MultiChipVis

Alexander Adrian-Hamazaki*, Rodrigo Conceição**, Yerin Kim*

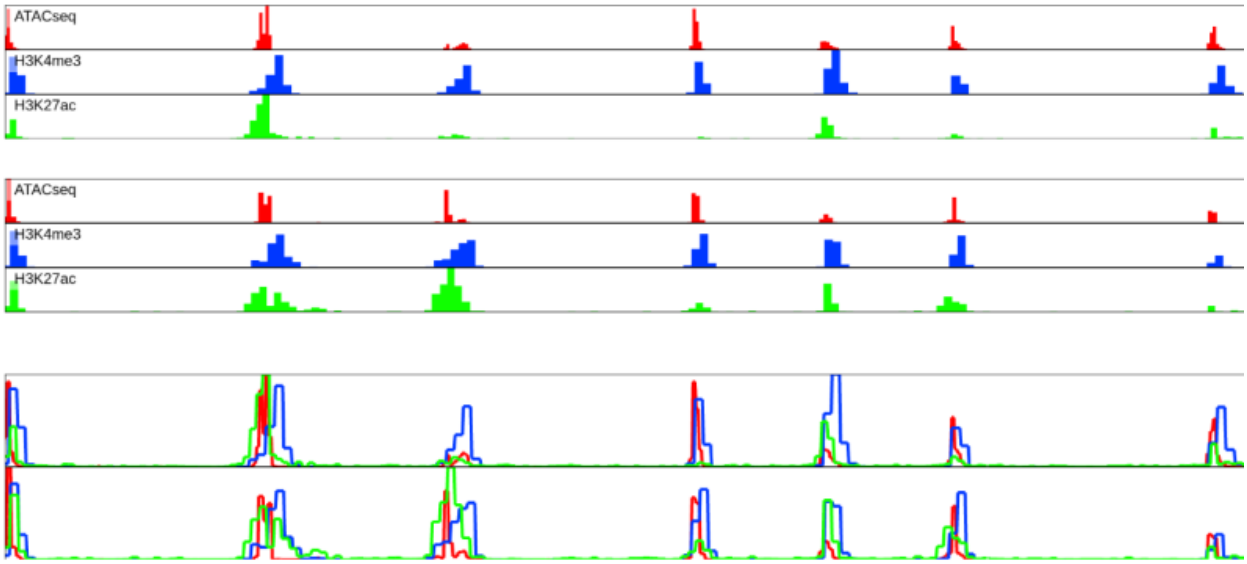

Fig. 1. Easily view ChIP-seq and ATAC-seq data aligned to a genome

Abstract— Every single one of our cells contains molecules of DNA that are unique to us. Together, these molecules, also called our genomes, make us who we are. Although every one of our cells has an identical genome, not all of our cells are identical. Indeed, our cells are able to differentiate into different cell types, such as Neurons, or Muscle cells. Our cells are able to do this because each cell type *regulates* its genes differently. Unraveling the mystery of gene regulation is essential for understanding the functions of these cells. Chromatin immuno-precipitation followed by sequencing (ChIP-seq), and The Assay for Transposase-Accessible Chromatin (ATAC-seq) are novel methods that researchers use to study gene regulation. Several tool-kits exist to visualize these types of data; however, there is a noticeable lack of a tool that enables researchers to compare data across experimental conditions. We present MultiChipViewer, a tool-kit that allows users to visualize ChIP-seq and ATAC-seq data, and empowers users to compare data from different experimental conditions.

Index Terms—ChIP-seq, Genome Regulation

1 Introduction

Our body is made of trillions of cells containing 48 identical molecules of DNA that store our unique genetic code [3]. In each one of these cells, there is a molecule of DNA with a genetic code unique to us. Although each one of these cells contains an identical molecule of DNA, we are somehow comprised of different cell types. Take Neurons, for example, the cell type that comprises the brain. Not only do Neurons look differently on the outside compared to skin cells, but it also looks different on the inside. On the inside of both of these cell types, there are several hundreds of thousands of molecules called mRNA [3]. These molecules are created by reading and copying the genetic code of the genes that are on their DNA. In this way, each mRNA is associated with a specific gene. Our different cell types look different on the outside because, on the inside, the amount of each of these mRNA molecules present in them is different [3]. This is possible because both cell types are able to regulate the amount of each mRNA they make. Investigating how the cells are able to regulate mRNA production has been a key problem in genomics that has caught a lot of recent attention.

In the past decade, new technologies such as Chromatin immuno-precipitation followed by sequencing (ChIP-seq), and The Assay for Transposase-Accessible Chromatin (ATAC-seq) have exponentially increased our understanding of how cells are able to regulate mRNA production [13]. These technologies have rapidly increased our understanding because they operate on the genomic level, meaning they are able to investigate the regulation of the entire genome at once. This compares dramatically to older methods, which were only capable of investigating regulation at specific locations that are pre-selected by experimenters.

In order to understand why ChIP-seq and ATAC-seq can investigate regulation, it is important to understand how gene regulation works. In brief, cells have hundreds of thousands of molecules called **Proteins** inside them. These proteins are the workhorse of the cell; they are the molecules that actually perform chemical functions for the cell, and give cells their overarching function [3]. On the other hand, genes have specific regions of interest around them called **Promoters** and **Enhancers**. Each gene has its own promoter and has its own enhancers [4]. These are special regions because proteins are able to interact with them. Depending on *what* proteins interact with these promoters and enhancers, different outcomes can occur to the regulation of their gene. For example, suppose Protein X interacts with the promoter of Gene A, causing more Gene A mRNA to be produced. However, if Protein Y binds to an enhancer of Gene A, this may reduce the amount of Gene A mRNA that is produced. Thus, we see that the regulation of

- *Department of Bioinformatics, UBC
- **Department of Applied Animal Biology, UBC

Gene A mRNA is a result of many complex combinations of proteins interacting to the gene's promoter and enhancers.

The biological knowledge to understand why ChIP-seq and ATAC-seq are able to investigate gene regulation is complex. In brief, ChIP-seq investigates regulation by scanning the *entire genome* to identify *where* one protein interacts with the genome, and how often it interacts with that area [13]. The frequency of its interaction is called the **ChIP-seq Signal**, or **ATAC-seq Signal** (Signal). Using our same example of Protein X, if we performed ChIP-seq *for* Protein X, we would find that it interacts with the promoter of Gene A. We might also get a very large ChIP-seq Signal, indicating that this interaction happens very frequently. ATAC-seq is similar in that it scans the entire genome; however, instead of scanning the genome for a specific protein, it scans the entire genome for the *lack* of a specific type of protein called **Histones** that are known *prevent* the expression of mRNA [2].

To allow researchers to cross-talk about locations in the genome, a standardized **Genomic Coordinate** system has been adopted [5]. All cells have their DNA packaged into units called Chromosomes [3]. For example, female mice have 20 chromosomes in each of their cells. Furthermore, each one of these chromosomes is made up of DNA. The smallest unit that a DNA molecule is made up of is called a Base Pair (BP) [3]. In the **genomic coordinate** system, every position in the genome is indexed with two values: 1) The chromosome number, 2) The number of Base Pairs (BPs) from the start of that chromosome. For example, the Hes1 gene (an arbitrarily selected gene) is on Chromosome 16 and begins at the 30,065,340 base pair from the start of Chromosome 16. One can also specify a range of genomic coordinates. For example, The Hes1 gene's location is Chromosome 16, which begins at 30,065,340 BPs from the start, and ends at 30,067,796 BPs from the start [7].

Although there are several fully functional tool-kits that allow the user to explore and visualize ChIP and ATAC-seq data, there lacks a tool-kit that allows researchers to compare data across different **experimental conditions**. *Experimental condition* is the jargon used to describe the different controlled experiments that researchers perform. For example, a researcher might be performing experiments in Heart and Liver tissue. In this case, the experimental conditions are *Heart* and *Liver*. In another case, a researcher might be exploring the development of an embryo. To study this development, they might sample the embryos throughout development. In this case, the experimental conditions are *every time-point that the embryos are sampled at*. With MultiChipViewer, we hope to empower researchers with a tool to compare ChIP-seq and ATAC-seq data across different experimental conditions.

2 Related Works

Several visualization tool-kits exist which can visualize ChIP and ATAC-seq data. The common paradigm between all of them is that ChIP and ATAC-seq Signals are vertically, or horizontally aligned to the genome of interest. These Signals are typically then encoded as line marks in the form of bars, where tall bars indicate a large Signal at that genomic coordinate.

The simplest of the tool-kits is the Genome Browser [8]. This browser is a web browser that is integrated with ENCODE, a leading consortium in the field [5]. It enables the user to visualize the ChIP or ATAC-seq Signals of one of the ENCODE data sets. Notably, this toolkit does not enable researchers to visualize their *own* data. The visualization primarily focuses on gaining insight into the ChIP and ATAC-seq Signals at and around a specific gene. This focus is supported by the ability to input a name of a gene of interest and jump the visualization to that gene. Oftentimes, researchers are interested in the area surrounding their gene of interest, so they can identify interesting features such as Enhancers they were unaware of. To support this, the user can pan and zoom into the data. The Genome Browser only visualizes *one* data set at a time. Despite this, the visualization is somewhat confusing at first glance. There are several different hues that are used to categorize the data; however, these hues do not aid in the separation of the data. Consequently, it is difficult to use Genome Browser to determine if specific ChIP or ATAC-seq peaks overlap with specific genomic regions, or if specific ChIP and ATAC-seq peaks overlap with other ChIP and ATAC-seq peaks.

SeqCode [1] is a command-line-based visualization tool with an associated web-based viewer. It is a simple method to view ChIP and ATAC-seq Signals, as well as other diagnostic data at the gene level [1]. SeqCode is somewhat unique because its choice of hue and saturation are carefully and thoughtfully implemented. As a result, the user is to able easily determine which ChIP and ATAC-seq peaks overlap with specific areas of the genome, and which ChIP and ATAC-seq peaks overlap with other ChIP and ATAC-seq peaks. SeqCode seeks to further enable the user to identify ChIP and ATAC-seq peaks that overlap by allowing the user to overlay ChIP and ATAC-seq Signals as line graphs. Although SeqCode is a low-barrier-to-entry tool, it is limited in that it is only able to generate images. It lacks the ability to scan through the genome, or to explore the region surrounding a gene of interest. This lack of interaction may degenerate the user's ability to explore the data. For example, they may fail to identify that a protein of interest is binding to an enhancer, and affecting gene regulation if they did not know that the enhancer existed beforehand.

EaSeq [9] is a software that supports comprehensive data exploration of genomic data such as ChIP and ATAC-seq [9]. EaSeq is unique in that it allows the user to compare ChIP or ATAC-seq data from two different experimental conditions. For example, with EaSeq, we could compare how the ChIP-seq Signals differ between a Cancer Cell and a Neuron Cell. EaSeq enables this by visualizing the *difference* in Signal between the two experimental conditions. This difference is then aligned to the genome. The ability to compare ChIP and ATAC-seq Signal between experimental conditions are underdeveloped in the majority of visualization toolkits. Oftentimes, researchers are not interested in the ChIP or ATAC-seq Signals in one experimental condition. Instead, they are typically interested in how Signals *change* between experimental conditions. In viewers like SeqCode, this is done by comparing the visualizations from different experimental conditions. EaSeq is forward-thinking because it allows users to directly compare experimental conditions within the toolkit itself; however, it is limited as it only enables comparisons between the two. Another limitation of EaSeq is the complexity afforded to it by its comprehensiveness. Because EaSeq offers so many tools to the user, it is not suitable for a user who wants a rapid understanding of their data and does not wish to invest the significant time required to learn how to use the tool.

Although there are several fully functional tool-kits that allow the user to explore and visualize ChIP and ATAC-seq data, there are several improvements that can be made. With MultiChipViewer, we hope to remedy some of the downsides of existing tools by creating a tool with 1) A low barrier to entry. 2) That allows users to upload their own data. 3) That allows the user to explore the surrounding area of their gene of interest. 4) That enables the user to *easily* explore the overlap of ChIP and ATAC-seq Signals. 5) That enables the user to directly compare multiple experimental conditions.

3 Data Abstraction

When a researcher performs a "ChIP-seq or ATAC-seq experiment" what exactly they mean often varies, and is contingent on their specific questions of interest. Researchers often say they "performed ChIP-seq on brain cells"; however, this claim is incomplete and lacks granularity. What the researcher means to say is that they "performed ChIP-seq on brain cells, looking for Protein X, Protein Y, and Protein Z". In reality, they have performed several experiments. For example, this researcher might have performed ChIP-seq on brain cells for the three proteins, H3K4me3, H3K27ac, and H3K36me3. Each protein requires an independent ChIP-seq experiment, and each ChIP-seq experiment yields one file and one data set. There are thousands of unique proteins that can be assessed with ChIP-seq; however, researchers are not typically interested in all proteins at a given time, instead focusing on only a handful. On the other hand, ATAC-seq is much simpler. Because ATAC-seq always measures the lack of histone proteins at genomic coordinates, one ATAC-seq experiment always yields one data set.

A researcher might also be interested in comparing ChIP-seq or ATAC-seq data sets across various experimental conditions. For ex-

ample, they might be interested in how the H3K4me3, H3K27ac, and H3K36me3 proteins, and ATAC-seq Signals change between the Heart and Liver. In this instance, the researcher would have to perform eight experiments. Three ChIP-seq experiments would be performed in the Heart, followed by one ATAC-seq experiment in the Heart. Then these four experiments would have to be performed in the Liver. This total project would result in eight data sets. A researcher might also be interested in how the HH3K4me3, H3K27ac, and H3K36me3 proteins, and ATAC-seq Signals change across several time points; for example 1 hour after drug treatment, 2 hours after drug treatment, and so on. In this instance the researcher would perform the same four experiments for each time point of interest, leading to 4 data files per time point.

ChIP-seq and ATAC-seq data sets are typically stored in one of two formats, BED-narrowpeak or BigWig. BED-narrowpeak files are comma-delimited files that contain 10 columns with specific names [12]. There are four fundamental BED-narrowpeak columns [12]. They indicate:

1. The Signal strength of an ATAC or ChIP-seq peak
2. The chromosome containing the peak
3. The starting genomic coordinate of the peak on the given chromosome
4. The end genomic coordinate the peak on the given chromosome

The remaining columns in BED-narrowpeak files pertain to quality metrics. These BED-narrowpeak files can be quite large, typically exceeding 2Gb. Consequently, the BigWig format is typically used when ChIP and ATAC-seq data are meant to be visualized. BigWig files contain the four fundamental BED-narrowpeak columns [12]. They are stored in an indexed binary format; when using this format, only the genomic regions that are being visualized are sent to the display [12]. This leads to a considerable performance increase [12].

The range in magnitude of the ChIP-seq Signal can vary depending on the protein of interest that the ChIP-seq was performed for. For example, ChIP-seq Signals for Protein X may vary from 0-1000 (arbitrary unit). Whereas the ChIP-seq Signals for Protein Y may vary from 0-50. ATAC-seq Signals do not vary as considerably, typically ranging from 0-1000.

The range of genomic coordinates of ChIP-seq and ATAC-seq are contingent on the organism that a ChIP or ATAC-seq experiment is performed on. For example, much of the science surrounding genomic regulation is performed in Mice and Human cells. Male and female mice have 21 and 20 unique chromosomes respectively [6]. This difference is due to males having one copy of the Y Chromosome, and one copy of the X Chromosome; whereas female mice have two copies of the X Chromosome. Thus, the possible unique chromosome values of a male mouse can range from Chromosome 1-19, Chromosome Y, and Chromosome X. In humans, males and females have 24 and 23 unique chromosomes respectively. Each chromosome in each organism varies in length. However, they are typically in the tens or hundreds of millions of base pairs [6].

When performing ChIP and ATAC-seq, researchers are typically interested in the regulation of *genes*. Mice and Humans both have approximately 25000 genes [7]. These genes are found on the genome, the genome is made up of DNA, and DNA is actually a double-stranded molecule. These two **strands** are often called the **Plus** and **Minus** strands. A given gene is found on the Plus *or* Minus strand. Furthermore, genes on the Plus strand all share the same *direction*. This *direction* gives an understanding of where we might expect certain special proteins to interact with the gene. The *directions* that the genes on a strand are arbitrary; however by convention, genes on the *Plus* strand are going to the *right*, and genes on the *Minus* strand go to the *left*. Finally, genes have *lengths*. The *length* of the gene is described by the difference between its genomic stop coordinate and its genomic start coordinate. For example, the Hes1 gene is found on the Plus strand [7]. Its genomic coordinates are Chr:30,065,340-30,067,796, giving it a gene length of 2456 BPs. To fully communicate the location of a gene in the genome, we need to know the genomic coordinates of the gene, and the strand that it is on. GenCode [7], is one of the leading groups that have annotated the genomic coordinates and strand information for Mice and Human genes.

To fully understand the regulation of a Gene, researchers need to know if ChIP and ATAC-seq peaks overlap with Promoters and Enhancers. Unfortunately, the field of genomics does not know the exact Genomic Coordinates of all Promoters and Enhancers. The SCREEN project by ENCODE [11], is the field's leading project to identify all Promoters and Enhancers. Their data is provided in a BigWig format. The files contain the three fundamental columns required for a Genomic Coordinate (chromosome number, start, and stop location), but do not have strand information. Similarly to Genes, Promoters and Enhancers have a *length* that can be described by taking the difference between the *start* and *stop* genomic coordinates. Contrarily to Genes, Promoters and Enhancers are thought to be located on *both* of the DNA strands [11]. Consequently, Promoters and Enhancers do not have a notion of *direction* Thus, the data provided by SCREEN fully describes where Promoters and Enhancers are found in the genome.

4 Task Abstraction

When researchers perform ChIP and ATAC-seq, the space of questions that they are trying to solve is large. For example, some questions may be related to cancer, while others may be related to embryonic development. They may be interested in specific questions pertaining to these genes such as, "does Protein X interact with Promoter Z"; however, they may be asking more general questions such as "what does Protein X interact with". What is common across most of these questions is that the researchers are often interested in what a *particular set of proteins* are doing. The set of proteins is typically fairly small, often ranging from 0-10, but it could be more. Recall that when you perform *one* ChIP-seq experiment, you can only determine the genomic coordinates of *one* protein. Thus, to add more proteins to your investigation is to perform another ChIP-seq experiment. Although some researchers may be interested in ten or more ChIP-seq protein targets, we are less ambitious with our visualization. MultiChIPViewer viewer will focus on visualizing ChIP and ATAC-seq data for researchers who are targeting six or fewer proteins. This is the first task that MultiChiPViewer must handle.

1. Allow users to upload up to six of their own ChIP-seq or ATAC-seq data-set types

Similarly to proteins, when researchers perform ChIP and ATAC-seq, they are typically interested in how a protein interacts with a set of genes. Because ChIP and ATAC-seq are methods that generate data for the *entire* genome, the set of genes can range in orders of magnitudes from tens of genes, to thousands, to the entire genome. Furthermore, researchers may also be interested in the relationship between two or more genes in their gene set of interest. They may ask questions such as "does my ChIP-seq Signal decrease at Gene X while increasing at Gene Y". Although researchers may be investigating more than one gene at a time, the focus of this visualization tool will be to investigate one gene at a time. This same heuristic is applied to all existing ChIP-seq viewing tools [9] [7] [8]. Focusing on only one gene at a time allows us to focus on comparing how one gene changes across experimental conditions. This focus highlights the second task that the user must be able to perform in MultiChIPViewer:

2. Allow navigation to specific genes of interest.

In order for ChIP and ATAC-seq Signals to be meaningful, the user must be able to identify if the Signals overlap with Genes, Enhancers, Promoters, and other Signals *within* the experimental condition. If a Signal overlaps with Genes, Enhancers, or Promoters, often the user is interested in how *strong* that Signal is. This can be represented by the magnitude of the Signal relative to the maximum Signal value for that ChIP-seq data set. Thus, the following task much is achievable through MultiChIPViewer:

3. Evaluate if a ChIP or ATAC-seq Signal is found at the same Genomic Coordinates of a Promoter, Enhancer, or Gene. If so, evaluate the magnitude of that Signal relative to the largest Signal in the data set.

If a Signal overlaps with another Signal *within* the same experimental condition, we are not interested in *comparing* magnitudes of the two Signals; however, we are still interested in the magnitude of both Signals relative to the maximum value of their respective data-sets. With this in mind, we have identified the following tasks that the user must be able to perform in MultiChIPViewer:

4. Evaluate if a ChIP or ATAC-seq Signal is found at the same Genomic Coordinates as another ChIP or ATAC-seq Signal *within* an experimental condition. If so, evaluate the magnitudes of each Signal relative to the largest Signal in their respective data sets.

A novel task that MultiChIPViewer is seeking to handle is allowing the user to compare ChIP and ATAC-seq data from multiple experimental conditions at once. Users will be most interested in determining if there are Signals that overlap between experimental conditions. If so, users will seek to evaluate if the magnitude of the *same type* of Signal has changed between conditions. For example, a user might be interested in determining if an ATAC-seq Signal at a specific genomic coordinate decreased between the Liver and Brain experimental conditions. The user might also then be interested in seeing if an H3K4me3 Signal at a specific genomic coordinate increased. The user will not be interested in comparing the ATAC-seq Signal in the Liver, and the H3K4me3 Signal in the Brain. This leads to MultiChIPViewer to its imperative task:

5. Evaluate if a ChIP or ATAC-seq Signal is found at the same Genomic Coordinates as the *same type* of ChIP or ATAC-seq Signal across experimental conditions. If so, evaluate if the Signal decreases or increases across experimental conditions.

Furthermore, we found that the main weakness of SeqCode was its lack of interaction. We believe that researchers may be interested in exploring the genomic coordinates surrounding a gene of interest. There are often elements of interest surrounding a gene of interest, such as an Enhancer, that the researcher is unaware of. If the user has to preselect the exact genomic coordinates to view, this makes it difficult for the user to discover these unknown artifacts. Thus, MultiChIPViewer must handle the sixth task:

6. Allow users to explore the surrounding genomic coordinates of a gene of interest

5 Solution

5.1 Data Loading

The first focus of our solution was to develop a solution to enable users to upload their own data. Although many options could have been taken, we decided to use a web interface, where users are able to specify experimental conditions, and ChIP or ATAC-seq data sets to add to those experimental conditions (Fig. 2A). This toolbar was kept as simple as possible. Notably, users must specify the color of a ChIP or ATAC-seq data set when they are loading it (Fig. 2B). Users are meant to choose one color for one type of ChIP-seq data set. For example, all ChIP-seq data sets across experimental conditions will be *Red*, and all ATAC-seq data sets across all experimental conditions will be *Blue*. The user can select from six, highly contrasting colors. All colors are highly saturated. Consequently, users can compare a total of six different ChIP or ATAC-seq data types. On the other hand, there is no upper limit for how many experimental conditions the user can compare.

5.2 Gene-Level Navigation

Because MultiChIPViewer is designed for users to investigate specific genes of interest, users are able to jump to their gene of interest by searching for it in the toolbar (Fig. 2C),

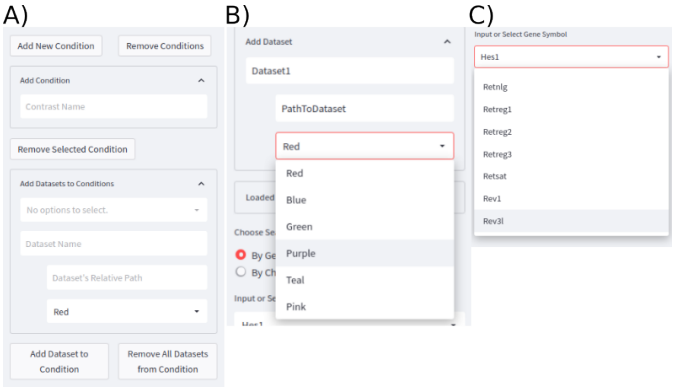

Fig. 2. MultiChIPViewer toolbar. Names of experimental conditions are added by typing them in the "Add Condition" box. And selecting "Add New Condition". ChIP-seq and ATAC-seq data sets can be added to an experimental condition by selecting the experimental condition, adding a Data-set Name, the path to the data file, selecting a color, and then clicking "Add Dataset to Condition. The user can navigate to genes of interest by typing in the name of the gene into the search bar.

5.3 Evaluating Signal Overlap

In tasks 2 and 3 and 4, the user must be able to determine if a Signal at specific genomic coordinates overlaps with a genomic region of interest or another Signal. To allows users to assess overlap, MultiChIPViewer takes a vertical-alignment approach.

5.3.1 Hard-Coded Genomic Coordinates

Genomic coordinates are laid out horizontally at the top of the visualization (Fig.3A). The genomic coordinates serve as a foundation to which all other marks are vertically aligned.

Underneath the genomic coordinates, there is another horizontally laid section containing Genes (Fig.3B). The genes are represented as horizontally laid line marks with some volume (a horizontal bar). All genes are colored grey. The genes are vertically aligned with the *foundational genomic coordinates* such that the genomic coordinates of the beginning of the gene are encoded by a rectangular edge. On the other hand, the genomic coordinates of the end of the gene are encoded by the tip of a triangular edge. In this way, the *direction* of the gene is also encoded by the triangular edge. As the rectangular end encodes the *start* and the triangular end encodes the *end*, the *length* of the line mark can be thought of as encoding the *length* of the gene. The genes are separated into two rows. Genes on the top row are on the Plus strand, and genes on the bottom row are on the Minus strand. Gene labels are placed on top of the grey bars.

Beneath the genes are horizontal sections which contain Enhancers and Promoters tracks (Fig.3C). These sections are also vertically aligned with the *foundational genomic coordinate system*. The Enhancers and Promoters are encoded as horizontal line marks with identical widths (horizontal bars). Enhancers are colored as a medium-saturated orange, whereas promoters are encoded with a medium-saturated yellow. Because Enhancers and Promoters are not on a specific Plus or Minus strand, they do not have a *direction* is no need to have two "rows". Because Enhancers and Promoters do not have a notion of *direction*, the *stop* genomic coordinate is not encoded by a triangular shape. Instead, the rectangular ends of the Enhancers and Promoter encode their *start*, and *stop* genomic coordinates. In this way, the length of the line marks encodes the *length* of the Enhancers and Promoters.

Because the tasks of assessing the ChIP and ATAC-seq overlap with Genes, Promoters, and Enhancers, the genomic coordinates, genes, enhancers, and promoters are hard-coded into the visualization. They cannot be removed, and their colors cannot be changed.

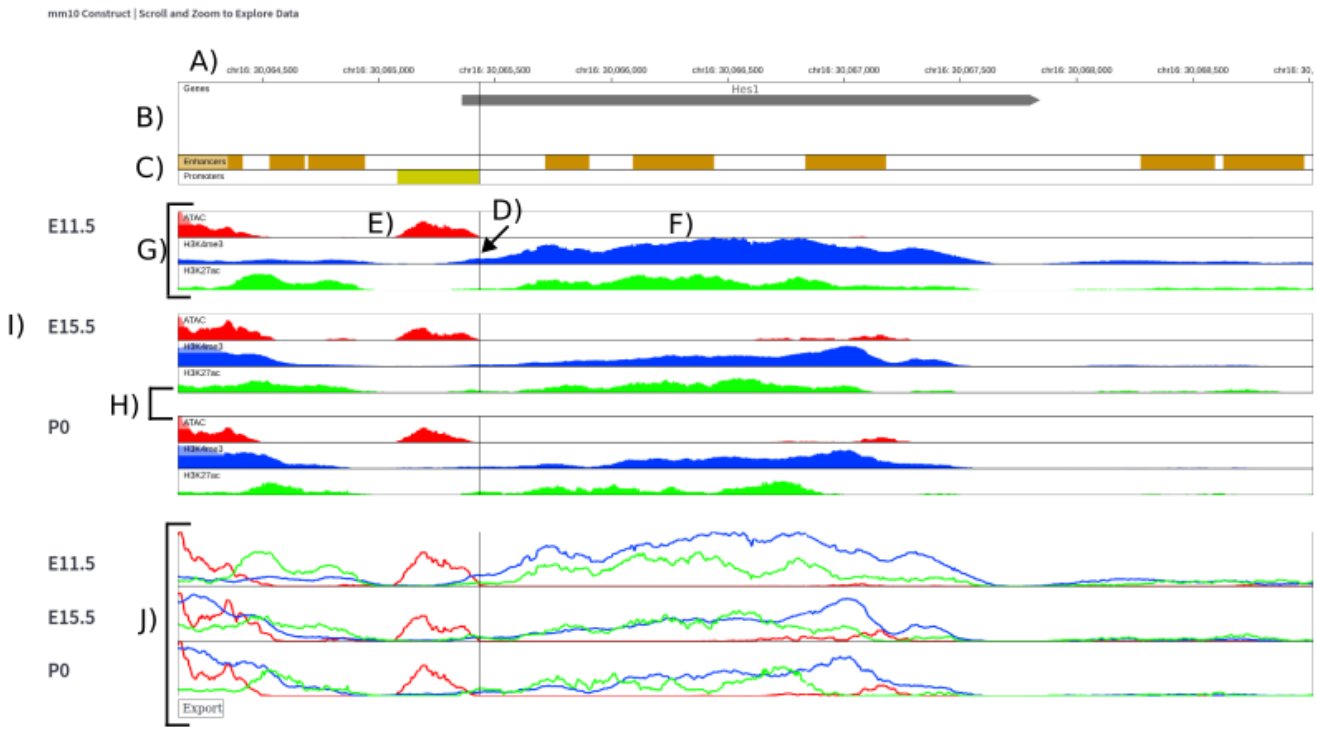

Fig. 3. An overview of MultiChIPViewer. **A**: The fundamental genomic coordinate system that all data is aligned to. **B**: A section containing genes at the current genomic coordinates. The Hes1 gene is on the top row, indicating that it is on the Plus strand of DNA. **C**: The Enhancers and Promoters at the current genomic coordinates. **D**: Moving the mouse cursor across the visualization creates a vertical line to help determine overlaps in the Signal. **E**: The red ATAC-seq peak *pops* due to its high saturation and contrast with the other hues. **G**: ChIP-seq and ATAC-seq data sets within an experimental condition are vertically aligned with no separating white space. **H**: Experimental conditions are vertically aligned with a small vertical white space. **I**: Labels of the experimental conditions. **J**: In the *merged view*, ChIP-seq and ATAC-seq data sets from each experimental condition are overlaid, and vertically stacked with no separating white space.

5.3.2 Vertical Scanning and Color

Following the vertical-alignment heuristic, ChIP or ATAC-seq data sets are automatically vertically aligned to the foundational genomic coordinate system (Fig.3D). Because data sets are vertically aligned, it enables the user to vertically scan the genomic coordinate system to identify if Signals of interest overlap with Genes, Enhancers, Promoters, or other Signals. To further enable vertical scanning, the user's mouse creates a vertical line across the entirety of the visualization (Fig.3D).

Color encoding is a major enabler in assessing the overlap of Signals. ChIP and ATAC-seq data sets are loaded into MultiChIPViewer with one of six highly saturated colors that are strongly contrasting hues. This high saturation and strong contrast allow specific peaks of interest to *pop-out* to the user. For example, if the user is interested in the ATAC-seq Signal at 30,065,250 (Fig.3E), the user can instantly locate Signal without becoming confused by non-ATAC-seq data by merely focusing on the *Red* data. The strong contrast in hues results in good *data separation*, allowing the users to easily determine that the *Green* Signal overlaps with the *Blue* Signal (Fig.3E). Notably, the saturation of the Genes, Enhancers, and Promoters is lower than the saturation of the ChIP and ATAC-seq data sets (Fig.3C). This decision was made to suggest to the user the data types are different. These color encodings were imperative, as exemplified by the Genome Browser, which suffers from a lack of data separation and pop-out due to poor implementation of color.

5.3.3 Assessing Signal Overlap Across Experimental Conditions

Task 4 requires that users are able to compare Signals from multiple experimental conditions. ChIP and ATAC-seq Signals from all experimental conditions are vertically aligned to the *foundational genomic coordinate system*.

The data sets in a given experimental condition are vertically aligned with no separating white space (Fig.3G). Then, each experimental condition is vertically aligned with white space (Fig.3H). The goal of this white space is to enable users to distinguish between experimental conditions. If the user is able to distinguish the different experimental conditions, then the user will be primed to compare the different experimental conditions. The names of the experimental conditions inputted by the user are also displayed (Fig.3I).

Ideally, the users have loaded their data such that all data sets of the same time are the same color. For example, all ATAC-seq data sets are Red, whereas all H3K4me3 data sets are Blue. If the user loads their data in this way, then color can be used to *pop-out* the *same type* of data-set from the different experimental conditions. For example, if users are interested in determining if an ATAC-seq Signal is present in E11.5, E15.5, and P0 they will be able to focus only on the *Red* data sets.

5.3.4 The Merged View

To help users identify if peaks are overlapping within an experimental condition, or across experimental conditions, they can look at the *merged view* (Fig.3J). In the merged view, all ChIP and ATAC-seq data from a given experimental condition are collapsed onto one horizontal section. The different experimental conditions are then vertically aligned with no separating white space. The types of ChIP and ATAC-seq data remain stratified by their selected color. The Signal strength of all of the ChIP and ATAC-seq tracks are normalized in the same way, such that they are directly comparable.

Using the merged view can also help users determine if Signals overlap *across* experimental conditions. The lack of white space reduces the time users spend searching for specific experimental conditions

or Signals of interest. Allowing for more direct comparisons across experimental conditions

5.4 Evaluating Signal Magnitude

Upon identifying a Signal of interest, a user will want to know the strength of the Signal. Furthermore, if a user is comparing across experimental conditions, the user will want to know how the strength of that signal changes. In MultiChIPViewer, the Signals from one ChIP-seq data set in one specific experimental condition are normalized to the maximum Signal magnitude of the ChIP-seq data set found within the surrounding area that the user is visualizing. For example, in figure3, the ATAC-seq Signals in E11.5 are normalized to the maximum Signal value of the data set found in the *surrounding 100,000 BPs* in both directions. This is a notable deviation from the initial Tasks (Tasks 2, 3, and 4), as the initial tasks sought to compare the Signal magnitudes relative to the maximum values in each data set. The decision was made to normalize to the surrounding area because this is the default in Gosling, and it still provides users with an idea of the Signal strength.

To further enable users to identify the magnitude of a Signal, users are able to mouse over a Signal of interest. This opens a tooltip, which provides the genomic coordinates, and raw magnitude of the Signal (not shown).

5.5 Exploration of Surrounding Genomic Coordinates

To enable the exploration of data, the visualization enables linked panning and zooming of the visualization. The Gene, Enhancer, Promoter, and ChIP or ATAC-seq tracks are all linked such that they zoom and pan together. This prevents the disruption of the much-required vertical alignment. To get an overarching view of a Gene, or even a larger region, the user can zoom out. Although there is no upper limit to zooming, it is not recommended that the user zooms out too much, as it results in too much data being loaded at once, significantly reducing the performance of the system. The user can zoom in until the viewer is several hundred BP's in width. This enables the user to get information about specific genomic coordinates of elements of interest. To further enable the identification of specific genomic coordinates of interest, mousing over an element of interest creates a toolkit that displays vital information such as Signal strength, and genomic coordinates.

6 Implementation

This visualization was created in Python using a package called Gosling [10]. To visualize our Gosling visualization, we used Streamlit [14]. Gosling has created a component to easily view Gosling visualizations in Streamlit [15]. Using Streamlit, we built an interactive web application to visualize our Gosling visualization.

Gosling is a package whose specific goal is to enable users to create genomic visualization tools. Specific marks are available, such as the line marks used to encode Gene length. Importantly, the ability to align files containing genomic coordinates to a *fundamental genomic coordinate system* is already implemented in Gosling. Much of the initial development of the visualization tool was understanding the lackluster documentation. Upon mastery of the Gosling syntax, utilizing the package to create the visualization came easily. To create the Hard-coded Gene, Enhancer, and Promoter sections, the Gene, Enhancer, and Promoter files are read by Gosling. Using a Gosling command (gos.vertical), the three data types are aligned to the genomic coordinate system, and they are linked for linked zooming and panning.

The toolbox on the left side of the visualization is created using Streamlit. Once the user inputs their experimental conditions, and data-sets of interest, and the colors they desire those data sets to be visualized as. Streamlit passes these values to Gosling. Gosling then begins the process of opening the data sets, and vertically aligning ChIP and ATAC-seq data sets within the same experimental condition, vertically aligning the experimental conditions to each other, and then vertically aligning the entire system to the *fundamental genomic coordinate system*. These vertical alignments were also performed using the gos.vertical command.

7 Milestones

Figures 4, 5, and 6 display the milestones for the group.

Task	Estimate	Actual
Figuring out Gosling Syntax	4	20
Creating Gene Track	8	8
Creating Promoter and Enhancer Track	8	6
Creating 1 Chipseq track	2	2
Ability to stack many Chipseq Tracks	4	3
Streamlit Data loader	8	15
Streamlit Adding Biological Contrast functionality	4	5
Creating Merged View	10	10
Gosling add ability to vertically stack multiple biological cont	10	20
Writing Update	5	5
re-writing update	8	16
Figuring out how to prperly label contrasts	1	4
Forcing different elements to pan and zoom the same	2	1
Witing Intro	2	2
Creating powerpoint	4	6
Finding dataset to use and processing it	2	8
Getting screenshots for powerpoint and paper	2	2
Writing Final report	30	30
Total	114	163

Fig. 4. Alex's Milestones

Task	Estimate	Actual
Understanging ChIP-seq and ATAC-seq	2	6
Exploring ChIP-seq data	3	5
Search for Visualization Tools for ChIP-seq	2	2
Test Visualization tools selected	6	6
Writing introduction	8	16
Writing Related Work	8	12
Exploring Gosling	1	3
re-writing introduction	6	10
re-writing related work	6	4
review of concepts of data visualization implemented	3	4
Exploring the MultiChIPViewer	3	2
Exploring datasets to use	1	1
Writing discussion abd future work	10	12
Writing conclusion	2	1
Writing abstract	1	1
Creating powerpoint	4	8
recording the presentation	2	2
Writing and review of Final report	8	6
Total	76	101

Fig. 5. Rodrigo's Milestones

	Estimated	Actual
Understanding and exploring ChipSeq data	5	5
Gosling syntax exploration	10	15
Gosling code testing	4	4
Checking ChipSeq GitHub pulling and code debugging	25	30
Streamlit data rendering	8	10
Debugging Gosling codes for rendering visualization	9	9
Data Abstraction and Task abstraction writing	5	6
Pulling Github repo and checking code changes	4	4
Rending visualization to see whether the changes have been conveyed	8	8
Testing tool to see how the application can handle more than six datasets	4	4
Creating slides for presentation	8	6
Adding and revising reports for the final report	8	6
Total	98	107

Fig. 6. Yerin's Milestones

8 Results

To exemplify how the user is to use MultiChipVis, we will walk through possible research scenarios. In the scenario, the researcher is investigating how ChIP and ATAC-seq Signals change at the Hes1 gene throughout the embryonic development of mice brains.

8.1 Example Walk-through of Embryonic Development

In this scenario, there are three experimental conditions, E11.5, E15.5, and P0. Data sets in the E11.5 and E15.5 experimental conditions are from the brains of mouse embryos when they are 11.5 and 15.5 days old respectively. Data sets in the P0 experimental condition are from mouse brains immediately after the birth of the mouse. At each of these time points, the user performed ATAC-seq, as well as ChIP-seq for H3K4me3, and H3K27ac.

Once the user has uploaded their data, they first decide they want to navigate to the Hes1 Gene, their gene of interest. To jump to the Hes1 gene, they simply input "Hes1" into the Gene Search bar (Fig.7A). Once the visualization has jumped to the genomic coordinates of Hes1, the user can begin to investigate their data. Suppose the user is interested in if the Hes1 gene is being produced in the mice's brains. The researcher knows that if the Hes1 gene is being produced, then there should be an ATAC-seq peak that overlaps with the Hes1 Promoter. The user moves their mouse to the Promoter of the Hes1 gene, creating a vertical black line across the entire visualization (Fig.7B). Using this line, the user is able to rapidly discern that there are Red ATAC-seq Signals that are overlapping with the Yellow Promoter. As a result, the user concludes that Hes1 is produced in embryonic mouse brains throughout the development

The user might then be interested in identifying if the strength of the ATAC-seq peak that overlaps with the Hes1 Promoter changes throughout development. Suppose a colleague informs the user that less Hes1 is produced in the brain as the mouse embryo ages. With this knowledge, the user hypothesizes that the reduction in Hes1 is being driven by a decrease in the ATAC-seq Signal at the Hes1 Promoter. In order to compare the magnitude of the ATAC-seq Signal throughout development, the user might look towards the merged view, as the lack of large separations between the experimental conditions may allow for easier comparison between the experimental conditions. Because all of the ATAC-seq Signals are in Red, the user is able to easily locate the ATAC-seq Signal of interest (Fig.7C). Unfortunately, because each ATAC-seq Signal is normalized within its experimental condition, the user is not able to compare the relative magnitudes of the Signals. The user mouses over the Signal in each experimental condition so that they are able to access the raw values (Fig.7D). They determine that the ATAC-seq Signal at the Hes1 Promoter is indeed decreasing.

9 Discussion and Future Work

With the creation of genome-wide techniques such as ChIP-seq and ATAC-seq, several visualization toolkits for genomic data have been created. However, there are no tools that allow users to visualize multiple experimental conditions at one time. The MultiChIPViewer fills this gap as an easy-to-learn tool for ChIP and ATAC-seq data exploration. In order to make our tool entry-level and intuitive to use, the layout of the tool was made as simple as possible, allowing users to declare experimental conditions, load data sets into those experimental conditions, and jump to genes of interest.

In MultiChIPViewer the key heuristic is that Genes, Enhancers, Promoters, ChIP-seq, and ATAC-seq data are vertically aligned to the same genomic coordinate system. This heuristic enables a user to assess the overlapping of Signals to relevant features in the genome, and it enables users to assess the overlapping of Signals to each other. MultiChipViewer was created with the intention of viewing one gene at a time. This design decision was based on the fact that researchers are often interested in how a particular gene changes in different experimental decisions. However, this decision is also a limiting factor of MultiChipViewer. There are many cases in which users may want to compare how ChIP and ATAC-seq Signals at different genes change together. Indeed, a change in ChIP or ATAC-seq Signal at one gene may directly impact the ChIP or ATAC-seq Signal at another gene. Future work could be done to allow MultiChIPViewer to better support this function. For example, the option could be provided to create several *fundamental genomic coordinate systems* so that users could navigate to different genes for comparisons.

In MultiChIPViewer the decision was made to limit the maximum number of ChIP or ATAC-seq data types that a user can display. By limiting the number of different colors to six, the user is limited to six different data-set types. This decision was made in order to have a clear interpretation of the data sets that are loaded. We believe that adding more data-sets results in having too many *similar* hues in the display. This reduces data separation and prevents users from being able to clearly distinguish Signals of interest.

A fundamental downfall of MultiChipViewer is how the ChIP and ATAC-seq data are normalized. Due to the order in which the tool was developed, we lacked the foresight to normalize the ChIP-seq and ATAC-seq data of the *same type* in the same way. This prevented users from being able to compare Signal magnitudes between experimental conditions - which was one of the primary goals of the tool. In future iterations of the tool, this normalization should be fixed to enable cross-experimental-condition comparisons.

10 Conclusions

Gene regulation is a key component to understanding molecular biology. ChIP-seq and ATAC-seq have revolutionized how gene activity is studied because they enable researchers to collect data for the entire genome. ChIP and ATAC-seq allow researchers to find the locations in the genome where a protein of interest is interacting. Several visualization tools are available to aid in the detection of the ChIP and ATAC-seq Signals, and they have several shortcomings. The largest shortcoming is the inability to compare ChIP and ATAC-seq data across experimental conditions. MultiChIPViewer is a low-barrier visualization tool for ChIP and ATAC-seq data that allow the user to easily compare ChIP-seq and ATAC-seq data from multiple experimental conditions.

11 Code Availability

The code for MultiChipViewer can be found at https://github.com/AlexAdrian-Hamazaki/MultiChipViewer

Acknowledgments

The authors wish to thank Tamara Munzer for her valuable feedback during the development of this tool.

References

[1] G.-R. M. . D. L. Blanco, E. Productive visualization of high-throughput sequencing data using the seqcode open portable platform. *Scientific Reports*, 11(19545), 2021. doi: 10.1038/s41598-021-98889-7

[2] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1):21.29.1–21.29.9, 2015. doi: 10.1002/0471142727.mb2129s109

[3] D. P. Clark, N. J. Pazdernik, and M. R. McGehee. *Chapter 2 - Basic Genetics*. Academic Cell, third edition ed., 2019. doi: 10.1016/B978-0-12-813288-3.00002-1

[4] D. P. Clark, N. J. Pazdernik, and M. R. McGehee. *Chapter 4 - Genes, Genomes, and DNA*. Academic Cell, third edition ed., 2019. doi: 10.1016/B978-0-12-813288-3.00004-5

[5] E. P. Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2012. doi: 10.1038/nature11247

[6] M. G. Database. Mouse facts, June 2022.

[7] A. Frankish and e. a. Diekhans. GENCODE 2021. *Nucleic Acids Research*, 49(D1):D916–D923, 12 2020. doi: 10.1093/nar/gkaa1087

[8] W. J. e. a. Kent. The human genome browser at ucsc. *Genome research*, 12:996–1006, 2002. doi: doi:10.1101/gr.229102

[9] J. J. A.-S. S. e. a. Lerdrup, M. An interactive environment for agile analysis and visualization of chip-sequencing data. *Natural Structural Molecular Biology*, 23:349–357, 2016. doi: 10.1038/nsmb.3180

[10] S. L'Yi, Q. Wang, F. Lekschas, and N. Gehlenborg. Gosling: A grammar-based toolkit for scalable and interactive genomics data visualization. 2021. doi: 10.31219/osf.io/6evmb

[11] P. M. e. a. Moore, J.E. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583:699–710, 2020. doi: 10.1038/s41586-020-2493-4

[12] T. R. of the University of California. Ibigwig track format, October 2022.

[13] P. P.J. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10:669–680, 2009. doi: 10.1038/nrg2641

[14] Streamlit. https://streamlit.io.

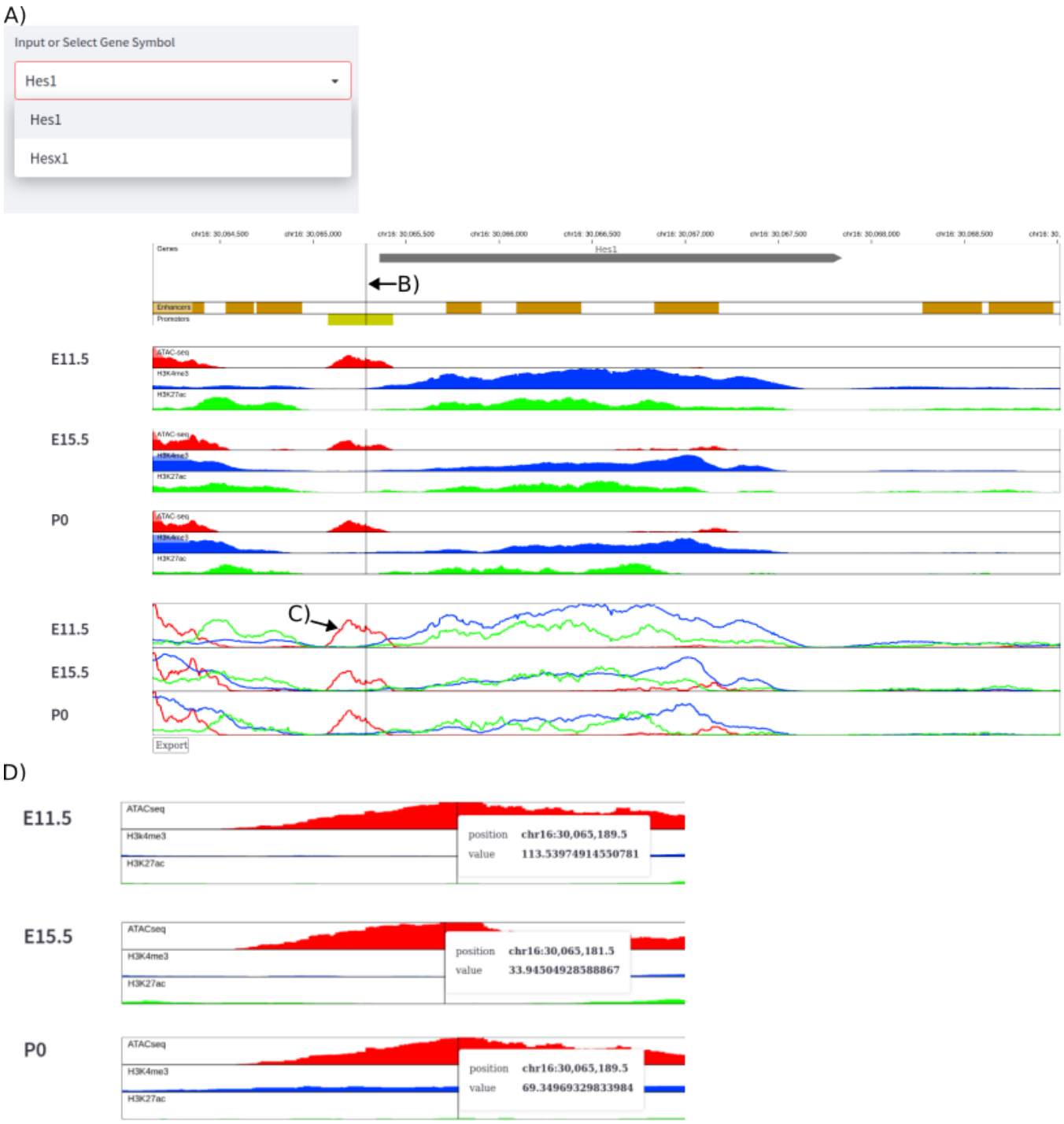

Fig. 7. An example scenario where the user is investigating change in ChIP-seq and ATAC-seq Signals at the Hes1 gene throughout the embryonic development of mouse brains. The user collected data from three developmental time points, E11.5, E15.5, and P0. At each time point, the user performed ATAC-seq and collected ChIP-seq data for H3K4me3 and H3K27ac. The user has successfully loaded their data into MultiChipVis **A**: To investigate the Hes1 gene, the user selects it from the toolbar. **B**: The user moves their mouse to the Hes1 Promoter, resulting in a vertical line throughout the visualization. The user uses this black line to determine that there is an ATAC-seq Signal at the Hes1 Promoter. **C**: The user attempts to use the *merged view* to determine if the ATAC-seq Signal at the Hes1 Promoter decreases throughput development. Because of the current normalization schema, the user is unable to view a decrease in value. **D**: The user mouses over the ATAC-seq Signals and a tool-tip appears giving the raw Signal magnitude. Using these tooltips, the user concludes that the ATAC-seq Signal at the Hes1 Promoter is decreasing.

[15] Streamlit-gos. https://github.com/gosling-lang/streamlit-gosling.