Check for updates

SOFTWARE TOOL ARTICLE

ChIPdig: a comprehensive user-friendly tool for mining multi-sample ChIP-seq data [version 1; peer review: 3 approved with reservations]

Ruben Esse

Department of Biochemistry, Boston University School of Medicine, Boston, MA, 02118, USA

V1 First published: 31 Jul 2019, 8:1295
<https://doi.org/10.12688/f1000research.20027.1>
 Latest published: 31 Jul 2019, 8:1295
<https://doi.org/10.12688/f1000research.20027.1>

Abstract

In recent years, epigenetic research has enjoyed explosive growth as high-throughput sequencing technologies become more accessible and affordable. However, this advancement has not been matched with similar progress in data analysis capabilities from the perspective of experimental biologists not versed in bioinformatic languages. For instance, chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) is at present widely used to identify genomic loci of transcription factor binding and histone modifications. Basic ChIP-seq data analysis, including read mapping and peak calling, can be accomplished through several well-established tools, but more sophisticated analyzes aimed at comparing data derived from different conditions or experimental designs constitute a significant bottleneck. We reason that the implementation of a single comprehensive ChIP-seq analysis pipeline could be beneficial for many experimental (wet lab) researchers who would like to generate genomic data.

Here we present ChIPdig, a stand-alone application with adjustable parameters designed to allow researchers to perform several analyzes, namely read mapping to a reference genome, peak calling, annotation of regions based on reference coordinates (e.g. transcription start and termination sites, exons, introns, and 5' and 3' untranslated regions), and generation of heatmaps and metaplots for visualizing coverage. Importantly, ChIPdig accepts multiple ChIP-seq datasets as input, allowing genome-wide differential enrichment analysis in regions of interest to be performed. ChIPdig is written in R and enables access to several existing and highly utilized packages through a simple user interface powered by the Shiny package. Here, we illustrate the utility and user-friendly features of ChIPdig by analyzing H3K36me3 and H3K4me3 ChIP-seq profiles generated by the modENCODE project as an example.

ChIPdig offers a comprehensive and user-friendly pipeline for analysis of multiple sets of ChIP-seq data by both experimental and computational researchers. It is open source and available at

Open Peer Review

Approval Status ? ? ?

	1	2	3
version 1	?	?	?
31 Jul 2019	view	view	view
1. Tao Liu , Roswell Park Comprehensive Cancer Center, Buffalo, USA			
2. Hinrich Gronemeyer , University of Strasbourg, Illkirch, France			
Marco Antonio Mendoza-Parra , University of Evry-val-d'Essonne, University Paris-Saclay, Évry, France			
3. Thomas Carroll , The Rockefeller University, New York City, USA			

Any reports and responses or comments on the article can be found at the end of the article.

<https://github.com/rmesse/ChIPdig>.

Keywords

ChIP-seq, read mapping, peak calling, genomic region annotation, differential enrichment analysis, heatmaps, metaplots

This article is included in the **RPackage** gateway.

Corresponding author: Ruben Esse (rmesse@bu.edu)

Author roles: **Esse R:** Conceptualization, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was supported by the GM107056 R01 grant awarded by the National Institutes of Health to Dr. Alla Grishok.

Copyright: © 2019 Esse R. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Esse R. **ChIPdig: a comprehensive user-friendly tool for mining multi-sample ChIP-seq data [version 1; peer review: 3 approved with reservations]** F1000Research 2019, **8**:1295 <https://doi.org/10.12688/f1000research.20027.1>

First published: 31 Jul 2019, **8**:1295 <https://doi.org/10.12688/f1000research.20027.1>

Abbreviations

bp: base pairs; ChIP-seq: chromatin immunoprecipitation followed by next-generation sequencing; cpm: counts per million; FDR: false discovery rate; GEO: Gene Expression Omnibus; GUI: graphical user interface; H3K36me3: histone H3 trimethylated at lysine 36; H3K4me3: histone H3 trimethylated at lysine 4; NCBI: National Center for Biotechnology Information; NGS: next-generation sequencing; PCR: polymerase chain reaction; PP: posterior probability; TMM: trimmed means of M values; TSS: transcription start site; UCSC: University of California Santa Cruz

Introduction

Interactions between nuclear proteins and DNA are vital for cell and organism function. They control DNA replication and repair, safeguard genome stability, and regulate chromosome segregation and gene expression. Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) is a powerful method for assessing such interactions and has been widely used in recent years to map the location of post-translationally modified histones, transcription factors, chromatin modifiers and other non-histone DNA-associated proteins in a genome-wide manner. This progress has been fostered by the increasing technical feasibility and affordability of this technology, with more documentation and technical support available to researchers, including commercial library preparation kits, aided by the plummeting costs of sequencing and the ease of multiplexing samples. In light of this progress, ChIP-seq data sets are continuously deposited in publicly-accessible databases, such as the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and the ENCODE consortium portal¹. Therefore, there is an unprecedented wealth of epigenomic data in the public domain that can be used for integrative and correlative analyses.

This important advancement has not been accompanied with similar progress in user-friendly post-sequencing data analysis pipelines, which is still a significant bottleneck often handled by skilled bioinformaticians. A key step in ChIP-seq data analysis is to map reads to a reference genome assembly. Programs such as BWA² or Bowtie/Bowtie2^{3,4} are frequently employed for this purpose. Aligned data are then processed to find regions of enrichment along the genome, thereby identifying potential loci of DNA binding by the target protein or of deposition of the histone post-translational modifications of interest. This process is known as peak calling and can be performed by using algorithms such as MACS/MACS2^{5,6} and SICER⁷. SAMtools is another popular software used in next-generation sequencing (NSG) analysis that provides various utilities for manipulating alignments, including sorting, merging, indexing and generating alignments in a per-position format⁸.

Following the initial and common steps in ChIP-seq data analysis mentioned above, downstream analysis is often more customizable and may present a hurdle, especially in the case of comparing data derived from different conditions or experimental settings. Artifacts arising from bias of DNA fragmentation, variation of immunoprecipitation efficiency, as well as polymerase chain reaction (PCR) amplification and sequencing depth bias, result in ChIP-seq experiments with distinct signal-to-noise

ratios and impose great challenges to the computational analysis⁹. Several methods addressing the differential enrichment analysis problem, i.e. the detection of genomic regions with changes in ChIP-seq profiles between two distinct samples or sets of replicate samples, have been proposed^{10–12}. Generally, these methods rely on the initial detection of candidate peak regions by a conventional peak calling algorithm, and then this peak-defining information is applied to analysis with methods tailored for the differential expression analysis of RNA-seq data such as edgeR¹³ or DESeq¹⁴. Downstream analysis of ChIP-seq data may also involve the annotation of genomic regions based on reference coordinates (e.g. distance from nearest transcription start site) and visualization of normalized coverage by means of heatmaps and metaplots.

Importantly, ChIP-seq data analysis typically relies, at least in part, on tools that have been designed to run primarily on Linux/Unix-based systems, while biologists who need to work with NGS may be unfamiliar with such operating systems. In the past few years, several packages designed to handle NGS data have been released for R, a popular platform-independent programming language and software environment for computing and graphics. Here we present ChIPdig¹⁵, an open-source application that leverages on several R packages to enable comprehensive and modular analysis of multi-sample ChIP-seq data sets through a user-friendly graphical user interface (GUI), allowing any experimental biologist with minimal computational expertise to use it easily.

Methods

Implementation

ChIPdig¹⁵ is developed in R using package Shiny and relies on multiple R/Bioconductor packages, including **QuasR** and **BSgenome**¹⁶, for mapping reads to a reference genome assembly¹⁷, **BayesPeak** for peak calling¹⁸, **csaw**¹² and **edgeR**¹³ for differential enrichment analysis, **ChIPseeker** for annotation of genomic regions of interest¹⁹, and **EnrichedHeatmap** for generating coverage heatmaps²⁰. Packages used in several analysis modules implemented in ChIPdig are **GenomicRanges**²¹, **valr**²², **shinyFiles**, **GenomicFeatures**, **ggplot2**, **ggsignif**, **reshape2** and **circize**. The GUI allows integrative and interactive usage of these powerful libraries without requiring programming or statistical experience from the user. The required packages may be installed manually by the user or automatically upon launching the tool for the first time. A step-by-step tutorial on how to use ChIPdig is provided at <https://github.com/rmesse/ChIPdig>.

Operation

R and RStudio versions 3.4.1 and 1.0.44 or higher, respectively, are required, and both memory usage and execution times are contingent on the specificities of the intended analysis (e.g. size of input files, sequencing depths and size of the reference genome assembly).

ChIPdig has the following capabilities: (1) alignment of reads to a reference genome; (2) normalization and comparison of distinct ChIP-seq data sets; (3) annotation of genomic regions; (4) generation of comparative heatmaps and metaplots for visualization of normalized coverage in user-defined specific

regions. Each capability corresponds to a specific analysis module which can be loaded separate from other modules and using different input files. For the alignment and annotation modules, the tool automatically fetches the desired reference assembly and gene model, respectively, from public repositories. Analysis of mammalian genome data is supported.

An earlier version of this article can be found on bioRxiv (DOI: <https://doi.org/10.1101/220079>).

Use cases

To illustrate each module of the tool, raw single-end ChIP-seq data for H3K4me3 and H3K36me3 in the model organism *Caenorhabditis elegans* were downloaded from NCBI GEO with series IDs [GSE28770](#) and [GSE28776](#), respectively. These two histone modification marks have different distribution profiles, namely promoter-proximal enrichment with punctuated peaks for H3K4me3, and gene body enrichment with broad peaks for H3K36me3²³. The analysis was performed on a computer with Windows 7 Enterprise operating system (Microsoft Corporation, USA), 3.4 GHz CPU and 8 GB memory. [Figure 1](#) shows the GUI displayed upon launching ChIPdig¹⁵ from R Studio. The left pane displays four radio buttons, each corresponding to one of the four analysis modules, as well as a clickable box for selection of the folder containing the input files for the analysis.

Alignment of ChIP-seq reads

The aligning module of ChIPdig relies primarily on the QuasR package, which supports the analysis of single-read and paired-end deep sequencing experiments¹⁷. If desired, read preprocessing can be performed to prepare the input sequence files prior to alignment, e.g. removal of sequence segments corresponding to adapters and low-quality reads. Sequence files in FASTQ format for each input DNA and ChIP sample corresponding to the H3K4me3 and H3K36me3 ChIP-seq experiments were aligned to the WS220/ce10 reference genome (available through the BSgenome package¹⁶) ([Figure 2](#)), producing an average of 10 million mapped reads per sample. The execution time was approximately 2 h 30 min.

ChIP-seq data normalization, peak calling and differential enrichment analysis

A recurrent problem in ChIP-seq data analysis lies at the comparison of multiple coverage profiles generated from different experiments or corresponding to different conditions. This problem is addressed in the second analysis module of ChIPdig. Upon its selection, the user is prompted to provide a tab-delimited text file with the names of each ChIP sample (treatment) mapped reads file in BAM format, the matching input DNA (control) file, a sample ID, the condition or target name, and a color designation for the output peak and coverage files ([Figure 3a](#)). Any of the 657 R built-in colors can be chosen. A bin size parameter set to 50 bp (base pairs) by default is used in both peak calling and genome-wide library normalization for differential enrichment analysis. If desired, duplicate reads can be removed prior to normalization and sequences can be extended to a median fragment length that can be estimated either computationally or experimentally (e.g. average size of fragments generated by chromatin shearing) ([Figure 3b](#)). The initial mapped read processing outputs a table with library sizes and median fragment sizes, a chromosome size list, and a multidimensional scaling plot representing the similarity of samples in the data set ([Figure 3c](#)). Data are normalized based on library sizes and on the trimmed means of M values (TMM) approach²⁴, which is implemented in the edgeR package¹³. Initial mapped read processing for the H3K4me3 and H3K36me3 ChIP-seq data took approximately 10 min.

Completion of the initial processing of mapped reads releases options for downstream processing which are posted to the sidebar panel, namely export of files representing sequencing coverage, peak calling and differential enrichment analysis ([Figure 4](#)). Coverage is expressed in log₂-transformed counts per million (cpm) for each genomic bin and, for each sample ID indicated by the user, three files with bedGraph format are generated: treatment (ChIP sample), control (input DNA sample) and the treatment-to-control ratio. Each file can be loaded onto a genome browser for visualization ([Figure 5](#)). Generation of coverage files for H3K4me3 and H3K36me3 data took approximately 8 min.

Figure 1. ChIPdig user interface displayed upon launch. The left pane displays four radio buttons for selection of the analysis module and a clickable box for selection of the folder containing the input files. Additional features are sequentially unlocked as the user progresses through the software suite. Outputs generated throughout the analysis and additional instructions are placed in the main panel under a tab corresponding to the analysis module selected by the user.

ChIPdig

Figure 2. Alignment module. The user supplies a text file listing the input files corresponding to unmapped reads in either uncompressed (e.g.: '.fq', '.fastq') or compressed (e.g.: '.gz', '.bz2', '.xz') format. Both single-end and paired-end data are supported. A scroll-down menu displays the reference genome assemblies available for alignment.

Figure 3. Initial processing of mapped reads. The user supplies a tab-delimited text file listing the input DNA and ChIP sample mapped read files (a). If desired, duplicate read removal and fragment size extension can be performed (b). The application outputs a table with library sizes and median fragment sizes, a chromosome size list, and a multidimensional scaling plot representing the similarity of samples in the data set (c).

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:

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):

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

Select regions to limit differential enrichment analysis to (BED format) (please do not include column names):

Choose condition to test:

Choose reference condition:

Figure 4. Downstream processing of mapped reads. The user may generate coverage files in bedGraph format, perform peak calling for each sample, and assess differential enrichment for a set of genomic regions of interest.

Figure 5. H3K4me3 and H3K36me3 ChIP-seq coverage. Values represent log2-transformed counts per million (cpm) for the ChIP sample subtracted by log2-transformed cpm for the input sample. Files in bedGraph format were exported from ChIPdip and loaded onto the UCSC Genome Browser. These profiles exemplify the well-known scenario whereby H3K4me3 accumulates around the transcription start site and H3K36me3 is enriched at the gene body.

Peak calling is performed via the BayesPeak package by using a hidden Markov model and Bayesian statistical methodology¹⁸. The user specifies a posterior probability (PP) threshold and genomic regions with PP above such threshold are identified as peaks. If replicates are indicated, commonly enriched genomic regions representing replicated peaks can be derived, and a track definition line can be added, allowing each file to be loaded onto the UCSC genome browser for visualization. In addition, a consensus peak set representing all candidate enriched regions across the full data set may be obtained. The user may be interested in choosing this option for posterior differential enrichment analysis. Peak calling was performed for the H3K4me3 and H3K36me3 ChIP-seq data sets, yielding 4994 replicated peaks for H3K4me3 and 25855 replicated peaks for H3K36me3. Peak calling execution time was approximately 12 h.

Differential enrichment analysis resorts to functions implemented in the csaw¹² and edgeR¹³ packages. Following library size and TMM normalization, bin-level coverage is computed and, if desired, bins in which coverage for input DNA samples exceeds that of the corresponding ChIP samples are filtered off. If the user is interested in differential enrichment analysis in a specific set of regions, the corresponding file in BED format has to be provided. To illustrate this feature of ChIPdig, differential enrichment analysis was performed for comparing H3K4me3 coverage with that of H3K36me3 using either

H3K4me3 replicated peak coordinates (Figure 6a) or those corresponding to H3K36me3 peaks (Figure 6b), with a false discovery rate (FDR) threshold of 0.1. As expected, at H3K4me3 genomic peak coordinates, H3K4me3 coverage is greater than that of H3K36me3, and the opposite is observed for H3K36me3 peaks. Differential enrichment analysis for both comparisons took approximately 5 min.

Annotation of genomic regions

Annotation of genomic regions of interest is performed via the ChIPseeker package¹⁹. The user supplies the file with regions in BED format and chooses the reference genome assembly, as well as the distance upstream and downstream of the annotated transcription start site to be considered for assignment of promoter regions (Figure 7a). Replicated H3K4me3 and H3K36me3 peaks were annotated and, characteristically of these marks²³, most H3K4me3 peaks were assigned to promoters (Figure 7b), whereas H3K36me3 peaks lie predominantly at gene bodies (Figure 7c). The execution time for both annotation operations was approximately 3 min.

Comparative heatmaps and metaplots

The user can supply multiple coverage files in bedGraph format and generate heatmaps and metaplots to visualize coverage in a specific region set in a comparative manner. This module of ChIPdig relies on a custom algorithm which builds a coverage

Figure 6. Differential enrichment analysis results. H3K4me3 ChIP-seq coverage was compared with that of H3K36me3 using either (a) H3K4me3 peak genomic coordinates or those of (b) H3K36me3 peaks with a false discovery rate threshold of 0.1. For each analysis, a mean-difference plot representing the library size-adjusted log2-transformed fold change (the difference) against the average log2-transformed coverage (the mean), as well as a box-and-whisker plot showing the global change in normalized coverage between the two conditions, were originated.

matrix based on bin size and reference coordinates (start, end or both) selected by the user. Such matrix is then plotted in the form of a comparative metaplot and a set of heatmaps. H3K4me3 and H3K36me3 coverage files were loaded onto ChIPdig, along with a BED file with *C. elegans* transcription units in the WS220/ce10 reference assembly. Coverage was computed in 25 bp bins and gene bodies were either expanded or compressed to 500 bp. Windows of 250 bp upstream and downstream of each region were selected (Figure 8a). Typical of H3K4me3, coverage is higher in the vicinity of the transcription start site, whereas H3K36me3 is enriched at transcription unit bodies (Figure 8b, c). Heatmap and metaplot generation took approximately 4 min.

Comparison with publicly available tools available through the Galaxy framework

In the past decade, several tools have been made available through Galaxy, an interactive system that provides a simple

Web portal enabling users to analyze genomic data derived from high-throughput sequencing techniques²⁵. We have compared results obtained by ChIPdig with those obtained within Galaxy using Bowtie³ for read alignment and MACS2^{5,6} for peak calling. The percentage of mapped reads was higher for ChIPdig, but it should be noted that no read pre-processing was performed for read mapping in Galaxy (Figure 9a). For H3K4me3 peak calling, the number of called peaks using ChIPdig was similar to that in Galaxy, and, for H3K36me3, more peaks were called in ChIPdig than in Galaxy (Figure 9a). This can be attributed to the fact that the peak calling algorithm used by ChIPdig, and which is implemented in the BayesPeak package¹⁸, generates smaller peaks than those generated by Bowtie. More importantly, the spatial localization of ChIPdig-generated peaks overlapped very significantly with that of Galaxy-generated peaks (Figure 9b). We have also compared coverage tracks generated by ChIPdig (background-subtracted) with those generated by the Galaxy version of

Figure 7. Annotation of genomic regions. (a) The region file in BED format is uploaded and reference genome assembly, as well as distances upstream and downstream of transcription start site, are chosen. (b) Replicated H3K4me3 peaks are mostly promoter-proximal, whereas (c) H3K36me3 peaks are found predominantly at gene bodies.

Figure 8. Generation of heatmaps and comparative metaplot for visualizing coverage. Coverage files for H3K4me3 and H3K36me3 were supplied to ChIPdig, as well as a file with coordinates of *C. elegans* transcription units. (a) The heatmaps and metaplot were generated by considering a 25-bp bin size, and gene bodies were resized to 500 bp. Windows of 250 bp upstream and downstream of each region were selected. As observed in both (b) the heatmaps and (c) the comparative metaplot, coverage for H3K4me3 is overall higher in the vicinity of transcription start sites, while that of H3K36me3 is enriched at transcription unit bodies.

Figure 9. Comparison of ChIPdig analysis results with those obtained through the Galaxy framework. (a) The percentage of mapped reads using the alignment module of ChIPdig is higher than that using Bowtie within Galaxy, but it should be noted that read pre-processing was not included in the latter analysis. Peak calling in ChIPdig and in Galaxy resulted in similar numbers of peaks called for H3K4me3, and more H3K36me3 peaks were called using ChIPdig compared with Galaxy-generated peaks. (b) Importantly, ChIPdig-generated peaks overlapped very significantly with Galaxy-generated peaks. (c) Coverage tracks generated by ChIPdig have a profile similar to those generated by the Galaxy version of DeepTools.

DeepTools and observed that the corresponding profiles are very similar (Figure 9c). These observations indicate that ChIP-seq data analysis using ChIPdig compares favorably with that in Galaxy.

Conclusions

ChIPdig¹⁵ is a user-friendly application for handling multiple ChIP-seq data sets and has diverse useful capabilities spanning a comprehensive analysis pipeline, namely: read alignment to a reference genome, ChIP-seq data normalization, peak calling, differential enrichment analysis, annotation of genomic regions, and generation of comparative heatmaps and metaplots for visualization of normalized coverage.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Source code available from: <https://github.com/rmesse/ChIPdig>.

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.3345788>¹⁵.

License: GNU General Public License 3.0.

Grant information

This research was supported by the GM107056 R01 grant awarded by the National Institutes of Health to Dr. Alla Grishok.

Acknowledgements

The author would like to acknowledge Dr. Alla Grishok (Department of Biochemistry, Boston University School of Medicine, Boston, MA, USA) for appraising the work at its earliest stages.

References

- Davis CA, Hitz BC, Sloan CA, *et al.*: **The Encyclopedia of DNA elements (ENCODE): data portal update.** *Nucleic Acids Res.* 2018; **46**(D1): D794–D801.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Langmead B: **Aligning short sequencing reads with Bowtie.** *Curr Protoc Bioinformatics.* 2010; **Chapter 11**: Unit 11.7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhang Y, Liu T, Meyer CA, *et al.*: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol.* 2008; **9**(9): R137.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Feng J, Liu T, Qin B, *et al.*: **Identifying ChIP-seq enrichment using MACS.** *Nat Protoc.* 2012; **7**(9): 1728–40.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zang C, Schones DE, Zeng C, *et al.*: **A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.** *Bioinformatics.* 2009; **25**(15): 1952–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Meyer CA, Liu XS: **Identifying and mitigating bias in next-generation sequencing methods for chromatin biology.** *Nat Rev Genet.* 2014; **15**(11): 709–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Liang K, Keles S: **Detecting differential binding of transcription factors with ChIP-seq.** *Bioinformatics.* 2012; **28**(1): 121–2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shao Z, Zhang Y, Yuan GC, *et al.*: **MANorm: a robust model for quantitative comparison of ChIP-Seq data sets.** *Genome Biol.* 2012; **13**(3): R16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lun AT, Smyth GK: **csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows.** *Nucleic Acids Res.* 2016; **44**(5): e45.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–40.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Anders S, Huber W: **Differential expression of RNA-Seq data at the gene level – the DESeq package.** *EMBL, Heidelberg, Ger.* 2012.
[Reference Source](#)
- rnesse/ChIPdip: **First release of ChIP-dip (Version v1.0.0).** *Zenodo.* 2019.
<http://www.doi.org/10.5281/zenodo.3345788>
- Bonhoure N, Bounova G, Bernasconi D, *et al.*: **An introduction to the Biostrings / BSgenome framework.** *Nature.* 2014; **11**(1): 1–13.
- Gaidatzis D, Lerch A, Hahne F, *et al.*: **QuasR: Quantification and annotation of short reads in R.** *Bioinformatics.* 2015; **31**(7): 1130–2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cairns J, Spyrou C, Stark R, *et al.*: **BayesPeak—an R package for analysing ChIP-seq data.** *Bioinformatics.* 2011; **27**(5): 713–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yu G, Wang LG, He QY: **ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization.** *Bioinformatics.* 2015; **31**(14): 2382–3.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gu Z: **EnrichedHeatmap: Making Enriched Heatmaps.** R package version 1.6.0. 2017.
[Publisher Full Text](#)
- Lawrence M, Huber W, Pagès H, *et al.*: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol.* 2013; **9**(8): e1003118.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Riemondy KA, Sheridan RM, Gillen A, *et al.*: **valr: Reproducible genome interval analysis in R [version 1; peer review: 2 approved].** *F1000Res.* 2017; **6**: 1025.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Liu T, Rechtsteiner A, Egelhofer TA, *et al.*: **Broad chromosomal domains of histone modification patterns in *C. elegans*.** *Genome Res.* 2011; **21**(2): 227–36.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol.* 2010; **11**(3): R25.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Giardine B, Riemer C, Hardison RC, *et al.*: **Galaxy: A platform for interactive large-scale genome analysis.** *Genome Res.* 2005; **15**(10): 1451–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 25 February 2020

<https://doi.org/10.5256/f1000research.21986.r59875>

© 2020 Carroll T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thomas Carroll

Bioinformatics Resource Center, The Rockefeller University, New York City, NY, USA

The author presents a R based toolset for the analysis of ChIP-seq data in a GUI framework. The construction of R based ChIP-seq analysis pipelines affords the potential for the use of wide range of tools from R and Bioconductor libraries while offering a low dependency piece of software.

ChIPdig uses QuasR, a wrapper for Bowtie, for the alignment of ChIP-seq data from a BSgenome object. The Rbowtie2 and Rsubread packages are now both available on Windows, Mac and Linux systems and should be considered alongside Bowtie. I believe they would offer significant speed and memory usage improvements over QuasR. Although these do not accept BSgenome objects, ChIPdig could easily generate the FASTA from these packages for use with indexing steps of both packages.

Blacklisted regions should be considered in this tool as they have been shown to have strong effects on the QC, fragment length estimation and between sample normalisation. Inclusion of methods of blacklist filtering from known sources (such as Encode) or in software derived blacklists (using GreyListChIP) should be performed.

The output of BedGraph instead of BigWigs may cause some problems for users when working with larger genomes such as human or mouse. BigWigs may not be able to be exported on Windows systems but users of Mac or linux should have this option available to them to make this feature worthwhile.

Peak calling is performed with BayesPeak. It is unclear how this performs on the different types of epigenetic marks used in this study. Some more options for peak calling could be included here to allow finer control of the stitching of peaks into larger peaks. A simple bin based peak calling approach such as implemented in the CSAW user guide would be useful here. How the identification of replicated peaks is not clear in text and could be expanded.

The example differential enrichment analysis compare H3k4me3 and H3K36me3 signals directly. This is a strange example as most differential ChIP-seq analysis is performed within the same

antibody. An example comparing the change in one histone mark over different conditions/treatment/tissue types would be a more useful and relevant comparison.

This differential enrichment example does highlight a potential pitfall with this approach where the majority of sites change. The user should be warned in these circumstances as conclusions are likely to be invalid. Alternative normalisations such as to total mapped reads in peaks or total mapped reads to genome could be provided as options (as in Diffbind).

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, ChIP-seq, ATAC-seq, Splicing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 12 November 2019

<https://doi.org/10.5256/f1000research.21986.r55692>

© 2019 Gronemeyer H et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hinrich Gronemeyer

Functional Genomics and Cancer, Institute of Genetics and of Molecular and Cellular Biology (IGBMC), CNRS, UMR 7104, French Institute of Health and Medical Research (INSERM), U 1258,

University of Strasbourg, Illkirch, France

Marco Antonio Mendoza-Parra

10UMR 8030 Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, University of Evry-val-d'Essonne, University Paris-Saclay, Évry, France

Technological advancement over the past two decades has generated a scientific landscape that has shifted from gene-centric to genome-centric research. As consequence, we are facing an ever-increasing amount of (functional) genomics datasets in publicly available repositories, which is in principle an extremely rich asset of knowledge. Indeed, any research project should start with an analysis of the existing data and analyzed the obtained results in view of this - sometimes hidden - information. However, the main caveats are presently (i) the availability of simple user-friendly computational biology tools to query, analyze and integrate the existing data and (ii) the absence of quality assessment indicators attached to each data set.

The author of the present article has addressed the first issue for assays involving chromatin immunoprecipitation followed by sequencing of the co-IPed DNA (ChIP-seq). This is widely used for mapping the cistromes of transcription factors (TFs) and other chromatin binding factors and for monitoring DNA and functional histone modifications along the genome. He describes the "ChIPdig" tool, an R pipeline for ChIP-seq data analysis from the alignment of sequenced reads to the differential analysis of multiple datasets. The goal of this pipeline is fully justified, namely to provide users with a single pipeline that covers most of the required analytical steps without the need of jumping between tools available in different programming languages.

Unfortunately, there are a number of issues which need to be addressed in order to make this tool attractive compared to others.

Conceptual issues. The author has assembled a pipeline which uses other R packages (described in 'Implementation') but does not consider their limitations.

1. **Peak calling.** ChIP-seq patterns reflect the specifics of TF binding and DNA/histone modifications, which can be highly variable. In simple terms ChIP-seq signals can be 'sharp' (e.g., TFs, H3K4me3) or broad (e.g., H3K27me3, H3K36me3). The various existing peak callers have very different characteristics and cannot be used without considering these aspects¹.
2. **Normalization.** ChIPdig uses the simplest way of normalizing data sets according to total mapped reads (TMRs). This is not state-of-the-art and can lead to problems when important differences exist between TMRs of different datasets. Thus, methodologies like quantile normalization perform significantly better^{2,3}.
3. **Comparison** of H3K4me3 and H3K36me3 (Figs. 5, 6). While this is an illustrative example based on current knowledge in the field, the data comparison *per se* can be criticized, as very different types of signals (sharp H3K4me3 and broad H3K36me3 signals) are compared using the same peak caller and sub-optimal normalization. In Fig. 5 it appears that the control (input?) is simply subtracted (after normalization?) resulting in negative and positive (peaks/landscapes). It thus appears that the variation between the input and H3K36me3 signal intensity profiles is as high as the signals shown for H3K36me3. In Fig. 6 the value of these displays for functional interpretation is not clear.

4. **ChIP-seq artifacts.** The author addresses correctly the potential artifacts that can be obtained for a multiplicity of reasons in ChIP-seq experiments. Unfortunately, however, he does not offer solutions in the context of the ChIPdig pipeline. This is one of the most important issues concerning ChIP-seq data sets, as significant number of datasets cannot be used due to quality issues and very few tools incorporate quality assessment (e.g. qcGenomics; <http://www.ngs-qc.org/qcgenomics/>).
5. **Challenging examples.** The examples that have been chosen to illustrate ChIPdig are the 'easy' ones, where good antibodies and thus, clean data sets exist. However, ChIPdig should be challenged with more 'difficult' TFs/histone marks and 'kinky' input profiles.
6. **Technical issues.** The main hindrance for non-bioinformatics trained scientists to exploit existing databases is the need for time-consuming data download and re-formatting. Currently only the qcGenomics database contains pre-computed data sets allowing for immediate access.
7. **Computational time.** In addition to downloading datasets for analysis, users face as main technical disadvantages of ChIPdig the time for computation. As stated by the authors, alignment of 10 million reads take 2.5 hrs. This is for the *C. elegans* genome, so it would be far too long for the human genome.
8. **Alignment.** It appears that the user has no choice on the alignment parameters and multiple alignment tools.
9. **Organisms.** A description is missing in the documentation whether ChIPdig is currently only available for *C. elegans* or whether it can be used for other reference genomes. In the present drop-down box, only *C. elegans* is available.
10. **Memory requirements.** Trials with an 8Gb PC to use human ChIP-seq raw fastq files for trimming, etc. gave error messages and the program aborted.

References

1. Mendoza-Parra MA, Nowicka M, Van Gool W, Gronemeyer H: Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. *BMC Genomics*. 2013; **14**: 834 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Mendoza-Parra MA, Sankar M, Walia M, Gronemeyer H: POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. *Nucleic Acids Res*. 2012; **40** (4): e30 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Saleem MM, Mendoza-Parra MA, Cholley PE, Blum M, et al.: Epimetheus - a multi-profile normalizer for epigenomic sequencing data. *BMC Bioinformatics*. 2017; **18** (1): 259 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Functional genomics and cancer

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 07 October 2019

<https://doi.org/10.5256/f1000research.21986.r53175>

© 2019 Liu T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tao Liu

Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA

Although the author explained the rationale of building the tool ChIPdig, I am not convinced why the author selected the exact tools to be included-- QuasR for alignment, BayesPeak for peak calling, differential calling with csaw+edgeR, ChIPseeker for genomic region annotation. To make a pipeline with pure R packages is an interesting idea. However, the end-users will miss the flexibility to choose tools they want -- usually the most popular tools in the field, compared with using the Galaxy or Snakemake framework.

The use case presented is only based on modENCODE C.elegans ChIP-seq datasets published in 2011, so I wonder how the pipeline can be scaled up to deal with bigger datasets or data from a larger genome such as human data (author can use recent ENCODE data as an example). Without more use cases on recent data, the tool seems yet another ChIP-seq pipeline tailored for reanalyzing specific datasets.

Lastly, ChIPdig uses QuasR qAlign function to align reads to reference genome, and qAlign is, in fact, a wrapper on Bowtie, which is included in Galaxy pipeline that the author compared. Therefore the difference of mapping results in figure 9a maybe just due to parameter settings. However, such information -- how the author generated Galaxy results -- is missing in the manuscript.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics algorithm development on NGS datasets

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research