Bryant University

HONORS THESIS

Chromatin Architecture: Mechanisms of Gene Regulation

BY Logan O'Donnell

ADVISOR • Steven Weicksel Ph.D.
EDITORIAL REVIEWER • Kristin Scaplen Ph.D.

Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program
April 2022

Table of Contents

ABSTRACT

The rapid growth and division of cells as they proliferate and penetrate to surrounding tissues defines the collection of disease states known as cancer. Abnormal gene expression drives this uncontrollable replication of cells. In recent studies, aberrant *HOX* gene expression has been noted in a multitude of cancer types such as myeloid leukemia, prostate cancer, and breast cancer [1]. Within cancerous cells with aberrant *HOX* expression, late expressing *HOX* genes are suppressed while early expressing *HOX* genes are reactivated. *HOX* genes are significant in controlling early phases of organismal development such as cell cycle, cell movement, and gene expression. Likewise, *HOX* genes are in later embryonic stages are important for regulation of limb development, body spatial plan formation, and apoptosis processes. As stated above, *HOX* genes that are expressed early in development tend to regulate cell cycle, differentiation, and migratory processes, mechanisms commonly manipulated by cancer. There are various essential factors that determine the regulation of these genes and their expression levels. One aspect that is considered a master regulator is organization of the chromatin containing genes. Specifically, gene expression is dependent on how tightly or loosely they are packaged in chromatin (DNA plus certain proteins). Genes that are expressed tend to be more open (loosely) allowing for regulatory proteins to bind to chromatin sequence driving generation of their encoded product. Moreover, the tightly packed chromatin results in subsequent gene product to not be produced. Many studies have demonstrated that alterations in the spatial architecture of chromatin results in improper regulation in cancer cells. Likewise, aberrant *HOX* genes have been readily present in sample of cancerous cells. Therefore, the hypothesis for the following thesis is that organization of chromatin is a critical regulator for *HOX* genes and mutations in the structure of the chromatin leads to abnormal *HOX* gene expression.

Using zebrafish embryos and cell lines we generated a map of contact points made between chromatin loops within topological associated domains of *hox* genes with the circular chromosome conformation capture (4C) technique (AIM1). These contact points were further analyzed with a comparative genomic profile to other fish species to identify the binding sites for key regulatory proteins (AIM2), which was tested functionally in future experimentation. Future studies can use the significant information collected through sequencing and analysis

techniques in this thesis to expose the connection between *HOX* expression and disease state formation.

INTRODUCTION

Cancer affects millions each year globally, with an incidence rate of 442.4 per 100,000 individuals in the US, leading to 158 deaths per 100,000 Americans yearly [2]. Modern medicine has resulted in a decline in mortality rate of cancer since the 20th century. The addition of early preventative screenings and educational resources have assisted in this downward trend of cancer deaths. Although this decline is a promising sign for advancements in medicine, the majority of the decline has been seen in the absence of new treatment development. New and increasingly complex cases of cancer are still prevalent due to the complicated and specific nature of cancer. Growth of cancers is reliant on the distinct characteristics of the cancer type as well as the unique genomics of the individual. Thus, the development of advanced and effective treatment is necessary in order to tackle cancers in a more specific manner to their origin or development. In order to implement these more effective treatment options, more research in the field of cancer genomics is crucial. There are a multitude of aspects that have thwarted efforts to refine the approach to combating cancer, including the lack of knowledge with regard to the specific mechanisms driving carcinogenesis (the processes of cancer progression). The following thesis explores one such mechanism at a molecular level, focusing on the organization of chromatin (DNA (deoxyribonucleic acid) and protein complexes), and its effects on gene regulation/expression.

Within the nucleus of every cell is the instruction manual, the genome, for all cellular function. DNA is the building block of this cell specific instructions, stored as a code of four different chemical base pairs. In humans, each genome consists of roughly three billion base pairs, resulting in nearly ~2 meters of a DNA which needs to fit in a nucleus roughly 10 um bit, a space roughly the size of a speck of dush. To accomplish this feat, DNA orients in a double stranded structure and takes on a helical shape similar to a twisted ladder. DNA passes genetic information from parent cell to two daughter cells during cellular division by DNA replication. DNA replication encompasses the separation of the DNA strands, priming of the template strand, and construction of the new DNA segment. Following replication is the transcription process of DNA.

Transcription's objective is to create a messenger molecule, ribonucleic acids (RNA), to hold a DNA sequence of interest. Genes code for different features or cellular components of the organism. Genes are read by the cell through the process of transcription, which creates a messenger molecule, ribonucleic acids (RNA). The information from the DNA segment is carried by the RNA copy, or transcript, and contains knowledge on how to properly build a protein for the cell. Transcription is a critical stage in gene expression since it controls which proteins are made and at what rate. Therefore, transcription is a highly regulated step by the cell and is specific to the needs of the cell. In the cell, not every gene is expressed at the same time point. Therefore, the cell must strictly manage the expression levels of the desired genes and repression of others.

In order to obtain the instructional material from the messenger RNA, the sequence has to be translated from the language utilized in DNA to terms of proteins. The sequential order of the instructions develops the framework of the protein. Protein molecules are the cell's "workhorses" and perform all of the processes required for life. Examples of protein's functions include building the body's structural layout, repairing organismal body tissue, and proper coordination of bodily functions.

Additionally, DNA complexes with additional proteins, such as histones, which aid in condensing the instructions into a threadlike structure known as chromatin. The process of condensing or uncoiling chromatin regulates gene expression by controlling whether a mechanism is turned on (expressive) or off (suppressed). Although each cell in every organism has the same DNA, not every gene is expressed. DNA regulation differentiates cells by activating specific sets of genes resulting in differential expression of proteins in order to carry out their prescribed job. The regulation of the expressive levels of a gene is likely due to selective evolutionary processes to conserve energy and space. If all cells were to express every gene at all periods of the day, it would take a quantity of energy that would not reasonably be consumed on a daily basis. Likewise, for expression, the DNA within the gene must be unwound, transcribed, and then translated. The constant uncoiling of the DNA would force cells to be huge and undifferentiated, resulting in a lack of complexity in the organism.

The structural components of chromatin in a cell are necessary to control the proper expression and activity levels of a gene [4]. When alterations occur to the three-dimensional structure of chromatin, misregulation of genes is often a consequence [5]. Misregulated genes, and their corresponding cis-factors and trans-factors, are commonly the root of cancer development in the body [3]. Cis-regulatory elements, such as promoters, enhancers, and silencers, are non-coding DNA sequences that control gene transcription [37]. Trans- regulatory factors are proteins that can ultimately recognize and bind to target sequences of cis-factors to control transcriptional expression levels and subsequent gene expression. In agreement with this, it is seen that in disease states there are problems with trans-regulatory factors and subsequent issues with regulatory processes. These improper modifications of trans-regulatory factors ultimately coincide with the lack of appropriate cell functioning. Utilizing chromosome imagery techniques, the understanding of the underlying principles that link chromosomal organization, trans-regulatory factors, and cis-factors to gene regulation. Furthering the comprehension of the chromosomal structure and subsequent gene regulatory relationships allows for better overall insight on healthy cell development which may act as a comparative model to diseased states.

Supplementary proteins pack the chromatin into loops which are methodically arranged into chromosomes. There is a common misconception that chromosomes bear resemblance to an "X" shape. However, chromosomal imaging techniques, reveal that a more accurate picture of the chromosome's shape is analogous to a spherical blob [41]. The shape of chromosomes changes with age and development, is changed in some illnesses, and can impact gene expression. The genome's organization and three-dimensional structure within the nucleus is dynamic, and conformation changes play a role in gene transcription control. As the chromosomal conformation is altered or modified, it results in subsequent changes of expression. The changes in structure can lead to condensing of genes, effectively turning them off, or loosing of chromatin loops containing genes resulting in expression. The orientation of the chromosomes and the three-dimensional (3D) folding of the eukaryotic genome is a highly ordered process that is closely tied to functional DNA-dependent activities, including DNA replication and transcription [39]. Chromosomes are divided into subregions, defined by chromatin boundaries mediated by cohesin (protein complex) known as topologically associated domains (TADs). Experimentation from chromosome confirmation capture (3C) techniques identify provide snapshots of these domains. Distinct segmentation of looped chromatin by an unlooped or insulator region defines these domains. Genomic sequences build TAD's organizational structure and conserve the same genes across all diverse cell types [6]. These domains contribute to gene control with many genes regulated together and clusters defined by their function located in the same TAD. Although, the capture systems support the notion that TAD's do not contain different genes across cell type, they may be regulated in different ways [6,7]. Ubiquitous factors, such as CTCF and cohesins, conserve TAD's clear boundaries. These ubiquitous factors are highly conserved proteins that can function by activating transcription, specifically a repressor, resulting in the blockage of communication to enhancers [6]. These factors may also recruit additional transcriptional factors to aid in the establishment of chromatin structure. Based on the literature, gene specific factors, unique to the cell's overall function, maintain the loop's organization for proper gene expression. [6]. Active or turned-on genes are more frequently found in the central interior of the nucleus than repressed areas, which are commonly found closer to the perimeter of the nucleus [39] (Figure 1). The repressed areas are pushed and condensed against the periphery, making the information in that section of DNA more difficult to access, thus, subsequently turned off.

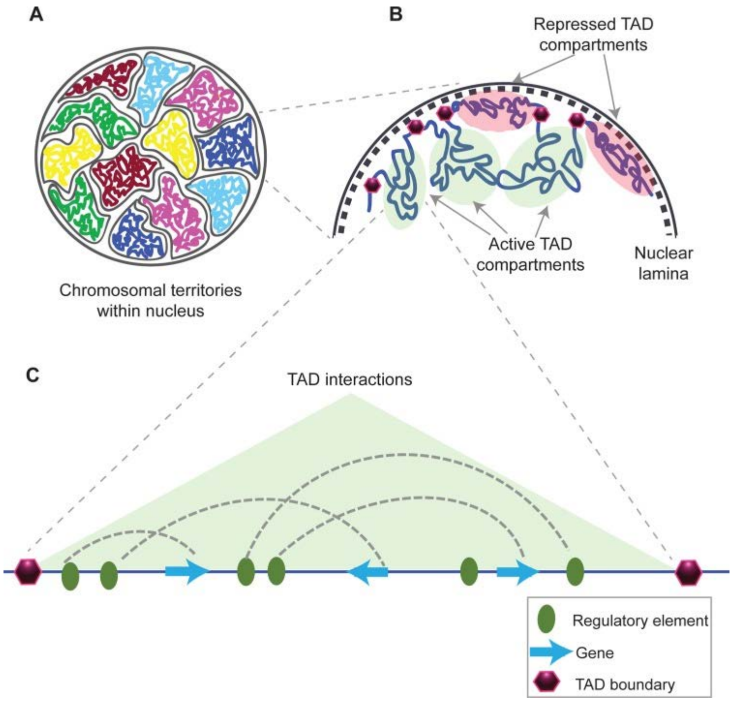

Figure 1. Organization of chromosomes within the nucleus [40]

Furthermore, previous studies illustrate that many cancer types, TAD boundaries have been disrupted resulting in changes in gene expression. In addition, recent transcriptome studies have identified the misregulation *HOX* genes in various cancer types (e.g., prostate, breast, and colorectal), however, as of yet, the role of *HOX* in cancer remains a mystery [44]. The *HOX* gene family is essential for all metazoans and encode transcription factors responsible of the activation and repression of proteins crucial for proper timing of growth and differentiation during organismal development, including functions such as the cell cycle and cellular movement. *HOX* genes function in embryonic and mature cells, both in the undifferentiated and differentiated cellular mechanisms. Undifferentiated *HOX* genes, are present in early and late embryonic development and manage the migration, differentiation, and advancement of cells in growth [8,9]. During the process of differentiation, which comes with maturity of the organism, activation of expression of later *HOX* genes occurs with simultaneous repression of immature *HOX* gene expression. The synchronization of stimulation and suppression of *HOX* genes depend on developmental signals, additional *HOX* encoded proteins, along with all modifiers which that provide their modification by altering chromatin organization. *HOX* genes that behave abnormally and do not follow the typical systematic pathway have been detected in various cancers. In the carcinogenic cells, instead of activation of late expressing *HOX* genes, there is a gain of early expressing *HOX* expression in the matured cell [9]. Subsequently, the early expression *HOX* genes stimulate the functionality of downstream targets meant for early developmental processes. This results in mechanisms that are commonly associated with cell cycle and cell differentiation to be reactivated. Since many of these processes are known to be manipulated in carcinogenesis, it implies the affiliation of *HOX* genes supporting cancer progression. Although these are logical conclusions of connection, there is a gap of knowledge on the exact role of *HOX* genes in cancer development due to their function as master regulators (which will be discussed in the latter part of the literature review) [8]. However, based on the previous observations in *HOX* gene studies displaying alterations of structure of chromatin resulting in changes in expression levels in cancerous cells, there is plausible linkage between chromatin modifications driving cancer growth due to *HOX* gene alteration.

The current information revolving around *HOX* genes exposes that the regulatory factors for this gene type are *HOX* function, ubiquitous insulating factors, developmental signals, and CCCT binding factor factors [10,11]. These components influence the expressive levels of *HOX* genes by modifying the structure of *HOX* gene chromatin, leading to a change from open (active) to closed (repressed) states for further downstream signaling [10,11]. The process of how these regulatory units regulate individual genes is unclear.

In addition, *HOX* expression has been shown to be regulated by chromatin organization [32]. Taken together it is possible that the misregulation of *HOX* expression observed in cancer is due to changes in chromatin organization during the disease state. Interestingly, *HOX* gene expression correlates with the inverse of normal gene expression found in non-diseased cells, meaning, genes normally expressed are repressed and vice-versa. up- or downregulated. For example, *HOX*C family gene expression has also been found to be upregulated in most solid tumor forms, including colon, lung, and prostate cancer [44]. *HOXA9* and *HOXB13* were the two *HOX* genes shown to be most often mutated or altered to some extent in solid tumors [44]. *HOXA* genes, *HOXB* genes, *HOXC* genes, and *HOXD* genes have displayed altered expression in breast and ovarian cancers, colon cancers, prostate and lung cancer, and colon and breast cancers, respectively [44]. Since *HOX* genes are associated with growth and development processes, a direct modification or change in expression levels of this gene family has the potential to result in improper cellular activity, particularly activities favorable for cancer progression. The following project intends to mitigate the limitations in understanding the relationship of chromatin organization and *HOX* gene expression, and consequently cancer development.

This thesis attempts to better understand the mechanisms responsible for controlling chromatin structure and gene regulation. For these experiments, the zebrafish *hox* genes was chosen to be studied as the model. This gene family encodes for a homeobox consisting of transcriptional factors that properly control embryonic development, cell differentiation, and organogenesis (organ formation) for the Metazoan Phyla [8]. Based on previous observations, my hypothesis was that aberrant *HOX* expression in cancerous cells is a result of modification in the three-dimensional structure of the chromatin which controls *HOX* genes. In order to identify the specific systematic pathways of chromatin organization that control expressive levels of *HOX* genes, the zebrafish will be utilized as a model. Zebrafish are an ideal model since they are small, robust, and the embryos can be easily and efficiently collected without harm to the mother. Zebrafish breed readily by producing between 50 to 300 embryos approximately every ten days. Since zebrafish produce such a large quantity of embryos and breed rapidly larger sample size for experimentation can be collected without difficulty. Furthermore, zebrafish embryos are laid and fertilized externally, thus manipulation and harvesting of the embryos can be done ethically without any consequence to the mother. Most importantly, zebrafish are one the simplest eukaryotic models that display genetic similarities to humans resulting in a reliable and predictive model.

The first aim was to identify the *HOX* chromatin structure in a normal cell. The organizational loops of the *HOX* clusters will be mapped though circular chromosome confirmation capture and high-throughput paired end sequencing (4C) in four hour post-fertilization zebrafish embryos. These sequencing results will be analyzed to identify chromatin contacts within the *hox* gene clusters in collaboration with the URI Bioinformatics Core utilizing pipe4C [12]. The identification of these contacts will indicate the location of looped regions in *hox* genes, effectively creating a map of the chromatin organization in the *hox* TADs. This map will aid future experimentation by providing a control, allowing for comparison to reads from different developmental timepoints when specific *HOX* genes are stimulated or repressed. It was hypothesized that as different *HOX* genes are activated, a subsequent change of the chromatin landscape will occur to reflect the regulatory sequences stimulated. In the context of the intent of this thesis, these maps can be used for comparison against those generated from cancer cells to illustrate differences in abnormal *HOX* gene expression.

The second aim for this project will be to determine the cis-regulatory factors that are necessary for proper *HOX* chromatin arrangement. The hypothesis was that the contact points of AIM1 will pinpoint the cis-regulatory sequences inside the loops resulting in chromatin interactions. The 4C maps will be inspected for cis-regulatory sequence motifs by employing TFBSTools packed with MEME-ChIP found in the URI Bioinformatics Core [14,15]. With the site of the sequences determined, future experiments can make use of the loss or gain-of-function approaches to conclude the function of trans-factors in *HOX* gene stimulation as they relate to cis-regulators [15]. These observations of interaction and sequential analysis will be significant in the comprehension of *HOX* activation and the relation it has to cancer in future studies.

RESEARCH QUESTIONS

(1) What is the organization of chromatin inside the *HOX* gene?

   a. Where are the chromatin interactions within *HOX* TADs?
   b. How are the loops arranged in the chromatin?
   c. Is there a systematic coordination of specific components in the gene based on function?
   d. What components regulate the proper structure formation?

(2) Where are the cis-regulatory sites located?

   a. Are these elements essentially for development of *HOX* chromatin structure?
   b. Are trans-factors bound to the cis-regulated sequences?
   c. Do the trans and cis elements have a major contribution to structure development and subsequent expression?

LITERATURE REVIEW

The literature that revolves around the field of chromosomal research are commonly split into two sections of thought: rigid factual knowledge of the architecture and the connection between this framework and human disease states. Therefore, to understand the relationship between the two approaches of information, the following literature review discusses the background information relevant to comprehend the processes that lead to human disease states, specifically cancer as it relates to the thesis, prior to mutations that can ultimately result in biological disorders.

Historical Research of the Genome

The human body must scale down approximately two meters of genomic material into a micrometer nucleus while maintaining access for transcription and replication. Insight into the mechanisms that dictate the organization and usage of the genome is an area of continuous research in the field of genetics [16]. Over the course of fifty years, analysts have put forth a concentrated effort to the creation of procedures and innovations to facilitate sequencing DNA and RNA [17]. The genome contains all information necessary for human life to grow, develop, and adapt to environmental surroundings, therefore, being of the utmost importance for health officials to fully comprehend. Sequential understanding of the human genome is significant; however, it is just a starting point for the molecular genetic field. Comprehensive analysis of the human genome not only provides insights for normal physiological development, but it also serves as a model for aberrant growths such as cancer. The next step for the health sector to further investigate is cancer genomics. Improved and expanded knowledge on cancer genomics by classifying cancer types and subtypes based on their genetics, helps to advance precision medicine. Extensive understanding of molecular cancer genomics can aid patients with a more exact diagnosis and, as a result, a more tailored treatment plan and information for likelihood of diagnosis in family members. Improvements for more extensive cancer prevention and treatment plans start with genomic sequencing. Genomic sequencing and testing can provide identification of inherited DNA mutations that may increase an individual's risk of cancer.

The technology of modern genomic sequencing of DNA began in the late 1960's when Ray Wu and Dale Kaiser manipulated the traditional mechanism utilized for RNA sequencing in order to partially map the DNA molecules [17]. The incomplete or partial mapping was due to the fact that DNA is significantly longer than RNA molecules. In a human chromosome, a DNA molecule may be up to 250 million nucleotide pairs long, however, most RNAs are just a few thousand nucleotide pairs long, and many are much less. Wu and colleagues initiated their research by sequential analysis of the cohesive ends of λ phage DNA from an Enterobacteria [17]. The molecular biologists attached radiolabeled primers and DNA polymerase to the sequence, allowing for small fragments of marked DNA to be produced. The substituent was then analyzed with ionophoresis and a two-dimensional electrophoresis method [17]. The sequencing of the twelve nucleotides from the linear bacteriophage's DNA resulted in an abundance of curiosity in the mapping of the entire DNA strand. In 1977, Frederick Sanger and his team intertwined Kaiser and Wu's processes along with the present data of the time to develop a technique which sequenced a full virus genome [17]. Utilizing an octanucleotide as a primer, DNA polymerase I, and radioactive nucleoside triphosphates, the team of biologists produced a significant length of DNA from a filamentous bacteriophage to start testing sequential methods [18]. The relatively intricate "plus and minus" process that Sanger established resulted in over 5,000 nucleotides synthesized, which marked a breakthrough in the scientific community [18]. Around the same year Sanger published his work on the filamentous bacteriophage, Americans Walter Gilbert and Allan Maxam constructed their own methodology for DNA sequencing. The Sanger and Maxam-Gilbert procedures seemingly overlap with similar systematic manipulation of the DNA. However, the Maxam-Gilbert sequential method altered the DNA with base specific chemical treatment and subsequent DNA cleavage [17]. The DNA is run on a polyacrylamide gel and the length of separated pieces (and hence location of explicit nucleotides) are distinguished and imply a specified sequence [17].

More recently, the National Human Genome Research Institute (NHGRI) established the communal foundation for gene sequencing and the overall exploration of human genomics from the 1980s to the early 2000s. Through countless experiments, NHGRI investigators intended to sequence the 3 billion DNA letters that construct the complete set of genetic instructions for human life [16]. Originally, the institute funded the Human Genome Project with the general mission of identifying the genes and DNA subunits the genome consisted of. The program began its genomic analysis in 1990 and developed its first linkage map of the genome four years later, which included subsequent data on DNA patterns on chromosomes [17]. In recent years the association has moved towards conducting studies to better understand the genomic correlation to human disease and disorders. The spatial conformation of the genome is directly connected to its purpose and integral role in the body. However, the current understanding of the complexity of genomic structure is unrefined and fragmented [19]. In order to interpret the causation of mutation, and consequently disease, behind the sequences, the majority of research has switched from the focus on the nucleotide sequence itself to the architecture of the chromatin within the genome.

Chromatin

The genome incorporates numerous subunits of chromosome regions, compartments, and topological domains, which are regularly separated by compositional proteins like CTCF, cohesin, and chromatin circles. In eukaryotic cells, DNA is bundled into chromosomes, which are composed of the nucleoprotein complex called chromatin [19, 21]. As previously discussed, the chromatin contains all of the instructions and functional components for all mechanisms to occur within an organism. The genome has specifically packed the instructional sequences in order to control regulation of expression and efficiency of the cells. Further, as previously discussed, the connection between the trans- factors and cis- regulators is essential. If these two elements cannot readily find each other within the genome, the regulatory role they provide would not be performed [15]. Thus, the architecture of the chromatin that places these factors and regulators close in proximity for an efficient cascade is important for cellular life. Chromatin works with associated protein molecules such as histones, DNA-binding factors (DBFs), and the basal transcription machinery, along with RNA to collectively allow a cell to properly function [20]. The transcription factor CCCTC-binding factor (CTCF) plays a significant role in ensuring the correct folding and structure of the chromatin. Likewise, CTCF's interaction with cohesin is imperative for boundary and loop development in TADs [33].

The fundamental proteins in chromatin are the histones, which ensure tight packaging of all essential elements for DNA. The most basic unit of chromatin is the nucleosome, constructed of histones with less than two complete rounds (160 base pairs) of DNA bound to protein [22]. The organization of the DNA into nucleosomes either produces a closed, heterochromatin, or open, euchromatin, structure. The structure of heterochromatin, found in the nuclear periphery, is considered transcriptionally repressive, permitting a basal expression level to be presented for a gene [22]. Contrarily, euchromatin architecture of chromatin results in advanced accessibility to specific transcription and replication proteins, thus being located more interiorly in the nucleus [21,22].

The significance of chromatin is to act as a controller of the expression in genes and the extent to which it transcribes the encoded product [21]. For instance, higher order chromatin organization is affiliated with long distance gene regulation that influences cell development and death [21]. Moreover, the level of condensation of chromatin is imperative to the segregation of the chromosome in the stages of mitosis and meiosis. Mutations and defects that occur in these higher order chromatin sequences may ultimately contribute to biological abnormalities and likelihood of disease in the human [20,21].

In recent years, chromosome conformation capture family (3C) techniques (e.g., Hi-C, 4C, and 5C) were developed in order to reveal the key features of the chromatin organization and how they relate to the entire three-dimensional structure within the cell [23]. Specifically, these methods identified general principles of structure and it has become apparent that chromosomes are further divided into specific compartments known as topologically associating domains (see topologically associating domains section).

Gene Regulation

Gene expression in eukaryotic cells is regulated by repressors as well as by transcriptional activators. Like their prokaryotic counterparts, eukaryotic repressors bind to specific DNA sequences and inhibit transcription [27]. In some cases, eukaryotic repressors simply interfere with the binding of other transcription factors to DNA. For example, the binding of a repressor near the transcription start site can block the interaction of RNA polymerase or general transcription factors with the promoter [27]. Other repressors compete with activators for binding to specific regulatory sequences and inhibit transcription. These repressors contain the same DNA-binding domain as the activator but lack its activation domain thus their binding to a promoter or enhancer blocks the binding of the activator, thereby inhibiting transcription [22]. In contrast to repressors that simply interfere with activator binding, many repressors (called active repressors) contain specific functional domains that inhibit transcription via protein-protein interactions. The first active repressor was described in 1990 during studies of a gene called Krüppel, which is involved in embryonic development in *Drosophila* (the common fruit fly) [27]. Molecular analysis of the Krüppel protein demonstrated that it contains a discrete repression domain, which is linked to a zinc finger DNA-binding domain [27]. The Krüppel repression domain could be interchanged with distinct DNA-binding domains of other transcription factors [27]. These hybrid molecules also repressed transcription, indicating that the Krüppel repression domain inhibits transcription via protein-protein interactions, irrespective of its site of DNA binding [27].

Many active repressors have since been found to play key roles in the regulation of transcription in animal cells, in many cases serving as critical regulators of cell growth and differentiation. As with transcriptional activators, several distinct types of repression domains have been identified. For example, the repression domain of Krüppel is rich in alanine residues, whereas other repression domains are rich in proline or acidic residues [27]. The functional targets of repressors are also diverse. Some repressors inhibit transcription by interacting with general transcription factors, such as TFIID; others are thought to interact with specific activator proteins [21].

The regulation of transcription by repressors as well as by activators considerably extends the range of mechanisms that control the expression of eukaryotic genes [20]. One important role of repressors may be to inhibit the expression of tissue-specific genes in inappropriate cell types. For example, as noted earlier, a repressor-binding site in the immunoglobulin enhancer is thought to contribute to its tissue-specific expression by suppressing transcription in nonlymphoid cell types [27]. Other repressors play key roles in the control of cell proliferation and differentiation in response to hormones and growth factors. The relationship between chromatin structure and transcription is evident at several levels. First, actively transcribed genes are found in decondensed chromatin, corresponding to the extended 10-nm chromatin fibers [27]. For example, microscopic visualization of the polytene chromosomes of *Drosophila* indicates that regions of the genome that are actively engaged in RNA synthesis correspond to decondensed chromosome regions [27]. Similarly, actively transcribed genes in vertebrate cells are present in a decondensed fraction of chromatin that is more accessible to transcription factors than the rest of the genome [23]. The relationship between the architecture and gene regulation was be explored in great depth beyond the review of the literature in the experimentation of the following thesis.

Topologically Associating Domains (TADs)

Understanding the systems that underlie chromosome folding inside the nucleus is fundamental to concluding the connection between genome design and functionality. Within the past decade, the chromosome conformation capture strategies have uncovered that the genome of numerous species is coordinated into areas of preferential internal chromatin sequences referred to as topologically associating domains (TADs) [10]. These domains are characterized by coordinated portions of circled looped chromatin isolated from distinct bundles of circled looped chromatin by an unlooped limit of chromatin [24]. TADs are dependent on genomic sequence and conserve the same genes from one cell to another [24]. Although TADs do not contain different genes, the regulation mechanism act differently and subsequently differentiate cell types [23]. The boundaries correspond to regulation; thus, genes are coregulated within cell differentiation if present in the same TAD [23]. Furthermore, TADs or contact domains can differ in size, chromatin features, and mechanisms underlying their formation. This suggests that TADs might be subdivided into different subtypes, each of them characterized by specific structural and functional properties [22,23].

The importance of TADs appears to be to provide another layer of gene control for the cell. Genes within TADs tend to be regulated together, and many gene clusters, such as some *HOX* genes, are found in the same TAD [9]. Maintenance of the loops within TADs appear to be regulated by gene specific factors, such as transcription factors and vary from cell type to cell type depending on gene expression (Figure 2).

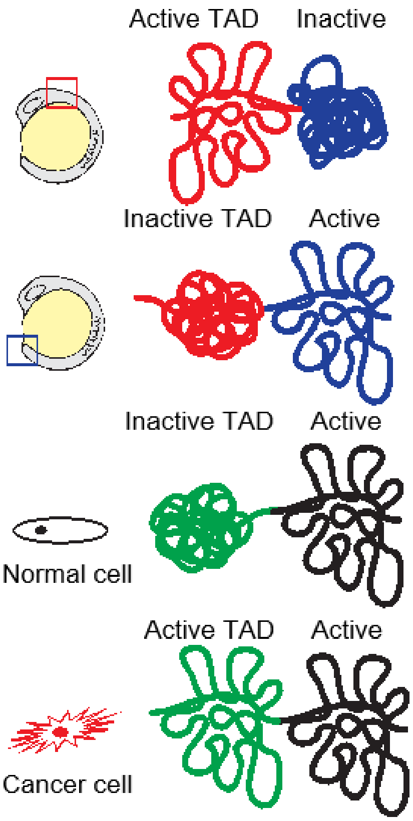

Figure 2. TAD regulation in different tissue types (red/blue) and healthy and cancer cells (green/black)

HOX Genes

A specifically important group of genes in the scope of this thesis are the *HOX* genes. The *HOX* gene family is highly conserved in evolution and determine the development of many internal structures of diverse animal species [9]. At a molecular level, *HOX* genes are often organized in TADs depending on their expression and regulatory constraints. Initial studies illustrated the importance of *HOX* genes in defining boundaries and cellular territories within an organism [8]. However, it is understood that the genes are functional components of organ development due to their role in regulation of differentiation, apoptosis, reproduction, and transcription [8]. In early embryonic development, *HOX* genes have been observed prior to differentiation and formation of germ layers to tissue layers and are suggested to control major migratory and timing properties of cells [8].

HOX genes code for proteins that regulate expression levels by binding to the DNA directly with the aid of cofactors and collaborators [9]. In mechanisms to *HOX* genes, the coexpression from the additional cofactors and collaborators results in the increase of affinity and specificity for targets further downstream in mechanisms to *HOX* genes [9]. In animal studies, TALE class of homeoproteins, encoded by Pbx and Meis gene, are most often found associated with the *HOX* genes in processes. The genes were originally investigated in *Drosophila* and mutations in the domain resulted in homeosis as entire structures were altered to resemble that of another in the body [9]. For example, when the *HOX* gene *Antennapedia* mutated, the functional aspects of antennae were transformed into additional legs [8,9]. These studies, along with those performed in mouse models, implicated that the *HOX* genes most likely are responsible for cellular function specifically associated with proper maturation and spatial body plan development. The theory was subsequently extended to animals where it was found that regardless of the divergence in morphologies, there was a consistent reliance on this genetic system to correctly form the spatial body composition [32]. Thus, the modern hypothesis for the *HOX*_genes proposes that by activation and regulation of a series of targets in downstream mechanisms, by use of realizator genes, *HOX* genes influence the organizational factors for organ growth [9].

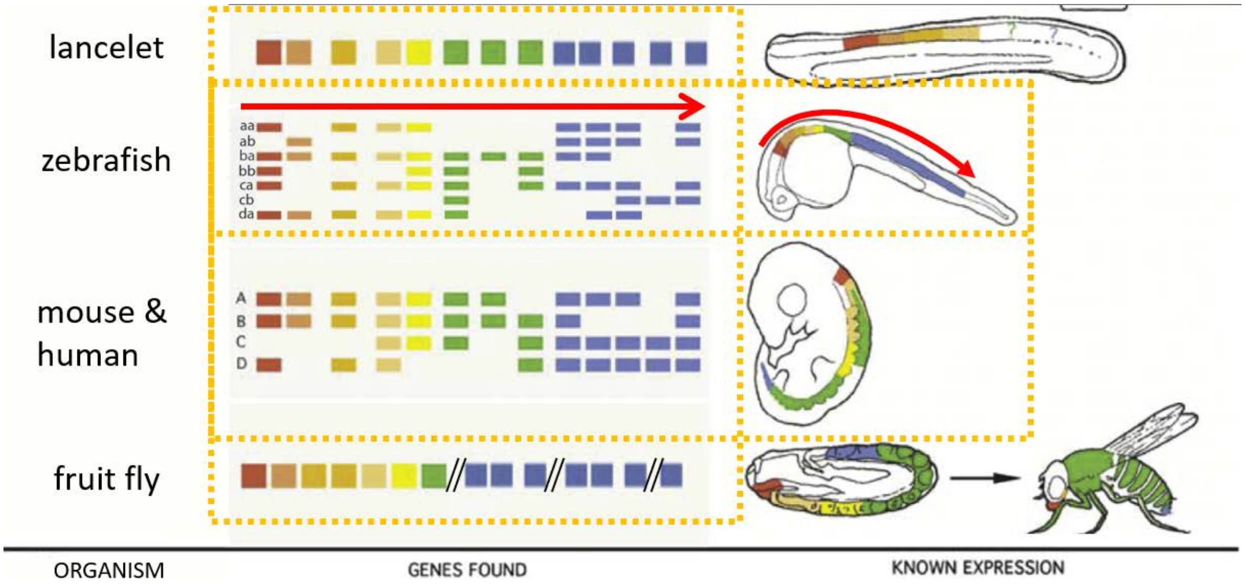

Figure 3. Evolutionary conservation of *HOX* genes in organismal development

Studies have also suggested that the significance of *HOX* genes in organismal development, as seen in Figure 3, may likely be the reason why they are apparent in both undifferentiated and differentiated cell types. However, as maturation occurs in the species, early expression of undifferentiated signals is extinguished in order to switch to late differentiated *HOX* gene expression. The maintenance of this key coordination of signaling patterns is dependent on the chromatin organization of chromatin modifiers, developmental signals (e.g. FGF and retinoic acid), and transcriptional repressors [8,9]. Most of these processes include the manipulation of trans- factors interacting with cis-regulators inside of the *HOX* sequence to proceed in regulation of the gene itself. Data asserts the notion that these supplementary tools used in conjunction with *HOX* genes are crucial for mechanisms such as opening, closing, and controlling the spatial and temporal gene clusters [9].

Although local modifications are essential for proper expression of *HOX* genes to occur, much of the more recent literature in the field suggests that the higher order chromatin architecture may oversee much of the transcriptional control [8,9,33]. The changes in the compartmentalization of the *HOX* clusters were first discovered microscopically by examining loci within the nuclei with fluorescent *in situ* hybridization (FISH). The observation exposed that *HOX* genes were less dense in regions of activation after the interaction with retinoic acid. In tissues where early embryonic *HOX* genes become silent, the clusters developed distinct three-dimensional structures defined by various gene loci [32]. In opposition, the active genes were clustered in a separate region of the chromatin structure with a different structure [32]. The spatial segregation of the *HOX* genes dependent on activation level appears to be substantial for appropriate advancement in a species maturity. The correlation between function and spatial location of the *HOX* genes remains to be established in the current research literature. Whether the subsequent organization of the *HOX* genes can be connected to mutations such as cancer and human disorders was explored in the data collection of this thesis.

Cancer

Cancer Development

Cancer is defined as uncontrolled cell development and procurement of metastatic properties [25]. Commonly, stimulation of oncogenes and additional deactivation of tumor suppressors corresponds to uncontrolled cell cycle development and inactivation of apoptotic systems. Unlike benign tumors, dangerous malignant growths can metastasize, which happens to some degree because of the downregulation of cell adhesion receptors, essential for tissue-explicit cell–cell connection, and up-regulation of receptors or activation of metalloproteases that improve cell mobility [25].

In normal cells, the products of proto-oncogenes act at different levels along the pathways that stimulate cell proliferation [25]. Mutated versions of proto-oncogenes or oncogenes can promote tumor growth. Inactivation of tumor suppressor genes like pRb and p53 results in dysfunction of proteins that normally inhibit cell cycle progression [26]. Cell cycle deregulation associated with cancer occurs through mutation of proteins important at different levels of the cell cycle. In cancer, mutations have been observed in genes encoding CDK, cyclins, CDK-activating enzymes, CKI, CDK substrates, and checkpoint proteins.

There are various processes by which these mutations and dysregulations are procured. Genetically, the modification in the DNA can be random due to chromosomal deletion or translocation, dysregulated gene expression or regulation, or preemptive intercellular signaling [25]. These occurrences may stimulate the genetic code that advances dysregulated cell cycling and can also inactivate apoptotic pathways.

The process of searching for new cancer drugs has undergone a major change: it has moved from a strategy identifying drugs that kill tumor cells towards a more mechanistic strategy acting on molecular targets that underly cell transformation [26]. The evidence that CDK, their regulators and substrates are targets of genetic alteration in different types of human cancer has stimulated the search for chemical CDK inhibitors [26]. Different strategies for therapeutic intervention can modulate CDK activity: targeting the major regulators of CDK activity (indirect strategy) or inhibiting the catalytic activity of the CDK kinases (direct strategy). Approaches for the indirect strategy include overexpression of CKI, synthesis of peptides mimicking the effects of CKI, decrease of cyclin levels, modulation of the proteasomal machinery, modulation of the phosphorylated state of CDK and of the enzymes regulating it [26]. The literature implies that more information about the structure and mechanism of the DNA from the unregulated cells that are encoded may aid in the production of corresponding therapies that may turn off their uncontrolled features.

Chromatin Mutations Associated with Cancer

Cancer can take advantage of all the critical information for cellular processes being centrally located and drive healthy cells to a disease state. The TAD boundaries and regulation have been observed as mutated in malignant cancerous growths [31]. Comparatively, the modifications cancers can cause to the chromatin lead to detrimental effects in regulation and expression throughout the organism. Therefore, the altered state of the chromatin structure resulting in dysregulation implies the relevance of the architecture in regard to conventional functionality of a cell. Furthermore, studies have shown that aberrant *HOX* genes are found in cancerous cells. *HOX* genes draw specific interest to pinpoint the understanding of their application to regulation because this gene family is known to be controlled by chromatin structure. The current hypothesis revolving around the specific connections between abnormal *HOX*, and its chromatin structure are still unknown in the cancer field, however, many presume the relationship by be moderated by TADs.

In recent studies the expression of *HOX* genes in specific TADs have been associated with dysregulation. It has been observed that embryonically expressed *HOX* genes consistently turn back on with loss of the later expressed *HOX* genes in cancer infected tissues [31]. This connection proposes that the addition of the early expressing *HOX* genes results in an activation of functionality to the downstream targets. The major features of *HOX* genes in undifferentiated embryonic cells are to ensure differentiation, proper progression in the cell cycle, and cell migration [31]. Considering that these are the exact genes being reactivated in cancer invaded cells, it suggests the significance of the genes in promoting the disease state mechanisms. The literature clearly illustrates the lack of clarity of the exact cellular role of *HOX* genes in cancer development due to the interconnection and overlapping processes that cancer is known to attack [28, 31]. With the prior observational studies of *HOX* genes' abundance in mutated tissues and the known data of the importance of chromatin organization ensuing manipulated gene expression in cancer cells, a linkage between the two are hypothesized in this thesis.

Theoretically, it may be the chromatin changing structure and thus modifying the embryonic *HOX* expression that allows for carcinogenesis to proceed without detection.

However, there is still much to learn about the regulation of specific TADs and gene groups by chromatin organization. Furthermore, there are a multitude of unknowns regarding three-dimensional structure within the TADS of *HOX* genes and the regulatory factors that maintain the correct orientation of loops within TADs. In the following thesis, in an attempt to fill the gaps of knowledge in cancer development, the three-dimensional structures of chromatin and subsequent TADS found in the *HOX* genes were mapped and analyzed to act as a control model for healthy cells.

METHODOLOGY

The proposed mechanism for the following research utilizes three major steps in order to identify the specific structure of *HOX* chromatin with a particular interest in the TAD regions. In this study, zebrafish embryos were exploited to recreate a similar cellular environment. Utilizing a quantitative polymerase chain reaction (qPCR) was the first major procedure to conduct, which amplified the desired DNA sequence by quantifying the nucleic acids [34]. According to preliminary research using the 3C mechanism, running the same fragments in through a 4C process should increase the knowledge of the structural looping interactions in the *HOX* gene regions. Thus, the next step in the process would include creating a 4C library on the *HOX* gene clusters. Subsequently, after library development, a collaboration with the URI Genomics and Sequencing Center allowed for complete sequencing of the samples with pipe4C pipeline. The mappable reads of the fragmented genome can then be analyzed using peakC to identify and pinpoint loops and significant chromatin interaction. Once the protocol was validated, three replications of the process occurred, lessening the uncertainties involved in the process and ensuring reproducibility.

By taking advantage of the lack of development in the embryos and the similarity zebrafish genetically have to humans, the zebrafish model allowed for results that are applicable and meaningful for this experiment. As previously mentioned, usage of zebrafish embryos directly should provide a system that is still reliant on the natural inputs and signals, reproducing unmodified genomic conditions. Furthermore, zebrafish embryos are easily collected for lab application needs, thus serving as a reliable and stable source of cells for manipulation. Based on previous studies, three biological replicates of four-hour post fertilization (hpf) zebrafish embryos were the subjects of the study due to the limited contamination from specific regulated cells [41].

The 4C process required the zebrafish cells to produce an environment that was uniform and resembles the organization of the genome. This homologous sample was best seen at four-hours post fertilization of the embryos. In earlier embryonic stages (prior to four hours), cells are rapidly dividing and display disorganized levels of chromatin with no gene regulation and minimal transcription [41]. Moreover, later timepoints, closer to six or nine hours, exhibit highly differentiated and mature cells as developmentally important *HOX* genes are condensed and repressed, making the *HOX* genes difficult to access. During the four-hour post fertilization timepoint, the embryos are in a premature or naive state, but genes begin to be grouped off in TADS and expressed. Likewise, at this time point, there is a large and increasing cell to embryo quantity ratio, and the cells within the embryos are all relatively in the same developmental period. Thus, the cell population at four-hpf balances having an extensive sample of cells and inclusion of chromatin structure [41].

The proceeding sections will further explain the series of measures and techniques that were implemented for this thesis and are based on literature recommendations along with information from the lead advisor, Professor Weicksel, of how to conduct the sequential analysis.

Creation of 4C Libraries

Many molecular techniques of chromatin rely upon the nuclear ligation of the chromosome as seen in the family of chromosome conformation capture methods (3C) [34]. The chromosome capture methods produce snapshots of the population-average interaction frequency between two loci [34]. The interactions are dependent on the spatial orientation and proximity of the loops in the chromatin. The 3C family in addition to derived analogous techniques are contingent on four identical biochemical steps [30]. First, to preserve loop formation, crosslink of the nucleus with a zero-crosslinker (formaldehyde) occurs. Chromatin fragmentation follows the linkage by way of a highly specific restrictive enzymatic digestion process. The fragmented portions of the chromatin are re-ligated in a dilute solution to ensure only the digested ends are ligated. The solution was intended to prefer intra-molecular interactions instead of inter- based interactions [34]. The fragments are consequently purified and using a PCR, the chimeric ligated products can be distinguished. The reads are then compared to the reference genome for analysis. The re-ligated strands are the most essential step in the 3C procedure for it places chromatin fragments found in close proximity in the three-dimensional structure directly adjacent.

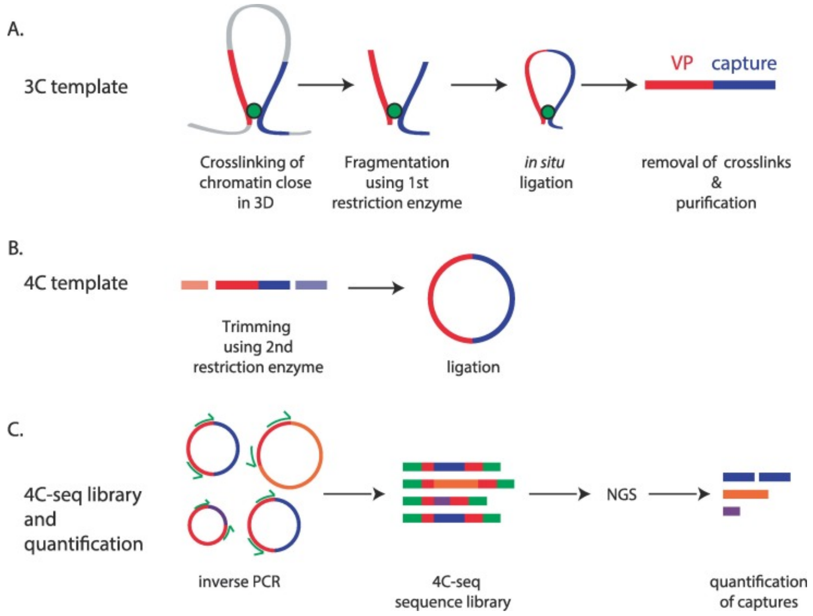
A.
3C template
VP
capture
Crosslinking of chromatin close in 3D
Fragmentation using 1st restriction enzyme
in situ ligation
removal of crosslinks & purification
B.
4C template
Trimming using 2nd restriction enzyme
ligation
C.
4C-seq library and quantification
inverse PCR
4C-seq sequence library
NGS
quantification of captures

Figure 4. Outline of the 3C, 4C, and 4C-seq experimental mechanisms [35]

Employing the circular chromosome confirmation capture (4C), the distinct interactions of the *HOX* gene clusters were be mapped. Utilizing circular chromosome confirmation capture allows for validation or rejection of hypotheses on the factors involving loop formation within in TADs for seven zebrafish *HOX* gene clusters. Exploiting primers unique to the *HOX* genes (known as "bait"), linear fragmentation transpires from the circular constructs.

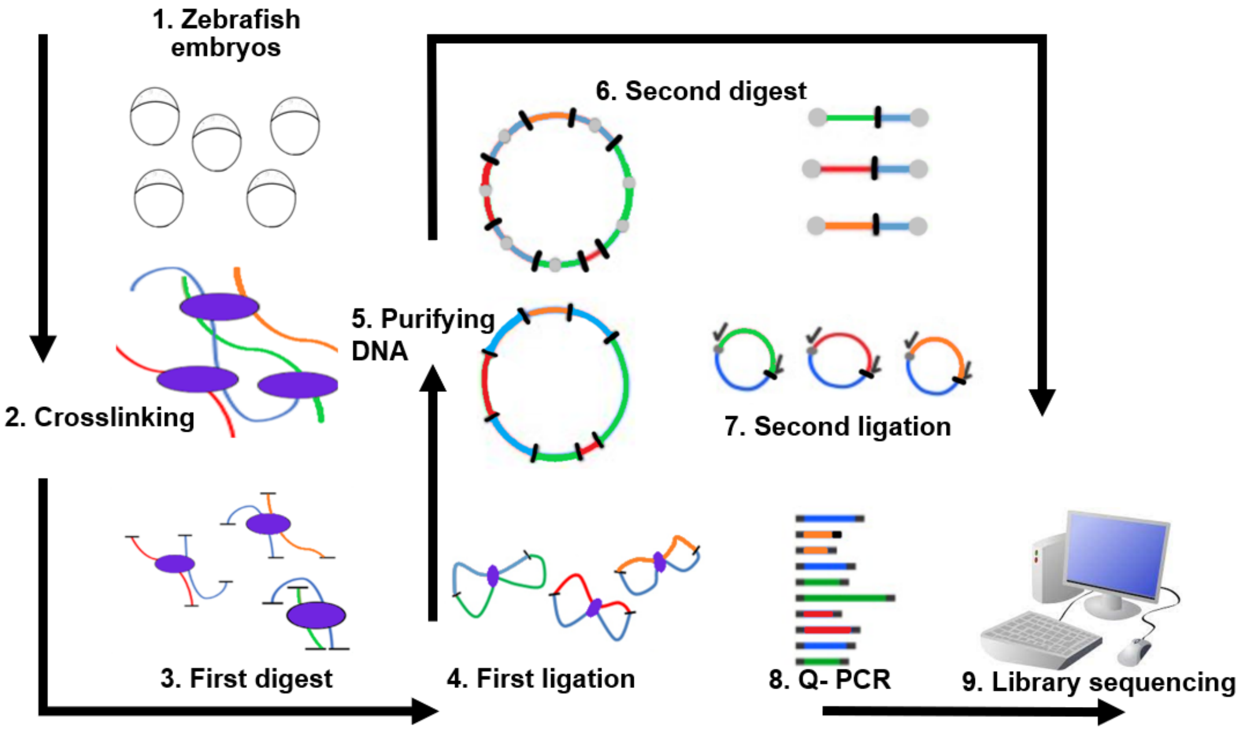

Figure 5. Methodology of 4C process

These fragments develop due to the amplification of the *HOX* gene sequence and the captured sequence from the opposite side of the established structure, as seen in step seven of the 4C diagram (Figure 5). This generates the library with caught sequential data that are located in close spatial space to the *HOX* bait sequence. The advantages of 4C over other fragmentation library collecting was the lack of bias in identification of *HOX* gene regions. This technology can detect the intra- and inter- interaction and connections of chromatin fragments with high levels of resolution [34]. Therefore, this explicit sequencing and analysis is becoming more widely used in studies regarding the correlation between genes and disease states [34].

There were two major controls for this laboratory experiment. One of which was a completely digested and subsequently ligated genomic DNA sample that is representative of the entirety of possibilities for chromatin interactions. Thus, the uncross-linked DNA acted as the positive control and was manipulated in a fashion that is known to produce a result comparatively to the negative control, which was not expected to change due to any variable. The negative control was a fully digested genomic DNA sample that did not proceed through the ligation process, representing zero chromatin interactions. Only ligation of a small division of the available fragment sample for close fragment ligation was induced by using a low concentration of ligase in the model. This differs entirely from the positive control, which consists of high concentrations of ligase because the intention is to ligate all the control sample.

Preceding the sequencing of the 4C libraries, validation of the libraries was be done through quantitative PCR (qPCR). This chain reaction tool evaluates the fragmented library, along with positive and negative controls by identifying interaction points. To perform the PCR process, DNA was combined with the four bases of DNA, DNA polymerase, and primers. The mixture was then heated, which separates the DNA and allows it to be copied. Following the heating step, the DNA sample was allowed to cool, thus permitting primers to connect to the corresponding locations on the DNA. The enzyme DNA polymerase binded to the primer and began copying the DNA in small pieces or genes to create the new DNA strand. This method was performed numerous times (25-35 cycles) to ensure proper amplification of the desired DNA segments. In short, one primer binded to the bait *HOX* sequential pattern, while the secondary primer attached to the capture sequence separated by the primary restriction site [29]. The results from the qPCR measure the signals of the negative control with respect to the positive control signals.

Dr. Weicksel's preliminary research with the 3C model indicates that the few *hox* primers ran created 3C libraries and analysis properly recognize *hox* interactions (Figure 6.). In Figure 6., the blue bar displays the control for the qPCR interactions, therefore, when the orange bar is equal or similar to the control demonstrates that those two *hox* genes do not interact or have a point of contact. However, when the orange bar is vastly greater than the blue control bar, it is indicated that there is a contact point between loops of the two designated *hox* genes. This exploratory data is illustrated in Figure 6, which exposes the potential for loop formation between *hox* sites *hoxb1a- hoxb4a, hoxb1a-hoxb13a,* and *hoxb2a-hoxb4a* (Figure 6). Due to the promising results from the 3C libraries, the next progression of the experimentation included increasing the quantity of bait primers and repeat the process on a 4C library.

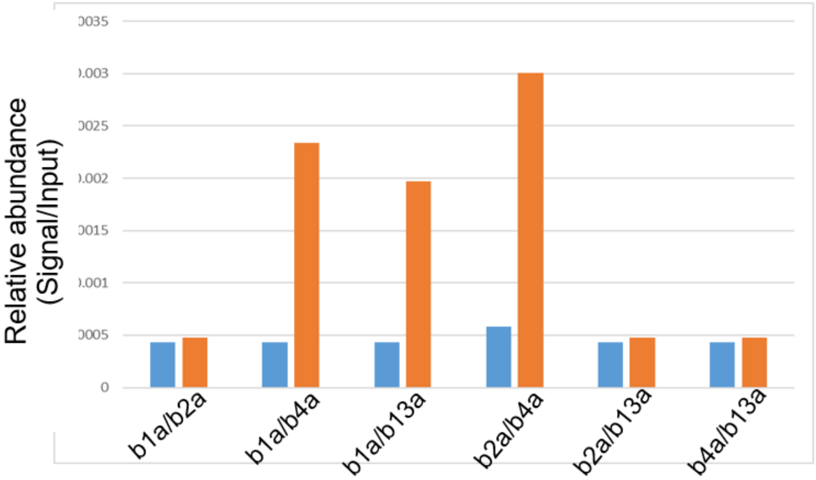

Figure 6. qPCR of *hox-hox* 3C interaction of 4-hour post fertilization zebrafish embryos (orange) or negative genomic DNA control (blue). Units relative to positive control. N=1

Sequential Analysis of 4C Libraries

The URI Genomics and Sequencing Center preformed the majority of the next generation sequencing analytical techniques from the fragmented libraries. The facility preformed the necessary library preparation, quality control, and sequencing of the 4C samples. The library of sequences was generated through the MiSeq 600-cycle reagent kit based upon previously specified instructions. To reduce the redundancy of the sample (from the bait sequences), a phiX DNA spike was introduced to the sample [36]. Once passed through quality control, sequencing of the samples was complete on a singular lane on an Illumina MiSeq system. Following, the analysis of the produced data occurred through direct collaboration with Dr. Chris Hemme from URI.

Together, Dr. Hemme and Dr. Weicksel have developed the pipeline system to prepare for the evaluation of the significance of the sequential maps produced. The pipeline software, specifically pipe4C pipeline, measured the complex 4C samples from FASTQ files existing in the R/Bioconductor environment [35]. The modified raw FASTQ files constituted the readable reference genome (danRer11) with aid from Bowtie2, which discarded the unmappable regions. In silico, the fragmented end library was generated from the reference genome to eliminate sequences that are parallel to the reference genome. The reads that correspond to the fragment end (restriction enzyme end) library are notable and are retained for further examination. The three biological replicate samples are normalized and smoothed in order to produce a total of 1 million mappable reads for analysis.

The pipe4C add on, peakC, was utilized to identify potential loop formation based upon the peaks representing the points of contact in the chromatin. The non-parametric statistics define the significant peaks, which are dependent on the sequential coverage of the 4C fragment samples relative to the background model. The background model was constructed by examining the upstream and downstream reads to the called peaks across all replicated biological samples. Pipe4c allows for visualization of the major peaks, although additional resources such as R-tools or the UCSC genome browser would be equally as successful.

In this project it was essential to recognize the problems in the previously suggested 4C protocol and develop resolutions. This responsibility encompassed care and cultivation of the zebrafish, collection of embryos, evaluating different restriction enzymes, and corroboration of proper primers for the capture process of 4C. Based on the replications performed of the established procedure, analysis of the data and generation of the background interactions was composed.

Identification of Cis- Regulatory Elements

The information from the 4C sequencing and mapping provide the important sequences of the loops and contact points. Once the sequences are recognized, the subsequent proteins that bind to these sequences and mechanisms which regulate these proteins were be disclosed. When validating the structure to function relationship, it was presumed that within or near contact points there is a presence of cis-regulatory sequences that bonded to trans-factors.

In coincidence with Dr. Hemme, the chromatin maps from the 4C samples were further investigated, explicitly through a 1kb window centered around the called peaks [45]. TFBSTools, an application within MEME-ChIP, interpret and subsequently distinguish the transcription factor binding sites from the rest. This software searched for motifs from supplemental databases or *de novo* (from a new set of processes). The TFBSTools include the JASPAR database, which consists of curated and non-redundant transcription factor (TF) binding profiles (JASPAR 2020). The data index was utilized to search the sequenced produced in the 4C mapping for known cis-elemental motifs. Furthermore, the *de novo* feature determined parallel series of sequences to identify clearly defined patterns within the windows. Thus, the noteworthy repetitive sequences can be tested to a further extent in future projects for their influence on the organization of *HOX* genes and the possible linkage to cancer development.

Alongside Dr. Weicksel and Dr. Hemme, development of the data analysis pipeline in addition to administering the protocol for the diagnostic tool was performed. Accountability was held for designing the pipeline from the multitude of software previously mentioned along with fine tuning the operation in the computational context. The standard for the search results included the recognition of the known transcriptional factors that regulate the expression of the *HOX* gene. These factors are mentioned amongst the review of the literature and are as follows: *HOX* proteins, MEIS, PBX, CTFC, and retinoic acid receptors. All of these transcriptional factors have definite DNA binding sites, and the identification would validate that the experimental approach was functioning correctly.

RESULTS

Application of gel electrophoresis allowed for visualization of the 4C library once fragments were amplified by the qPCR process. The electrophoresis process verifies and tests if the 4C library was constructed properly. An electrical field drove the sample through the agarose gel containing small pores. The gel electrophoresis process separated the 4C DNA sample based on molecular size/length. Since DNA is negatively charged, the sample migrated to the positively charged electrode when the electric current was supplied to the gel. As well, the shorter strands of DNA traveled faster (found at the bottom of the gel) across the gel than longer strands, thus displaying the fragments ordered by size in the agarose gel. The ordered fragments were visualized with a blue light transilluminator, which excited the dye added to the 4C sample and displayed the DNA lengths in the blue light spectrum. The florescent blue light produced from the agarose gel electrophoresis illustrated a smear of sequences around 1 kb in length (Figure 7) [45]. The triangle in Figure 7. indicates an increased concentration the 4C library sample in those reactions, thus exhibited by the brighter smear shown on the agarose gel. Each lane on the agarose gel us a different titration of the DNA sample. The smear result in the last lane of each gene was consistent with other reports and indicated that a 4C library was successfully constructed from chromatin in zebrafish embryos [45]. Proceeding this check-point step, sequencing primers for next generation sequencing were added to the 4C sample. Once this instrument specific primer was combined with the 4C library, the sample was sent out for sequential analysis of the chromatin interactions.

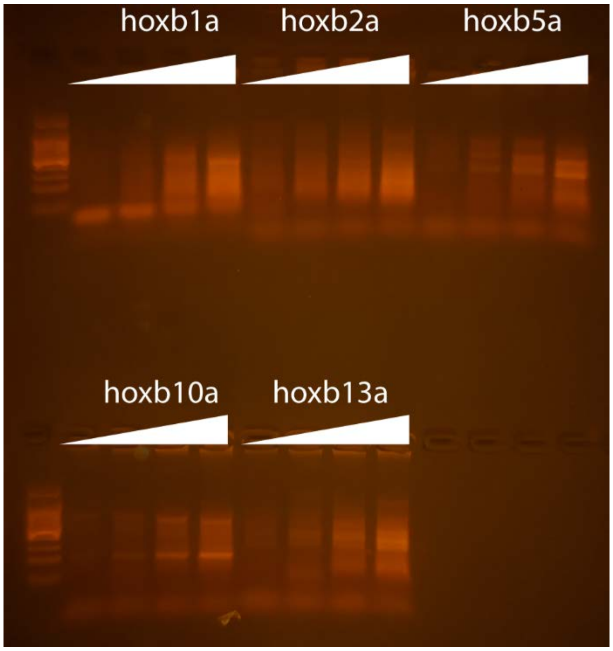

Figure 7. Electrophoresis data from test of 4C *hox* ba cluster library

The current analysis of results from the *hox* ba gene cluster are still waiting full analysis in the collaboration with URI Genomics and Sequencing Center. Preliminary sequential analysis of the 4C data supported the notion that the sequences from the 4C library map back to the *hox* ba clusters of the zebrafish (Figure 8). Based on the primers and restriction enzymes employed, reads of the 4C library sample yielded expected results based on the literature. The *hoxb1a* cluster was used an example of the reads shown from sequential analysis in Figure 8. Each of the bars from Figure 8. display a sequence/result which were reads derived from the intended *hoxb1a* primer. As shown in Figure 8., the sequence from the 4C library was located around the restriction enzymes used (DpnII and NlaIII) and mapped back to the intended genome loci of the zebrafish genome. Based on the preliminary analysis, these data suggested that a 4C library was develop and were amplified to gene specific *hoxb1a* primers. The identification of the points of contact and the elements attached to the *hox* cluster fragments are still not established. Thus, the investigation of the sequences close together in three-dimensional space is being continued.

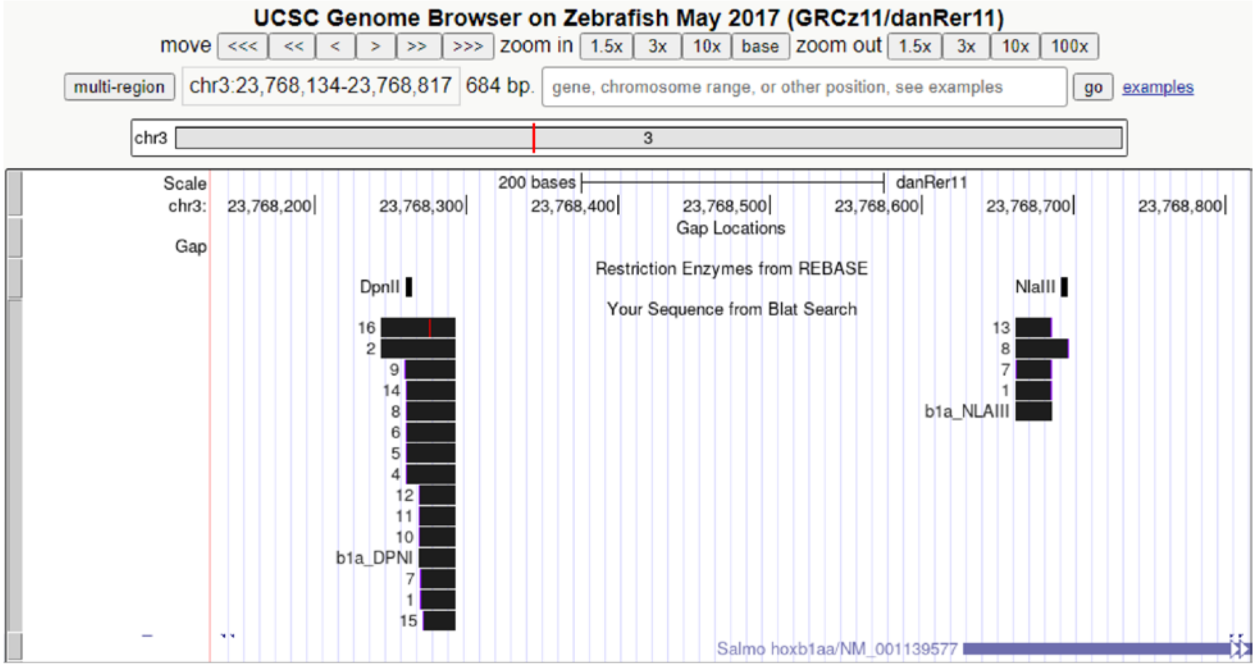
UCSC Genome Browser on Zebrafish May 2017 (GRCz11/danRer11)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
chr3:23,768,134-23,768,817
684 bp.
gene, chromosome range, or other position, see examples
go
examples
chr3
3
Scale
chr3:
200 bases
danRer11
23,768,200
23,768,300
23,768,400
23,768,500
23,768,600
23,768,700
23,768,800
Gap Locations
Gap
Restriction Enzymes from REBASE
DpnII
NlaIII
Your Sequence from Blat Search
16
2
9
14
8
6
5
4
12
11
10
b1a_DPNI
7
1
15
13
8
7
1
b1a_NLAIII
Salmo hoxb1aa/NM_001139577

Figure 8. Sequential mapping of *hoxb1a* 4C library to zebrafish genome

DISCUSSION AND FUTURE DIRECTIONS

The lack of sequential data results in an inability to conclude information regarding the contact points and what subsequent gene sequence are locating in these regions. The collaboration with URI's Genomics and Sequencing Center will continue in order to collect more information regarding the identification of where the *hox* genes are connected to other sequences in the genome. Recognizing the specific gene sequence's location, the function that gene codes for, and the gene products consequential involvement in the proper functioning will allow for further comprehension of expression mechanisms. Moreover, Correlations and relationships regarding gene regulation and expression can be drawn from this data concerning chromatin architecture of specific gene in future studies.

There are three major future directions for the following thesis. The first of these future goals is to identify important sequences that regulate loop formation within multiple *HOX* clusters. The important sequences are in reference to the specific sequences found within the points of contact. The data for significant sequences of the loops will allow for subsequent determination of proteins that bind to these sequences and how those proteins are regulated. All sequential information and coinciding proteins are crucial to explore as a whole to provide a comprehensive model of the three-dimensional structure of the chromatin in *HOX* genes. The second aim for additional research would be to conduct a comparative analysis of additional timepoints. The four post-fertilization time point chosen only represents a snapshot of chromatin architecture in the cell. By comparing a multitude of timepoints, the natural progression and development of chromatin folding can be examined. The differentiation in three-dimensional structure between timepoints will provide insight on expression levels of various genes. The last direction for future research would be producing a comprehensive healthy model of chromatin interaction to compare to disease states such as cancer. In the epidemiological field, comparative analysis between the healthy cell model and disease state cells highlights structural chromatin discrepancies. Extensive research on the contrasts on cell states may provide relevant information to treat cancers in a more specific manner based on genomics. Specifically, the knowledge on how and why *hox* gene mutations are commonly found in disease state cells, such as cancer, may aid in treatment methods by looking at a particular modification in a gene as a target. However, prior to focus on cancer genomics, an extensive model of chromatin interactions in a healthy state is required. The mechanisms within a disease state cannot be understood without comprehension of the proper development and architecture for the entirety of genes. Overall, future studies can use the significant information collected from sequencing and analysis techniques of the *hox* gene chromatin structure to expose important relationships between chromatin architecture and *hox* expression.

REFERENCES

1. Shah N, Sukumar S (2010) The *HOX* genes and their roles in oncogenesis. Nat Rev Cancer 10:361–371 . https://doi.org/10.1038/nrc2826
2. National Cancer Institute (2017) Statistics at a Glance: The Burden of Cancer in the United States. Cancer Stat OMB No.: 0925-0642
3. Bailey-Wilson J (2020) Glossary of Genetic Terms. In: Natl. Hum. Genome Res. Institue. https://www.genome.gov/genetics-glossary/Diploid
4. Bickmore WA, Van Steensel B (2013) Genome architecture: Domain organization of interphase chromosomes. Cell
5. Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, Demichelis F, Melnick AM, Elemento O, Rubin MA (2012) Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci U S A 109:9083–9088 . https://doi.org/10.1073/pnas.1112570109
6. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science (80-) 295:1306–1311 . https://doi.org/10.1126/science.1067799
7. Smith EM, Lajoie BR, Jain G, Dekker J (2016) Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. Am J Hum Genet. https://doi.org/10.1016/j.ajhg.2015.12.002
8. Rezsohazy R, Saurin AJ, Maurel-Zaffran C, Graba Y (2015) Cellular and molecular insights into *HOX* protein action. Dev. 142:1212–1227
9. Sánchez-Herrero E (2013) *HOX* Targets and Cellular Functions. Artic ID 2013: . https://doi.org/10.1155/2013/738257
10. Bickmore WA, Mahy NL, Chambeyron S (2004) Do higher-order chromatin structure and nuclear reorganization play a role in regulating *HOX* gene expression during development? Cold Spring Harb Symp Quant Biol 69:251–257 . https://doi.org/10.1101/sqb.2004.69.251
11. Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, Trono D, Spitz F, Duboule D (2013) A switch between topological domains underlies *HOX*D genes collinearity in mouse limbs. Science (80-) 340:1234167 . https://doi.org/10.1126/science.1234167

12. Krijger PHL, Geeven G, Bianchi V, Hilvering CRE, de Laat W (2020) 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. Methods 170:17–32 . https://doi.org/10.1016/j.ymeth.2019.07.014
13. Geeven G, Teunissen H, de Laat W, de Wit E (2018) peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. Nucleic Acids Res 46:e91 . https://doi.org/10.1093/nar/gky443
14. Tan G, Lenhard B TFBSTools: an R/bioconductor package for transcription factor binding site analysis. https://doi.org/10.1093/bioinformatics/btw024
15. Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. Bioinforma Appl NOTE 27:1696–1697 . https://doi.org/10.1093/bioinformatics/btr189
16. Fraser J, Williamson I, Bickmore WA, Dostie J (2015) An Overview of Genome Organization and How We Got There: from FISH to Hi-C. Microbiol Mol Biol Rev 79:347–372. doi: 10.1128/mmbr.00006-15
17. Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107:1–8. doi: 10.1016/j.ygeno.2015.11.003
18. Sanger F (1988) SEQUENCES, SEQUENCES, AND SEQUENCES. Annu Rev Biochem
19. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376–380 . https://doi.org/10.1038/nature11082
20. Van Steensel B (2011) Chromatin: Constructing the big picture. EMBO J 30:1885–1895. doi: 10.1038/emboj.2011.135
21. Zheng H, Xie W (2019) The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol 20:535–550. doi: 10.1038/s41580-019-0132-4
22. Lee JY, Orr-Weaver TL (2001) Chromatin. Encycl Genet 340–343
23. Szabo Q, Bantignies F, Cavalli G (2019) Principles of genome folding into topologically associating domains. Sci Adv 5. doi: 10.1126/sciadv.aaw1668
24. Weicksel SE, Gupta A, Zannino DA, Wolfe SA, Sagerström CG (2014) Targeted germ line disruptions reveal general and species-specific roles for paralog group 1 *HOX* genes in zebrafish. BMC Dev Biol 14: . https://doi.org/10.1186/1471-213X-14-25

25. Sarkar S, Horn G, Moulton K, Oza A, Byler S, Kokolus S, Longacre M (2013) Cancer development, progression, and therapy: An epigenetic overview. Int J Mol Sci 14:21087–21113. doi: 10.3390/ijms141021087
26. Vermeulen K, Van Bockstaele DR, Berneman ZN (2003) The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer. Cell Prolif 36:131–149. doi: 10.1046/j.1365-2184.2003.00266.x
27. Cooper G (2000) Regulation of Transcription in Eukaryotes. In: The Cell: A Molecular Approach, 2nd ed. Sinauer Associates
28. Morgan MA, Shilatifard A (2015) Chromatin signatures of Cancer. Genes Dev 29:238–249. doi: 10.1101/gad.255182.114
29. Saha A, Wittmeyer J, Cairns BR (2006) Chromatin remodelling: The industrial revolution of DNA around histones. Nat Rev Mol Cell Biol 7:437–447. doi: 10.1038/nrm1945
30. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, 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. Nat Genet 38:1341–1347. doi: 10.1038/ng1891
31. Luo Z, Rhie SK, Farnham PJ (2019) The Enigmatic *HOX* Genes: Can We Crack Their Code? Cancers (Basel) 11: . https://doi.org/10.3390/cancers11030323
32. Montavon T, Duboule D (2013) Chromatin organization and global regualtion of *HOX* gene clusters. Philos Trans R Soc Lond B Biol Sci 368:20120367
33. Cuadrado A, Gimenez-Llorente D, Kojic A, Rodriguez-Corsino M, Cuartero Y, Martin-Serrano G, Gomez-Lopez G, Marti-Renom MA, Losada A (2019) Specific Contributions of Cohesin-SA1 and Cohesin-SA2 to TADs and Polycomb Domains in Embryonic Stem Cells. Cell Rep 27:3500-3510 e4 . https://doi.org/10.1016/j.celrep.2019.05.078
34. Barutcu A, Fritz A, Zaidi S, VanWijnen A, Lian J, Stein J, Nickerson J, Imbalzano A, Stein G (2016) C-ing the genome: A compendium of chromosome conformation

capture methods to study higher-order chromatin organization. J Cell Physiol 231:31–35. doi: 10.1002/jcp.25062.C-ing

35. Krijger PHL, Geeven G, Bianchi V, Hilvering CRE, de Laat W (2020) 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. Methods 170:17–32. doi: 10.1016/j.ymeth.2019.07.014
36. van den Boogaard M, van Weerd JH, Bawazeer AC, Hooijkaas IB, van de Werken HJG, Tessadori F, de Laat W, Barnett P, Bakkers J, Christoffels VM (2019) Identification and Characterization of a Transcribed Distal Enhancer Involved in Cardiac Kcnh2 Regulation. Cell Rep 28:2704-2714 e5 . https://doi.org/10.1016/j.celrep.2019.08.007
37. Wang Q, Jia Y, Wang Y, Jiang Z, Zhou X, Zhang Z, Nie C, Li J, Yang N, Qu L (2019) Evolution of cis- And trans-regulatory divergence in the chicken genome between two contrasting breeds analyzed using three tissue types at one-day-old. BMC Genomics 20:1–10. doi: 10.1186/s12864-019-6342-5
38. Brown T (2002) Chapter 1: The Human Genome. In: Genomes, 2nd edition. Oxford
39. Szabo Q, Bantignies F, Cavalli G (2019) Principles of genome folding into topologically associating domains. Sci Adv 5. doi: 10.1126/sciadv.aaw1668
40. Matharu N, Ahituv N (2015) Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. PLoS Genet 11:1–14. doi: 10.1371/journal.pgen.1005640
41. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502:59–64. doi: 10.1038/nature12593
42. Weicksel SE, Xu J, Sagerström CG (2013) Dynamic Nucleosome Organization at *hox* Promoters during Zebrafish Embryogenesis. PLoS One 8. doi:10.1371/journal.pone.0063175
43. Dekker J, Valton A-L (2016) TAD disruption as oncogenic driver. Curr Opin Genet Dev 36:34–40. doi: 10.1016/j.gde.2016.03.008.TAD
44. Bhatlekar S, Fields JZ, M Boman B (2014) *HOX* genes and their role in the development of human cancers. J Mol Med 92:811–823

45. Detrich H, Zon L, Westerfield M (2016) Assay for transposase-accessible chromatin and circularized chromosome conformation capture, two methods to explore the regulatory landscapes of genes in zebrafish. In: The Zebrafish: Genetics, Genomics, and Transcriptomics, 4th ed. pp 414–428