ChIPVis: A Simple ChIP-Seq Visualization Tool

(Name subject to improvement)

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 unravelling 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. 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]. This technology has revolutionised high throughput genomic studies as it can map *in vivo* genome-wide binding sites of histones and proteins. It has enhanced researchers' capability to investigate more complete genomic datasets [2]. ChIP-seq also allows us to analyse multiple samples simultaneously, elucidating cooperative interactions and unravelling important biological regulations in the genome. The advances in this technique made it more accessible to investigators and thus, making this technology widely implemented in whole genomics field [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 enhancer regions, H3 lysine 4 trimethylation (H3K4me3) associated to promoter regions and H3 lysine 36 trimethylation (H3K36me3) associated to transcribed regions in gene bodies [3]. Pieces of genome sequences are read individually through ChIP-seq technique, thus, to investigate complete gene it is necessary to map each sequence.

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.

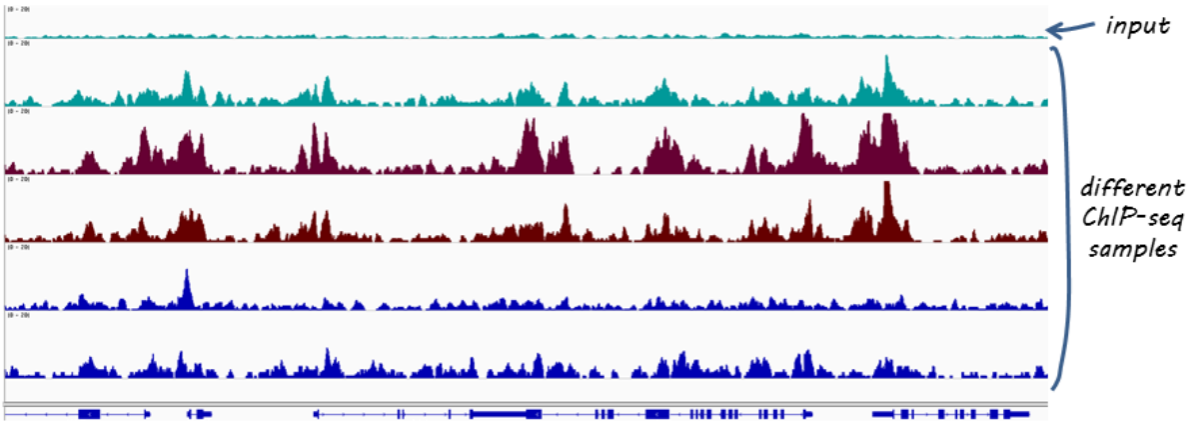

Figure 1: A Common ChIP-Seq Visualization (Image from [11])

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, experimentalist 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 level of data detail, however the data interpretation is not often intuitive. Thus, experimentalist would benefit from interactive genome-wide visualization designs incorporating enriched detail targeting important biological functions or highlighting possible biological significance in some gene modification.

3. **Data Abstraction**

We will visualize 3 types of ChIP-seq datasets, H3K4me3, H3K27ac, and H3K36me3. These datasets are from ENCODE, a consortia that creates, processes, and stores ChIP-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.

In our visualization we will treat the various chromosomes in the genome as a Field. 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 contains information from 20 or 21 chromosomes in the case of males with Y chromosomes. Each dataset 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.

Table 1 - Data Abstraction of H3K4m3 Data

Attributes	**Column name**	**H3K4m3**
Attribute 1	CHROM	• Meaning: The name of the chromosome • Type: Categorical • Cardinality/Range: Cardinality = 27 (mice) • Further Notes: Mice have 21 unique chromosomes (including Y). Some strange values will need to be cleaned
Attribute 2	CHROMSTART	• Meaning: The starting position of the CHIP signal in the chromosome • Type: Categorical • Cardinality/Range: Cardinality = number of items (34577) • Further Notes: All Values are unique. The first base in a chromosome is numbered 0

Attribute 3	CHROMEND	• Meaning: The ending position of the CHIP signal in the chromosome • Type: Categorical • Cardinality/Range: Cardinality = number of items (34577) • 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
Attribute 4	NAME	• Meaning: Defines the name of the peak • Type: Categorical • Cardinality/Range: Cardinality = number of items (34577)
Attribute 5	SCORE	• Meaning: the score value will determine the level of gray in which this feature is displayed • Type: Ordered Quantitative Sequential • Cardinality/Range: 0-1000 • Further Notes: I do not think this has a use for us
Attribute 6	STRAND	• 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
Attribute 7	signalValue	• Meaning: Measurement of enrichment in field • Type: Ordered Quantitative Sequential • Cardinality/Range: 1.42-76.35 • Further Notes: This is our main signal I think
Attribute 8	pValue	• Meaning: statistical significance (-log base 10) • Type: Ordered Quantitative Sequential • Cardinality/Range: 2.00 -256.77 • Further Notes: P value cut-off was 0.01 (value of 2 in this log transformed data)
Attribute 9	qValue	• Meaning: Significance using FDR (-logbase10) • Type: Ordered Quantitative Sequential • Cardinality/Range: 0.33-253.86
Attribute 10	peak	• Meaning: bp distance from the start site to the peak summit • Type: Quantitative Sequential • Cardinality/Range: 11-9836 • Further Notes: 0 is based on chromstart

Table 2 - Data Abstraction of H3K27ac Data

Attributes	**Column name**	**H3K27ac**

Attribute 1	CHROM	• Meaning: The name of the chromosome • Type: Categorical • Cardinality/Range: Cardinality = 25 (mice) • Further Notes: Mice have 21 unique chromosomes (including Y). Some strange values will need to be cleaned
Attribute 2	CHROMSTART	• Meaning: The starting position of the CHIP signal in the chromosome • Type: Categorical • Cardinality/Range: Cardinality = 91830 • 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
Attribute 3	CHROMEND	• Meaning: The ending position of the CHIP signal in the chromosome • Type: Categorical • Cardinality/Range: Cardinality = 91836 • Further Notes: Has some non-unique values, 21. This is becoming values that can be duplicated for multiple Fields/chromosomes. 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
Attribute 4	NAME	• Meaning: Defines the name of the peak • Type: Categorical • Cardinality/Range: Cardinality = number of items (91857)
Attribute 5	SCORE	• Meaning: the score value will determine the level of gray in which this feature is displayed • Type: Ordered Quantitative Sequential • Cardinality/Range: 0-1000 • Further Notes: It is not informative enough
Attribute 6	STRAND	• 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
Attribute 7	signalValue	• Meaning: Measurement of enrichment in field • Type: Ordered Quantitative Sequential • Cardinality/Range: 1.5-48.37 • Further Notes: This is our main signal I think
Attribute 8	pValue	• Meaning: statistical significance (-log base 10) • Type: Ordered Quantitative Sequential • Cardinality/Range: 2-406

		• Further Notes: P value cut-off was 0.01 (value of 2 in this log transformed data)
Attribute 9	qValue	• Meaning: Significance using FDR (-logbase10) • Type: Ordered Quantitative Sequential • Cardinality/Range: 0.68-397.46
Attribute 10	peak	• Meaning: bp distance from the start site to the peak summit • Type: Quantitative Sequential • Cardinality/Range: 0-13420 • Further Notes: 0 is based on chromstart

Table 3 - Data Abstraction of H3K36me3 Data

Attributes	**Column name**	**H3K36me3**
Attribute 1	CHROM	• Meaning: The name of the chromosome • Type: Categorical • Cardinality/Range: Cardinality = 21 • Further Notes: Mice have 21 unique chromosomes (including Y)
Attribute 2	CHROMSTART	• Meaning: The starting position of the CHIP signal in the chromosome • Type: Categorical • Cardinality/Range: Cardinality = 126932 • 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
Attribute 3	CHROMEND	• Meaning: The ending position of the CHIP signal in the chromosome • Type: CategoricalCardinality/Range: Cardinality = 126934 • Further Notes: Has some non-unique values, 21. This is becoming values that can be duplicated for multiple Fields/chromosomes. 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
Attribute 4	NAME	• Meaning: Defines the name of the peak • Type: Categorical • Cardinality/Range: Cardinality = number of items 126992
Attribute 5	SCORE	• Meaning: the score value will determine the level of gray in which this feature is displayed • Type: Ordered Quantitative Sequential

		• Cardinality/Range: 0-1000 • Further Notes: It is not informative enough
Attribute 6	STRAND	• 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
Attribute 7	signalValue	• Meaning: Measurement of enrichment in field • Type: Ordered Quantitative Sequential • Cardinality/Range: 1.52-17.84 • Further Notes: Presumably, the main signal
Attribute 8	pValue	• Meaning: statistical significance (-log base 10) • Type: Ordered Quantitative Sequential • Cardinality/Range: 2-159.69 • Further Notes: P-value cut-off was 0.01 (value of 2 in this log-transformed data)
Attribute 9	qValue	• Meaning: Significance using FDR (-logbase10) • Type: Ordered Quantitative Sequential • Cardinality/Range: 0.81-150.25
Attribute 10	peak	• Meaning: bp distance from the start site to the peak summit • Type: Quantitative Sequential • Cardinality/Range: 0-30445 • Further Notes: 0 is based on chromstart

4. **Task Abstraction**

Actions

- **Analyze:**
  - **Individuals should be able to discover what genes are active/not active**
    Viewing changes (increase/decrease) in:
    - Methylation of promoter
    - Acetylation of Enhancer
    - ATACseq At enhancer/Promoter regions

    Viewing overlap in:
    - ATACseq Signal and Acetylation signal
    - ATACseq signal and promoter signal
- **Produce:**
  - **Annotate:** Provide users the option to "flag" genes as "Active" or "NonActive"
  - **Record:** Provide users the option to make publication-quality figures. Especially for a user interested in a specific range of the genome
- **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 other comparisons. However, one type of categorical comparison comparison will be allowed at one time.

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.

Main Window:

Genome/Field marks and channels

- Line Mark: The Genome
  - Channel: Horizontal Position (How the "Field" system is represented)
- Line Mark: Enhancers
  - Channel: Grey, Dark Saturation
  - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the enhancer in the field)
  - Channel: Length (How much of the field is taken up by the Enhancer)
- Line Mark: Promoters
  - Channel: Black
  - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the Promoter in the field)
  - Channel: Length (How much of the field is taken up by the Promoter)
- Line Mark: Genes
  - Channel: Grey, Light Saturation
  - Channel: Annotation channel (Gene name is annotated within the Line Mark itself)
  - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the gene in the field)
  - Channel: Length (How much of the field is taken up by the Gene)

ChIP-seq signal marks and channels

- Line Mark: Peak (Methylation, Acetylation)
    - Channel: Horizontal Position (Where the peak is in the field)
    - Channel: Frequency (How strong the peak is for a given region of the field)
        - Strong peak = Solid color
        - Weak peak = Observable frequency with drop off
        - Scaled to 1
    - Channel: Amplitude (How strong the peak is for a given region of the field)
        - Scaled to 1
    - Channel: Hue (The categorical distinction of different signal types)
    - Channel: Vertical Position (The categorical distinction of different signal types)
        - The different signal types will be vertically stacked
        - Each mark will have an identical Y length.
    - Channel: Vertical Position
        - 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
        - Annotated with sample ID
    - Channel: Saturation
        - Horizontal gray and white striping to distinguish the different samples more easily

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 to depending on the region of the genome that is selected.

Genome/Field marks and channels

- Line Mark: The Genome
    - Channel: Horizontal Position (How the “Field” system is represented)
- Line Mark: Enhancers
    - Channel: Grey, Dark Saturation
    - Channel: Horizontal Position (Overlayed onto the Genome Line mark to represent the position of the enhancer in the field)
    - Channel: Length (How much of the field is taken up by the Enhancer)
- Line Mark: Promoters
    - Channel: Black
    - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the Promoter in the field)
    - Channel: Length (How much of the field is taken up by the Promoter)
- Line Mark: Genes

    - Channel: Grey, Light Saturation
    - Channel: Annotation channel (Gene name is annotated within the Line Mark itself)
    - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the gene in the field)
    - Channel: Length (How much of the field is taken up by the Gene)

ChIP-seq signal marks and channels: Line graph

- Line Mark: Peak (Data with Smooth Function)
    - Channel: Horizontal Position (Where the signal amplitude is in the field)
    - Channel: line graph amplitude (The amount of each signal scaled to 1)
    - Channel: Hue (The categorical distinction of different signal types)
    - Channel: Saturation (All colors will have lower saturation to enable some opacity to view overlaps

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 cagetorical 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:

- Line Mark: The Genome
    - Channel: Horizontal Position (How the "Field" system is represented)
- Line Mark: Enhancers
    - Channel: Grey, Dark Saturation
    - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the enhancer in the field)
    - Channel: Length (How much of the field is taken up by the Enhancer)
- Line Mark: Promoters
    - Channel: Black
    - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the Promoter in the field)
    - Channel: Length (How much of the field is taken up by the Promoter)
- Line Mark: Genes
    - Channel: Grey, Light Saturation
    - Channel: Annotation channel (Gene name is annotated within the Line Mark itself)
    - Channel: Horizontal Position (Overlayed onto The Genome Line mark to represent the position of the gene in the field)
    - Channel: Length (How much of the field is taken up by the Gene)

ChIP-seq signal marks and channels: Original data:

- Line Mark: Peak
    - Channel: Height (Value of CHIP-seq Signal)
    - Channel: Hue (The categorical distinction of different signal types)
    - Channel: Vertical Position (The categorical distinction of different signal types)
        - The different signal types will be vertically stacked
        - Each mark will have an identical Y length
    - Channel: Vertical Position
        - 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
    - Channel: Saturation: Horizontal Grey and white striping to distinguish the different samples more easily

ChIP-seq signal marks and channels: Frequency graph:

- Line Mark: Peak (ATACseq, Methylation, Acetylation)
    - Channel: Horizontal Position (Where the peak is in the field)
    - Channel: Frequency (How strong the peak is for a given region of the field)
        - Strong peak = Solid color
        - Weak peak = Observable frequency with drop off
        - Scaled to 1
    - Channel: Amplitude (The How strong the peak is for a given region of the field)
        - Scaled to 1
    - Channel: Hue (The categorical distinction of different signal types)
    - Channel: Vertical Position (The categorical distinction of different signal types)
        - The different signal types will be vertically stacked
        - Each mark will have an identical Y length
    - Channel: Vertical Position
        - 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
    - Channel: Saturation: Horizontal Grey and white striping to distinguish the different samples more easily

Sketches of ChIPVis

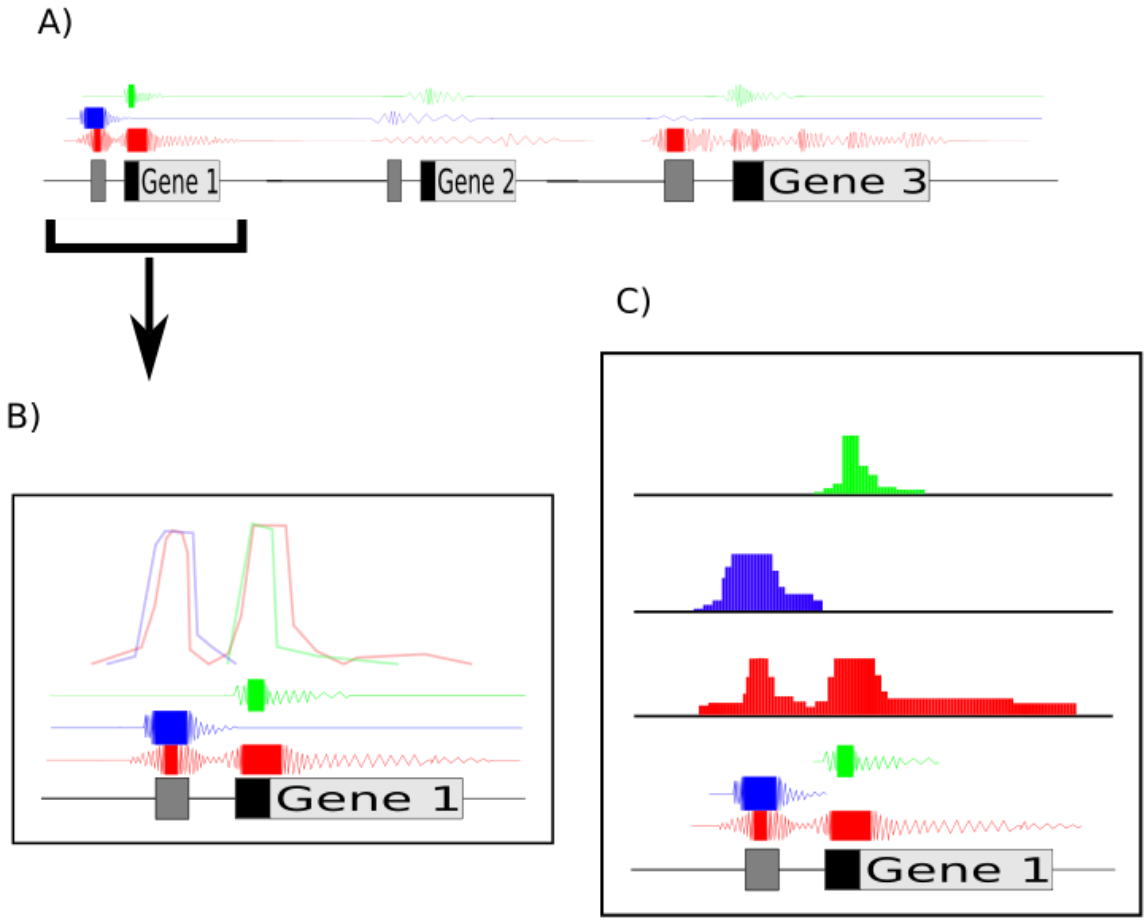

Fig. 2 An Example of the ChIPVis with One Sample:

The three windows of the proposed visulization 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 waves 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

The three windows of the proposed visulization 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 labled, 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.

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 signal 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 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 enhanced 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.

Implementation

Gosling is a grammar for interactive and scalable genomics data visualization. Gosling balances expressiveness for comprehensive multi-scale genomics data visualizations with accessibility for domain scientists. The accompanying JavaScript toolkit called Gosling.js provides scalable and interactive rendering. Gosling.js is built on top of an existing platform for web-based genomics data visualization to further simplify the visualization of common genomics data formats. Gosling can demonstrate the expressiveness of various usage scenarios and supports the design of novel genomics visualizations using semantic zooming.

6. Milestones

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

Table 4 - Project development chronogram and task distribution

Task Timeline	Proposal Oct 10th – 21st	Update Oct 22nd – Nov 15th	Presentation Dec 14th	Final Report Nov 16th – Dec 16th
Introduction	*Rodrigo*	*Rodrigo*	*All group members will collaborate during the development of the presentation*	*Rodrigo*
Related Work	*Rodrigo*	*Rodrigo*		*Rodrigo*
Data Abstraction	*Alex*	*Alex*		*Alex*
Task Abstraction	*Yerin*	*Yerin*		*Yerin*
Solution / Implementation	*Alex/Yerin*	*Alex/Yerin*		*Alex/Yerin*
Results	*Alex*	*Alex/Yerin*		*Alex/Yerin*
Discussion	-	*Rodrigo*		*Rodrigo*
Future work	-	*Rodrigo*		*Rodrigo*
Milestones	*Rodrigo*			-
Conclusion	-	-		*All*
Bibliography	*All*	*All*		*All*

7. **Discussion**
8. **Future work**
9. **Conclusion**

10. **Bibliography**

[1] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics,* vol. 10, pp. 669-680, 2009. https://doi.org/10.1038/nrg2641

[2] R. Nakato and K. Shirahige, "Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation," *Briefings in Bioinformatics,* vol. 18, no. 2, pp. 279-290, 2017. https://doi.org/10.1093/bib/bbw023

[3] R. Nakato and T. Sakata, "Methods for ChIP-seq analysis: A pratical workflow and advanced applications," *Methods,* vol. 187, pp. 44-53, 2021. https://doi.org/10.1016/j.ymeth.2020.03.005

[4] S. pepke, B. Wold and A. Mortazavi, "Computation for ChIP-seq and RNA-seq studies," *Nature Methods,* vol. 6, pp. 22-32, 2009. https://doi.org/10.1038/nmeth.1371

[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. https://doi.org/10.1038/jhg.2011.43

[6] S. Hino, T. Sato and M. Nakao, "Chromatin Immunoprecipitation Sequencing (ChIP-seq) for Detecting Histone Modifications and Modifiers," *Epigenomics,* vol. 2577, pp. 55-64, 2022. https://doi.org/10.1007/978-1-0716-2724-2_4

[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. https://doi.org/10.1007/s40484-013-0006-2

[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. https://doi.org/10.1109/TVCG.2009.116.

[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. https://doi.org/10.1038/nsmb.3180

[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, no. 18, pp. 2847-2852, 2014. https://doi.org/10.4161/15384101.2014.949201

[11] [A common ChIP-seq Visualization] adapted from *HBC Training* https://hbctraining.github.io/Intro-to-ChIPseq/lessons/10_data_visualization.html