DOI: 10.1111/cgf.12096
Eurographics Conference on Visualization (EuroVis) 2013
B. Preim, P. Rheingans, and H. Theisel
(Guest Editors)

Volume 32 (2013), Number 3

An Interactive Analysis and Exploration Tool for Epigenomic Data

H. Younesy[1] , C. B. Nielsen[†2] , T. Möller[1,3] , O. Alder[4] , R. Cullum[4] , M. C. Lorincz[5] , M. M. Karimi[5] and S. J. M. Jones[2]

[1]School of Computing Science, Simon Fraser University
[2]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency
[3]Faculty of Computing Science, University of Vienna
[4]Terry Fox Laboratory, British Columbia Cancer Agency
[5]Department of Medical Genetics, Life Sciences Institute, The University of British Columbia

Abstract

In this design study, we present an analysis and abstraction of the data and tasks related to the domain of epigenomics, and the design and implementation of an interactive tool to facilitate data analysis and visualization in this domain. Epigenomic data can be grouped into subsets either by k-means clustering or by querying for combinations of presence or absence of signal (on/off) in different epigenomic experiments. These steps can easily be interleaved and the comparison of different workflows is explicitly supported. We took special care to contain the exponential expansion of possible on/off combinations by creating a novel querying interface. An interactive heat map facilitates the exploration and comparison of different clusters. We validated our iterative design by working closely with two groups of biologists on different biological problems. Both groups quickly found new insight into their data as well as claimed that our tool would save them several hours or days of work over using existing tools.

Categories and Subject Descriptors (according to ACM CCS): H.5.m [Information Systems]: Information Interfaces and Presentation—Miscellaneous; I.3.8 [Computing Methodologies]: Computer Graphics—Applications

1. Introduction

Most cells in an organism share the same underlying DNA sequence (genome) and yet they display a great diversity of physical properties and functions. This diversity largely comes from differences in which genes are active (expressed) or silent (repressed) in each cell type. Changes in gene expression caused by mechanisms other than changes in the underlying DNA sequence are broadly referred to as epigenetic changes, where "epi" indicates a change "above" the genome. Examples of such mechanisms include chemical modifications to the DNA itself or to its associated proteins. We will refer to these chemical modifications as "epigenetic marks".

Techniques such as ChIP (chromatin immunoprecipitation) coupled with innovative DNA sequencing technology (ChIP-seq) have revolutionized our ability to measure the abundance of epigenetic marks across the genome, giving rise to so called "epigenomic data". Many large consortia such as ENCODE [The12] and the NIH Epigenomics Roadmap Project [BSC*10] have convened to exploit these technologies and perform hundreds of ChIP-seq experiments involving diverse cell types. The key challenge for new biological insight lies in integrative analysis, in which different data are combined and interpreted together, for example as patterns of epigenetic marks across different cell types.

While computational methods to interpret these data continue to evolve and improve, there is great value in data exploration and many questions remain too ill-defined to be addressed in an automated fashion. Visualization is thus a valuable tool in this domain. In addition, the rapidly changing computational tool set for data analysis often requires significant computational expertise to use. Many of the biologists who possess the detailed knowledge needed to interpret these data must rely on programming experts. As a result, interactive visualization holds great promise in being able to lower the computational barrier to analysis and en-

[†] Joint first author

gage biology experts more directly in data processing and interpretation.

In this paper, we address the need for visual analysis tools and present an interactive tool for visual exploration and analysis of epigenomic data. Our first contribution is a characterization of the data (see Section 2) and a discussion and abstraction of the related domain tasks (see Section 3). Second, we provide our design including an interactive heat map explorer and approaches for querying combinations and subsets (see Section 5). Third, we validate our approach by presenting two detailed case studies with two groups of domain experts working on different biological problems, and reporting their insights (see Section 6). We also comment on lessons learned in our design process and provide suggestions for other researchers working in this domain.

2. Biological Background and Data

2.1. Epigenetic Marks

Due to noise introduced at various stages of the ChIP-seq procedure, the resulting measurements of epigenetic marks are not binary values corresponding to the presence/absence of a given chemical modification at each position in the genome. Rather, ChIP-seq provides measurements of epigenetic mark enrichment across the genome and filtering methods are used to distinguish signal "peaks" from background noise. Each peak has a chromosome start and an end position and enrichment values across the interval. A commonly used format to store such peaks is the Wiggle (WIG) format [WIG]. Researchers may refer to peak data as wig files, track data, samples, or experiments, but we refer to them as "epigenetic marks" or simply "marks" throughout the paper.

2.2. Region Sets

In order to make sense of epigenetic marks across the genome, researchers often focus on genomic regions defined by features of biological interest. One common example is the set of start positions of known genes referred to as transcription start sites (TSS). A region set is a collection of genomic intervals and is usually described in GFF (General Feature Format) [Ste10] or BED (Browser Extensible Data) [BED] file formats. These formats capture the genomic locations of the regions and support inclusion of additional information, such as external database identifiers for each feature. The number of regions within the region set depends on the type of the analysis, but it usually varies between a few hundred to tens of thousands.

Often, the genomic intervals are of a fixed length centred on features of interest, for example, $\pm 1,000$ nucleotides (nt) around a TSS. It is also quite possible for them to be of different lengths, for example, the boundaries of annotated genes. Our users all asked for a fixed length interval and requested that variable length regions either be extended or truncated to this fixed length.

2.3. Data Abstraction

Many analysis tasks involve investigation of multiple epigenetic marks across a single region set. This allows us to consider only the subset of epigenetic mark values that fall within the target regions. Since the epigenetic marks and the region set use a common reference genome coordinate system, the mapping is straightforward. The result is a set of high-dimensional vectors, where each vector contains the values of a given epigenetic mark across a single region. If we define m as the number of epigenetic marks under consideration, r as the number of regions in the set, and l as the region length (same for all regions in the set), then this process will produce $m \times r$ vectors of length l.

In order to perform computationally efficient analysis on these vectors and to reduce artifacts such as signal spikes, it is common practice to accumulate multiple positions of each vector into a single bin. Our users typically used bin sizes of 50 to a few hundred, which have biological meaning in terms of nucleotide length. So, for example, regions of size 2,000 nt and a bin size of 50 nt, will result in vectors containing only 40 values. The binned values are normalized with methods such as the sigmoid function used by ChromaSig [HRW08] to enable comparison between multiple epigenetic marks.

3. Task Analysis

Data peaks are a very common starting substrate for analysis (see Section 2.1) and there are a great number of different questions biologists might wish to ask that can help reveal the functional role of the epigenetic marks. Common ones include "are there genes nearby and what are they?" or "do these peaks lie in regions with characteristic patterns?".

Through our discussions with several biologists, we have identified several tasks that are not well served by current tools. While solutions exist, they are awkward, do not immediately produce an interactive visual, or do not capture all the requested functionality. Here we categorize these tasks and highlight important considerations gleaned from discussion with analysts.

3.1. Task 1 - Signal Query

Considerations: (1) It remains open for debate whether the signal height from an epigenetic mark has biological significance. Given that they are acquired through an enrichment process, height is important for distinguishing signal from noise; however it is unclear whether more subtle height differences are meaningful. Biologists therefore tend to reason about peaks as being either present or absent. (2) An individual epigenetic mark is almost never considered in isolation. Part of the power of the ChIP-seq technology is to profile the positions of several modifications in the same cell or tissue type in parallel and integrate the results into a meaningful picture of the larger system. (3) While many analyses are exploratory, biologists very often have a particular signal pattern in mind. A common workflow is to query for regions with a pattern of interest and then explore from this starting point.

Example Question: "Show me the regions where there are peaks in marks A and B, but not in C or D. I don't care about the peak status in E through G."

3.2. Task 2 - Cluster

Considerations: (1) Clustering is a powerful exploration tool and is best used in cases where no precise query can be formulated. (2) It is also used in categorizing differences in signal position or distribution. (3) Researcher often want to explore the output of Task 1 through clustering.

Example Question: "In my target region set, what are the classes of data patterns in marks E through G?"

3.3. Task 3 - Quality Control

Considerations: (1) *k*-means clustering is widespread and well-known in the biology community. It has the well-known drawback that it requires the number of clusters as input (which requires informed guess work). *k*-means is also using random seeds, creating different clusters each time it is run. (2) Datasets "in the wild", including sequencing data, often do not have an obvious cluster structure, i.e. clusters often overlap. (3) Biologists desire to visually inspect the clusters to either assure the reliability of the clusters, try a different cluster number or do further downstream analysis of the found clusters.

Example Question: "Do most regions in the first cluster follow the trend of signal presence in mark A and absence in mark C?"

3.4. Task 4 - Comparison

Considerations: (1) Biologists very often want to find the intersection of sets. For example, clusters of interest could be obtained through iterations of Task 1 and 2. A biologist may then want to compare the clusters from different workflows to determine whether they contain the same or different regions. (2) Generation of many different intersections can be laborious and comparison of the output in visualization modules is cumbersome.

Example Question: "Do subset 1 and subset 2 contain the same or different regions?"

3.5. Task 5 - Downstream analysis

Considerations: Biologists will frequently need to generate visual and text outputs of their results and findings either to (1) read them into other tools for further analysis, (2) to communicate them with their peers or (3) to include them in manuscripts.

Example: "I want to use another tool to check the functional similarities of the regions in this subset."

4. Related Work

Genome browsers are a popular approach for visualizing genome-scale data [NCD*10] and play an important role in increasing the accessibility of large public data sets, such as the ENCODE data resource currently hosted by the UCSC Genome Browser [RDL*12]. Each epigenetic mark is displayed as a separate heat map or histogram plot, often called a "track", and then multiple marks can be viewed simultaneously by vertically stacking these tracks. Part of the power of this arrangement is that data from diverse marks are anchored to the same horizontal reference coordinate and can thus be readily compared.

Genome browsers are optimized for viewing one local region at a time. While this makes them valuable for detailed data inspection and exploration, it prevents them from aiding in global pattern analysis. Several techniques have emerged to facilitate global pattern discovery in epigenomic data. These include probabilistic methods for the discovery of epigenetic signatures *de novo*, such as ChromaSig [HRW08], and more recently, Hidden Markov Model approaches [EK10] and Bayesian network approaches [HBW*12] to uncover recurrent epigenetic states. However, these methods require significant computational skill to use and in most cases remain inaccessible to most biologists.

There are a handful of applications that attempt to bridge this computational gap in epigenomic data analysis. For example, CisGenome [JJM*08] contains a graphical interface for running analyses such as peak detection, false discovery rate computation and sequence analysis. Similarly, seqMINER [YKC*11] offers a range of data processing capabilities including an implementation of *k*-means clustering and a corresponding heat map display. This type of clustering and heat map view have been widely accepted for epigenomic data ever since their appearance in early analysis papers [HSH*07, HHH*09]. Most recently, Cistrome [LOT*11] provides integrative analysis and visualization tools for ChIP-seq data, taking advantage of the Galaxy platform [GNTT10]. The strength of these tools lies in their ability to connect diverse analysis methods in a single application. While they provide visualization components, the emphasis is on chaining tools into a workflow rather than on optimizing the visual representation and there is very little linking between the different visual displays.

Spark [NYO*12] is a recent tool that provides a visual workflow to address clustering. While it does a good job in helping to explore different clusters (Task 2), it doesn't provide a way to understand the variance of the clusters (Task 3) nor does it allow the user to query on/off combinations (Task 1) nor does it facilitate comparisons (Task 4).

Understanding combinatorial combinations (Task 1) has been hard and does not scale well. Practical implementations are typically constrained to just very few sets and are often visualized using Venn- or Euler-diagrams [CR04]. Alternative representations of combinatorial queries use iconic representations or a Karnaugh map [Huo08].

More recently StratomeX [LSS*12] provides visual subset comparison using ribbons of varying width drawn between neighboring columns. While this visual encoding works for comparison of multiple different cluster results, we specifically focus on just comparing two different results here. While we can envision to integrate some of its functionality into a future version of our tool, our focus was the integration of Tasks 1-5 in one simple tool at this point.

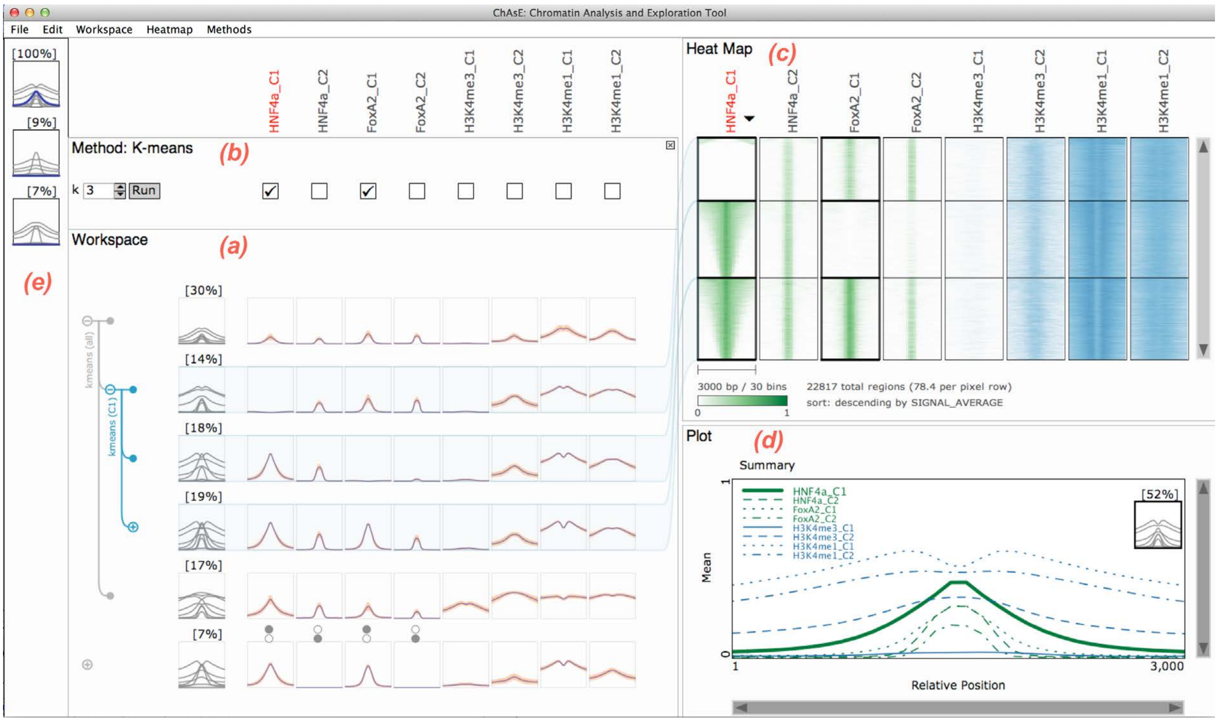


Figure 1: Interface: (a) Workspace Pane, (b) Method Pane, (c) Heat Map Pane, (d) Plot Pane, (e) Favourites Pane.

Finally, standard visual and interaction concepts such as Brushing and Linking [BC87] as well as Dynamic Queries [AWS92] are commonplace in today's visualization tools such that they are well understood by our users. Hence, our tool is making extensive use of these concepts.

5. ChAsE

We now describe our tool, called ChAsE (Chromatin Analysis and Exploration), and outline how our current approach addresses the analysis tasks discussed in Section 3.

5.1. Data Input

A graphical user interface allows specifying one or more epigenetic marks and one region set of genomic intervals. Processing parameters, such as the normalization options, or visualization parameters, such as heat map colour or ordering, can be specified per epigenetic mark. A visibility option was added after we observed that users preferred to load a larger set of epigenetic marks upfront and then modify it depending on their immediate analysis goals. The region size and number of bins need to be specified only once as they will be identical for preprocessing all epigenetic marks.

The processing time depends on the size of the input files but usually takes a few minutes per data file. The results of the processing are stored in the output directory specified by the user, so future data loading times will be much faster (a few seconds). Users can reopen the input dialog during analysis and modify the input parameters or add or remove marks without losing the current state of their analysis.

5.2. Interface

The ChAsE interface consists of five linked panes as shown in Figure 1. Data from a single region set and one or more epigenetic marks is first loaded into the Workspace Pane (Figure 1(a)). We will refer to this as the "full set". It can then be divided into various subsets, which we will simply call "set" or "sets", using functionality within the three alternate Method Panes (Figure 1(b), Figure 3). Once created, a set can be stored in the Favourites Pane (Figure 1(e)) for later use. The plots can be inspected in a zoomed view in the Plot Pane (Figure 1(d)). Closer inspection of data across individual regions is reserved for the Heat Map Pane (Figure 1(c)).

5.3. Workspace Pane

The Workspace Pane (Figure 1(a)) shows a snapshot of the current sets and is organized as a matrix. Each row of this matrix corresponds to one set and a column corresponds to a particular epigenetic mark. We chose a data representation commonly used by biologists in the field, called a "profile plot". The *x*-axis captures offsets from the region start (e.g. position relative to a TSS) and the *y*-axis is used to express a summary statistic for all values at these relative positions. A profile plot summarizes the data for each epigenetic mark in each set (i.e. for each cell in the matrix) providing the user with a quick visual summary of the data patterns.

Offset from the main matrix, the leftmost column displays a summary of all the epigenetic marks in one row as overlaid profile plots which we call a "summary plot". Comparing signal distributions across many columns can be challenging and the summary plot offers a valuable mechanism for spotting subtle differences in signal distributions between epigenetic marks. To aid this comparison, the summary plot and profile plots are linked, such that when a user mouses over a column, its corresponding curve in the summary plot is highlighted. The size of each set is shown in square brackets as a percentage of the full set or the actual number of regions when the size drops below 1%. Each set can also have a user specified title or descriptive note allowing the users to keep track of their history. Clicking on either the summary plot or any individual profile plot automatically displays it in the Plot Pane (Figure 1(d), lower right) and Heat Map Pane (Figure 1(c), upper right) for closer inspection.

5.3.1. Profile Plot Views

A user can alternate between different profile plot visualizations and their size through a context menu. We provide four choices of summary statistics for display in the profile plots (Figure 2): The *Mean and standard deviation view* (2a) shows the average signal profile of the region set surrounded by the $+/-$ standard deviation range. The *Continuous box plot view* (2b) shows the median signal surrounded by the quartile boundaries as an indication of the range and frequency of signal heights. The *Mean and signal scatter view* (2c) shows the average signal profile as well as a scatter of all profiles for the region set accumulated and rendered with a log scale. The *Mean and peak scatter view* (2d) accumulates only the max peak value per region for each epigenetic mark rather than the entire profile, addressing users' expressed interest in the distribution of the peaks in a set in terms of their height and location.

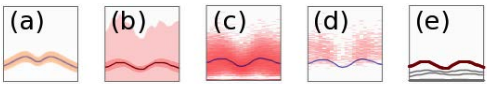


Figure 2: Profile plot views: (a) mean +/- standard deviation, (b) continuous box plot, (c) mean and signal scatter, (d) mean and peak scatter, and (e) summary plot.

5.4. Method Panes

There are three Method Panes (Figure 3) to address Tasks 1, 2, and 4 outlined in Section 3: Signal Query, Cluster, and Comparison. Only one of these three Method Panes is displayed at a time and they always appear at the top of the Workspace Pane. We found, this minimized confusion and allowed the user to focus on a single method.

5.4.1. Signal Query Pane

We explored several different data encodings and interaction schemes to help the user specify a particular signal query. Displaying all possible on/off (i.e. present/absent) combinations across all epigenetic marks would quickly lead to an overwhelming number of options and was impractical. So, it was important to enable our users to limit the combinations by specifying whether a signal should be on or off or either in each epigenetic mark.

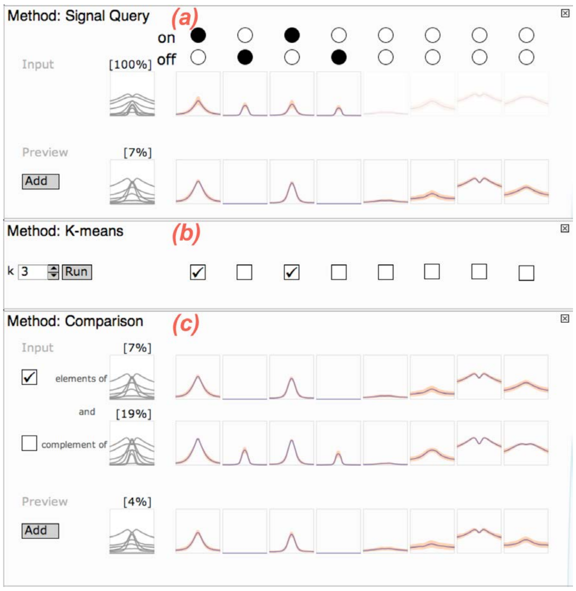


Figure 3: Method Panes: (a) Signal Query (b) Cluster (c) Comparison.

We initially tried providing constraints that allowed the user to express multiple combinations at once. For example, a user could specify both the on and off state for mark A and just the on state for mark B. This would give rise to three sets: A-on and B-on, A-off and B-on, and the rest. However, this approach of expressing combinations was not intuitive to our users. Instead, during our discussions, they would often simply draw out the combinations of interest, one at a time. The number of combinations our users wished to generate tended to be small compared to the space of possibilities. The process could be thought of as querying for individual signal sets and we therefore decided to support this one-at-a-time querying more directly.

Figure 3a shows our final Signal Query Pane design. A user opens this pane by selecting a target set and choosing "Signal Query" from the top “Methods” menu. A pair of check boxes appears above each column and allows the user to specify on (top box checked), or off (bottom box checked), or indifference (neither checked). As the user modifies the query, a preview of the resulting set is shown at the bottom of the pane as a row of profile plots as well as on the heat map to the right of the pane. In cases where the resulting set is empty no plot or heat map will be shown. The resulting set is only imported into the Workspace once the user clicks the "Add" button. This allows the user to accumulate sets in their Workspace Pane when there are multiple desired combinations, as well as to make the sets available for further analysis (i.e. clustering or comparison). Annotations are shown above the created sets in the workspace showing the query used to create the subset (an example shown at the bottom of the Workspace Pane in Figure 1(a)).

5.4.2. Cluster Pane

A user initiates clustering by selecting the target set and selecting "*k*-means clustering" from the "Methods" menu. In addition to the number of clusters, the Cluster Pane allows specifying the epigenetic marks to be included in the clustering step using the check boxes above each mark (Figure 3b). Clicking the "Run" button commences the clustering run. Once the process is complete, the resulting clusters appear in the Workspace Pane as the children of the input set in a tree structure. The heat map view shows the clusters separated by horizontal lines with a thicker stroke used to indicate the marks included in the clustering.

Because clusters can be subsequently subclustered, we needed to manage potentially large tree structures. Leaf nodes (i.e. clusters with no sub-clusters) are represented with solid circles, whereas parent nodes (i.e. clusters with sub-clusters) are represented by either a ⊖ sign, to indicate an expanded node, or with a ⊕ sign, to indicate a collapsed node. Allowing the user to toggle between expanded and collapsed states by clicking on the parent nodes made the tree structure manageable.

The user can explore the newly created clusters or choose to rerun *k*-means clustering with the same or different parameters. It is a known fact that the result of *k*-means will depend not just on the value *k*, but also on the initial seeding consisting of *k* randomly selected members of the input set. Thus each run of *k*-means can result in a different clustering. We chose not to fix this seeding to artificially hide this drawback of the *k*-means algorithm. Our users were aware of this fact and tended to run *k*-means until an interesting clustering is observed or until they could assess the reproducibility of a cluster (part of Task 3).

5.4.3. Comparison Pane

Comparisons can be formulated as queries for the intersection across multiple sets (Task 4). A comparison is initiated by selecting two or more sets from the Workspace Pane while pressing the Shift key and then selecting "Cluster Comparison" from the "Methods" menu. As shown in Figure 3c, the Comparison Pane displays the input sets and a preview of the intersection using the same profile plot display found in the Workspace Pane. The check boxes on the left of the summary plots allows the user to specify either inclusion (checked) or exclusion (unchecked) of the set and a label is shown for clarification of the set operation. Initially all check boxes are checked, thus the result is the intersection of all sets.

5.5. Heat Map Pane

Heat maps are one of the most widely used visual encodings for biological data [WF09]. They encode the values of a data matrix as shades of colour. Heat maps are used in different stages of the research from data analysis to presentation, but despite their popularity, there are valid arguments against their use [Won10]. It is much harder to compare signal variations and the overall signal shape from colour variations alone, so other encodings such as profile curves [MWS*10, MMDP10], are used alternatively. In addition, the resolution of the data is usually higher than the resolution of the heat map, and therefore the pixels represent an average and can hide certain data characteristics such as local peaks. We partially address these concerns with our interactive heat map.

Figure 1(c) depicts the Heat Map Pane. Each column corresponds to a single epigenetic mark and the rows correspond to genomic regions. To render the heat map, the values of the data matrix are mapped to the colour specified by the user. Users tend to use different colors for marks with different biological nature. We provide a set of six sequential colour schemes with the same perceived intensity as well as two diverging colour schemes which we picked using ColorBrewer [Bre12]. Regions can be sorted by the signal in one mark at a time. All columns are coordinated such that the row order is the same across columns. An arrow above a column indicates the mark currently dictating the sorting. The direction of the arrowhead indicates the sort order and can be flipped when clicked. Underneath the heat map, a legend provides the total number of the regions, the display density (regions/pixel row), and the current sorting criteria.

Regions are initially sorted by their order in the input regions file, but different sorting criteria, such as signal average, signal max, or signal peak offset, can be chosen by the user through a context menu or the top "Heatmap" menu. Sorting is commonly used to get an overview of the distribution of the signal value and shape across regions of an epigenetic mark, comparing the correlations between multiple epigenetic marks, as well as visually assessing the quality and variations within the clusters (Task 3). Regions belonging to a collapsed parent, will all be sorted together. For an expanded parent, regions belonging to different children will be sorted separately.

As stated above, heat maps suffer from at least two major short comings. First, detecting the shape of the signal from the colour variation is not straightforward. To address this, as the user drags the mouse pointer on the heat map we show a profile plot of the regions underneath the heat map over the legend information. Second, when too many regions are overlapped and averaged within one row of pixels, we allow users to interactively zoom and pan through the heat map. This is realized by a resizable scrollbar on the right side of the heat map.

5.6. Plot Pane

The plot pane (Figure 1(d)) shows a zoomed version of the selected profile plot and includes additional legends and axis labels. One of our users showed most interest in a zoomed summary plot where the profiles for all epigenetic marks for a region set are overlaid together. Hence, the user can adjust the horizontal and vertical view range through resizable scrollbars. A small view of the entire plot is shown on the top right corner with the invisible view range shaded.

5.7. Favorites Pane

We were often asked by the users to be able to save the current results of their analysis or partial hypothetical findings before performing different tasks. Although it is possible to have all of them in the workspace view and save them to file, this would have cluttered the workspace view in the long run competing with the goal of quick and easy access to the current working sets. We thus provided a favourite pane (Figure 1(e)), where region sets within the workspace could be added for future reference and brought back to the workspace as needed. Regions in the favourite pane are shown with a summary plot and their size.

5.8. Common Functionality

In addition to the functionality specific to each view specified above, most views share some common functionality, which are available through contextual menus. This functionality includes operations such as annotation, removal, and export of region sets. Further, the user can save images of the heat map or profile plots as high resolution PDF files (Task 5).

6. Case Studies

Our design process had three phases: (1) iterations of interface sketches based on feedback we received from domain experts, (2) implementation of an initial prototype interface based on these refined sketches, and (3) an iterative refinement of the prototype based on feedback from biologists after using the prototype. During this last phase, we first gave the users a tutorial on the use of the prototype. We then loaded their data and observed them using our tool. During this session, the users offered out loud descriptions of their thoughts while using the tool. We then collected more reflective feedback after deployment of ChAsE for several weeks. Here we present illustrated walkthroughs of two case studies with two groups of collaborators.

6.1. Case Study 1: Signal Querying and Clustering

Our first group of collaborators were two biologists who were researching the co-localization of patterns across four marks in human liver cells under two conditions C1 and C2.

6.1.1. Analysis 1: Filtering Using the Signal Query

Using a set of regions centred on peaks collected from four marks, labelled HNF4a_C1, HNF4a_C2, FoxA2_C1 and FoxA2_C2, our collaborators' first step was to filter for regions containing signal from two or more of the four marks. This corresponds to 11 out of the possible 16 combinations of presence or absence of signal in four marks. Using the Signal Query Pane, they first identified the four sets in which signal is present for only one of the four marks (4(a)). Next, they used the Comparison Pane to exclude these sets from the full set by intersecting their complements (Figure 4(b)). While this took less than a minute, a similar workflow with their previous tools would have required them to extract each of the 11 combinations, taking them tens of minutes.

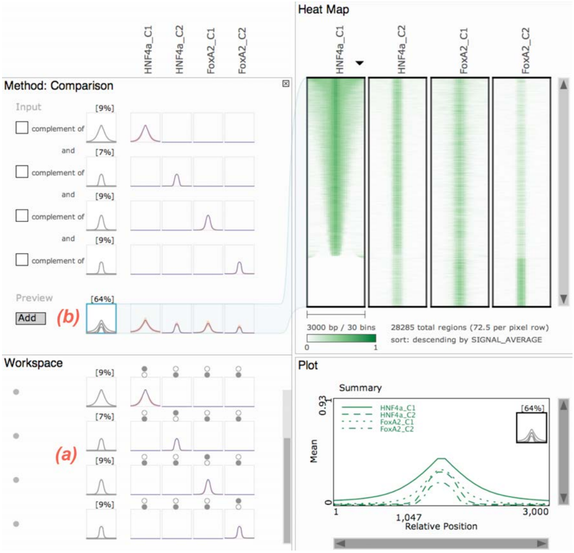


Figure 4: Analysis 1: Comparison Pane is used to exclude regions in which only one mark out of four showed a signal.

6.1.2. Analysis 2: Finding Patterns Using the Signal Query

Our collaborators then wanted to identify those regions with signal in HNF4a_C1 and FoxA2_C1, but not in HNF4a_C2 and FoxA2_C2. This would have been frustrating, if not impossible, to achieve with *k*-means clustering alone, but it was readily performed using the Signal Query Pane (Figure 5).

Our collaborators then scanned across the resulting profile plots and inspected the data patterns for seven additional marks not used in the query step. Several observations resulted that confirmed their predictions for the regions in the set created using signal query:

1. H3K4me3_C1 and H3K4me3_C2 are only weakly associated with these regions. This is consistent with previous observations [HRZ*10, HSH*07]
2. H3K4me1_C1 and H3K4me1_C2 differ in their distributions across the regions; H3K4me1_C1 displays a distinct bimodal (two peaks) distribution, whereas H3K4me1_C2 appears unimodal (single peak). This observation is also consistent with previous reports [HRZ*10].
3. These regions have very low levels of H3K27me3_C1 and H3K9me3_C1, which is expected for transcriptionally active sites.
4. H4ac_C1 mimics the bimodal distribution pattern of H3K4me1_C1, also reported previously [HSH*07].

Identification of such subsets based on data presence or absence can otherwise be done using an intersection tool, such as that provided in Galaxy [GNTT10] or written in custom code. Each intersection must be performed separately and stored for subsequent loading into different tools that provide profile plot or heat map views. This makes the generation of multiple sets laborious and the comparison of profile plots from other marks difficult. Our tool markedly shortened the time our collaborators needed to generate filtered sets of interest and also provided instant feedback regarding the corresponding data patterns in other marks.

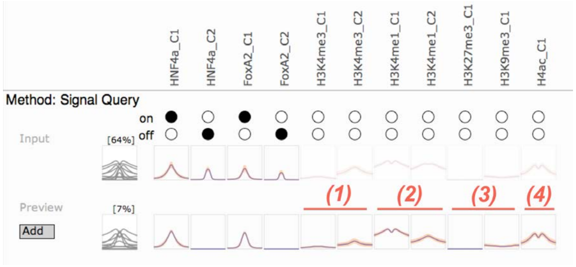


Figure 5: Signal Query Pane used in Case Study 1. Numbers correspond to the observations in Analysis 2.

6.1.3. Analysis 3: Chaining Querying and Clustering

A more detailed inspection of the heat map and brief exploratory sorting of the columns revealed a small set of H3K4me1_C1 with a unimodal rather than a bimodal distribution. Subsequent *k*-means clustering on that mark isolated the unimodal set (highlighted in Figure 6). While unimodal profiles for H3K4me1 have been observed previously in other conditions [HRZ*10], the unimodal pattern for H4ac is undocumented and warrants further investigation.

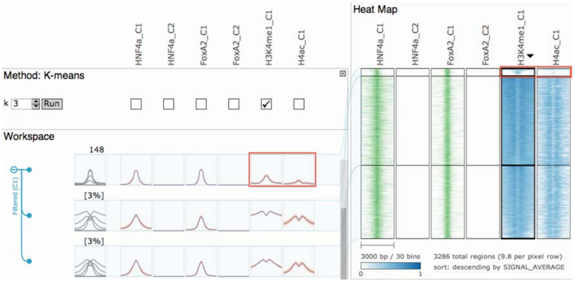


Figure 6: *k*-means clustering of the set shown in Figure 5. The unimodal distribution of H3K4me1_C1 and H4ac_C1 is highlighted with a border.

This analysis illustrated the value of enabling users to interleave their steps of analysis while providing them with visualizations to support quality control in the process.

6.2. Case Study 2: Exploration with the Interactive Heat Map

Our second group of collaborators were a biologist and a bioinformatician who were studying the relationship between several different marks in mouse embryonic stem cells. For this analysis, the region set consisting of about 30,000 regions in the neighbourhood of characterized genes (TSS +/- 1,000 base pairs) and a total of six different marks were loaded into the tool and labelled CpG, 5-mC, 5-hmC, H3K4me3, HeK27me3, and TET1.

6.2.1. Analysis 1: Initial Exploration using the Heat Map

Unlike in Case Study 1, this group of collaborators wanted to use their original unfiltered data, which was guaranteed to have some low and noisy signals in most regions. This prevented them from taking advantage of the Signal Query or the Comparison Panes effectively and they only employed the clustering and heat map browsing in their analysis. To further support this task, we introduced a divergent colour scheme to make it easier to judge whether the data values are low, medium or high (blue, yellow, and red, respectively).

Initial browsing of the data in the Heat Map Pane while sorting the regions by the average value of different marks, showed co-localization of CpG and H4K4me3, but an anti-correlation with 5-mC and 5-hmC. This is shown in Figure 7 and is consistent with the previous published studies [BHE*02].

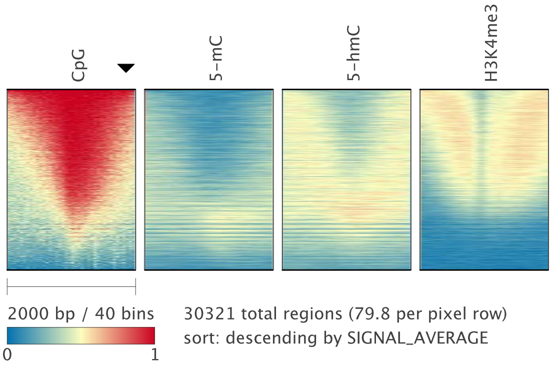
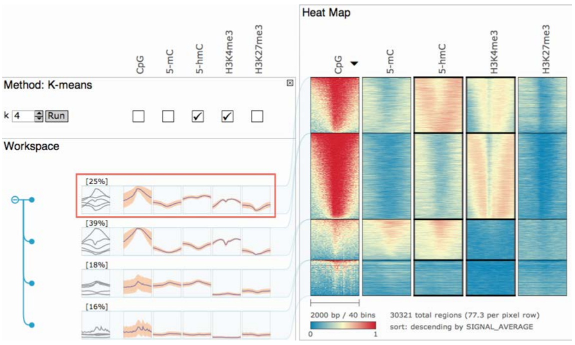


Figure 7: Heat map sorted by CpG. This figure is a direct PDF export from the tool.

6.2.2. Analysis 2: Coupling clustering with the Interactive Heat Map

Our collaborators then added a fifth mark H3K27me3 and experimented with different clusterings using the Cluster Pane. Their initial hypothesis was that when 5-hmC and H3K4me3 are present, H3K27me3 should be absent. In biological terms, this would indicate that 5-hmC is present at transcriptionally active genes, where H3K4me3 is high and H3K27me3 is low. By clustering on 5-hmC and H3K4me3 only, our collaborators noticed that the cluster with both 5-hmC and H3K4me3 also unexpectedly showed some H3K27me3 signal (top cluster in Figure 8).

This observation led them to explore the pattern using a different clustering based on H3K4me3 and H3K27me3 alone. As shown in Figure 9, they were able to identify regions where 5-hmC and H3K4me3 are present and H3K27me3 is absent (top), as originally predicted, but also uncovered another class of regions in which all three are present at high/moderate levels (middle), which enabled them to rule out their original hypothesis. This observation was later confirmed to be consistent with recently published results [YHS*12].



Method: K-means

k 4 Run ☐ ☐ ☑ ☑ ☐

Figure 8: Clustering by H3K4me3 and 5-hmC into four clusters. A cluster with a high level of H3K4me3 and 5-hmC, but low level of 5-mC is highlighted with a border (top cluster).

To investigate the possible biological reasons for these patterns, our collaborators added a sixth mark, TET1. They observed that the top cluster despite having medium to low values of 5-mC and 5-hmC had a high value of TET1. In biological terms, TET1 is a protein that facilitates a chemical change from 5-mC to 5-hmC. Thus our collaborators were able to conclude that any 5-hmC produced from 5-mC in the presence of TET1 is present only transiently and is presumably rapidly further processed. This was a valuable insight.

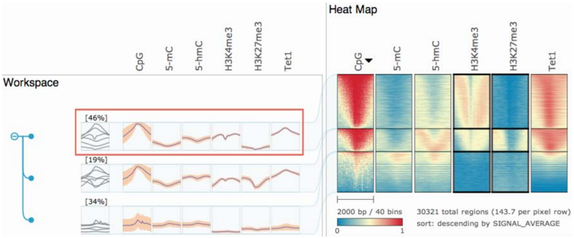


Figure 9: k-means clustering by H3K4me3 and H3K27me3 into three clusters. A cluster with low levels of both 5-mC and 5-hmC is highlighted with a border (top cluster).

The Cluster Pane facilitated these analyses by providing a simple interface to specify the number of clusters and the marks to be included in the clustering. Once the clustering was complete, a preview of the result was immediately shown as a tree view of profile plots in the Workspace Pane and as sub-clusters sorted individually in the Heat Map Pane. This allowed our users to quickly observe the variation within the clusters and check for existence of interesting patterns, and to rerun the clustering to test if it was stable.

This study showed the value of supporting gradual exploration. In the past, our collaborators had used scripting and Matlab for analysis and similar steps had taken them much more effort to accomplish. This use case provides an example of how clustering is best used for pattern discovery at the point when the researcher wishes to perform an exploratory analysis or wants to isolate a set of regions based on data distribution and not pure presence/absence of signal.

This study also showed the usefulness of the heat map view to reveal variation and spatial patterns within clusters. For instance, in Figure 9 the difference between top and middle clusters is much more visible from the heat map compared to just the profile plots.

7. Lessons Learned

Perhaps one of the most difficult aspects to get right was the ability to deal with a quickly expanding set of possible combinations. In several previous iterations of our design, our signal querying pane enabled the user to create combinatorial combinations of on/off behaviours. This often resulted in too many combinations being displayed of which the user was simply interested in a subset. Only after restricting this interface to query exactly one of these combinations at a time did we resolve the usability issues. This was possible after realizing that our users really only needed to analyze a very few and very specific combinations and hence, it was best to have them query them one-by-one. Although some of these tasks could be done in other tools, these tools were sufficiently complicated to use that the effort was not in balance with the payoff.

Further, besides assuring flexible output formats enabling a proper downstream-analysis, all of our users were very keen on functionality that would let them produce high-resolution figures for their publications and communication of their results to peers. There is perhaps a key set of functions that should be provided with most tools, including, but not limited to the export of high-resolution images in standard formats, annotation of features of interest, and customization of colour maps, labels and fonts. It is important to note, that the image resolutions for publications should often be higher than the typical screen resolution.

8. Future Work

Through discussions with our users, we have identified a number of possible extensions to ChAsE. These include providing improved guidance to the user on the choice of k used during the clustering step, in addition to providing metrics of cluster stability. Being able to reproduce a particular clustering would also be desirable (current clustering is sensitive to initial cluster seeds). There is also potential to provide visual feedback on some of the upstream processing steps such as peak-calling, not addressed here. For example, many peak-calling tools remove peaks with a max height below some threshold. Being able to interactively tune that threshold and visualize the results would be of great value.

9. Acknowledgements

Funding for H.Y. was provided by NSERC PGS-D scholarship. C.B.N. is a CIHR and MSFHR post-doctoral fellow. M.C.L. is supported by CIHR grant 20R91610. M.K. is a MSFHR post-doctoral fellow. S.J.M.J. is a senior scholar of MSFHR. We thank Brad Hoffman, Arthur Kirkpatrick and Michael Sedlmair for helpful comments.

References

[AWS92] Ahlberg C., Williamson C., Shneiderman B.: Dynamic queries for information exploration: an implementation and evaluation. In *Proc. of the SIGCHI Conf. on Human Factors in Comp. Sys.* (1992), CHI '92, ACM, pp. 619–626. 4

[BC87] Becker R. A., Cleveland W. S.: Brushing Scatterplots. *Technometrics 29*, 2 (May 1987), 127–142. URL: http://www.jstor.org/stable/1269768. 4

[BED] UCSC Genome Bioinformatics – Browser Extensible Data (BED) [online]. URL: http://genome.ucsc.edu/FAQ/FAQformat.html [cited 1 Dec 2012]. 2

[BHE*02] Bernstein B., Humphrey E., Erlich R., Schneider R., Bouman P., Liu J., Kouzarides T., Schreiber S.: Methylation of histone H3 Lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences 99*, 13 (2002), 8695–8700. 8

[Bre12] Brewer C.:. ColorBrewer 2.0 [online]. Dec 2012. URL: http://www.ColorBrewer2.org [cited 1 Dec 2012]. 6

[BSC*10] Bernstein B., Stamatoyannopoulos J., Costello J., Ren B., Milosavljevic A., et al.: The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology 28*, 10 (2010), 1045–1048. 1

[CR04] Chow S., Ruskey F.: Drawing area-proportional Venn and Euler diagrams. In *Graph Drawing*, Liotta G., (Ed.), vol. 2912 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2004, pp. 466–477. 3

[EK10] Ernst J., Kellis M.: Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol 28*, 8 (Aug 2010), 817–825. 3

[GNTT10] Goecks J., Nekrutenko A., Taylor J., The Galaxy Team: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Bio. 11*, 8 (2010), R86. 3, 7

[HBW*12] Hoffman M., Buske O., Wang J., Weng Z., Bilmes J., Noble W.: Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods 9*, 5 (2012), 473–476. 3

[HHH*09] Heintzman N., Hon G., Hawkins R., Kheradpour P., Stark A., Harp L., et al.: Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature 459*, 7243 (2009), 108–112. 3

[HRW08] Hon G., Ren B., Wang W.: ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Computational Biology 4*, 10 (Oct 2008), e1000201. 2, 3

[HRZ*10] Hoffman B., Robertson G., Zavaglia B., Beach M., Cullum R., et al.: Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome research 20*, 8 (2010), 1037–1051. 7, 8

[HSH*07] Heintzman N., Stuart R., Hon G., Fu Y., Ching C., Hawkins R., et al.: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics 39*, 3 (2007), 311–318. 3, 7

[Huo08] Huo J.: KMVQL: a visual query interface based on Karnaugh map. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2008), AVI '08, ACM, pp. 243–250. 3

[JJM*08] Ji H., Jiang H., Ma W., Johnson D. S., Myers R. M., Wong W. H.: An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol 26*, 11 (Nov 2008), 1293–1300. 3

[LOT*11] Liu T., Ortiz J., Taing L., Meyer C., Lee B., Zhang Y., Shin H., Wong S., Ma J., Lei Y., et al.: Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biology 12*, 8 (2011), R83. 3

[LSS*12] Lex A., Streit M., Schulz H., Partl C., Schmalstieg D., Park P. J., Gehlenborg N.: StratomeX: Visual analysis of Large-Scale heterogeneous genomics data for cancer subtype characterization. *Computer Graphics Forum (EuroVis '12) 31*, 3 (2012), 1175–1184. 3

[MMDP10] Meyer M., Munzner T., DePace A., Pfister H.: MulteeSum: A tool for comparative spatial and temporal gene expression data. *IEEE Transactions on Visualization and Computer Graphics 16* (2010), 908–917. 6

[MWS*10] Meyer M., Wong B., Styczynski M., Munzner T., Pfister H.: Pathline: A tool for comparative functional genomics. *Computer Graphics Forum 29*, 3 (2010), 1043–1052. 6

[NCD*10] Nielsen C. B., Cantor M., Dubchak I., Gordon D., Wang T.: Visualizing genomes: techniques and challenges. *Nature Methods 7*, 3 Suppl (Mar 2010), S5–S15. 3

[NYO*12] Nielsen C., Younesy H., O'Geen H., Xu X., Jackson A., Milosavljevic A., Wang T., Costello J., Hirst M., et al.: Spark: A navigational paradigm for genomic data exploration. *Genome Research* (2012). 3

[RDL*12] Rosenbloom K., Dreszer T., Long J., Malladi V., Sloan C., et al.: ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic acids research 40*, D1 (2012), D912–D917. 3

[Ste10] Stein L.:. Generic Feature Format version 3 [online]. Dec 2010. URL: http://www.sequenceontology.org/gff3.shtml [cited 1 Dec 2012]. 2

[The12] The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. *Nature 489* (2012), 57–74. 1

[WF09] Wilkinson L., Friendly M.: The history of the cluster heat map. *The American Stat. 63*, 2 (2009), 179–184. 6

[WIG] UCSC Genome Bioinformatics – Wiggle Track Format (WIG) [online]. URL: http://genome.ucsc.edu/goldenPath/help/wiggle [cited 1 Dec 2012]. 2

[Won10] Wong B.: Points of view: Color coding. *Nature Methods 7*, 8 (Aug 2010), 573–573. 6

[YHS*12] Yu M., Hon G., Szulwach K., Song C., Zhang L., Kim A., Li X., Dai Q., Shen Y., Park B., et al.: Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* (2012). 8

[YKC*11] Ye T., Krebs A. R., Choukrallah M. A., Keime C., Plewniak F., Davidson I., Tora L.: seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res 39*, 6 (Mar 2011). 3