Chapter 8

Interpreting and Visualizing ChIP-seq Data with the seqMINER Software

Tao Ye, Sarina Ravens, Arnaud R. Krebs, and Làszlò Tora

Abstract

Chromatin immunoprecipitation coupled high-throughput sequencing (ChIP-seq) is a common method to study in vivo protein–DNA interactions at the genome-wide level. The processing, analysis, and biological interpretation of gigabyte datasets, generated by several ChIP-seq runs, is a challenging task for biologists. The seqMINER platform has been designed to handle, compare, and visualize different sequencing datasets in a user-friendly way. Different analysis methods are applied to understand common and specific binding patterns of single or multiple datasets to answer complex biological questions. Here, we give a detailed protocol about the different analysis modules implemented in the recent version of seqMINER.

Key words Protein–DNA interactions, Chromatin immunoprecipitation coupled high-throughput sequencing (ChIP-seq), seqMINER, Genome-wide, Visualization, Multiple datasets, k-means clustering

1 Introduction

Chromatin immunoprecipitation (ChIP) is a powerful technique for mapping of protein–DNA interactions inside a cell [1]. In combination with high-throughput sequencing (ChIP-seq), the localization of posttranslationally modified histone proteins, histone variants, transcription factors, or histone modifying enzymes can be determined at a genome-wide scale [2, 3]. The method is based on formaldehyde cross-linking of protein–DNA complexes in living cells, following cell lysis and shearing of DNA into 200–700 base pair (bp) fragments. In case the protein of interest (POI) is very stably associated with the DNA, such as histone proteins, the cross-linking step can be skipped. Next, an antibody targeting the POI will pull down the actual DNA binding sites. After reverse cross-linking, the ChIP-ed DNA fragments are purified and quantitatively analyzed by high-throughput sequencing. For further bioinformatics analysis of all experiments, it is important to include a control sample. This can

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_8,

be either DNA, treated like the immunoprecipitated DNA (Input), or "mock" ChIP-ed DNA, using a nonspecific antibody for the ChIP (i.e., IgG control).

Each ChIP-seq run leads to millions of short reads (tags) and it is challenging to process, analyze, and interpret the large amount of data. The first part of high-throughput sequencing analysis uses common processing pipelines, which involves the alignment of raw reads to the genome, data normalization, and identification of enriched signal regions (Peak calling) [4]. In the second stage, individual programs allow detailed analysis, biological interpretation, and visualization of ChIP-seq results.

To provide a more complex picture of biological processes in a cell, many studies aim to compare different datasets obtained by ChIP-seq. Since the protein–DNA interactions studied by ChIP represent only POI binding at the moment of cross-linking, studies apply cellular differentiation models to compare different factors or chromatin modifications in a system with dynamic transcriptional changes. All these require analytical and computational modeling techniques, which compare multiple sequencing datasets in one cell type or the differential binding of factors in various cell types.

seqMINER has been designed to analyze multiple, or single, ChIP-seq datasets of different factors like transcription factors, chromatin-modifying enzymes, or histone modifications [5]. It is a user-friendly software, which can be used to analyze specific and common binding patterns of different factors. In addition, seqMINER helps to understand differential binding patterns of one factor in more than one cell type. Here, we provide a detailed protocol for the usage of the seqMINER platform.

1.1 Overview of Software Tools Required for the Generation of Appropriate seqMINER Input Files

To successfully run the seqMINER software, we will first give a small overview about available software tools required for the generation of appropriate seqMINER input files. The mapping of short reads against the reference genome is typically the first step to analyze the obtained ChIP-seq data. The most popular software tools are BOWTIE [6] and BWA [7]. For the manipulation and storage of read alignments or the generation of SAM/BAM formats, SAMtools [8] are recommended. To identify high-confidence binding sites of a ChIP sample, peak calling algorithms calculating the enriched tag densities over the background noise are applied. The peak calling methods normally include the normalization between the ChIP and control samples. Since the identification of signal peaks is a central task in interpreting ChIP-seq results, many peak calling algorithms have been established. The most common methods are MACS [9], SICER [10], or FindPeaks 4.0 [11].

The seqMINER software applies different analysis methods to highlight general as well as specific patterns in a given dataset.

All methods require a set of reference coordinates, which can be either a list of ChIP-seq enrichment clusters (peaks) of a particular factor. Alternatively, transcription start sites (TSSs) of genes, transcription termination sites (TTSs), or whole gene coordinates can be used. This might be of interest to analyze, for instance, the binding patterns of RNA polymerase II (Pol II) at the start or the end of genes. For the data collection, all middle points of the reference coordinates are calculated and the read densities of multiple aligned read dataset are collected in a defined window around the reference coordinates. The signal enrichment status of these multiple tracks can be analyzed through two different algorithms. In the Density Array Method, the created matrix of tag densities around the reference coordinates is reorganized by k-means clustering. This will create different groups with similar genomic features, which can be visualized and further analyzed in a heatmap graphical interphase. However, this clustering method does not allow the comparison of quantitative changes between multiple datasets. The enrichment-based method allows the one to one comparison of enrichment values between two different datasets. In this case the enrichment values are presented in a scatter plot and a table, which can be exported for further analysis. Besides, additional options have been integrated into the platform to allow individual analysis of datasets or clustering results. Since these options have been described previously, the newest seqMINER release implements an Annotation system.

2 Materials

2.1 Operating System and Software Requirements

The seqMINER software is suitable for any operating system which has a Java Runtime Environment (version 6 or above) installed, such as Linux, OS X (≥10.5), or Microsoft Windows.

For 4–5 datasets of ten million reads, seqMINER can be used on a local computer with a 32-bit operating system having 2–4GB random access memory (RAM) or on a 64-bit operating system, which will have almost no limit for memory usage.

Still it is recommended to run seqMINER over a server since it equipped normally with more RAM. This will decrease the analysis time and allow increasing the amount of data to be analyzed. In this case, a local X-windows service should be installed (e.g., Xming for Windows).

2.2 Installation

Users should first check or download the Java Runtime Environment from http://www.java.com/. The last version of seqMINER can be downloaded from http://sourceforge.net/projects/seqminer/. More information is available at http://bips.u-strasbg.fr/seqminer/, which is under General Public License (GPL3). The downloaded file needs to be unzipped.

To launch seqMINER under windows, double-click the seqMINER.bat file (*see* **Note 1**).

To launch seqMINER by a terminal:

Go to the seqMINER folder

>cd seqMINER_folder

Launch the application

>java -Xmx1500m –jar seqMINER.jar

2.3 Files and Formats

seqMINER supports BED, SAM/BAM, or default Bowtie output file formats as input files (for file format information: https://genome.ucsc.edu/FAQ/FAQformat.html). Two different types of input files are required: (1) a reference coordinate file and (2) aligned read files. The reference coordinate file can be generated with any peak detection algorithm, e.g., MACS [9], SICER [10], or FindPeaks 4.0 [11]. It is recommended to directly use the summits file of the peak detection analysis, which represents the summit points of identified peaks. On the other hand, it is possible to take RefSeq or Ensembl transcription start sites or whole RefSeq or Ensembl genes as reference coordinates. These files can be extracted from annotation databases or downloaded from http://sourceforge.net/projects/seqminer/files/Reference%20coordinate/. The second input files are the aligned raw read files of all datasets, which can be analyzed through different methods described in the following section.

3 Method

seqMINER is designed to carry out basic ChIP-seq data analysis in an easy and fast way to answer biological questions. The platform allows the comparison between a reference set of genomic positions (reference coordinate file) and multiple ChIP-seq datasets (aligned raw read file).

3.1 Standard Analysis

The standard analysis is separated in three steps, which can be followed at the java interface from the left to the right. A screenshot of the java interface with already uploaded datasets is presented in Fig. 1.

3.1.1 Step 1: File Selection

Over the browser button or under “File,” the files are chosen.

1. One reference coordinate file is required. As an example, the refGene mm9 seqMINER files is taken (Fig. 1).
2. Several aligned read files can be selected. Here we used already published datasets for the histone marks H3K27me3 (GSM307619) and H3K36me3 (GSM307619).

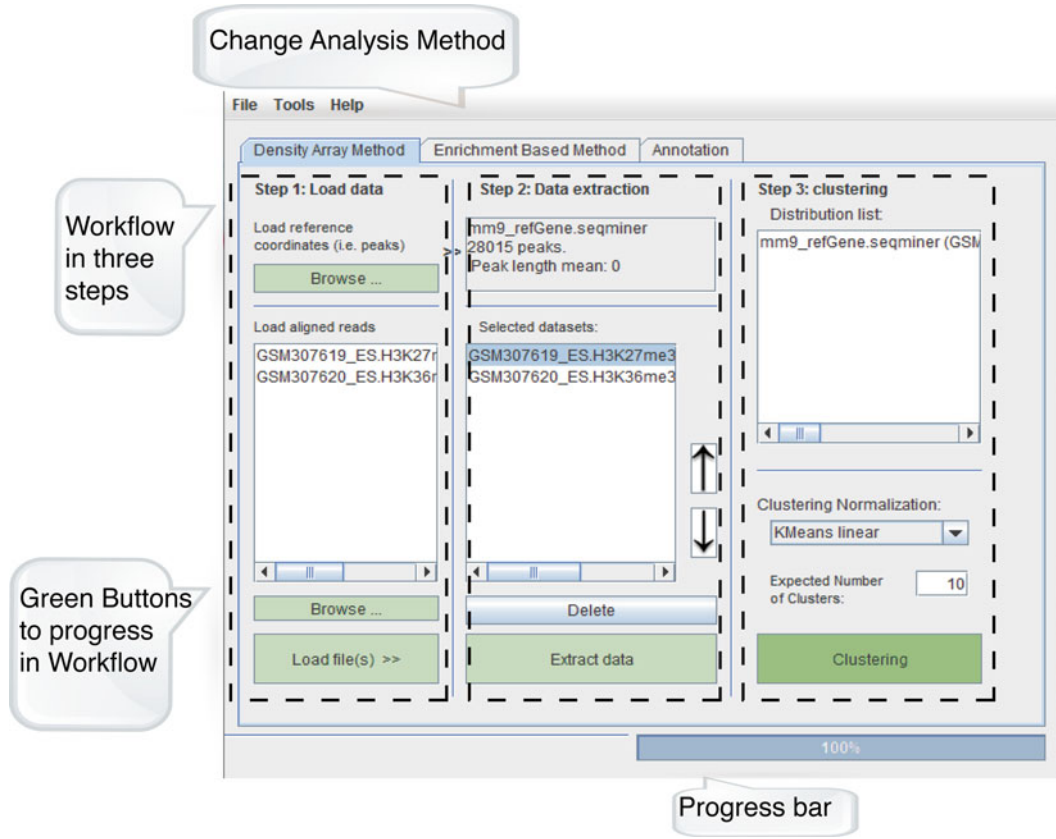

Fig. 1 seqMINER interface with analysis methods and workflow. The *green* buttons are required for data upload and progressing in the workflow from the *left* to the *right*. In the *upper* part the analysis methods can be changed to Density Array Method, Enrichment-Based Method, or Annotation. As reference coordinate, the mm9_refGene.seqminer file with 28015 peaks is uploaded. Aligned read files are taken from the Gene Expression Omnibus database (GSM307619, GSM307620)

All required files can be loaded simultaneously or one by one through the "Load file(s)" button (*see* **Note 2**).

3.1.2 Step 2: Extract Data

After loading all files, it is possible to change their order, which will be later presented in the heatmap, through the flash buttons. In addition, the "Delete" button erases already loaded datasets. At this step optional parameters for the different methods should be defined (*see* Subheadings 3.3 and 3.4). Finally, the green "Extract data" button extracts the tag densities of the datasets according to the reference coordinates for further analysis (*see* **Note 3**).

3.1.3 Step 3: Clustering and Data Visualization

The 3rd step varies depending on the method to be used to analyze the signal enrichment in multiple tracks. They can be chosen in the upper part of the java interface (Fig. 1). The methods are discussed in the following sections.

3.2 Density Array Method

This method collects the read densities over a window around a reference coordinate. The identification of groups with similar features is conducted through k-means clustering.

seqMINER proposes two normalization methods (linear and ranked normalization), which are directly applied in the clustering step. Of note, the non-normalized data is represented in the final visualization step. The k-means clustering method of interest can be chosen under clustering normalization. The default algorithm is set to k-means raw (*see* **Note 4** below). It is recommended to use k-means raw for a single datasets and k-means linear or ranked for multiple datasets. The number of expected clusters is 10 by default. It is possible to define a higher or lower value.

By clicking the "Clustering" button, the analysis is started (*see* **Note 5** below). Visualization of the results is achieved through a heatmap (Fig. 2), which will automatically appear in a separated interface (*see* **Note 6**). Based on their biological meaning, the generated clusters can be reorganized. In Fig. 2 for this the given cluster needs to be selected and shifted with the flash buttons.

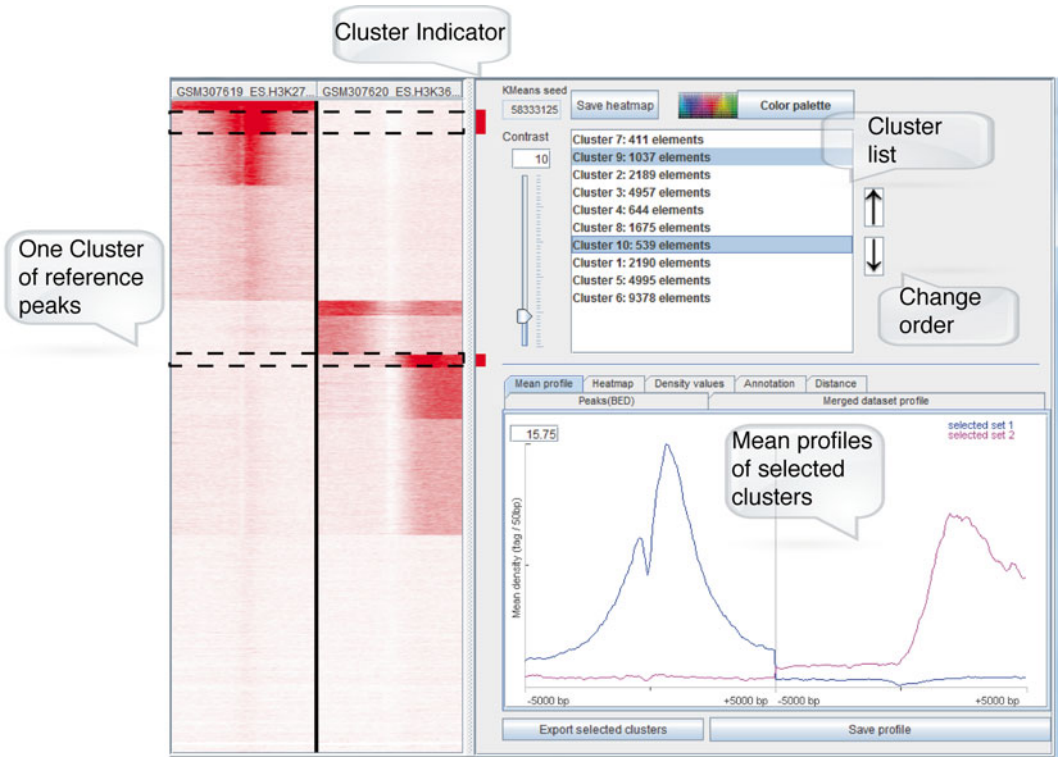

Fig. 2 Visualization of data after clustering with the Density Array Method. The *left* site represents a heatmap with different clusters of reference peaks. These clusters are listed on the *right*, whereas the order can be changed according to the biological relevance. In addition, the mean profiles of the selected clusters (two) are shown on the lower *right panel*

By using the shift or control (Ctrl) key, many clusters can be selected at the same time. The generated heatmap can be exported with "Save heatmap" as a png file.

It is possible to investigate each generated cluster or functional group. Under "Peaks" (bed) the reference coordinated for the selected cluster is found and can be saved with export selected cluster as a bedfile. Following the lower right panel in the Cluster Heatmap interface, a merged dataset profile of all selected clusters is depicted. This average profile represents the mean tag density for one, or multiple, selected cluster over the defined analysis window around the peaks. Mean profiles compare distributions around the peak middle points of the selected cluster(s). Different clusters will be presented in different curves. In Fig. 2, an example of the H3K27me3 and H3K36me3 mean profiles is depicted. H3K27me3 has a peak around the promoter, whereas H3K36me3 localizes more downstream of the promoters. In addition, the average profile can be shown as a heatmap. The heatmap represents the raw tag densities in the defined window around the reference coordinates of the selected cluster(s). "Save profile" will export the given graph as a png file. Additionally, the density values used for the generation of profiles are presented in a table (*see* **Note 7**). After copying these values into an Excel or a text file, they can be used for further analysis or plotting. The order of the density values corresponds with the clustered Peak (bed) coordinates (*see* **Note 8**).

A ChIP-seq peak annotation system is implemented in the seqMINER platform. However, it is necessary to select a genome assembly under Annotation panel (*see* **Note 9** below). For each reference coordinate, the closest gene will be annotated. Moreover, the distance of peaks to the closest TSS is represented in a bar chart under "Distance." The window size and bin size are configurable.

3.3 Enrichment-Based Method

This method allows the one to one comparison of datasets in a quantitative way. It calculates the total number of reads for each dataset at the reference coordinates. By default the analysis window is set to the peak interval. It can also be defined as a fixed interval around the calculated middle points of the reference coordinates. It is optional to load a control dataset (*see* **Note 10**). The analysis is launched with the "Calculate density array" command and a Dot Plot interface will appear. The files are compared in a Scatter Plot. The data presented at the x and y axis, as well as the coordinates can be modified manually. In the right panel, a table with all the calculated values is generated. The scatter plot and the table could be exported with "Save Image" or "Export table" commands for further analyses.

3.4 General Options

The default parameters for the peak extensions, read and clustering options are found in Tools with the shortcut Alt –O or under Tool → Options → General.

1. Peak extension: The read densities are collected around each calculated middle point of the reference coordinates, peak summits, or TSSs in a defined analysis window. The default value is 5,000 base pairs up and downstream from the middle point.
2. Read options: If there is strand information in the reference peak file, the reverse strand reference genes are automatically turned to forward strands. After calculating the peak middle point and prior to analysis, all reads are extended to 200 bp by default. In the read options, "Enable reads extensions" can be inactivated or the size of extendable reads can be modified.
3. Clustering options: These options are applied in the different algorithms and normalization procedures applied by the Density Array Method. The Wiggle step defines the clustering resolution. By default the Wiggle step will generate, for example, 10,000(bp)/50(step length) = 200 values per reference coordinate. In case there is not enough memory, the "Wiggle step" can be increased. An increase in the "Wiggle step" to, for example, 100 will result in half of the memory consuming of the clustering step. The "Max runs" indicates the numbers of clustering steps in the algorithms. The "Percentile threshold" is used in the k-means linear normalization method. All intensity values are divided by the percentile of the threshold. The "Percentile threshold" is by default 75 %, which is the third quartile of the distribution. The background values which are smaller than the T threshold will be excluded from the previous distribution. The T threshold is also applied in the ranked-based normalization method. All intensity values are sorted in ascending order. The T threshold defines the rank of the sorted intensity values, which will then be replaced by 0. Thus, all background values are considered as equal during the clustering process.

3.5 Gene Profile Option

seqMINER conducts the Density Array Method with the reference gene body. The gene profile options are found under Tools → Options → Gene profile. It is important to activate these setting before Subheading 3.1.2 (Extract data) (*see* **Note 11**).

As an example dataset, we used a Pol II dataset (GSM307823), whereas the seqMINER refGene mm9 was uploaded as a reference coordinate in Subheading 3.1.1. Usually, the reference file should contain the gene start and end coordinates. The annotation data can be downloaded from annotation databases or http://sourceforge.net/projects/seqminer/files/Reference%20coordinate/. By default each gene (reference gene body) is equally divided in 160 bins, whereas 20 extra bins are added to the upstream and downstream regions. These extra bins are by default 5,000 bp long, which can be modified through peak extensions. Then, the read

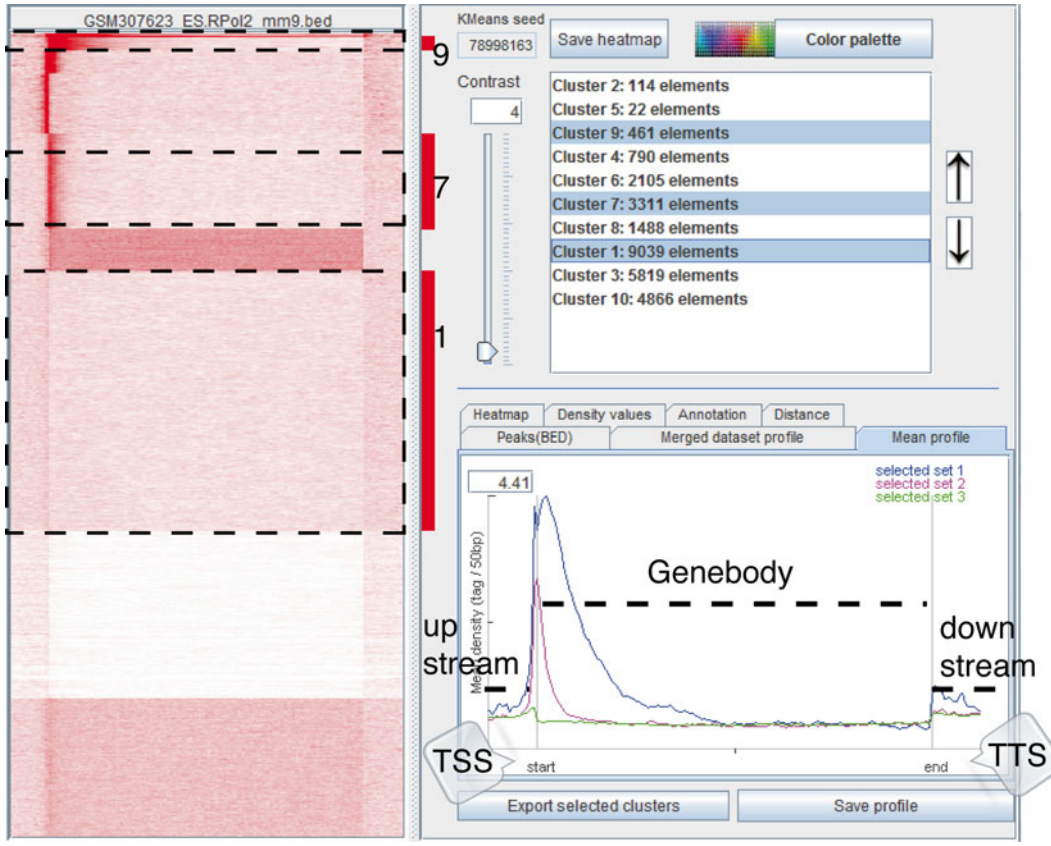

Fig. 3 Visualization of data after gene profile analysis through the Density Array Method. An RNA polymerase II aligned read file is used as an example file (GSM307623) to analyze the gene profile at the reference genes (mm9_refGene.seqminer file). On the *right panel*, the resulting heatmap of the different clusters is shown. The reference file is divided into an upstream region, the TSS, gene body, TTS, and downstream region

densities can be collected within these bins. After the clustering step, the results of the gene profile analysis are shown in a separated java interface (Fig. 3).

3.6 Re-clustering of Data

There are two possibilities of re-clustering data with the Density Array Method:

1. In case the user would like to get the same clustering results through re-clustering of the same dataset, the "Run k-means with a given seed" option under Tools → Options → General should be activated or defined. By default this k-means seed value is randomly generated by seqMINER.
2. For a better resolution of genomic features in particular clusters, obtained through the Density Array Method, seqMINER proposes a re-clustering of data. First, the reference peak bedfile of the generated cluster (s) in Subheading 3.1.3 needs to

be exported as described under Subheading 3.3. Afterward the exported bedfile can be loaded under Subheading 3.1.1 (Load) data as the reference coordinate file, following data extraction. The clustering normalization method should be set to k-means enrichment (linear) before the "Clustering" button is pressed. The data can be analyzed as described before.

3.7 Visualization of Data Without Clustering

seqMINER provides the possibility to visualize and analyze the data using the Density Array Method without prior clustering. For this the heatmap interface can be activated with the Visualization button after data extraction (Subheading 3.1.2). The "Visualization of HeatMap" button is found through a right-mouse click at the highlighted dataset in the Distribution list. In addition, all extracted read densities can be saved as a text file with "Export data." Another advantage of skipping the clustering step is that already analyzed datasets, or existing results, can be uploaded for visualization and reanalysis. In addition, it is possible to annotate all identified peaks (reference coordinates) of a dataset with a peak annotation method.

4 Notes

1. Depending on the quantity of data and the computer configuration, the maximum RAM memory, which is attributed to the Java virtual machine, requires to be increased or decreased using the option –Xmx. For a 32-bit operation system, the maximum available memory is 1.5GB, so the default parameter is -Xmx1500m. There is almost no limitation for memory usage, and it is possible to set a value higher than the physical memory for a 64-bit operating system. To change the memory usage under windows, the seqMINER.bat file needs to be opened with the text editor. Thus, the value in red can be modified: java -Xmx1500m -jar seqMINER.jar.
2. This step takes the most of the analysis time. Therefore, it is recommended to load only the required files. In addition, it is better to load the files one by one or at most two at the same time for a PC with few RAM. Under Tools → "Statistic": Information about the reference coordinate file, loaded and aligned read files, and memory usage can be found. If the button "Garbage collection" is pressed, the memory usage can be reduced.
3. Under Tools → "Statistic all extracted": Distributions with information about the elements per line and estimated memory usage are listed.
4. Different clustering normalization methods can be applied. It is recommended to use k-means raw for single dataset clustering, since there is no normalization between datasets

included. The k-means linear and ranked clustering methods should be used for the analysis of multiple ChIP-seq datasets. These methods implement normalization between datasets as described by Ye et al. [5]. The first method applies linear normalization, whereas the percentile (P) is chosen by the user. The second method is a ranked-based normalization method. The minimum threshold (T) is by default 10 and can be modified. Both parameters (P and T) can be defined under Tools → Options → General, before data extraction.

5. In case there is not enough available memory, the clustering step will take a long time or an error will be indicated. To overcome this problem, it is recommended to clear non-used datasets at the Distribution list (Subheading 3.1.3). The "Delete" option is found at the highlighted/activated datasets with the right-mouse click. In addition, the memory usage can be reduced under Tools → Statistic and "Garbage collection."
6. seqMINER normalizes the datasets before the k-means clustering. The obtained clustering results (Heatmap, density profiles) are presented with the raw sequencing files, which are not normalized.
7. To export the extracted read densities of all generated clusters, it is recommended to highlight the given dataset at the Distribution list in the analysis Subheading 3.1.3. Using the right-mouse button, "Export data" is found. This will generate a text file.
8. Since the order of lines corresponds to the Peak(bed) file, we suggest to create an excel table including the reference coordinates and density values. The columns of the density values represent the defined bins of the dataset(s). In case there are multiple datasets, five columns with the value −1 separate the datasets one by one.
9. From the seqMINER 1.3, we have added a new panel named "Annotation." This function helps us to do the peak annotation with the public databases as Refseq or Ensembl directly in seqMINER. A genome assembly could be selected before the clustering analysis. Recent human and mouse assembly annotations are already provided as a combo box. Customized annotation table could be extracted with Ensembl-biomart tool: (http://www.ensembl.org/biomart/martview/). After selecting the assembly, the attributes should be selected in the following order:

   Chromosome Name
   Gene Start (bp)
   Gene End (bp)
   Ensembl Gene ID
   Associated Gene Name
   Strand

Finally the result can be exported in text format. User can browse the file within the "Advanced" button popup panel or copy the file into the "lib" folder under the unzipped seqMINER folder by adding an extension of ".seqminer" manually for permanent usage.

10. When a control dataset (c) is added, the enrichment values (q) of the analyzed dataset (d) are calculated as described by Ye et al. [5]. This q value can be log(2) transformed.
11. seqMINER does not normalize the average gene profile to the length of the genes. seqMINER divides the distance before and after the gene body in 20 bins, which is fixed for each reference coordinate. In contrast, the distance in the gene body is not in a fixed interval. Thus, the 160 bins of each reference coordinate differ in the number of base pairs. Therefore, small genes like ribosomal or histone genes will appear with artificial high values in the Heatmap.

References

1. O'Neill LP, Turner BM (1996) Immunoprecipitation of chromatin. Meth Enzymol 274:189–197
2. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680
3. Ku CS, Naidoo N, Wu M et al (2011) Studying the epigenome using next generation sequencing. J Med Genet 48:721–730
4. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6(11 Suppl):S22–S32
5. Ye T, Krebs AR, Choukrallah MA et al (2011) seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 39(6):e35
6. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
7. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
8. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
9. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
10. Zang C, Schones DE, Zeng C et al (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25: 1952–1958
11. Fejes AP, Robertson G, Bilenky M et al (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24:1729–1730