Cactus: a user-friendly and reproducible ATAC-Seq and mRNA-Seq analysis pipeline for data preprocessing, differential analysis, and enrichment analysis

Salignon Jérôme[1*], Lluís Millan-Ariño[1], Maxime Garcia[2,3], Christian G. Riedel[1*]

[1] Department of Bioscience and Nutrition, Karolinska Institutet, Blickagången 16, Huddinge, SE-141 83, Sweden
[2] National Genomics Infrastructure, Science for Life Laboratory, Tomtebodavägen 23A, Solna, SE-171 65, Sweden
[3] Department of Oncology-Pathology, Karolinska Institutet, Visionsgatan 4, Solna, SE-171 64, Sweden
*To whom correspondence should be addressed.

Abstract:

Summary: With the recent rapid advancements in sequencing technology, the analysis of omics data has become increasingly important for biomedical research. However, many laboratories lack the time or expertise to thoroughly analyze such data. Among omics assays, ATAC-Seq (Assay for transposase-accessible chromatin using sequencing) and mRNA-Seq (messenger RNA sequencing) are easy and cost-effective methods that are commonly used to define chromatin accessibility and gene expression levels, respectively. Here we introduce Cactus (Chromatin accessibility and transcriptomics unifying software), a pipeline that streamlines the individual or integrated analysis of ATAC-Seq and mRNA-Seq data. The pipeline is written in Nextflow for efficient scaling, caching, and parallelization, and all tools are packaged in Singularity/Docker containers or conda/Mamba virtual environments for simple installation and high reproducibility. Cactus conducts preprocessing on raw sequencing reads, followed by differential analysis between conditions. Results are split into Differential Analysis Subsets (DASs) based on significance threshold, direction of change, annotated genomic regions, and experiment type. Then, Cactus computes enrichments of entries from internal (i.e., DASs) and external (i.e., ontologies, pathways, DNA binding motifs, ChIP-Seq (Chromatin immunoprecipitation sequencing) binding sites, and chromatin states) databases in DASs and presents results in barplots and customizable heatmaps. The visualization of the results is simplified by the generation of merged PDFs, merged tables, and formatted Excel tables. In conclusion, Cactus can assist researchers in gaining comprehensive insights from chromatin accessibility and gene expression data in a quick, easy, and reproducible manner.

Availability and implementation: Cactus can be installed via Nextflow pull jsalignon/cactus. Source code and documentation are available online (https://github.com/jsalignon/cactus).

Contact: jerome.salignon@ki.se or christian.riedel@ki.se

1. Introduction

The decreasing cost of next-generation sequencing has resulted in omics-scale assays slowly becoming as affordable as conventional assays that only query a few molecules or loci (Alpern *et al.*, 2019). However, the high dimensionality of omics data remains a challenge. To address this issue, analysis pipelines are useful tools as they reduce the time, expertise, and effort needed by laboratories to analyze their omics-scale data. Specifically, modern pipelines often provide a high degree of reproducibility and efficiency (Strozzi *et al.*, 2019) through the use of workflow engines such as Nextflow (Di Tommaso *et al.*, 2017) or SnakeMake (Köster and Rahmann, 2012) and package managers, such as Docker (Merkel, 2014), Singularity (Kurtzer *et al.*, 2017), or conda (Anaconda Inc., 2015).

mRNA-Seq (Nagalakshmi *et al.*, 2008) and ATAC-Seq (Buenrostro *et al.*, 2013) are two commonly used omics assays because they are affordable, do not require large amounts of sample material, and provide a broad overview of the transcriptional-regulatory landscape of a sample. mRNA-Seq measures gene expression, while ATAC-Seq measures chromatin accessibility, both without the need to optimize the assay for a specific set of molecules or genomic regions. Many pipelines exist to analyze mRNA-Seq and ATAC-Seq data, but they commonly only go as far as peak calling for ATAC-Seq (Luo *et al.*, 2022; Grandi *et al.*, 2022) or differential gene expression analysis for mRNA-Seq (Pavlovich and Cauchy, 2022), leaving many researchers in need of deeper biological insights. Particularly, combining omics results from different methods and performing enrichment analyses is often done in a low-throughput, time-consuming, and hard-to-reproduce manner, creating a strong need for pipelines that offer easy automation of these steps.

To address this, we present Cactus, a pipeline coded in Nextflow for cache-friendly, efficient, and scalable deployment, with all tools encapsulated in containers (Docker or Singularity) or in virtual environments (Conda or Mamba (QuantStack, 2020)) for hassle-free installation and high reproducibility (Fig. 1a). Cactus is a versatile software that can be employed to analyze mRNA-Seq or ATAC-Seq data separately or jointly, allowing for the investigation of interrelated changes in chromatin accessibility and gene expression. The pipeline preprocesses data, performs differential analysis, and splits results into DASs. Enrichment analysis is then conducted on each DAS using several internal and external datasets, providing comprehensive molecular insights (Fig. 1b-d).

2. Creation of references

All the references required by Cactus are parsed by an efficient Nextflow script that uses containers to improve ease of use, speed, and reproducibility. Currently, Cactus supports the four ENCODE/modENCODE species: *H. sapiens*, *M. musculus*, *D. melanogaster,* and *C. elegans* (Dunham *et al.*, 2012; Stamatoyannopoulos *et al.*, 2012; THE MODENCODE CONSORTIUM *et al.*, 2010; Gerstein *et al.*, 2010). Genomic sequences and annotations were downloaded from Ensembl (Cunningham *et al.*, 2022). The ENCODE API (Luo *et al.*, 2020) was used to download the ChIP-Seq binding sites of 2,714 Transcription Factors (TFs) (Landt *et al.*, 2012; Boyle *et al.*, 2014) and chromatin states in the form of 899 ChromHMM profiles

(Boix *et al.*, 2021; van der Velde *et al.*, 2021) and 6 HiHMM profiles (Ho *et al.*, 2014). Slim annotations (cell, organ, development, and system) were parsed and used to create groups of ChIP-Seq binding sites that share the same annotations, allowing users to analyze only ChIP-Seq binding sites relevant to their study. 2,779 TF DNA binding motifs were obtained from the Cis-BP database (Lambert *et al.*, 2019). The R packages AnnotationHub (Morgan and Shepherd, 2021) and clusterProfiler (Yu *et al.*, 2012; Wu *et al.*, 2021) were used to obtain GO terms and KEGG pathways, respectively.

3. Data analysis workflow

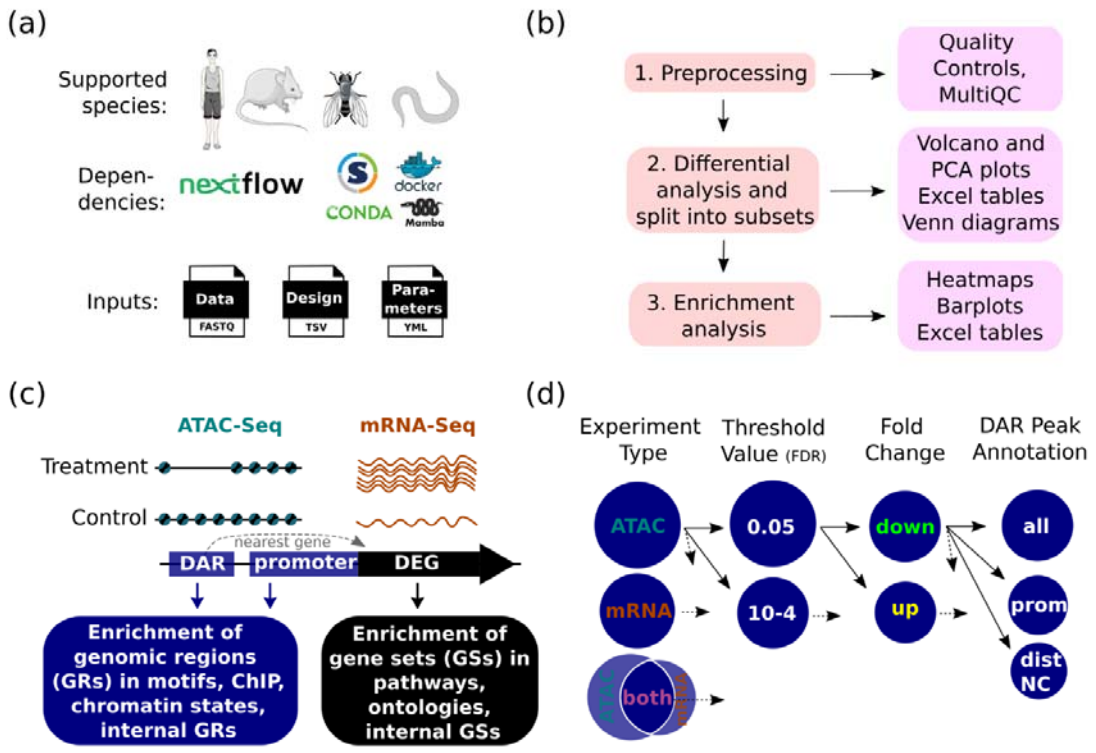

Figure 1. Overview of Cactus. (a) Key features. Species icons were adapted from (Servier; DBCLS). (b) Simplified workflow. (c) Example of enrichment analysis performed for a gene showing an increase in both chromatin accessibility and gene expression upon treatment. Enrichment of internal GRs and GSs indicates enrichment of GRs and GSs in other GRs and GSs generated by the pipeline. Black lines and blue circles represent DNA and nucleosomes, respectively. Orange lines represent mRNA molecules. (d) Sub-workflow showing the creation of DASs. Dotted arrows indicate optional additional filters. Abbreviations: DAR, differentially accessible region; DEG, differentially expressed gene; ChIP, ChIP-Seq binding sites; motifs, DNA binding motifs; FDR, false discovery rate; prom, promoter; distNC, distal non-coding region.

3.1 Preprocessing

Quality controls on raw reads are performed using FastQC (Simon Andrews, 2018) and are presented in multiQC reports (Ewels *et al.*, 2016) for both ATAC-Seq and mRNA-Seq data. For ATAC-Seq, reads are trimmed with skewer (Jiang *et al.*, 2014), compressed in parallel using pigz (Fig. 1b, Fig. S1), and aligned with bowtie2 (Langmead and Salzberg, 2012). Randomly sampled raw and aligned reads are used to rapidly generate various quality control figures that are included in the multiQC report. Low-quality, duplicated, and mitochondrial reads are removed, and insert size distributions are plotted. Bigwig tracks and correlations between samples are generated using Deeptools (Ramírez *et al.*, 2016). Aligned reads are converted to bed format and shifted to account for the shift of the transposase (Buenrostro *et al.*, 2013). Peaks are called using macs2 (Zhang *et al.*, 2008). Saturation curves are generated to indicate if re-sequencing is necessary. Peaks are split into sub-peaks and filtered for blacklisted regions, genomic DNA (i.e., input control) peaks if available, and custom hotspot regions if specified. Then, ATAC-Seq peaks are annotated with ChIPSeeker (Yu *et al.*, 2015) and summary figures are made with clusterProfiler. For mRNA-Seq, gene expression levels are quantified using kallisto (Bray *et al.*, 2016).

3.2 Differential analysis

For ATAC-Seq data, differentially accessible regions (DARs) are estimated by DiffBind (Ross-Innes et al., 2012) and these regions are then annotated to the closest gene with ChIPseeker (Fig. 1c, Fig. S2). For mRNA-Seq data, differentially expressed genes (DEGs) are determined with sleuth (Pimentel et al., 2017). Standardized tables, PCA, volcano plots, and Venn diagrams are generated. Differential analysis results are then split into subsets (i.e., DASs) (Fig. 1d, Fig. S2) to perform enrichment analyses from various angles and extract the most insights out of the data. Filters used to split data are: experiment type (ATAC, mRNA, or both) (Fig. S2), fold change type (up or down), significance level (adjusted p-value cutoff or top N entries), and DAR Peak Annotation (PA) to genomic region (currently 17 options, such as: all, promoter, gene, intergenic, etc.). The significance level of DARs by PA filters is plotted to help identify PA filters to use for downstream enrichment analysis.

3.3 Enrichment analysis

For each DAS, Cactus performs enrichment analyses among experimental conditions as well as with external databases of gene ontologies, pathways, DNA binding motifs, ChIP-Seq binding sites, and chromatin states (Fig. 1c). For all enrichment analyses, the overlaps of the target (e.g., a pathway) with the DAS and the remaining entries (differential analysis results not in the DAS) are computed. Overlaps are computed using clusterProfiler for functional annotations, Homer (Heinz *et al.*, 2010) for TF DNA binding motifs, and custom scripts for other enrichment categories. Then, two-sided Fisher's exact tests are conducted to identify enriched and depleted entries (Rivals *et al.*, 2007), and corrections for multiple testing are done using Benjamini and Hochberg's False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). Results are presented in tables, barplots, and heatmaps. Heatmaps, which display DASs that pass the same filters except the Fold Change filter, are particularly useful for

quickly identifying patterns in complex experimental designs. They can be customized by specifying user-defined groups of comparisons to plot together.

4. Implementation

Cactus is a computational tool primarily written in the Nextflow, shell, and R programming languages. Most of the tools utilized by Cactus are packaged within Biocontainers (da Veiga Leprevost *et al.*, 2017; Grüning *et al.*, 2018), which ensures consistency across container images and virtual environments. One of the key advantages of Cactus is its ability to run offline. This feature ensures stable runtime performance and allows for the analysis of sensitive patient data without any security concerns. Installation is easy, as the necessary containers or virtual environments are automatically downloaded the first time the pipeline is run. To help users become familiar with the pipeline, test datasets for each species have been created by downloading publicly available datasets with fetchngs (Ewels *et al.*, 2020) and subsampling the reads to the minimum required for running the pipeline. These test datasets and references are available on Figshare (Salignon, 2022). Before using Cactus, users should download the necessary references and, optionally, test datasets. Cactus can be executed by providing it with a YML file that specifies the run parameters, including the location of TSV files outlining the experimental design and the path to the raw FASTQ.gz data files (Fig. 1a).

5. Discussion

Cactus is a user-friendly and highly reproducible pipeline that allows researchers to perform in-depth analyses of their mRNA-Seq and/or ATAC-Seq datasets by comparing them internally and with various external datasets. The pipeline is designed for ease of use, with features such as the creation of formatted Excel tables, merged PDFs, and merged tables, while also allowing for a high degree of customization. When we applied Cactus to published datasets containing samples from *C. elegans* worms and human cells (Kolundzic *et al.*, 2018), our pipeline confirmed the main findings of the study, provided additional observations, and generated new molecular hypotheses (Supplementary Materials), arguing that Cactus performs well in real-life scenarios. We believe that Cactus will be particularly useful for laboratories with limited time or bioinformatics expertise, and that it can help to reduce the reproducibility crisis seen in many omics studies (Ma *et al.*, 2021) and science in general (Baker, 2016).

Funding sources

This work was supported by the Swedish Research Council (VR) grants 2015-03740, 2017-06088, and 2019-04868; the Swedish Cancer Society (Cancerfonden) grant 20 1034 Pj; and the Novo Nordisk Foundation grants NNF21OC0070427 and NNF22OC0078353 to C.G.R.

Acknowledgement

We would like to thank Aaron C. Daugherty and the Brunet lab for sharing the code of their 2017 *C. elegans* ATAC-Seq manuscript, which was a valuable asset during the initial development of our ATAC-Seq analysis process. Additionally, we would like to thank all the scientists who have created the databases and tools used by Cactus, i.e., contributors to ENSEMBL, members of the ENCODE and modENCODE consortia, and creators of the Cis-BP database.

Author Contributions

J.S. and C.G.R. conceived the project. J.S. developed Cactus and all related scripts, created the test datasets, the references, the documentation, and conducted the case study analysis. L.M.A. provided initial test data and feedback on the design of Cactus. M.G. implemented support for Docker, conda, and Mamba, and provided guidance on the architecture of the pipeline. J.S. and C.G.R. wrote the manuscript.

References

Alpern,D. *et al.* (2019) BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biology*, **20**, 71.

Anaconda Inc. (2015) Anaconda.

Baker,M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Boix,C.A. *et al.* (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.

Boyle,A.P. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**, 525–527.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, **10**, 1213–1218.

Cunningham,F. *et al.* (2022) Ensembl 2022. *Nucleic Acids Research*, **50**, D988–D995.

DBCLS nematode icon by DBCLS https://togotv.dbcls.jp/en/pics.html is licensed under CC-BY 4.0 Unported https://creativecommons.org/licenses/by/4.0/.

Di Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol*, **35**, 316–319.

Dunham,I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Ewels,P.A. *et al.* (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*, **38**, 276–278.

Gerstein,M.B. *et al.* (2010) Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science*, **330**, 1775–1787.

Grandi,F.C. *et al.* (2022) Chromatin accessibility profiling by ATAC-seq. *Nat Protoc*, **17**, 1518–1552.

Grüning,B. *et al.* (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*, **15**, 475–476.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576–589.

Ho,J.W.K. *et al.* (2014) Comparative analysis of metazoan chromatin organization. *Nature*, **512**, 449.

Jiang,H. *et al.* (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**, 182.

Kolundzic,E. *et al.* (2018) FACT Sets a Barrier for Cell Fate Reprogramming in Caenorhabditis elegans and Human Cells. *Developmental Cell*, **46**, 611-626.e12.

Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Kurtzer,G.M. *et al.* (2017) Singularity: Scientific containers for mobility of compute. *PLOS ONE*, **12**, e0177459.

Lambert,S.A. *et al.* (2019) Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet*, **51**, 981–989.

Landt,S.G. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357–359.

Luo,L. *et al.* (2022) Bibliometric review of ATAC-Seq and its application in gene expression. *Briefings in Bioinformatics*, **23**, bbac061.

Luo,Y. *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, **48**, D882–D889.

Ma,L. *et al.* (2021) NPARS—A Novel Approach to Address Accuracy and Reproducibility in Genomic Data Science. *Frontiers in Big Data*, **4**.

Merkel,D. (2014) Docker: lightweight Linux containers for consistent development and deployment. *Linux J.*, **2014**, 2:2.

Morgan,M. and Shepherd,L. (2021) AnnotationHub: Client to access AnnotationHub resources.

Nagalakshmi,U. *et al.* (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, **320**, 1344–1349.

Pavlovich,P.V. and Cauchy,P. (2022) Sequences to Differences in Gene Expression: Analysis of RNA-Seq Data. In, Christian,S.L. (ed), *Cancer Cell Biology: Methods and Protocols*, Methods in Molecular Biology. Springer US, New York, NY, pp. 279–318.

Pimentel,H. *et al.* (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, **14**, 687–690.

QuantStack (2020) Open Software Packaging for Science. *Medium*.

Ramírez,F. *et al.* (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, **44**, W160–W165.

Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Ross-Innes,C.S. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

Salignon,J. (2022) References and test datasets for the Cactus pipeline. DOI: 10.17044/scilifelab.20171347.v4.

Servier rickets-man-clothed, mouse-gray, and drosophila-blackeyes icons by Servier https://smart.servier.com/ are licensed under CC-BY 3.0 Unported https://creativecommons.org/licenses/by/3.0/.

Simon Andrews (2018) FastQC.

Stamatoyannopoulos,J.A. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, **13**, 418.

Strozzi,F. *et al.* (2019) Scalable Workflows and Reproducible Data Analysis for Genomics. In, Anisimova,M. (ed), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology. Springer, New York, NY, pp. 723–745.

THE MODENCODE CONSORTIUM *et al.* (2010) Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

da Veiga Leprevost,F. *et al.* (2017) BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, **33**, 2580–2582.

van der Velde,A. *et al.* (2021) Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol*, **4**, 1–15.

Wu,T. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.

Yu,G. *et al.* (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.

Yu,G. *et al.* (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, **16**, 284–287.

Zhang,Y. *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, **9**, R137.