ChIPdig: a comprehensive user-friendly tool for mining multi-sample ChIP-seq data

Ruben Esse[1,*] and Alla Grishok[1,*]

[1]Department of Biochemistry, Boston University School of Medicine, Boston, MA, 02118, USA

*To whom correspondence should be addressed.

Abstract

Summary: In this article we present ChIPdig, a tool for analyzing and comparing multiple ChIP-seq data sets. ChIPdig is written in R and enables access to powerful R-based functions and packages through a simple user interface powered by the shiny package. It allows users to align reads to a reference genome, perform peak calling and differential enrichment analysis on regions of interest, annotate such regions based on available coordinates of transcription start and termination sites, exons, introns, 5' UTRs and 3' UTRs, and generate heatmaps and metaplots for visualizing coverage.

Availability and implementation: ChIPdig is open source and available at https://github.com/rmesse/ChIPdig.

Contact: rmesse@bu.edu or agrishok@bu.edu

1 Introduction

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a widely used technique to assess protein-DNA interactions in a genome-wide manner, generating binding profiles of proteins, such as transcription factors and modified histones, under various conditions. Decryption of key biological features from such data relies heavily on a series of computational analytical procedures, but these are accessible mostly to skilled computational researchers. As such, data processing is at present a considerable bottleneck in ChIP-seq data mining, particularly when multiple samples and conditions are to be analyzed.

To help experimentalists process multi-sample ChIP-seq data, we developed a user-friendly tool named ChIPdig. Four independent analysis modules were built, corresponding to different steps of the analysis pipeline. The first module allows the user to map reads to a reference genome. The sequences may be pre-processed to remove adapters and filter-out low-quality reads. The second module was designed to normalize and compare ChIP-seq data sets corresponding to different targets (e.g. H3K4me3 *versus* H3K36me3) or conditions (e.g. drug *versus* vehicle control). An option to call peaks in each ChIP sample with a matching input DNA sample is provided, and peaks can be further processed to obtain replicated peaks (in case replicate samples are used) and a consensus peak set with peaks called in all samples in the set. The differential enrichment analysis tool serves to obtain regions significantly enriched or depleted in a condition to test by comparison with a control condition. Lastly, the fourth module allows visualization of coverage in a set of regions, either provided by the user or obtained by the peak calling command, by means of heatmaps and a comparative metaplot.

ChIPdig provides a comprehensive solution to address the challenges that ChIP-seq data analysis presents. It is easy to use and requires no programming expertise.

2 Implementation

ChIPdig is developed in R using package shiny and relies on multiple R/Bioconductor packages, including QuasR for mapping (Gaidatzis *et al.*, 2015), BayesPeak for peak calling (Cairns *et al.*, 2011), csaw (Lun and Smyth, 2015) and edgeR (Robinson *et al.*, 2010) for differential enrichment analysis, ChIPseeker for annotation of genomic regions (Yu *et al.*, 2015), and EnrichedHeatmap for generating coverage heatmaps (Gu, 2017). Operations on genomic intervals in all modules is performed using either GenomicRanges (Lawrence *et al.*, 2013) or valr (A. Riemondy *et al.*, 2017).

3 Results

To test ChIPdig performance for each module, raw single-end ChIP-seq data for H3K4me3 and H3K36me3 in *Caenorhabditis elegans* were downloaded from NCBI GEO with series IDs GSE28770 and GSE28776, respectively. These two histone modifications have different distribution profiles, namely promoter-proximal enrichment and punctuated peaks for H3K4me3, and gene body enrichment and broad peaks for H3K36me3 (Liu *et al*., 2011), and therefore serve as a suitable and clear illustration of the usefulness of ChIPdig.

3.1 Alignment of reads to a reference genome

The aligning module of ChIPdig relies primarily on the QuasR package, which supports the analysis of single read and paired-end deep sequencing experiments. Sequence files in FASTQ format for each input DNA and ChIP sample were aligned to the WS220/ce10 reference genome, producing an average of 10 million reads per sample. A report with several quality metrics for each alignment was automatically generated and is available in Supplementary Figure S1.

3.2 Normalization and comparison of ChIP-seq data sets

The alignment files were subjected to the peak calling tool via the BayesPeak package, which uses a hidden Markov model and Bayesian statistical methodology (Cairns *et al*., 2011). Peaks called in the two replicas were obtained for both marks and are comparable to the publicly available peak sets for the same data (Supplementary Tables S1 and S2 for H3K4me3 and H3K36me3, respectively). Differential enrichment analysis was performed on either H3K4me3 peaks or H3K36me3 peaks to assess regions enriched or depleted in H3K36me3 in comparison with H3K4me3 (Figure 1A).

3.3 Annotation of regions

Annotation of the peaks called for both marks was based on the ChIPseeker package. In line with previous well-known observations (Liu *et al*., 2011), H3K4me3 was enriched at promoters (< 1500 bp upstream of annotated transcription start site of each gene), whereas H3K36me3 was enriched at gene bodies (Figure 1B). The full annotation data for each peak set can be found in Supplementary Tables S3 and S4, respectively.

3.4 Visualization of coverage

Visualization of background-normalized coverage for each modification was performed using coordinates corresponding to all transcription units in assembly WS220/ce10 (n = 52444). 500 bp windows upstream and downstream of each gene were visualized in 10 bp segments, and gene body lengths were either compressed or expanded to 1000 bp. Typical of H3K4me3, the maximum cumulative coverage coincides with the transcription site, while H3K36me3 coverage is higher along gene bodies (Figures 1C and 1D).

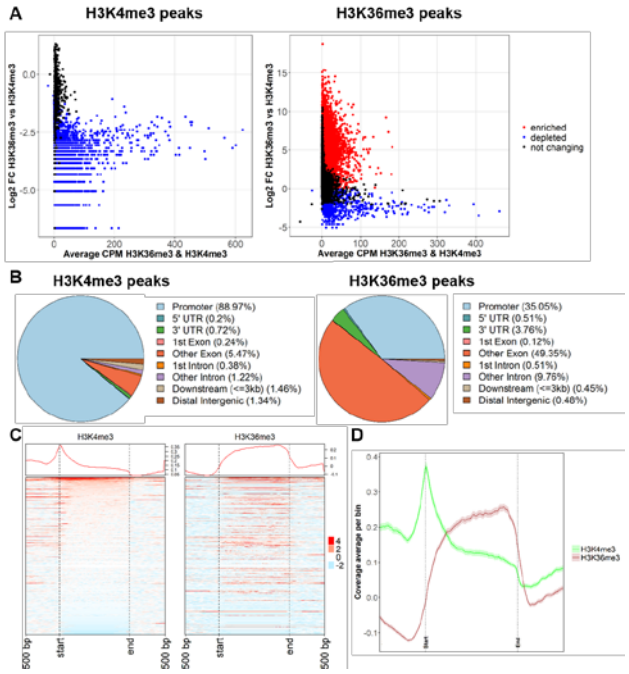

Fig. 1. ChIPdig results for example ChIP-seq datasets (NCBI GEO with series IDs GSE28770 and GSE28776). (A) Differential enrichment analysis in H3K36me3 *vs* H3K4me3 in peaks defined by H3K4me3 enrichment (left) or H3K36me3 enrichment (right). **(B)** Annotation of H3K4me3 and H3K36me3 obtained by the peak calling command. **(C,D)** Visualization of background-normalized coverage of H3K4me3 and H3K36me3 in *C. elegans* transcription units and 500 bp windows upstream and downstream.

Funding

This work has been supported by the NIH grant R01 GM107056 awarded to A.G..

Conflict of Interest: none declared.

References

A. Riemondy,K. *et al.* (2017) valr: Reproducible genome interval analysis in R. *F1000Research*, **6**, 1025.

Cairns,J. *et al.* (2011) BayesPeak-an R package for analysing ChIP-seq data. *Bioinformatics*, **27**, 713–714.

Gaidatzis,D. *et al.* (2015) QuasR: Quantification and annotation of short reads in R. *Bioinformatics*, **31**, 1130–1132.

Gu,Z. (2017) EnrichedHeatmap: Making Enriched Heatmaps. R package version 1.6.0. *https://github.com/jokergoo/EnrichedHeatmap*.

Lawrence,M. *et al.* (2013) Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.*, **9**.

Liu,T. *et al.* (2011) Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res.*, **21**, 227–236.

Lun,A.T.L. and Smyth,G.K. (2015) Csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.*, **44**, 1–10.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.

Yu,G. *et al.* (2015) ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.