Analysis of multi-sample ChIP-seq data using ChIPdig: tutorial

Ruben Esse (rmesse@bu.edu)

05/06/2018

Introduction to ChIPdig

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an invaluable technique to assess genome-wide localization data for DNA-associated proteins. However, several steps in ChIP-seq data analysis, from read mapping to assessment of differential enrichment, often rely on command line utilities that are not readily available to the majority of wet lab researches who wish to generate and process genomic data. ChIPdig provides a multi-modular analysis pipeline designed to facilitate ChIP-seq data analysis. The analysis modules of ChIPdig are:

- Read alignment
- Normalization and comparison of ChIP-seq data sets
- Annotation of genomic regions
- Visualization of coverage (heatmap and metaplot generation)

This vignette explains a common workflow of ChIP-seq data analysis using each of the aforementioned ChIPdig modules. All analyzes described here are performed using the toy data provided, corresponding to ChIP-seq data for the H3K4me3 and H3K36me3 marks (GEO accessions GSE28770 and GSE28776, respectively) in the model organism *C. elegans*. To simplify the workflow and due to file size limitations in the repository hosting the application, only a fraction of the full data is provided.

Installation and Loading

ChIPdig is developed in R programming language and relies on multiple packages provided through Bioconductor, an open source project for the analysis and comprehension of genomic data. The user interface was built using package "shiny". ChIPdig requires R and RStudio versions 3.4.1 and 1.0.44, respectively, or higher.

The application is available at: https://github.com/rmesse/ChIPdig. To be able to use it, you will first need to download the folder to your computer, and then either manually install the necessary packages or let the application do it for you when executed for the first time. Bioconductor packages and their dependencies can be installed by typing the following commands in the R studio console:

```
source("https://bioconductor.org/biocLite.R")
biocLite("BSgenome")
```

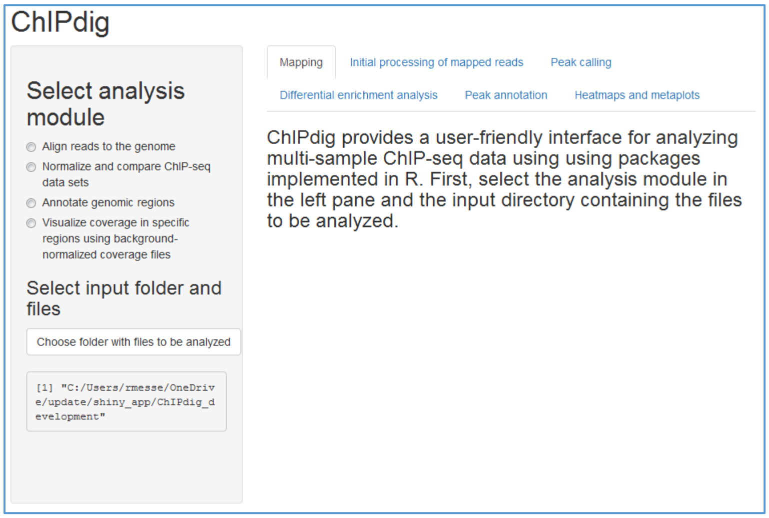
In the example above, the first command establishes communication with the Bioconductor repository and the second command installs package "BSgenome". Repeat the second command for the remaining Bioconductor packages used by ChIPdig:

```
biocLite("QuasR")
biocLite("csaw")
biocLite("edgeR")
biocLite("GenomicAlignments")
biocLite("BayesPeak")
biocLite("ChIPseeker")
biocLite("GenomicFeatures")
biocLite("EnrichedHeatmap")
```

Other necessary packages may be installed with these commands:

```
install.packages("shiny")
install.packages("shinyFiles")
install.packages("ggplot2")
install.packages("ggsignif")
install.packages("reshape2")
install.packages("valr")
install.packages("circlize")
```

To load ChIPdig, simply navigate to the folder downloaded from the repository and open any of the following files using RStudio: “server.R”, “global.R”, or “ui.R”. Then, in the RStudio console, click on the “Run App” button. If any of the packages necessary to run ChIPdig is not installed, this will be done automatically. The user interface then pops us with the following screen:



ChIPdig

Select analysis module

○ Align reads to the genome
○ Normalize and compare ChIP-seq data sets
○ Annotate genomic regions
○ Visualize coverage in specific regions using background-normalized coverage files

Select input folder and files

[1] "C:/Users/rmesse/OneDrive/update/shiny_app/ChIPdig_development"

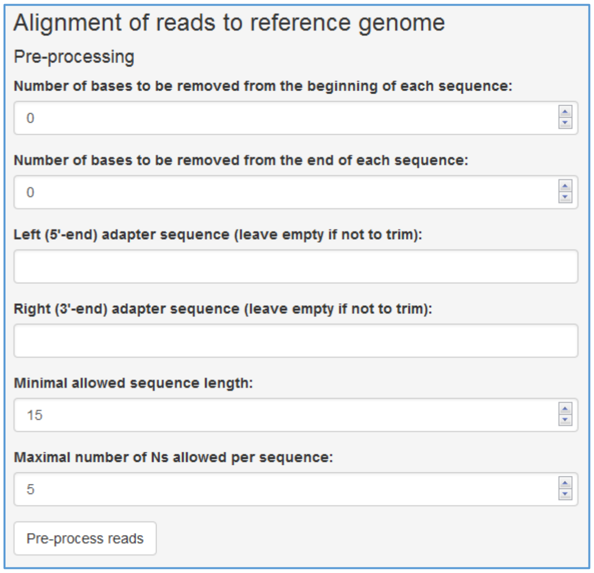
ChIPdig provides a user-friendly interface for analyzing multi-sample ChIP-seq data using using packages implemented in R. First, select the analysis module in the left pane and the input directory containing the files to be analyzed.

Mapping
Initial processing of mapped reads
Peak calling
Differential enrichment analysis
Peak annotation
Heatmaps and metaplots
Choose folder with files to be analyzed

Under “Select input folder and files”, click on “Choose folder with files to be analyzed” and select the folder containing the toy data used in this tutorial.

Read Alignment

In this first section, we will map reads to the reference genome assembly. Each toy data file used in this section of the vignette corresponds to 1 million reads randomly selected from the original files. To load the files, browse to the toy data folder and select file “mapping_toy_data.txt”. This file lists 8 samples: 4 ChIP samples (H3K4me3 and H3K36me3, 2 replicates) and 4 corresponding control samples (no immunoprecipitation).



Alignment of reads to reference genome

Pre-processing

Number of bases to be removed from the beginning of each sequence: 0

Number of bases to be removed from the end of each sequence: 0

Left (5'-end) adapter sequence (leave empty if not to trim):

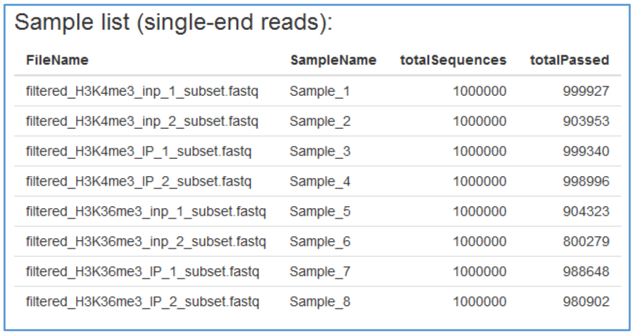
Right (3'-end) adapter sequence (leave empty if not to trim):

Minimal allowed sequence length: 15

Maximal number of Ns allowed per sequence: 5

Reads can be pre-processed. In this example, we use this option to filter out sequences with length less than 15 bases and more than 5 ambiguous (N) bases:

After read pre-processing, the following table is displayed in the screen, showing that at least 80% reads passed the filtering parameters selected.



Sample list (single-end reads):

FileName	SampleName	totalSequences	totalPassed
filtered_H3K4me3_inp_1_subset.fastq	Sample_1	1000000	999927
filtered_H3K4me3_inp_2_subset.fastq	Sample_2	1000000	903953
filtered_H3K4me3_IP_1_subset.fastq	Sample_3	1000000	999340
filtered_H3K4me3_IP_2_subset.fastq	Sample_4	1000000	998996
filtered_H3K36me3_inp_1_subset.fastq	Sample_5	1000000	904323
filtered_H3K36me3_inp_2_subset.fastq	Sample_6	1000000	800279
filtered_H3K36me3_IP_1_subset.fastq	Sample_7	1000000	988648
filtered_H3K36me3_IP_2_subset.fastq	Sample_8	1000000	980902

The files with reads that satisfy the pre-processing parameters are saved in the directory specified by the user (in this case, the toy data folder).

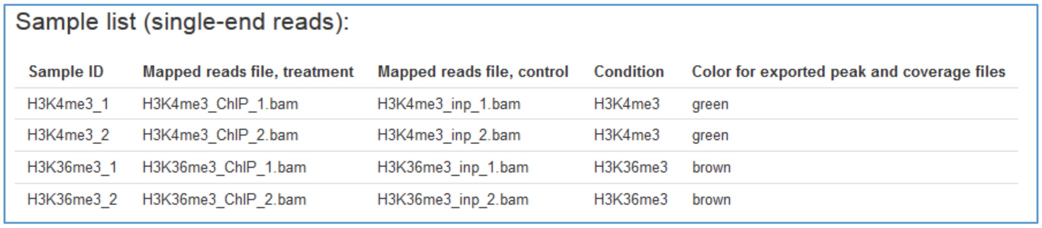
To align the reads to the reference genome assembly, select the latter from the scroll-down menu. For this example, select a recent *C. elegans* reference genome assembly (e.g. "BSgenome.Celegans.UCSC.ce10") and click on "Align reads to reference genome" to start the alignment process. Upon completion of this process, the aligned read files in BAM format are saved to the input directory, along with the corresponding index (BAI) files and text files describing the mapping parameters. A report in PDF format ("qc_report.pdf") listing quality parameters (e.g. percentage of reads mapped) is also generated.

Normalization and Comparison of ChIP-seq data sets

This is the core module of ChIPdig and it serves to normalize different ChIP-seq data sets, perform peak calling, generate coverage files compatible with publicly available genome browsers and assess significant changes in abundance of the target protein between different conditions. The toy data corresponds to reads aligned to 10% of chromosome I (chrI, positions 1 to 1507242) of the *C. elegans* ce10 reference assembly using the full data set (FASTQ format files downloaded from NCBI GEO, accessions GSE28770 and GSE28776).

1. File Loading

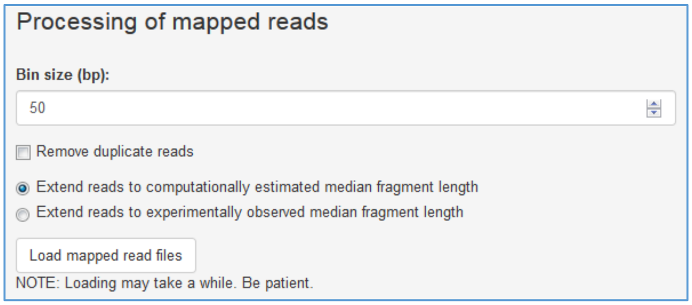
Under "Processing of mapped reads", click on the browse button and select file "mapped_read_processing_toy_data.txt". The latter is a tab-delimited text file describing the ChIP and control (no immunoprecipitation) files to be analyzed. Upon selection of this file, a table describing the files is displayed in the screen:



Sample list (single-end reads):

Sample ID	Mapped reads file, treatment	Mapped reads file, control	Condition	Color for exported peak and coverage files
H3K4me3_1	H3K4me3_ChIP_1.bam	H3K4me3_inp_1.bam	H3K4me3	green
H3K4me3_2	H3K4me3_ChIP_2.bam	H3K4me3_inp_2.bam	H3K4me3	green
H3K36me3_1	H3K36me3_ChIP_1.bam	H3K36me3_inp_1.bam	H3K36me3	brown
H3K36me3_2	H3K36me3_ChIP_2.bam	H3K36me3_inp_2.bam	H3K36me3	brown

The left pane displays the following options:



Processing of mapped reads

Bin size (bp):

50

□ Remove duplicate reads

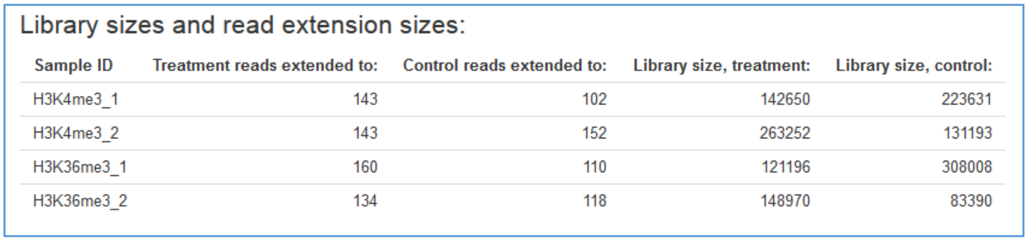
● Extend reads to computationally estimated median fragment length

○ Extend reads to experimentally observed median fragment length

Load mapped read files

NOTE: Loading may take a while. Be patient.

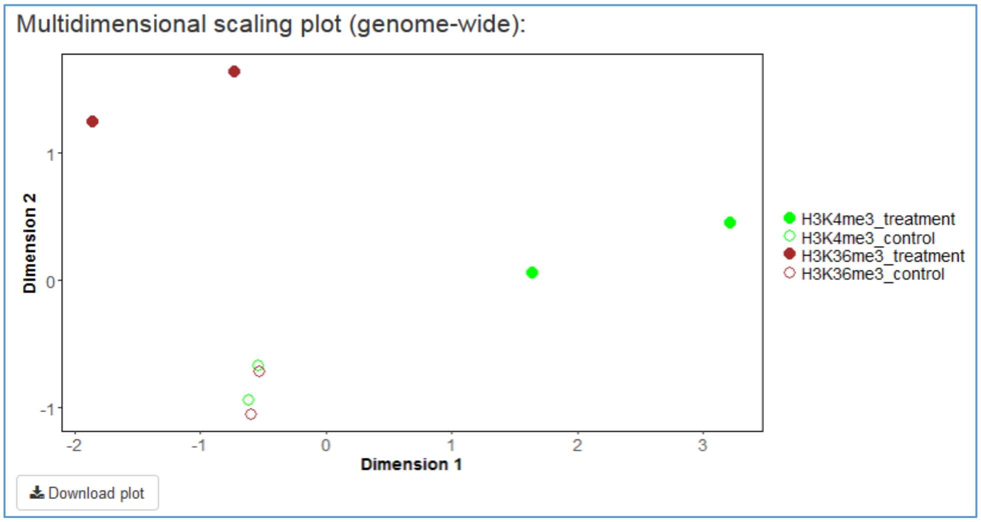
The bin size will be used throughout the module. In ChIP-seq data analysis, it might be important to extend reads prior to further processing. This is because each read corresponds to the sequence at the end of the fragment, i.e., it does not span the entire length of the fragment. Therefore, we need to extend the reads in order to have a better representation of the fragments obtained experimentally. In this example, we opt to extend reads to a computationally estimated median fragment length. After files are loaded, a table of library sizes (i.e. number of reads) and read extension sizes is displayed:



Library sizes and read extension sizes:

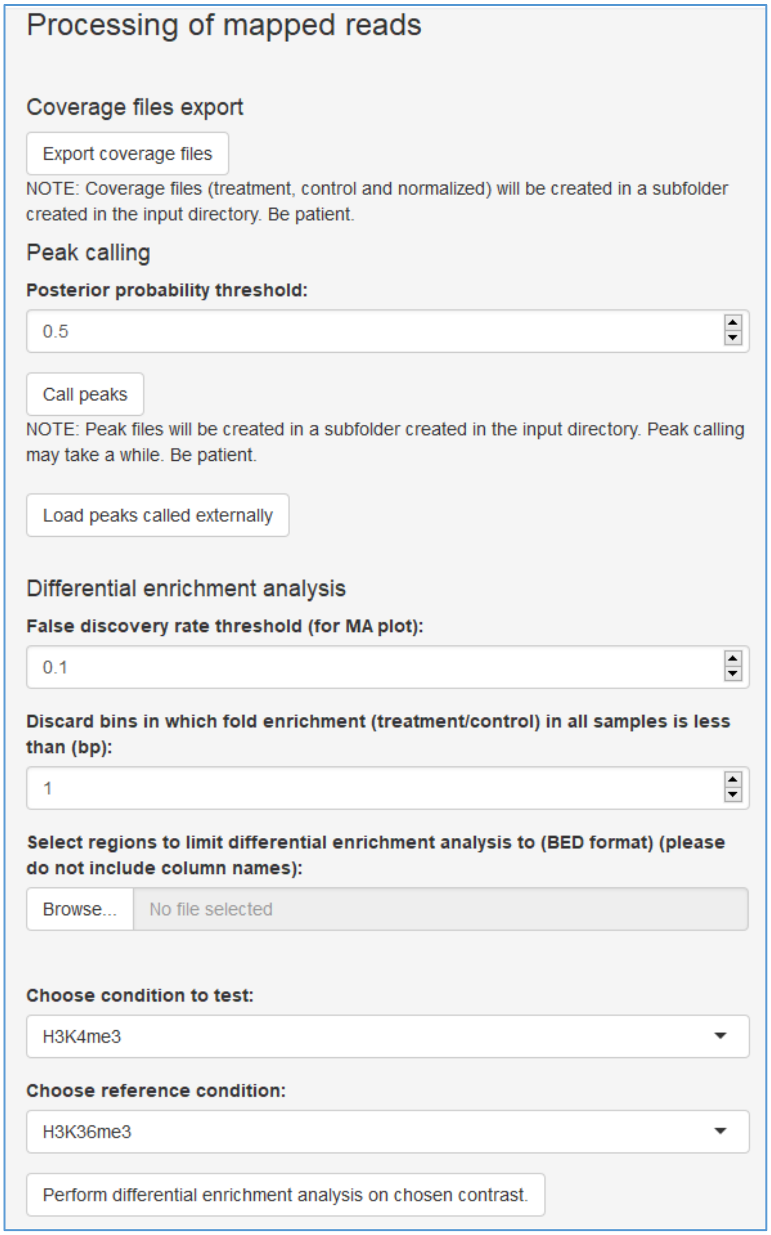
Sample ID	Treatment reads extended to:	Control reads extended to:	Library size, treatment:	Library size, control:
H3K4me3_1	143	102	142650	223631
H3K4me3_2	143	152	263252	131193
H3K36me3_1	160	110	121196	308008
H3K36me3_2	134	118	148970	83390

A PCA plot is also generated:



Multidimensional scaling plot (genome-wide):

Note that the replicates corresponding to the H3K4me3 (filled green) and H3K36me3 (filled brown) treatment samples form two separate clusters, whereas the 4 control samples form a single cluster. This may be expected, since control DNA (no immunoprecipitation) is obviously not biased towards the target of the ChIP experiment (either no antibody is used, or mock antibody). After initial processing of input reads, the following options for downstream processing are toggled:



Processing of mapped reads

Coverage files export

NOTE: Coverage files (treatment, control and normalized) will be created in a subfolder created in the input directory. Be patient.

Peak calling

Posterior probability threshold: 0.5

NOTE: Peak files will be created in a subfolder created in the input directory. Peak calling may take a while. Be patient.

Differential enrichment analysis

False discovery rate threshold (for MA plot): 0.1

Discard bins in which fold enrichment (treatment/control) in all samples is less than (bp): 1

Select regions to limit differential enrichment analysis to (BED format) (please do not include column names): Browse... No file selected

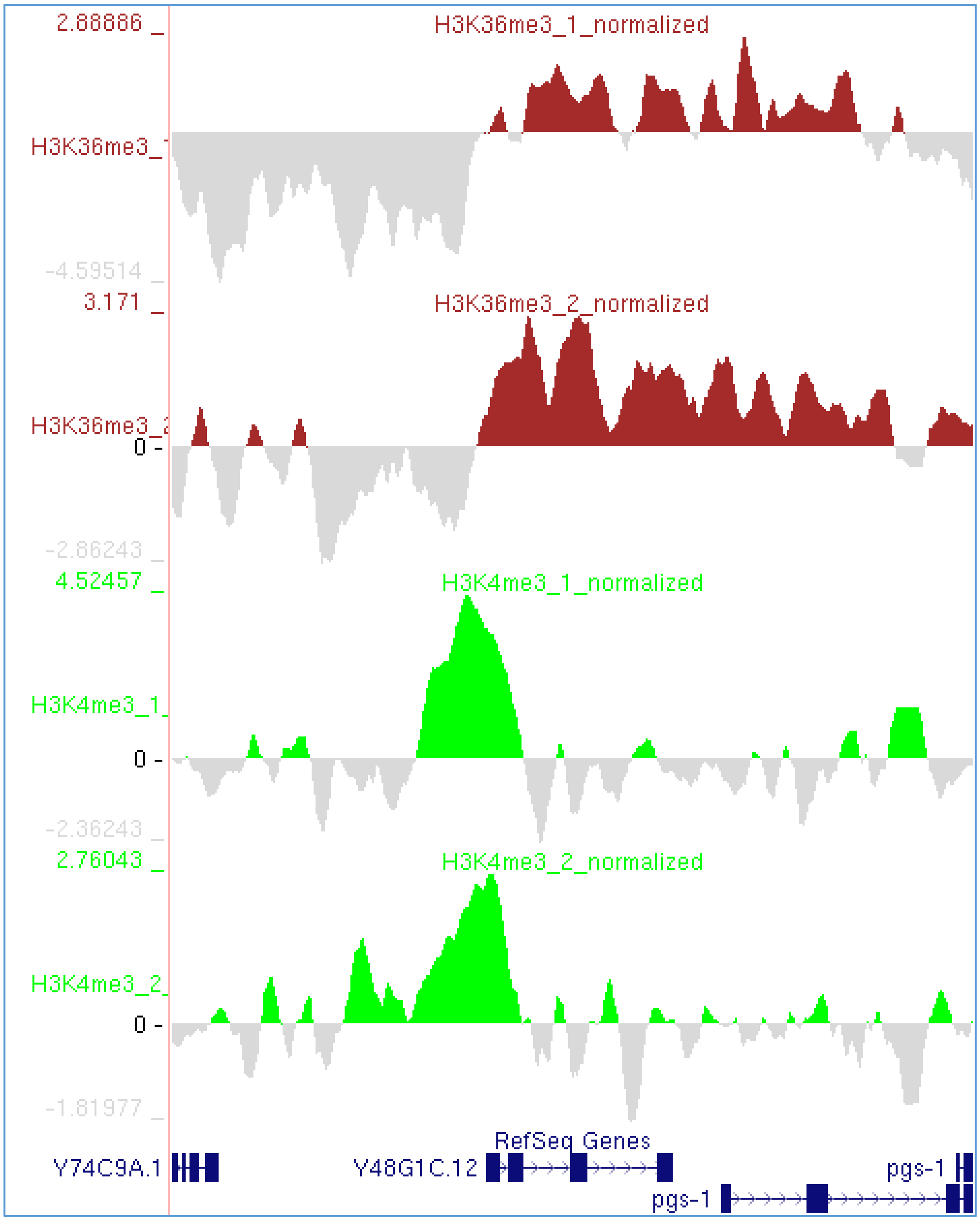
Choose condition to test: H3K4me3

Choose reference condition: H3K36me3

Perform differential enrichment analysis on chosen contrast.

2. Generation of Coverage Files

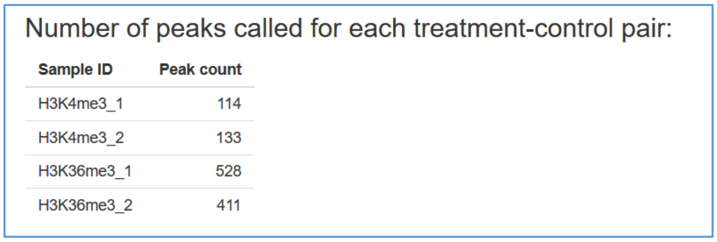
This option generates coverage files in BedGraph format (see https://genome.ucsc.edu/goldenpath/help/bedgraph.html for more information) which can be uploaded to a genome browser. The following is a UCSC genome browser snapshot with H3K4me3 and H3K36me3 control-subtracted coverage tracks generated by ChIPdig.



Note the higher density of coverage for both replicates of H3K4me3 at the region upstream of gene Y48G1C.12, and, conversely, the accumulation of H3K36me3 at the coding region. This profile is typical of these modification marks for several organisms, from yeast to humans.

3. Peak Calling

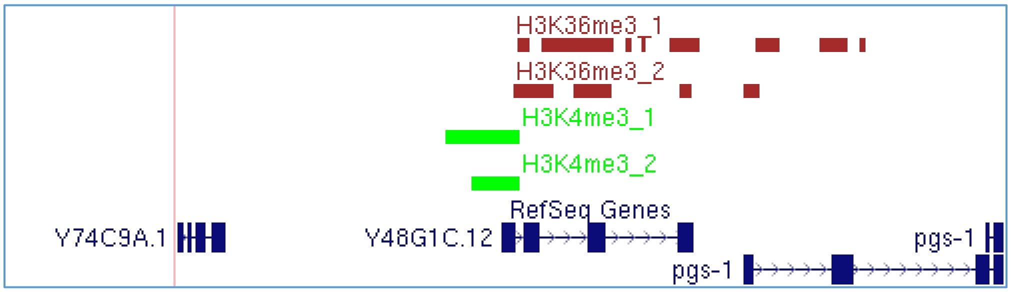
Peak calling is performed to identify areas in the genome that are enriched in the protein of interest compared with the background (i.e. no immunoprecipitation). Peak calling with a higher posterior probability threshold will be more stringent. In this example, we call peaks with the default value (0.5). Upon completion of peak calling, the following table is displayed in the screen, as well as optional commands to generate replicated peak files (i.e. regions enriched in both replicates for each target) and a consensus peak file representing peaks observed in the two conditions (“consensus_peaks.bed”). All files are saved to a subfolder located in the toy data folder.



Number of peaks called for each treatment-control pair:

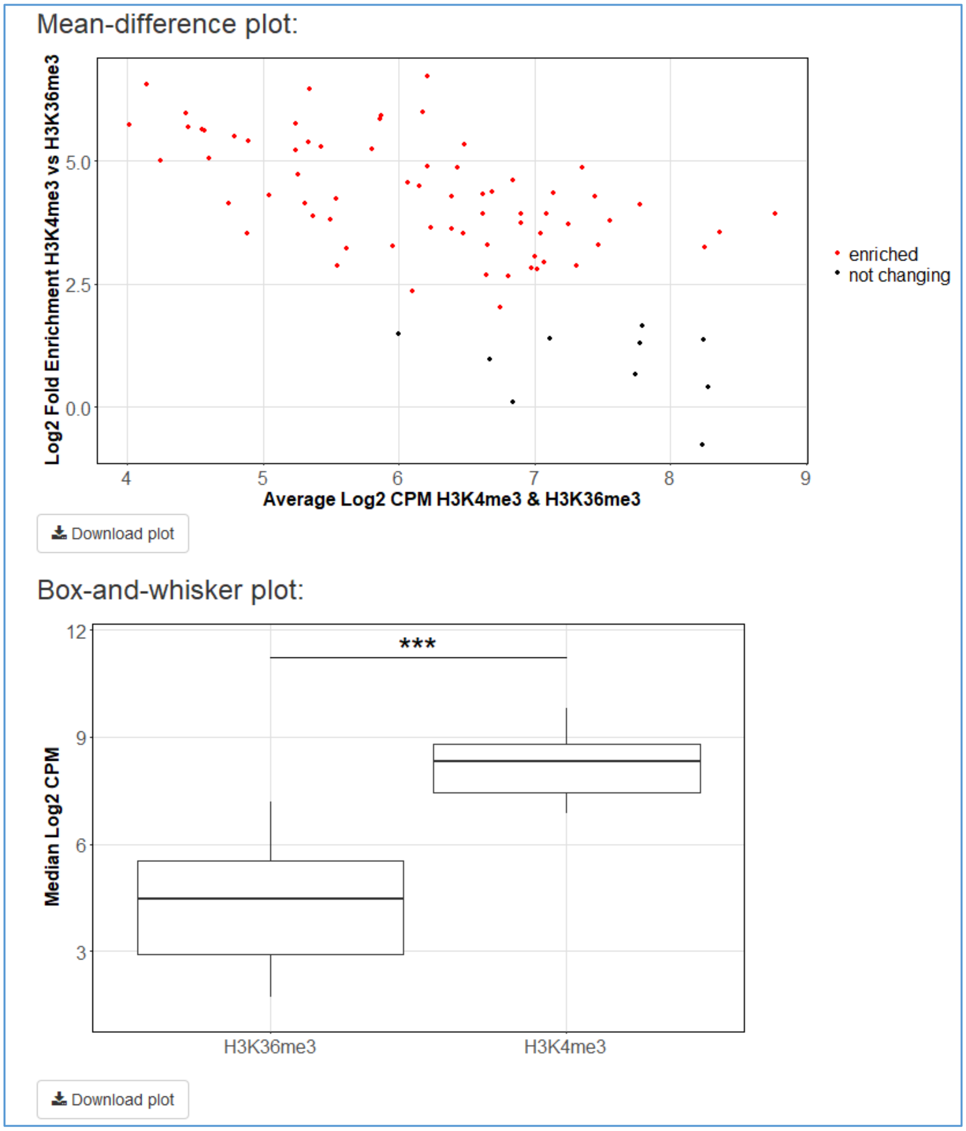
Sample ID	Peak count
H3K4me3_1	114
H3K4me3_2	133
H3K36me3_1	528
H3K36me3_2	411

An option to add a track definition line to each peak file is also provided, for which the color selected at input is used (i.e. green for H3K4me3 peaks and brown for H3K36me3 peaks). As expected for these histone modification marks, H3K4me3 has a peak at the promoter region of gene Y48G1C.12, whereas peaks of H3K36me3 are located at the coding region:



4. Differential Enrichment Analysis

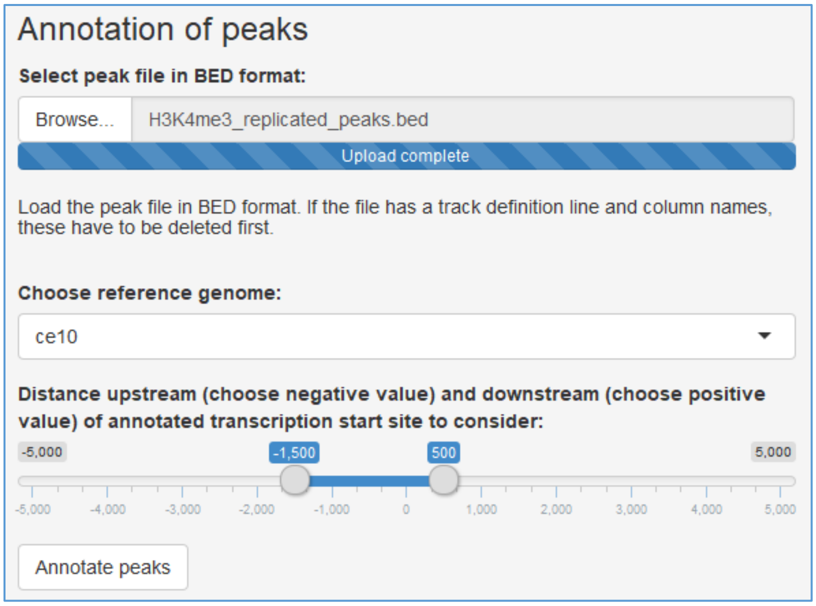
The core capability of ChIPdig lies at the identification of differentially bound regions between two sample groups. This analysis can span the entire genome, or set to user-defined regions. Bins in which the coverage of the control sample is greater or equal to that of the treatment (immunoprecipitated) sample are discarded. In this example, we select the "H3K4me3_replicated_peaks.bed" file as input, representing replicated H3K4me3 peaks. As expected, coverage for the H3K4me3 samples at these regions is higher than that of the H3K36me3 samples.



A data table with fold enrichment for each region, as well as average log2-transformed coverage, P-value and false discovery rate is also displayed and can be downloaded.

Annotation of Genomic Regions

One may be interested in knowing whether a set of regions (e.g. those obtained by peak calling) is enriched in a certain type of regions in the genome. The H3K4me3 mark, for example, is enriched at promoters. The annotation module of ChIPdig overlaps user-defined regions with annotation information available for the genome of interest. In this example, we load the “H3K4me3_replicated_peaks.bed” file and choose as promoter region an arbitrary window spanning 1500 bp upstream of the transcription start site to 500 bp downstream of the transcription start site.



Annotation of peaks

Select peak file in BED format:

H3K4me3_replicated_peaks.bed

Load the peak file in BED format. If the file has a track definition line and column names, these have to be deleted first.

Choose reference genome:

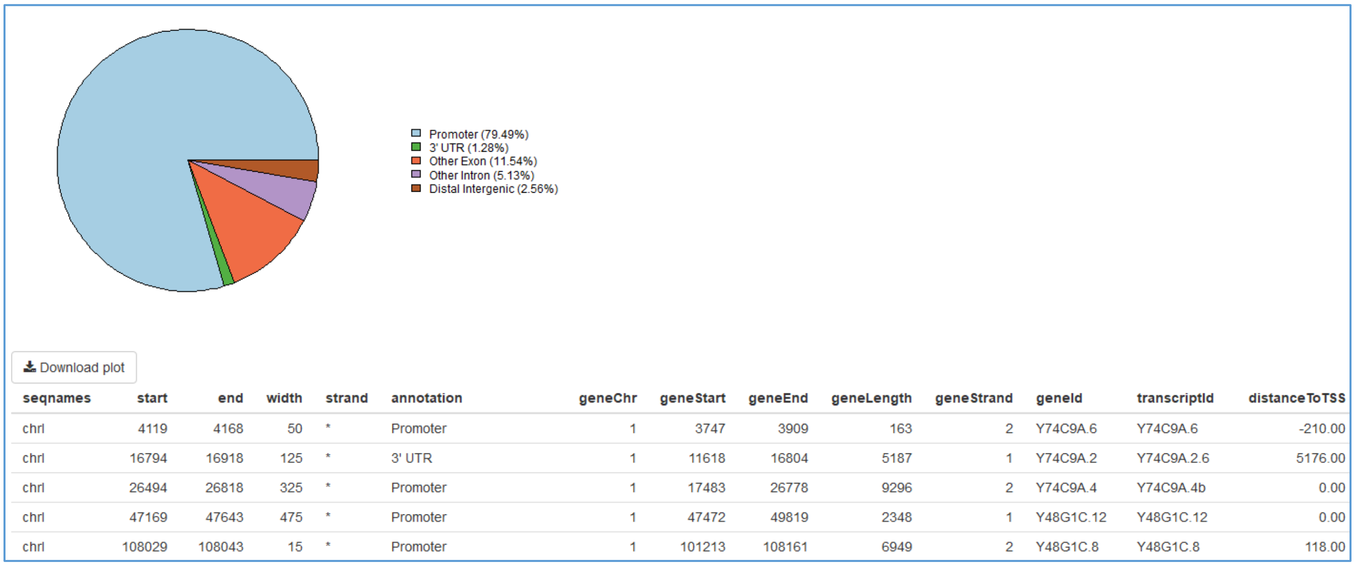
ce10

Distance upstream (choose negative value) and downstream (choose positive value) of annotated transcription start site to consider:

-1,500

500

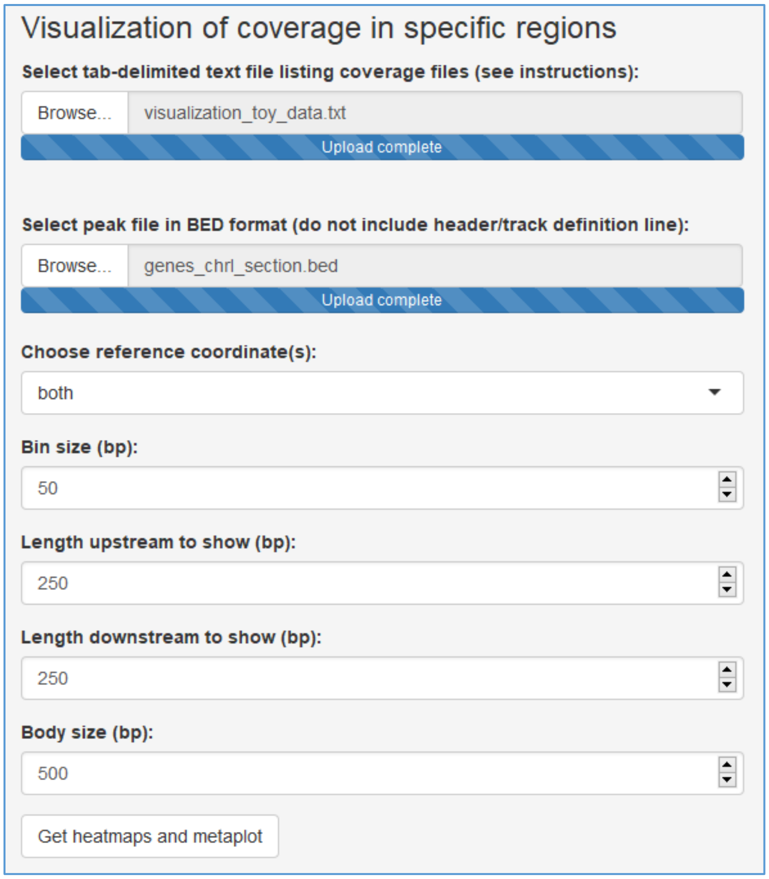
The output is a pie plot representing percentages of regions at annotation classes, as well as a table in which each regions is annotated. For illustration purpose, only the first 5 rows are shown:



seqnames	start	end	width	strand	annotation	geneChr	geneStart	geneEnd	geneLength	geneStrand	geneId	transcriptId	distanceToTSS
chrI	4119	4168	50	*	Promoter	1	3747	3909	163	2	Y74C9A.6	Y74C9A.6	-210.00
chrI	16794	16918	125	*	3' UTR	1	11618	16804	5187	1	Y74C9A.2	Y74C9A.2.6	5176.00
chrI	26494	26818	325	*	Promoter	1	17483	26778	9296	2	Y74C9A.4	Y74C9A.4b	0.00
chrI	47169	47643	475	*	Promoter	1	47472	49819	2348	1	Y48G1C.12	Y48G1C.12	0.00
chrI	108029	108043	15	*	Promoter	1	101213	108161	6949	2	Y48G1C.8	Y48G1C.8	118.00

Visualization of coverage

This module of ChIPdig provides a graphical representation of ChIP-dig coverage in the context of regions defined by the user (e.g. genes). The heatmaps and comparative metaplot can be centered at the start coordinates (e.g. transcription start sites), the end coordinates (e.g. transcription termination sites), or show both. In the latter situation, region lengths are expanded or compressed to a specific value. In this tutorial, we choose to visualize at genes present in the first 1.5 kbp of chromosome I. We show 250 bp upstream of the transcription start site and 250 bp downstream of the transcription start site. Please load file “visualization_toy_data.txt”, found in the toy data subfolder, as well as the regions file (“genes_chrI_section.bed”).



Visualization of coverage in specific regions

Select tab-delimited text file listing coverage files (see instructions):

Browse... visualization_toy_data.txt

Upload complete

Select peak file in BED format (do not include header/track definition line):

Browse... genes_chrl_section.bed

Upload complete

Choose reference coordinate(s): both

Bin size (bp): 50

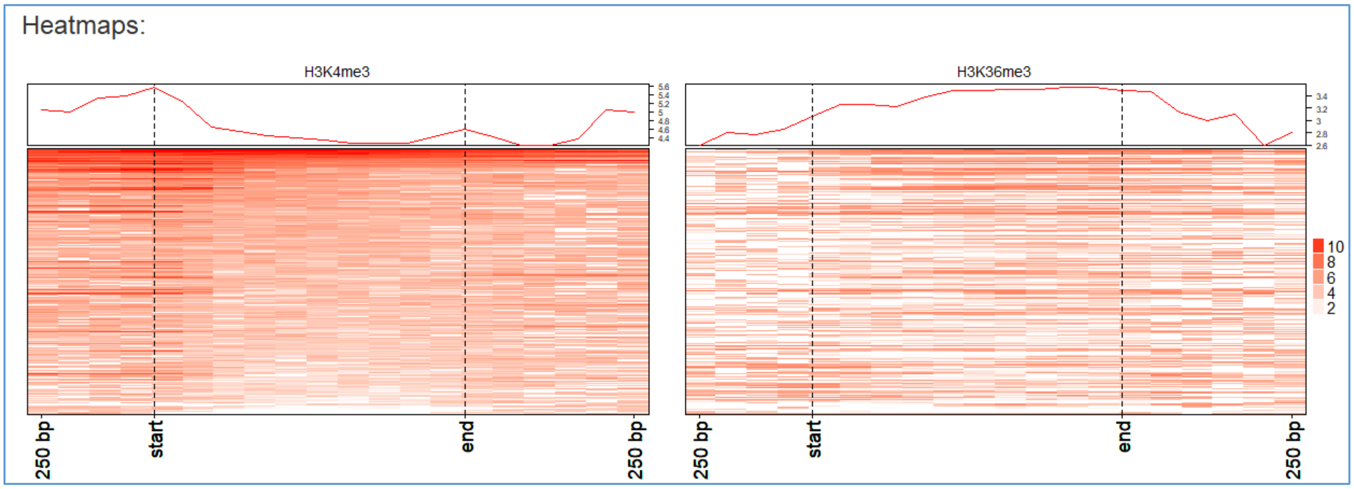
Length upstream to show (bp): 250

Length downstream to show (bp): 250

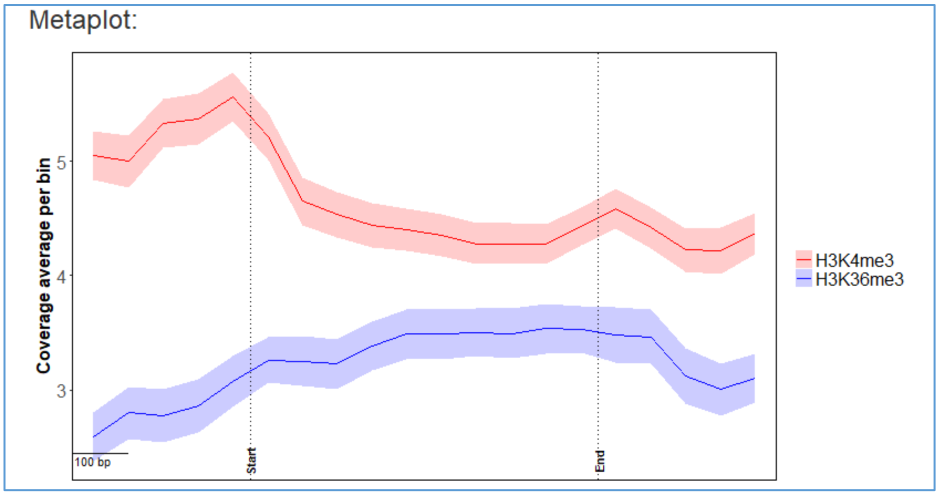
Body size (bp): 500

Get heatmaps and metaplot

As expected, there is higher density of H3K4me3 around gene transcription start sites, whereas H3K36me3 is distributed along genes bodies.



Heatmaps:



Metaplot:

Coverage average per bin
5
4
3
100 bp
Start
End
H3K4me3
H3K36me3