ChIP-Seq Data Visualization Made Simple

Alex-Adrian, Rodrigo S. Conceição, Yerin Kim

1. Introduction

The use of technologies to map of proteins to specific regions of DNA has been essential in unraveling gene activity and regulatory processes in multiple biological contexts. In order to efficiently store our massive amount of DNA, mammalian cells tightly wind the DNA around proteins called histones. These histones are able to be modified by the cell, and these modifications are vital in determining what genes become activated or repressed. Furthermore, genes that are tightly wrapped around histones tend to be *innacessable,* and are consequently repressed. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a technique that is able to measure histone modifications as well as other proteins bound to DNA, and it has been widely implemented in genome studies [1]. The Assay for Transposase-Accessible Chromatin (ATAC-seq) has been used to identify regions of the DNA that are *not* tightly wound by histones and are *accessible*. These technologies have revolutionized high throughput genomic studies. By identifying *in vivo* genome-wide binding sites of histones and proteins they have enhanced researchers' ability to investigate more complete genomic datasets [2]. They also allows us to analyze multiple samples simultaneously, elucidating cooperative interactions and unraveling important genomic regulatory relationships. Advances in these techniques have made it fairly accessible to investigators and thus, this technology has been widely implemented in the field of genomics [3].

Data generated and quality of ChIP-seq analysis depends heavily on the specificity and sensitivity of antibodies used. In general, specific antibodies will give more detailed information of binding sites, whilst non-specific antibodies generate less detailed information and greater level of background noise [4]. Common markers involved in gene regulation used in ChIP-seq analysis are H3 lysine 4 monomethylation (H3K4me1) associated to active enhancer regions, H3 lysine 4 trimethylation (H3K4me3) associated to

active promoter regions and H3 lysine 36 trimethylation (H3K36me3) associated to activly transcribed regions in gene bodies [3]. When a gene contains ATAC-seq signal that overlaps with regions containing these active ChIP-seq histone markers, the gene is often interpreted as active.

Visualizing Chip-seq signal of histones and other DNA-binding proteins in few gene sequences can be less complex and laborious. In a typical analysis, the interpretation is based the location of the signal (encoded as a peak) of histone markers used in the ChIP-seq [5]. Sharp peaks of H3K4me3 and H3K4me1 are usually expected at promoters and enhancers respectively. Whereas broader peaks are frequently observed with gene body histones (e.g., H3K36m3) [6]. Different tools are available to identify peaks in ChIP-seq data, a commonly used is model-based analysis of ChIP-seq (MACS) [7].

In general, ChIP-seq data analysis is not intuitive or user friendly. It typically involves the user investigating 1) The strength of a peak at several pre-defined genomic regions of interest, and 2) The degree to which the peak overlaps with other peaks. Current method of analysis frequently resort to vertically stacking multiple bar plots, and comparing readout between them by in essence, tracing your finger down the plots to compare the overlap of peaks (Fig 1). Although this is feasible in the case of three histone markers, we believe it has room for improvement. Furthermore, many experiments include performing ChIP-seq on categorically different samples, such as Heart vs Liver vs Lung. When this complexity is introduced, it becomes extremely difficult to compare differences in signal strength, and changes in peak overlap across categories. Our aim is to develop an interactive design to be user-friendly that ameliorate these weaknesses.

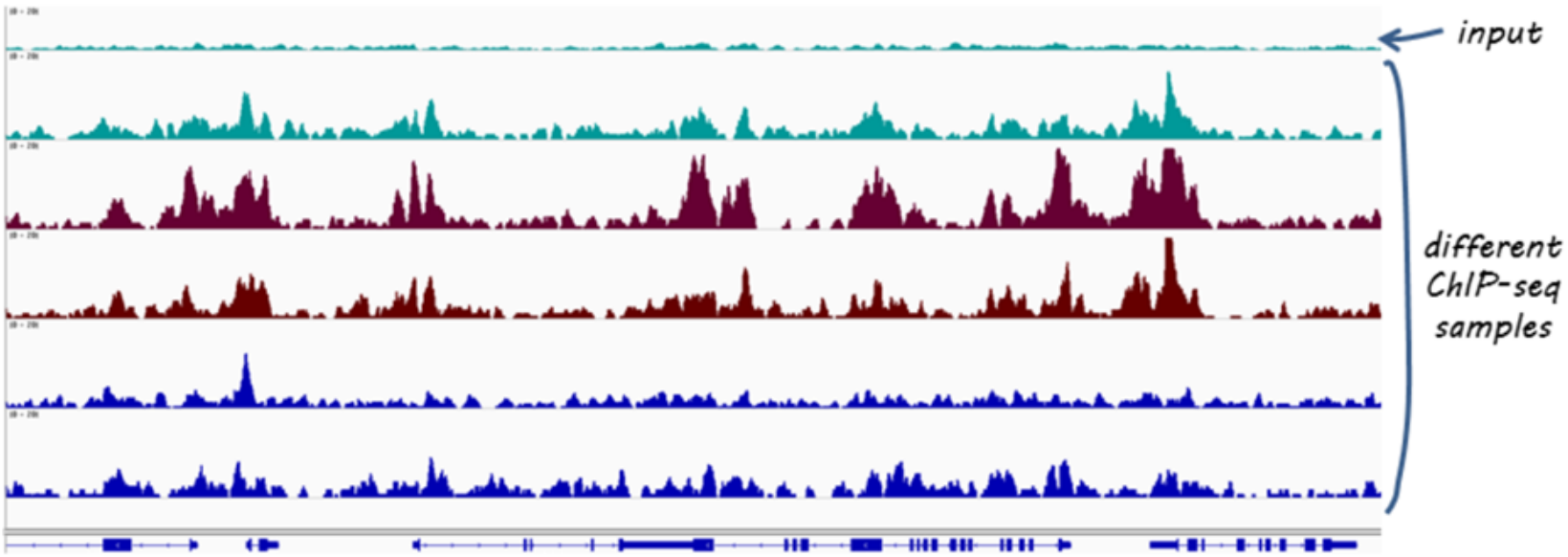


Figure 1: A Common ChIP-Seq Visualization (Image adapted from *HBC Training* https://hbctraining.github.io/Intro-to-ChIPseq/lessons/10_data_visualization.html)

2. **Related Work**

There are different tools available that are capable to quickly map short reads generated from ChIP-seq analysis, such as Bowtie, Eland, and BWA [5]. The fragments of DNA sequence can be grouped into contigs and then in assemblies to build the original DNA sequence [8]. The DNA sequence can be validated using the comparative method, which uses genome databanks to identify the sequences [5]. Another method commonly used for new gene sequences is *de novo,* which is based on finding overlaps between reads [5]. The decision of which tool to use as well as limitations often come to computational capacity, speed and time to

perform each task. Although these tools are available, often their capacity are limited to distribution analysis or to maps of smaller portions of the genome, resulting in investigators having to manipulate the data in multiple software [2]. Tools combining analysis techniques in one platform have been developed to address these limitations such as ABySS-Explorer, Easeq, Spark and Epiviz [7]. Moreover, along with technology advances it increases the possibilities and capacity to develop powerful genome-wide analytical tools [9]. The ABySS-Explorer is an example of design that uses related gene sequences assemblies to interactive display large datasets [8]. More recently developed, the EaSeq is a gene visualization tool that combines interactivity to user-friendly tools for genome-wide analysis and visualization. EaSeq can be more accessible as tasks can be performed by less powerful machines, and data is instantly transformed to interactive plots that can be modified and exported by the user [9]. Although these tools are being largely implemented in genome studies as they facilitate multi-scale analysis, experimentalists still need to go through peak activity to access quality of reads, peak region and annotation to interpretate biological functions in these gene modifications [10].

Powerful tools to local and detailed sequence read are available to display greater levels of data detail, however the data interpretation is not often intuitive. Thus, experimentalists would benefit from interactive genome-wide visualization designs incorporating enriched detail targeting important biological functions or highlighting possible biological significance in some gene modification. LYi et. al more recently developed gosling which is a grammar-based toolkit combining interactivity and scalability to multiple visualization types used in genomics data visualization [11].

3. **Data Abstraction**

We will visualize 3 types of ChIP-seq datasets, H3K4me3, H3K27ac, and H3K36me3, and ATAC-seq data. These datasets are from ENCODE, a consortium that creates, processes, and stores ChIP-seq and ATAC-seq data among other data types. H3K4me3 is an epigenetic modification to the DNA packaging protein Histone H3 that indicates tri-methylation at the 4th lysine residue of the histone H3 protein. When found at promoter regions, it is indicative that ced promoter is active. H3K27ac is an epigenetic modification to the

DNA packaging protein histone H3. It is a mark that indicates the acetylation of the lysine residue at N-terminal position 27 of the histone H3 protein. When found at enhancer regions, it is indicative that ced enhancer is active. H3K36me3 is an epigenetic modification to the DNA packaging protein Histone H3. It is a mark that indicates the tri-methylation at the 36th lysine residue of the histone H3 protein and is used to indicate an active gene body. ATAC-seq signal indicates that a given genomic region is not tightly wound by histones, and could be active.

In our visualization we will treat the various chromosomes in the genome as an ordered list. The data we have selected is from Mice. Mice have 40 chromosomes, but since 19 chromosomes are duplicates, only unique chromosomes are given in the datasets. As a result, the datasets contain information from 20 chromosome for females, or 21 chromosomes for males (as they have a Y chromosome). Each dataset is provided as a ".csv" and has 10 attributes. The three datasets are within a reasonably similar range; the H3K4me3 dataset has 34577 rows, the H3K27ac dataset has 91857 rows, and H3K36me3 dataset has 126992 rows. As these datasets can be very large and dense, they are often downloadable as a condensed file named a ".bigwig" file. These files contain three attributes. The ATAC-seq file we are currently working with is in this format, and will likely eventually use the ".bigwig" file formats for the ChIP-seq datasets for efficiency.

Table 1 - Data Abstraction of H3K4m3, H3K27ac, and H3K36me3

Attributes		**H3K4m3**	**H3K27ac**	**H3K36me3**
CHROM	Common	• Meaning: The name of the chromosome • Type: Categorical • Further Notes: Mice have 21 unique chromosomes (including Y). Some strange values will need to be cleaned		
	Diffrences	• Cardinality/Range: Cardinality = 27 (mice)	• Cardinality/Range: Cardinality = 25 (mice)	• Cardinality/Range: Cardinality = 21 (mice)

CHROMSTART	Common	• Meaning: The starting position of the CHIP signal in the chromosome • Type: Categorical		
	Diffrences	• Cardinality/Range: Cardinality = 34577 (number of items) • Further Notes: All Values are unique. The first base in a chromosome is numbered 0 • Further Notes:	• Cardinality/Range: Cardinality = 91830 (number of items) • Further Notes: Has some non-unique values, 27. This becomes values that can be duplicated for multiple Fields/chromosomes The first base in a chromosome is numbered 0	• Cardinality/Range: Cardinality = 126932 (number of items) • Further Notes: Has some non-unique values. This becomes values that can be duplicated for multiple Fields/chromosomes The first base in a chromosome is numbered 0
CHROMEND	Common	• Meaning: The ending position of the CHIP signal in the chromosome • Type: Categorical • Further Notes: For example, the first 100 bases of chromosome 1 are defined as chrom=1, chromStart=0, chromEnd=100, and span the bases numbered 0-99 in our software (not 0-100), but will represent the position notation chr1:1-100		
	Diffrences	• Cardinality/Range: Cardinality (number of items = 34577)	• Cardinality/Range: Cardinality (number of items = 91836) • Further Notes: Has some non-unique values, 21. This is becoming values that can be duplicated for multiple Fields/chromosomes.	• Cardinality/Range: Cardinality (number of items = 126934) • Further Notes: Has some non-unique values, 21. This is becoming values that can be duplicated for multiple Fields/chromosomes.
NAME	Common	• Meaning: Defines the name of the peak • Type: Categorical		
	Diffrences	• Cardinality/Range: Cardinality (number of items = 34577)	• Cardinality/Range: Cardinality (number of items = 91857)	• Cardinality/Range: Cardinality (number of items = 126992)
SCORE	Common	• Meaning: the score value will determine the level of gray in which this feature is displayed • Type: Ordered Quantitative Sequential		

		• Cardinality/Range: 0-1000		
	Diffrences	• Further Notes: I do not think this has a use for us	• Further Notes: It is not informative enough	• Further Notes: It is not informative enough
STRAND	Common	• Meaning: Defines the strand. Either "." (=no strand) or "+" or "-". • Type: Categorical • Cardinality/Range: 1 • Further Notes: ChIP-seq is all ".". So, this is redundant info		
signalValue	Common	• Meaning: Measurement of enrichment in field • Type: Ordered Quantitative Sequential • Further Notes: The main signal of the project		
	Diffrences	• Cardinality/Range: 1.42-76.35	• Cardinality/Range: 1.5-48.37	• Cardinality/Range: 1.52-17.84
pValue	Common	• Meaning: statistical significance (-log base 10) • Type: Ordered Quantitative Sequential • Further Notes: P value cut-off was 0.01 (value of 2 in this log transformed data)		
	Diffrences	• Cardinality/Range: 2.00 -256.77	• Cardinality/Range: 2-406	• Cardinality/Range: 2-159.69
qValue	Common	• Meaning: Significance using FDR (-logbase10) • Type: Ordered Quantitative Sequential		
	Diffrences	• Cardinality/Range: 0.33-253.86	• Cardinality/Range: 0.68-397.46	• Cardinality/Range: 0.81-150.25
peak	Common	• Meaning: bp distance from the start site to the peak summit • Type: Quantitative Sequential • Further Notes: 0 is based on chromstart		
	Diffrences	• Cardinality/Range: 11-9836	• Cardinality/Range: 0-13420	• Cardinality/Range: 0-30445

Table 2 - Data Abstraction of ATAC-seq

Attributes	Interpretation
Value	The value of a peak in a given range of genomic coardinates
Start	The start position of a peak.
End	The end position of a peak

4. Task Abstraction

4.1. Actions

Analyze:

Users should be able to discover what genes are active or not active. This will be done by viewing the decrease or increase of ChIP-seq, and ATAC-seq signals in genomic regions. This will also be done by viewing the degree of overlap between the signals at given genomic regions.

Produce:

Once users find interesting genomic regions, have the option to “flag” genes as “Active” or “NonActive”. Especially for users interested in a specific range of genome, can make publication-quality figures.

Search:

- **Lookup:** Allow users to be brought to a specific gene symbol or a genomic coordinate.
- **Browse:** Allow users to look at larger regions of the genome, not just specific genes. For example, we might allow a browsing window of 100,000 base pairs.
- **Query:** Allow users to compare the signals at/around the gene with other categorical samples. For example, comparing Heart vs Liver vs Lung. We can set a goal of 10 comparisons. In our project, these will be called “Contrasts”.

Targets

- Size of peaks
- Overlap of ATACseq peaks with other peaks
- Trends in Size and Overlap over different categorical datasets

5. Solution

Visual Encodings and Idioms

Our goal is to have three windows in which different information is displayed. For aid in understanding the encodings please view Figures 2 and 3.

5.1. Main Window

- **Genome/Field marks and channels:**

The Genome attribute will be encoded as an ordered list, it can be visualized by horizontal position channel with a horizontal line mark. Genes will be encoded in the horizontal channel as line marks corresponding to their position and length. Because DNA is double stranded, and each strand the different strands will be encoded by two hues. Enhancers and promoters will be encoded by line marks with length corresponding to their length, and they will be aligned horizontally to the genome. Enhancers and promoters will be encoded with hues.

- **ChIP-seq signal marks and channels:**

Chemical modifications (Methylation, Acetylation) will be encoded with peaks on horizontal position. Peaks will be encoded with frequency scaled to 1. Strong peaks will show solid color whereas weak peaks will show observable frequency with drop off. Different signals will be separately encoded with different hues. Different signal types will be vertically stacked. Each mark will have an identical y-axis length. Vertical positions are applicable for comparison only. The categorical distinction of different comparison conditions (time, tissue, etc.). The different signal types that are vertically stacked will be vertically stacked with respect to comparison samples. Sample IDs will be annotated. Horizontal gray and white striping to distinguish the different samples more easily.

5.2. Second Window

Upon interacting with a specific region via a sliding selection, the second window will contain a higher-resolution window. This window live updates depending on the region of the genome that is selected.

- **Genome/Field marks and channels:**

  Identical to Genome/Field marks and channels of main window

- **ChIP-seq signal marks and channels:**

  Peak signals will be smoothed out in a horizontal position where the signal amplitude is in the field. Line graph amplitude will

be scaled to 1. The different signal types will be categorized by different hues. All colors will have lower saturation to enable some opacity to view overlaps.

5.3. Third Window

Because scientists often require the viewing of the raw data, upon interacting with a genome users can open a third window can see the raw data for a specific categorical sample (ie. Liver) for a specific genomic location. The user can open many of these third windows for different specific categorical samples (ie. Windows for Liver, Heart, and Lung) to allow comparisons between different groups. This would be more akin to a "typical" ChIP-seq visualization.

- **Genome/Field marks and channels:**
  Identical to Genome/Field marks and channels of main window
- **ChIP-seq signal marks and channels:**
  Original data is encoded with peaks. The categorical distinction of different signal types will be encoded on vertical position. The different signal types will be vertically stacked and each mark will have an identical y-axis length. The categorical distinction of different comparison conditions (time, tissue, etc.). The different signal types that are vertically stacked will be vertically stacked with respect to comparison samples annotated with the sample ID. Horizontal Grey and white striping to distinguish the different samples more easily.

Frequency graphs of each signal (ATACseq, Methylation, Acetylation) will be encoded the same way as the original data.

6. Results

6.1. Sketches of ChIPVis

Fig. 2 - An Example of the ChIPVis with One Sample

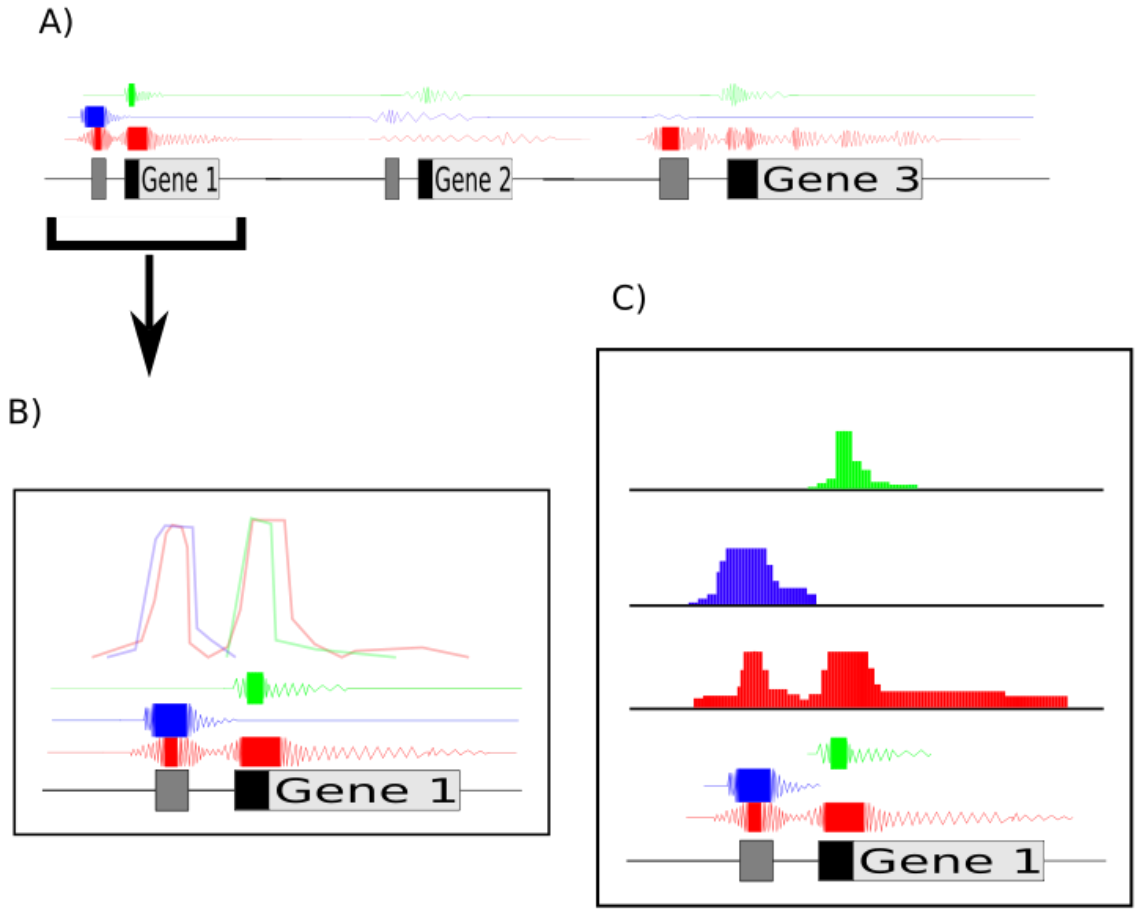


Fig. 2 - The three windows of the proposed visualization are shown. **A)** Window 1 of the proposed visualization. The genome is displayed as a field at the bottom of the vis. ChIP-seq signals are overlaid on top of the regions in the field that they correspond to.

Amplitude, and frequency of the waves redundantly encode the signal of a ChIP-seq signal. **B)** Window 2 of the proposed visualization. The window is opened by selecting a region of the genomic field in window 1. The wave forms are retained in window 2, and a line graph is added. The height of the line graph encodes ChIP-seq signal strength. **C)** Window 3 of the proposed visualization. Users have the ability to view raw data of a selected region.

Fig. 3 - An Example of the ChIPVis with Several Samples

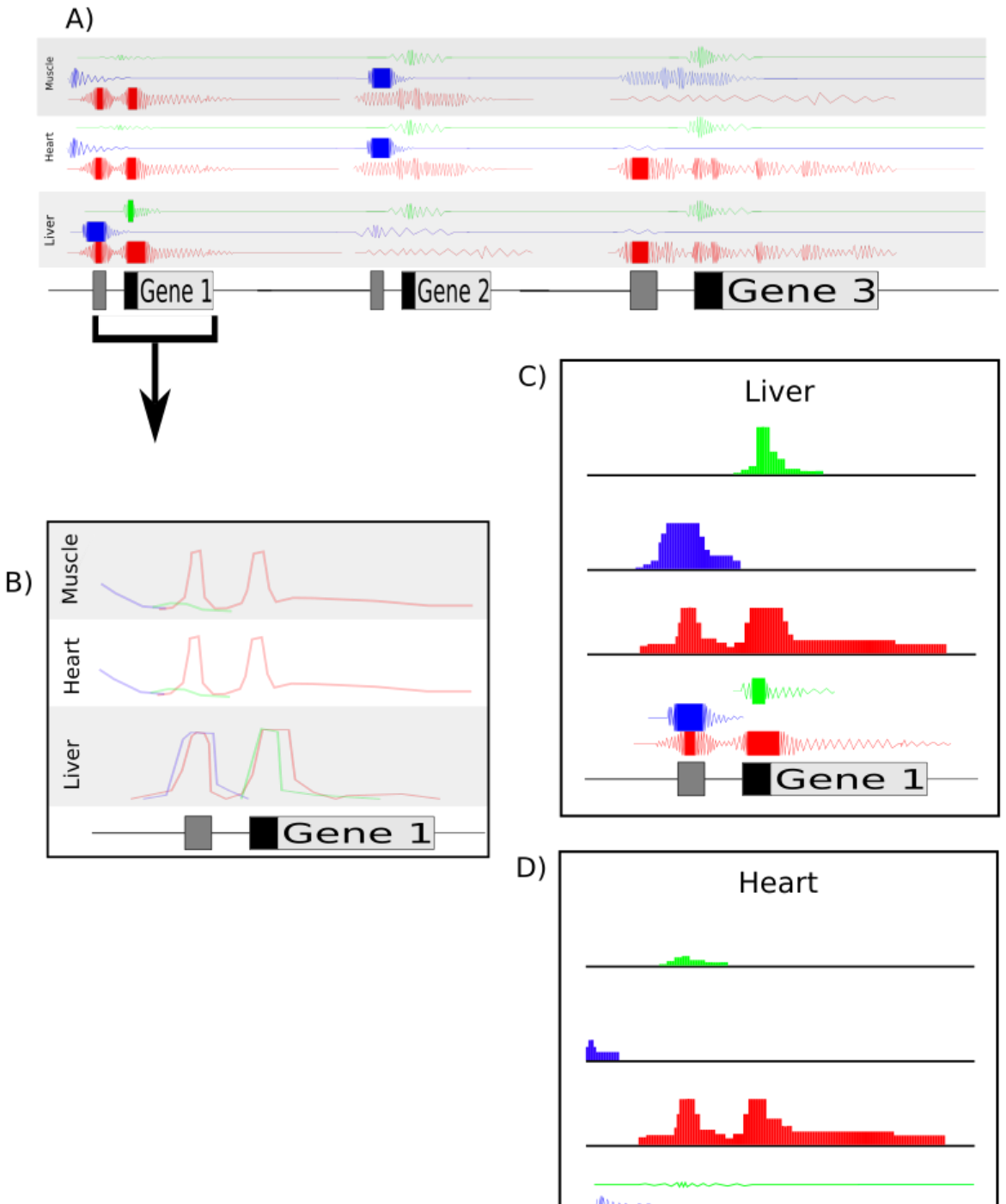

A)
Muscle
Heart
Liver
Gene 1
Gene 2
Gene 3
B)
Muscle
Heart
Liver
Gene 1
C)
Liver
Gene 1
D)
Heart

Fig. 3 - The three windows of the proposed visualization are shown. **A)** Window 1 of the proposed visualization. The genome is displayed as a field at the bottom of the vis. ChIP-seq signals are overlaid on top of the regions in the field that they correspond to. Amplitude, and frequency of the waves redundantly encode the signal of a ChIP-seq signal. Signals of various samples are labeled, and separated by grey-white striping. **B)** Window 2 of the proposed visualization. The window is opened by selecting a region of the genomic field in window 1. Line graphs are added, and different samples are separated by grey-white striping. **C)** Window 3 of the proposed visualization. Users have the ability to view raw data of a selected region. Many window 3's can be opened at once.

6.2. Simulated Results

Figures 2 and 3 represent an example case for ChIPVis. Figure 2 displays an example of ChIP-seq data from one dataset, such as a dataset from the Heart. Suppose, Green, Blue, and Red represent H3K4me3, H3K27ac, and H3K36me3 ChIP-seq signals respectively. In the proposed workflow, the user would be scanning through the genome, or lookup a particular gene of interest, for example "Gene 1" . In window 1, they would be able to rapidly identify several facts. 1) Gene 1's enhancer (indicated by light grey) has H3K27ac, and H3K36me3 histone marks. 2) Gene 1's Promoter has H3K4me3 and H3K36me3 histone marks 3) Gene 1's Gene body has H3K36me3 marks. 4) There is significant overlap between the H3K36me3 marks and the other marks. From these, the user would rapidly conclude that Gene 1 is likely being actively expressed. The user might then be interested in Gene 3. The user would rapidly identify that the gene body contains H3K36me3. However, there is a lack of H3K4me3 marks at the promoter, and H3K27ac marks at the enhancer. The user would likely conclude that the gene is not actively expressed.

If the user required more information on Gene 1, they would highlight this genomic region, and it would be opened in Window 2. The user could continue to inspect the presence of specific histone marks, and their overlap. However, if the user wanted to view the raw data, a method of selection would enable them to open Window 3, which contains a more typical ChIP-seq visualization containing the raw data.

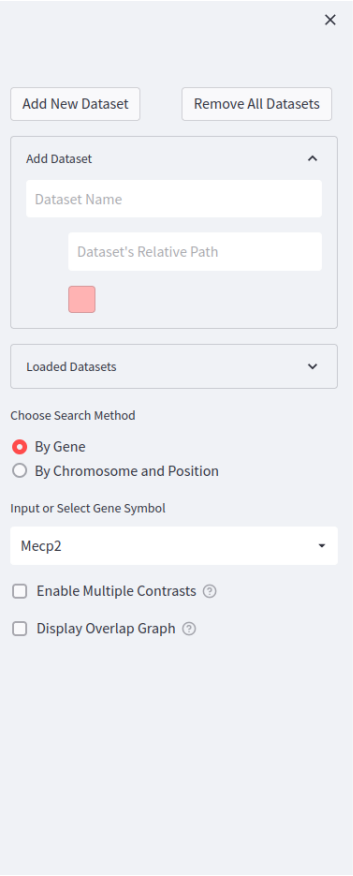
Figure 3 represents an example where various ChIP-seq datasets are being analyzed. In this case, data from Muscle, Heart, and Liver is being analyzed. To investigate Gene 1, the user would locate it in Window 1. They would hopefully discern several facts about the histone marks at Gene 1. 1) The H3K4me3, and H3K27ac marks are present at the promoter and enhancer respectively in Liver. However, these signals disappear in Heart and Muscle. 2) As the signals are lost in Heart and Muscle, there is no overlap between them and the H3K36me3 that is retained. This would indicate that Gene 1 is active in Liver, but not active in Heart and Muscle.

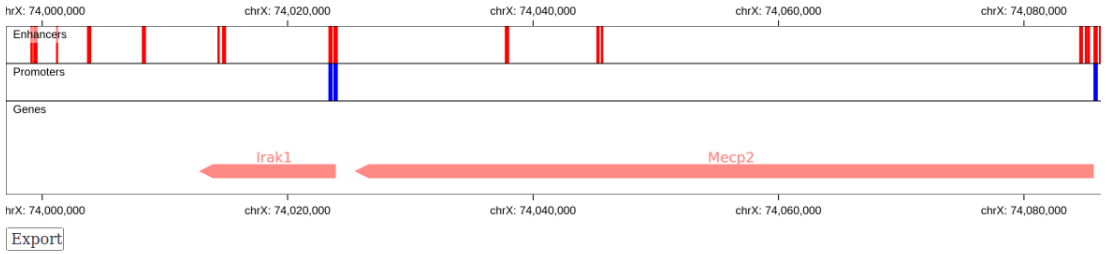
If the user required higher resolution on Gene 1, they would highlight this genomic region and view it in Window 2. The user would be able to rapidly view the lack of H3K4me3, and H3K27ac signal in Heart and Muscle. To view the raw data, the user could open a Window 3 for the Liver data, or the Heart data. These data would strengthen the conclusion that Gene 1 is not expressed in Heart and Muscle, but it is in Liver.

6.3. Screenshots of Current Application:

6.3.1. On Startup:

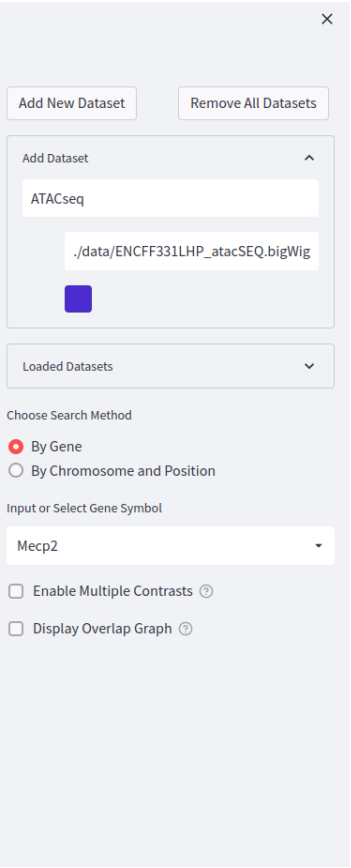
On startup, the genes, promoters, and enhancers are displayed. The program automatically starts at the “Mecp2” gene. On startup, no ChIP-seq or ATAC-seq datasets have been added. To add a dataset, the user must give a dataset a name, describe its path, and choose a color. Then the user must click “Add New Dataset”.

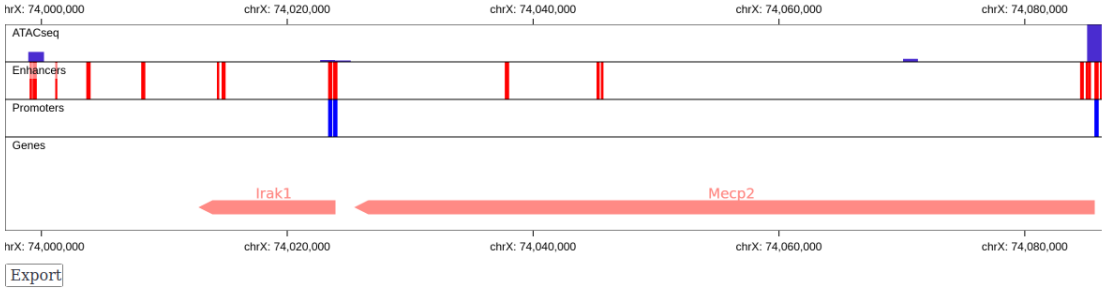




6.3.2. Adding a Datset:

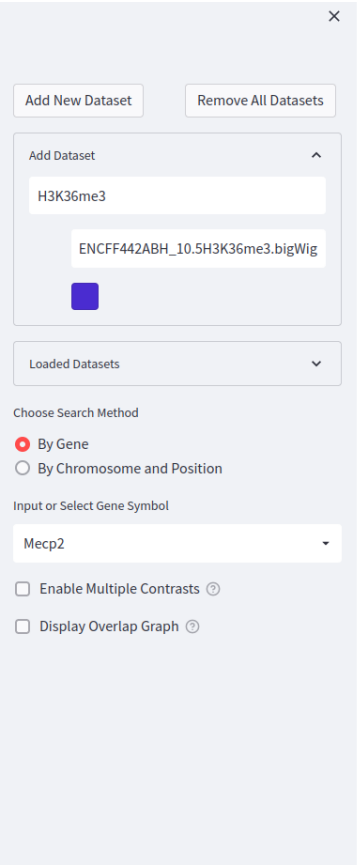
The user has added a dataset. The dataset is displayed above the Enhancers, Promoters, and Genes.

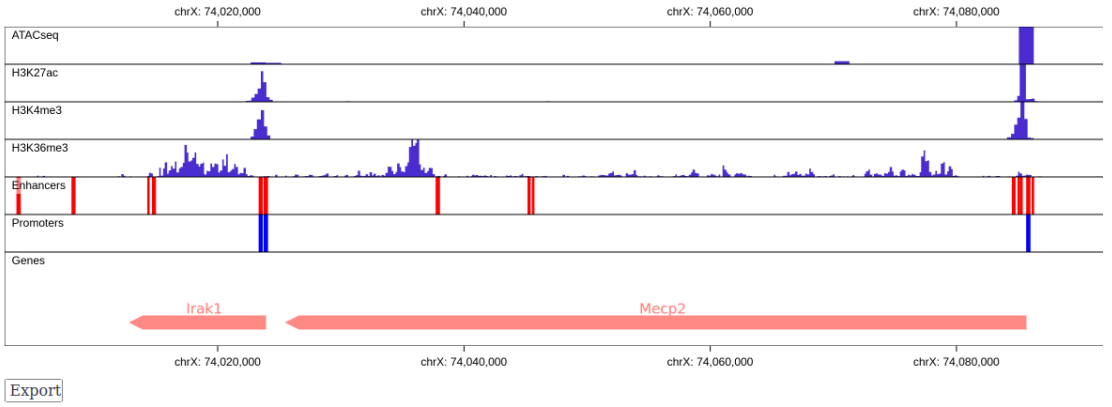




6.3.3. Adding Several Datsets

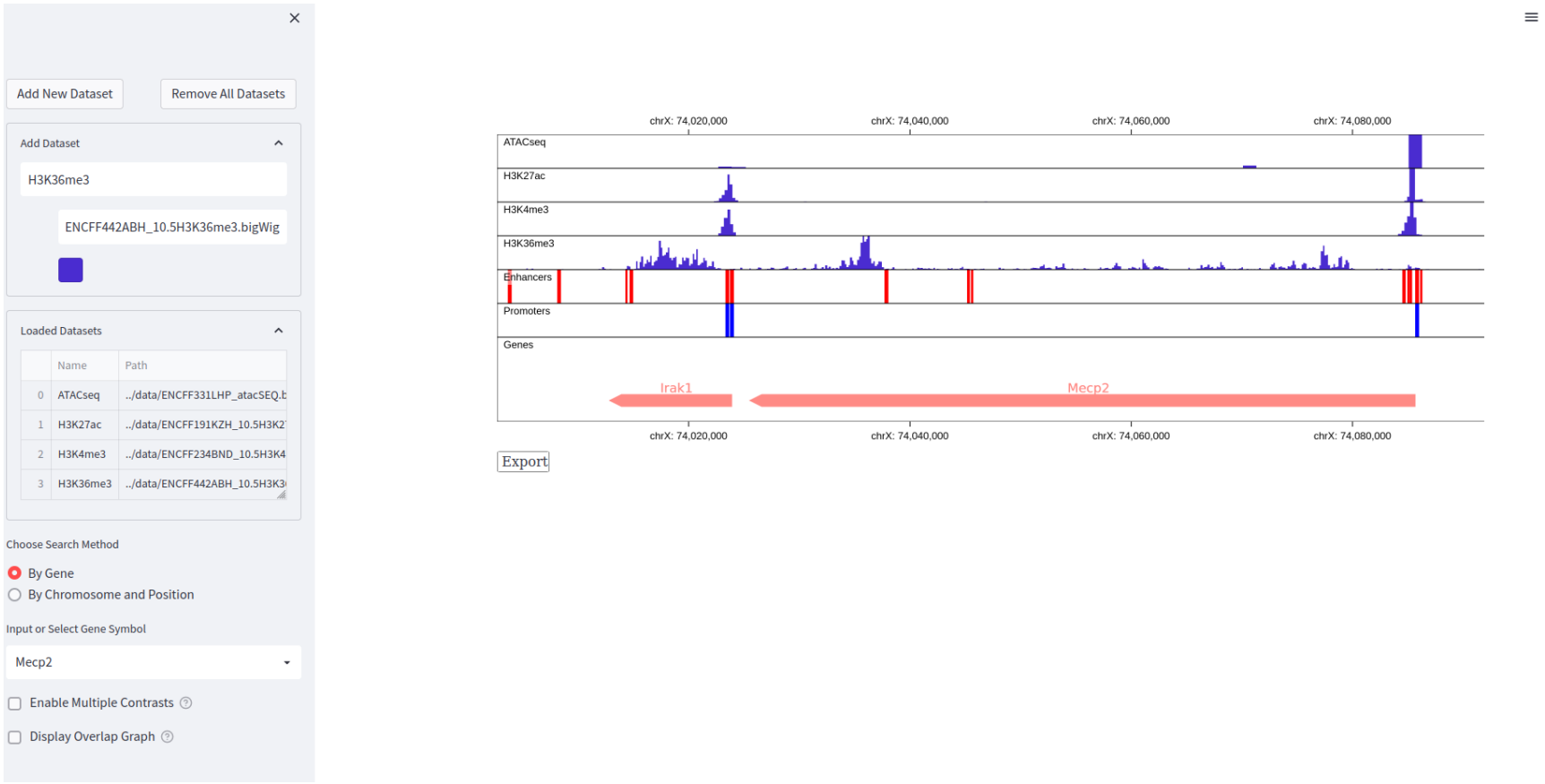
The user has added an ATACseq, H3K27ac, H3K4me3, and H3K36me3 dataset. For simplicity, I have made them all dark blue.





6.3.4. Keeping Track of Loaded Datasets

The user can expand the “Loaded Datasets” widget to view the datasets that are currently loaded.



6.3.5. Lookup of Genes

The user can search for a gene of interest. Selecting a new gene will jump the visualization to that gene of interest.

Add New Dataset | Remove All Datasets

Add Dataset

H3K36me3

ENCFF442ABH_10.5H3K36me3.bigWig

Loaded Datasets

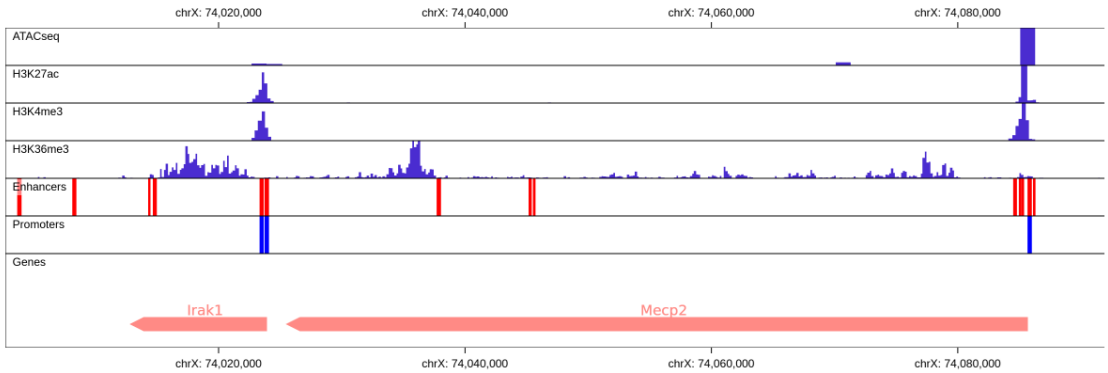
Name	Path

Map3k2
Map3k20
Map3k21
Map3k3
Map3k4
Map3k5
Map3k6
Map3k7

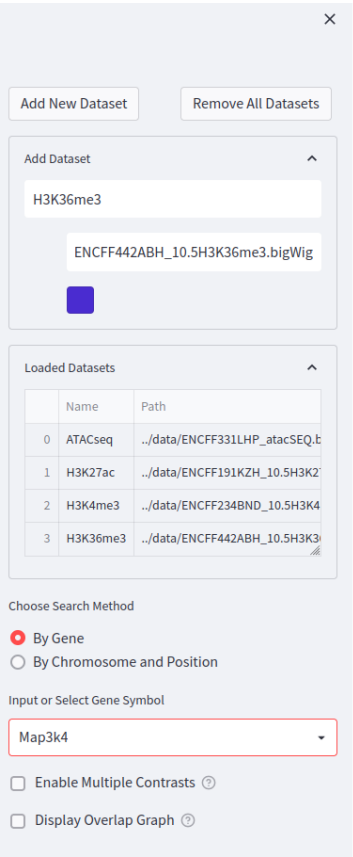
Mecp2

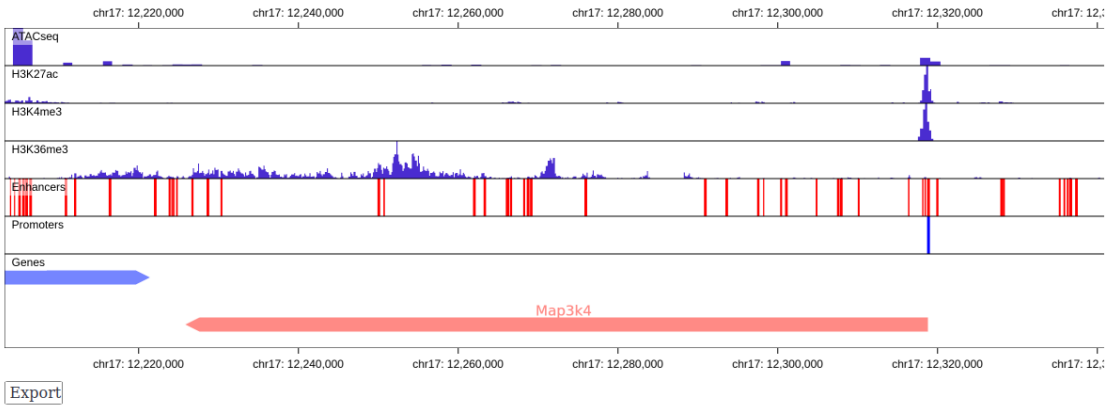
Enable Multiple Contrasts

Display Overlap Graph



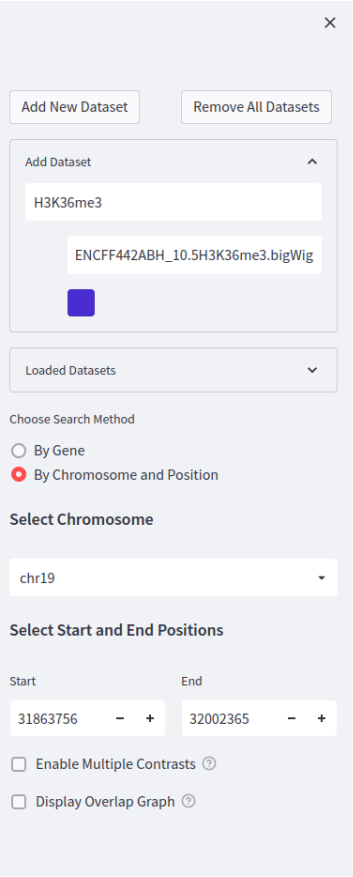
Export

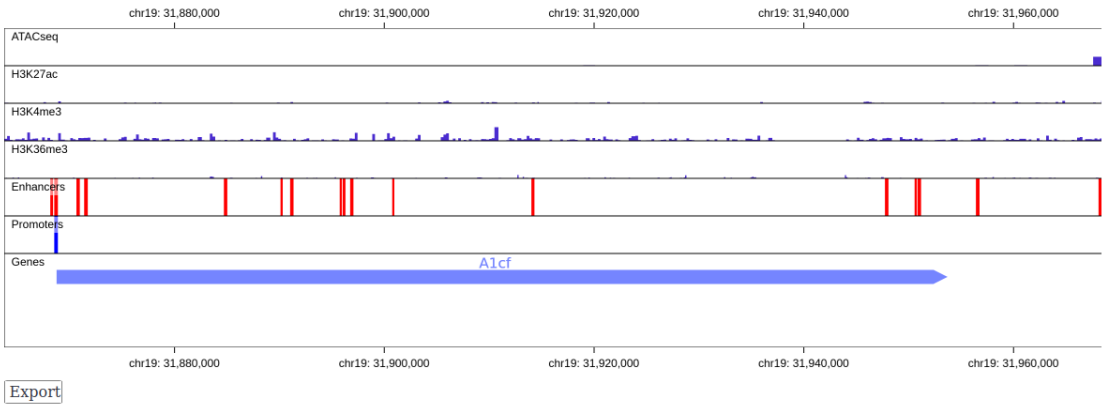




6.3.6. Scanning Coardinates

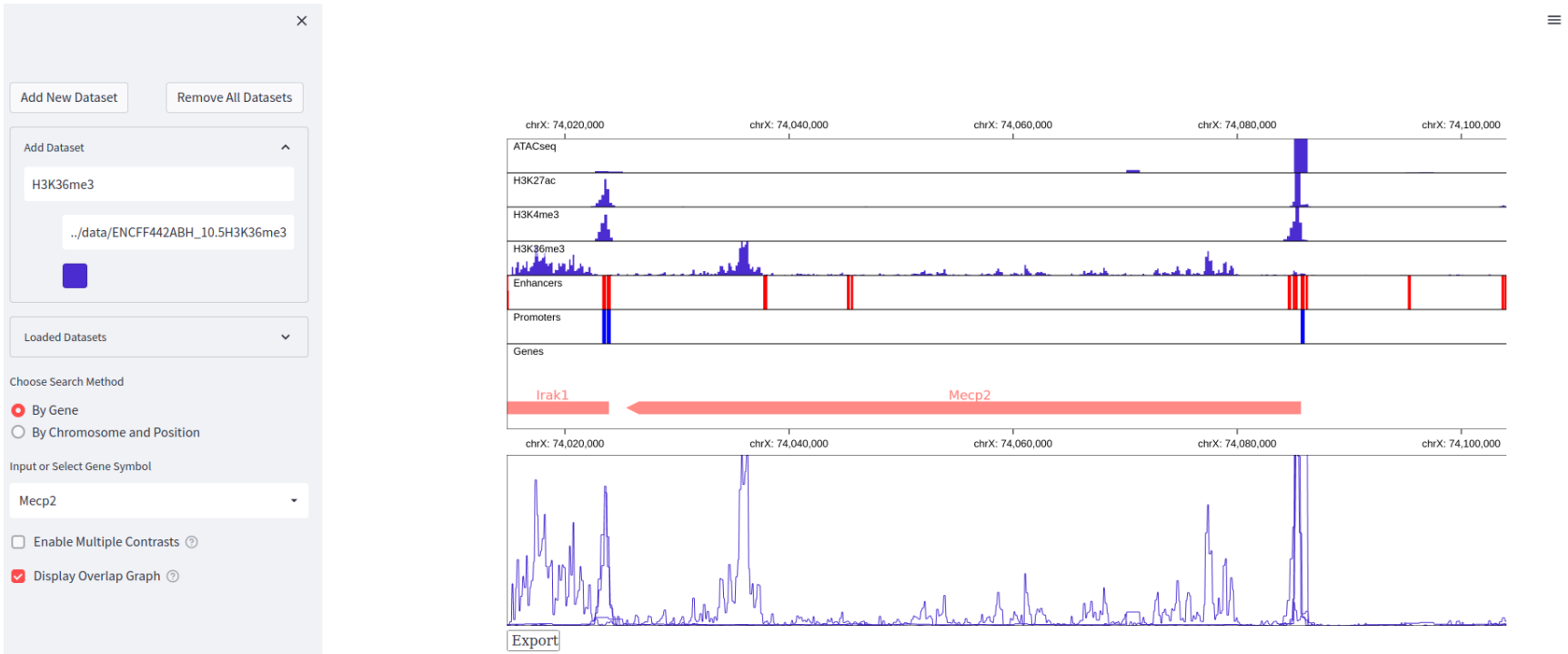
By clicking “By Chromosome and Position” under “Choose Search Method”, the user can scan regions of the genome by chromosome, and position.





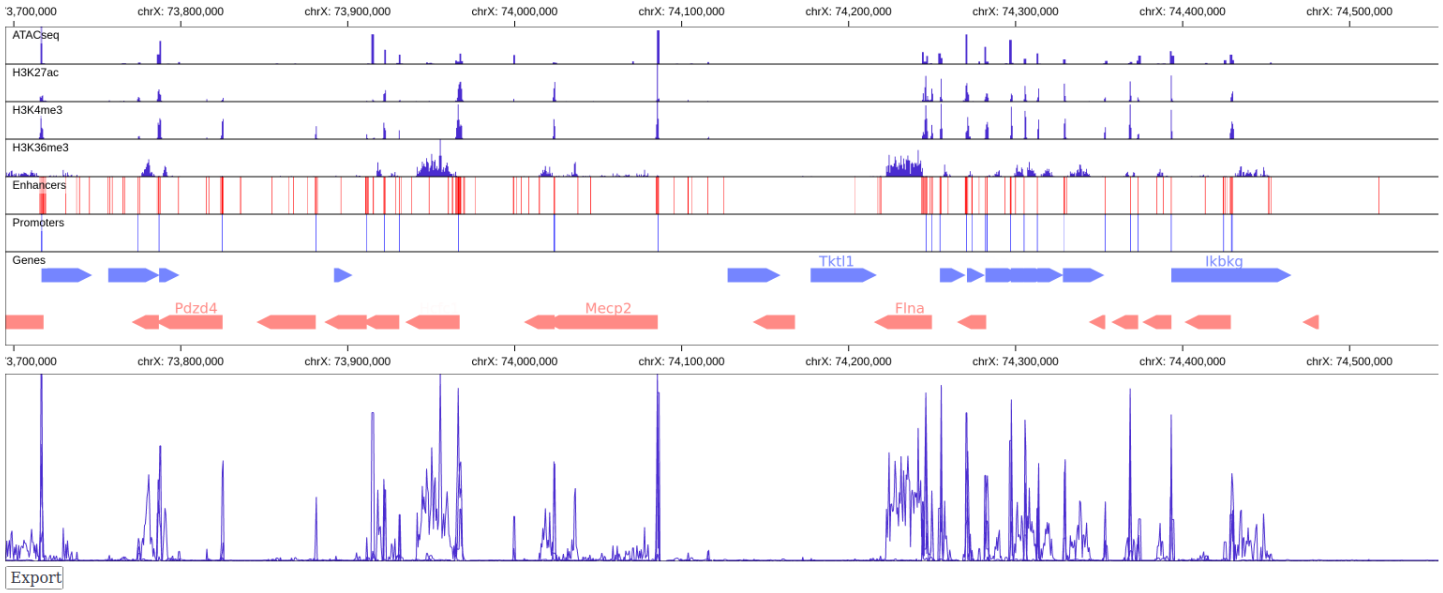
6.3.7. Displaying Overlap Graph

By clicking “Display Overlap Graph” the user creates a new track where the ChIP-seq and ATAC-seq signals are merged onto one track. Please note, because I selected the same color for all the datasets they are not differentiable. In the future, these colors will also be reduced in saturation to enable some opacity so one can more easily view overlaps.



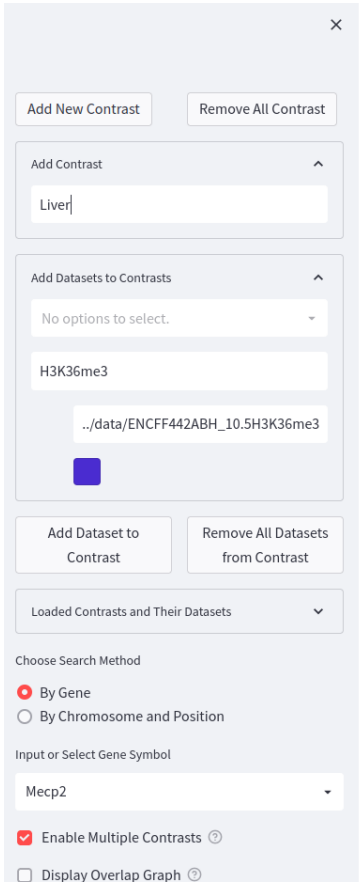
6.3.7. Removal of Toolbar, Zooming and Panning

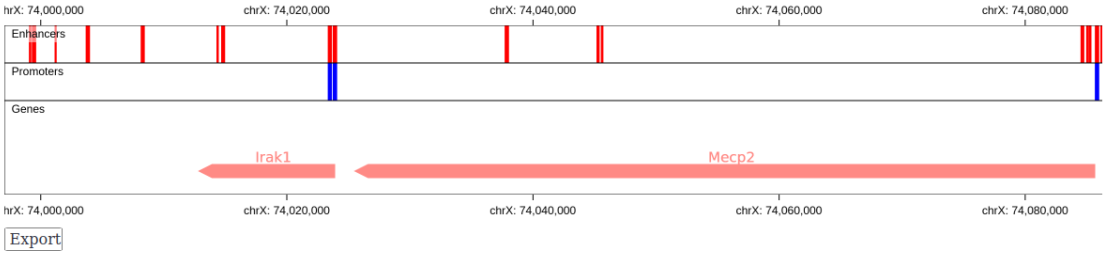
The sidebar can be removed. The visualization can be zoomed in or out on by scrolling the mouse wheel forwards or backwards respectively. The visualization can be panned left and right by dragging the visualization left or right respectively (not shown in static photo).



6.3.7. Enabling Multiple Contrasts

By clicking “Enable Multiple Contrasts” the user can visualize multiple contrasts simultaneously. Clicking this button resets the visualization and changes the options available in the toolbar.





6.3.8. Adding Multiple Contrasts, and Adding Datasets to Contrasts

The user can name multiple contrasts, and add datasets to selected contrasts. Every second contrast has a “grey” background to help distinguish them from eachother. The Contrasts and their datasets can be viewed by expanding the “Loaded Contrasts and their Datasets” widget. Future work requires implementing the “Overlap View” that is available when viewing one dataset (see 6.3.7). Future work also requires placing the Contrast Names as labels on the Y axis so that contrasts can be easily identified.

Add New Contrast | Remove All Contrast

Add Contrast

Heart

Add Datasets to Contrasts

Liver

Liver

Heart

../data/ENCFF442ABH_10.5H3K36me3

Add Dataset to Contrast | Remove All Datasets from Contrast

Loaded Contrasts and Their Datasets

Choose Search Method

By Gene

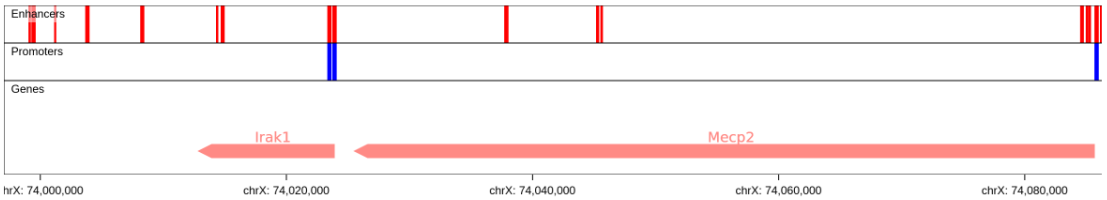
By Chromosome and Position

Input or Select Gene Symbol

Mecp2

Enable Multiple Contrasts

Display Overlap Graph



Export

Add New Contrast | Remove All Contrast

Add Contrast

Heart

Add Datasets to Contrasts

Heart

ATACseq

./data/ENCFF331LHP_atacSEQ.bigWig

Add Dataset to Contrast | Remove All Datasets from Contrast

Loaded Contrasts and Their Datasets

Choose Search Method

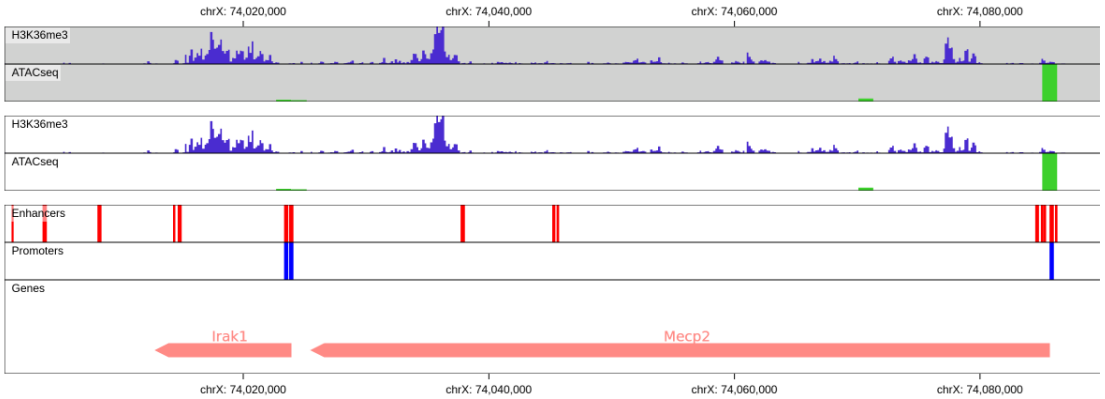
- By Gene
- By Chromosome and Position

Input or Select Gene Symbol

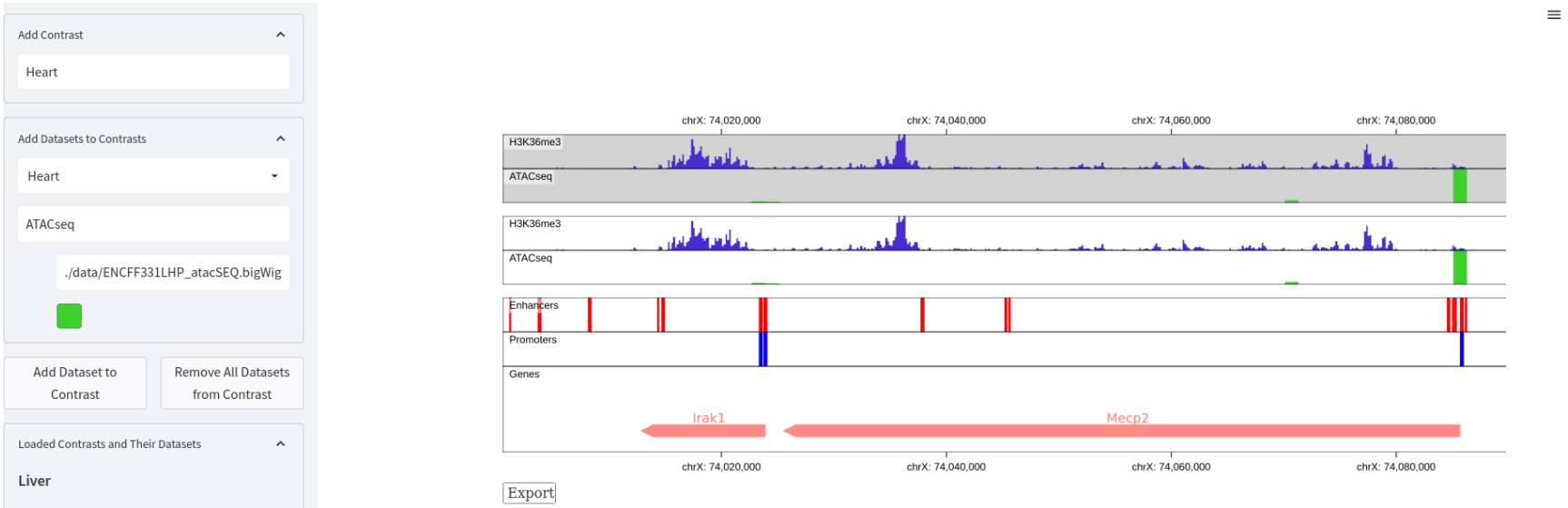
Mecp2

Enable Multiple Contrasts

Display Overlap Graph



Export



7. Implementation

To create a scalable and interactive genomic visualization we will be using the Python implementation of Gosling. Using Gosling, we are able to easily create a genomic coordinate system and align it with genomic data. Many common data genomic formats can be used directly in Gosling, such as BAM, BigWig, CSV, BED, GFF, and JSON [11]. To visuzalize our Gosling implementation, we will be using Streamlit (https://streamlit.io). Gosling has creating a component to easily view Gosling

visualizations in Streamlit (https://github.com/gosling-lang/streamlit-gosling). Using Streamlit, we are able to build an interactable web application to visualize our Gosling implementation.

8. **Milestones**

Project milestones are described in table 4. Weekly meetings will be carried by the group members to discuss the project development.

Table 4 - Project development chronogram and task distribution

Project Members	Project task	Timeline	Expected Work Hours	Actual Work Hours	Status	Deadline
Alex	Data abstraction	Oct 10th - Oct 21st			Completed	Oct 21st
	Defining possible solutions	Oct 10th - Oct 21st			Completed	Oct 21st
	Visualization encoding - Promoter,Enhancer,Gene Track ✅ - Ability to add Chipseq and ATACseq datasets✅ - Ability to create “Merged View” (half done) - Ability to view multiple “Contrasts”✅ - Make the page pretty - Add Y axis labels for different contrasts - Add ability to flag genes as active	Oct 21st - Nov30	120	70	In Progress	Nov 30
	Results				In Progress	

Rodrigo	Introduction	Oct 10th - Oct 21st	8	16	Completed	Oct 21st
	Related work	Oct 10th - Nov 15th	12	14	In Progress	Dec 16th
	Discussion	Dec 1st - Dec 10th	12	-	Not Started	Dec 16th
	Conclusion	Dec 10th	2	-	Not Started	Dec 16th
Yerin	Task abstraction	Oct 10th - Oct 21st			Completed	Oct 21st
	Implementation	Nov 15th			Completed	Oct 21st
	Visualization encoding	Oct 21st -			In Progress	
	Results				In Progress	
Collaborative	Follow up meetings to brainstorm ideas, assess project development, and overcome challenges	Throughout the project development	6-8	3	In Progress	-

9. **Discussion**
10. **Future work**
11. **Conclusion**

12. **Bibliography**

[1] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics,* vol. 10, pp. 669-680, 2009.

[2] R. Nakato and K. Shirahige, "Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation," *Briefings in Bioinformatics,* vol. 18, pp. 279-290, 2017.

[3] R. Nakato and T. Sakata, "Methods for ChIP-seq analysis: A practical workflow and advanced applications," *Methods,* vol. 187, pp. 44-53, 2021.

[4] S. Pepke, B. Wold and A. Mortazavi, "Computation for ChIP-seq and RNA-seq studies," *Nature Methods,* vol. 6, pp. 22-32, 2009.

[5] S. Bao, R. Jiang, W. Kwan, B. Wang, X. Ma and Y.-Q. Song, "Evaluation of next-generation sequencing software in mapping and assembly," *Journal of Human Genetics,* vol. 56, pp. 406-414, 2011.

[6] S. Hino, T. Sato and M. Nakao, "Chromatin Immunoprecipitation Sequencing (ChIP-seq) for Detecting Histone Modifications and Modifiers," *Epigenomics. Methods in Molecular Biology,* vol. 2577, pp. 55-64.

[7] H. Shin, T. Liu, X. Duan, Y. Zhang and X. S. Liu, "Computational methodology for ChIP-seq analysis," *Quantitative Biology,* vol. 1, pp. 54-70, 2013.

[8] C. B. Nielsen, S. D. Jackman, I. Birol and S. J. Jones, "ABySS-Explorer: Visualizing Genome Sequence Assemblies," *IEEE Transactions on Visualization and Computer Graphics,* vol. 15, pp. 881-888, 2009.

[9] M. Lerdrup, J. V. Johansen, S. Agrawal-Singh and K. Hansen, "An interactive environment for agile analysis and visualization of ChIP-sequencing data," *Nature Structural & Molecular Biology,* vol. 23, pp. 349-357, 2016.

[10] R. Mundade, H. G. Ozer, H. Wei, L. Prabhu and T. Lu, "Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond," *Cell Cycle,* vol. 13, pp. 2847-2852, 2014.

[11] S. L'Yi, Q. Wang, F. Lekschas and N. Gehlenborg, "Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization," *IEEE Transactions on Visualization and Computer Graphics,* vol. 28, pp. 140-150, 2022.

https://streamlit.io/