Methods for Identifying and Analyzing Cis-Regulatory Elements in *Petromyzon marinus*

By Cameron Bellian

Department of Ecology and Evolutionary Biology, University of Colorado at Boulder

Defended on March 28, 2022

Thesis Advisor:

Dr. Daniel Medeiros, Department of Ecology and Evolutionary Biology

Additional Committee Members:

Dr. Edward Chuong, Department of Molecular, Cellular, and Developmental Biology

Dr. Barbara Demmig-Adams, Department of Ecology and Evolutionary Biology

Abstract

Every day we interact with numerous organisms, all deriving their form and function from their own specific genetic code. With the great diversity in forms, also known as organismic phenotype, early biologists expected to see a great diversity in genetic makeup. However, with the advent of sequencing technologies, it has been discovered that the genetic makeup of each of these organisms is far more similar than one would expect. Most organisms have the same types of genes, with small changes in sequence relative to the change in phenotype. This raises the question of where the variation comes from. The answer lies in how and when these genes are turned on or off, which is called gene regulation. Cis-regulatory elements (CRE) are a specific type of gene regulation. They are regions of DNA on the same chromosome as the gene they regulate; they can either activate the transcription of the gene into an mRNA and then protein, or they can repress the expression of a gene. New methods in both genome sequencing and lab-based methods of CRE identification have made it easier to assess the functional differences of CREs in the genome and determine what makes them work. Currently, we understand much less about these regions than protein-coding regions. While we can predict exactly what protein a protein-coding sequence will make, we cannot always predict what function a CRE will have within an organism. Rules that dictate this function are referred to as CRE grammar and are an area of significant research interest currently. With a better understanding of CRE grammar, we can predict their function and determine how they contribute to both the function of individual organisms as well as the evolutionary history of a clade, and how they have evolved their specific phenotypic features. In this thesis, I used

Petromyzon marinus, the sea lamprey, as my model organism. They are part of one of only two living groups of jawless vertebrates, and they have significant difference in their adult morphology relative to all other vertebrates. For this reason, we can use their uniqueness within the vertebrate group for assessing the changes in cis-regulation that lead to changes in phenotype, especially when they are compared to other, more well studied organisms, such as Danio rerio, the zebrafish. In this thesis, I have adapted a combination of lab-based and computational methods to create a concerted method for identifying and analyzing CREs and their position within the context of the genome, called the cis-regulatory landscape. By developing a unified, standardized method for identifying CREs, future studies can use this method to identify and test CREs, helping to elucidate the CRE grammar that dictates CRE function, and therefore guides, on a small scale, individual phenotypes and, on a large scale, the evolution of species phenotype as a whole.

1 Introduction

1.1 Background

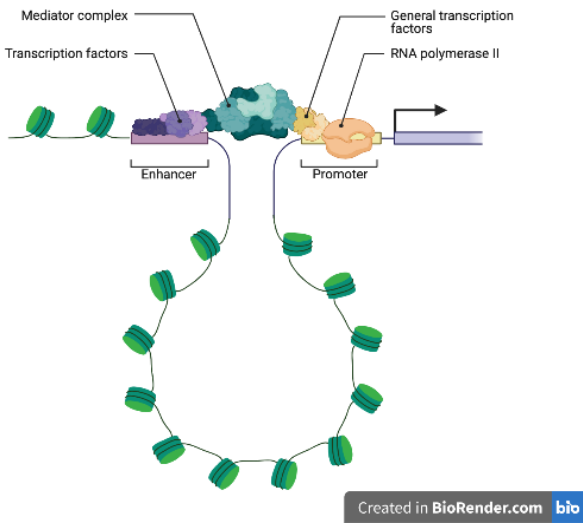

Figure 1. Generalized depiction of an active cis-regulatory element. Created with BioRender.com

How genotypes become phenotypes is one of the founding questions of many fields of biology, and certainly one of the most actively researched questions, including in the field of evolutionary developmental biology. We are surrounded by diversity on a daily basis, whether it be interspecific variation between species or intraspecific variation between individuals within a species, and these differences must be derived by some combination of their environment and their genetic makeup. Early biologists assumed that the great diversity of species was derived from vast differences in protein-coding sequences between different species. However, once sequencing technology and its corresponding computational methods were developed to allow for whole-genome sequencing, the corresponding discoveries were not expected. Instead, it was found that most organisms have largely the same number and type of genes (Prachumwat & Li, 2008). The changes in organismic phenotype must instead lie in the timing, location, and duration with which these protein-coding genes are expressed, referred to as gene regulation. This has led to the development of one of the big questions in the field of evolutionary developmental biology:

how does the regulation and evolution of gene expression lead to heritable changes in organismic phenotype and how does gene expression function? The answer lies in gene regulation, and more specifically the network of gene regulatory elements that control gene expression and allow for changes in morphology (Murrell et al., 2005; Prachumwat & Li, 2008; Wittkopp, 2006).

The functional units of gene regulation involve both trans-regulatory elements and cis-regulatory elements, but here I will focus on the cis-regulatory elements (CRE). These CREs can be broken down into three distinct types, based on their function. There are promoters found at the beginning of every gene that help initiate transcribing the protein-coding sequence into an mRNA before being converted into a protein, enhancers which upregulate the expression of a gene, and finally silencers that downregulate the expression of a gene (Wittkopp & Kalay, 2012). I will be focusing on enhancers here as one of the crucial ways in which evolution of an organism is modulated by fine-tuning the spatial and temporal expression of proteins. The rules that dictate the function of a CRE, such as how and what activates it, how flexible its sequence identity can be while still maintaining function, and its ability to subtly change phenotype, can be referred to as CRE grammar. This area of CRE research is still not well understood, and yet it can unlock the understanding of the many ways in which CREs function to create active gene regulatory networks. If we can understand the ways in which CRE grammar works, we can begin to understand how it shapes both the diversity of all organisms as well as how CRE function can shape the slight changes in phenotype seen across a population.

To study this, the model organism of choice was Petromyzon marinus, the sea lamprey. They are part of one of only two extant groups of jawless vertebrates and have significantly different adult morphology from most other vertebrates. They also have a crucial, vertebrate-specific feature, the neural crest cells. This cell lineage is a pluripotent, embryonic cell type that differentiates into many diverse adult tissues, such as facial skeleton and melanocytes. This cell type does not exist in any other group, so comparing neural crest cells with their presumed closest-relative cell type in invertebrate chordates can allow for interesting comparative analyses of CRE function and location within the vertebrate family.

1.2 Evolutionary Ideas

There are several theories about how these elements evolve, but the dominant theory is that cis-regulatory evolution is mostly driven by the modification of previously existing CREs. These CREs are modified in similar ways to protein coding sequences, especially after duplication events (Gompel et al., 2005; Jimenez-Delgado et al., 2009; Kvon et al., 2016). There is some evidence of *de novo* CRE formation, but this appears to be a much rarer occurrence, similar to what is seen in gene evolution (Arnold et al., 2014; Tsai et al., 2012). However, there is less agreement on how the exact layout and composition of these elements is conserved (Farley et al., 2015; Jindal & Farley, 2021). In fact, it is quite common to see divergent CRE sequences with maintained regulatory function (Farley et al., 2015; Kvon et al., 2016; Wong et al., 2020).

How this function is conserved is not as well studied, and therefore, not as well understood. One possibility is that the location of these enhancers within the spatial context of genic introns and exons, or the cis-regulatory landscape, may be conserved (Acemel et al., 2016; Arnold et al., 2014). One study has shown potential for spatial conservation across the entire

animal kingdom guiding CRE function more than sequence (Wong et al., 2020). However, this study focused their analysis to only a few regions that were predisposed to high degrees of conservation, limiting the assumptions that can be drawn. The other possibility is that CREs can move within the genome with relative fluidity if key sequence identities are maintained (Farley et al., 2015; Jindal & Farley, 2021). Regardless, it appears that CREs have a much greater degree of fluidity in their form than protein-coding sequences while still maintaining their function.

1.3 Neural Crest Cells

Neural crest cells are a vertebrate-specific cell type that appears during neurulation of vertebrate embryos. Their multipotent nature and role in developing much of the facial skeleton, melanocytes, and neurons means that they are viewed as a critical step in the evolution of vertebrates (Gans & Northcutt, 1983; Glenn Northcutt, 2005). They also provide an interesting area of study for CRE and cis-regulatory landscape evolution because they quickly arose as a new cell type in stem vertebrates. The closest relatives of vertebrates that are used as model organisms, *Ciona intestinalis* and *Branchiostoma floridae*, do not have any neural crest cells, but they potentially have precursor cell lineages (Horie et al., 2018; Yu, 2010). Crucially, this means that neural crest cells are an ideal cell type for studying cis-regulatory element and landscape evolution because these new cell types must have arisen via the development of new regulatory networks. The evolutionary history can be identified by comparing the cis-regulatory element and landscape composition of invertebrate chordates, *Ciona intestinalis* and *Branchiostoma floridae,* with a vertebrate chordate, *Petromyzon marinus* (sea lamprey) (Delsuc et al., 2006). These comparative studies have the potential to shed much light on the ways in which CREs have evolved, elucidating their grammar.

1.4 Key Questions

The aim of this thesis is to identify and implement an unbiased method for identifying CREs within a given genome using a combination of lab methods and computational methods. Once tested and proven, future goals of this pipeline include testing identified, putative orthologous enhancers and their functional output. From there, we can begin to draw assumptions as to how these elements derive their function from either their primary or secondary structure, such as sequence, transcription factor binding site motifs, and contribution to genome structure or enhancer RNAs.

The model organism of focus for the lab methods in this research is the sea lamprey (*Petromyzon marinus)*. This organism provides a great intermediate between the basic invertebrate chordates of Ciona and Branchiostomae, and the more complex vertebrate morphology of an organism like the teleost zebrafish. Lamprey also provides an idea of what a basic neural crest cell gene regulatory network might look like *in vivo*. By focusing our efforts on sea lamprey, we can fill in the gap between these two groups, creating a clearer picture of how CREs have evolved and to what extent the cis-regulatory landscape matters with regards to function.

2 Methods

2.1 Lab Methods

The focus of the lab methods is best summarized in three distinct parts. First, I developed a new method for dissociating live embryos into individual live cells without reducing cell viability. Next, I developed a method for taking isolated live cells and rapidly stripping away the cellular membrane and any organelles, leaving behind an intact cell nucleus. Finally, I optimized

a recently-developed lab method, Cleavage Under Targets and Tagmentation, or CUT&Tag, for extracting target DNA associated with nucleosomes that had specific post-translation modifications of their histone tails. These steps unify into a single pipeline within the lab, allowing for the staging, collection, dissociation, and nuclear extraction within the span of a single workday. This rapidly increased the potential throughput of this protocol. However, to start, first I had to identify the developmental stages of interest of *Petromyzon marinus* (Sea Lamprey) embryos.

2.1.1 Staging

Previous studies have identified the developmental stages of sea lamprey, including when neural crest cells are originally formed, when they are migrating, and when they are differentiating (Tahara, 1988). These sources were used to identify Tahara stages 21 and 22 of sea lamprey as the focus of the analytical methods. This corresponds to stages that involve migratory neural crest cells. The hyperactivity of the neural crest cell differentiation genes at these stages provides an ideal time to probe for changes in cis regulation that may allow for this cell type to be both hyper-mobile and pluripotent. The methods for staging, as described by Tahara, focusing on morphology and not just time since fertilization, to fully determine the developmental stage of the embryo.

2.1.2 Post-Translational Modifications

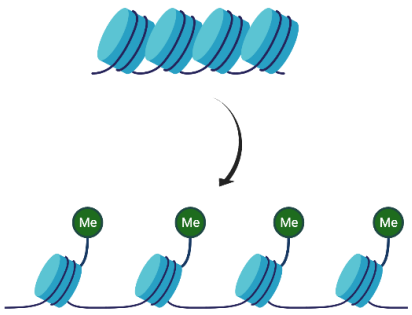

Figure 2. A modification of histones via H3K4me3 and H3K4me2 loosens the chromatin structure, allowing for easier access to the CREs. Created with BioRender.com.

Histone modifications are often used as flags to identify active regulatory elements, actively read genes, and regions of suppressed gene expression (Henikoff & Shilatifard, 2011; Kaya-Okur et al., 2019; Shah et al., 2018;

Soares et al., 2017). Histone 3 is of interest for studying enhancers, and there are two copies in each nucleosome. The methylation of lysine 4 and the acetylation of lysine 27 on histone 3, which are post-translational modifications, are identifiers of active regulatory elements within the genome (Shah et al., 2018; E. Smith & Shilatifard, 2014; Soares et al., 2017). Once two copies of each of the four histones (H2A,H2B,H3,H4) are combined into a protein-complex, a nucleosome is formed (McGhee & Felsenfeld, 1980). A strand of approximately 150 base pairs of DNA wraps around the nucleosome complex due to electrostatic attractions (McGhee & Felsenfeld, 1980). This allows for the compact storage of DNA in the cell. Fortunately for researchers, it also can allow for the identification of regions of DNA based on the post-translational modifications found and the activities with which they are normally correlated. These post-translational modifications limit the ability of DNA to bind tightly to the nucleosome, increasing the accessibility of the DNA to the environment of the nucleus, including transcription factors, polymerases, and other protein complexes.

2.1.3 CUT&Tag

To identify the active cis-regulatory elements in the model organisms, a new molecular biology technique, called Cleavage Under Targets and Tagmentation, or CUT&Tag, was used. This method uses a primary antibody to bind to a post-translational modification of interest of histone 3, and then a secondary antibody is bound to this primary antibody (Kaya-Okur et al., 2019). Next, a Tn5 transposase that is conjugated with both protein A and protein G is bound to the secondary antibody (Kaya-Okur et al., 2019). By varying the salt concentration, the Tn5 complex then cuts the region of DNA on either side of the nucleosome that it is bound to and adds adapters to each end of the sample (Kaya-Okur et al., 2019). These small fragments can

then be washed away and saved, with the adapters eliminating any need for sequencing library preparation (Kaya-Okur et al., 2019).

The benefit of this protocol over other similar techniques, such as CHiP-Seq, is due to the apparent low signal-to-noise ratio, the need for far less sample to get statistically significant results, and the reduced cost due to a simplified library prep and the need for less starting material. Therefore, I believed that CUT&Tag was an ideal method due to the nature of my samples. Lamprey only spawn for approximately 6 weeks in mid-summer every year, so all embryo collection and dissociation happened within that time frame. This limits the amount of material that can be collected; therefore, a more sensitive test was a benefit to my study design.

2.1.4 Embryo Dissociation and Nuclear Isolation

Figure 3. Initial, yolky dissociation of Stg 21 Lamprey Embryos

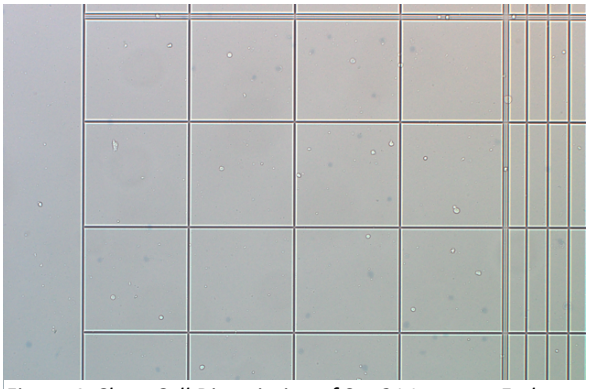

Figure 4. Clean Cell Dissociation of Stg 21 Lamprey Embryos

There are, however, some complications in beginning the CUT&Tag protocol. The most important is that the organism of interest is reduced to only intact nuclei, with no remaining extra-cellular matrix, organelles, or other cellular components. The first step was to isolate individual live cells from each of the developmental stages of interest. The first experiments were using zebrafish embryos due to their relative ease in acquisition and the lack of seasonality in their reproductive cycle, which allowed for us to develop a sound protocol before moving to working with lamprey. The embryo dissociation protocol was adapted from the Adam et al.

protocol and the Quillien et al. protocol (Adam et al., 2017; Quillien et al., 2017). Zebrafish embryos up to the prim-5 stage, the last tested, were easy to dissociate with little cell death. Therefore, I started each new lamprey dissociation experiment by trying the zebrafish protocol. This provided us with a clear, sound starting point that I could easily and rapidly modify as needed based on the outcomes.

However, the sea lamprey embryos provided many more technical challenges. Their developmental morphology includes a more integrated yolk that dissociates into granules that intermingle with any remaining cells, thus contaminating the samples. To mitigate this, the zebrafish protocol was supplemented with steps from the Gentsch and Smith protocol (Gentsch & Smith, 2019). These steps focused on a thorough wash with several surfactants to remove the fat-based yolk granules while leaving intact cells behind (Gentsch & Smith, 2019). This step had to be optimized because of the ability of surfactants to disrupt the phospholipid bilayer and lysing the cells of interest too early in the protocol. To mitigate this, short washes with minimal agitation were used.

Isolating nuclei was then performed using a protocol adapted from Henikoff et al. (Henikoff et al., 2020). The usage of surfactants and agitation were maintained, as per the protocol, but the type and strength of the surfactant was changed based on the sensitivity of the sample cells. For zebrafish, Triton 100-X was used, but for sea lamprey, Tween-20 was used because the sea lamprey nuclei are fragile and were lysing with the use of the stronger Triton 100-X.

2.1.5 Experimental Focus

The focus of these experiments has been on H3K4 methylation states, and especially H3K4me3 and H3K4me2. These post-translational modifications are most strongly correlated

with a combination of promoters for H3K4me3 and with enhancers and promoters for H3K4me2 (Henikoff & Shilatifard, 2011; Shah et al., 2018; Soares et al., 2017). By comparing the location of these respective post-translational modifications, it is possible to pinpoint the sequence and location of putative cis-regulatory elements for further study. This does, however, limit how powerful the lamprey CRE identification methods are. Because I used two antibodies for similar post-translational modifications, it is possible that there was cross-reactivity between my antibodies and their respective epitopes. Also, it would be best practice to use two very different types of post-translational modifications to identify CREs. These changes will be implemented in future trials of this study.

2.1.6 Sequencing

The pre-sequencing quality check of my samples was completed using the Bioanalyzer at the University of Colorado (CU) Boulder Biocore facility. The focus here was to ensure that the sample contained DNA fragments of the appropriate length, which are approximately 150 base pairs plus both adapters. Sequencing of my samples, which included Lamprey stage 21 H3K4me3, stage 21 H3K4me2, stage 22 H3K4me2, and stage 22 H3K4me3, was completed at the CU Anschutz Sequencing Core. My samples were sequenced to a depth of between 5 million and 48 million reads, dependent on the sample. All pre-sequencing sample prep, such as concentration matching and pooling, was done by the CU Anschutz Sequencing Core faculty. The selected parameters were paired end 150 cycle 2x150 on a NovaSEQ 6000.

2.2 Computational Methods

To analyze the data, a pipeline was developed to process from raw sequence reads to read depth at a particular locus. The computational pipeline developed was adapted from Henikoff et al., with minor changes based on the needs of these specific experiments and the intended uses

(Henikoff et al., 2020). After this pipeline was completed, I developed my own algorithm to look for overlapping or otherwise interacting peaks within the presumed topologically associating domain of a gene of interest. This allowed for the rapid testing of my data to identify any putative regulatory elements and ability to store their length and location relative to the gene body and other exons. This algorithm was validated using existing CHiP-Seq data from zebrafish and used to create predictions annotated to the lamprey genome assembly.

2.2.1 Filtering and Alignment

Once sequenced, the reads were first filtered by quality of Phred 20 or higher using Cutadapt (Barske et al., 2018; Martin, 2011). This correlates to a minimum 99% confidence in each individual base called by the sequencer. Any reads with scores below that were thrown out to mitigate noise or false results. At the same time, any remaining adapter sequence was removed, also using Cutadapt (Martin, 2011).

Alignments of these sorted reads was accomplished using Bowtie2, a Burrows-Wheeler based algorithm (Langmead et al., 2019; Langmead & Salzberg, 2012). The first step creates a reference file from the current reference genome. I used the 2017 germline genome assembly because it has been used for previous ATAC-Seq and CHiP-Seq experiments. Therefore, it is well proven as a functional genomics tool and allows for better comparison between my data and outside data (J. J. Smith et al., 2018). The paired-end sequencing reads were fed in to the Bowtie2 aligner, producing a SAM file of all the aligned reads, their location, and the depth of reads at that location. This SAM file can then be converted to many different formats based on the necessary use case.

In this case, the data was converted to a bigWig file using the bedGraphtoBigWig utility from UCSC (Kent et al., 2010). This allows the genome browser to show the variability in read

depth as a continuous variable, unlike bedGraph formats, and compresses the data to allow for easier upload to various sources and browsers. Finally, peak-calling software was also used to verify the qualitative identification of peaks with an algorithmic, quantitative method. I used MACS2 for its simple interface, large share of users within the bioinformatics field, and accessibility as open-source software (Zhang et al., 2008).

2.2.2 Algorithm

The next focus was on developing an algorithm that objectively identifies the location of cis-regulatory elements by inputting previous CUT&Tag or CHiP-Seq data sets. The conversion to discretized peak data allows for vary simple overlap calculations. The algorithm asks a user for the chromosome of interest, the location of the gene on the chromosome of interest, and the desired search distances on either side of the gene. It then finds all peaks that coexist on the chromosome of interest and that are within the bounds provided by the user before doing an all-for-all comparison to determine if any of the peaks overlap, or conversely, do not overlap at all. Currently, the overlap characteristic is set so that for any given peak, if there is at least a single base pair overlap, it will collapse both these peaks into one larger peak with the longest possible length. This means that not all of the new peak is necessarily part of the overlap. However, because of the flexibility of cis-regulatory function and the inexactness of the peak formation due to histone location limiting cut site to up to 150 bp periodicity, this method helps guarantee that more of the enhancer will be caught by the algorithm.

To test this algorithm and its usability, I implemented two types of tests. The first was a visual test to determine how closely the reported peak from the algorithm corresponded with a peak that a trained researcher would agree was significant. The second test was to compare to previously identified enhancers as chosen by humans and determine which peaks accurately

picked a region of interest, where humans were too generous or stringent, and where the algorithm was too generous or stringent.

3 Results

3.1 Lab Methods

The embryo dissociation experiments provided a great number of challenges throughout the trials. The presence of yolk granules around the size of the individual cells prevented proper separation or preservation of these cells without first removing the granules. Fortunately, by utilizing their difference in lipid content, I was able to strip away the yolk granules. Using high-powered microscopy at the CU Boulder Light Microscopy Core, I was able to confirm that these methods leave intact cells, but the samples are devoid of any yolk. I repeated imaging after freezing at -70ºC and then thawing gently in a 20ºC water bath, followed by staining with DAPI to identify nuclei based on nucleic acid content. I found that when preserved in a 10% DMSO solution, nuclei that were in yolk free solutions were much more likely to remain intact. Those nuclei that were frozen with yolk remnants in their sample were more likely to be lysed. DAPI imaging showed loose nucleic acids in my sample, instead of well-bound nuclei.

My initial CUT&Tag experiments eventually resulted in a large enrichment of DNA strands approximately 250 base pairs in length, as shown by an Agilent Bioanalyzer 2100. This provides confidence that most of the cuts made by the Tn5 enzyme were centered on a nucleosome and that the adapters were successfully added to the samples. This is verified because the nucleosome can carry 148 base pairs of DNA plus the addition of adapters on either side of the cut DNA. When sequenced at the CU Anschutz Sequencing Core, there was significant variation of the total number of reads per sample and although not the most efficient,

and with the potential for added noise, the extra sequence depth did not hinder the data analysis steps.

3.2 Computational Methods

The alignment rate of my sequence data was lower than expected, at about fifty percent alignment rate, but there are several possible explanations. One potential is that a significant portion of the sequenced segments were shorter than the overall read length from the sequencer. In this case, the machine reads through the end of the sequence and then starts reading the other adapter. If the read of the adapter is not long enough, the trimming algorithm does not recognize it as adapter sequence and leaves it in place. To mitigate this problem, I used a less strict alignment protocol with Bowtie2. This allowed for the ends of sequences to be removed when aligning if they did not fit, but the core of the sequenced region did fit. This had a noticeable effect in increasing the overall alignment percentage of the sequence data.

The next issue is that the lamprey genome has historically been difficult to assemble. It has both a high GC content and a high degree of repeats within it. Finally, lamprey undergo significant genome rearrangements as they develop, eliminating up to 0.5 gigabases of DNA from their 2.3 gigabase genome by the end of the third day of development (J. J. Smith et al., 2018; Timoshevskiy et al., 2016). The genome assembly that I have been utilizing is from germline cells, with the goal of potentially eliminating any of these issues. It is unclear, however, how these rearrangements change the structure of the genome, and therefore the potential alignment of sequence data to the known genome.

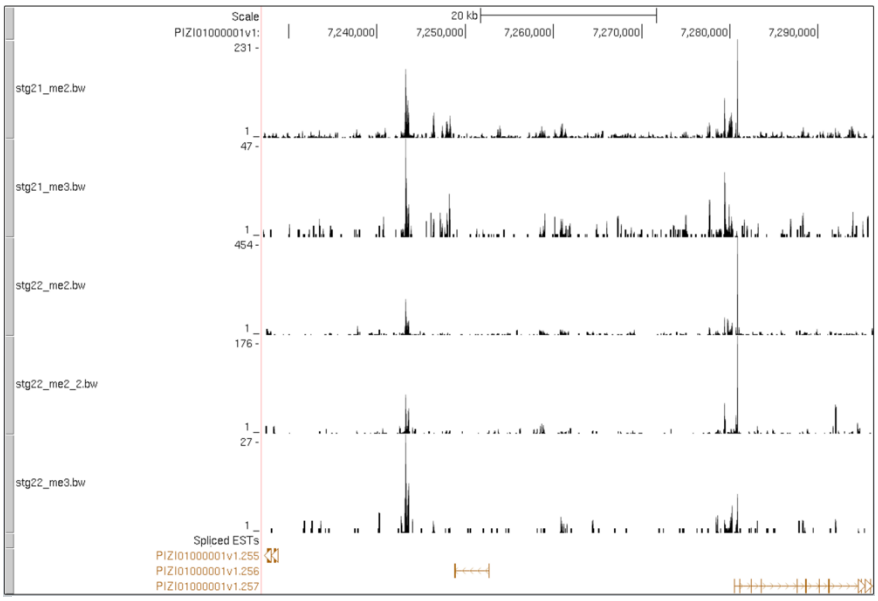

Figure 5. bigWig tracks on the UCSC Genome Browser for Stg 21 and 22 Lamprey sequence results

After aligning the sequence data, I was left with interpreting the data. A combination of tools from the University of California Santa Cruz Genomics Institute and personally developed software were utilized to continue to trim and interpret the data. The resulting data shows that the methods developed here do result in easily identifiable regions of enriched histone post-translation modifications that associate as expected around certain developmental regulator genes that are active at the stages of development tested. While the number of significant peaks, as designated by my algorithm, varied between each of the stages, all the stages and genes tested had at least 30 regions identified as putative cis-regulatory elements within 0.5 gigabase in either direction of the center of the gene body. This provides confidence that the algorithm is identifying potential enhancers that are worth further study. Finally, when compared with researcher-selected peaks in the zebrafish genome, there was some overlap in peaks. However, there was also a significant difference in what each tool, human, or computer, chose as significant. It appears that human selected peaks are better at identifying peaks that are prominent relative to their surroundings, even if they are not globally prominent. The computer was more consistent while choosing peaks, using all the selection criteria each time while not skipping any peaks. Human selections, however, were much more subjective.

4 Discussion

Cis-regulatory elements are a critical part of every gene regulatory network that guides embryogenesis and development or maintains organismal function (Gompel et al., 2005). The ability to identify and test CRE function with the express goal of determining what rules dictate the function of CREs has been an intractable problem until recent advances in high-throughput, modern biological methods. Put forward here is a combined lab and computational pipeline that allows for the unbiased identification of CREs based on the post-translational modifications of histones within the developing embryo. I have shown first that the new lab method, CUT&Tag, is a tractable and efficient method for identifying putative cis-regulatory elements within the developing lamprey embryo, despite the complications that this species provides due to their distinct, and at times divergent, biology. Second, I have developed a computational pipeline for identifying putative cis-regulatory elements in an objective, repeatable way. However, there are areas in which this lab-based and computational pipeline could be improved.

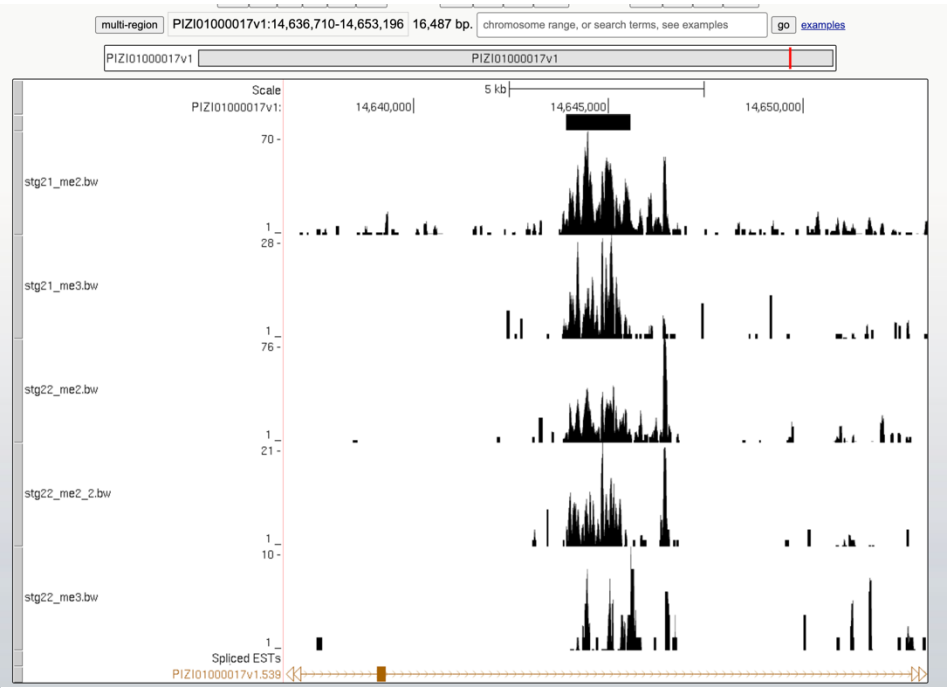

Figure 6. Results of peak calling algorithm are shown as a black bar at the top of the page. Here is an example of how the bar correlates to peaks found.

Because they have both neural crest cells, which is a vertebrate invention, and a relatively simple body plan, lampreys provide an useful model for studying the evolution of gene regulatory networks (York et al., 2020, 2021). Lamprey have undergone fewer whole-genome duplications than most other vertebrates, such as the teleosts (J. J. Smith et al., 2018). The lack of research on lamprey's regulatory landscape and function has impeded our understanding of gene regulation because their position in the vertebrate phylogenetic tree would allow for investigating the regulatory function of a new cell type before it gained too many diverse functions. This is especially useful in comparison to invertebrate chordates, such as the amphioxus, *Branchiostomae floridae*, which have relatively similar body plans and a similar evolutionary history as lamprey's but lack neural crest cells.

Difficulties in procuring and manipulating these embryos has limited my studies, but I show here that with modifications to existing embryo dissociation protocols, I am able to extract intact and usable cells or nuclei that can be stored long-term and then used when needed. While currently I have focused on post-translational modifications using CUT&Tag, especially because little data exists from lamprey CHiP-Seq experiments, these methods are also highly transferrable to other lab methods, such as scRNA-Seq. (York et al., 2021). The lack of data in both post-translational modifications and scRNA-Seq in this organism could rapidly be remedied now that the methods for dissociation and extraction have been optimized. This would allow for much greater research into the functional aspects of the lamprey genome.

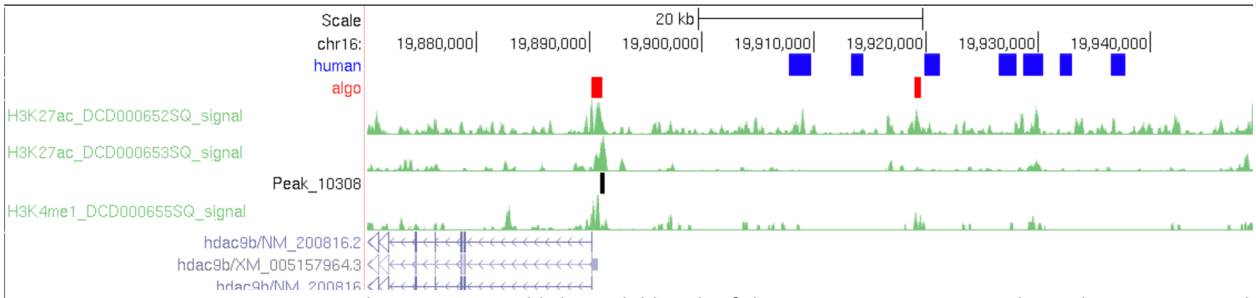

Figure 7. UCSC Genome Browser tracks comparing publicly available Zebrafish CHiP-Seq, putative CREs chosen by an expert, and algorithmically determined peaks.

A weakness exposed by the implemented computational methods is the fact that the majority of peak-calling to identify active cis-regulatory elements is done subjectively by an individual scanning across a genome browser with many tracks of different datasets. This was best exemplified when the algorithm developed here was tested on a zebrafish CHiP-Seq dataset that a trained researcher had analyzed for peaks of interest. Here I discovered that the human-identified peaks were very inconsistent in their stringency. While some peaks that were clearly prominent and significant were identified, many were also chosen that did not appear to be much more significant that the region next to them. This is problematic when trying to standardize

peak-calling and cis-regulatory element identification because the methodology becomes individual-dependent. It was with these concerns in mind that the cis-regulatory element identification algorithm was developed. By combining statistical tests to identify what peaks are significant with a simple search and combine algorithm, the goal was to develop a tool that allows for an unbiased method for identifying active cis-regulatory elements from relatively unbiased data, such as ChiP-Seq or CUT&Tag. The output of this tool is intended to be an easy-to-read list of all putative cis-regulatory elements, which can then be filtered by function as wanted and then tested in a way appropriate for the elements *in vivo* function.

Future directions of this project will focus on taking this unbiased dataset of cis-regulatory elements and using gateway cloning methods to create a GFP construct activated by the CRE provided by the algorithm. The expression of GFP can then be imaged at different time series to understand the ways in which each of these CREs function. Finally, the elements can be grouped by function and then compared, looking at sequence, transcription factor binding motifs, 3-dimensional structure, or other components to determine what features functionally equivalent elements have in common, without having to use one of those features as a search criteria, biasing the data.

References

1. Acemel, R. D., Tena, J. J., Irastorza-Azcarate, I., Marlétaz, F., Gómez-Marín, C., de la Calle-Mustienes, E., Bertrand, S., Diaz, S. G., Aldea, D., Aury, J.-M., Mangenot, S., Holland, P. W. H., Devos, D. P., Maeso, I., Escrivá, H., & Gómez-Skarmeta, J. L. (2016). A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nature Genetics*, *48*(3), 336–341. https://doi.org/10.1038/ng.3497
2. Adam, M., Potter, A. S., & Potter, S. S. (2017). Psychrophilic proteases dramatically reduce single cell RNA-seq artifacts: A molecular atlas of kidney development. *Development*, dev.151142. https://doi.org/10.1242/dev.151142
3. Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C., & Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genetics*, *46*(7), 685–692. https://doi.org/10.1038/ng.3009
4. Barske, L., Rataud, P., Behizad, K., Del Rio, L., Cox, S. G., & Crump, J. G. (2018). Essential Role of Nr2f Nuclear Receptors in Patterning the Vertebrate Upper Jaw. *Developmental Cell*, *44*(3), 337-347.e5. https://doi.org/10.1016/j.devcel.2017.12.022
5. Delsuc, F., Brinkmann, H., Chourrout, D., & Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, *439*(7079), 965–968.

6. Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., & Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science*, *350*(6258), 325–328. https://doi.org/10.1126/science.aac6948
7. Gans, C., & Northcutt, R. G. (1983). Neural Crest and the Origin of Vertebrates: A New Head. *Science*, *220*(4594), 268–273. https://doi.org/10.1126/science.220.4594.268
8. Gentsch, G. E., & Smith, J. C. (2019). Mapping Chromatin Features of *Xenopus* Embryos. *Cold Spring Harbor Protocols*, *2019*(4), pdb.prot100263. https://doi.org/10.1101/pdb.prot100263
9. Glenn Northcutt, R. (2005). The new head hypothesis revisited. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *304B*(4), 274–297.
10. Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A., & Carroll, S. B. (2005). Chance caught on the wing: Cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*, *433*(7025), 481–487. https://doi.org/10.1038/nature03235
11. Henikoff, S., Henikoff, J. G., Kaya-Okur, H. S., & Ahmad, K. (2020). Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *ELife*, *9*, e63274. https://doi.org/10.7554/eLife.63274
12. Henikoff, S., & Shilatifard, A. (2011). Histone modification: Cause or cog? *Trends in Genetics*, *27*(10), 389–396. https://doi.org/10.1016/j.tig.2011.06.006
13. Horie, R., Hazbun, A., Chen, K., Cao, C., Levine, M., & Horie, T. (2018). Shared evolutionary origin of vertebrate neural crest and cranial placodes. *Nature*, *560*(7717), 228–232. https://doi.org/10.1038/s41586-018-0385-7
14. Jimenez-Delgado, S., Pascual-Anaya, J., & Garcia-Fernandez, J. (2009). Implications of duplicated cis-regulatory elements in the evolution of metazoans: The DDI model or how

simplicity begets novelty. *Briefings in Functional Genomics and Proteomics*, *8*(4), 266–275. https://doi.org/10.1093/bfgp/elp029

15. Jindal, G. A., & Farley, E. K. (2021). Enhancer grammar in development, evolution, and disease: Dependencies and interplay. *Developmental Cell*, *56*(5), 575–587. https://doi.org/10.1016/j.devcel.2021.02.016
16. Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, *10*(1), 1930. https://doi.org/10.1038/s41467-019-09982-5
17. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17), 2204–2207. https://doi.org/10.1093/bioinformatics/btq351
18. Kvon, E. Z., Kamneva, O. K., Melo, U. S., Barozzi, I., Osterwalder, M., Mannion, B. J., Tissières, V., Pickle, C. S., Plajzer-Frick, I., Lee, E. A., Kato, M., Garvin, T. H., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Rubin, E. M., Dickel, D. E., Pennacchio, L. A., & Visel, A. (2016). Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell*, *167*(3), 633-642.e11. https://doi.org/10.1016/j.cell.2016.09.028
19. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923
20. Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, *35*(3), 421–432. https://doi.org/10.1093/bioinformatics/bty648

21. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, 10--12. http://dx.doi.org/10.14806/ej.17.1.200

22. McGhee, J. D., & Felsenfeld, G. (1980). *Nucleosome Structure*. 42.

23. Murrell, A., Rakyan, V. K., & Beck, S. (2005). From genome to epigenome. *Human Molecular Genetics*, *14*(suppl_1), R3–R10. https://doi.org/10.1093/hmg/ddi110

24. Prachumwat, A., & Li, W.-H. (2008). Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. *Genome Research*, *18*(2), 221–232. https://doi.org/10.1101/gr.7046608

25. Quillien, A., Abdalla, M., Yu, J., Ou, J., Zhu, L. J., & Lawson, N. D. (2017). Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq. *Cell Reports*, *20*(3), 709–720. https://doi.org/10.1016/j.celrep.2017.06.070

26. Shah, R. N., Grzybowski, A. T., Cornett, E. M., Johnstone, A. L., Dickson, B. M., Boone, B. A., Cheek, M. A., Cowles, M. W., Maryanski, D., Meiners, M. J., Tiedemann, R. L., Vaughan, R. M., Arora, N., Sun, Z.-W., Rothbart, S. B., Keogh, M.-C., & Ruthenburg, A. J. (2018). Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. *Molecular Cell*, *72*(1), 162-177.e7. https://doi.org/10.1016/j.molcel.2018.08.015

27. Smith, E., & Shilatifard, A. (2014). Enhancer biology and enhanceropathies. *Nature Structural & Molecular Biology*, *21*(3), 210–219. https://doi.org/10.1038/nsmb.2784

28. Smith, J. J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M. C., Parker, H. J., Cook, M. E., Hess, J. E., Narum, S. R., Lamanna, F., Kaessmann, H., Timoshevskiy, V. A., Waterbury, C. K. M., Saraceno, C., Wiedemann, L. M., Robb, S. M. C., Baker, C.,

Eichler, E. E., Hockman, D., … Amemiya, C. T. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nature Genetics*, *50*(2), 270–277.

29. Soares, L. M., He, P. C., Chun, Y., Suh, H., Kim, T., & Buratowski, S. (2017). Determinants of Histone H3K4 Methylation Patterns. *Molecular Cell*, *68*(4), 773-785.e6. https://doi.org/10.1016/j.molcel.2017.10.013
30. Tahara, Y. (1988). *Normal Stages of Development in the Lamprey, Lampetra reissued (Dybowski)*. 10.
31. Timoshevskiy, V. A., Herdy, J. R., Keinath, M. C., & Smith, J. J. (2016). Cellular and Molecular Features of Developmentally Programmed Genome Rearrangement in a Vertebrate (Sea Lamprey: Petromyzon marinus). *PLOS Genetics*, *12*(6), e1006103. https://doi.org/10.1371/journal.pgen.1006103
32. Tsai, Z. T.-Y., Tsai, H.-K., Cheng, J.-H., Lin, C.-H., Tsai, Y.-F., & Wang, D. (2012). Evolution of cis-regulatory elements in yeast de novo and duplicated new genes. *BMC Genomics*, *13*(1), 717. https://doi.org/10.1186/1471-2164-13-717
33. Wittkopp, P. J. (2006). Evolution of cis-regulatory sequence and function in Diptera. *Heredity*, *97*(3), 139–147. https://doi.org/10.1038/sj.hdy.6800869
34. Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. https://doi.org/10.1038/nrg3095
35. Wong, E. S., Zheng, D., Tan, S. Z., Bower, N. I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B. M., Kikuchi, K., McGlinn, E., Francois, M., & Degnan, B. M.

(2020). Deep conservation of the enhancer regulatory code in animals. *Science*, *370*(6517), eaax8137.

36. York, J. R., Thresher, R. E., & McCauley, D. W. (2021). Applying functional genomics to the study of lamprey development and sea lamprey population control. *Journal of Great Lakes Research*, *47*, S639–S649. https://doi.org/10.1016/j.jglr.2020.03.010
37. York, J. R., Yuan, T., & McCauley, D. W. (2020). Evolutionary and Developmental Associations of Neural Crest and Placodes in the Vertebrate Head: Insights From Jawless Vertebrates. *Frontiers in Physiology*, *11*, 986. https://doi.org/10.3389/fphys.2020.00986
38. Yu, J.-K. S. (2010). The evolutionary origin of the vertebrate neural crest and its developmental gene regulatory network – insights from amphioxus. *Zoology*, *113*(1), 1–9. https://doi.org/10.1016/j.zool.2009.06.001
39. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Appendix

The code used for this project, including both the pipeline and overlap algorithm, are available at https://github.com/cbellian/Honors-Thesis. The bigWig track files are available at https://genome.ucsc.edu/s/CBellian/petMar3. The zebrafish CHiP-Seq data was accessed from https://danio-code.zfin.org/dataExport/?view=table&selected_facet=ChIP-seq-assay_type and visualized using the public track hub published on the UCSC Genome Browser. Cutadapt can be accessed at https://cutadapt.readthedocs.io/en/stable/. The bowtie2 alignment tool is available at http://bowtie-bio.sourceforge.net/bowtie2/index.shtml. The bedGraphtoBigWig tool can be downloaded at http://hgdownload.soe.ucsc.edu/admin/exe/macOSX.x86_64/. MACS2 can be derived from the MACS3 repository at https://github.com/macs3-project/MACS/. The genome browser used was the UCSC genome browser, available at https://genome.ucsc.edu/. Finally, the Petromyzon marinus reference genome can be downloaded at https://www.ncbi.nlm.nih.gov/assembly/GCA_002833325.1.