Li et al. BMC Bioinformatics 2019, 20(Suppl 15):515
https://doi.org/10.1186/s12859-019-3090-0

BMC Bioinformatics

SOFTWARE Open Access

CSA: a web service for the complete process of ChIP-Seq analysis

Min Li[1*], Li Tang[1], Fang-Xiang Wu[2], Yi Pan[3] and Jianxin Wang[1]

From 14th International Symposium on Bioinformatics Research and Applications (ISBRA'18)
Beijing, China. 8-11 June 2018

Abstract

Background: Chromatin immunoprecipitation sequencing (ChIP-seq) is a technology that combines chromatin immunoprecipitation (ChIP) with next generation of sequencing technology (NGS) to analyze protein interactions with DNA. At present, most ChIP-seq analysis tools adopt the command line, which lacks user-friendly interfaces. Although some web services with graphical interfaces have been developed for ChIP-seq analysis, these sites cannot provide a comprehensive analysis of ChIP-seq from raw data to downstream analysis.

Results: In this study, we develop a web service for the whole process of ChIP-Seq Analysis (CSA), which covers mapping, quality control, peak calling, and downstream analysis. In addition, CSA provides a customization function for users to define their own workflows. And the visualization of mapping, peak calling, motif finding, and pathway analysis results are also provided in CSA. For the different types of ChIP-seq datasets, CSA can provide the corresponding tool to perform the analysis. Moreover, CSA can detect differences in ChIP signals between ChIP samples and controls to identify absolute binding sites.

Conclusions: The two case studies demonstrate the effectiveness of CSA, which can complete the whole procedure of ChIP-seq analysis. CSA provides a web interface for users, and implements the visualization of every analysis step. The website of CSA is available at http://CompuBio.csu.edu.cn

Keywords: ChIP-seq, Quality control, Peak calling, Downstream analysis, Visualization

Background

Next-generation sequencing technologies have produced a large amount of raw data, lots of computational methods have been developed to solve the problem of genome assembly [1–6], variation detection and annotation [7, 8], which had given rise to the release of unknown reference genome and helped interpret the complex genome structure. Based on the complete reference genome, the analysis of NGS data has become reasonable, the chromatin immunoprecipitation sequencing (ChIP-seq) [9] is an important technology for functional genomics research [10], and brought a qualitative leap for related biological experiments. The real value of the ChIP-seq technology lies not only in obtaining information about the distribution of DNA-related proteins in the genome, but also in digging deeper esoteric secrets behind such information [11].

The process of ChIP-seq contains mapping, peakcalling, and downstream analysis. Mapping is the most memory-consuming step, and lots of mapping methods are proposed to align the sequenced reads to reference genome. BWA [12] is a software package that maps low divergence sequences to a large reference genome. Bowtie [13] is a short read aligner, which is ultrafast speed and memory-efficiency. Bowtie2 [14] is used to align sequencing reads to long reference sequences, with the features of ultrafast and memory-efficiency. SOAP [15] is a faster and efficient alignment tool for short sequence reads against reference sequences. BLAST [16] is used to find the similar regions between biological sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify

* Correspondence: limin@mail.csu.edu.cn
[1]School of Computer Science and Engineering, Central South University, Changsha, China
Full list of author information is available at the end of the article

members of gene families. Subread [17] also finds regions of local similarity between sequences, which aligns nucleotide or protein sequences against sequence databases and calculates the statistical significance of matches. NGM [18] has the ability to process higher mismatch rates than comparable algorithms while still performing better than them in terms of runtime, and is a flexible and highly sensitive short read mapping tool, which requires SSE enabled 64 bit dual-core. The step of peakcalling is to detect the protein modification and identify the transcription factor binding sites. MACS [19] can evaluate the significance of enriched ChIP regions by capturing the influence of genome complexity, and MACS [19] combines the information of sequencing tag positions and orientations to improve the spatial resolutions. MACS2 is an updated version of MACS [19]. PeakSeq [20] is used to identify and rank the peak regions in ChIP-Seq experiments. PeakRanger [21] takes a while for user's browser to parse the generated HTML file. The lc tool needs about 1.7G ram per 10 million aligned reads. SICER [22] is to identify the enriched domains from histone modification ChIP-Seq data by a clustering method. The focus of Fin.

dPeaks [23] is on post-alignment analysis. This program includes interpreters for most common aligners and SNP callers and is able to use input from a wide variety of formats. Fseq [24] is to intuitively summarize and display individual sequence data as an accurate and interpretable signal. In the method of AREM [25], reads are modeled using a mixture model corresponding to K enriched regions and a null genomic background. BroadPeak [26] is abroad peak calling algorithm for diffuse ChIP-seq datasets. BCP can search the input file, and find the enrichment of peaks. PePr [27] uses a negative binomial distribution to model the read counts among the samples in the same group, and looks for consistent differences between ChIP and control group or two ChIP groups run under different conditions. The method diffReps [28] takes into account the biological variations within a group of samples and uses that information to enhance the statistical power. SISSRs [29] identifies the binding sites from short reads which are generated from ChIP-Seq experiments precisely.

In recent years, several platforms have been developed to analyze ChIP-seq experiment data. These platforms can be divided into three categories: command line, GUI, and web service. One of the most popular command line-based platform is HOMER [30], which provides NGS analysis and motif finding. ChIPseeker [31] is an R package, having both the command line and GUI version for ChIP peak annotation, comparison and visualization, while it is demands the system environment and requires installation in users' servers. Other platforms are based on web services, such as Nebula [32] and ChIPseek [33]. Nebula integrates several peak calling methods and provides motif findings. ChIPseek is a web server based on HOMER, which also provides peak calling, motif finding and KEGG analysis. However, most of these web-based tools can neither cover the whole process of ChIP-seq analysis, nor provide the visualization of results. The downstream analysis usually includes motif finding, Gene Ontology Analysis, and pathway analysis. The algorithm findMotifs in HOMER can find the de novo motifs and known motifs. The algorithm annotatePeaks in HOMER can perform Gene Ontology Analysis, associate peaks with gene expression data, calculate ChIP-Seq tag densities from different experiments, and find motif occurrences in peaks. iPAGE [34] provides a complete meta-analysis of whole-genome datasets in cooperation with FIRE, and a *P*-value heatmap with significant categories is generated.

Here, we develop a web-based ChIP-Seq Analysis tool (CSA), which provides a comprehensive analysis of ChIP-seq data by integrating seven mapping algorithms, thirteen peak calling methods, and three downstream analysis methods. CSA places great emphasis on the workflow, which helps finish the whole analysis through several easy steps. In addition, CSA provides the visualization of the entire process. Table 1 shows a comprehensive comparison between CSA and several other typical platforms for ChIP-seq analysis including HOMER [30], ChIPSeqWorkflow [35], ChIPseeker [31], CisGenome [36], ChIP-seq tool [37], Nebula [32], and ChIPseek [33]. Table 1 also lists the systems on which the platforms rely, the installation requirement, the interface, and the functions.

The major contributions of CSA include 1) CSA integrates more comprehensive functions, from mapping to downstream analysis, and the tools used to convert formats are also integrated; 2) CSA provides a guideline for users to choose appropriate tools, and allows users to define their own workflows, which can help them complete their analysis through several easy steps; 3) CSA also provides the visualization of the entire process, including the results of mapping, peak calling, motif finding, and pathway analysis.

Implementation

CSA provides the whole process of ChIP-seq analysis, and the pipeline of CSA for analyzing ChIP-seq data is shown in Fig. 1. In this pipeline, we take ChIP-seq raw data, a reference genome, and a control file as inputs. The step of mapping aligns short reads to reference sequences. Seven popular mapping tools: BWA [12], Bowtie [13], Bowtie2 [14], SOAP [15], BLAST [16], Subread [17], and NGM [18] are integrated in CSA. After mapping, CSA provides the step of quality control to check the correlation between replicates and published datasets

Table 1 Current typical platforms for ChIP-Seq analysis

Platform	system	installation	Interface	Functions & tools									
				work flow	Mapping	Quality control	Peak calling	Format convert	Data visualization	Find motifs	GO analysis	Peaks annotation	Pathway analysis
HOMER[22]	UNIX LINUX MacOS	Perl installation scripts	Command line	–	–	makeTagDirectory	findPeaks	pos2bed.pl, bed2pos.pl	–	findMotifsGenome.pl	annotatePeaks.pl	annotatePeaks.pl	–
ChIPSeqWorkflow[27]	UNIX LINUX	compilation from source installer	Command line	–	bowtie	–	MACS	Samtools, Bedtools	–	MEME	–	–	GAGE
ChIPseeker[23]	UNIX LINUX	R package manager	GUI/Command line	√	–	ChIPseeqerRead CountDistribution	ChIPseeqer	–	peaks coverage	ChIPseeqerFIRE, ChIPseeqerMotifMatch	–	ChIPseeqer Annotate	iPAGE
CisGenome[28]	no limitation	compilation from source installer (for Windows)	GUI/command line	–	–	–	tilemapv2	file_bed2cod, fasta_soft2hardmask, etc.	–	motifmap_matrixscan_genome,etc.	refgene_getnearestgene	–	–
ChIP-Seq tool[29]	no limitation/ UNIX LINUX	not needed	Web server/ Command line	–	Chipcor, Chipextract, chipscore	–	Chippeak, chippart	Compactsga, featreplace, etc.	–	–	–	–	–
Nebula[24]	no limitation	not needed	Web server	√	Bowtie	FASTQC	HMCan, MICSA, MACS, PeakSplitter, CCAT, FindPeaks	Samtools, Bamtools	–	ChIPmunk, AhoPro	Get peak distribution around TSS/ histone	√	–
ChIPseek[25]	no limitation	not needed	Web server	–	–	–	HOMER	BEDTools	peak location distribution	HOMER	√	–	KEGG
CSA	no limitation	not needed/ compilation from script	Web server/ local web server	√	BWA, Bowtie, Bowtie2, SOAP, BLAST, Subread, NGM	deepTools	MACS, MACS2, PeakSeq, PeakRanger, SICER, PePr, BCP, diffReps, SISSRs, FindPeaks, AREM, Fseq, BroadPeak	Samtools, bamCoverage	Mapping results,peak calling results, motif finding, Go annotation, pathway analysis	findMotifsGenome.pl	annotatePeaks.pl	annotatePeaks.pl	iPAGE

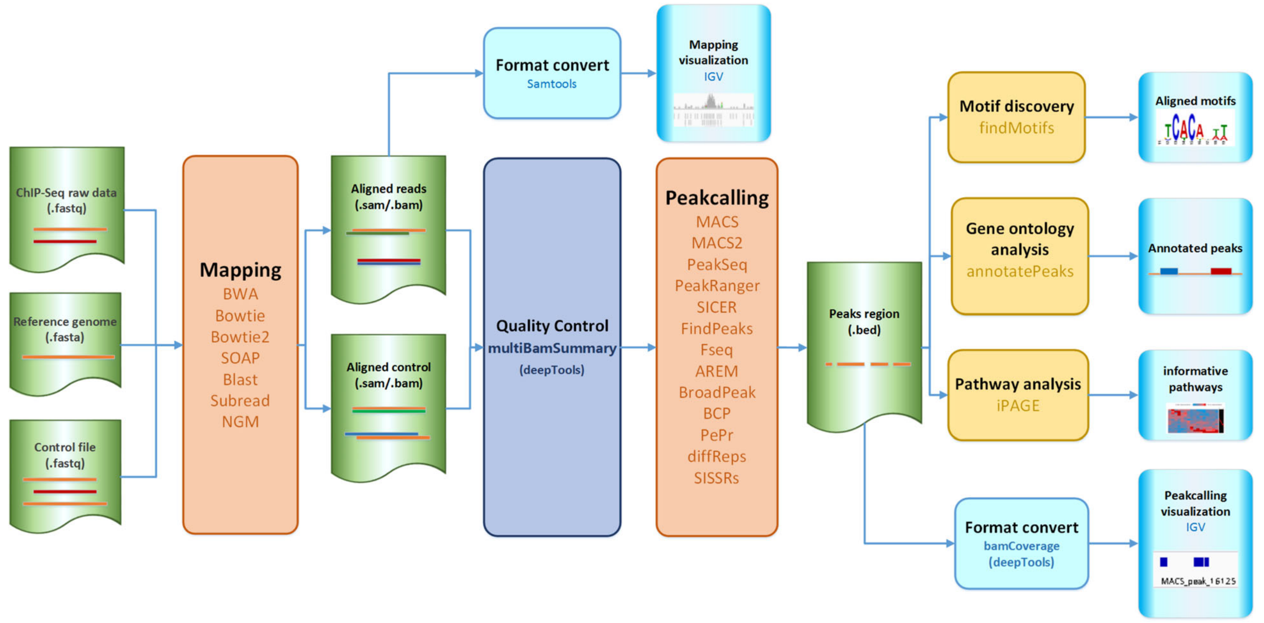


Fig. 1 The pipeline of CSA, which includes four stages: mapping, quality control, peak calling and downstream analysis. Samtools and bamCoverage are used to convert the format. The visualization of mapping, peak calling, and downstream analysis are also provided

by integrating multiBamSummary. Peak calling is the most important step which finds the enrichment of peak regions. Thirteen peak calling methods: MACS [19], MACS2, PeakSeq [20], PeakRanger [21], SICER [22], FindPeaks [15], Fseq [16], AREM [17], BroadPeak [18], BCP, PePr [19], diffReps [20], and SISSRs [29] are integrated in CSA. Moreover, three downstream analysis tools are integrated for motif analysis, GO analysis, and pathway analysis, to help users conduct further analysis and discover interesting results behind these data.

Mapping and quality control

Mapping aligns short reads to long reference sequences, and is the most computationally intensive step in the overall data analysis process. Therefore, it is important to select the appropriate alignment strategy in this step. CSA integrates seven mapping tools, while each tool has its own advantages and disadvantages. To our best knowledge, no software systems can be applied to all cases. These tools are broadly based on two approaches: hash table and Burrows- Wheeler. Burrows-Wheeler is more common, and several tools, like BWA [12], Bowite [13], and SOAP [15], have been developed based on Burrows- Wheeler indexing. If the length of reads is greater than 100 bp, it's better to use BWA. If the reads is short and single-end, Bowtie would get high accuracy. In addition, SOAP is suitable for both single-end and paired-end alignment, it reduces the usage of computer memory and improves the speed of processing reads.

Quality control is performed by the method of multiBamSummary, which is involved in the package of deepTools [38]. This tool is useful to find the correlation between published data sets and the files generated by the step of mapping. The result of this tool is an array of correlation coefficients which are displayed as a clustered heatmap. Users can judge how "strong" the relationship between the published data set and their own files.mapping and quality control.

Peakcalling

Peakcalling detects the enrichment of peak regions in ChIP-seq analysis, and thirteen methods are integrated. SAM or BAM files generated by mapping along with the control file used as the input of Peakcalling. Peak signals are generally classified into three categories according to the shape of peaks and the type of raw data. These three types are: sharp, broad and mixed. The sharp peak signals usually presented at the protein-DNA binding sites or on the histone modification sites of the regulatory elements. The broad type of peak signals generally has relationship with transcription factors and the histone modification in the gene expression region. Most current tools are suitable for the analysis of sharp peaks, such as MACS [19]. In addition, SICER [22] is designed for broad peaks [39].

Downstream analysis

We implemented three downstream analysis modules: motif analysis (findMotifs), GO analysis (annotatePeaks), and Pathway analysis (iPAGE [34]). Motif analysis module uses the BED file as input, and finds out whether the identified binding sites defined by the BED file contain the previously established consensus binding sequences for the respective proteins. Gene Ontology analysis module looks for the enrichment of various genomic annotations in peaks or regions described in the BED files. Pathway analysis module results in a *P*-value heatmap with significant categories.

Visualization

Visualization provides users with display of sequence and peak distributions. CSA integrates IGV [40] to show the results of mapping and Peakcalling. After mapping, users can get SAM format files and the alignments of reads could be visualized with these files. In the figure of alignment, the gray arrows represent reads, while the arrow indicates the orientation of the mapping. The nucleotides marked in different colors indicate mismatches between the reads and the reference. Light gray areas and white blocks display the alignments. After Peakcalling, users can get the reports about the enrichment of peaks in which a BED file is involved. IGV [40] could display the regions of enrichment through the BED file. In the figure of Peakcalling, the blue lines represent the peaks, and the length of blue lines indicates the width of peaks.

Results

Case study 1: genome-wide co-localization of several transcription regulators on enhancers

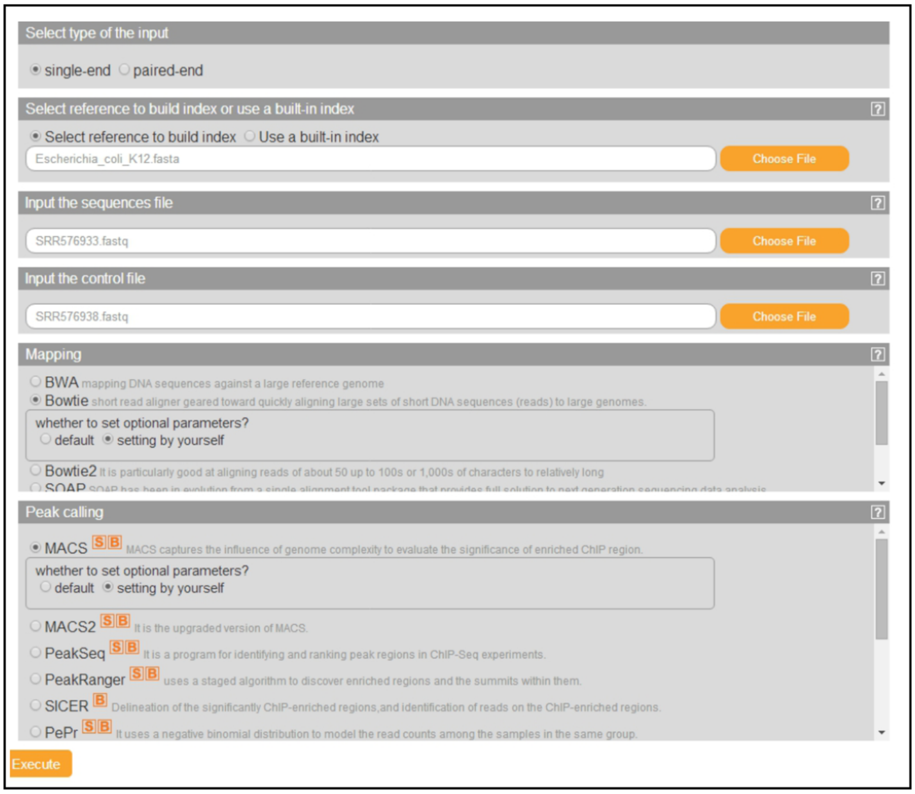
This case study describes the approach reported in Nature Cell Biology [41]. We just perform the mapping and peak calling part of their ChIP-seq analysis. YAP and TAZ are potent inducers of cell proliferation, regulating organ growth and tumorigenesis. In their analysis, YAP and TAZ antibodies were used to perform the ChIP-seq experiment in MDA-MB-231 breast cancer cells. A list of tools were used for analysis, uniquely mapped reads were retained by using Bowtie [13] (version 0.12.7), and the reference genome was hg19. Samtools was used to remove the redundant reads. IDR (Irreproducible Discovery Rate) framework was used to evaluate the consistency of the replicate experiment. Peaks were detected by MACS2 version 2.0.10, and IgG ChIP-seq was used as the control sample. The IDR threshold of 0.01 was regarded as the standard to identify the best peaks number for all datasets. At last, the enrichment of each peak could be displayed by using IGV [40].

Preparing the input data file. Here we used the “WorkFlow” module to repeat this analysis process. Firstly, ChIP-seq dataset was downloaded from Gene Expression Omnibus (GEO) [42] with accession number of GSE66083. We can get the raw sequences of YAP/TAZ/TEAD/IgG in the format of SRA, and all these data files should be converted into FASTQ format by sratoolkit so that the files could fit the input format of “WorkFlow” module.

Performing “WorkFlow”. On the page of “WorkFlow”, we selected “single-end” as the type of input, and then chose the sequences file of YAP in the format of FASTQ. CSA contained the references of genome hg19 and hg38, the reference was built in advance to save time, we clicked “Use a built-in index” to select the hg19 as the reference. In the field of control files, the FASTQ file of IgG should be input here. The mapping box contained 7 alignment tools integrated in CSA, here we chose Bowtie, and used the default parameters. The peak calling box contained 13 peak detection tools, we chose MACS2, and also used the default setting. The last step, after clicking the “Execute” button, the workflow started. We repeated the steps for the analysis of TAZ and TEAD. The definition of the workflow is shown in Fig. 2.

Viewing the output. When the operation was finished, the web jumped to the page of “Results visualization”. We learned from the analysis of Zanconato et al. that the region of promoters and enhancers here were defined by the genomic locations and overlaps of H3K4me1 and H3K4me3 peaks [43]. We selected one promoter region, and one enhancer region. Filling the file input field of scope with “chr4:41,518,010-41,541, 509”, it took a while for the visualization tool to deal with the scope. After processing, the graph of peaks binding to promoters would display on the page, and users can also download the result files through the web page in one month. Additional file 1: Fig. S1. (A) in the supplementary material shows YAP/TAZ/TEAD binding to promoters with the scope of “chr4:41,518,010-41,541, 509”. Then we input the scope of “chr4:41,118,180-41, 141,679” to view the peaks binding to enhancers. Additional file 1: Fig. S1. (B) in the supplementary material shows YAP/TAZ/TEAD binding to enhancers. We recommend using “Mapping visualization” to view the enrichment. Because the visualization of peak calling is based on the bed format file, peaks are described with a lot of blue horizontal lines. Although we can get the number and the region of peaks from this graph, it is still not clearly to identify the correlation between transcriptomes.

We performed the analysis of YAP, TAZ, and TEAD separately, and input two scopes mentioned above for these three transcription factors independently. During these analysis, CSA generated 6 figures totally. For each transcription factors, two figures were created and represented binding to promoters and enhancers respectively. In order to compare these results more obviously, we



Select type of the input

● single-end ○ paired-end

Select reference to build index or use a built-in index [?]

● Select reference to build index ○ Use a built-in index

Escherichia_coli_K12.fasta

Input the sequences file [?]

SRR576933.fastq

Input the control file [?]

SRR576938.fastq

Mapping [?]

○ BWA mapping DNA sequences against a large reference genome

● Bowtie short read aligner geared toward quickly aligning large sets of short DNA sequences (reads) to large genomes.

whether to set optional parameters?
○ default ● setting by yourself

○ Bowtie2 It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long

○ SOAP SOAP has been in evolution from a single alignment tool package that provides full solution to next generation sequencing data analysis

Peak calling [?]

● MACS S B MACS captures the influence of genome complexity to evaluate the significance of enriched ChIP region.

whether to set optional parameters?
○ default ● setting by yourself

○ MACS2 S B It is the upgraded version of MACS.

○ PeakSeq S B It is a program for identifying and ranking peak regions in ChIP-Seq experiments.

○ PeakRanger S B uses a staged algorithm to discover enriched regions and the summits within them.

○ SICER B Delineation of the significantly ChIP-enriched regions,and identification of reads on the ChIP-enriched regions.

○ PePr S B It uses a negative binomial distribution to model the read counts among the samples in the same group.

Fig. 2 The definition of the workflow

intercepted the core region of these figures, and spliced them together, as shown in Additional file 1: Fig. S2.

From this case, we carried out genome-wide analyses of YAP/TAZ-binding sites through ChIP-seq, and found that most YAP/TAZ-bound cis-regulatory regions coincided with enhancer elements, located distant from TSSs. This finding can help researchers to capture new and essential aspects of YAP/TAZ mediated transcriptional regulation.

Case study 2: motif discovery in ChIP-seq peaks

In this case, we used the dataset obtained from the study of Nature Cell Biology [41], which was described above. In their research, motif finding was crucial to find the correlation between variant transcription factors. The De novo motif finding and known motif finding were operated by the tool of findMotifs in HOMER [30]. In this study, 500 bp windows were used to search the motifs at the peak summits. The enrichment of known motifs was detected by screening the reliable motifs in HOMER motif database [44] and JASPAR database [45].

Data acquisition and processing. We reproduced the motif discovery following the method integrated in the CSA. The analysis processes were as follows. First, Supplementary Table 1 from Zanconato et al. was downloaded, the shared YAP/TAZ and TEAD4 binding sites. Second, the forth column (Chromosome), sixth column (start position), and seventh column (end position) were collected into a text file called “peak_mix.bed”. Then we used this file as the input of CSA, the appropriate genome should be hg19, and we used the default region size for motif finding: 200, and the optional parameters were chosen with the default setting.

Results visualization. Although several files were generated, here we concentrated on homerResults.html (showing the output of de novo motif finding in the form of web pages) and knownResults.html (showing the output of known motif finding in the form of web pages). From the page of homerResults.html, as shown in Additional file 1: Fig. S3, 18 de novo motifs were found, and there were two possible false positives, and motifs were ranked in accordance with the *p*-value in ascending order. The detail information of each motif was obtained by clicking the link “More Information”. On the detail information page, as shown in Additional file 1: Fig. S4, the logo of the motif and several numerical metrics were presented, and top ten known motifs that match best to this motif were listed, where the discovered de novo motif can be compared with the known motif database. Known motif databases here are the HOMER motif database and JASPAR database. From the page of knownResults.html, we can view the known motif discovery. Different from the known motifs found on the detail information page mentioned above, the known motifs here were found by comparing the regions that were contained in the bed format file to the known motif database. In addition, we also take GO enrichment analysis and KEGG pathway analysis, the results figures are shown in Additional file 1: Figs. S5 and S6.

Conclusion

In this study, we have presented the CSA web server for the whole process of ChIP-seq analysis, including the step of mapping, quality control, peak calling, and downstream analysis. CSA also provides the function of workflow, which allows users to define their own procedure. In addition, CSA visualizes mapping, peak calling and motif finding results. For the common type of ChIP-seq datasets, including histone modifications and transcription factor, CSA can provide the corresponding tool for processing them. In addition, CSA can detect differences in ChIP signals between ChIP samples and controls to identify absolute binding sites. What's more, for general ChIP-seq analysis, biologists need to perform multiple analysis steps, and each step needs different tools. Switches between different tools may take a lot of time for biologists to learn the usage of tools and convert the formats of data. Here we provide the modular design of workflows in CSA, through which users only need to provide raw data files, and select the appropriate tools and parameters, CSA can complete data analysis automatically.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3090-0.

Additional file 1: Fig. S1. YAP/TAZ/TEAD (A) binding to promoter; (B) binding to enhancers. **Fig. S2.** YAP/TAZ/TEAD binding comparison of promoters and enhancers. **Fig. S3.** CASE STUDY 2: Motif discovery in ChIP-Seq peaks. Finding de novo motifs. **Fig. S4.** CASE STUDY 2: Motif discovery in ChIP-Seq peaks. Detail information about the motif 1. **Fig. S5.** CASE STUDY 2: GO enrichment analysis for ChIP-Seq peaks. **Fig. S6.** CASE STUDY 2: KEGG pathway analysis for ChIP-Seq peaks.

Abbreviations

ChIP-seq: Chromatin immunoprecipitation sequencing; IDR: Irreproducible Discovery Rate; NGS: Next generation of sequencing technology

Acknowledgements

Not applicable.

About the Supplement

This article has been published as part of BMC Bioinformatics, Volume 20 Supplement 15, 2019: Selected articles from the 14th International Symposium on Bioinformatics Research and Applications (ISBRA-18): bioinformatics. The full contents of the supplement are available at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-15

Authors' contributions

ML carried out the integration studies of ChIP-seq analysis, participated in the design of the platform and draft the manuscript. LT carried out the coding of the platform and drafted the manuscript. FXW participated in the design of the study and helped to draft the manuscript. YP and JXW conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Funding

The support of this work and the publication costs are funded by the National Natural Science Foundation of China under Grants (No. 61732009 and No. 61622213), the 111 111 Project (No. B18059), and the Hunan Provincial Science and Technology Program (2018WK4001).

Availability of data and materials

The supplementary materials are provided, and the website of CSA is available at http://CompuBio.csu.edu.cn. The datasets used in case study are available in accession GSE66083.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer Science and Engineering, Central South University, Changsha, China. [2]Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, SKS7N5A9, Saskatoon, Canada. [3]Department of Computer Science, Georgia State University, GA30303, Atlanta, USA.

Received: 28 August 2019 Accepted: 10 September 2019
Published: 24 December 2019

References

1. Gnerre S, MacCallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci. 2011;108(4):1513–8.
2. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.
3. Liao X, Li M, Luo J, et al. Improving de novo assembly based on reads classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019. https://doi.org/10.1109/TCBB.2018.2861380.
4. Liao X, Li M, Zou Y, et al. An efficient trimming algorithm based on multi-feature fusion scoring model for NGS data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019. https://doi.org/10.1109/TCBB.2019.2897558.
5. Li M, Tang L, Wu FX, et al. SCOP: a novel scaffolding algorithm based on contig classification and optimization. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty773.
6. Li M, Tang L, Liao Z, et al. A novel scaffolding algorithm based on contig error correction and path extension. IEEE/ACM transactions on computational biology and bioinformatics. 2019;16(3):764–73.
7. Neuman JA, Isakov O, Shomron N. Analysis of insertion–deletion from deep-sequencing data: software evaluation for optimal detection. Brief Bioinform. 2012;14(1):46–55.
8. Yohe S, et al. Clinical validation of targeted next-generation sequencing for inherited disorders. Arch Pathol Lab Med. 2015;139(2):204–10.
9. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823.
10. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669.
11. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. Nat Rev Genet. 2012;13(12):840–52.

12. Li H, Durbin R. Fast and accurate long-read alignment with burrows–wheeler transform. Bioinformatics. 2010;25(5):1754–60.
13. Langmead B. Aligning short sequencing reads with bowtie. Current protocols in bioinformatics. John Wiley & Sons. Inc.:Unit. 2010;11:7.
14. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357.
15. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. Soap2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966–7.
16. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
17. Liao Y, Smyth GK, Shi W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41(10):89–94.
18. Sedlazeck FJ. Nextgenmap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013;29(21):2790–1.
19. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (macs). Genome Biol. 2008;9(9):R137.
20. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, et al. Peakseq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66–75.
21. Feng X, Grossman R, Stein L. Peakranger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics. 2011;12(1):139.
22. Xu S, Grullon S, Kai G, Peng W. Spatial clustering for identification of chip-enriched regions (sicer) to map regions of histone methylation patterns in embryonic stem cells. Methods Mol Biol. 2014;1150:97–111.
23. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM. Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics. 2008;24(15):1729.
24. Boyle AP, Guinney J, Crawford GE, Furey TS. F-seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008;24(21):2537–8.
25. Newkirk D, et al. AREM: aligning short reads from chip-sequencing by expectation maximization. International Conference on Research in Computational Molecular Biology Springer-Verlag. 2011;2011:283–97.
26. Wang J, Lunyak VV, Jordan IK. Broadpeak: a novel algorithm for identifying broad peaks in diffuse chip-seq datasets. Bioinformatics. 2013;29(4):492.
27. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA. Pepr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-seq data. Bioinformatics. 2014;30(18):2568–75.
28. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. Diffreps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. PLoS One. 2012;8(6):e65598.
29. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-dna binding sites from ChIP-seq data. Nucleic Acids Res. 2008;36(16):5221.
30. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Mol Cell. 2010;38(4):576.
31. Yu G, Wang LG, He QY. Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382.
32. Boeva V, Lermine A, Barette C, Guillouf C, Barillot E. Nebula--a web-server for advanced ChIP-seq data analysis. Bioinformatics. 2012;28(19):2517.
33. Chen TW, Li HP, Lee CC, Gan RC, Huang PJ, Wu TH, et al. Chipseek, a web-based analysis tool for chip data. BMC Genomics. 2014;15(1):539.
34. Goodarzi H, Elemento O, Tavazoie S. Revealing global regulatory perturbations across human cancers. Mol Cell. 2009;36(5):900–11.
35. Cormier N, Kolisnik T, Bieda M. Reusable, extensible, and modifiable r scripts and kepler workflows for comprehensive single set ChIP-seq analysis. BMC Bioinformatics. 2016;17(1):270.
36. 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):1293–300.
37. Ambrosini G, Dreos R, Kumar S, Bucher P. The ChIP-seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data. BMC Genomics. 2016;17(1):938.
38. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. Deeptools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42: Web Server issue), 187–91.
39. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential chip-seq analysis. Brief Bioinform. 2016;17(6):953.
40. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
41. Zanconato F, Forcato M, Battilana G, Azzolin L, Quaranta E, Bodega B, et al. Genome-wide association between yap/taz/tead and ap-1 at enhancers drives oncogenic growth. Nat Cell Biol. 2015;17(9):1218.
42. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. Ncbi geo: archive for functional genomics data sets-10 years on. Nucleic Acids Res. 2011;39(Database issue):1005–10.
43. Rhie SK, Hazelett DJ, Coetzee SG, Yan C, Noushmehr H, Coetzee GA. Nucleosome positioning and histone modifications define relationships between regulatory elements and nearby gene expression in breast epithelial cells. BMC Genomics. 2014;15(1):331.
44. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime -regulatory elements required for macrophage and b cell identities. Mol Cell. 2010;38(4):576.
45. Portalescasamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2009;38(Database issue):D105–10.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.