ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data

Lihua J Zhu1,2*, Claude Gazin3, Nathan D Lawson1,2, Hervé Pagès4, Simon M Lin5, David S Lapointe6, Michael R Green1,2

Abstract

Background: Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) or ChIP followed by genome tiling array analysis (ChIP-chip) have become standard technologies for genome-wide identification of DNA-binding protein target sites. A number of algorithms have been developed in parallel that allow identification of binding sites from ChIP-seq or ChIP-chip datasets and subsequent visualization in the University of California Santa Cruz (UCSC) Genome Browser as custom annotation tracks. However, summarizing these tracks can be a daunting task, particularly if there are a large number of binding sites or the binding sites are distributed widely across the genome.

Results: We have developed ChIPpeakAnno as a Bioconductor package within the statistical programming environment R to facilitate batch annotation of enriched peaks identified from ChIP-seq, ChIP-chip, cap analysis of gene expression (CAGE) or any experiments resulting in a large number of enriched genomic regions. The binding sites annotated with ChIPpeakAnno can be viewed easily as a table, a pie chart or plotted in histogram form, i.e., the distribution of distances to the nearest genes for each set of peaks. In addition, we have implemented functionalities for determining the significance of overlap between replicates or binding sites among transcription factors within a complex, and for drawing Venn diagrams to visualize the extent of the overlap between replicates. Furthermore, the package includes functionalities to retrieve sequences flanking putative binding sites for PCR amplification, cloning, or motif discovery, and to identify Gene Ontology (GO) terms associated with adjacent genes.

Conclusions: ChIPpeakAnno enables batch annotation of the binding sites identified from ChIP-seq, ChIP-chip, CAGE or any technology that results in a large number of enriched genomic regions within the statistical programming environment R. Allowing users to pass their own annotation data such as different ChIP preparations and a dataset from literature, or existing annotation packages, such as GenomicFeatures and BSgenome, provides flexibility. Tight integration to the biomaRt package enables up-to-date annotation retrieval from the BioMart database.
conserved elements. However, searching the UCSC Genome Browser can be a daunting task for the user, particularly if there are a large number of binding sites or the binding sites are distributed widely across the genome.

Several useful web applications have been developed for managing and annotating ChIP-chip data [12-14] and ChIP-seq data [14]. However, there is a need for technology platform-independent and genome-independent batch annotation tools. Here we describe a Bioconductor package called ChIPpeakAnno that facilitates batch annotation, using a variety of annotation sources, of binding sites identified by any technologies which result in large datasets of enriched regions, such as ChIP-chip, ChIP-seq and CAGE. ChIPpeakAnno leverages the statistical environment R/Bioconductor with various sources of annotations, such as Ensembl, the UCSC genome database and others. In addition, users have the flexibility to label enriched regions with any annotation of interest such as a dataset from the literature. This package is available from Bioconductor, an open source and open development software project specializing in biological data analysis and integration based on R, a system for statistical computation and graphics [15,16]. Bioconductor tools are distributed as separate but interoperable packages, each specializing in different areas of biological data analysis, such as the limma package for analyzing microarray data [17] and the biomaRt package for retrieving genomic annotation from the federated query system BioMart Ensembl [11,18,19]. The ChIPpeakAnno package contains various functionalities to batch-annotate enriched regions identified from ChIP-seq, ChIP-chip or CAGE experiments.

ChIPpeakAnno emphasizes flexibility, integration and ease of use. Users are supplied with functionalities for annotating peaks from ChIP-seq, ChIP-chip, CAGE or any experiment resulting in a list of chromosome coordinates with any annotation track users are interested in. Even though some of the functionalities such as the retrieval of neighbouring sequences for a set of peaks are available in other software programs, the complete set of tools and the flexibility offered by ChIPpeakAnno are not available in any other software. The main differentiating point from CEAS, CisGenome and other software is that ChIPpeakAnno enables comparison between a set of peaks with any annotation feature objects, for example comparing to CpG islands, to conserved elements (or other annotated features not captured by CEAS http://ceas.cbi.pku.edu.cn/submit.htm or CisGenome http://www.biostat.jhsph.edu/~hji/cisgenome/index.htm) (survey results) or comparing two sets of peaks between replicates (personal communication with Ivan Gregoretti at NIH) or transcription factors within a complex (unpublished data). In addition, unlike ChIPpeakAnno, CEAS or CisGenome does not have overlapping significance testing or Gene Ontology (GO) enrichment testing implemented. GO is a system for describing the molecular function, biological process and cellular compartmentalization of gene products [20]. Another main advantage of ChIPpeakAnno is the ability/flexibility to plug in with other annotation packages, such as biomaRt [17] and GO.db, ChIP-chip analysis packages such as Ringo [21] and ACME [22], other fast moving deep-sequencing analysis capabilities and infrastructure (Table 1) such as ShortRead [23], DESeq [24], edgeR [25], BayesPeak [26], chipseq, ChIP-seqR, Rolexa [27], BGenome, IRanges, Biostrings, rtracklayer [28], GenomeGraphs [29] and statistical analysis tools such as mutlitest and limma [17] in Bioconductor (survey results).

Usability is the top priority for ChIPpeakAnno. Once the package is loaded, one line of code (annotatePeakBatch) enables users to find nearest or overlapping features such as gene, exon, miRNA, 5’ utr, 3’ utr, peaks from another dataset or any annotation track of interest. Users are also provided with the flexibility and functionality to get the annotation on the fly (getAnnotation). Two lines of code (getEnrichedGO) allow users to find enriched gene ontology terms. One line of code (makeVennDiagram) allows users to draw a Venn diagram and provide a p-value for determining the significance of the overlapping between datasets. Repeated calling of function findOverlappingPeaks enables users to find the overlapping among peaks from several different experiments, which will help users to determine how peaks from different replicates overlap and how peaks from different transcription factors within a complex overlap.

Implementation

ChIPpeakAnno implements a common annotation workflow for ChIP-seq or ChIP-chip data in R, a system for statistical computation and graphics [15,16]. To promote component reuse and compatibility among Bioconductor packages, ChIPpeakAnno utilizes the IRanges package and represents the peak list as RangedData to efficiently find the nearest or overlapping gene, exon, 5’ utr, 3’ utr, microRNA (miRNA) or other custom feature(s) supplied by the user, such as the most conserved non-coding element, CpG island or transcription factor binding sites. All peak-calling software produces a file containing at least a list of chromosome coordinates that is all ChIPpeakAnno package needs. Both BED http://genome.ucsc.edu/FAQ/FAQformat#format1 and GFF (General Feature Format, http://genome.ucsc.edu/FAQ/FAQformat#format3) are common file formats that provide a flexible way to define peaks or annotations as data lines. Therefore, conversion functions BED2RangedData and GFF2RangedData were
implemented for converting these data formats to a RangedData object. Since the genome annotations are updated periodically/frequently, we have leveraged the biomaRt package from Bioconductor to enable retrieval of annotation data on the fly from Ensembl. For fast access, transcription start sites from common genomes such as TSS.human.NCBI36, TSS.human.GRCh37, Exon-PlusUtr.human.GRCh37, TSS.mouse.NCBIM37, TSS.rat.RGSC3.4 and TSS.zebrafish.Zv8 were included as pre-built annotation data packages. Users also have the flexibility to pass annotation data from the GenomicFeatures package as well as their own annotation data, such as a list of binding sites from other transcription factors, a different ChIP preparation or a different peak-calling algorithm. We have also utilized the BSgenome package to implement functions that allow retrieval of flanking sequences associated with peaks of interest. This facilitates subsequent PCR amplification, cloning and/or motif discovery using algorithms such as MEME [3,30]. To ascertain whether the identified peaks are enriched around genes with certain GO terms, we have implemented a GO enrichment test. This test applies the

Table 1 An overview of Bioconductor packages for analyzing high-throughput sequencing data.

| Package          | Classification | Functionalities                                                                 |
|------------------|----------------|---------------------------------------------------------------------------------|
| ShortRead        | Input/Output   | Provides methods for reading, quality assessment (QA) and basic manipulation of high-throughput sequencing data. |
| Rolexa           | Base Calling   | Supports probabilistic base calling, quality checks and diagnostic plots for Solexa sequencing data. |
|_ranges           | Infrastructure | Provides infrastructure for representing and manipulating sets of integer ranges, and implements algorithms for range-based calculations such as intersect, union, disjoint, overlap and coverage. |
| BSgenome         | Whole Genome   | Supplies infrastructure for efficiently representing, accessing and analyzing whole genome. |
| Biostrings       | String manipulation | Provides functions for pattern matching, sequence alignment and string manipulation. |
| rtracklayer      | Visualization  | Provides an interface between R and genome browsers and implements functions to import, create, export, and display track data by linking R with existing genome browsers. |
| GenomeGraphs     |                | Integrates Ensembl annotation obtained using the biomaRt package and the grid graphic package to facilitate visualization, plotting and analysis of a diverse genomic datasets. |
| ChIPpeakAnno     | Annotation     | Implements a common annotation workflow for ChIP-seq data such as finding nearest or overlapping features and obtaining enriched GO terms. In addition, it contains functions for determining the significance of the overlap and visualizing the overlap as a Venn diagram among different datasets. |
| Genominator      | Annotation     | Offers an interface for storing and retrieving genomic data in SQLite database. |
| ChiPsim          | Simulation of ChIP-seq experiments | Provides a framework for the simulation of ChIP-seq experiments such as nucleosome positioning and transcription factor binding sites. |
| chipsseq*        | Analysis of ChIP-seq data | Implements basic workflow for analyzing ChIP-seq experiments, including functions to extend reads, calculating genomic coverage, and identifying peaks. |
| CSR*             |                | Contributes methods to normalize the count data and detect protein-bound genomic regions with controlled false discovery rate through random permutation. Models the sequence counts as poison distribution. |
| BayesPeak*       |                | Identifies peaks using hidden Markov models and Bayesian statistical methodology. Models the sequence counts as the negative binomial distribution. |
| ChiPseqR         | Analysis of nucleosome ChIP-seq data | Furnishes functions to analyze nucleosome ChIP-seq data and may be adapted to handle other types of ChIP-seq experiments. |
| edgeR            | Analysis of RNA-seq data | Provides statistical routines for determining differential expression in count-based expression data such as RNA-seq, SAGE and CAGE. The RNA-seq data are modelled as the negative binomial distribution and applied with empirical Bays procedure. |
| DESeq            |                | Implements functions for identifying differentially expressed genes from RNA-seq data by modelling the RNA-seq data as the binomial distribution. |
| baySeq           |                | Contains methods to determine differential expression in count based expression data with more complex experimental designs using Bayesian methods. |
| DESeq*           |                | Provides functions for identifying differentially expressed genes from RNA-seq data by modelling the RNA-seq data as the negative binomial distribution. |
| goseq*           | Enrichment testing of RNA-seq data | GO enrichment testing for RNA-seq data. |

*Available in BioC 2.6 in R 2.11.0.
Results

Example 1: Finding the nearest gene and the distance to the transcription start site of the nearest gene

The output from ChIP-seq or ChIP-chip analysis is a list of binding sites (as chromosome locations) that are significantly enriched in the ChIP sample(s) compared with the corresponding control sample(s). The example below details how to find the nearest gene and the distance to the transcription start site (TSS) of the nearest gene in the human genome for a list of binding sites (named myPeakList) of type RangedData. The distance is calculated as the distance between the start of the binding site and the TSS, which is the gene start for genes located on the forward strand and the gene end for genes located on the reverse strand.

The first step is to load the ChIPpeakAnno package, an example dataset and an annotation dataset. In this example, the example dataset contains putative STAT1-binding regions identified in un-stimulated cells [2], and the annotation dataset contains the TSS coordinates and strand information from human GRCh37.

> library(ChIPpeakAnno)
> data(myPeakList)
> data(TSS.human.GRCh37)

In the next step, the function a annotatePeakInBatch is called to find the gene with nearest TSS or overlapping gene that is not the nearest TSS and corresponding distance for the list of binding regions. Sometimes, a peak is within a gene but far from the gene’s TSS. Setting the parameter output to “both” outputs both the genes with nearest TSS and the overlapping gene. The parameter maxgap sets the maximum gap to be considered as overlapping. The parameter multiple indicates whether multiple overlapping features should be returned for one peak.

> annotatedPeak = annotatePeakInBatch(myPeakList, AnnotationData = TSS.human.GRCh37, output = "both", multiple = F, maxgap = 0)

The annotated peaks can be saved as an Excel file for biologists to view easily.

> write.table(as.data.frame(annotatedPeak), file = "annotatedPeakList.xls", sep = "\t", row.names = FALSE)

Plotting the distribution of the peaks relative to the TSS gives a birds-eye view of the peak distribution relative to the genomic features of interest.

> y = annotatedPeak$distanceToFeature[!is.na(annotatedPeak$distanceToFeature)] & annotatedPeak$fromOverlappingOrNearest == "NearestStart"
> hist(y, xlab = "Distance To Nearest TSS", main = "", breaks = 1000, xlim = c(min(y)-100, max(y) + 100))

Such a plot generated from the putative STAT1-binding regions identified in un-stimulated cells [2] shows that the STAT1-binding sites are enriched in regions near TSSs (Figure 1). A pie chart was also generated as follows to show the distribution of relative position of the peaks to the nearest gene (Figure 2).

> temp = as.data.frame(annotatedPeak)
> pie(table(temp[as.character(temp$fromOverlappingOrNearest == "Overlapping" | (as.character(temp$fromOverlappingOrNearest == "NearestStart") & !temp$peak %in% temp[as.character(temp$fromOverlappingOrNearest == "Overlapping",temp%insideFeature)]))

It is also possible to obtain the annotation on-line from Ensembl using the getAnnotation function as follows prior to calling annotatePeakInBatch. Please refer
 Peaks in text format from peak-calling algorithms can be easily imported to R as `data frame` then converted to `RangedData`. For binding sites represented in BED or GFF format, `BED2RangedData` or `GFF2RangedData` were provided for converting these data formats to `RangedData`.

**Example 2: Determining the significance of the overlapping and visualizing the overlap as a Venn diagram among different datasets**

The second example describes how to determine the significance of the overlap, visualize the overlap in a Venn diagram and obtain merged peaks from different datasets such as different biological replicates, different peak-calling algorithms or different proteins functioning as a transcription complex. Here we give examples using different biological replicates.

The first step is to load the `ChiPeakAnno` package and three example datasets as `RangedData` that contains putative Ste12-binding regions identified in yeast from three biological replicates [31].

```r
>library(ChiPeakAnno)
>data(Peaks.Ste12.Replicate1)
>data(Peaks.Ste12.Replicate2)
>data(Peaks.Ste12.Replicate3)
```

In the next step, the function `makeVennDiagram` is called to generate a Venn diagram to visualize the overlap among the three replicates. In addition, pair-wise overlapping significance tests were performed with hypergeometric test and p-values were generated. The parameter `NameOfPeaks` indicates the names of the dataset for labeling the Venn diagram. The parameter `maxgap` indicates the maximum distance between two peak ranges for them to be considered overlapping. The parameter `totalTest` indicates how many potential peaks in total that is used in hypergeometric test.

```r
>makeVennDiagram(RangedDataList(Peaks.Ste12.Replicate1, Peaks.Ste12.Replicate2, Peaks.Ste12.Replicate3), NameOfPeaks = c("Replicate1", "Replicate2", "Replica-3"), maxgap = 0, totalTest = 1580)
```

As a result, a Venn diagram was generated for visualizing the overlap among the above three replicates. The pair-wise overlap comparisons show that the peaks identified from the replicates overlap significantly (Figure 3, p-value < 0.0001). The same analysis was applied to the three biological replicates of Cse4 and the overlap between replicate 1 and 2 is significant at p-value < 0.01 while the other two overlapping is significant at p-value < 0.05 (Figure 4).

The peak ranges from replicates do not overlap perfectly. It is desirable to combine all the overlapping peaks across replicates into merged peaks that cover all the overlapping peaks from the replicates. Calling the function `findOverlappingPeaks` can generate the merged
peaks. Besides the parameters mentioned previously, an additional required parameter *multiple* indicates whether to return merged peaks from multiple overlapping peaks.

```
>MergedPeaks = findOverlappingPeaks(findOverlappingPeaks(Peaks.Ste12.Replicate1, Peaks.Ste12.Replicate2, maxgap = 0, multiple = F, NameOfPeaks1 = "R1", NameOfPeaks2 = "R2")$MergedPeaks, Peaks.Ste12.Replicate3, maxgap = 0, multiple = F, NameOfPeaks1 = "Replicate1Replicate2", NameOfPeaks2 = "R3")$MergedPeaks
```

Next, nearest genes and distances between peak location and nearest TSS can be obtained by annotating the merged peaks with SGD1.01 using `annotatePeakInBatch` function illustrated in example 1 (Figure 5 & 6). The same analysis was performed with Cse4 binding-sites (Figure 7 & 8). The un-equal variance t-test shows that the distribution of the distance of Ste12-binding sites to nearest TSSs (Figure 5, 264 ± 36 bases) is very different from that of Cse4-binding sites (Figure 6, 311 ± 160 bases) (p-value = 0.001). Ste12-binding sites are distributed more towards the upstream of the gene (Figure 5 & 6) while Cse4-binding sites are distributed more inside and downstream of the gene (Figure 7 & 8). This result is consistent with what has been observed previously [31]. The annotated datasets are available in Additional file 1 and Additional file 2.

**Example 3: Obtaining the sequences around the binding sites for PCR amplification or motif discovery**

The third example describes how to obtain the sequences surrounding binding sites (in this example,
100 bp of upstream and downstream sequence) for PCR amplification, cloning or motif discovery [3,30].

The first step is to load the ChIPpeakAnno package and create an example peak dataset as RangedData. Next, the organism-specific BSgenome package is loaded followed by calling the function getAllPeakSequence. The function available.genomes shows a list of available organism-specific BSgenome data packages. In this example, E. coli data package is used due its light-weight.

```r
>library(ChIPpeakAnno)
>peaks = RangedData(IRanges(start = c(100, 500), end = c(300, 600), names = c("peak1", "peak2")), space = c("NC_008253", "NC_010468"))
>library(BSgenome.Ecoli.NCBI.20080805)
>peaksWithSequences = getAllPeakSequence(peaks, upstream = 100, downstream = 100, genome = Ecoli)
```

To convert the sequences to a common FASTA file format, the following function is called.

```r
>write2FASTA(peaksWithSequences, file="test.fa", width = 50)
```

Sequences for merged Ste12 binding sites were obtained from package BSgenome.Scerevisiae.UCSC.sacCer2 (Additional file 3). Significant motifs (E-value < 0.0000001) were identified by running MEME [30] with motif occurrence set as ZOOP, minimum width as 8, maximum width as 20 and other parameters as default (Figure 9).

```r
>peaks = RangedData(IRanges(start = c(100, 500), end = c(300, 600), names = c("peak1", "peak2")), space = c("NC_008253", "NC_010468"))
>library(BSgenome.Ecoli.NCBI.20080805)
>peaksWithSequences = getAllPeakSequence(peaks, upstream = 100, downstream = 100, genome = Ecoli)
>write2FASTA(peaksWithSequences, file="test.fa", width = 50)
```

Significant motifs (E-value < 0.0000001) were identified by running MEME [30] with motif occurrence set as ZOOP, minimum width as 8, maximum width as 20 and other parameters as default (Figure 9).
Example 4: Obtaining enriched GO terms near the peaks

The fourth example describes how to obtain a list of enriched GO terms associated with adjacent genes using a hypergeometric test.

The first step is to load the TSS annotated peak, which is the result returned from calling the function `annotatePeakInBatch`, and the organism-specific GO gene mapping package (e.g., `org.Hs.eg.db` for the GO gene mapping for human; for other organisms, please refer to http://www.bioconductor.org/packages/release/data/annotation/ for additional `org.xx.eg.db` packages).

```r
> data(annotatedPeak)
> library(org.Hs.eg.db)
```

The next step is to call the function `getEnrichedGO`. The parameter `maxP` is the maximum p-value required to be considered to be significant, `multiAdj` indicates whether to apply multiple hypothesis testing adjustment, `minGOterm` is the minimum count in a genome for a GO term to be included, and `multiAdjMethod` is the multiple testing procedure to be applied (for details, see `mt.rawp2adjp` in the `multtest` package).

```r
> enrichedGO <- getEnrichedGO (annotatedPeak[1:6], orgAnn="org.Hs.eg.db", maxP = 0.01, multiAdj = TRUE, minGOterm = 10, multiAdjMethod="BH")
```

Where `enrichedGO$bp` contains a list of enriched GO biological process, `enrichedGO$mf` contains a list of enriched GO molecular functions and `enrichedGO$cc` contains a list of enriched GO cellular components.

Table 2 shows a list of enriched GO terms for yeast transcription factor Ste12 [31].

### Conclusions

ChIPpeakAnno enables batch annotation of binding sites identified from ChIP-seq, ChIP-chip, CAGE or any technology that results in a large number of enriched genomic regions for any species with existing annotation data within the statistical programming environment R. Allowing users to pass their own annotation data such as different ChIP preparation and a dataset from literature, or existing annotation packages, such as Genomic-Features and BSgenome, provides flexibility while the tight integration to the biomaRt package enables up-to-date annotation retrieval from the BioMart database. The main advantage of ChIPpeakAnno is the ability/flexibility to plug in with other annotation packages, ChIP-chip analysis packages, other fast moving deep-sequencing analysis capabilities and infrastructure and statistical analysis tools in Bioconductor. Another advantage of ChIPpeakAnno is that it enables comparison between a set of peaks with any annotation feature objects, between two sets of peaks from replicate experiments or transcription factors within a complex and determination of the significance of the overlap.

The ChIPpeakAnno package provides documentation in the form of an interactive manual illustrating the usage of individual functions as well as a vignette containing executable code snippets demonstrating a case-oriented help session. The vignette is run at package building and installation time, and thus also serves as a testing suite. Some of the examples described in the paper are also demonstrated in the vignette.

### Availability and requirement

ChIPpeakAnno is an open source software package under the GNU General Public Licence v2.0 and has been contributed to the Bioconductor Project. The software, source code and documentation are available for download from http://www.bioconductor.org or installed from R by typing source http://bioconductor.org/biocLite.R and biocLite (“ChIPpeakAnno”). The package has been tested and run on OS X, Windows and various Linux systems.

### Table 2 Enriched GO terms of Ste12-binding sites in yeast.

| GO ID   | GO Term                  | GO Definition                                                                 | Category | FDR  |
|---------|--------------------------|-------------------------------------------------------------------------------|----------|------|
| GO:0055114 | oxidation reduction         | The process of removal or addition of one or more electrons with or without the concomitant removal or addition of a proton or protons. | BP       | 0.018 |
| GO:0008270 | zinc ion binding          | Interacting selectively and non-covalently with zinc (Zn) ions.                 | MF       | 0.047 |
| GO:0043167 | ion binding               | Interacting selectively and non-covalently with ions, charged atoms or groups of atoms. | MF       | 0.046 |
| GO:0043169 | cation binding            | Interacting selectively and non-covalently with cations, charged atoms or groups of atoms with a net positive charge. | MF       | 0.046 |
| GO:0043565 | sequence-specific DNA binding | Interacting selectively and non-covalently with DNA of a specific nucleotide composition, e.g. GC-rich DNA binding, or with a specific sequence motif or type of DNA e.g. promoter binding or rDNA binding. | MF       | 0.046 |
| GO:0046914 | transition metal ion binding | Interacting selectively and non-covalently with a transition metal ions; a transition metal is an element whose atom has an incomplete d-subshell of extranuclear electrons, or which gives rise to a cation or cations with an incomplete d-subshell. Transition metals often have more than one valency state. Biologically relevant transition metals include vanadium, manganese, iron, copper, cobalt, nickel, molybdenum and silver. | MF       | 0.046 |

The enriched GO terms were obtained from the putative Ste12-binding regions merged from three biological replicates identified in yeast [31]. The parameter used to generate the list is `maxP = 0.05, multiAdj = TRUE, minGOterm = 5` and `multiAdjMethod = "BH"`. 
ChiPpeakAnno depends on R version 2.10. 0 or later and the following Bioconductor packages: biomartR, multtest, iRanges, limma, Biostrings, BSgenome, and GO.db. In addition, the lightweight organism-specific package BSgenome. Ecoli.NCBI.20080805 and org.Hs.eg.db were installed during build time for testing the code snippets in the vignette. All these packages can be downloaded from Bioconductor or installed from R using the http://bioconductor.org/biocLite/R script.

Additional file 1: Annotated Ste12-binding sites. An Excel file contains the annotated Ste12-binding sites merged from the three biological replicates in yeast [31].

Additional file 2: Annotated Cse4-binding sites. An Excel file contains the annotated Cse4-binding sites merged from the three biological replicates in yeast [31].

Additional file 3: Sequence file of Ste12-binding sites used for MEME input. FASTA formatted sequence file from Ste12-binding sites merged from the three biological replicates in yeast [31].

Acknowledgements
We would like to thank the support from the Program in Gene Function and Expression (PGFE) at UMass Medical School (UMMS). We are grateful for the constructive suggestions from the manuscript editors and anonymous reviewers, and the Bioconductor package reviewers Nishant Gopalak, Marc Carlson and other anonymous reviewers. We are indebted to the users of ChiPpeakAnno who provided great ideas and feedbacks to enhance the features of the software. We also thank Zhiping Weng in the Program of Bioinformatics and Integrative Biology at UMSM for reviewing the manuscript, Ivan Gregoretti at National Institute Health for helping with the revision, Sara Evans at PGFE for editorial assistance, Alan Ritacco at UMMS Academic Research and Computing Services for providing computational support, and Glenn Maston at PGFE. Ping Wan at Capital Normal University and Ellen Kettler at the UMMS Deep Sequencing Core Facility for helpful discussion.

Author details
1Program in Gene Function and Expression, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. 2Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. 3UMR217 CNRS/CEA, IRCM-CEA, Evry, ile de France 91057, France. 4Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109-1024, USA. 5The Biomedical Informatics Center, Northwestern University, Chicago IL 60611, USA. 6Information Services, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.

Authors’ contributions
LJZ drafted the manuscript. LJZ and HP developed the software package. CG, NDL, MRG, SML and DSQL provided scientific advice. DSL performed the MEME analysis. All authors participated in writing and approved the final manuscript.

Received: 9 December 2009 Accepted: 11 May 2010
Published: 11 May 2010

References
1. Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of in vivo protein-DNA interactions. Science 2007, 316(5830):1497-1502.
2. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Vahir R, Delaney A, et al: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007, 4(8):651-657.
3. Valouev A, Johnson DS, Sundquist A, Medina C, Anten E, Batzoglou S, Myers RM, Sidow A: Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. Nat Methods 2008, 5(9):829-834.
4. Johnson DS, Li W, Gordon DB, Bhattacharjee A, Curry B, Gish J, Brzulj L, Carroll JS, Brown M, Flicek P, et al: Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Res 2008, 18(3):393-403.
5. Fejes AP, Robertson G, Bilenky M, Vahid R, Bainbridge M, Jones SJ: FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 2008, 24(15):1729-1730.
6. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 2008, 26(11):1293-1300.
7. Zhang Y, Li T, Meyer CA, Ekehottie J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al: Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008, 9(9):R137.
8. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M: Modeling ChIP sequencing in silico with applications. PLoS Comput Biol 2008, 4(8):e1000158.
9. Sharon E, Lubliner S, Segal E: A feature-based approach to modeling protein-DNA interactions. PLoS Comput Biol 2008, 4(8):e1000159.
10. Rozowsky J, Luskin C, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 2009, 27(1):66-75.
11. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melucco C, Hammond M, Rocca-Serra P, Cox T, Birney E: Ensembl: a generic system for fast and flexible access to biological data. Genome Res 2004, 14(1):169-170.
12. Ryder E, Jackson R, Ferguson-Smith A, Russell S: MAMMOT—a set of tools for the design, management and visualization of genomic tile datasets. Bioinformatics 2006, 22(7):883-884.
13. Cesaroni M, Cittaro D, Brozzi A, Pelicci PG, Luzi L: CARPET: a web-based database for the analysis and expression profiling of the genome. Bioinformatics 2006, 22(5):2918-2920.
14. Shin H, Liu T, Mannai AK, Liu XS: CEAS: cis-regulatory element annotation system. Bioinformatics 2009, 25(19):2605-2606.
15. Ihaa RG: R: a language for data analysis and graphics. J Comput Stat 1996, 5:5.
16. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Duotod S, Ellis B, Gautier L, Ge Y, Gentry J, et al: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004, 5(10):R80.
17. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2005, 3, Article3.
25. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26(1):139-140.

26. Spyrou C, Stark R, Lynch AG, Tavare S: BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics 2009, 10:299.

27. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F: Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 2008, 9:431.

28. Lawrence M, Gentleman R, Carey V: rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 2009, 25(14):1841-1842.

29. Durinck S, Bullard J, Spellman PT, Dudoit S: GenomeGraphs: integrated genomic data visualization with R. BMC Bioinformatics 2009, 10:2.

30. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994, 2:28-36.

31. Lefrancois P, Euskirchen GM, Auerbach RK, Rozowsky J, Gibson T, Yellman CM, Gerstein M, Snyder M: Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. BMC Genomics 2009, 10:57.

doi:10.1186/1471-2105-11-237
Cite this article as: Zhu et al: ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 2010 11:237.