*Bioinformatics*, 31(14), 2015, 2382–2383
doi: 10.1093/bioinformatics/btv145
Advance Access Publication Date: 11 March 2015
Applications Note

OXFORD

Genome analysis

# ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization

**Guangchuang Yu[1,2,*], Li-Gen Wang[3] and Qing-Yu He[1,*]**

[1]Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, College of Life Science and Technology, Jinan University, Guangzhou 510632, China, [2]State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China and [3]Guangdong Information Center, Guangzhou 510031, China

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on February 2, 2015; revised on March 4, 2015; accepted on March 6, 2015

## Abstract

**Summary:** ChIPseeker is an R package for annotating ChIP-seq data analysis. It supports annotating ChIP peaks and provides functions to visualize ChIP peaks coverage over chromosomes and profiles of peaks binding to TSS regions. Comparison of ChIP peak profiles and annotation are also supported. Moreover, it supports evaluating significant overlap among ChIP-seq datasets. Currently, ChIPseeker contains 15 000 bed file information from GEO database. These datasets can be downloaded and compare with user's own data to explore significant overlap datasets for inferring co-regulation or transcription factor complex for further investigation.
**Availability and implementation:** ChIPseeker is released under Artistic-2.0 License. The source code and documents are freely available through Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html).
**Contact:** guangchuangyu@gmail.com or tqyhe@jnu.edu.cn
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) has become standard technology to investigate genome wide binding sites of transcription factors (TFs). To analyze these DNA binding sites to determine their potential regulatory functions, proximal genes/transcripts and regulatory elements are considered to annotate these loci. Several applications were developed for annotating ChIP-seq data. ChIPpeakAnno (Zhu *et al.*, 2010) is an R package that designed for ChIP-seq and ChIP-chip data annotation. Because ChIPpeakAnno does not consider strand information, it misclassifies peaks with wrong orientation (see Supplemental File). Other tools including PAVIS (Huang *et al.*, 2013) and PeakAnalyzer (Salmon-Divon *et al.*, 2010), they all have similar functions of annotating peaks by nearest genes and proximal genomic features. Existing tools are all designed for single dataset. ChIP-seq is rapidly becoming a common technique and there are a large number of datasets available in the public domain. Results from individual experiments provide a limited understanding of chromatin interactions, as there is many factors cooperate to regulate transcription. Existing software lack of functionalities for comparing profiles of ChIP-seq datasets in different levels, including profiles of peaks binding to TSS regions, annotation and enriched functional profiles.

There are increasing evidences shown that combinations of TFs are important for regulating gene expression (Perez-Pinera et al., 2013; Zhu et al., 2008). However, systematically identification of TF interactions by ChIP-seq is still not available. Even if a specific TF binding is essential for a particular regulation was known, we do not have prior knowledge of all its co-factors. There are no systematic strategies available to identified un-known co-factors by ChIP-seq. To fill this gap, we present an R/Bioconductor package, ChIPseeker, for ChIP peak annotation, comparison and visualization. It incorporates statistical testing of co-occurrence of difference ChIP-seq datasets and can be used to identify co-factors by exploring publicly available ChIP-seq datasets.

## 2 Functions

ChIPseeker implements *annotatePeak* function for annotating peaks with nearest gene and genomic region where the peak is located. Almost all annotation software calculates the distance of a peak to the nearest TSS and annotates the peak to that gene. This can be misleading as binding sites might be located between two start sites of different genes or hit different genes, which have the same TSS location in the genome. The *anntatePeak* function provides parameters to annotate genes with a max distance cutoff and all genes within this distance will be reported for each peak. For annotating genomic region, *annotatePeak* function reports detail information when genomic region is Exon or Intron. For instance, 'Exon (uc002sbe.3/9736, exon 69 of 80)', means that the peak is overlaps with the 69th exon of the 80 exons that transcript uc002sbe.3 possess and the corresponding Entrez gene ID is 9736. ChIPseeker provides *plotAnnoBar*, *plotAnnoPie*, *vennpie* and *plotDistToTSS* functions to visualize ChIP peak annotation. *plotAnnoBar* and *plotDistToTSS* functions support comparing among different ChIP-seq data. ChIPseeker supports annotating ChIP-seq data of a wide variety of species if they have transcript annotation *TxDb* object available. Users can generate a *TxDb* object from UCSC Genome Browser, BioMart database or GFF/GTF3 file by *makeTranscript DbFromUCSC*, *makeTranscriptDbFromBiomart* and *makeTran scriptDbFromGFF* functions respectively. These functions are available via the GenomicFeatures package (Lawrence *et al.*, 2013).

For functional enrichment analysis, ChIPseeker works fine with in house developed Bioconductor package, clusterProfiler (Yu *et al.*, 2012), DOSE (Yu *et al.*, 2015) and ReactomePA. Once the peaks are annotated, users can use these packages to identify predominant biological themes through Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, Disease Ontology and Reactome Pathway. Functional profile comparison is also supported by clusterProfiler.

ChIPseeker provides *covplot* to visualize the peak locations and intensities over the whole genome. The *plotAvgProf2* function visualizes the average profile of ChIP peaks binding to TSS region, while *peakHeatmap* function generate a heatmap of ChIP peaks binding to TSS regions. These two functions can accept a list of bed files and plot profiles and heatmaps among several ChIP experiments, making it easy to compare among different ChIP experiments based on the profiles of peaks binding to TSS regions.

For comparing overlap of peaks and annotated genes, ChIPseeker provides *vennplot* to calculate the overlap and visualize the result. This method is widely used to compare different ChIP experiments and to infer cooperative in regulation. However simple overlap calculation is insufficient. In ChIPseeker, we provide *enrichPeakOverlap* function to perform statistically rigorous comparisons of ChIP-seq datasets. ChIPseeker provides *shuffle* function to randomly permute the genomic locations of ChIP peaks. The enrichment analysis of peak overlap is based on permutation test. A thousand (nShuffle = 1000, by default) of random ChIP data were generated to estimate the background null distribution of the overlap. The *P*-value is then calculated by the probability of observing more extreme overlap. Multiple comparison corrections including Bonferroni, Benjamini and False Discovery Rate (FDR) are also incorporated. With the *enrichPeakOverlap* function, we can estimate how well two replicated experiments are and also compare overlap of two ChIP datasets from experiments of two binding proteins. If the overlap is significant, then these two proteins may cooperate in regulations.

To make this co-factor inference available to the community, we incorporated the ChIP seq data deposited in GEO database. With these datasets, we can compare our own dataset to those deposited in GEO to identified co-occurrence binding proteins that maybe cooperated with the one we are interested in. Hypothesis can be generated by this inference and serve as a starting point for further investigations. Within the ChIPseeker package, we have collected information of about 15 000 bed files deposited in GEO database, which covers 68 species. User can access detail information by *getGEOInfo* function for specify genome version. ChIPseeker provides *downloadGEObedFiles* function to download all the bed files of a particular genome. Users can also use *downloadGSMbedFiles* to download all bed files with a specific GSM accession number list. After downloading the bed files, users can pass the folder to *enrichPeakOverlap* function, which will parse the folder and compare all the bed files. It is also possible to test the overlap significance among bed files that are mapping to different genome versions and even bed files from different species if chain file is passed to *enrichPeakOverlap*.

## 3 Results

ChIPseeker is developed as an R package within the Bioconductor (Gentleman *et al.*, 2004) project and is released under Artistic-2.0 License. ChIPseeker integrates ChIP annotation, comparison and visualization and serves as a toolbox for analysis of ChIP-seq data. It can visualize genomic coverage of ChIP-seq data, annotate genomic features of ChIP peaks and their nearest genes. ChIPseeker supports visualize intensities of ChIP peaks binding to TSS sites and comparison based on these profiles. It supports statistical test for overlap among ChIP-seq datasets. More importantly, it removes a major obstacle for scientists to query and utilize ChIP-seq data in public domains.

## Funding

This work was supported by the National Natural Science Foundation of China (21271086 to Q.Y.H) and Guangdong Natural Science Research Grant (32209003 to Q.Y.H.). The fundings are all received from China.

*Conflict of Interest*: none declared.

## References

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Huang,W. *et al.* (2013) PAVIS: a tool for peak annotation and visualization. *Bioinformatics*, **29**, 3097–3099.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

Perez-Pinera,P. *et al.* (2013) Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat. Methods*, **10**, 239–242.

Salmon-Divon,M. *et al.* (2010) PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*, **11**, 415.

Yu,G. *et al.* (2012) clusterProfiler: an R Package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.*, **16**, 284–287.

Yu,G. *et al.* (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.

Zhu,J. *et al.* (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.*, **40**, 854–861.

Zhu,L.J. *et al.* (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, **11**, 237.