A peer-reviewed version of this preprint was published in PeerJ on 26 April 2016.

Chabbert CD, Steinmetz LM, Klaus B. 2016. *DChIPRep*, an R/Bioconductor package for differential enrichment analysis in chromatin studies. PeerJ 4:e1981 https://doi.org/10.7717/peerj.1981

DChIPRep, an R/Bioconductor package for differential enrichment analysis in chromatin studies

Christophe D Chabbert, Lars M Steinmetz, Bernd Klaus

The genome–wide study of epigenetic states requires the integrative analysis of histone modification ChIP–seq data. Here, we introduce an easy–to–use analytic framework to compare profiles of enrichment in histone modifications around classes of genomic elements, e.g. transcription start sites (TSS). Our framework is available via the user–friendly R/Bioconductor package *DChIPRep*. *DChIPRep* uses biological replicate information as well as chromatin Input data to allow for a rigorous assessment of differential enrichment. *DChIPRep* is available for download through the Bioconductor project at http://bioconductor.org/packages/*DChIPRep.* **Contact** DChIPRep@gmail.com

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016

DChIPRep, an R/Bioconductor package for differential enrichment analysis in chromatin studies

Christophe D. Chabbert[1], Lars M. Steinmetz[2], and Bernd Klaus[3]

[1]**European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany**
Current address: Astra Zeneca, Oncology iMed, CRUK-Cambridge Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom.
[2]**European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany,**
Stanford Genome Technology Center, Palo Alto, CA 94304, USA and
Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]**European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany, corresponding author, Email: bernd.klaus@embl.de**

ABSTRACT

The genome–wide study of epigenetic states requires the integrative analysis of histone modification ChIP–seq data. Here, we introduce an easy–to–use analytic framework to compare profiles of enrichment in histone modifications around classes of genomic elements, e.g. transcription start sites (TSS). Our framework is available via the user–friendly R/Bioconductor package *DChIPRep*. *DChIPRep* uses biological replicate information as well as chromatin Input data to allow for a rigorous assessment of differential enrichment. *DChIPRep* is available for download through the Bioconductor project at http://bioconductor.org/packages/DChIPRep
Contact. DChIPRep@gmail.com

Keywords: Bioinformatics, Computational Biology, Genomics, ChiP–Seq, Chromatin, Histone–Modifications

INTRODUCTION

The elementary component of eukaryotic chromatin, the nucleosome, is composed of 147bp DNA fragments wrapped around an octamer comprising two copies of 4 of the histone proteins. The N–terminal tails of these proteins are subject to multiple post–translational modifications (PTM) including acetylation, phosphorylation and methylation. Recent studies have highlighted the importance of these PTM in key cellular processes such as transcription, DNA replication and repair. Protocols based on chromatin immunoprecipitations followed by deep sequencing (ChIP–seq) allow for a genome–wide mapping of these modifications. Such endeavors have resulted in the generation of complex sequencing datasets that require appropriate bioinformatics tools to be analyzed. From this data, profiles of enrichment in histone modifications around classes of genomic elements, e.g. transcription start sites (TSS) are routinely computed. Once these enrichment profiles have been obtained, a common analysis task is to compare them between experimental conditions. However, due to a lack of tools tailored to the assessment of differential enrichment, these comparisons are often performed in a purely descriptive manner (e.g. by comparing plots of enrichment profiles around transcription start sites). In this article, we present a workflow to assess differential enrichment in a statistically rigorous way. This workflow is implemented in a user–friendly package named *DChIPRep* that is available via the Bioconductor project (Huber et al., 2015).

Review of existing tools and approaches

Several software tools designed to analyze certain aspects of histone modification data are already available. They mostly focus on the genome–wide determination of nucleosome positions and on the identification of genomic loci enriched in the modifications of interest. Diverse statistical and numerical approaches have been concurrently implemented, including Fourier transform (*nucleR*, Flores and Orozco, 2011), Gaussian filtering (*Genetrack*, Albert et al., 2008), wavelets (*NUCwave*, Quintales et al., 2014) as well as probabilistic or Bayesian approaches (*NucleoFinder* Becker et al., 2013, *PING 2.0* Woo et al., 2013, *NOrMAL* Polishko et al., 2012).

Some algorithms proposed recently go beyond the determination of nucleosome positions and aim at assessing differential enrichment. However, they commonly rely on the identification of regions of interest (e.g. around called peaks) using the ChIP–seq datasets themselves e.g. *DiffBind*, (Ross-Innes et al., 2012; Stark and Brown, 2011). Notably, *csaw* (Lun and Smyth, 2014) allows for a genome wide identification of differential binding events without an a priori specification of regions of interest. It uses a windowing approach and implements strategies for a post hoc aggregation of significant windows into regions. peer However, to the best of our knowledge, no direct approach to compare enrichment profiles of histone modifications around classes of genomic elements exists so far. Furthermore, most existing tools do not offer the possibility to directly correct for biases using the Input chromatin samples. Commonly, these profiles are analyzed in a purely descriptive manner and conclusions are drawn solely from plots of metagene/metafeature (e.g. transcription start site plots).

Here we present *DChIPRep*, an R/Bioconductor package designed to compute and compare histone modification enrichment profiles from ChIP–seq datasets at nucleotide resolution. The workflow implemented in *DChIPRep* uses both the biological replicate and the chromatin Input information to assess differential enrichment. By adapting an approach for the differential analysis of sequencing count data (Love et al., 2014), *DChIPRep* tests for differential enrichment at each nucleotide position of a metagene/metafeature profile and determines positions with significant differences in enrichment between experimental groups. An overview of the complete workflow is given next.

Overview of the implemented framework

The framework implemented in *DChIPRep* consists of three main steps:

1. The chromatin Input data is used for positionwise–normalization.
2. The methodology of Love et al. (2014) is used to perform positionwise testing. A minimum $\log_2$–fold–change greater than zero is set during the testing procedure to ensure that called positions show an non–spurious differential enrichment.
3. Finally, in order to assess statistical significance, local False Discovery Rates (local FDRs, Strimmer, 2008) are computed from the p–values obtained as a result of the testing step.

Real data analysis

We apply *DChIPRep* and a modified version of its framework using methodology inspired by the *csaw* and *edgeR* (Lun and Smyth, 2014; McCarthy et al., 2012) packages to yeast MNase–seq data and compare the enrichment profiles around TSS in wild–type and mutant strains, demonstrating how our package can derive biological insights from large–scale sequencing datasets.

PACKAGE OVERVIEW

General architecture

DChIPRep uses a single class `DChIPRepResults` that wraps the input count data and stores all of the intermediate computations. The testing and plotting functions are then implemented as methods of the `DChIPRepResults` object. The plotting functions return *ggplot2* (Wickham, 2009) objects than can subsequently be modified by the end–user.

DChIPRep's analytical method uses histone modification ChIP–Seq profiles at single nucleotide resolution around a specific class of genomic elements (e.g. annotated TSS). In the case

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016

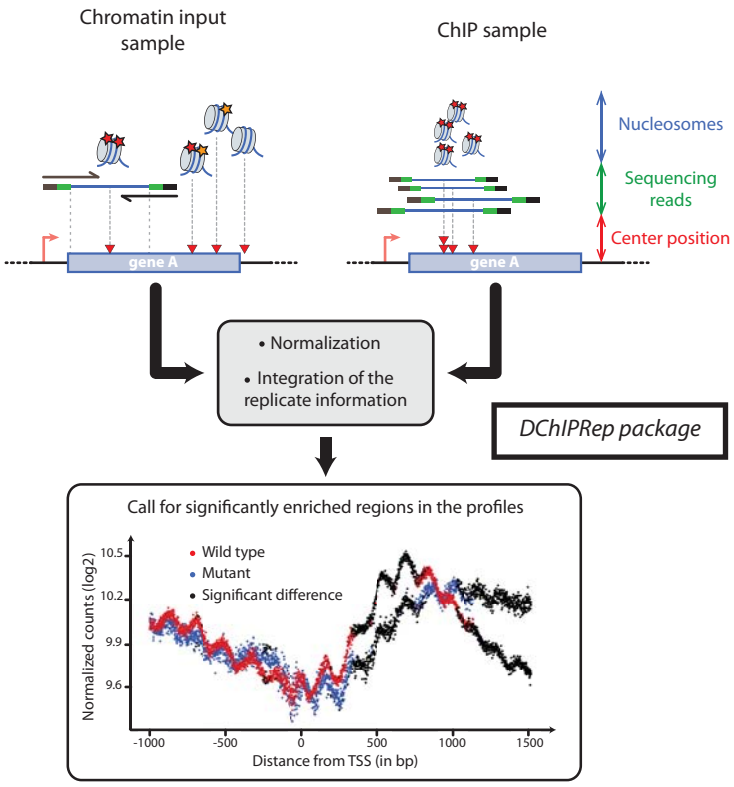

Figure 1. Illustration of the *DChIPRep* workflow. Chromatin Input– and ChIP–data are analyzed jointly and positions showing significantly different enrichment are identified using the replicate information.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016

of paired–end MNAse–seq reads, such profiles can be obtained using the middle position of the genomic interval delimited by the DNA fragments (Fig. 1).

Thus, the variables characterizing the samples are the genomic positions relative to a specific class of genomic elements (e.g. TSS). These variables take the values given by the number of sequenced fragments with their center at these specific positions. The data is summarized across genomic features (e.g. genes or transcripts) at each of these nucleotide positions, so that metagene/metafeature profiles are obtained. The input data for DChIPRep can be alignment files in the SAM format or already processed count data.

Data import

DChIPRep has two possible data input formats. The input data can be two count tables per sample (for ChIP and Input), with the genomic features used (e.g. genes or transcripts) in the rows and the position wise counts per genomic feature in the columns. Alternatively, one can provide two count tables for ChIP and Input that contain the data at the metafeature level, such that the data is summarized across individual genomic features. These tables then have one row per position relative to the genomic element (e.g. TSS) studied and one column per sample.

DChIPRep can either import tab-separated .txt files (two files per experimental sample with data at the level of the individual genomic features) or two R count matrices for ChIP and Input data, which contain the data already summarized at a metafeature level (summarized across features per position). A table containing the experimental conditions and other sample specific annotation is needed as well. A Python script (DChIPRep.py) is also provided along with the package to generate suitable tab-separated input files from SAM alignment files and a gff annotation. The script may be customized via multiple parameters.

Further details on the data import can be found in the package vignette, which is available via Bioconductor and as Supplemental_Item_1_vignette_DChIPRep.html.

Computation of the metafeature profile

Once the data is summarized on the feature level (i.e. count tables), we can compute the metafeature profiles with the function `summarizeCountsPerPosition` for each of the ChIP–Seq and chromatin Input samples.

The function first filters out features with very low counts. Then, in order to summarize the data across features, a trimmed mean of the counts at each position is computed.

Finally, these positionwise mean values are multiplied by the number of features retained at each position. This way, a raw metafeature profile for each individual sample is obtained.

Call for enriched regions

The statistical approach implemented in *DESeq2* is used to call for significantly deferentially enriched positions (Love et al., 2014).

Here, the chromatin input is used to compute normalization factors that correct for potential local biases in chromatin solubility, enzyme accessibility or PCR amplification. After specifying a minimum fold change, Wald tests are performed to assess significant changes in the metagene/metafeature profiles.

Finally, local FDRs estimated by the `fdrtool` (Strimmer, 2008) package are used to assess statistical significance based on the p–values obtained from the Wald–test.

All of these steps are implemented in the `runTesting` function (Fig. 1).

Plotting functions

DChIPRep provides two plotting functions to represent and inspect the final results of the analysis. The `plotProfiles` function summarizes the biological replicates by taking a position wise mean and then plots a smoothed enrichment profile around the genomic element class of interest (e.g. TSS).

The `plotSignificance` function plots the unsmoothed enrichment profile and highlights positions with a significant difference in enrichment as returned by the runTesting function (Fig. 1). The plotting functions return *ggplot2* objects that can be easily customized.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016

A CASE STUDY

We applied *DChIPRep* to a paired–end MNAse–seq dataset for which biological replicates are available (Chabbert et al., 2015). Using the annotation from Xu et al., 2009, we compared the enrichment of the H3K4me2 mark in annotated ORFs (5170 items) in the wild type strain of Saccharomyces cerevisiae and the *set2*Δ mutant. We have called a significant enrichment (local FDR < 0.2) in the mutant for 906 positions located within 1500bp downstream of the transcription start site (Fig. 1).

Analysis steps for the case study

In order to illustrate the usage of *DChIPRep* we document the series of simple commands that are needed to be to run a typical analysis.

After the data has been preproccesed, we first need to import a table that contains the annotation information for our samples. This table contains information on the count table file names and the desired number of up– and downstream positions to be compared, as well as the experimental group a sample belongs to. As mentioned above, details on the required format of the annotation table can be found in the package vignette in the supplement (Supplemental_Item_1_vignette_DChIPRep.html).

We can then import the data using the function `importData`.

Listing 1. Data Import

```
sampleTable_K4me2 <- read.csv("sampleTable_K4me2.csv")
importedData <- importData(sampleTable_K4me2)
```

After then data import, we can perform the positionwise testing with the `runTesting` function, extract the results using the `resultsDChIPRep` function and finally obtain the significance plot in Fig. 1 via a call to the `plotSignificance` function.

Listing 2. Results and Firgure

```
testResults <- runTesting(importedData)
testResults <- resultsDChIPRep(testResults)
plotSignificance(testResults)
```

A comparison to an *csaw*/*edgeR*–based pipeline

The framework implemented in *DChIPRep* uses the *DESeq2*–package (Love et al., 2014) to perform the statistical testing. The *csaw*–package (Lun and Smyth, 2014) implements a strategy based on methods implemented in *edgeR* (McCarthy et al., 2012) to assess differential binding in ChIP–Seq data sets genome–wide. While *csaw* and *DChIPRep* are not directly comparable, we can adapt the *csaw* framework to assess the differential enrichment (for a summary of the *csaw* framework, see section 1.3. of the *csaw* user guide at Bioconductor).

Specifically, we used the log–normalization factors computed from the chromatin–input as offsets for the GLM–model and then applied the quasi–likelihood (QL) methods of Lund et al. (2012) to perform a dispersion shrinkage and an appropriate F–test to assess the differential enrichment. Note that since *edgeR* does not allow for an a priori specification of a fold change threshold, we had to specify it post hoc. The complete analysis can be found in supplementary file 2 – ReproduceFiguresDChIPRepPaper.zip.

Figure 2 shows the results of this approach. The modified pipeline identified 1127 positions as significantly deferentially enriched located within 1500bp downstream of the transcription start site. Comparing Fig. 2 to Fig. 1, we see that *DChIPRep* identifies deferentially enriched regions more consistently, while the *edgeR*–based pipeline often calls positions with relatively small fold changes as significant. This might be due to the fact that a post hoc fold change thresholding had to be performed. *DChIPRep* would therefore be less prone to calling false positive as it is less sensible to weak enrichment (which might be resulting from intrinsic variability in the performance of immunoprecipitation for example).

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016

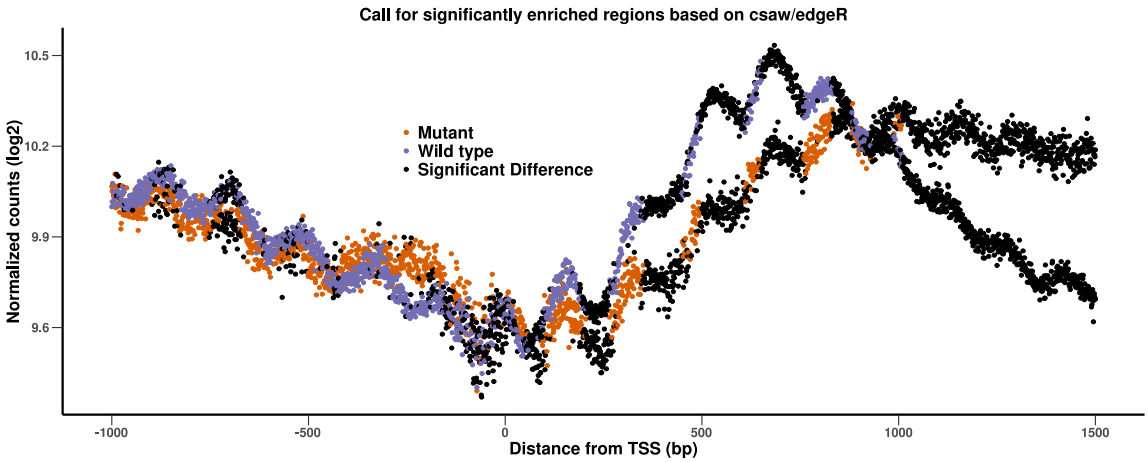

Figure 2. Results of the *csaw/edgeR*–based calling for enriched regions. We applied an *edgeR*–based testing to the data (instead of using *DESeq2*). This included a post hoc thresholding of the fold–changes. The figure shows that this pipeline often calls positions with moderate fold–changes as significant in our example data set.

Reproducible research

The complete code and the data used for the case study can be found in the supplementary material (Supplemental_Item_2_ReproduceFiguresDChIPRepPaper.zip).

DISCUSSION AND CONCLUSION

The package *DChIPRep* provides an integrated analytical framework for the computation and comparison of enrichment profiles from replicated ChIP–seq datasets at nucleotide resolution.

Starting from the primary alignment of paired–end reads, the software allows a rapid identification of significantly differentially enriched positions relative to classes of genomic elements and provides straightforward plotting of the enrichment profiles.

We also applied the *DChIPRep*–package to a published data set. This case study demonstrates *DChIPRep*'s favourable performance when compared to a pipline inspired by the *csaw*–package for differntial binding analysis.

Acknowledgements

We thank Sophie Adjalley, Vicent Pelechano, Aleksandra Pekowska and Alejandro Reyes for helpful discussions and critical comments on the manuscript.

Funding

This work was supported by a PhD fellowship from Boerhinger Ingelheim Fonds [to C.D.C]; and the Deutsche Forschungsgemeinschaft [1422/3-1 to L.M.S.].

REFERENCES

Albert, I., Wachi, S., Jiang, C., and Pugh, B. F. (2008). GeneTrack–a genomic data processing and visualization framework. *Bioinformatics*, 24(10):1305–1306.

Becker, J., Yau, C., Hancock, J. M., and Holmes, C. C. (2013). NucleoFinder: a statistical approach for the detection of nucleosome positions. *Bioinformatics*, 29(6):711–716.

Chabbert, C. D., Adjalley, S. H., Klaus, B., Fritsch, E. S., Gupta, I., Pelechano, V., and Steinmetz, L. M. (2015). A high-throughput ChIP-seq for large-scale chromatin studies. *Molecular Systems Biology*, 11(1):777–777.

Flores, O. and Orozco, M. (2011). nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, 27(15):2149–2150.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016

Huber, W., Carey, J., V., Gentleman, R., Anders, S., Carlson, M., Carvalho, S., B., Bravo, C., H., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, D., K., Irizarry, A., R., Lawrence, M., Love, I., M., MacDonald, J., Obenchain, V., Ole's, K., A., Pag'es, H., Reyes, A., Shannon, P., Smyth, K., G., Tenenbaum, D., Waldron, L., Morgan, and M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12):550.

Lun, A. T. L. and Smyth, G. K. (2014). De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Research*, 42(11):e95–e95.

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5).

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

Polishko, A., Ponts, N., Roch, K. G. L., and Lonardi, S. (2012). NORMAL: accurate nucleosome positioning using a modified gaussian mixture model. *Bioinformatics*, 28(12):i242–i249.

Quintales, L., Vázquez, E., and Antequera, F. (2014). Comparative analysis of methods for genome-wide nucleosome cartography. *Brief Bioinform*, 16(4):576–587.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C., and Carroll, J. S. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*.

Stark, R. and Brown, G. (2011). Diffbind: differential binding analysis of chip-seq peak data. *Bioconductor - http://bioconductor.org/packages/release/bioc/html/DiffBind.html*.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Woo, S., Zhang, X., Sauteraud, R., Robert, F., and Gottardo, R. (2013). PING 2.0: an r/bioconductor package for nucleosome positioning using next-generation sequencing data. *Bioinformatics*, 29(16):2049–2050.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L. M. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1723v2 | CC-BY 4.0 Open Access | rec: 9 Feb 2016, publ: 9 Feb 2016