapter ligation step requires further optimization

The sonication, size selection, quantification and biotin pulldown were performed without any discernable issues. An example of a size selection gel is shown in Figure 6. An example of Quant-iT Hi-C DNA quantification is shown in Figure 7. Notably, during the Quant-iT picogreen quantification, the supernatant of the biotin pulldown was quantified as well. Using the average of the three measurements of the total DNA in this specific Hi-C library before adding Streptavidin beads to pull down the ligation junction fragments, 2.465 ug, only about 0.385 ug or 15.6% of the library remained on beads after pulldown.

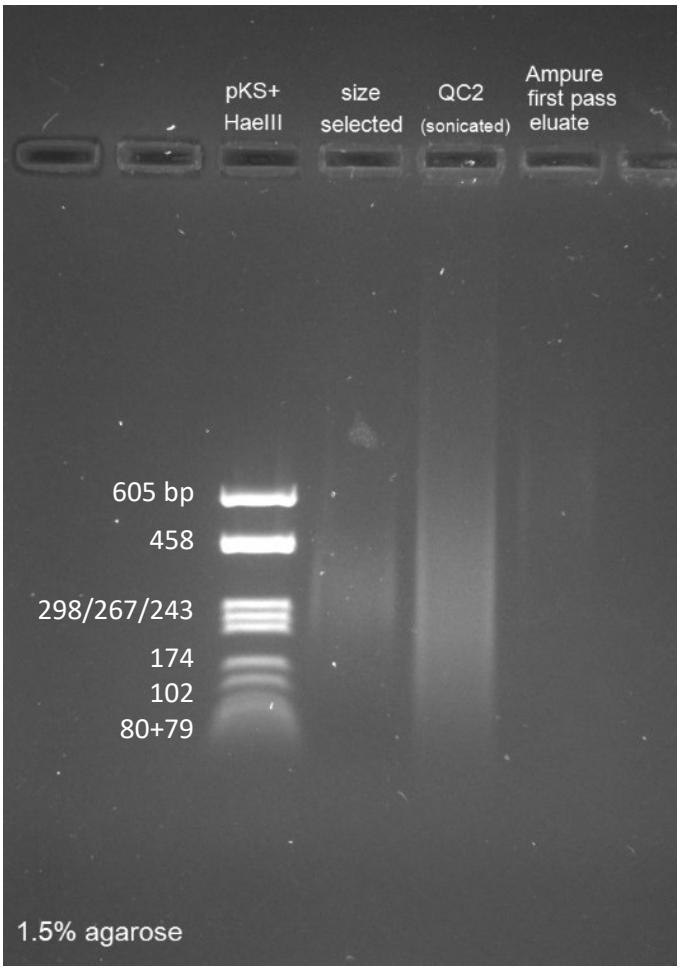
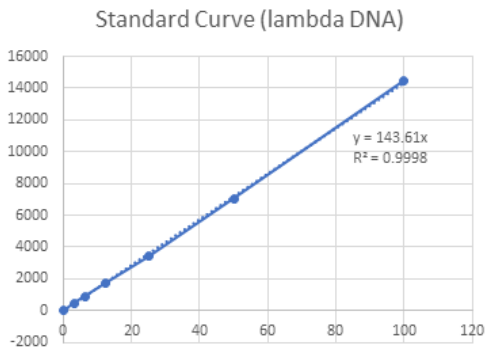


Figure 6: An example of a size-selected Hi-C library. The sonicated library is shown in the QC2 lane. The rightmost lane shows the eluate from the first pass of Ampure XP beads purification, recapitulating that the longer fragments were filtered out successfully.

A

Fluorescence readings

Row/Column	1 (Standard curve)	3 (Sample)
A	14619	47036
B	7195	21924
C	3651	12677
D	1894	4978
E	1037	
F	624	
G	143	



B

Fluorescence readings

DNA	Fluorescence reading	pg/uL picogreen	total pg picogreen	pg/ul in sample	ug/uL in sample	ug in 300uL Hi-C DNA
4uL Hi-C lib.	47036	327.5259	32752.59	8188.148	0.008188	2.456445
2uL Hi-C lib.	21924	152.6635	15266.35	7633.173	0.007633	2.289952
1uL Hi-C lib.	12677	88.2738	8827.38	8827.38	0.008827	2.648214
1uL Streptavidin beads runoff	4978	34.66332428	3466.332428	3466.332428	0.003466332	2.079799457

Figure 7: An example calculation using the Quant-iT Picogreen ds-DNA quantification kit. The coefficient obtained from the standard curve, constructed from the fluorescence measurements of fixed amounts of lambda DNA, is used to calculate the total mass of DNA in a given Hi-C sample. Three samples of varying volumes (4uL, 2uL, 1uL) of the Hi-C library are quantified, along with 1uL of the buffer left over from biotin pulldown with Streptavidin beads.

After optimizing the majority of the Hi-C protocol, difficulties were encountered with the end repair and adapter ligation steps. Initially, TruSeq adapters were ordered directly from Illumina [Set A, #20015960] and used as outlined in Methods. At the quality control Phusion PCR step, however, instead of obtaining the expected result of a smear spanning 350-550bp, a single band at approximately 120bp was observed (Figure 8). This size corresponds with the approximate length of two pairs of adapters ligated together. The band was not eliminated after using a Lithium Chloride wash to remove any non-biotinylated DNA from the Streptavidin beads, indicating that the adapters were indeed attached to the Hi-C DNA on the beads. To further confirm that the PCR product represents an adapter dimer, the product was digested with MboI; the dimer sequence contains a MboI site.

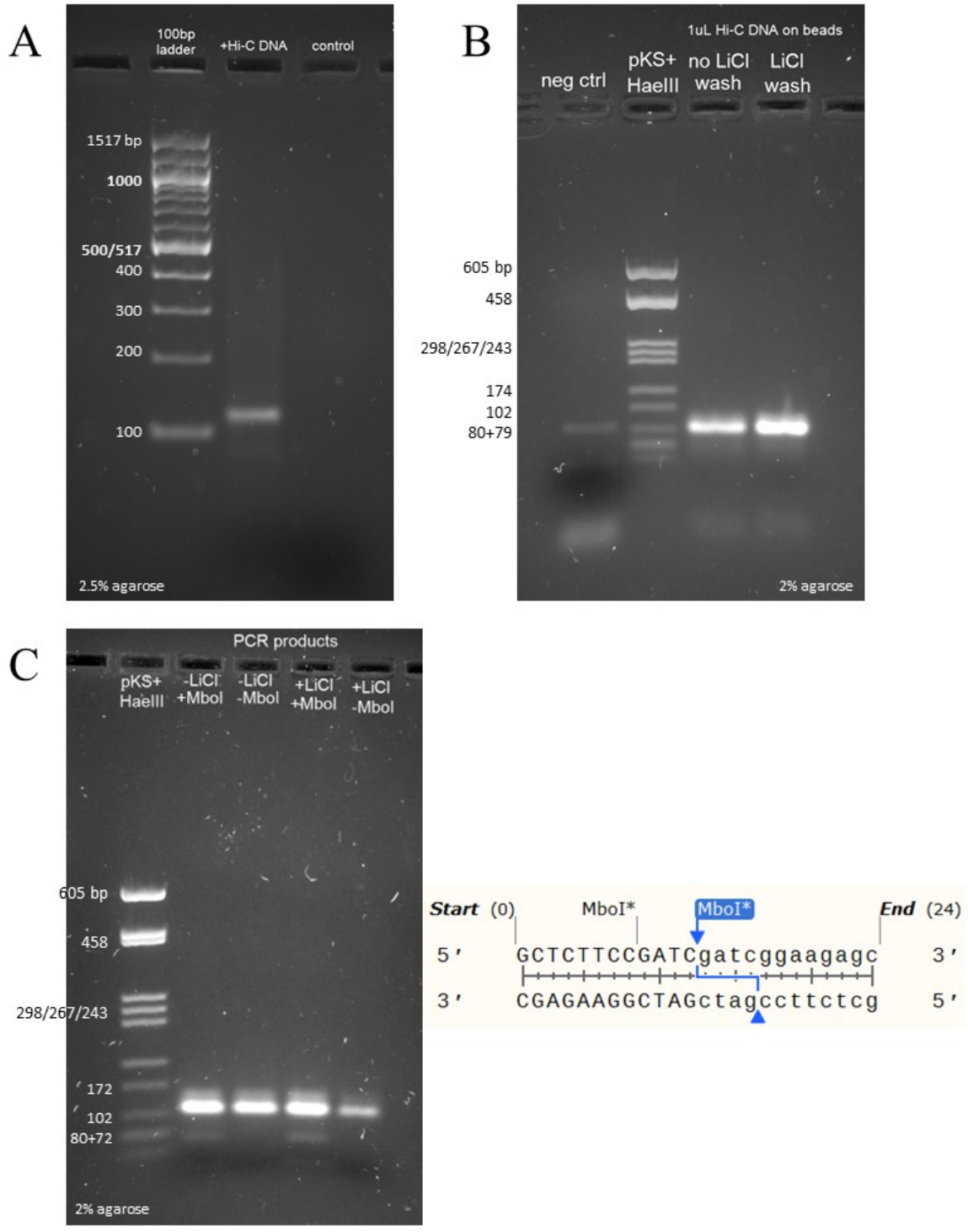


Figure 8. A) Phusion PCR of Hi-C DNA on Streptavidin beads. **B)** Phusion PCR to compare Hi-C DNA on Streptavidin beads before and after a LiCl wash. **C)** ***Top panel:*** PCR products cut with MboI, suggesting that the PCR product is a dimer of TruSeq adapters. A limited quantity of MboI was used, likely explaining the low ratio of cut to uncut DNA. ***Bottom panel:*** the sequence

of the complimentary segment of an adapter dimer, assuming the singular-T overhang was lost. MboI cut sites are shown in blue.

To avoid any potential problems with the manufacturing or shipping conditions, which we suspect led to adapter dimers, single-stranded oligodeoxynucleotides were ordered following the sequences of each component of annealed adapters: the Illumina Universal Adapter and Adapter Index #2 (sequences shown in Supplementary Materials file S33). The oligos were resuspended in TE and annealed immediately prior to ligation as described in Methods. The use of oligos in place of Illumina kit components resulted in a Phusion PCR gel which showed no bright band at 120bp, but no expected product was seen either (Figure 9). To test the efficiency of the end repair and adapter ligation protocol, test reactions were carried out using PCR product fragments in place of Hi-C DNA as described in Methods. Using Taq DNA polymerase and following the full end repair and adapter ligation protocol resulted in approximately 25-30% of PCR fragments shifting upwards by 120bp on gel, presumably indicating that adapters were ligated onto each end of the fragments. Another 30-35% of the fragments shifted up by approximately 60bp, indicating that adapters were successfully ligated to one end. The remaining fragments did not shift up, indicating that ligation efficiency must be optimized. Representative examples of these experiments are shown in Figure 10.

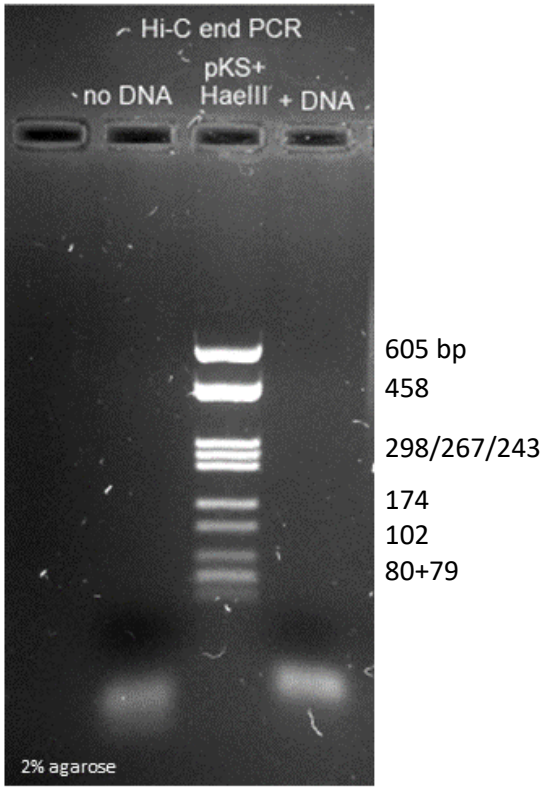


Figure 9. Phusion PCR result after annealed adapter oligos were ligated onto Hi-C DNA (stabilized on Streptavidin beads).

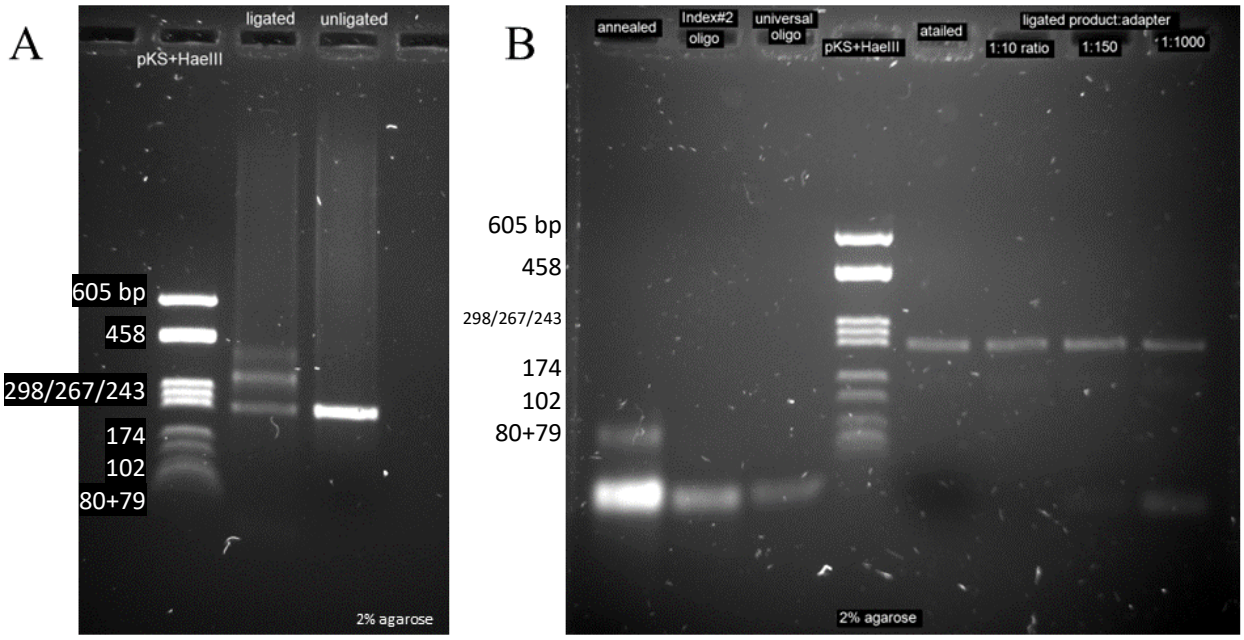


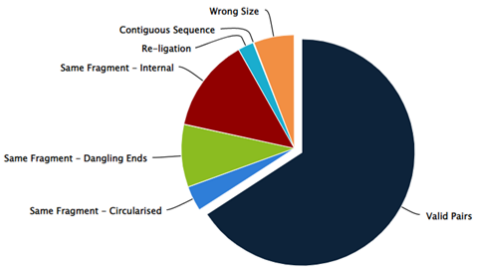
Figure 10. A) Taq DNA polymerase-derived PCR product which has been end-repaired, A-tailed, and ligated with annealed adapter oligos. **B)** Leftmost lanes show the annealed adapter mix and each of the two oligos in order; notably, the ratio of annealed (higher band) to single oligos is relatively low. Rightmost lanes show Taq DNA polymerase-derived PCR products ligated with various ratios of annealed adapter mix. None of the reactions appear successful. Ratios were calculated based on gel quantification of PCR product bands and the known concentration of annealed adapter mix without taking the annealed complex to single stranded oligo ratio into account.

Juicer and Juicer tools can be used to observe differences in Hi-C contacts

While optimizing the Hi-C protocol at the bench, I started exploring and implementing available computational tools for Hi-C analysis using Hi-C sequencing data generated by another student from our lab along with publicly available Hi-C datasets. Two pipelines were used to align and filter raw Hi-C sequencing data: Juicer and HiCUP with further processing in HIFI and Juicer Tools [10, 11, 13]. Initially, the Juicer pipeline encountered errors on the Compute Canada cluster Graham wherein fragments were improperly assigned intra-chromosomal, and the sex chromosomes were improperly assigned to chromosome number 23 in both mouse and human libraries. For this reason, HiCUP was used to align Hi-C data, resulting in a BAM [63] file type. BAM files were processed using HIFI and Juicer Tools as described in Methods to produce .hic files viewable with Juicebox. This method was used to compare basic statistics between an in-solution mouse Hi-C library and an in-situ mouse Hi-C library generated by students Kris Cao and Christopher Cameron; the results generated by HiCUP showed that the proportion of trans contacts to cis contacts was drastically higher in the in-solution ligation library. These results, shown in Figure 11, align with a prior study comparing the biases encountered in in-solution versus in-situ Hi-C libraries [39]. In particular, the high proportion of trans contacts in the in-solution Hi-C library indicate the likelihood that many unique contact pairs arise from random ligation. The in-situ library also contains a higher number of unique read pairs, allowing for higher resolution in a Hi-C contact matrix compared to the in-solution dataset.

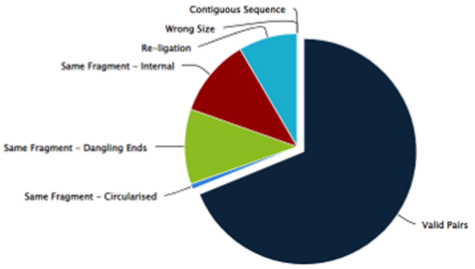
A **Filtering**

	Di-Tag Count
Valid Pairs	29,838,507
Invalid Pairs	15,464,319
Same Circularised	1,634,276
Same Fragment Dangling Ends	4,053,423
Same Fragment Internal	6,063,205
Re-ligation	1,019,688
Contiguous Sequence	37,916
Wrong Size	2,655,811
Total Pairs	45,302,826



B **Filtering**

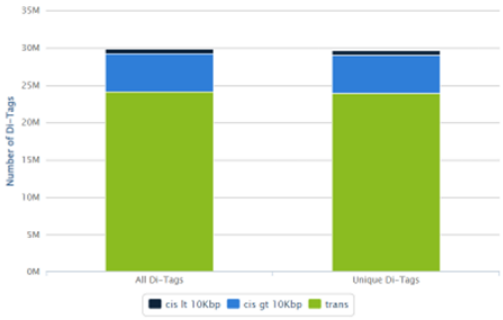
	Di-Tag Count
Valid Pairs	84,919,653
Invalid Pairs	38,391,597
Same Circularised	941,442
Same Fragment Dangling Ends	13,270,608
Same Fragment Internal	13,733,384
Re-ligation	10,446,163
Contiguous Sequence	0
Wrong Size	0
Total Pairs	123,311,250



C **De-duplication**

Percentage uniques: 99.1

	All Di-Tags	Unique Di-Tags
Read Pairs	29,838,507	29,584,137
Cis-close (< 10Kbp)	616,670	611,229
Cis-far (> 10Kbp)	5,147,480	5,103,579
Trans	24,074,357	23,869,329



D **De-duplication**

Percentage uniques: 96.65

	All Di-Tags	Unique Di-Tags
Read Pairs	84,919,653	82,078,381
Cis-close (< 10Kbp)	11,904,126	11,496,040
Cis-far (> 10Kbp)	59,451,839	57,465,657
Trans	13,563,688	13,116,684

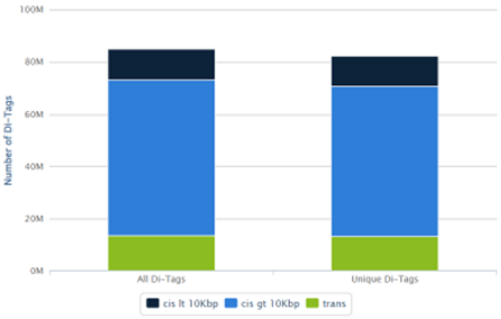


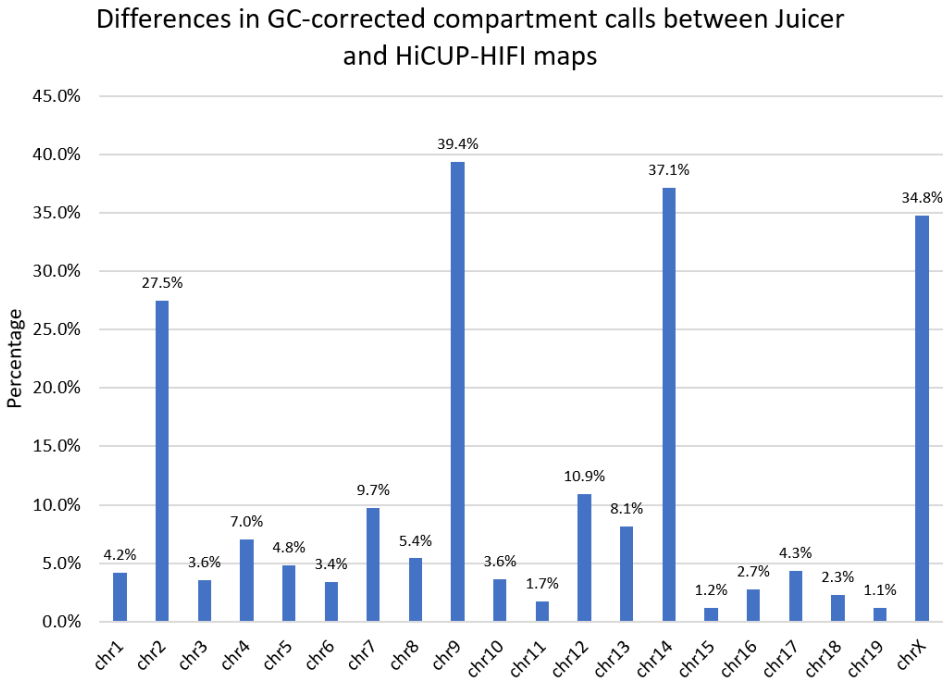
Figure 11. A) Filtering results from the mouse ESC in-solution ligation Hi-C library. All terms are elaborated on in Table 1. **B)** Filtering results from the mouse ESC in-situ ligation Hi-C library. **C)** De-duplication results from the mouse ESC in-solution ligation Hi-C library. D) De-duplication results from the mouse ESC in-situ ligation Hi-C library.

Table 1. HiCUP di-tag classifications.

truncated	Reads are truncated at the modified restriction site to remove bases that may inhibit a read from mapping to the reference genome.
unique	After duplicate reads are discarded, the percentage of di-tags kept is expressed.
re-ligation	Di-tags formed by proximal ligation between two adjacent restriction fragments are discarded.
same	Read pairs which map to the same restriction fragment are discarded.
same circularized	Read pairs which map to the same restriction fragment wherein the orientation is inverse, indicating a circularized fragment, are discarded.
same dangling end/internal	Read pairs which fall between two restriction sites or wherein only one end contains a restriction site are discarded.
wrong size	Di-tags which fall outside of the range targeted during the size selection step of the Hi-C protocol are discarded.
contiguous	Fragments which are determined to be incompletely digested or re-ligated are discarded.

After an update to the Standard Environment of Graham, the Juicer pipeline no longer encountered errors. As the Juicer pipeline performs mapping and matrix generation in one step for the user and is capable of mapping inter-chromosomal contacts, its usage was preferred for subsequent libraries. To determine whether Juicer results were comparable to HiCUP and HIFI results, the same mouse ESC Hi-C sequencing data was processed using both pipelines and the Eigenvector values compared. All Eigenvector values were calculated with Juicer Tools. As shown in Figure 12, the same Hi-C sequencing data processed through different pipelines results in differences in compartment calls when compartment types are corrected by GC content. Using the same Juicer-generated contact matrix, different methods of correcting A and B compartments were tested. It was found that four chromosomes (3, 7, 9, and 12) had their compartment calls completely reversed when comparing GC content correction to gene density correction.

A



B

Chromosome	Number of 100kb bins with different A/B classification between GC% and GD corrections	Total 100kb bins in chromosome	Total 100kb bins in chromosome excluding eigenvector = 0
chr3	1569	1601	1569
chr7	1424	1455	1424
chr9	1216	1246	1216
chr12	1171	1202	1171

Figure 12. A) Mouse ESC Hi-C sequencing data was processed through Juicer or HiCUP and HIFI; when Eigenvector values are corrected using the GC content per bin as described in Methods, differences arise. The percentage of bins with different compartment calls are shown per chromosome. **B)** The four chromosomes for which GC content-corrected (GC%) and gene density (GD) resulted in reversed compartment calls in the same Juicer-derived contact matrix. Raw sequencing data: Fraser et al., 2015 [66].

Based on these results, it was decided that Juicer would be used going forward to streamline the processing protocol. Focus was shifted to Hi-C data from human cell samples. After generating contact matrices using Juicer from publicly available H1-hESC and definitive endoderm raw Hi-C sequencing data, it was found that in certain chromosomes the RNA-seq expression values were higher (average and/or total values) in B compartments than in A compartments when compartments were corrected using GC content (Figure 13).

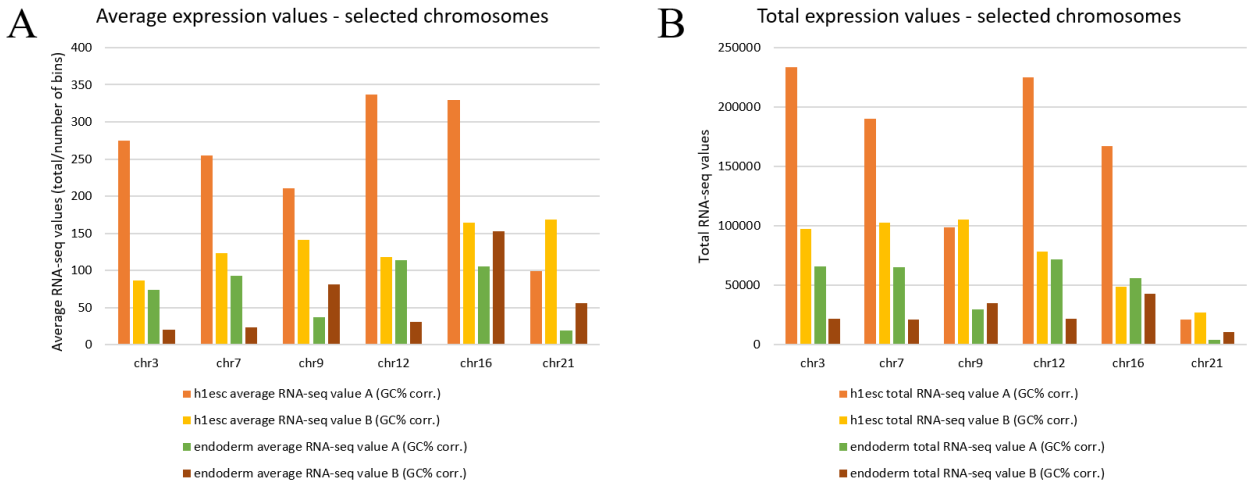


Figure 13: When compartments were called by GC content (GC%) on H1-hESC and definitive endoderm Hi-C data, the compartment types do not necessarily align with RNA-seq expression data; A compartments are expected to have higher expression than B compartments. **A)** RNA-seq data per compartment type per lineage expressed as the average expression value per bin of each compartment type. Chromosomes 9, 16, and 21 have unexpected results in the endoderm dataset. **B)** RNA-seq data per compartment type per lineage expressed as the total expression value per compartment type. In this view, chromosomes 9 and 21 have unexpected patterns in the H1-hESC dataset as well as the endoderm dataset.

Based on these results, it was decided that a more reliable method of correcting Eigenvector values to reflect A and B compartment types was needed. While correcting by gene density was appealing, it was decided that correcting by gene expression data would be more reliable for comparison between Hi-C data derived from differentiated and undifferentiated cells as gene density is a static value per bin while gene expression data is specific to each condition. Following compartment assignment, Excel spreadsheets were constructed to show characteristics of each bin for easy storage of data; for each 100kb bin, one line in a spreadsheet contains the average GC content, the gene density, and the Eigenvector values, compartment type, and RNA-seq expression value of one or more cell types.

Using a simple Python script, attributes of entire compartments were concatenated as described in Methods. Comparing concatenated compartment attributes of H1-hESC and definitive endoderm Hi-C datasets [67] showed that across the genome, H1-hESCs had a greater quantity of large (≥100 bins) compartments of both A and B type compared to endoderm cells, as shown in Figure 14. In order to compare the characteristics of individual genes or clusters, Python scripts were written to take as arguments two genomic positions and two .csv files containing information on the chromosome of interest. Two versions of the script were written to accommodate single-bin or concatenated compartment data. Examples of single-bin and whole compartment comparisons of genes of interest in a study of *HOTAIRM1* between undifferentiated and differentiated cell datasets is shown in Table SI and Table SII, respectively. Notably, compartment shifts before and after differentiation are visible through the starting and ending positions of compartments where certain genes of interest are located. While future studies will focus on differentiation in *HOTAIRM1* polyA-knock in cell lines rather than

unaltered cell lines, this comparison exercise illustrates the types of comparisons which are possible.

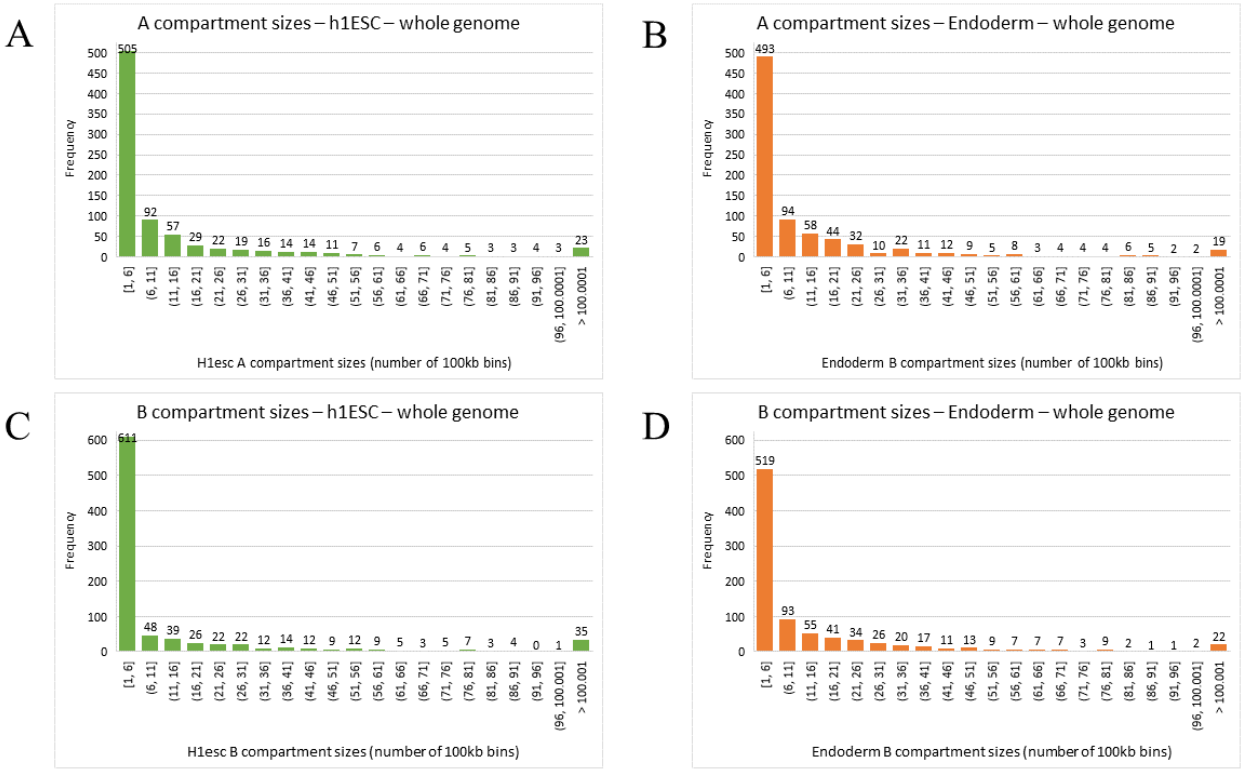


Figure 14: The distribution of compartment lengths across the genome derived from Hi-C data from H1-hESC and Endoderm cells. **A)** H1-hESC A compartment lengths. **B)** Endoderm A compartment lengths. **C)** H1-hESC B compartment lengths. **D)** Endoderm B compartment lengths.

IX. Discussion

In the future of this project, following the transfection of the CRISPR-Cas9 plasmid containing sgRNA targeting *HOTAIRM1* and the repair template containing a polyA site, several steps must be taken before Hi-C libraries in various conditions can be generated and analyzed. Currently, plasmid-transfected NCCIT cells are being selected using puromycin so that single cells can be selected for cell line generation. Once viable cells are available, the efficacy of the CRISPR-Cas9 mediated gene editing should be tested first by designing PCR primers flanking the insert region and performing a diagnostic PCR on DNA extracted from the bulk population. As the insert is rather short, the amplified fragment could be sent for sequencing for further confirmation. RT-qPCR should then be used to calculate knock-in efficiency and to test the effects of *HOTAIRM1*'s premature transcription termination on its target genes. Not only would we expect *HOTAIRM1* to be depleted, but other genes in *HOXA*'s 3' end, particularly *HOXA1*, should be depleted.

If the knock-in is found to be successful, cells should be prepared in different conditions for Hi-C library preparation. Ideally, four distinct libraries will be prepared to obtain a more complete picture of the effect of the knock-in on chromatin architecture: knock-in RA-induced NCCIT cells, uninduced knock-in NCCIT cells, RA-induced control NCCIT cells, and uninduced control NCCIT cells. The control undifferentiated and differentiated heatmaps will provide a baseline to which the knock-in datasets can be compared. In comparing polyA tail knock-in Hi-C data to control data, we would expect the previously reported unfolding of the *HOXA* cluster's 3' end during RA-induced differentiation to be diminished. In addition, future researchers should pay attention to any differences in contact frequency and compartment types in genes belonging to pathways which are reportedly regulated by *HOTAIRM1* in differentiation

or human disease, such as the *HOXA1*-Nanog regulatory loop, which contains the genes of interest *SOX2* and *POU5F1*, and is currently under investigation by our laboratory's PhD student, Dana Segal.

During the course of the Hi-C protocol, the end repair step may need further optimization. End repair testing experiments indicate room for improvement in the efficiency of adapter ligation, whether adapters are sourced from the Invitrogen kits or oligonucleotides ordered from IDT. Care was taken to order the universal adapter oligo from IDT with a phosphorothioate bond between the C and T nucleotides on the 3' end to minimize the chances of losing the singular T overhang; the Index #2 oligo was also ordered with a 5' phosphate group to allow for ligation, following guidelines found in the Illumina TruSeq Adapters Demystified Rev. A document from Tufts University [68]. Thus, issues were unlikely to be caused by loss of the singular T overhang. One possible problem is shown in Figure 9B, which shows in the leftmost lane that a relatively small proportion of the adapter mixture remains annealed after it has been through the annealing protocol and run on gel. While it is fraught quantifying single-stranded DNA or a mostly single-stranded DNA complex with Ethidium Bromide, as it is an intercalating agent and binds to double-stranded nucleic acids, it is estimated that 12.6% of the adapter mix remains in the annealed complex. Naturally, this conclusion would be more convincing if the solution was run on an acrylamide gel, but due to time constraints it was attempted to alleviate this possible issue by adding a very high ratio of adapters to DNA during the ligation reaction; notably, the Hi-C protocol used a very high ratio of adapters to DNA even before this problem was encountered. The underwhelming results from test experiments using PCR products in place of Hi-C DNA still indicate that the protocol likely requires further improvement in order to be useful. It must be noted that PCR product fragments are not

stabilized on Streptavidin beads during these experiments as they do not contain biotin, thus the comparison between these tests and Hi-C DNA results is not entirely comparable; that said, future experiments will possibly require additional tweaks to the protocol, such as variations in ligation conditions, or ordering new materials, such as a new Illumina adapter kit.

As a more drastic option, updated variations on the Hi-C protocol have been published in recent years. As an example, a low-input biotin-free variation termed eHi-C (easy Hi-C) [69], which is inspired by 4C and enrichment of ligation products (ELP) [70], has been established. The methodology uses an initial digestion with the 4-cutter DpnII, which results in fragment circularization from a self-ligation reaction; the circularized products are re-digested with the 6-cutter HindIII. Using this method may result in changes in the data analysis pipeline, as the researchers outline their own Hi-C data analysis pipeline named HiCorr for libraries derived from this method due to differences in biases from traditional Hi-C. That said, should the low recovery exemplified in Figure 6 be a contributing factor to the lack of success in the NCCIT Hi-C experiments covered by this thesis, no longer relying on biotin pulldown is an appealing option.

Precisely how the Hi-C library will be generated is yet to be determined, however the foundation established in this thesis can serve as a starting point for comparative data analysis following Hi-C sequencing. One benefit of generating stable knock-in cell lines is that successful clones can ideally be used for many different experiments through propagation. As various members of our laboratory work on ChIP, RNA extraction and RT-qPCR, expression and histone modification data can be used in conjunction with data obtained from Hi-C contact matrices to form a more complete picture of *HOTAIRM1*'s roles in chromatin organization in human pluripotent cells. As existing research has focused on the role of *HOTAIRM1* in conformational

changes in the *HOXA* locus [4] or in specific pathways of interest, the ability to observe whole-genome changes in chromatin conformation allows us to further understand any chromatin remodeling mechanisms by which *HOTAIRM1* affects gene expression. To achieve consistently with prior work from our laboratory examining chromatin conformation changes in NT2-D1, cells should be collected 72 hours after RA induction in order to perform comparisons [4]. These experiments may validate established target genes of *HOTAIRM1* and be used to speculate putative targets which have yet to be fully investigated.

The analysis method proposed in this thesis should be effective even with any future changes to the Hi-C library generation protocol. Testing of two different pipelines, Juicer or HiCUP with HIFI, showed that Juicer is the easier of the two to run successfully. This is valuable for reproducibility as the user makes fewer elective decisions when using a one-click pipeline. For future experiments, regardless of how the library was generated, Juicer can be used to align and filter raw paired-end reads. Once reads are aligned and a contact matrix is generated, Juicer tools can easily be used to calculate the principal component (Eigenvector) at the bin resolution of interest. Since all novel scripts used in the methods outlined in this thesis are dependent on semi-manually constructed Excel files, the method is simple to use provided the user is able to edit the scripts' input to reflect their preferred column headers. The three columns of each sheet are derived from the .bedGraph files which result from using Juicer Tools to calculate the principal component of each contact matrix, which need only be imported to Excel as a .tsv file. The remaining columns, at this stage of development, are manually copied into each sheet from other output files. For future usage, perhaps when comparing the characteristics of many Hi-C libraries, it may be worth writing an additional script which reads all of the necessary input files

for each chromosome and writes a .csv in one step, only relying on correct filenames and directory structure. Not only would this save time, but it reduces the chance of user error.

There is additionally room for improvement in the user-friendliness of each script explored in this process. As of now, while calculations can be easily performed as each script reads Excel files as Pandas DataFrames [71], they each require the user to open a Python environment and manually enter filenames. Streamlining each script such that each is usable on the command line would save time and allow the user to run each one in batches, improving the reproducibility by reducing human error. Additionally, the scope of this thesis has focused mainly on chromatin compartment types. In cases where it is also desirable to study TAD boundary locations, the currently existing scripts can serve as a foundation for additional scripts to compare TAD boundaries, ChIP-seq tracks, or any other positional datapoints of interest between two libraries. With a few changes and precise DataFrame construction, the total or average of any numerical attribute across any individual bin or multi-bin area can be easily reported and compared. For a genome-wide analysis such as the method proposed here, integrating the use of an additional DataFrame containing positional information of all genes of interest may be useful going forward.

Certainly, establishing this Hi-C library generation and data analysis method is done in the hopes that it can be used to compare the differences in chromatin organization between *HOTAIRM1* polyA-knock in cells and control cells. Despite challenges encountered along the way, this methodology is appealing for a variety of reasons. PolyA knock-in is attractive over other methods of interfering with *HOTAIRM1*'s expression such as deleting the gene entirely or CRISPRi. As *HOTAIRM1* is not insignificant in genomic length, deleting the region may alter the chromatin architecture in an unforeseen way, leading to possibly unexpected results.

CRISPRi stops RNA Polymerase from transcribing along the genome, and as chromatin architecture is altered during the transcription process, ideally transcription would proceed as usual for the most accurate Hi-C results. Additionally, polyA knock-in is appealing as the same cell line can be used for a variety of experiments. Members of our lab may be interested in performing ChIP, ChIP-seq, or RNA-seq protocols using the cell line generated for the purpose of Hi-C, deepening our overall understanding of *HOTAIRM1*'s effects on gene expression and epigenetic modifications.

X. Summary

This thesis has detailed advances in comparative analysis of Hi-C data for the purpose of observing differences in chromatin architecture between control NCCIT cells and NCCIT cells containing a knocked-in polyadenylation sequence near the transcription start site of the *HOTAIRM1* gene. While Hi-C libraries must be generated in future studies using clones successfully transfected with edited lentiCRISPRv2 and the repair template, this method lays the foundation for future investigations.

XI. Copyright

XII. Reference List

1. Fatica, A., & Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature reviews. Genetics, 15*(1), 7–21.
2. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature, 485*(7398), 376–380.
3. Ferraiuolo, M. A., Rousseau, M., Miyamoto, C., Shenker, S., Wang, X. Q., Nadler, M., Blanchette, M., & Dostie, J. (2010). The three-dimensional architecture of Hox cluster silencing. *Nucleic acids research, 38*(21), 7472–7484.
4. Wang, X. Q., & Dostie, J. (2017). Reciprocal regulation of chromatin state and architecture by *HOTAIRM1* contributes to temporal collinear *HOXA* gene activation. *Nucleic acids research, 45*(3), 1091–1104.
5. Wang, L., Wang, L., Wang, Q., Yosefi, B., Wei, S., Wang, X., & Shen, D. (2021). The function of long noncoding RNA *HOTAIRM1* in the progression of prostate cancer cells. *Andrologia, 53*(2), e13897.
6. Li, Y. Q., Sun, N., Zhang, C. S., Li, N., Wu, B., & Zhang, J. L. (2020). Inactivation of lncRNA *HOTAIRM1* caused by histone methyltransferase RIZ1 accelerated the proliferation and invasion of liver cancer. *European review for medical and pharmacological sciences, 24*(17), 8767–8777.
7. Sanjana, N. E., Shalem, O., & Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nature methods, 11*(8), 783–784.
8. Pradeepa, M. M., McKenna, F., Taylor, G. C., Bengani, H., Grimes, G. R., Wood, A. J., Bhatia, S., & Bickmore, W. A. (2017). Psip1/p52 regulates posterior *Hoxa* genes through activation of lncRNA Hottip. *PLoS genetics, 13*(4), e1006677.
9. Levitt, N., Briggs, D., Gil, A., & Proudfoot, N. J. (1989). Definition of an efficient synthetic poly(A) site. *Genes & development, 3*(7), 1019–1025.
10. Pleasure, S. J., & Lee, V. M. (1993). NTera 2 cells: a human cell line which displays characteristics expected of a human committed neuronal progenitor cell. *Journal of neuroscience research*, *35*(6), 585–602.
11. Damjanov, I., Horvat, B., & Gibas, Z. (1993). Retinoic acid-induced differentiation of the developmentally pluripotent human germ cell tumor-derived cell line, NCCIT. *Laboratory investigation; a journal of technical methods and pathology*, *68*(2), 220–232.
12. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems, 3*(1), 95–98.
13. Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). *HiCUP: pipeline for mapping and processing Hi-C data. F1000Research*, *4*, 1310.
14. Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems*, *3*(1), 99–101.
15. Cameron, C. J., Dostie, J., & Blanchette, M. (2020). HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome biology, 21*(1), 11.

16. Miura, H., Poonperm, R., Takahashi, S., & Hiratani, I. (2018). Practical Analysis of Hi-C Data: Generating A/B Compartment Profiles. Methods in molecular biology (Clifton, N.J.), 1861, 221–245.
17. Krumlauf R. (2018). Hox genes, clusters and collinearity. *The International journal of developmental biology*, *62*(11-12), 659–663.
18. Lewis E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature, 276*(5688), 565–570.
19. Izpisúa-Belmonte, J. C., Falkenstein, H., Dollé, P., Renucci, A., & Duboule, D. (1991). Murine genes related to the Drosophila AbdB homeotic genes are sequentially expressed during development of the posterior part of the body. *The EMBO journal, 10*(8), 2279–2289.
20. Mallo, M., Wellik, D. M., & Deschamps, J. (2010). Hox genes and regional patterning of the vertebrate body plan. *Developmental biology, 344*(1), 7–15.
21. Brosnan, C. A., & Voinnet, O. (2009). The long and the short of noncoding RNAs. *Current opinion in cell biology, 21*(3), 416–425.
22. Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., & Shen, F. (2019). Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *International journal of molecular sciences*, *20*(22), 5573.
23. Novikova, E. L., & Kulakova, M. A. (2021). There and Back Again: Hox Clusters Use Both DNA Strands. *Journal of developmental biology*, *9*(3), 28.
24. Zhang, X., Lian, Z., Padden, C., Gerstein, M. B., Rozowsky, J., Snyder, M., Gingeras, T. R., Kapranov, P., Weissman, S. M., & Newburger, P. E. (2009). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human *HOXA* cluster. *Blood*, *113*(11), 2526–2534.
25. Zhang, X., Weissman, S. M., & Newburger, P. E. (2014). Long intergenic non-coding RNA *HOTAIRM1* regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. *RNA biology*, *11*(6), 777–787.
26. Rea, J., Menci, V., Tollis, P., Santini, T., Armaos, A., Garone, M. G., Iberite, F., Cipriano, A., Tartaglia, G. G., Rosa, A., Ballarino, M., Laneve, P., & Caffarelli, E. (2020). *HOTAIRM1* regulates neuronal differentiation by modulating NEUROGENIN 2 and the downstream neurogenic cascade. *Cell death & disease*, *11*(7), 527.
27. Schuettengruber, B., Bourbon, H. M., Di Croce, L., & Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell*, *171*(1), 34–57.
28. Yuan, W., Wu, T., Fu, H., Dai, C., Wu, H., Liu, N., Li, X., Xu, M., Zhang, Z., Niu, T., Han, Z., Chai, J., Zhou, X. J., Gao, S., & Zhu, B. (2012). Dense chromatin activates Polycomb repressive complex 2 to regulate H3 lysine 27 methylation. *Science (New York, N.Y.)*, *337*(6097), 971–975.
29. Li, Q., Dong, C., Cui, J., Wang, Y., & Hong, X. (2018). Over-expressed lncRNA *HOTAIRM1* promotes tumor growth and invasion through up-regulating *HOXA1* and sequestering G9a/EZH2/Dnmts away from the *HOXA1* gene in glioblastoma multiforme. *Journal of experimental & clinical cancer research : CR*, *37*(1), 265.
30. Xie, P., Li, X., Chen, R., Liu, Y., Liu, D., Liu, W., Cui, G., & Xu, J. (2020). Upregulation of *HOTAIRM1* increases migration and invasion by glioblastoma cells. *Aging*, *13*(2), 2348–2364.

31. Cohen, M. E., Yin, M., Paznekas, W. A., Schertzer, M., Wood, S., & Jabs, E. W. (1998). Human SLUG gene organization, expression, and chromosome map location on 8q. *Genomics*, *51*(3), 468–471.
32. Ahmadov, U., Picard, D., Bartl, J., Silginer, M., Trajkovic-Arsic, M., Qin, N., Blümel, L., Wolter, M., Lim, J., Pauck, D., Winkelkotte, A. M., Melcher, M., Langini, M., Marquardt, V., Sander, F., Stefanski, A., Steltgens, S., Hassiepen, C., Kaufhold, A., Meyer, F. D., … Remke, M. (2021). The long non-coding RNA *HOTAIRM1* promotes tumor aggressiveness and radiotherapy resistance in glioblastoma. *Cell death & disease*, *12*(10), 885.
33. Li, Y. Q., Sun, N., Zhang, C. S., Li, N., Wu, B., & Zhang, J. L. (2020). Inactivation of lncRNA *HOTAIRM1* caused by histone methyltransferase RIZ1 accelerated the proliferation and invasion of liver cancer. *European review for medical and pharmacological sciences*, *24*(17), 8767–8777.
34. Li, D., & Zeng, Z. (2019). Epigenetic regulation of histone H3 in the process of hepatocellular tumorigenesis. *Bioscience reports*, *39*(8), BSR20191815.
35. Hamilton, M. J., Young, M., Jang, K., Sauer, S., Neang, V. E., King, A. T., Girke, T., & Martinez, E. (2020). *HOTAIRM1* lncRNA is downregulated in clear cell renal cell carcinoma and inhibits the hypoxia pathway. *Cancer letters*, *472*, 50–58.
36. Wan, L., Kong, J., Tang, J., Wu, Y., Xu, E., Lai, M., & Zhang, H. (2016). *HOTAIRM1* as a potential biomarker for diagnosis of colorectal cancer functions the role in the tumour suppressor. *Journal of cellular and molecular medicine*, *20*(11), 2036–2044.
37. Liu, J., Liu, J., & Lu, X. (2019). *HOXA1* upregulation is associated with poor prognosis and tumor progression in breast cancer. *Experimental and therapeutic medicine*, *17*(3), 1896–1902.
38. Kim, C. Y., Oh, J. H., Lee, J. Y., & Kim, M. H. (2020). The LncRNA *HOTAIRM1* Promotes Tamoxifen Resistance by Mediating *HOXA1* Expression in ER+ Breast Cancer Cells. *Journal of Cancer*, *11*(12), 3416–3423.
39. Chang M. (2012). Tamoxifen resistance in breast cancer. *Biomolecules & therapeutics*, *20*(3), 256–267.
40. Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews. Genetics*, *2*(4), 292–301.
41. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.
42. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380.
43. Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–1311.
44. Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., & Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature genetics, 38(11), 1341–1347.

45. Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, *16*(10), 1299–1309.
46. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–293.
47. Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, *43*(11), 1059–1065.
48. Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B. M., Wingett, S. W., & Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome biology*, *16*(1), 175.
49. Cao, J., Luo, Z., Cheng, Q., Xu, Q., Zhang, Y., Wang, F., Wu, Y., & Song, X. (2015). Three-dimensional regulation of transcription. *Protein & cell, 6*(4), 241–253.
50. Schmid, M. W., Grob, S., & Grossniklaus, U. (2015). HiCdat: a fast and easy-to-use Hi-C data analysis tool. *BMC bioinformatics, 16*(1), 277.
51. Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology, 16*, 259.
52. Sauria, M. E., Phillips-Cremins, J. E., Corces, V. G., & Taylor, J. (2015). HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome biology, 16*, 237.
53. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760.
54. Philip A. Knight, Daniel Ruiz, A fast algorithm for matrix balancing, IMA Journal of Numerical Analysis, Volume 33, Issue 3, July 2013, Pages 1029–1047
55. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M., Li, Y., Hu, M., Hardison, R., Wang, T., & Yue, F. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome biology, 19*(1), 151.
56. Koziol, Quincey, & Robinson, Dana. (2018, March 25). HDF5. [Computer software]. https://github.com/HDFGroup/hdf5.
57. Abdennur, N., & Mirny, L. A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics (Oxford, England), 36*(1), 311–316.
58. Stojic, L., Lun, A., Mascalchi, P., Ernst, C., Redmond, A. M., Mangei, J., Barr, A. R., Bousgouni, V., Bakal, C., Marioni, J. C., Odom, D. T., & Gergely, F. (2020). A high-content RNAi screen reveals multiple roles for long noncoding RNAs in cell division. Nature communications, 11(1), 1851.
59. Houldsworth, J., Reuter, V. E., Bosl, G. J., & Chaganti, R. S. (2001). ID gene expression varies with lineage during differentiation of pluripotential male germ cell tumor cell lines. *Cell and tissue research, 303*(3), 371–379.
60. Seidman, C. E., Struhl, K., Sheen, J., & Jessen, T. (2001). Introduction of plasmid DNA into cells. *Current protocols in molecular biology, Chapter 1*, Unit1.8.

61. Bourgey, M., Dali, R., Eveleigh, R., Chen, K. C., Letourneau, L., Fillon, J., Michaud, M., Caron, M., Sandoval, J., Lefebvre, F., Leveque, G., Mercier, E., Bujold, D., Marquis, P., Van, P. T., Anderson de Lima Morais, D., Tremblay, J., Shao, X., Henrion, E., Gonzalez, E., … Bourque, G. (2019). GenPipes: an open-source framework for distributed and scalable genomic analyses. *GigaScience, 8*(6), giz037.
62. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England), 26*(6), 841–842.
63. Murakami Y. (2013). Heterochromatin and Euchromatin. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) *Encyclopedia of Systems Biology*. Springer, New York, NY.
64. Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., & Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (New York, N.Y.)*, *343*(6166), 84–87. https://doi.org/10.1126/science.1247005
65. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078–2079.
66. Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R., FANTOM Consortium, Semple, C. A., … Nicodemi, M. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular systems biology, 11*(12), 852.
67. Akgol Oksuz, B., Yang, L., Abraham, S., Venev, S. V., Krietenstein, N., Parsi, K. M., Ozadam, H., Oomen, M. E., Nand, A., Mao, H., Genga, R., Maehr, R., Rando, O. J., Mirny, L. A., Gibcus, J. H., & Dekker, J. (2021). Systematic evaluation of chromosome conformation capture assays. *Nature methods, 18*(9), 1046–1055.
68. Schiemer, J. (2011). *Illumina TruSeq Adapters Demystified Rev. A*. [PDF]. Retrieved from http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf.
69. Lu, L., Liu, X., Huang, W. K., Giusti-Rodríguez, P., Cui, J., Zhang, S., Xu, W., Wen, Z., Ma, S., Rosen, J. D., Xu, Z., Bartels, C. F., Kawaguchi, R., Hu, M., Scacheri, P. C., Rong, Z., Li, Y., Sullivan, P. F., Song, H., Ming, G. L., … Jin, F. (2020). Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. *Molecular cell, 79*(3), 521–534.e15.
70. Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z., & Noma, K. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research, 38*(22), 8164–8177.
71. McKinney W. (2010). Data Structures for Statistical Computing in Python [Web]. In van der Walt & Millman (Eds.), *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-00a