ISSN 1213-6972 Journal of WSCG http://www.wscg.eu/ Vol.26, 2018, No.1

Analyzing Histone Modifications Using Tiled Binned Clustering and 3D Scatter Plots

Dirk Zeckzer[1]
zeckzer@informatik.uni-leipzig.de

Daniel Wiegreffe[1,2]
daniel@bioinf.uni-leipzig.de

Lydia Müller[2,3]
lydia@bioinf.uni-leipzig.de

[1]Image and Signal Processing Group, Leipzig University

[2]Bioinformatics, Leipzig University

[3]Natural Language Processing Group, Leipzig University

ABSTRACT

A major goal in epigenetics is understanding how cells differentiate into different cell types. Besides the increase of individual data sets, the amount of replicated experiments generating a tremendous amount of data is ever increasing. While biologists primarily analyze their data on the highest level using statistical correlations or on the lowest level analyzing nucleotide sequences, determining the fate of histone modifications during cell specification necessitates improved analysis capabilities on one or more intermediate levels. For this type of analysis, it proved to be very useful to use tiled binned scatter plot matrices showing binary relationships or to use tiled binned 3D scatter plots showing ternary relationships. Quarternary or general n-ary relationships are not easily analyzable using visualization techniques like scatter plots, only. Therefore, we augmented existing clustering methods with the tiling and binning idea enabling the analysis of n-ary relationships. Analyzing the changes of histone modifications comparing two cell lines using tiled binned clustering, we found new, unknown relations in the data.

Keywords

Clustering, Binning, Tiles, 3D Scatterplot, Biological Visualization, Histone Modifications

1 INTRODUCTION

Epigenetics, "The study of heritable changes in gene expression that are not mediated at the DNA sequence level" [5] is a very important research area. As part of epigenetics, researcher study histone modifications. Histone modifications are important for the development of the cells regulating the transcription states of genes and thereby their overall expression patterns. The evolution of certain histone modifications plays an important role during the differentiation of cells, e.g., from embryonic stem cells to embryonic fibroblasts or to neural progenitor cells [23, 27, 28].

In order to analyze this fate of histone modifications, one or more intermediate levels of data analysis are needed that go beyond the high-level statistical correlations and the low-level sequence-level analyses. Steiner et al. [23] proposed a visualization based on self-organizing maps (SOMs) on an intermediate level. Moreover, Zeckzer et al. proposed tiled binned scatter plot matrices (TiBi-SPLOM) [28] and tiled binned 3D scatter plots (TiBi-3D) [27] fostering analyzing the fate of epigenetic marks during differentiation. While our previous approaches [23, 27, 28] provide effective and efficient visualizations of histone modification data supporting these intermediate level analyses, there is still room for improvements.

To go beyond *binary* and *ternary* relationship analyses supported by these approaches and thus supporting general *n-ary* relationship analysis, we propose tiled binned clustering. The results of this processing step are then displayed using TiBi-3D-like visualizations. Using visualizations based on TiBi-SPLOM could be used as an alternative but are not yet implemented.

Concretely, the contributions of this paper are:

- *Visual Analysis*
  - A tiled binned clustering method that together with an adapted tiled binned 3D scatter plot fosters analyzing *n*-ary relations on histone modification data.
  - A complete work flow from the raw data to the final visualization and interaction implemented in TiBi-Cluster.

https://doi.org/10.24132/JWSCG.2018.26.1.1

- *Biology*
  - Comparison of 6 histone modifications between mesendodermal cells and mesenchymal stem cells based on data from the NIH Roadmap Epigenomics project [20].

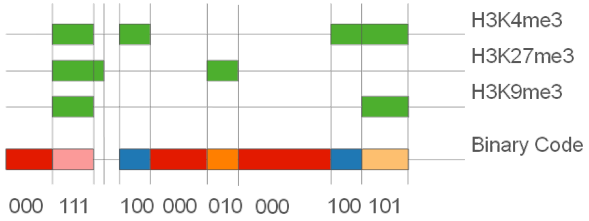
2 BIOLOGICAL BACKGROUND

2.1 Data Sets

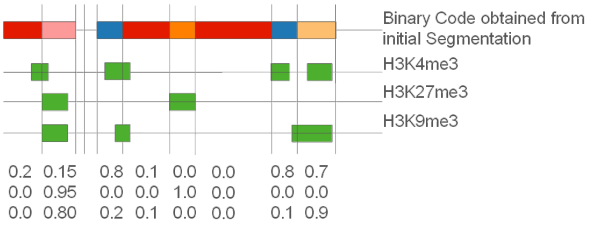
As data sets, we chose mono-, di-, and trimethylation of histone H3 at lysine K4 (H3K4me1, H3K4me2, H3K4me3), acetylation and trimethylation of histone H3 at lysine K27 (H3K27ac, H3K27me3), and trimethylation at histone H3 at lysine K36 (H3K36me3). Those data sets are available on the NIH Roadmap Epigenomics project's web portal (`http://egg2.wustl.edu/`, Release 9) [20]. The first three modifications are associated with enhancers and transcription. H3K4me1 is found at enhancers in the promoter regions. It is associated with silencing and even demarcates H3K4me3 [4]. H3K4me2 is a mark for transcription factor binding sites and increases the binding of transcription factors to their binding site. H3K4me3 is found at promotors and enhances the binding of the polymerase complex. Acetylation at H3K27 changes the charge of the histone, and thus, leads to an open chromatin formation enabling transcription. It is furthermore found with H3K4me2 at transcription factor binding sites. H3K27me3 is a mark indicating repressed transcription. H3K36me3 is known to enrich at splice sites. In this way, it indicates not only active transcription but also splicing events at specific splice sites.

2.2 Preprocessing

To measure those epigenetic marks between multiple cell types and cell lines, ChIP-seq (chromatin immunoprecipitation sequencing) is performed for every replicate. The results of this procedure consist of billions of short DNA sequences distributed over the whole genome marking the associated histone. We downloaded the raw data of these ChIP-seq experiments. During the next step, the data is mapped against the human reference genome version hg19 with the mapper segemehl [11] with 80% accuracy. Afterwards, we used the Picard Tools [1] to remove PCR duplicates. These steps result in the short DNA sequences being annotated to specific regions in the human genome. As ChIP-seq data contains small errors and some DNA sequences are falsely aligned to regions within the genome, only the significant peaks within the genome-wide distribution of the mapped DNA sequences are treated as a valid histone modification mark (the method being applied is called 'peak-calling'). We used the peak-caller Sierra Platinum [18] to extract those significant regions having support by multiple replicates.



(a) Example of the binary code resulting from segmentating the reference cell line (or reference cell type) based on 3 histone modifications (marks). The Binary Code represents the type of modification pattern present in the chromatin segment (here: 8 different codes).



(b) Example of the additional data annotation based on the segmentation of the reference data. The vectors represent the type of modification pattern present in this segment relative to the reference. The same three histone modifications for a different cell type are used here.

Figure 1: Overview of the two-stage segmentation procedure: Segmentation of the reference data and annotation of the additional data.

Based on these peak calls we calculated a whole genome segmentation. This method was described in detail by Steiner et al. [23]. During this process, all peaks are projected onto the genome (see Figure 1(a)). Nucleotide sequences with the same combination of peaks from different histone modifications as well as nucleotide sequences without any peaks are the *segments*. The approximated length of DNA wrapped around a histone complex together with the linker between two of them is approximately 200 nucleotides. As shorter segments are most likely noise, they are discarded during segmentation.

The generated segments are described by their location using a unique identifier. Besides this identifier, each segment stores a multi-dimensional data vector. The first data column contains the code from the reference cell type used during segmentation (see Figure 1(a)). In our case, the histone modification peaks for H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, and H3K36me3 of the mesendodermal stem cells were used. The subsequent columns contain the overlap of each histone modification for all other cell types (Figure 1(b)). In our case, six columns are created overlapping the histone modification peaks for H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, and H3K36me3 of the mesenchymal stem cells with the previously created segments. Furthermore, the CG (cytosine-guanine dinucleotide) density and the length of each segment are stored in each segment. This

reference-based segmentation is possible, since the mesenchymal stem cells arise from the mesendodermal cells and changes of histone marks indicate variances due to cell differentiation. The segmentation results in 1,419,149 data points.

3 RELATED WORK

Due to size of the available data set, the visualization of chromatin data is mainly based on two approaches: a coarse grained high level analysis and a low level analysis at single gene level. The high level analysis is supported by heatmaps that can visualize the correlations between multiple modifications over the whole genome (used by e.g., Kuang et al. [15]). It is possible to visualize correlations between functional groups of the genome either for multiple modifications at a single point in time or for a single cell type, or for a single modification at multiple points in time or for multiple cell types. The low level analysis can be supported by using a genome browser like WashU [29] or UCSC [14]. A genome browser annotates data tracks linearly to a specific genomic region based on the chromosome positions. With these tracks it is possible to analyze multiple modifications, time points, and cell types within a single visualization but only for a limited region. The comparison between these regions is left to the expert and is tedious. Both visualization techniques do not provide a semi-global trend analysis that supports both the comparison between multiple modifications, and multiple time points or multiple cell types.

Recently, Steiner et al. [23] and Zeckzer et al. [27, 28] presented three new approaches for the analysis of chromatin data supported by interactive visualizations. Moreover, Gerighausen et al. [9] applied a combination of k-means clustering and principal component analysis (PCA) to chromatin data. To our knowledge, these four approaches are the only ones available for intermediate level analysis of chromatin data and the fate of histone modifications.

Previous work related to the visualization of multivariate data in general was reviewed by them, too, including purely visual methods as well as dimensionality reduction and clustering approaches combined with the visualizations of their outcome [10]. The latter methods were tested by us before (unpublished, see also Section 5 for the clustering methods and the clustering result assessments considered) but without conclusive results. These and other unrelated techniques are also described in the recent state-of-the-art report by Liu et al. [16]. The comparison of different scatter plot techniques and dimension reduction methods were extensively reviewed by Sedlmair et al. [22]. They propose the usage of 2D scatter plots for most cases except for artificial or grid data sets for which 3D scatter plots outperform the alternative 2D scatter plot techniques and scatter plot matrices. The data obtained after tiled-binned clustering (and also after tiling and binning alone) falls exactly into the latter category, namely grid data, where—according to Sedlmair et al. [22]—3D scatter plots outperform their 2D counterparts.

4 TASK AND GOALS

4.1 Task

The task of the analyst—biologist or bioinformatician—comprises the analysis of the fate of histone modifications. The underlying data consists of $10^5 - 10^7$ data points (genome segments) with multiple dimensions per cell type (in our case: 6 different histone modifications for mesendodermal cells and mesenchymal stem cells, respectively) and additional dimensions characterizing the segments (here: CG density and segment length). Visualization and interaction facilities as well as the methodology for creating the visualization are developed such that the analyst is supported in creating and verifying hypotheses as well as in gaining additional insights into the changes in epigenetic marks during cell differentiation. Creating hypotheses and gaining additional information are exploratory tasks that are complemented by the verification task.

4.2 Goals

The major goal of this methodology including the visualization and the interaction facilities is supporting the analyst in performing the task. Additionally, visualization as well as interaction facilities are designed following the recommendations from information visualization literature [19,25,26]. The literature provides a set of guidelines any visualization should adhere to (see also Zeckzer et al. [27,28]):

1. The methodology and the visualization should be independent of data specific properties the only assumption being that the number of dimensions is fixed for all data points. This allows for a flexible work flow and flexible visualizations that in principle can also handle data sets from other domains.

2. The approach should use real screen estate efficiently.

3. The approach should provide a good overview of the data.

4. The approach should ease the identification of pattern in the data.

5. The visualizations should be fast to create and fast to interact with.

5 TIBI-CLUSTER—METHODOLOGY

Given tabular data consisting of data points (rows) and attributes (columns), the amount of data values (cells) is the product of the number of data points and the number of attributes. Frequently, the size of the data to be analyzed is too large for manual inspection. Visualization and interaction associated with the visualization can mitigate this problem and support analyses of small, medium, and large data. If the size of the data is too large, however, the amount of pixels on the screen is no longer sufficient to display all data points. To be able to show a summary or all of the data, three principal methods are often applied, among others: using tiled displays (i.e., 2, 3, or more displays), reducing the number of attributes (dimensionality reduction), or clustering the data points.

In the case of ChIP-Seq data preprocessed as described in the biological background (Section 2), clustering [13] failed to provide any useful insights. Testing the divisive clustering methods k-means++ [2, 17], k-median [13], and even consensus clustering [7, 8] in conjunction with clustering assessment strategies and indices like Davies Bouldin Index [6], Hubert Statistics [12], Dunn Index [3], and silhouettes [21] (see also Jain and Dubes [13]) lead to a partitioning of the data with limited expressiveness [9]. Applying the visualization strategies TiBi-SPLOM [28] and TiBi-3D [27] showed why: the characteristics of the data makes the direct application of clustering unsuitable. The data is very noisy. Even applying hierarchical clustering methods would not be beneficial in this context.

As a consequence of clustering methods not being suitable for the original data and the success of the idea of tiled-binning of the data for visualizing the data, a combination of both was conceived. The resulting methodology applies tiling and binning to the original data *before* applying clustering methods. The benefit is that the noise in the data is smoothed and that the inhomogeneity of the data is not a problem any more. Overall it shows, that tiling and binning the original data as a first step benefits both visualization and clustering methods.

The methodology for creating visualizations supporting the analysis of histone modification fate during cell differentiation is based on the following pipeline:

1. Loading the data (Section 6.1).
2. Assigning the individual data points to tiles of bins (Section 6.1).
3. Filtering bins (Section 6.2).
4. Clustering the filtered bins (Section 6.3).
5. Mapping the resulting d-dimensional data to the three dimension of a 3D tiled binned scatter plot (Section 6.4).

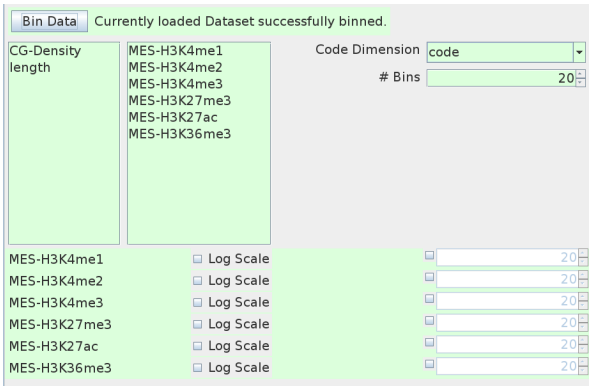


Figure 2: Selecting the dimensions for binning and the number of bins. The two list boxes show the available (left) and the selected (right) dimensions, respectively (middle left part). Clicking on one of the entries moves it to the opposite side, i.e., selected dimensions are deselected and deselected ones are selected. Moreover, the code dimension can be selected using the drop-down box and the number of bins for all selected dimensions can be chosen (middle right). Further, for each of the selected dimensions, the number of bins and whether the dimension should be scaled logarithmically can be selected individually (bottom). After selecting all items, binning can be started by pressing the "Bin Data" button (top). After binning, the background changes from red (changes pending) to green (changes applied), the latter state being shown in this screen shot.

The user interface for setting up and performing each step is described next (Sections 6.1–6.3). It is complemented by the visualization of the tiled binned clustering results (Section 6.4), interaction mechanisms allowing to change the setup of the visualization (Section 6.4), as well as tables showing detailed information about the tiles, the segments of a tile, and the centroids of the final cluster division (Section 6.5). Each step of the workflow and the tables are represented as consecutively ordered tabs within the GUI.

6 TIBI-CLUSTER—USER INTERFACE

All panels provided by TiBi-Cluster that were used to obtain the biological insights reported (Section 7) are also shown and explained in the accompanying video.

6.1 Creating Tiled-Bins of the Data

First, the data is loaded using the menu-item "File – Open". Each data point is a segment. Let n_s be the number of segments and let s_i, $i \in \{1,\ldots,n_s\}$ be a segment. Each segment has assigned n_a attributes a_j, $j \in \{1,\ldots,n_a\}$, one of them being its segmentation code $c(s_i)$ and one of them being its length. Altogether there are $n_s \cdot n_a$ values in the table. Subsequently, each attribute is referred to as a *dimension*.

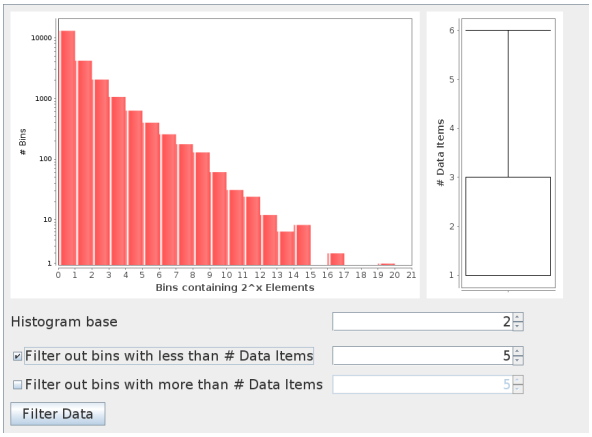


Figure 3: Filtering panel. Top-Left: histogram showing a double-logarithmic plot of the number of elements per bin (x-axis) versus the number of bins having that much elements (y-axis). Top-Right: box plot showing the distribution of elements per bin. Bottom: interaction items for changing the base of the logarithm of the number of elements per bin (x-axis), as well as the lower and the upper boundary of the filter range (number of elements per bin). Pressing the "Filter Data" button applies the filter to the data and the next step can be parameterized and performed.

Next, binning is performed. Figure 2 shows the 'Binning' tab that enables the user selecting the dimensions that should be analyzed. On the right hand side (middle), the dimension representing the segment codes can be chosen. On the left hand side (middle), the dimensions to be analyzed can be selected. In the lower part, the number of bins can be specified either uniformly for all selected dimensions or individually for each selected dimension separately. Additionally, logarithmic scaling can be chosen for each dimension selected. Binning is started by pressing the "Bin Data" button (top).

Binning itself divides the range of values of each dimension a_j into the specified number of intervals. The cross product of the intervals of all attributes gives the set of all possible bins. Then, each segment s_i is assigned to the bin b it belongs to.

Further, each bin is divided into several tiles. Each tile represents one of the n_c possible segment codes. Thus, each segment s_i is assigned to the tile $t_k, k \in \{1, \ldots n_c\}$ of its bin b according to its code $c(s_i) =: k$. In other words, each bin contains n_c tiles and thus can be written as an n_c-dimensional vector $\vec{b} = (t_1, \ldots, t_{n_c})^T$. Each $t_k, k = 1, \ldots, n_c$ is the k-th tile of bin b and represents the number of segments put into that tile.

The binning step can be repeated any time. Then, the subsequent steps have to be repeated, too.

6.2 Filtering the Bins

The second step after binning is filtering. Filtering is performed in the 'Filtering' tab before clustering the data to enable the analyst to remove bins that are not in the focus of the current analysis early. Moreover, filtering before clustering removes "noise" from the data that otherwise might influence the quality of the resulting clustering. The step needs to be under the control of the analyst as domain knowledge is necessary to select the interesting bins and to remove the remaining ones. The removed bins are marked as filtered. Thus, while they are not included in the clustering process, they can still be displayed in the visualization forming their own "filtered bins" cluster. Of course the analyst can choose to keep all bins, perform the subsequent steps, and come back to filtering after having a first idea of the data being analysed.

Selecting the bins to keep and those to remove is supported by two simple and effective visualizations (Figure 3, top). The first one (displayed on the left side) is a histogram with double logarithmic scales showing the size of the bins on the x-axis and the number of bins of the respective size on the y-axis. To improve the understanding of the bin-size distribution, a box plot is shown (on the right side). It gives the median, the lower and upper quartiles as well as the whiskers that represent the 1.5 IQR variability of the distribution of the number of segments per bin. At the bottom, the base of the logarithm used for the number of bin elements (x-axis) can be chosen. Moreover, the minimum and the maximum of segments per bin can be chosen. Bins with less than the minimum and more than the maximum of segments per bin are removed. The same can be achieved by graphically selecting an interval in the histogram. However, the spinners allow to provide exact values for the borders of the selected range.

Filtering is performed by pressing the "Filter Data" button. The filtering range can be changed and applied any time. Then, the clustering step has to be repeated, too.

6.3 Clustering the Filtered Bins

The most complex step in the methodology—algorithmically as well as with respect to the computational resources needed—is the clustering of the filtered bins. This is reflected by the interface in the 'Clustering' tab for setting the clustering parameters (Figure 4, description top-down). First the analyst can decide whether to use a single clustering method or to use consensus clustering based on the selected single clustering method. Next, the analyst chooses the parameters for the single clustering method. Currently, k-means and k-median can be selected as single clustering method. For both, the empty-cluster strategy—the strategy to apply if an empty cluster is produced during the process—determines the outcome of the clustering. The available empty-cluster strategies are to split the cluster with the largest variance or with the largest number of data points. Alternatively, a new

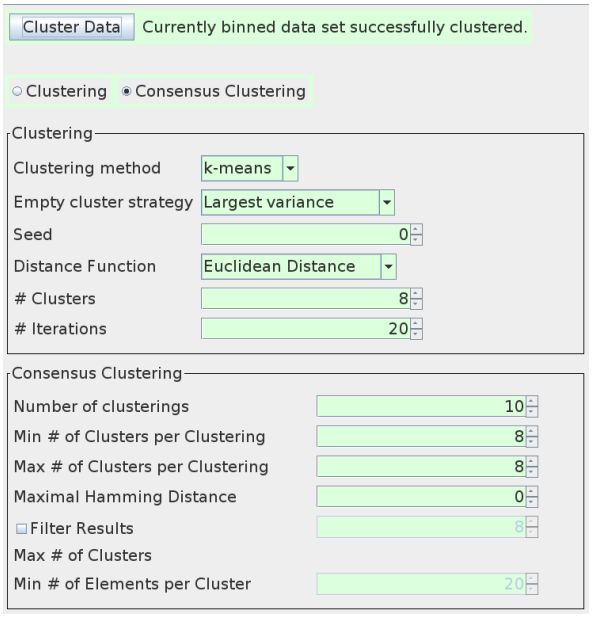


Figure 4: Interface for selecting the clustering method and its associated parameters. The clustering is started by pressing the "Cluster Data" button. Green backgrounds represent unchanged parameters, while red backgrounds represent parameters that changed. If the current parameter setting was used for the current clustering, button and label background are green (as shown here), and red otherwise.

centroid can also be determined by locating the farthest data point from the centroid of the empty cluster and then assigning it as a new centroid. As both clustering methods are based on random selection of the first point, the seed for the random number generator can be selected, This allows to obtain different results while all results can be reproduced.

Next, a distance function has to be selected. As defined in Section 6.1 we use vectors that represent the tiled and binned data for the clustering. The available functions for computing the distance d between two such vectors $\vec{b_1}$ and $\vec{b_2}$ are listed in Table 1. In addition, the value p can be specified when using the "LP-Norm".

K-means as well as k-median clustering require the number of clusters k to be specified before clustering. Moreover, it is useful to provide a maximum number of iterations.

If consensus clustering is selected, further parameters can be chosen by the analyst. Consensus clustering is based on the repeated clustering of a data set followed by computing the final cluster assignment based on a consensus function. The analyst can thus choose the number of times clustering is performed as well as the minimum and the maximum number of clusters created during each of the individual clusterings. The consensus function is based on computing the Hamming distance between the assignments of data points to clusters. The analyst can choose, what the maximum hamming distance for joining two clusters is. Finally, the consensus clustering results can be filtered according to two criteria. First of all, the maximum number of clusters generated by the consensus clustering can be specified. Moreover, selecting the minimum number of elements per cluster allows to remove outliers that are assigned to different clusters each time clustering is performed compared to the other members of the cluster.

Table 1: The different distance measures avaiable for clustering.

Name	$d(\vec{b_1},\vec{b_2}) =$
Angular distance	$cos^{-1}\left(\frac{\vec{b_1}\circ\vec{b_2}}{\lvert\vec{b_1}\rvert\cdot\lvert\vec{b_2}\rvert}\right)$
Euclidean distance	$\sqrt{\sum_{i=0}^{n}\left(\vec{b_{1,i}}-\vec{b_{2,i}}\right)^2}$
Manhattan distance	$\sum_{i=0}^{n}\left\lvert\vec{b_{1,i}}-\vec{b_{2,i}}\right\rvert$
LP-Norm	$\sqrt[p]{\sum_{i=0}^{n}\left\lvert\vec{b_{1,i}}-\vec{b_{2,i}}\right\rvert^p}$
Mahalanobis distance	$\sqrt{\left(\vec{b_1}-\vec{b_2}\right)^T S^{-1}\left(\vec{b_1}-\vec{b_2}\right)}$
Bin size difference	$\left\lvert\sum_{i=0}^{n}\vec{b_{1,i}}-\sum_{i=0}^{n}\vec{b_{2,i}}\right\rvert$

After selecting the clustering strategy and setting all parameters to the desired values, the analyst starts the clustering of the filtered bins by pressing the "Cluster Data" button. Upon termination, the background changes its color to green. Should any of the parameters be changed after clustering, its respective background color is set to red. At the same time, the background of the button for starting the clustering and the background of the message associated to the button are changed to red, too. The clustering parameters can be changed and applied any time.

The result of the clustering step is a set of clusters representing those bins whose tiles have a similar distribution with respect to the number of segments they contain. Thus, bins are not necessarily similar if they are close to each other with respect to their attribute values. However, if they are, this implies that bins that are close together share a similar distribution. The biological interpretation of clustering results and the biological insights gained are presented in Section 7.

6.4 Mapping the Tiled Binned Clusters to 3D Tiled Binned Scatter Plots

The visualization is based on the tiled binned 3D scatter plots introduced by Zeckzer et al. [27]. However, the mapping of information to visual elements is different. We redesignate the meaning of the spheres in the original 3d scatter plot and assign each cluster to one

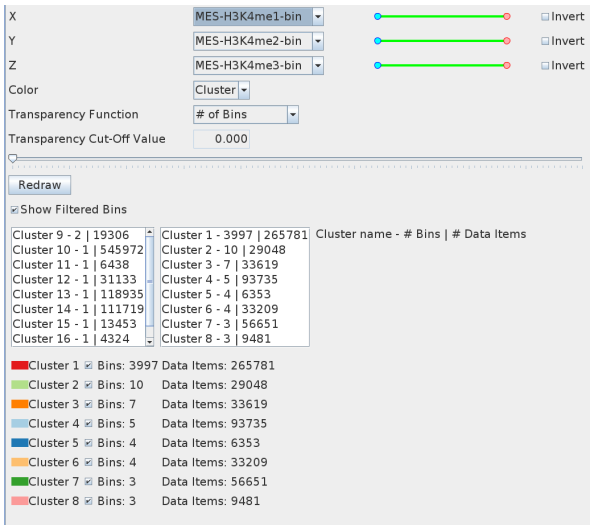


Figure 5: Mapping panel. Top: The analyst can choose which binning dimension is mapped onto *x*-, *y*-, and *z*-axis, respectively. Cutting planes and inverted cutting planes can be used to restrict the information shown and such eases the analysis of the data focused upon. Either the number of data items or the number of bins per cluster is mapped onto the transparency of each sphere. Bottom: The analyst selects up to eight clusters for the current investigation. Information about the selected clusters is displayed and each selected cluster can be disabled. Each cluster for each bin is represented by a sphere. Each cluster number is mapped onto color and position of the sphere.

sphere in our adopted visualization. Thus, eight clusters can be analyzed in parallel as shown in Figure 6. Those can be selected from the left list box in the lower part of the 'Mapping' tab (Figure 5) by clicking on an entry. Those clusters selected for display are shown in the right list box, which additionally can be used for deselecting clusters by clicking on the respective entry. Each cluster is assigned to a tile and thus will be represented by a sphere having the position corresponding to that tile in each bin, as well as the color corresponding to that tile. Cluster color, cluster number, and number of bins as well as number of data items per cluster are listed below those two list boxes. Moreover, a check box allows selecting and deselecting each cluster for display. If the analyst wishes, the filtered bins can be included into the display by marking the check box above the lists. Each filtered bin is represented by a *gray* sphere centered inside the area of its bin.

The number of bins per cluster or the number of data items per cluster can be mapped to the transparency of the spheres. Additionally, a transparency cut-off value can be selected showing only spheres having a higher transparency than the cut-off value selected.

The dimensions that are mapped onto the *x*-, *y*-, and *z*-axes are selected by the analyst using the respective drop-down boxes at the top of the interface. Moreover, cutting planes and inverted cutting planes can be activated by the sliders and the checkboxes to the right of the axes selection, respectively.

6.5 Tile Table, Segment Table, and Centroid Table

The 'Tile Table' tab contains an entry for each tile shown, i.e., for each sphere in the 3D scatter plot. Besides the cluster number, the number of segments of the cluster having a specific modification or a specific code are shown. The code was created by the segmentation process, whereas the modifications were chosen by the analyst while selecting the modifications used for binning (Section 6.1).

Selecting a sphere or selecting a row in the tile table results in showing all segments of that tile in the 'Segment Table' tab (linked views). Here, for each segment its long ID, its short ID as well as modification coverage or other information from the additional data selected during segmentation are shown. Moreover, CG densities computed during segmentation are displayed. Finally, the length of each segment is provided.

Selecting a row in the tile table highlights the respective sphere in the 3D scatter plot. Selecting a sphere in the 3D scatter plot leads to showing only the related tile entries in the tile table. Thus, tile table and 3D scatter plot are linked, too.

The 'Centroid Table' tab provides information about the centroids of each cluster. Hereby, each row represents one modification selected for binning, whereas each column represents a cluster. The centroid of each cluster is then given by the entries in the cells connecting the cluster to the respective modifications.

6.6 Visual Control Tab

It is possible to change the *point of view* to any angle. For exploring the visualization with any degree of detail, it is possible to *zoom* in and out. Thus, it is possible to navigate in any direction required to obtain the best views on the data. Additionally, an *auto-rotation* of the plot eases getting an overview of the data set facilitating the choice of suitable viewpoints. Any perspective can be *saved* and *loaded*, which supports comparing different data sets using the same perspective without effort. Most importantly, snapshots of the visualization can be saved to file.

7 BIOLOGICAL INSIGHTS

We demonstrate the functionality and benefits of TiBiCluster using a data set with 2 cell types and 6 histone modifications per cell type. It is compiled as described in Steiner et al. [23] but extends the code calculation to 6 dimensions yielding $2^6 = 64$ codes ranging from

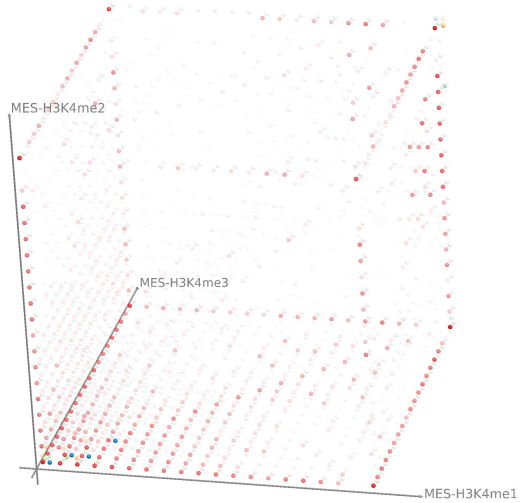
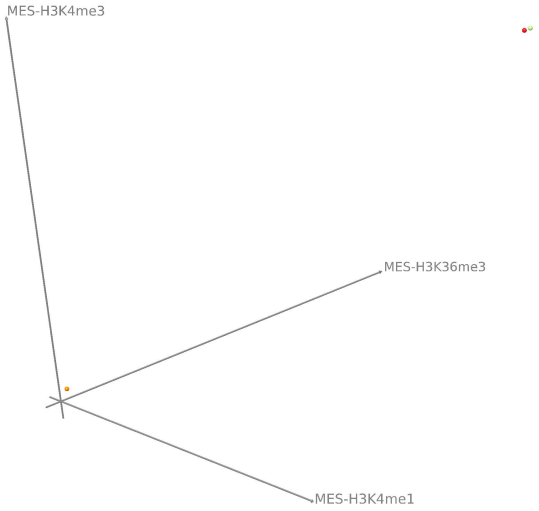


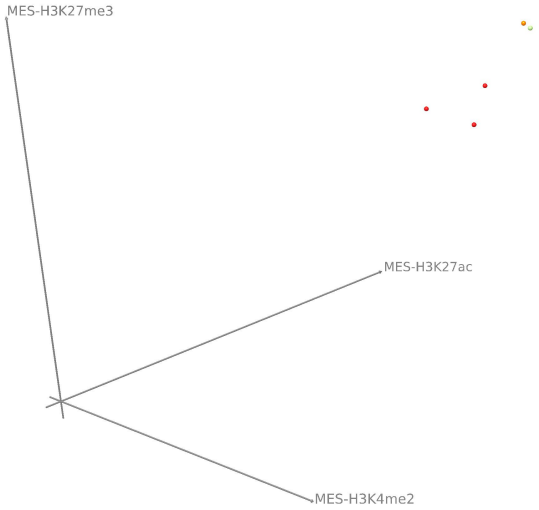
Figure 6: Results obtained after binning (Figure 2), filtering (Figure 3), and clustering (Figure 4). The parameter settings for the visual mapping are shown in Figure 5.

0 to 63. The reference cell type is the mesendodermal cell line. Data from a mesenchymal stem cell line is compared to it. Mesendodermal cells arise early in the embryogenesis, i.e., the development of an embryo out of a fertilized egg. They differentiate into mesodermal and endodermal cells, and finally into various organs. Mesenchymal stem cells differentiate out of successors of the mesendodermal cells. However, several differentiation steps lie in-between them.

The mesenchymal stem cell data is clustered using TiBi-Cluster. We choose the six modifications as the six dimensions for further analysis (Figure 2). Each dimension is binned with 20 bins. Most bins contain only a few elements, while only a few contain many (Figure 3). Bins with less than 5 segments are filtered out to focus more on coordinated changes. The data is clustered using consensus clustering (Figure 4). Ten instances of k-means clustering with euclidean distance and 8 clusters each is performed. Bins with Hamming distance 0 between the 10 cluster assignments obtained are summarized to one cluster. No clusters are filtered. The clustering results in 17 clusters. The parameters of the original mapping and selected clusters (Figure 5) yield a visualization showing the first eight clusters (Figure 6). Cluster 1 (red spheres) contains bins distributed over the whole hyper cube, i.e., segments with any epigenetic state in mesenchymal cells. All these segments are unmodified in mesendodermal cells and are located in bins containing only a few modification. These segments are thus genomic regions which become newly modified in mesenchymal cells but carry an unusual epigenetic state. Only few other clusters exist near the H3K4me1 – H3K4me3 plane (bottom: blue, green, and yellow spheres) and near



(a) Dimensions: H3K4me1, H3K4me3, H3K36me3



(b) Dimensions: H3K27ac, H3K27me3, and H3K4me2

Figure 7: 3D tiled bin scatter plots for cluster 8 (red spheres), cluster 14 (light green spheres), and cluster 17 (orange spheres).

the H3K4me1 – H3K4me3 – H3K4me2 corner (upper right: desaturated spheres).

Changing the selection of clusters displayed, the analyst finds clusters 8, 12, 13, 14, and 17 particularly interesting.

Two clusters, cluster 8 and cluster 14 are genes marked with all modifications in both cell types (see Figure 7). In total, about 120,000 segments belong to the two clusters. Please note, that this state, i.e., all marks studied, is theoretically possible but very unlikely. Mono-, di-, and trimethylation of H3K4 may only be observed in a single cell at the same position if distributed to the four copies of H3K4 in the histone complexes attached to the two alleles. Acetylation and trimethylation of H3K27 have an contrary effect, namely, formation of open and

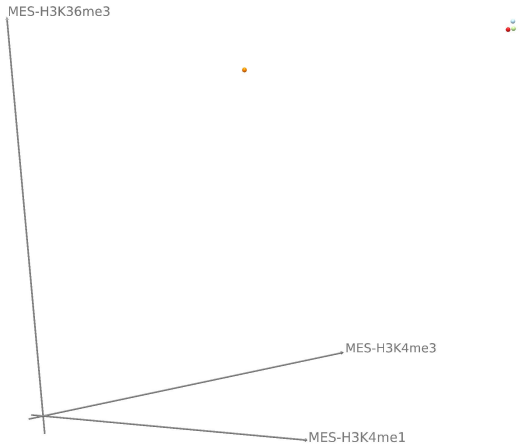


Figure 8: 3D tiled bin scatter plot for cluster 4 (red spheres), cluster 6 (light green spheres), cluster 12 (orange spheres), and cluster 13 (light blue spheres). Dimensions: H3K4me1, H3K4me3, H3K36me3

closed chromatin. They therefore do not occur at the same histone complex but only at the two different histone complexes attached to the two alleles. More likely, this is a mixed state in the population of cells. This indicates that the epigenetic state of these genomic regions is either not important for the cells identity and function or undergoes changes due to ongoing differentiation.

Segments in cluster 17 (almost 15,000) are in a mixed, undetermined state in mesendodermal cells, i.e., all or almost all marks. However, in mesenchymal stem cells, only three marks are present: H3K4me2, H3K27me3, and H3K27ac (see Figure 7) Again, it is unlikely that methylation and acetylation occur at the same histone complex due to their antagonistic effects. The combination of H3K4me2 and H3K27ac was shown to mark transcription factor binding sites and to recruit cell type specific transcription factors to them [24]. Since the third mark, H3K27me3, likely does not occur at the same allele or cell, the observed combination may indicate either an epigenome-driven recruitment of transcription factors in some cells or repression of the same loci in other cells. Alternatively, the recruitment is specific for one allele while the other allele is silenced by H3K27me3. Since ChIP-seq does not allow single cell measurements nor simultaneous measurement of combination of marks, we cannot distinguish between those alternatives.

In several clusters, we observe the combination of H3K4me1 and H3K36me3 (see Figure 8). These are cluster 12, cluster 13, as well as cluster 4 in combination with H3K4me3 and cluster 6 in combination with H3K4me3 and H3K27me3. Based on the current knowledge, this combination is unexpected. H3K4me1 is supposed to localize to promoters and represses transcription. H3K36me3 is associated with active splicing and thus with transcription and localized to splice sites. Even though one would expect those marks at different genomic loci and not in combination, we observe a strong co-occurrence of both marks.
We conclude therefore that their function is more diverse than known so far.

8 CONCLUSION

We introduced TiBi-Cluster, a methodology that combines tiling and binning with clustering. The results are then shown in tiled binned 3D scatter plots introduced before and adapted to the current methodology. Altogether, the methodology and the tool supporting the methodology allow analyzing the fate of chromatin modifications during cell differentiation. Our method proved to be vastly beneficial while analyzing and comparing six histone modifications in two cell lines. New, unknown relations were uncovered using TiBi-Cluster.

9 ACKNOWLEDGMENTS

We thank all the unknown reviewers of previous versions of our paper for their valuable comments. This work was partially funded by the German Federal Ministry of Education and Research (BMBF) within the project Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig (BMBF grant 01IS14014B).

10 REFERENCES

[1] A set of tools (in Java) for working with next generation sequencing data in the BAM (`http://samtools.sourceforge.net`) format. `http://broadinstitute.github.io/picard/`.

[2] D. Arthur and S. Vassilvitskii. k-means++: The Advantages of Careful Seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.

[3] J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):301–315, Jun 1998. doi: 10.1109/3477.678624

[4] J. Cheng, R. Blum, C. Bowman, D. Hu, A. Shilatifard, S. Shen, and B. Dynlacht. A Role for H3K4 Monomethylation in Gene Repression and Partitioning of Chromatin Readers. *Molecular Cell*, 53(6):979–992, 2014. doi: 10.1016/j.molcel.2014.02.032

[5] P. Cheung and P. Lau. Epigenetic regulation by histone methylation and histone variants. *Mol Endocrinol*, 19(3):563–73, Mar 2005.

[6] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979. doi: 10.1109/TPAMI.1979.4766909

[7] A. L. N. Fred. *Finding Consistent Clusters in Data Partitions*, pp. 309–318. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. doi: 10.1007/3-540-48219-9_31

[8] A. L. N. Fred and A. K. Jain. Data clustering using evidence accumulation. In *Object recognition supported by user interaction for service robots*, vol. 4, pp. 276–280, 2002. doi: 10.1109/ICPR.2002.1047450

[9] D. Gerighausen, D. Zeckzer, L. Müller, and S. J. Prohaska. ChromatinVis: a tool for analyzing epigenetic data, 2014. Poster presented at 2nd EMBO Conf. on Visualizing Biological Data, Heidelberg, Germany.

[10] G. Grinstein, M. Trutschl, and U. Cvek. High-Dimensional Visualizations. In *Proceedings of the VII Data Mining Conference KDD Workshop 2001*, pp. 7–19. ACM Press, New York, San Francisco-CA, USA, 2001.

[11] S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology*, 5(9):e1000502, Sep 2009. doi: 10.1371/journal.pcbi.1000502

[12] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075

[13] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[14] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[15] Z. Kuang, L. Cai, X. Zhang, H. Ji, B. P. Tu, and J. D. Boeke. High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. *Nature structural & molecular biology*, 21(10):854–863, 2014.

[16] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. In *Proc. Eurographics Conf. Visualization*, pp. 20151115–127, 2015.

[17] J. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297. University of California Press, 1967.

[18] L. Müller, D. Gerighausen, M. Farman, and D. Zeckzer. Sierra Platinum: A Fast and Robust Multiple-Replicate Peak Caller With Visual Quality-Control and -Steering. *BMC Bioinformatics*, 17(1):1–13, 2016. doi: 10.1186/s12859-016-1248-6

[19] T. Munzner. *Visualization Analysis and Design: Principles, Techniques, and Practice*. A K Peters Visualization Series, 2014.

[20] Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, Feb 2015. doi: 10.1038/nature14248

[21] P. Rousseeuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.*, 20(1):53–65, nov 1987. doi: 10.1016/0377-0427(87)90125-7

[22] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE transactions on visualization and computer graphics*, 19(12):2634–2643, 2013.

[23] L. Steiner, L. Hopp, H. Wirth, J. Galle, H. Binder, S. J. Prohaska, and T. Rohlf. A Global Genome Segmentation Method for Exploration of Epigenetic Patterns. *PLOS ONE*, 7(10):e46811, 2012.

[24] Y. Wang, X. Li, and H. Hu. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics*, 103(2-3):222–228, 2014. doi: 10.1016/j.ygeno.2014.02.002

[25] M. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A K Peters, 2010.

[26] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 ed., 2004.

[27] D. Zeckzer, D. Gerighausen, and L. Müller. Analyzing Histone Modifications in iPS Cells Using Tiled Binned 3D Scatter Plots. In *2016 Big Data Visual Analytics (BDVA)*, pp. 1–8, Nov 2016. doi: 10.1109/BDVA.2016.7787042

[28] D. Zeckzer, D. Gerighausen, L. Steiner, and S. J. Prohaska. Analyzing Chromatin Using Tiled Binned Scatterplot Matrices. *2014 IEEE Symposium on Biological Data Visualization (BioVis)*, abs/1407.2084, 2014.

[29] X. Zhou, R. F. Lowdon, D. Li, H. A. Lawson, P. A. Madden, J. F. Costello, and T. Wang. Exploring long-range genome interactions using the WashU Epigenome Browser. *Nature methods*, 10(5):375–376, 2013.