Title:

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:

The widespread use of ATAC-Seq and mRNA-Seq generates a need for methods to thoroughly analyze the resulting data. 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 significances thresholds, 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, user-friendly, and reproducible manner.

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 (1). 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 (2) through the use of workflow engines such as Nextflow (3) or SnakeMake (4) and package managers, such as Docker (5), Singularity (6), or conda (7).

mRNA-Seq (8) and ATAC-Seq (9) 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 transcriptome and the 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 most of them only go as far as peak calling for ATAC-Seq (10–12) or differential gene expression analysis for mRNA-Seq (13, 14), 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 (15)) 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 Differential Analysis Subsets (DASs). Enrichment analysis is then conducted on each DAS using several internal and external datasets, providing comprehensive molecular insights (Fig. 1b-d).

We tested Cactus by running it on ATAC-Seq and mRNA-Seq data generated from *C. elegans* whole animals and human cells. Examples of outputs generated by Cactus derived from this analysis are presented in Figures 3 and 4 to illustrate Cactus's capabilities. The biological interpretation of the analysis is presented in the subsequent figures. Importantly, we showed that Cactus could recapitulate the main findings of the original study and even provide some novel findings. Cactus successfully identified the key roles played by FOS and Polycomb-Repressive Complexes (PRC) in regulating reprogramming, which resonated with recently published studies (16–24). Further, Cactus identified interesting new candidates regulating reprogramming and delineated their effects in different tissues.

Material and Methods

1. 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* (25–28). Genomic sequences and annotations were downloaded from Ensembl (29). The ENCODE API (30) was used to download the ChIP-Seq binding sites of 2,714 Transcription Factors (TFs) (31, 32) and chromatin states in the form of 899 ChromHMM profiles (33, 34) and 6 HiHMM profiles (35). 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 (36). The R packages AnnotationHub (37) and clusterProfiler (38, 39) were used to obtain GO terms and KEGG pathways, respectively.

2. Data analysis workflow

2.1 Preprocessing

Quality controls on raw reads are performed using FastQC (40) and are presented in multiQC reports (41) for both ATAC-Seq and mRNA-Seq data (Fig. 2). For ATAC-Seq, reads are trimmed with skewer (42), compressed in parallel using pigz6, and aligned with bowtie2 (43). Randomly sampled raw and aligned reads are used to rapidly generate various quality control figures that are included in the multiQC report (Fig. 3a). Low-quality, duplicated, and mitochondrial reads are removed, and insert size distributions are plotted. Bigwig tracks and correlations between samples are generated using Deeptools (44). Aligned reads are converted to bed format and shifted to account for the shift of the transposase (9). Peaks are called using mac7 (45). Saturation curves are generated to indicate if re-sequencing is necessary (Fig. 3b). 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 (46) and summary figures are made with clusterProfiler. For mRNA-Seq, gene expression levels are quantified using kallisto (47).

2.2 Differential analysis

For ATAC-Seq data, differentially accessible regions (DARs) are estimated by DiffBind (48) and these regions are then annotated to the closest gene with ChIPseeker (Fig. 1c, Fig. S1). For mRNA-Seq data, differentially expressed genes (DEGs) are determined with sleuth (49).

Standardized tables, PCA, volcano plots (Fig. 3c), and Venn diagrams (Fig. 3d,e) are generated. Differential analysis results are then split into subsets (i.e., DASs) (Fig. 1d, Fig. S1) 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. S1), 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 (Fig. 3f).

2.3 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 and Homer (50) for TF DNA binding motifs. We employed custom scripts for other enrichment categories, such as pathways, ontologies, chromatin states or DASs. Then, two-sided Fisher's exact tests are conducted to identify enriched and depleted entries (51), and corrections for multiple testing are done using Benjamini and Hochberg's False Discovery Rate (FDR) (52). Results are presented in tables (Fig. 4a), barplots (Fig. 4b), and heatmaps (Fig. 4d). 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 (Fig. 4c) and by multiple options (e.g., adding dots proportional to log2 odds ratios or overlap size, plotting both enrichment and depletion or enrichment only, etc…).

3. 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 (53, 54), 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 user-friendly, 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 (14) and subsampling the reads to the minimum required for running the pipeline. Test datasets can be run with one-liners as shown in the online documentation (in the Quick Start section). These test datasets and references are available on Figshare (55). 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).

Results and discussion

Comparisons with existing tools

We set out to compare Cactus with recently published ATAC-Seq analysis tools (Table 1). We found that while several tools provide motif enrichment analysis, this analysis step is almost always done on raw peaks from a single condition. However, biological experiments typically involve comparison of a given condition to another condition, such as a control. Therefore, enrichment analysis is more informative when performed on entries selected from differential analysis. To our knowledge, CoBRA (56) is the only published pipeline, besides Cactus, to provide this important functionality. Furthermore, Cactus provides several unique functionalities not provided by any other tools such as integration with mRNA-Seq data, creation of Venn diagrams, splitting DA results into subsets, enrichment in chromatin states and ChIP-Seq binding sites, creation of customized heatmaps, reporting of results in merged pdfs or tables, and the possibility to include a genomic DNA control.

Category	Feature	esATAC	Alfred	ATAC-pipe	snakePipes	nfcore/atac	AIAP	CoBRA	PEPATAC	Cactus
Pipeline architecture	Workflow manager	—	—	—	Snakemake	Nextflow	—	Snakemake	Pypiper	Nextflow
	Containers	—	✓	—	—	✓	✓	✓	✓	✓
	Virtual environments	—	✓	—	✓	✓	—	✓	✓	✓
	Fully automated analysis	✓	—	✓	—	✓	✓	✓	✓	✓
	Quality controls on reads	✓	✓	✓	✓	✓	✓	—	✓	✓
Preprocessing ATAC-Seq data	Peak annotation	✓	✓	—	✓	✓	✓	✓	✓	✓
	Saturation analysis	—	—	—	—	—	✓	—	—	✓
	Enrichment of motifs in raw ATAC-Seq peaks	✓	✓	✓	—	✓	—	—	✓	—
	Differential accessibility analysis	—	—	✓	✓	✓	✓	✓	—	✓
Differential analysis	Venn diagrams	—	—	—	—	—	—	—	—	✓
	Volcano plots	—	—	✓	—	—	—	✓	—	✓
	Integration with mRNA-seq data	—	—	—	—	—	—	—	—	✓
	Splitting DA results into subsets	—	—	—	—	—	—	—	—	✓
	Enrichment of	—	—	—	—	—	—	✓	—	✓

	motifs									
	Enrichment of CHIP-Seq binding sites	—	—	—	—	—	—	—	—	✓
	Enrichment of chromatin states	—	—	—	—	—	—	—	—	✓
	Enrichment of ontologies and pathways	—	—	—	—	—	—	✓	—	✓
	Enrichment of comparisons	—	—	—	—	—	—	— (clustering)	—	✓
	Creation of customized heatmaps	—	—	—	—	—	—	—	—	✓
Additional modules	Reporting results in merged pdfs or tables	—	—	—	—	—	—	—	—	✓
	Optional genomic DNA control	—	—	—	—	—	—	—	—	✓
	Test dataset(s)	✓	—	✓	✓	—	✓	✓	✓	✓
	Automatic download of references	✓	—	✓	✓	—	—	✓	—	✓
PMID		**29522192**	**30520945**	**29982337**	**31134269**	**32055031**	**34273560**	**34284136**	**34859208**	—

Table 1. Comparison of the main features and implementation of recently published ATAC-Seq pipelines. Abbreviations: DA, Differential Analysis.

Outputs files

The outputs of Cactus are organized in 5 directories: Figures_Individual, Figures_Merged, Tables_Individual, Tables_Merged, and Processed_data; each of which contains the following folders: 1_Preprocessing, 2_Differential_Analysis, and 3_Enrichment analysis. Cactus outputs contain figures for the pre-processing of reads (e.g., insert size, correlation, PCA) and peaks (e.g., average profile, coverage, saturation curve), differential analysis (e.g., PCA and volcano, Venn diagrams), and enrichment analysis (e.g., barplots and heatmaps for the various enrichment categories). Produced tables include detailed and filtered tables for ATAC-Seq, mRNA-Seq and DASs. Processed data include reads, peaks, and quality controls at various steps of the analysis as well as R objects (e.g., gene sets, DiffBind, ChIPseeker, sleuth, GRange, ggplot2), and genomic regions for results of differential and enrichment analysis.

Case study

Introduction to the original study: Identification of FACT as a complex regulating reprogramming in *C. elegans* and human samples

To demonstrate the utility of Cactus, we reanalyzed data from a study by the Tursun laboratory, in which a histone chaperone complex, called FAcilitates Chromatin Transcription (FACT), was identified as a cellular reprogramming barrier in *C. elegans* (57). The authors overexpressed CHE-1, a C2H2 zinc finger Transcription Factor (TF) that is essential for specifying ASE gustatory neurons (58). Overexpression of CHE-1 alone is insufficient to induce the reprogramming of non-neuronal cells into ASE neurons. Therefore, the authors screened a whole-genome RNA interference library for knockdowns that would overcome this reprogramming barrier and induce expression of an ASE neuronal reporter gene. The most important hits from this screen were *C. elegans* orthologs of FACT subunits (HMG-3, HMG-4, SPT-16) and genes known to functionally interact with FACT in other species. Interestingly, *C. elegans* orthologs of FACT subunits showed tissue-specific effects, with HMG-3 and SPT-16 acting as barriers to reprogramming in germ cells, while HMG-4 and SPT-16 acted as barriers in the intestine. This discovery was further reinforced by the subsequent finding that human FACT subunits, SSRP1 and SUPT16H, acted as barriers for reprogramming human fibroblasts into induced pluripotent stem cells.

ATAC-Seq and mRNA-Seq results described in the original study

To investigate the molecular mechanisms behind these reprogramming barriers, the authors conducted ATAC-Seq and mRNA-Seq experiments under normal or reduced levels of FACT subunits in both *C. elegans* and human fibroblasts. The results showed a strong correlation between chromatin accessibility and transcriptional changes caused by the depletion of FACT subunits in both organisms. Even though FACT was previously known to activate transcription (59), the authors found an equal number of chromatin regions opening and closing and observed equal numbers of gene induction and repression events. Moreover, they observed expression changes in genes involved in cellular reprogramming in human fibroblasts, with nine reprogramming-preventive genes being downregulated and eight reprogramming-promotive genes, including CEBP (CCAAT/Enhancer-Binding Protein), being upregulated upon FACT depletion. The authors conducted a *de novo* motif analysis on the FACT-dependent ATAC-Seq peaks to identify TFs that may bind these regions and hence act downstream. Upon FACT depletion in *C. elegans,* binding motifs of ELT-2/7, CEH-20/40, and UNC-62 were enriched in closing chromatin regions, while binding motifs of JUN-1 were enriched in opening chromatin regions. In human fibroblasts, the binding motif of RUNX1/3 was enriched in closing regions, while the binding motifs of CEBPA, CEBPE, JUN, and FOSL2 were enriched in opening regions. Interestingly, both JUN and CEBP TFs were previously shown to promote reprogramming in *C. elegans* and human cells, respectively. The authors concluded that FACT acts as both an activator and a repressor of gene expression, and that JUN and CEBP may play important roles in the FACT-dependent repression of reprogramming-promoting genes.

Cactus could recapitulate the main findings of the original study

We ran Cactus on *Kolundzic et al.*'s data using standard parameters. Thus, all results shown below represent differential analysis subsets (DASs) with adjusted p-value cutoffs of 0.05. Our reanalysis confirmed the strong correlation between changes in chromatin accessibility and changes in transcription following knockdown of FACT subunits (Fig. 5, False Discovery Rate < 10^{-100} for all comparisons). We also observed a similar number of up- and down-regulated genes in both organisms (Fig. 5a,b), thereby confirming that FACT has both gene-activatory and -repressive effects. We then specifically examined the effect of FACT depletion on the transcription of key reprogramming genes in human fibroblasts, as defined in Figure 6L of *Kolundzic et al.* Looking at the direction of change in expression for 16 genes (7 reprogramming-promotive genes and 9 reprogramming-preventive genes) upon knockdown of FACT subunits we obtained results consistent with those of the original study for 29 out of 32 entries (Table S1). This confirmed the regulation of key reprogramming genes by the FACT complex. Lastly, our analysis confirmed most of the TF binding motif enrichments in regions opened or closed by FACT, namely enrichments for ELT-2/7, JUN-1, and RUNX1/3. Only the enrichments of UNC-62 and CEH-20/40 binding motifs in *C. elegans* could not be recapitulated (Fig. 6a,c). We conclude that our Cactus analysis was able to recapitulate all major ATAC-Seq- and mRNA-Seq-derived results and conclusions of the original study.

Cactus could confidently predict a role of FOS in regulating reprogramming

Notably, the advanced analysis pipeline of Cactus allowed us to make several additional observations not described in the original study. For example, we obtained additional evidence that CEBP and JUN/FOS bind to regions opening upon FACT depletion. Kolundzic *et al.* had shown that the DNA binding motif of CEBP-1 was highly enriched in FACT-dependent regions of human cells. While Cactus reproduces this observation, it also finds the same for *C. elegans* (Fig. 6a). Additionally, Cactus found ChIP-Seq binding sites of both CEBP and FOS (a binding partner of JUN, together forming the AP-1 complex) highly enriched in FACT-dependent regions (Fig 6d). In Cactus, genes with increased chromatin accessibility and increased gene expression in one condition versus another (i.e., FACT versus wild-type control) are called High chromatin Accessibility and High Expression (HA-HE) genes (60) (Fig. S1). Strikingly, we found that FOS and CEBP were the most enriched ChIP-Seq binding sites in ATAC-Seq peaks assigned to HA-HE genes (Fig. 7a). Taken together, our observations strongly suggest that FACT represses regions that are bound by JUN/FOS and CEBP. Kolundzic *et al.* reported the enrichment of the FOSL2 motif in chromatin that opens upon SUPT16H depletion. However, this result was only reported in one supplementary figure, for one organism and one FACT subunit. In contrast, Cactus could robustly detect such pattern in both species and for all FACT subunits. This result was even more striking when jointly analyzing ATAC-Seq and mRNA-Seq data. Interestingly, recent reports have confirmed that

FOS plays a key role in the regulation of reprogramming (16–19), supporting the results from our analysis.

Cactus could identify interesting TF candidates that may cooperate with FACT to regulate reprogramming in multiple tissues

In their study, *Kolundzic et al.* provided several ATAC-Seq and mRNA-Seq datasets related to FACT and their analysis, including RAW data generated upon knockdown of the germline-specific FACT subunit HMG-3 in *C. elegans*. However, their combined analysis was missing. We now analyzed this data and found that HMG-3 depletion results in similar chromatin accessibility and gene expression changes as the depletion of other FACT subunits, e.g., HMG-4 and SPT-16, with the exception of an additional ~20,000 regions where chromatin opened only upon HMG-3 depletion (Fig. 5c). Unexpectedly, we found that the JUN/FOS and CEBP-1 DNA binding motifs were highly depleted in these regions (Fig. 6a), implying that JUN/FOS and CEBP-1 show tissue-specific differences when partnering with FACT. This raises the question of whether FACT partners with other TFs in both the gut and the germline.

In *C. elegans*, the DNA-binding motifs (Fig. 6a) or ChIP-Seq binding sites (Fig. 6b) of several TFs involved in neuronal development (ALR-1, ZTF-16, ZTF-6, LSY-27) were enriched in chromatin regions that opened upon FACT depletion. Notably, of these, LSY-27 is involved in regulating the asymmetry of ASE neurons (61), which are the neurons specified by CHE-1 overexpression. Similarly, the DNA binding motif of the forkhead transcription factor FKH-10 was highly enriched in chromatin regions closing upon FACT depletion (Fig. 6a), which resonates with recent reports showing that Forkhead TFs are versatile regulators of cellular reprogramming to pluripotency (62). In human cells, Cactus identified not only FOS and CEBP but also STAT3 (Signal Transducer and Activator of Transcription 3) as TFs whose ChIP-Seq binding sites were highly enriched in chromatin regions opening upon FACT depletion (Fig. 6d). This is remarkable because STAT3 is a well-known promoter of reprogramming (63–65).

Cactus could successfully predict a role of Polycomb-Repressive Complexes in the regulation of reprogramming

Our analysis of chromatin states showed that opened chromatin was enriched for Polycomb-repressed regions in the L3 larvae of *C. elegans* (Fig. 8a), in a human foreskin fibroblast primary cell line (Fig. 8b), and in a human lymphoblastoid cell line (Fig. 8c). In addition, we found EZH2, the core enzymatic subunit of the Polycomb Repressor Complex 2 (PRC2), SUZ12, another PRC2 subunit, and RNF2, a PRC1 subunit, to be the only TF ChIP-Seq binding sites enriched in the promoters of HA-HE genes upon depletion of either of the two FACT subunits in human cells (Fig. 7b), with the PRC2 subunits being the most strongly enriched. Altogether, these results indicate that the repressive function of FACT is achieved via Polycomb-mediated gene repression. Interestingly, a previous *C. elegans* study identified PRC2 as another barrier that hinders the conversion of germ cells to somatic neurons upon

CHE-1 overexpression (66). No connection was described between FACT and Polycomb in the Kolundzic *et al.* study. This result resonates with reports from multiple studies indicating a key role of PRC1 and PRC2 in the reprogramming process (20–24). In conclusion, genes with high expression and high chromatin accessibility upon FACT depletion include key reprogramming-promoting genes and are located in Polycomb-repressed regions.

Our analysis showed that the DNA binding motif of SNPC-4 was highly enriched in chromatin opening upon FACT depletion in *C. elegans* (Fig. 6a). Interestingly, SNPC-4 is a component of the Upstream Sequence Transcription Complex (USTC) that binds to the H3K27-methylated promoter of piRNA genes, inducing their expression in a PRC2-dependent manner (67). Furthermore, piRNAs have been recently shown to be differentially expressed specifically in reprogrammed stem cells and not in embryonic stem cells (68, 69). SNPC-4 has not been described previously as playing a role in reprogramming, and therefore, it could be an interesting candidate to evaluate for a potential role in the FACT and PRC-2-mediated regulation of reprogramming.

Conclusions of the case study

In summary, we reproduced the main findings of Kolundzic *et al.*'s study using Cactus. Our analysis confirmed the role of JUN and CEBP in reprogramming and identified new TF candidates such as FKH-10, LSY-27, and SNPC-4. Importantly, our joint analysis of ATAC-Seq and mRNA-Seq data allowed us to characterize genes with consistent changes in chromatin accessibility and expression levels upon FACT knockdown. This allowed us to discover that FACT prevents the opening of chromatin at PRC2-repressed promoters nearby regions bound by the reprogramming-promoting factors CEBP and FOS. These results are consistent with the key role of PRC2 and FOS for regulating reprogramming, as reported in recently published studies (16–24). Finally, panels from Figures 3 to 8 were generated with just three calls to Cactus, making it a user-friendly and valuable tool for researchers to identify transcription factors that may be relevant to their conditions of interest.

Limitations

The end-to-end capabilities of Cactus allow users with limited bioinformatics expertise to analyze their data in depth. However, the pipeline architecture currently has a limited flexibility in the choice of tools to use and the analysis steps that are performed. For instance, more advanced users may find it useful to perform batch correction or to use differential analysis tools other than the one used in Cactus. Nevertheless, it can be argued that advanced users may still run Cactus to effortlessly obtain a comprehensive overview of the patterns present in their data, and then develop analysis scripts customized to their dataset specificities that can recapitulate the interesting patterns they observed. Indeed, Cactus can streamline fine-grained enrichment analysis by splitting differential analysis results into various subsets, which helps in analyzing the data from various angles.

Another limitation of Cactus is that it supports only four commonly used species, which limits its use in other species. This design was adopted to leverage the comprehensive datasets generated by the ENCODE and modENCODE consortia, such as blacklisted regions, chromatin states, and ChIP-Seq binding sites. To remedy this, future versions of Cactus could include additional supported species, for which the use of blacklisted regions, as well as the analysis of chromatin states and ChIP-Seq binding sites, will be disabled.

Conclusions

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. We showed that Cactus stands out among similar tools by having several unique and important features, notably with its capacity to conduct various enrichment analyses on DASs and to report results in customizable heatmaps. When we applied Cactus to published datasets containing samples from *C. elegans* worms and human cells (57), our pipeline confirmed the main findings of the study, provided additional observations, and generated new molecular hypotheses, arguing that Cactus performs well in real-life scenarios. We anticipate 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 (70) and science in general (71).

Data availability

The data used in this article are available in the Gene Expression Omnibus (GEO) database at https://www.ncbi.nlm.nih.gov/geo under the GEO accession GSE98758. The Cactus code and documentation are available on Github at https://github.com/jsalignon/cactus as well as in the Zenodo open repository as a release archive at 10.5281/zenodo.10226966 . Test datasets and references are available as a release archive in the Figshare SciLifeLab Data Repository at 10.17044/scilifelab.20171347.v4.

Funding

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. Finally, we would like to thank Paolo Di Tommaso for providing technical support and guidance when developing Nextflow scripts.

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

1. Alpern,D., Gardeux,V., Russeil,J., Mangeat,B., Meireles-Filho,A.C.A., Breysse,R., Hacker,D. and Deplancke,B. (2019) BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biology*, **20**, 71.

2. Strozzi,F., Janssen,R., Wurmus,R., Crusoe,M.R., Githinji,G., Di Tommaso,P., Belhachemi,D., Möller,S., Smant,G., de Ligt,J., *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.

3. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol*, **35**, 316–319.

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

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

6. Kurtzer,G.M., Sochat,V. and Bauer,M.W. (2017) Singularity: Scientific containers for mobility of compute. *PLOS ONE*, **12**, e0177459.

7. Anaconda Inc. (2015) Anaconda.

8. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, **320**, 1344–1349.

9. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (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.

10. Luo,L., Gribskov,M. and Wang,S. (2022) Bibliometric review of ATAC-Seq and its application in gene expression. *Briefings in Bioinformatics*, **23**, bbac061.

11. Grandi,F.C., Modi,H., Kampman,L. and Corces,M.R. (2022) Chromatin accessibility profiling by ATAC-seq. *Nat Protoc*, **17**, 1518–1552.

12. Freeman,T.F., Zhao,Q., Surya,A., Rothe,R. and Cenik,E.S. (2023) Ribosome biogenesis disruption mediated chromatin structure changes revealed by SRAtac, a customizable end to end analysis pipeline for ATAC-seq. *BMC Genomics*, **24**, 512.

13. 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.

14. Ewels,P.A., Peltzer,A., Fillinger,S., Patel,H., Alneberg,J., Wilm,A., Garcia,M.U., Di Tommaso,P. and Nahnsen,S. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*, **38**, 276–278.

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

16. Gomes,A.M., Kurochkin,I., Chang,B., Daniel,M., Law,K., Satija,N., Lachmann,A., Wang,Z., Ferreira,L., Ma'ayan,A., *et al.* (2018) Cooperative Transcription Factor Induction Mediates Hemogenic Reprogramming. *Cell Reports*, **25**, 2821-2835.e7.

17. Pecce,V., Verrienti,A., Fiscon,G., Sponziello,M., Conte,F., Abballe,L., Durante,C., Farina,L., Filetti,S. and Paci,P. (2021) The role of FOSL1 in stem-like cell reprogramming processes. *Sci Rep*, **11**, 14677.

18. Kamimoto,K., Adil,M.T., Jindal,K., Hoffmann,C.M., Kong,W., Yang,X. and Morris,S.A. (2023) Gene regulatory network reconfiguration in direct lineage reprogramming. *Stem Cell Reports*, **18**, 97–112.

19. Buckberry,S., Liu,X., Poppe,D., Tan,J.P., Sun,G., Chen,J., Nguyen,T.V., de Mendoza,A., Pflueger,J., Frazer,T., *et al.* (2023) Transient naive reprogramming corrects hiPS cells functionally and epigenetically. *Nature*, **620**, 863–872.

20. Pereira,C.F., Piccolo,F.M., Tsubouchi,T., Sauer,S., Ryan,N.K., Bruno,L., Landeira,D., Santos,J., Banito,A., Gil,J., *et al.* (2010) ESCs Require PRC2 to Direct the Successful Reprogramming of Differentiated Cells toward Pluripotency. *Cell Stem Cell*, **6**, 547–556.

21. Yamaji,M., Ueda,J., Hayashi,K., Ohta,H., Yabuta,Y., Kurimoto,K., Nakato,R., Yamada,Y., Shirahige,K. and Saitou,M. (2013) PRDM14 Ensures Naive Pluripotency through Dual Regulation of Signaling and Epigenetic Pathways in Mouse Embryonic Stem Cells. *Cell Stem Cell*, **12**, 368–382.

22. Macrae,T.A. and Ramalho-Santos,M. (2021) The deubiquitinase Usp9x regulates PRC2-mediated chromatin reprogramming during mouse development. *Nat Commun*, **12**, 1865.

23. Zijlmans,D.W., Talon,I., Verhelst,S., Bendall,A., Van Nerum,K., Javali,A., Malcolm,A.A., van Knippenberg,S.S.F.A., Biggins,L., To,S.K., *et al.* (2022) Integrated multi-omics reveal polycomb repressive complex 2 restricts human trophoblast induction. *Nat Cell Biol*, **24**, 858–871.

24. Cohen,L.R.Z. and Meshorer,E. (2023) H3K36 methylation is a reprogramming barrier. *Nat Cell Biol*, **25**, 1077–1078.

25. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

26. Stamatoyannopoulos,J.A., Snyder,M., Hardison,R., Ren,B., Gingeras,T., Gilbert,D.M., Groudine,M., Bender,M., Kaul,R., Canfield,T., *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, **13**, 418.

27. THE MODENCODE CONSORTIUM, Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L., *et al.* (2010) Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

28. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K., *et al.* (2010) Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science*, **330**, 1775–1787.

29. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Bennett,R., *et al.* (2022) Ensembl 2022. *Nucleic Acids Research*, **50**, D988–D995.

30. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K., *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, **48**, D882–D889.

31. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P., *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

32. Boyle,A.P., Araya,C.L., Brdlik,C., Cayting,P., Cheng,C., Cheng,Y., Gardner,K., Hillier,L.W., Janette,J., Jiang,L., *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.

33. Boix,C.A., James,B.T., Park,Y.P., Meuleman,W. and Kellis,M. (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.

34. van der Velde,A., Fan,K., Tsuji,J., Moore,J.E., Purcaro,M.J., Pratt,H.E. and Weng,Z. (2021) Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol*, **4**, 1–15.

35. Ho,J.W.K., Jung,Y.L., Liu,T., Alver,B.H., Lee,S., Ikegami,K., Sohn,K.-A., Minoda,A., Tolstorukov,M.Y., Appert,A., *et al.* (2014) Comparative analysis of metazoan chromatin organization. *Nature*, **512**, 449.

36. Lambert,S.A., Yang,A.W.H., Sasse,A., Cowley,G., Albu,M., Caddick,M.X., Morris,Q.D., Weirauch,M.T. and Hughes,T.R. (2019) Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet*, **51**, 981–989.

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

38. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, **16**, 284–287.

39. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.

40. Simon Andrews (2018) FastQC.

41. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

42. Jiang,H., Lei,R., Ding,S.-W. and Zhu,S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**, 182.

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

44. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, **44**, W160–W165.

45. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, **9**, R137.

46. Yu,G., Wang,L.-G. and He,Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.

47. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**, 525–527.

48. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R., *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

49. Pimentel,H., Bray,N.L., Puente,S., Melsted,P. and Pachter,L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, **14**, 687–690.

50. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576–589.

51. Rivals,I., Personnaz,L., Taing,L. and Potier,M.-C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

52. 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.

53. da Veiga Leprevost,F., Grüning,B.A., Alves Aflitos,S., Röst,H.L., Uszkoreit,J., Barsnes,H., Vaudel,M., Moreno,P., Gatto,L., Weber,J., *et al.* (2017) BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, **33**, 2580–2582.

54. Grüning,B., Dale,R., Sjödin,A., Chapman,B.A., Rowe,J., Tomkins-Tinch,C.H., Valieris,R. and Köster,J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*, **15**, 475–476.

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

56. Qiu,X., Feit,A.S., Feiglin,A., Xie,Y., Kesten,N., Taing,L., Perkins,J., Gu,S., Li,Y., Cejas,P., *et al.* (2021) CoBRA: Containerized Bioinformatics Workflow for Reproducible ChIP/ATAC-Seq Analysis. *Genomics, Proteomics & Bioinformatics*, **19**, 652–661.

57. Kolundzic,E., Ofenbauer,A., Bulut,S.I., Uyar,B., Baytek,G., Sommermeier,A., Seelk,S., He,M., Hirsekorn,A., Vucicevic,D., *et al.* (2018) FACT Sets a Barrier for Cell Fate Reprogramming in Caenorhabditis elegans and Human Cells. *Developmental Cell*, **46**, 611-626.e12.

58. Etchberger,J.F., Lorch,A., Sleumer,M.C., Zapf,R., Jones,S.J., Marra,M.A., Holt,R.A., Moerman,D.G. and Hobert,O. (2007) The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes Dev.*, **21**, 1653–1674.

59. Reddy,B.A., Jeronimo,C. and Robert,F. (2017) Recent Perspectives on the Roles of Histone Chaperones in Transcription Regulation. *Curr Mol Bio Rep*, **3**, 1–10.

60. Nair,V.D., Vasoya,M., Nair,V., Smith,G.R., Pincas,H., Ge,Y., Douglas,C.M., Esser,K.A. and Sealfon,S.C. (2021) Differential analysis of chromatin accessibility and gene expression profiles identifies cis-regulatory elements in rat adipose and muscle. *Genomics*, **113**, 3827–3841.

61. Zhang,F., O'Meara,M.M. and Hobert,O. (2011) A left/right asymmetric neuronal differentiation program is controlled by the Caenorhabditis elegans lsy-27 zinc-finger transcription factor. *Genetics*, **188**, 753–759.

62. Fu,M., Chen,H., Cai,Z., Yang,Y., Feng,Z., Zeng,M., Chen,L., Qin,Y., Cai,B., Zhu,P., *et al.* (2021) Forkhead box family transcription factors as versatile regulators for cellular reprogramming to pluripotency. *Cell Regeneration*, **10**, 17.

63. Yang,J., van Oosten,A.L., Theunissen,T.W., Guo,G., Silva,J.C.R. and Smith,A. (2010) Stat3 Activation Is Limiting for Reprogramming to Ground State Pluripotency. *Cell Stem Cell*, **7**, 319–328.

64. Tang,Y., Luo,Y., Jiang,Z., Ma,Y., Lin,C.-J., Kim,C., Carter,M.G., Amano,T., Park,J., Kish,S., *et al.* (2012) Jak/Stat3 Signaling Promotes Somatic Cell Reprogramming by Epigenetic Regulation. *Stem Cells*, **30**, 2645–2656.

65. Yu,Z., Dmitrieva,N.I., Walts,A.D., Jin,H., Liu,Y., Ping,X., Ferrante,E.A., Qiu,L., Holland,S.M., Freeman,A.F., *et al.* (2020) STAT3 modulates reprogramming efficiency of human somatic cells; insights from autosomal dominant Hyper IgE syndrome caused by STAT3 mutations. *Biology Open*, **9**, bio052662.

66. Patel,T., Tursun,B., Rahe,D.P. and Hobert,O. (2012) Removal of Polycomb Repressive Complex 2 Makes C. elegans Germ Cells Susceptible to Direct Conversion into Specific Somatic Cell Types. *Cell Reports*, **2**, 1178–1186.

67. Huang,X., Cheng,P., Weng,C., Xu,Z., Zeng,C., Xu,Z., Chen,X., Zhu,C., Guang,S. and Feng,X. (2021) A chromodomain protein mediates heterochromatin-directed piRNA expression. *Proceedings of the National Academy of Sciences*, **118**, e2103723118.

68. Zhu,Y., Fan,C. and Zhao,B. (2019) Differential expression of piRNAs in reprogrammed pluripotent stem cells from mouse embryonic fibroblasts. *IUBMB Life*, **71**, 1906–1915.

69. Salloum-Asfar,S., Abdulla,S.A., Taha,R.Z., Thompson,I.R. and Emara,M.M. (2022) Combined Noncoding RNA-mRNA Regulomics Signature in Reprogramming and Pluripotency in iPSCs. *Cells*, **11**, 3833.

70. Ma,L., Peterson,E.A., Shin,I.J., Muesse,J., Marino,K., Steliga,M.A. and Johann,D.J. (2021) NPARS—A Novel Approach to Address Accuracy and Reproducibility in Genomic Data Science. *Frontiers in Big Data*, **4**.

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

72. 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/.

73. 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/.

Figure legends

Figure 1. Overview of Cactus. (a) Key features. Species icons were adapted from (72, 73). (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 genomic regions (GRs) and gene sets (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.

Figure 2. Detailed workflow. The ellipses show the different analysis steps, with the utilized tool(s) written in italics. For each analysis step, the incoming arrows indicate the sources of the input files, and the outgoing arrows indicate the next analysis step(s) to which the output files are passed. Dotted arrows indicate optional connections. The numbers in the legend indicate Cactus' three main parts: 1. Preprocessing, 2. Differential Analysis, and 3. Enrichment Analysis. Abbreviations: ATAC, ATAC-Seq; mRNA, mRNA-Seq; DA, differential analysis; QC, quality controls; DASs, differential analysis subsets; ET, experiment type; TV, threshold value; FC, fold change; DAR, differentially accessible regions; PA, DAR peak annotation; KEGG, Kyoto Encyclopedia of Genes and Genomes; and GO, gene ontology.

Figure 3. Example of outputs for the preprocessing and differential analysis steps. (a) Percentage of reads aligned to exons for 6 samples. (b) Saturation analysis for one replicate of the control condition. Panels (c-f) show figures for the comparison SSRP1 knockdown vs

control. (c) Volcano plot. (d) Four-way Venn diagram showing the overlap between DEG- and DAR-associated genes. Please observe that multiple DARs of opening and closing chromatin can be associated to the same gene which explains the non-empty overlaps of the atac_up and atac_down sets. (e) Proportional two-way Venn diagram showing the overlap between mRNA-Seq and ATAC-Seq DASs. (f) Significance level of DARs assigned to each of the 17 available DAR Peak Annotation filters. Abbreviations: ctl, control.

Figure 4. Example of outputs for the enrichment analysis steps. Panels (a-b) show KEGG enrichment analysis results for the comparison SSRP1 knockdown vs control. (a) Formatted Excel table. (b) Barplot. (c) Content of the TSV file used to define the comparisons to make between conditions and the groups of comparisons to plot together in heatmaps. (d) Heatmaps of the enrichment between comparisons for the "all" group and the "supt16h" group defined in (c). The "ctl" group is shown in Figure 5d.

Figure 5. Enrichment of DASs in other DASs. Heatmaps showing the overlap between DEGs from mRNA-Seq (a,b) and DARs from ATAC-Seq (c,d) for *C. elegans* (a,c) and human (b,d) samples. Numbers and colors indicate the size and significance of the overlap, respectively. Colors show the signed minus $\log_{10}$ adjusted p-values (from two-sided Fisher tests), with positive values (green color) indicating enrichment and negative values (purple color) indicating depletion. The bin [0.699, 1.3] corresponds to the adjusted p-value interval [0.2, 0.05]. Abbreviation: ctl, control.

Figure 6. Enrichment of TFs CHIP-Seq binding sites and DNA binding motifs in DARs. Colors indicate the signed minus $\log_{10}$ adjusted p-values (from two-sided Fisher tests), with positive values (green color) indicating enrichment and negative values (purple color) indicating depletion. The lowest enrichment bin [0.699, 1.3] corresponds to the adjusted p-value interval [0.2, 0.05]. For (b) and (d), the y-axis labels are in the following format: protein_*C. elegans* developmental stage (b) or protein_cell line (d). Details about the cell lines in which ChIP-Seq experiments were conducted are available in the reference section of the Cactus documentation. Abbreviation: ctl, control.

Figure 7. Enrichment of TFs CHIP-Seq binding sites in DASs from HA-HE and LA-LE genes. Enrichment of DARs (a) and promoters (b) of HA-HE (High chromatin Accessibility and High Expression) and LA-LE (Low chromatin Accessibility and Low Expression) genes upon FACT knockdown in TFs CHIP-Seq binding sites. Colors indicate the signed minus $\log_{10}$ adjusted p-values (from two-sided Fisher tests), with positive values (green color) indicating enrichment and negative values (purple color) indicating depletion. The lowest enrichment bin [0.699, 1.3] corresponds to the adjusted p-value interval [0.2, 0.05]. The y-axis labels are in the following format: Protein_Cell line. Details about the cell lines in which ChIP-Seq

experiments were conducted are available in the reference section of the Cactus documentation. Abbreviation: ctl, control.

Figure 8. Enrichment of chromatin states in DARs. The chromatin state files used are a 17-state HiHMM model (35) for L3 *C. elegans* worms (a), a 17-state HiHMM model for B lymphocyte human cell line (b, cell line GM12878), and an 18-state ChromHMM model (33) for human primary foreskin fibroblast (c, ENCODE accession ENCSR251QDX). Colors indicate the signed minus $\log_{10}$ adjusted p-values (from two-sided Fisher tests), with positive values (green color) indicating enrichment and negative values (purple color) indicating depletion. The lowest enrichment bin [0.699, 1.3] corresponds to the adjusted p-value interval [0.2, 0.05]. Abbreviation: ctl, control; PC, Polycomb.

Figure 1

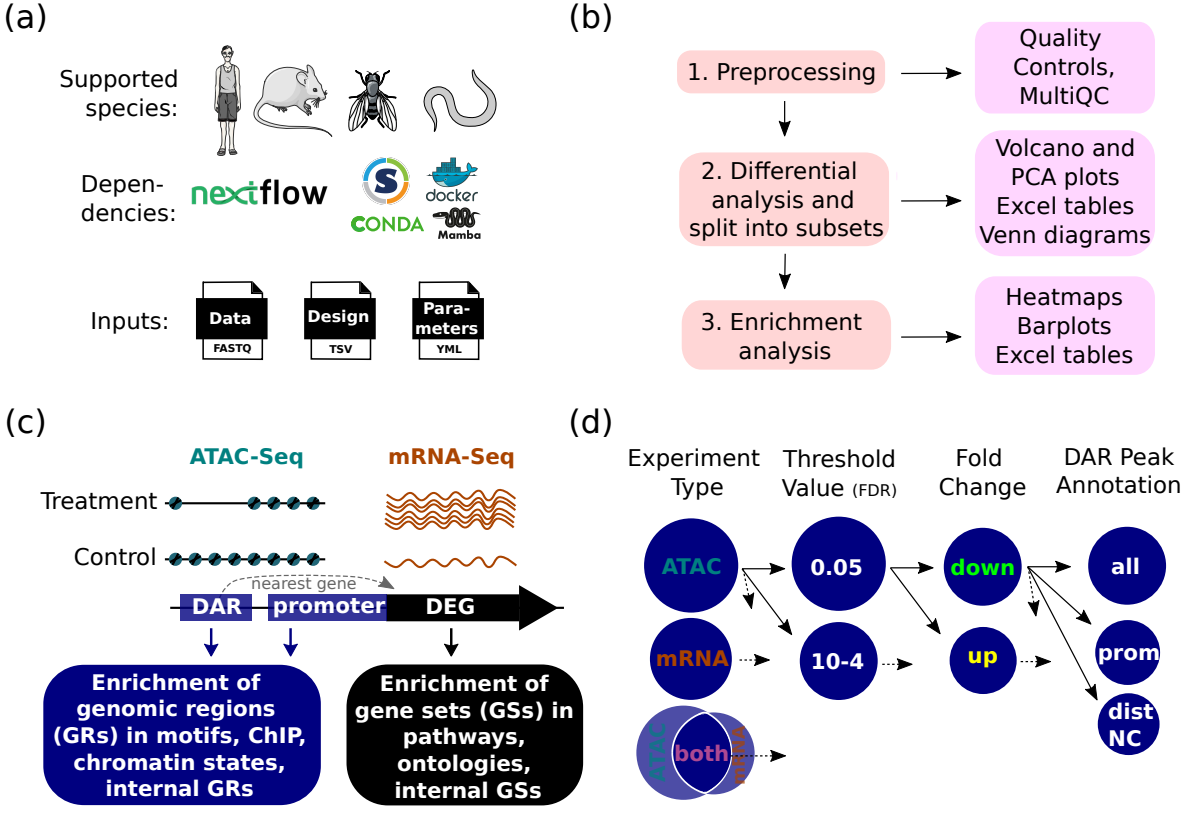
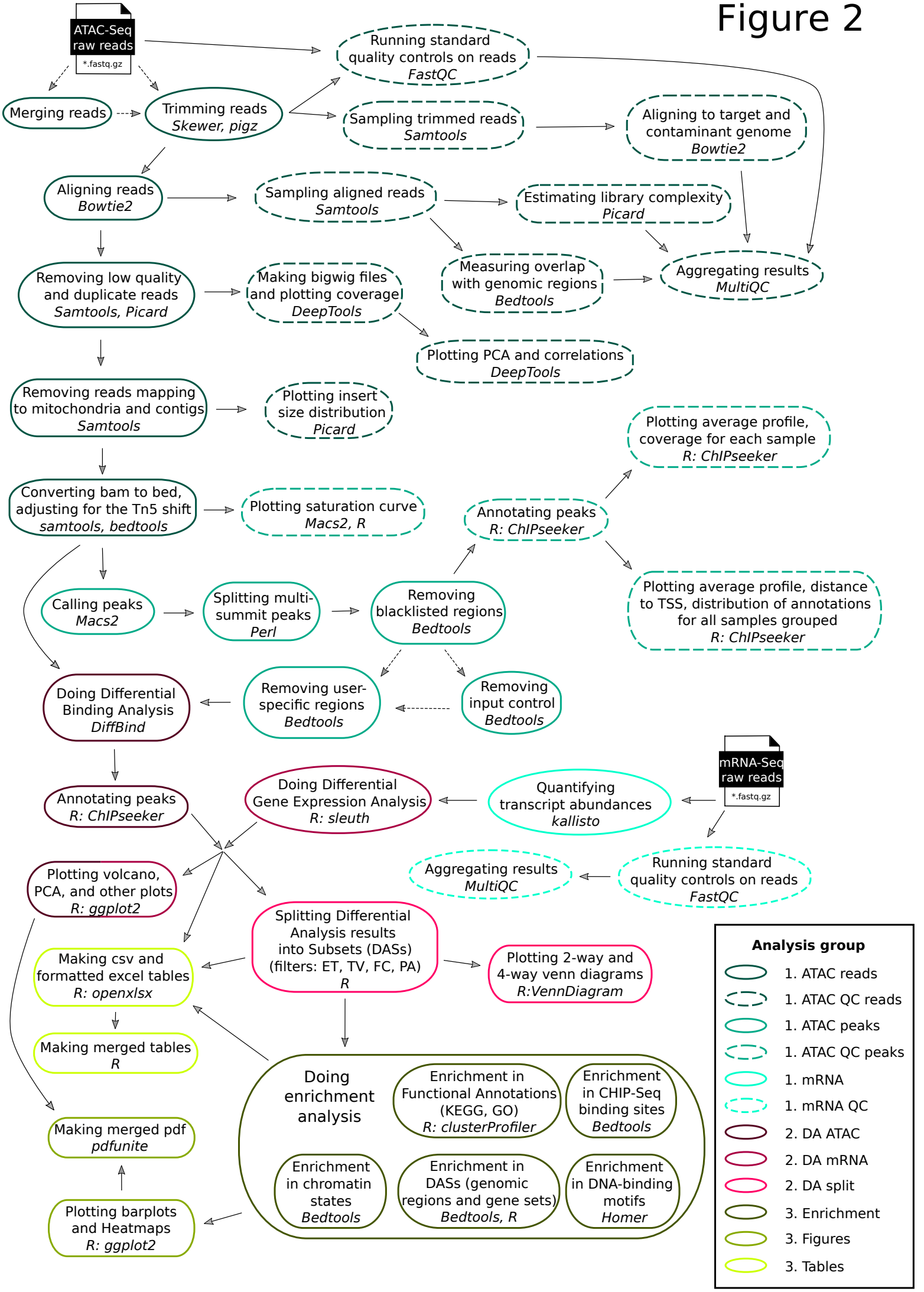

(a)
Supported species:
Depen-dencies:
nextflow
CONDA
docker
Mamba
Inputs:
Data
FASTQ
Design
TSV
Para-meters
YML
(b)
1. Preprocessing
Quality Controls, MultiQC
2. Differential analysis and split into subsets
Volcano and PCA plots
Excel tables
Venn diagrams
3. Enrichment analysis
Heatmaps
Barplots
Excel tables
(c)
ATAC-Seq
mRNA-Seq
Treatment
Control
nearest gene
DAR
promoter
DEG
Enrichment of genomic regions (GRs) in motifs, ChIP, chromatin states, internal GRs
Enrichment of gene sets (GSs) in pathways, ontologies, internal GSs
(d)
Experiment Type
Threshold Value (FDR)
Fold Change
DAR Peak Annotation
ATAC
mRNA
both
0.05
10-4
down
up
all
prom
dist
NC

Figure 2


ATAC-Seq raw reads
*.fastq.gz
Running standard quality controls on reads FastQC
Merging reads
Trimming reads Skewer, pigz
Sampling trimmed reads Samtools
Aligning to target and contaminant genome Bowtie2
Aligning reads Bowtie2
Sampling aligned reads Samtools
Estimating library complexity Picard
Removing low quality and duplicate reads Samtools, Picard
Making bigwig files and plotting coverage DeepTools
Measuring overlap with genomic regions Bedtools
Aggregating results MultiQC
Plotting PCA and correlations DeepTools
Removing reads mapping to mitochondria and contigs Samtools
Plotting insert size distribution Picard
Plotting average profile, coverage for each sample R: ChIPseeker
Converting bam to bed, adjusting for the Tn5 shift samtools, bedtools
Plotting saturation curve Macs2, R
Annotating peaks R: ChIPseeker
Calling peaks Macs2
Splitting multi-summit peaks Perl
Removing blacklisted regions Bedtools
Plotting average profile, distance to TSS, distribution of annotations for all samples grouped R: ChIPseeker
Doing Differential Binding Analysis DiffBind
Removing user-specific regions Bedtools
Removing input control Bedtools
mRNA-Seq raw reads
*.fastq.gz
Annotating peaks R: ChIPseeker
Doing Differential Gene Expression Analysis R: sleuth
Quantifying transcript abundances kallisto
Aggregating results MultiQC
Running standard quality controls on reads FastQC
Plotting volcano, PCA, and other plots R: ggplot2
Splitting Differential Analysis results into Subsets (DASs) (filters: ET, TV, FC, PA) R
Plotting 2-way and 4-way venn diagrams R:VennDiagram
Making csv and formatted excel tables R: openxlsx
Making merged tables R
Doing enrichment analysis
Enrichment in Functional Annotations (KEGG, GO) R: clusterProfiler
Enrichment in CHIP-Seq binding sites Bedtools
Enrichment in chromatin states Bedtools
Enrichment in DASs (genomic regions and gene sets) Bedtools, R
Enrichment in DNA-binding motifs Homer
Making merged pdf pdfunite
Plotting barplots and Heatmaps R: ggplot2
Analysis group
1. ATAC reads
1. ATAC QC reads
1. ATAC peaks
1. ATAC QC peaks
1. mRNA
1. mRNA QC
2. DA ATAC
2. DA mRNA
2. DA split
3. Enrichment
3. Figures
3. Tables

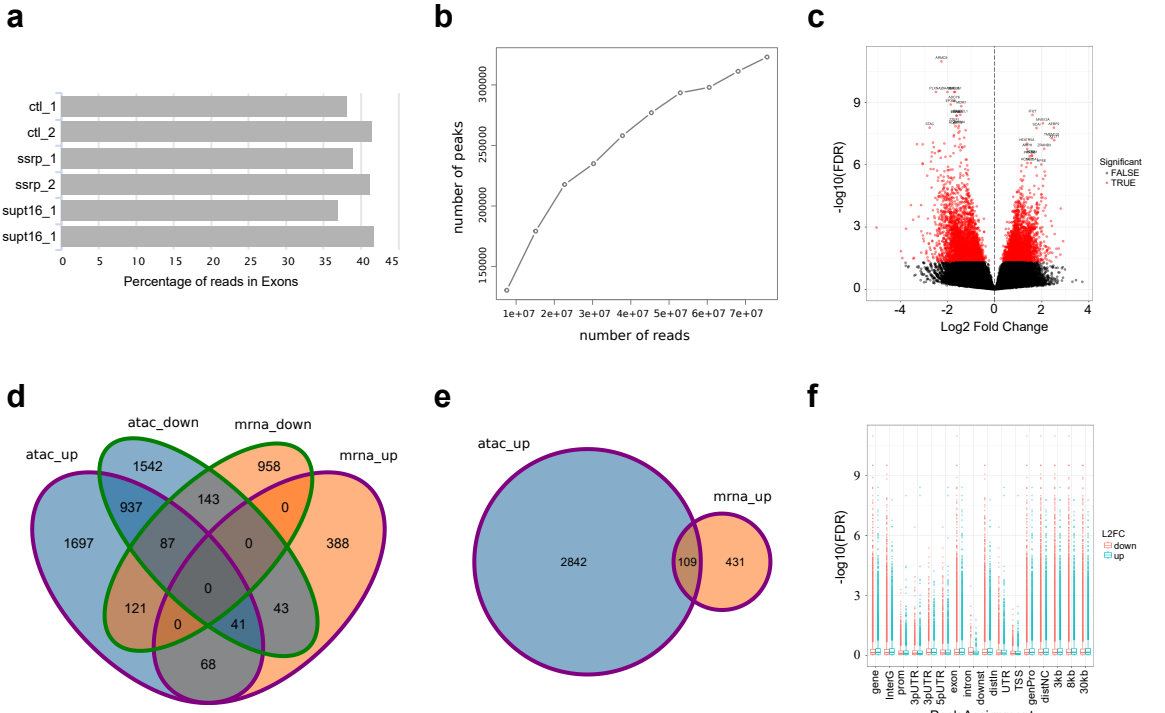

a
ctl_1
ctl_2
ssrp_1
ssrp_2
supt16_1
supt16_1
0 5 10 15 20 25 30 35 40 45
Percentage of reads in Exons
b
number of peaks
150000 200000 250000 300000
1e+07 2e+07 3e+07 4e+07 5e+07 6e+07 7e+07
number of reads
c
-log10(FDR)
0 3 6 9
-4 -2 0 2 4
Log2 Fold Change
Significant
FALSE
TRUE
d
atac_up
atac_down
mrna_down
mrna_up
1697
1542
958
388
937
143
0
87
0
0
121
0
41
43
68
e
atac_up
mrna_up
2842
109
431
f
-log10(FDR)
0 3 6 9
gene
IntergC
prom
3pUTR
3pUTR
5pUTR
exon
intron
downst
distin
UTR
TSS
genPro
distNC
2kb
8kb
30kb
Peak Assignment
L2FC
down
up

Figure 3

a

GE	ET	PA	FC	TV	COMP	tgt	pval	padj	L2OR	pt_da	ov_da	tot_da	pt_nda	ov_nda	tot_nda	tgt_id
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Olfactory transduction	1.4E-13	4.5E-11	-2.35	1.09	13	1188	5.34	439	8216	hsa04740
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	TGF-beta signaling pathway	2.9E-07	3.3E-05	1.54	3.28	39	1188	1.16	95	8216	hsa04350
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Focal adhesion	4E-07	3.3E-05	1.16	5.3	63	1188	2.45	201	8216	hsa04510
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Proteoglycans in cancer	3.2E-07	3.3E-05	1.15	5.39	64	1188	2.5	205	8216	hsa05205
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Axon guidance	1.5E-06	9.7E-05	1.15	4.8	57	1188	2.22	182	8216	hsa04360
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Adherens junction	2.6E-06	0.00014	1.62	2.61	31	1188	0.86	71	8216	hsa04520
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Wnt signaling pathway	3.3E-06	0.00016	1.14	4.46	53	1188	2.07	170	8216	hsa04310
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Rap1 signaling pathway	4.3E-06	0.00018	1.04	5.13	61	1188	2.56	210	8216	hsa04015
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Thyroid hormone signaling pathway	1.1E-05	0.00041	1.26	3.45	41	1188	1.47	121	8216	hsa04919
KEGG	ATAC	all	down	1.3	ssrp1_vs_ctl	Small cell lung cancer	5.6E-05	0.0019	1.29	2.69	32	1188	1.12	92	8216	hsa05222

b

c

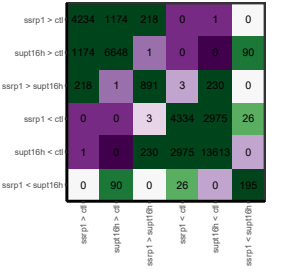
comparisons.tsv:

ssrp1	ctl
supt16	ctl
ssrp1	supt16

groups.tsv:

all	ssrp1_vs_ctl supt16_vs_ctl ssrp1_vs_supt16
ctl	ssrp1_vs_ctl supt16_vs_ctl
supt16	supt16_vs_ctl ssrp1_vs_supt16

d



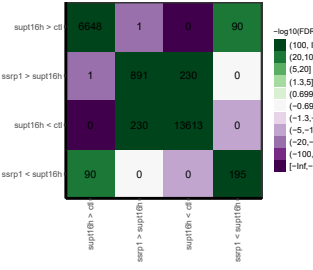


Figure 4

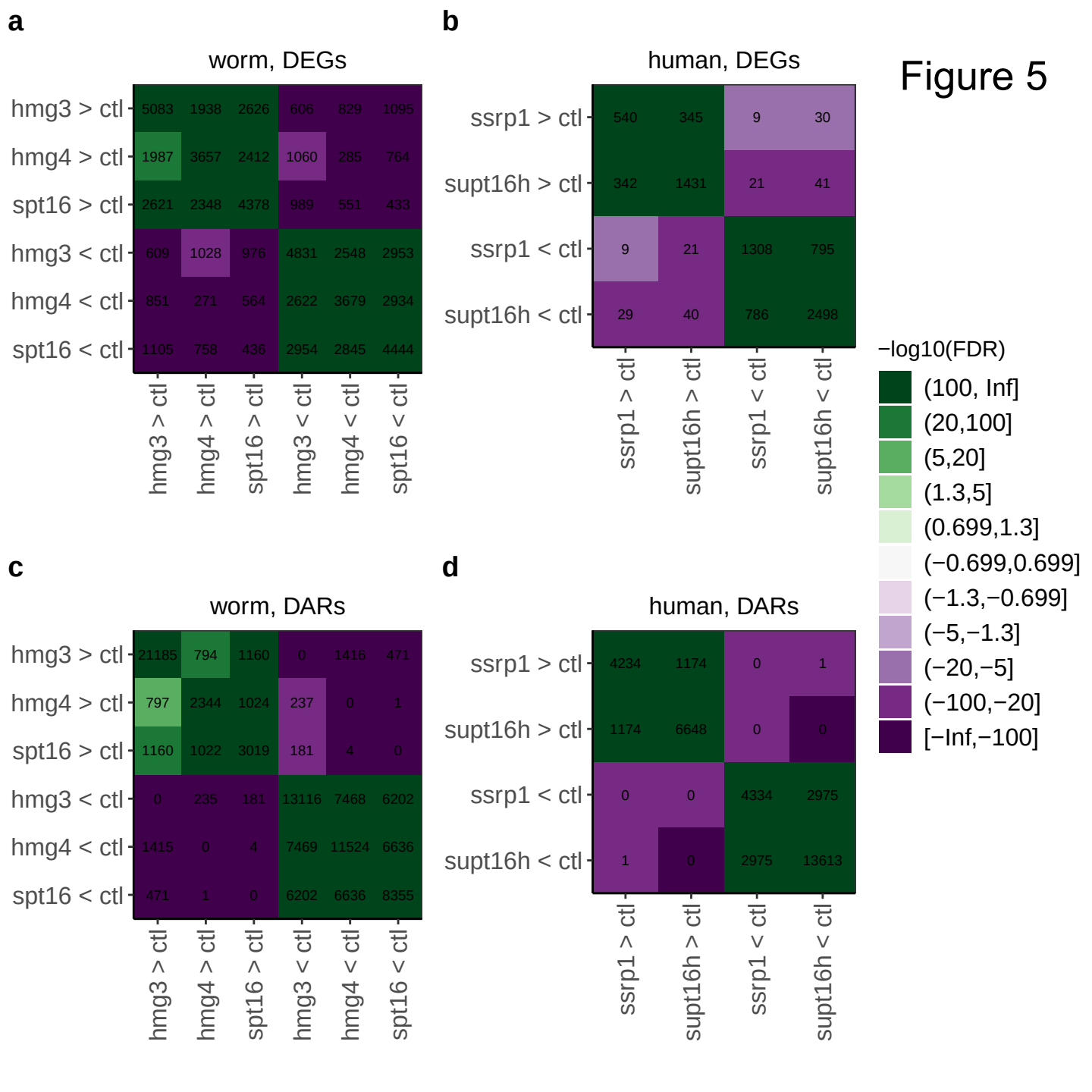


Figure 5
a
worm, DEGs
hmg3 > ctl
hmg4 > ctl
spt16 > ctl
hmg3 < ctl
hmg4 < ctl
spt16 < ctl
5083 1938 2626 606 829 1095
1987 3657 2412 1060 285 764
2621 2348 4378 989 551 433
609 1028 976 4831 2548 2953
851 271 564 2622 3679 2934
1105 758 436 2954 2845 4444
b
human, DEGs
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
540 345 9 30
342 1431 21 41
9 21 1308 795
29 40 786 2498
c
worm, DARs
21185 794 1160 0 1416 471
797 2344 1024 237 0 1
1160 1022 3019 181 4 0
0 235 181 13116 7468 6202
1415 0 4 7469 11524 6636
471 1 0 6202 6636 8355
d
human, DARs
4234 1174 0 1
1174 6648 0 0
0 0 4334 2975
1 0 2975 13613
−log10(FDR)
(100, Inf]
(20,100]
(5,20]
(1.3,5]
(0.699,1.3]
(−0.699,0.699]
(−1.3,−0.699]
(−5,−1.3]
(−20,−5]
(−100,−20]
[−Inf,−100]

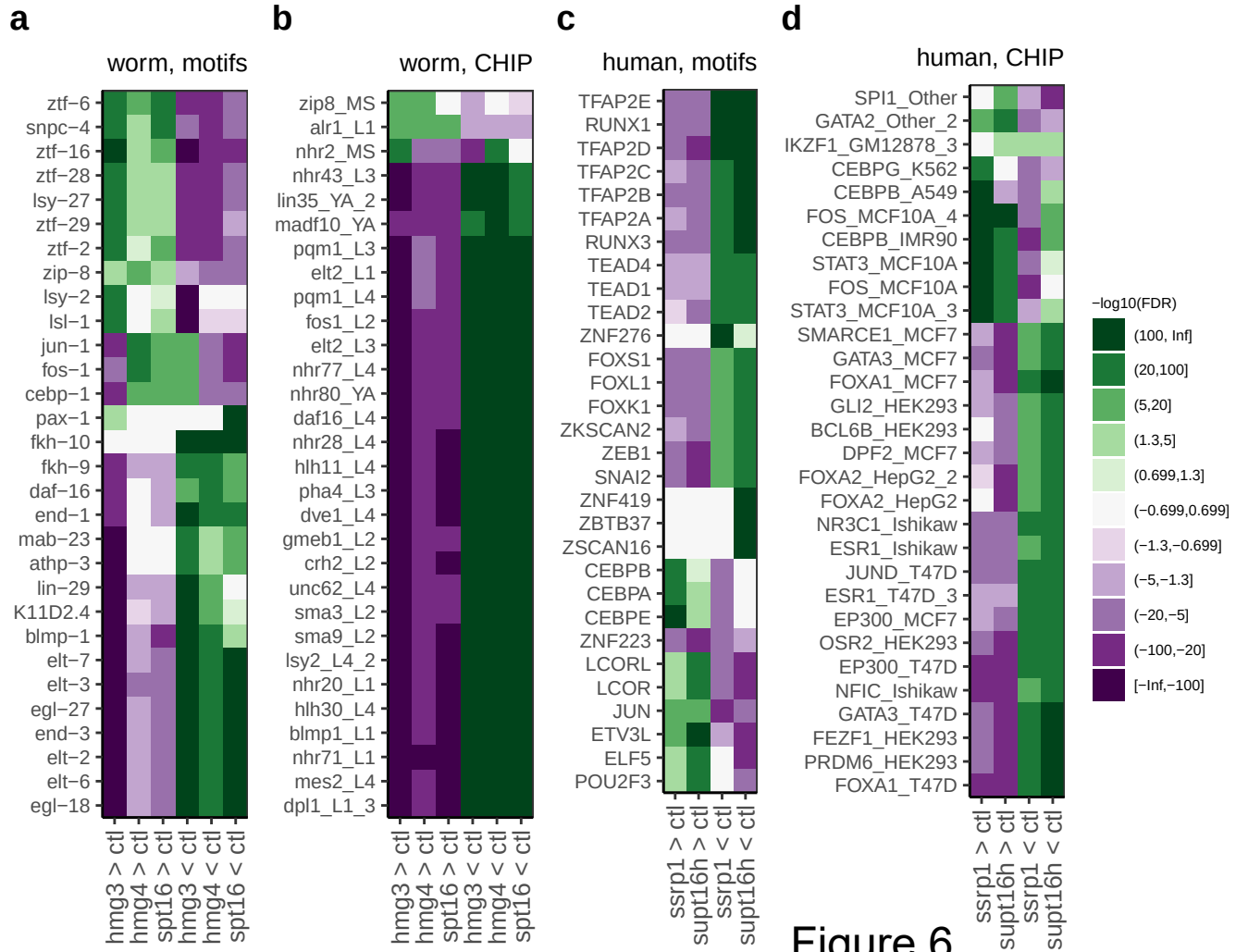

a
worm, motifs
ztf−6
snpc−4
ztf−16
ztf−28
lsy−27
ztf−29
ztf−2
zip−8
lsy−2
lsl−1
jun−1
fos−1
cebp−1
pax−1
fkh−10
fkh−9
daf−16
end−1
mab−23
athp−3
lin−29
K11D2.4
blmp−1
elt−7
elt−3
egl−27
end−3
elt−2
elt−6
egl−18
hmg3 > ctl
hmg4 > ctl
spt16 > ctl
hmg3 < ctl
hmg4 < ctl
spt16 < ctl
b
worm, CHIP
zip8_MS
alr1_L1
nhr2_MS
nhr43_L3
lin35_YA_2
madf10_YA
pqm1_L3
elt2_L1
pqm1_L4
fos1_L2
elt2_L3
nhr77_L4
nhr80_YA
daf16_L4
nhr28_L4
hlh11_L4
pha4_L3
dve1_L4
gmeb1_L2
crh2_L2
unc62_L4
sma3_L2
sma9_L2
lsy2_L4_2
nhr20_L1
hlh30_L4
blmp1_L1
nhr71_L1
mes2_L4
dpl1_L1_3
hmg3 > ctl
hmg4 > ctl
spt16 > ctl
hmg3 < ctl
hmg4 < ctl
spt16 < ctl
c
human, motifs
TFAP2E
RUNX1
TFAP2D
TFAP2C
TFAP2B
TFAP2A
RUNX3
TEAD4
TEAD1
TEAD2
ZNF276
FOXS1
FOXL1
FOXK1
ZKSCAN2
ZEB1
SNAI2
ZNF419
ZBTB37
ZSCAN16
CEBPB
CEBPA
CEBPE
ZNF223
LCORL
LCOR
JUN
ETV3L
ELF5
POU2F3
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
d
human, CHIP
SPI1_Other
GATA2_Other_2
IKZF1_GM12878_3
CEBPG_K562
CEBPB_A549
FOS_MCF10A_4
CEBPB_IMR90
STAT3_MCF10A
FOS_MCF10A
STAT3_MCF10A_3
SMARCE1_MCF7
GATA3_MCF7
FOXA1_MCF7
GLI2_HEK293
BCL6B_HEK293
DPF2_MCF7
FOXA2_HepG2_2
FOXA2_HepG2
NR3C1_Ishikaw
ESR1_Ishikaw
JUND_T47D
ESR1_T47D_3
EP300_MCF7
OSR2_HEK293
EP300_T47D
NFIC_Ishikaw
GATA3_T47D
FEZF1_HEK293
PRDM6_HEK293
FOXA1_T47D
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
−log10(FDR)
(100, Inf]
(20,100]
(5,20]
(1.3,5]
(0.699,1.3]
(−0.699,0.699]
(−1.3,−0.699]
(−5,−1.3]
(−20,−5]
(−100,−20]
[−Inf,−100]

Figure 6

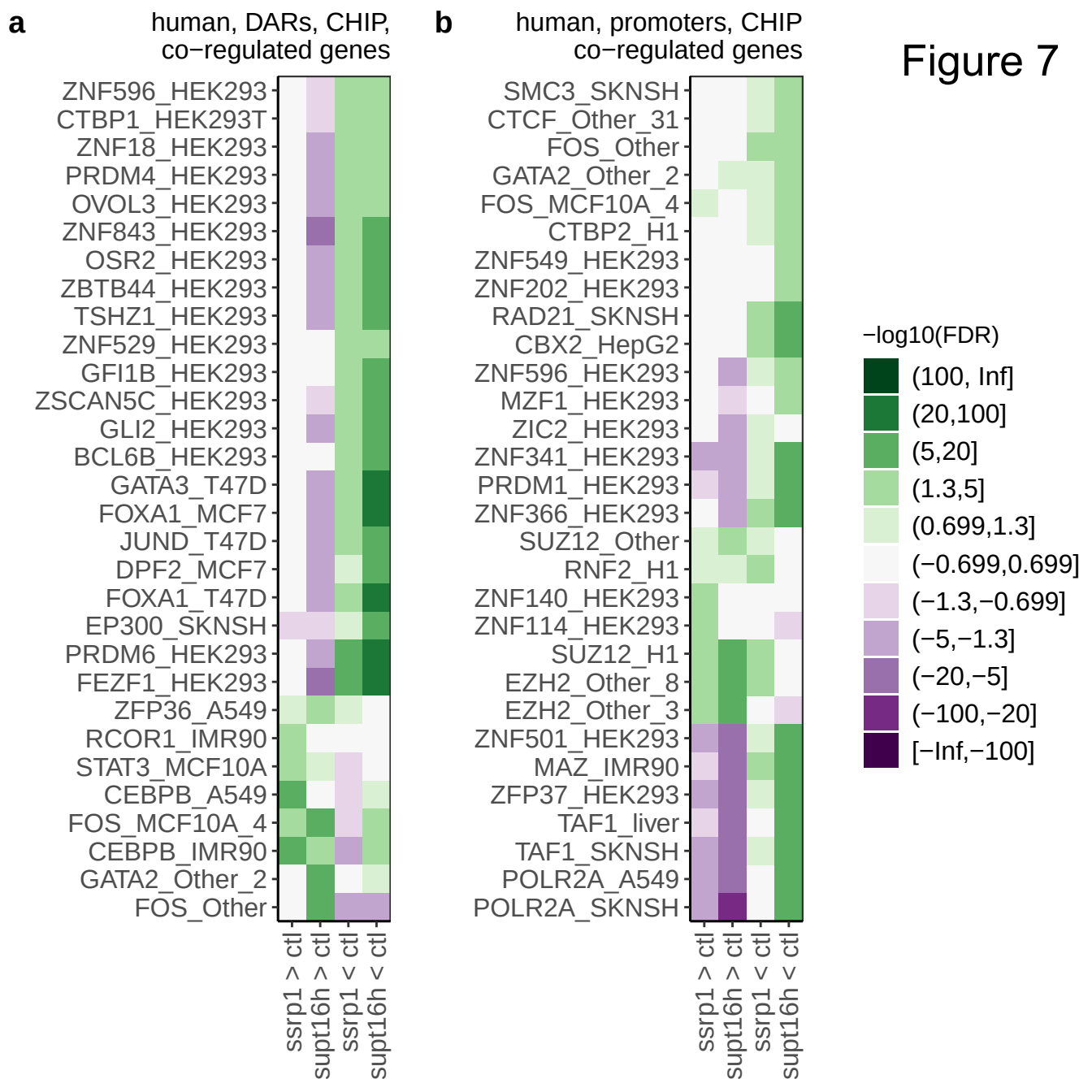


Figure 7
a
human, DARs, CHIP,
co−regulated genes
ZNF596_HEK293
CTBP1_HEK293T
ZNF18_HEK293
PRDM4_HEK293
OVOL3_HEK293
ZNF843_HEK293
OSR2_HEK293
ZBTB44_HEK293
TSHZ1_HEK293
ZNF529_HEK293
GFI1B_HEK293
ZSCAN5C_HEK293
GLI2_HEK293
BCL6B_HEK293
GATA3_T47D
FOXA1_MCF7
JUND_T47D
DPF2_MCF7
FOXA1_T47D
EP300_SKNSH
PRDM6_HEK293
FEZF1_HEK293
ZFP36_A549
RCOR1_IMR90
STAT3_MCF10A
CEBPB_A549
FOS_MCF10A_4
CEBPB_IMR90
GATA2_Other_2
FOS_Other
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
b
human, promoters, CHIP
co−regulated genes
SMC3_SKNSH
CTCF_Other_31
FOS_Other
GATA2_Other_2
FOS_MCF10A_4
CTBP2_H1
ZNF549_HEK293
ZNF202_HEK293
RAD21_SKNSH
CBX2_HepG2
ZNF596_HEK293
MZF1_HEK293
ZIC2_HEK293
ZNF341_HEK293
PRDM1_HEK293
ZNF366_HEK293
SUZ12_Other
RNF2_H1
ZNF140_HEK293
ZNF114_HEK293
SUZ12_H1
EZH2_Other_8
EZH2_Other_3
ZNF501_HEK293
MAZ_IMR90
ZFP37_HEK293
TAF1_liver
TAF1_SKNSH
POLR2A_A549
POLR2A_SKNSH
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
−log10(FDR)
(100, Inf]
(20,100]
(5,20]
(1.3,5]
(0.699,1.3]
(−0.699,0.699]
(−1.3,−0.699]
(−5,−1.3]
(−20,−5]
(−100,−20]
[−Inf,−100]

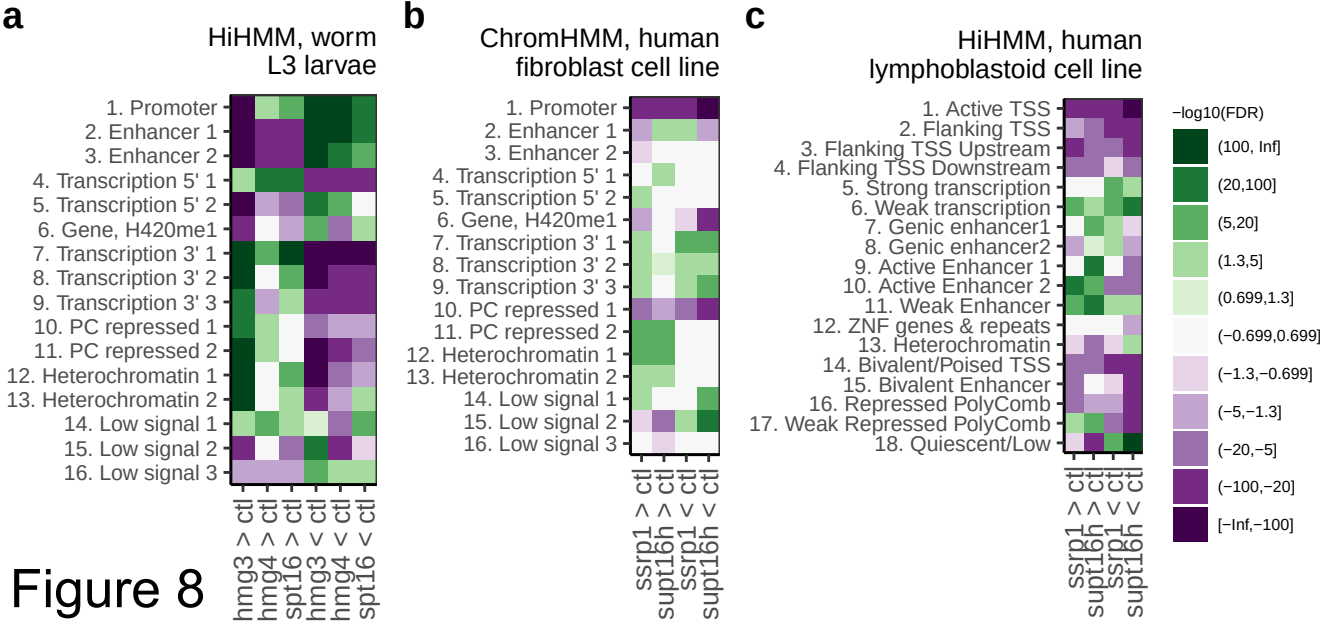

a
HiHMM, worm
L3 larvae
1. Promoter
2. Enhancer 1
3. Enhancer 2
4. Transcription 5' 1
5. Transcription 5' 2
6. Gene, H420me1
7. Transcription 3' 1
8. Transcription 3' 2
9. Transcription 3' 3
10. PC repressed 1
11. PC repressed 2
12. Heterochromatin 1
13. Heterochromatin 2
14. Low signal 1
15. Low signal 2
16. Low signal 3
hmg3 > ctl
hmg4 > ctl
spt16 > ctl
hmg3 < ctl
hmg4 < ctl
spt16 < ctl
b
ChromHMM, human
fibroblast cell line
1. Promoter
2. Enhancer 1
3. Enhancer 2
4. Transcription 5' 1
5. Transcription 5' 2
6. Gene, H420me1
7. Transcription 3' 1
8. Transcription 3' 2
9. Transcription 3' 3
10. PC repressed 1
11. PC repressed 2
12. Heterochromatin 1
13. Heterochromatin 2
14. Low signal 1
15. Low signal 2
16. Low signal 3
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
c
HiHMM, human
lymphoblastoid cell line
1. Active TSS
2. Flanking TSS
3. Flanking TSS Upstream
4. Flanking TSS Downstream
5. Strong transcription
6. Weak transcription
7. Genic enhancer1
8. Genic enhancer2
9. Active Enhancer 1
10. Active Enhancer 2
11. Weak Enhancer
12. ZNF genes & repeats
13. Heterochromatin
14. Bivalent/Poised TSS
15. Bivalent Enhancer
16. Repressed PolyComb
17. Weak Repressed PolyComb
18. Quiescent/Low
ssrp1 > ctl
supt16h > ctl
ssrp1 < ctl
supt16h < ctl
−log10(FDR)
(100, Inf]
(20,100]
(5,20]
(1.3,5]
(0.699,1.3]
(−0.699,0.699]
(−1.3,−0.699]
(−5,−1.3]
(−20,−5]
(−100,−20]
[−Inf,−100]

Figure 8