Adetunji and Abraham *BMC Bioinformatics* 2022, **23**(1):77
https://doi.org/10.1186/s12859-022-04588-z

BMC Bioinformatics

SOFTWARE

Open Access

SEAseq: a portable and cloud-based chromatin occupancy analysis suite

Modupeore O. Adetunji and Brian J. Abraham*

*Correspondence:
brian.abraham@stjude.org
Department
of Computational Biology,
St. Jude Children's Research
Hospital, Memphis, TN 38105,
USA

Abstract

Background: Genome-wide protein-DNA binding is popularly assessed using specific antibody pulldown in Chromatin Immunoprecipitation Sequencing (ChIP-Seq) or Cleavage Under Targets and Release Using Nuclease (CUT&RUN) sequencing experiments. These technologies generate high-throughput sequencing data that necessitate the use of multiple sophisticated, computationally intensive genomic tools to make discoveries, but these genomic tools often have a high barrier to use because of computational resource constraints.

Results: We present a comprehensive, infrastructure-independent, computational pipeline called SEAseq, which leverages field-standard, open-source tools for processing and analyzing ChIP-Seq/CUT&RUN data. SEAseq performs extensive analyses from the raw output of the experiment, including alignment, peak calling, motif analysis, promoters and metagene coverage profiling, peak annotation distribution, clustered/stitched peaks (e.g. super-enhancer) identification, and multiple relevant quality assessment metrics, as well as automatic interfacing with data in GEO/SRA. SEAseq enables rapid and cost-effective resource for analysis of both new and publicly available datasets as demonstrated in our comparative case studies.

Conclusions: The easy-to-use and versatile design of SEAseq makes it a reliable and efficient resource for ensuring high quality analysis. Its cloud implementation enables a broad suite of analyses in environments with constrained computational resources. SEAseq is platform-independent and is aimed to be usable by everyone with or without programming skills. It is available on the cloud at https://platform.stjude.cloud/workflows/seaseq and can be locally installed from the repository at https://github.com/stjude/seaseq.

Keywords: ChIP sequencing, CUT&RUN, Peak calling, Motif analysis, Cloud, Data analysis, Analysis pipeline, Computational genomics, Platform independent, GEO, SRA

Background

Understanding where proteins localize on chromatin is an important component in many study designs. Selectively purifying the chromatin fragments with antibodies and sequencing the resulting material affords detection of these protein-DNA interactions. Chromatin occupancy data from ChIP-Seq or CUT&RUN-Seq experiments can support a range of biological conclusions using different analytical approaches, but each analysis

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

usually requires intense computation and bioinformatics skills to execute. Basic processing of a chromatin binding dataset consists of several steps, including sequence read alignment to the reference genome, peak calling, annotation and visualization of peaks, coverage profiling, motif analysis, and most importantly quality control and assessment at each step. In general, the sequencing reads from a successful experiment are mapped to a reference genome, where read coverage roughly correlates with occupancy, and read-enriched regions are statistically determined. These regions or "peaks" can then be further mined to understand their functions and characteristics [1–3].

Multiple pieces of software exist for each analysis step, and attempts have been made to link these tools together in cohesive pipelines [4–10], as has been done for other analysis types [11–15]. However, these pipelines; summarized in Fig. 1, are limited by the types of analyses they can perform, often requiring substantial in-house computational infrastructure or expertise, which typically restricts their use to select research settings. Stemming from the vast number of existing tools and workflows, it remains challenging for new or seasoned researchers to decipher their usefulness and confidently determine relevant analyses.

To address these problems, we introduce a comprehensive and easy-to-use computational pipeline, "Single-End Antibody sequencing" (SEAseq), which takes advantage of enterprise-level workflow management tools to facilitate synonymous deployment across different computing infrastructures. SEAseq is a fully automated open-source pipeline that performs all the major analysis needed to process chromatin binding datasets ensuring high quality and useful results. It can be applied to data from any organism, and can automatically access publicly available data from the NCBI Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) repositories at the user's request [16, 17]. The pipeline is designed to be platform-independent and scalable: it can be executed on a personal computer or in high-performance computing (HPC) environments. Additionally, to enable broader utility in resource-poor environments, SEAseq can also be accessed in a reliable and reasonably priced parallel cloud computing environment, available at https://platform.stjude.cloud/workflows/seaseq.

Pipeline	Bioinformatics Software Installation Requirement	Pipeline Dependencies	Cloud Availability	Publication Year	Supported Peak Type(s)	Organism	Automated Download from GEO/SRA	Quality Assessment	Metagenes and Peak Annotation	Motif Analysis
Cistrome[8]	None	None (Web service)	No	2011	Narrow only	Restricted to provided organisms	No	None	Average signal profiles and gene annotation at a set of genomic locations	No
HiChIP[7]	Yes	Perl, R, Python	No	2014	Narrow & Broad	Restricted to provided organisms	No	- Sequencing library - Alignment performance	Average signal profiles and gene annotation at a set of genomic locations	Motif discovery
piPipes[6]	Yes	Perl, R, Python	No	2015	Narrow only	Restricted to provided organisms	No	None	Distribution of peaks at a set of genomic locations	No
ChiLin[9]	Yes (for Python version)	Python or Docker	No	2016	Narrow only	Restricted to provided organisms	No	- Sequencing library - Alignment performance - Enrichment profiles	Distribution of peak summits at a set of genomic locations	Motif discovery
pyflow-ChIPseq[10]	Yes	Snakemake (Python), R	No	2017	Narrow only	Human & Mouse	Yes; A separate R script	- Sequencing library - Alignment performance	No	No
Cut&Run Tools[5]	Yes	Python	No	2019	Narrow & Broad	Restricted to provided organisms	No	- Alignment performance - Enrichment profiles	No	Motif discovery
ENCODE ChIPseq	Yes	Cromwell (Java)	No	N/A	Narrow only	Supports all organisms	No	- Sequencing library - Enrichment profiles	No	No
SEAseq	**None**	**Docker, Cromwell (Java)**	**Yes**	**2021**	**Narrow & Broad**	**Supports all organisms**	**Yes**	**- Sequencing library - Alignment performance - Enrichment profiles Results are collated in an easy to interpret HTML**	**Average signal profiles and signal heatmaps at multiple sets of genomic locations. Per-gene/-peak annotation tables. Code to customize output visualizations.**	**Motif discovery & Motif enrichment in peaks and peak summits.**

Fig. 1 Features of the available pipelines for ChIP-seq or Cut&Run sequencing analysis

Implementation

SEAseq architecture

We were motivated to broaden the user base capable of performing chromatin sequencing analysis independent of computing infrastructure and expertise. To maximize the flexibility and portability of our pipeline, we adopted a workflow management system, Workflow Description Language (WDL) [18], and containerized the requisite tools, programs and SEAseq custom scripts to using Docker [19] (see Additional file 1 for the list of Docker images built for SEAseq).

The pipeline itself is not in a container platform due to its complexity; rather each step was compartmentalized into individual WDL tasks and sub-workflows to allow for independent execution of each task in an isolated environment with the use of Docker containers. Using WDL enables reorganization of individual tasks without having to redesign the entire pipeline. WDL also ensures efficient utilization and scaling of computing resources in a replicable and repeatable manner. Using Docker containers maximizes control and management of version configurations, software requirements and dependencies for improved portability and reproducibility. In addition, the choice to use WDL and Docker facilitates a standardized but customizable deployment across computing platforms, independent of the hardware infrastructure used to run SEAseq. The modular design of the pipeline allows experienced users to modify tools or steps in the pipeline, providing these modifications match the input/output schema formats used, by modifying the WDL code available on GitHub (https://github.com/stjude/seaseq).

SEAseq can be executed on a single computing node, such as a personal computer, or a parallel computing infrastructure including field-standard high-performance computing (HPC) clusters using compatible workflow execution engines such as Cromwell [20]. Complete usage instructions are outlined in the SEAseq documentation (https://github.com/stjude/seaseq/#readme). Three dependencies are required to run SEAseq: Java, Cromwell and an engine able to run Docker containers (such as Docker, Singularity [21]). Each of these packages can be installed with minimal user expertise.

SEAseq functionality

SEAseq performs the several fundamental ChIP-Seq or CUT&RUN analyses in a single execution. Figure 2 shows a schematic overview of the analyses performed by SEAseq, which are briefly described in the following paragraphs (see Additional file 2 for the expanded list of SEAseq pipeline steps and parameters).

FASTQ sequencing reads are stringently aligned to the reference genome provided using Bowtie [22]. The mapped reads are then further processed by removal of redundant reads [23] using SAMtools [24], and removal of reads in problematic regions or regions with significant background noise or artificially high signal [25] using BEDTools [26]. In addition, SEAseq characterizes the global binding preferences of the antibody using read density profiling in relevant genomic regions such as promoters and gene bodies using our custom version of BAMToGFF (https://github.com/stjude/BAM2GFF).

To compensate for the background noise intrinsic to chromatin binding assays, many analyses rely on identification of highly covered regions or peaks [3]. It is important to choose the appropriate peak-calling algorithm based on the type of protein targeted

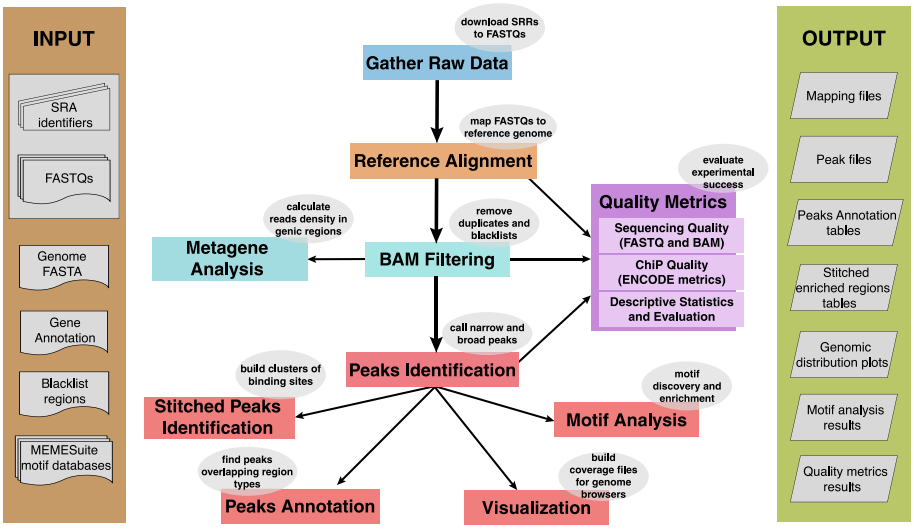

Fig. 2 High-level SEAseq schematic. The flowchart shows the input files (left), a top-down overview of the analysis steps executed by SEAseq (center), and the outputs (right). The input consists of the specified files needed to utilize SEAseq; these files include FASTQ files (in compressed gzip [.gz] format) and/or SRA identifiers (SRR), Genome FASTA [.fa], Gene Annotation [.gtf], and optionally UHS/ DER/ DAC blacklist regions [.bed] and one or more of the MEME suite position weight matrix databases. The output consists of the analysis results files generated from SEAseq: mapping files [.bam], peaks [.bed] and peak coverage files [.wig;. tdf;.bw], per-promoter and metagene average coverage and heatmaps plots, peak annotation distribution tables, motif discovery and enrichment results, and quality metrics results [.html]

[27]. After read alignment and filtering steps are completed, SEAseq identifies enriched regions for two binding profiles: MACS [28] for factors that bind shorter regions, e.g. many sequence-specific transcription factors, and SICER [29] for broad regions of enrichment, e.g. some histone modifications. The choice of these peak callers for the identification of narrow peaks and broad peaks was based on extensive review of published benchmarking evaluations and our own previous work [30–34]. Normalized and unnormalized coverage files are generated for visualization on multiple genome browsers such as the UCSC genome browser [35] and IGV [36]. The pipeline also identifies stitched clusters of enriched regions and separates exceptionally signal-rich regions, e.g. super-enhancers, from typical enhancers using ROSE [37, 38]. SEAseq also performs motif discovery and enrichment analysis to characterize overrepresented sequences using tools from the MEME Suite [39], and performs genic annotation of the various peaks and quantifies regions of abundance in promoters, gene bodies, gene-centric windows, and proximal genes using BEDTools and custom Python scripts.

SEAseq quality metrics and dashboard

Though other pipelines perform coverage-based analyses, most generally lack a comprehensive means of assessing the overall quality of the experiment. SEAseq uniquely calculates an extensive set of quality metrics for detecting experimental issues, including ChIP-Seq metrics recommended by the ENCODE consortium [40]. The SEAseq quality metrics include the percentage of reads mapped, nonredundant fraction (NRF), fraction of reads in peaks (FRiP), strand correlation scores (NSC, RSC), library complexity (PCR bottleneck) and many other important metrics used to infer

Table 1 SEAseq quality metrics performed and their definitions

Quality metric	Definition
Aligned percent	Percentage of mapped reads
Base quality	Per-base sequence quality distribution
Estimated fragment width	Average fragment size of the peak distribution
Estimated tag length	Sequencing read length
Fraction of reads in peaks (FRiP)	The fraction of reads within coverage-enriched regions
Linear stitched peaks (enhancers)	Total number of clustered enriched regions
Non-redundant fraction (NRF)	Fraction of uniquely mapped sequencing reads
Normalized peaks[a]	Peaks identified after input/control correction
Normalized strand-correlation coefficient (NSC)	The ratio of the maximum cross-correlation value divided by the background cross-correction
Sequence diversity	Sequence overrepresentation; if reads/sequences are overrepresented in the library
PCR bottleneck coefficient (PBC)	It is a measure of library complexity determined by the fraction of genomic locations with exactly one unique read versus those covered by at least one unique read
Peaks	Total number of enriched regions (peaks)
Raw reads	Total number of sequencing reads
Read length[b]	Average FASTQ read length
Relative strand-correlation coefficient (RSC)	A strand cross-correlation ratio between the fragment-length cross-correlation and the read-length peak
SE-like enriched regions (super enhancers)	Total number of SE-like clustered enriched regions
Overall quality	Average score rank of all metrics calculated

[a] Applicable when input/control is provided

[b] Applicable if multiple FASTQs are inputted

Rank Score	Color Name	Definition
POOR	RED	Bad or poor performance of the metric or overall sample, consider investigating the datasets.
BELOW-AVERAGE	ORANGE	Below-average performance of given metric or overall sample, recommend investigating further if unexpected.
AVERAGE	YELLOW	The metric or Sample displayed average performance.
GOOD	GREEN YELLOW	Above average or acceptable performance of stated metric or overall sample.
EXCELLECT	GREEN	Outstanding or should be expected performance for given metric or overall sample dataset.

Fig. 3 The Rank Score scheme for the SEAseq quality metrics. The rank score is flagged in a color scale for easy interpretation of metrics performance

quality as listed in Table 1. To facilitate integration and easy interpretation of these metrics, we devised a five-scale color-rank flag system to visually inspect the performance of each metric, and provide a cross-metric averaged rank score to easily intuit the performance of the overall analysis (Fig. 3). The rank flag system is estimated based on recommended thresholds from the ENCODE consortium, literature review and criteria provided in Table 2. These quality metrics can be broadly grouped into two subsets: quality assessment on the sequencing library and assessment on the enrichment profiles observed, providing users an easy way to intuitively explore the performance or quality of their data. The results are exported in a tab-delimited file

Table 2 Description of each quality metric color-rank scheme

Quality metrics	Rank				
	Excellent	Good	Average	Below-average	Poor
Aligned percent (*A*)	$A \geq 80\%$	$A \geq 70\%$	$A \geq 60\%$	$A \geq 50\%$	$A < 50\%$
Base quality (*B*)	*B*="pass"	-	*B*="warn"	-	*B*="fail"
Estimated fragment width	–	–	–	–	–
Estimated tag length (E)[a]	$E < \pm 10$	–	–	–	$E > \pm 10$
FRiP (*F*)[c]	$F \geq 0.05$	$F \geq 0.02$	$F \geq 0.01$	$F \geq 0.0075$	$F < 0.0075$
Linear stitched peaks (*L*)	$L \geq 10000$	$L \geq 5000$	$L \geq 2000$	$L \geq 1000$	$L < 1000$
NRF	$NRF \geq 0.8$	$NRF \geq 0.7$	$NRF \geq 0.6$	$NRF \geq 0.5$	$NRF < 0.5$
Normalized Peaks (N)	$N \geq 10000$	$N \geq 5000$	$N \geq 2000$	$N \geq 1000$	$N < 1000$
NSC[c]	$NSC \geq 1.045$	–	–	–	$NSC < 1.045$
Sequence diversity (*D*)	*D*="pass"	–	*D*="warn"	–	*D*="fail"
PBC (*C*)[c]	$C \geq 0.9$	$C \geq 0.75$	$C \geq 0.66$	$C \geq 0.5$	$C < 0.5$
Peaks (*P*)	$P \geq 10000$	$P \geq 5000$	$P \geq 2000$	$P \geq 1000$	$P < 1000$
Raw reads (*R*)	$R \geq 30M$	$R \geq 25M$	$R \geq 20M$	$R \geq 15M$	$R < 15M$
Read length	–	–	–	–	–
RSC[c]	$RSC \geq 1$	$RSC \geq 0.75$	–	–	$RSC < 0.75$
Super stitched peaks (*S*)[b]	$S \geq 0.2$	$S \geq 0.1$	$S \geq 0.05$	$S \geq 0.02$	$S < 0.02$
Overall quality (Q)	$Q \geq 2$	$Q \geq 1$	$Q \geq 0$	$Q \geq -1$	$Q < -1$

M millionreads

[a] *E* is difference between predicted tag length and average read length

[b] *S* is the ratio of superstitched regions divided by the linear stitched regions identified

[c] Extrapolated based on ENCODE recommended thresholds

and color flagged in html format. Additional file 3 shows a typical quality statistics report produced by SEAseq.

SEAseq on the cloud

To facilitate the broadest usage of our pipeline, and to empower researchers with little to no computational skills or resources, we offer a cloud-based version of SEAseq, called SEAseq Cloud, that is hosted on the St. Jude Cloud Genomics Platform [41]. The St. Jude Cloud Genomics Platform leverages Microsoft Azure and DNAnexus (https://www.dnanexus.com) to provide a secure and privacy compliant framework for analysis, storage and distribution of genomic data [41]. The DNAnexus platform provides an easy-to-navigate graphical user interface for exploration and analysis of user data, which can be quickly and securely uploaded, downloaded, and shared with collaborators at a reasonably low cost. For more advanced operations, DNAnexus also provides a command-line client. SEAseq cloud is available from https://platform.stjude.cloud/workflows/seaseq.

The user needs a DNAnexus account to use SEAseq Cloud. SEAseq requires that the user data, such as the FASTQs and genome files, be uploaded to a project folder. Navigating the SEAseq Cloud web interface after logging in, as shown in Fig. 4, involves selecting the "*Start*" button for first-time users, then the "*Launch This Tool*" button (Fig. 4a). The user will then be navigated to the SEAseq Cloud *Run Analysis* Page to start an analysis (Fig. 4b). DNAnexus requires the user input data files, such as the FASTQ files and required genome files (if applicable), be uploaded to a Project before proceeding. Uploading data files can be done on a Projects page by

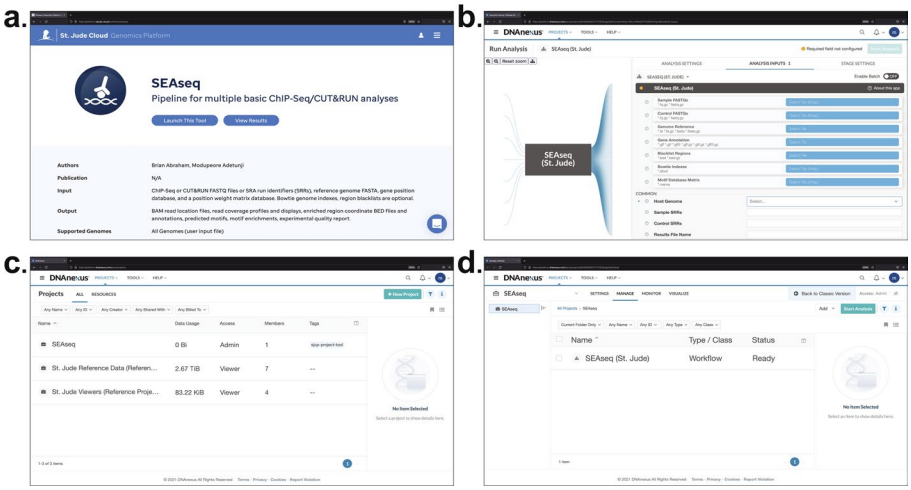

Fig. 4 SEAseq Cloud. **a** Screenshot of SEAseq workflow Landing Page. **b** SEAseq cloud Run Analysis "Analysis Inputs" Page. **c** DNAnexus Projects Page containing SEAseq Project. **d** SEAseq cloud Project page consisting of the SEAseq workflow and where user data files can be uploaded

selecting the *Projects* tab then the *All Projects* option to be directed to the Projects/Home page (Fig. 4c). The DNAnexus Projects page (also accessible from https://platform.dnanexus.com) will contain the vended SEAseq workflow in its self-titled project and St. Jude Reference Project files (Fig. 4c). The St. Jude Reference Project files contains some genome files acceptable by SEAseq for the user's convenience. To begin analysis, select on the SEAseq Project to access the SEAseq workflow (Fig. 4d) and/or to upload user data files by clicking the "*Add*" button. It is recommended to use the New (also called the "Pannexin") User Interface (UI) version, instead of the Classic/Legacy ("Membrane") version. Further descriptions will be based on the New UI as shown in Fig. 4b, d.

To start an analysis using SEAseq, select the "*SEAseq (St. Jude)*" workflow, and the user will be navigated to the *Run Analysis* page (Fig. 4b), where one can specify their preferred "*Execution Name*" and "*Execution Output Folder*" in the "*Analysis Settings*" input fields. The previously uploaded user data will then be available for selection in the requested "*Analysis Inputs*" input fields. The required fields are the sequencing FASTQs or SRA accession numbers (SRRs), "*Genome Reference*" and "*Gene Annotation*" files. SEAseq also provides a list of genome files, "*Host Genomes*", to select from (Fig. 4b). Optionally, one or more SRRs can be inputted into the "*Sample SRRs*" field for sample SRRs and/or the "*Control SRRs*" field for Input/Control SRRs. After the analysis inputs and settings has been specified to the user's preference, the SEAseq analysis can be started by selecting "*Start Analysis*". Once started, the status of the job can be monitored from the project folder under the "*Monitor*" tab (Fig. 4d) or SEAseq landing page by selecting the "*View Results*" button (Fig. 4a). When the analysis is completed, a notification message will be sent to the user's registered email and the analysis results files can be viewed and downloaded from the initially specified "*Execution Output Folder*". More information on SEAseq Inputs and Output directories and files can be found in Additional file 4.

Results and discussion

Replication of two ChIP-Seq studies using SEAseq

We re-analyzed ChIP-Seq data from two previously published studies to showcase the utility of SEAseq in performing relevant analyses in a single execution.

The first study probed functions of LIN28B in neuroblastoma and discovered it binds chromatin via interaction with the transcription factor ZNF143 (GEO accession: GSE138742) [42]. The dataset included LIN28B, ZINF143, and doxycycline-inducible engineered LIN28B ChIP-Seq data of BE2C cells. The analysis was executed using the SEAseq Cloud workflow. All required hg19 genome files and position weight matrices were first uploaded into a DNAnexus project. Each ChIP antibody SRR along with the corresponding Input DNA SRR were used as input, and the relevant genome files were selected to their required fields in SEAseq (see Additional file 5 for the accession numbers and genome files used). Analysis using SEAseq demonstrated the expected satisfactory overall scores for these samples' read quality and quality of peaks (> 1) (Fig. 5a). Motif analysis identified the major binding consensus motif in both LIN28B and ZNF143 as reported in the original publication (Fig. 5b). In addition, given the ability of SEAseq to perform a large array of advanced downstream analysis, we also observed the expected genome-wide occupancy results, demonstrating preferential

a. **ZNF143**

DATA	Overall Quality	Raw Reads	Base Quality	Sequence Diversity	Aligned Percent	Estimated Fragment Width	Estimated Tag Length	NRF	PBC	NSC	RSC	FRiP	Total Peaks	Normalized Peaks*	Linear Stitched Peaks	SE-like Enriched Regions
SAMPLE	GOOD	35796441	pass	pass	81.890	110	75	0.8575	0.8714	1.0541	0.9913	0.0278	10764	10764	1752	97
CONTROL		46608378	pass	pass	77.016	200	75	0.9425	0.9713	1.0132	0.7821	0.0015	605			

LIN28B

DATA	Overall Quality	Raw Reads	Base Quality	Sequence Diversity	Aligned Percent	Estimated Fragment Width	Estimated Tag Length	NRF	PBC	NSC	RSC	FRiP	Total Peaks	Normalized Peaks*	Linear Stitched Peaks	SE-like Enriched Regions
SAMPLE	GOOD	56053120	pass	pass	79.118	121	75	0.8183	0.8378	1.0266	0.8804	0.0218	11248	11248	540	52
CONTROL		46608378	pass	pass	77.016	200	75	0.9425	0.9713	1.0132	0.7821	0.0015	605			

yDox-LIN28Bwt

DATA	Overall Quality	Raw Reads	Base Quality	Sequence Diversity	Aligned Percent	Estimated Fragment Width	Estimated Tag Length	NRF	PBC	NSC	RSC	FRiP	Total Peaks	Normalized Peaks*	Linear Stitched Peaks	SE-like Enriched Regions
SAMPLE	GOOD	48240553	pass	pass	79.046	125	75	0.9039	0.9238	1.0359	0.8891	0.0525	19039	19039	1025	122
CONTROL		36795254	pass	pass	80.437	200	75	0.9190	0.9662	1.0187	0.9162	0.0072	2798			

p53

DATA	Overall Quality	Raw Reads	Base Quality	Sequence Diversity	Aligned Percent	Estimated Fragment Width	Estimated Tag Length	NRF	PBC	NSC	RSC	FRiP	Total Peaks	Normalized Peaks	Linear Stitched Peaks	SE-like Enriched Regions
SAMPLE	AVERAGE	7679779	pass	pass	65.959	161	36	0.7076	0.6982	1.2747	1.1565	0.0097	600	600	2131	64
CONTROL		12110109	pass	pass	74.473	35	36	0.8758	0.8812	1.0318	0.5044	0.0077	434			

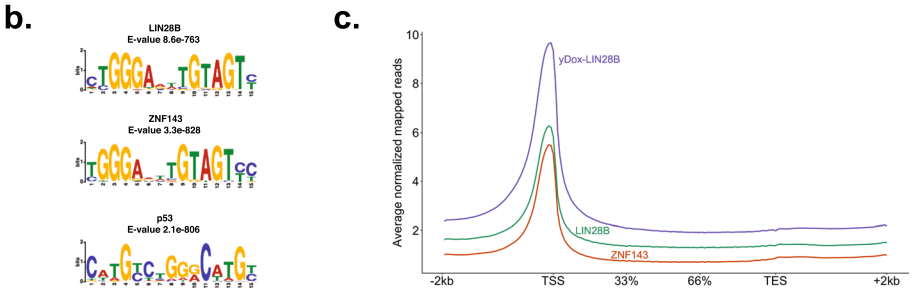

Fig. 5 SEAseq case study. **a** Screenshot of the SEAseq Quality Evaluation Report for all studies. **b** Sequence logo depicting the top predicted motif identified for each study. **c** Genome-wide occupancy analysis of LIN28B and ZNF143 reveal enriched binding at active promoters

Table 3 SEAseq Cloud performance results

ChIP	Number of reads in sample	Number of reads in input	Runtime (HH:MM)	Cost
LIN28B	54,894,988	45,603,300	32:07	$26.38
ZNF143	35,108,844	45,603,300	25:49	$22.35
yDox-LIN28B	47,203,903	35,682,203	28:20	$23.68
p53	7,473,100	11,760,480	10:58	$8.33

binding at promoter regions by LIN28B, ZNF143, and significantly with the doxycycline-induced LIN28B (Fig. 5e). The entire analysis of all three datasets was computed using the SEAseq Cloud platform, which successfully completed in an average of 28.45 wall clock hours costing an average of $24.14. Table 3 summarizes the total time and cost of analysis performed using SEAseq Cloud.

The second study profiled binding of the tumor suppressor p53 in normal human cells (GEO accession: GSE31558) [43]. To maximize comparability with the published results, we uploaded available hg18 reference genome data files, and the deposited study files, which include ChIP-Seq for p53 in IMR90 fibroblasts (see Additional file 5). SEAseq analysis showed an "AVERAGE" overall quality score; indicative from the sequencing depth, FRiP and the corresponding poor number of peaks identified (Fig. 5a). From this, we show that the analysis may benefit from higher sequencing coverage and depth. Nevertheless, we were able to obtain comparable results to the original publication, including the significantly enriched p53 binding motif (Fig. 5b). This analysis was successfully completed in 11 h costing $8.33 (Table 3).

These recapitulated results demonstrate the utility of SEAseq in analysis of antibody purification data.

SEAseq performance evaluation

To assess SEAseq performance and resource consumption, the aforementioned datasets using SEAseq Cloud were also analyzed on our IBM Spectrum LSF (v10.1.0.9) HPC cluster with Cromwell (v52) and Singularity (v3.8.0). Per-task analysis revealed the most time-consuming component of the pipeline is the metagene analysis step (completed in under 16 h). Most other steps were completed in under 1 h, with the exception of the read alignment and optional genome index construction steps, which each completed in 4–6 h. Figure 6 shows the time and memory consumption of the different steps. In addition, steps displaying the highest memory consumption were the metagene analysis, broad peaks identification, alignment, and genome index construction. Given the genome index construction step is computationally intensive, we recommend that users avoid this optional step by providing the bowtie genome indexes when analyzing multiple datasets. Overall, the different steps show a maximum memory usage of under 7 GB. This performance showcases the efficiency of SEAseq in appropriating tasks in a timely and memory efficient manner, and thus makes the pipeline deployable on most computing environments. Furthermore, the analysis also reveals the size of the FASTQs has a proportional effect on time and memory consumption.

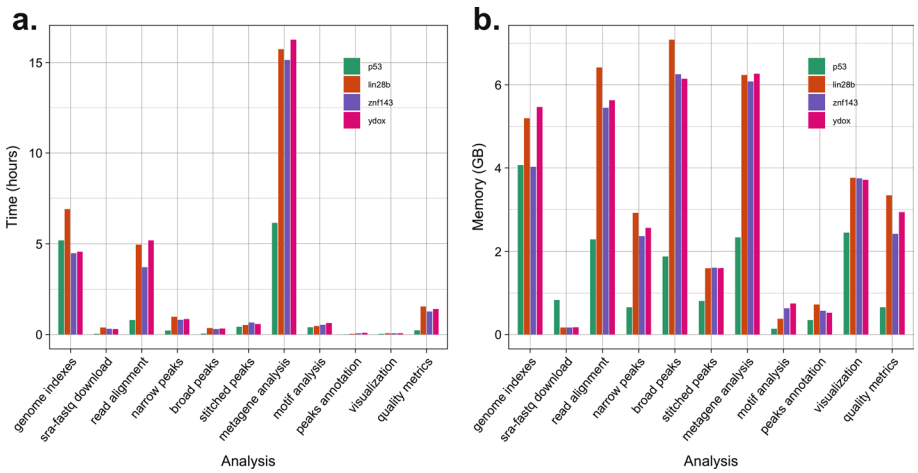

Fig. 6 SEAseq performance results. **a** Time elapsed in hours by analysis step. **b** Memory consumption in Gigabytes (GB) by analysis step

Conclusion

To our knowledge, SEAseq is the first portable, all-inclusive analysis pipeline for ChIP-Seq and CUT&RUN data. SEAseq is easy to use and provides a modular architecture for quick integration and customization for efficient data analysis on multiple computing infrastructures. Having proved highly comparable results to published datasets, we believe SEAseq fulfills a critical requirement as an efficient and reliable one-stop computational pipeline for high quality analysis results.

Availability and requirements

Project name: SEAseq
Project home page: https://github.com/stjude/seaseq
Project cloud page: https://platform.stjude.cloud/workflows/seaseq
Operating system(s): Platform independent
Programming language: WDL, Docker
Other requirements: Java, Cromwell (when using the GitHub version)
License: Apache License 2.0
Any restrictions to use by non-academics: None

Abbreviations

GEO: Gene Expression Omnibus; SRA: Sequence Read Archive; WDL: Workflow Description Language; UI: User Interface; HPC: High-Performance Computing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04588-z.

Additional file 1. SEAseq Docker images. List of Docker images built for SEAseq.

Additional file 2. SEAseq pipeline steps and parameters. Detailed description of SEAseq pipeline steps and parameters.

Additional file 3. Case study LIN28B. The complete HTML quality statistics report for LIN28B ChIP-seq analysis.

Additional file 4. SEAseq inputs and outputs. Input files used for SEAseq analysis and detailed descriptions of the Output directories and files generated using SEAseq.

Additional file 5. Case study datasets. Case Study Datasets and Genome Files information.

Acknowledgements
We would like to thank Brian Curran, Andrew Thrasher, Andrew Frantz, and the broader St. Jude Cloud team for their generous support in making the workflow accessible on the St. Jude cloud platform and DNAnexus. We wish to thank all the users who provided critical feedback.

Authors' contributions
M.O.A. designed and developed SEAseq and drafted the manuscript. B.J.A. supervised the work, gave feedback, and revised the manuscript. Both authors performed extensive testing of the pipeline and approved the final manuscript.

Funding
This work has been supported by funding from the American Lebanese Syrian Associated Charities (ALSAC) and the St. Jude Children's Research Hospital Collaborative Research Consortium on Chromatin Regulation in Pediatric Cancer. The funding body had no role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

Availability of data and materials
The developed software is freely available on GitHub athttps://github.com/stjude/seaseq. All datasets and genome files analyzed during this study are publicly available and information on how they were obtained are included in this published article (Additional file 5).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
M.O.A. has no competing interests. B.J.A. is a shareholder in Syros Pharmaceuticals.

Received: 9 November 2021 Accepted: 28 January 2022
Published online: 23 February 2022

References

1. Nakato R, Sakata T. Methods for ChIP-seq analysis: a practical workflow and advanced applications. Methods. 2021;187:44–53. https://doi.org/10.1016/j.ymeth.2020.03.005.
2. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform. 2016;18:bbw023. https://doi.org/10.1093/bib/bbw023.
3. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10:669–80. https://doi.org/10.1038/nrg2641.
4. Orlova NN, Bogatova OV, Orlov AV. High-performance method for identification of super enhancers from ChIP-Seq data with configurable cloud virtual machines. MethodsX. 2020. https://doi.org/10.1016/j.mex.2020.101165.
5. Zhu Q, Liu N, Orkin SH, Yuan G-C. CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. Genome Biol. 2019;20:192. https://doi.org/10.1186/s13059-019-1802-4.
6. Han BW, Wang W, Zamore PD, Weng Z. piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics. 2015;31:593–5. https://doi.org/10.1093/BIOINFORMATICS/BTU647.
7. Yan H, Evans J, Kalmbach M, Moore R, Middha S, Luban S, et al. HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. BMC Bioinform. 2014;15:280. https://doi.org/10.1186/1471-2105-15-280.
8. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, et al. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 2011;12:R83. https://doi.org/10.1186/gb-2011-12-8-r83.
9. Qin Q, Mei S, Wu Q, Sun H, Li L, Taing L, et al. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC Bioinform. 2016;17:404. https://doi.org/10.1186/s12859-016-1274-4.
10. Tang M. pyflow-ChIPseq: a snakemake based ChIP-seq pipeline. 2017. https://zenodo.org/record/819971.
11. Zhang X, Jonassen I. RASflow: an RNA-Seq analysis workflow with Snakemake. BMC Bioinform. 2020;21:1–9. https://doi.org/10.1186/S12859-020-3433-X/TABLES/2.
12. Garrido-Rodriguez M, Lopez-Lopez D, Ortuno FM, Peña-Chilet M, Muñoz E, Calzado MA, et al. A versatile workflow to integrate RNA-seq genomic and transcriptomic data into mechanistic models of signaling pathways. PLoS Comput Biol. 2021;17: e1008748. https://doi.org/10.1371/JOURNAL.PCBI.1008748.

13. D'Antonio M, De Meo PDO, Pallocca M, Picardi E, D'Erchia AM, Calogero RA, et al. RAP: RNA-Seq analysis pipeline, a new cloud-based NGS web application. BMC Genom. 2015;16:1–11. https://doi.org/10.1186/1471-2164-16-S6-S3/FIGURES/2.
14. Cameron CJF, Cameron CJF, Wang XQD, Dostie J, Blanchette M. LAMPS: an analysis pipeline for sequence-specific ligation-mediated amplification reads. BMC Res Notes. 2020;13:1–4. https://doi.org/10.1186/S13104-020-05106-1/FIGURES/1.
15. Banerjee S, Bhandary P, Woodhouse M, Sen TZ, Wise RP, Andorf CM. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. BMC Bioinform. 2021;22:205. https://doi.org/10.1186/s12859-021-04120-9.
16. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(Database):D19-21. https://doi.org/10.1093/nar/gkq1019.
17. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10. https://doi.org/10.1093/NAR/30.1.207.
18. OpenWDL. https://openwdl.org/.
19. Docker. https://www.docker.com/.
20. Cromwell. https://cromwell.readthedocs.io/en/stable/.
21. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS ONE. 2017;12:e0177459. https://doi.org/10.1371/JOURNAL.PONE.0177459.
22. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25. https://doi.org/10.1186/gb-2009-10-3-r25.
23. Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, et al. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinform. 2015;16:1–11. https://doi.org/10.1186/1471-2105-16-S13-S10.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
25. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9:9354. https://doi.org/10.1038/s41598-019-45839-z.
26. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
27. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013. https://doi.org/10.1371/journal.pcbi.1003326.
28. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008. https://doi.org/10.1186/gb-2008-9-9-r137.
29. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009;25:1952–8. https://doi.org/10.1093/bioinformatics/btp340.
30. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform. 2016;17:953–66. https://doi.org/10.1093/BIB/BBV110.
31. Starmer J, Magnuson T. Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. BMC Bioinform. 2016;17:1–10. https://doi.org/10.1186/S12859-016-0991-Z/FIGURES/4.
32. Laczik M, Hendrickx J, Veillard AC, Tammoh M, Marzi S, Poncelet D. Iterative fragmentation improves the detection of ChIP-seq peaks for inactive histone marks. Bioinform Biol Insights. 2016;10:209. https://doi.org/10.4137/BBI.S40628.
33. Jeon H, Lee H, Kang B, Jang I, Roh TY. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. Genom Inform. 2020;18:1–9. https://doi.org/10.5808/GI.2020.18.4.E42.
34. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-Seq peak detection. PLoS ONE. 2010. https://doi.org/10.1371/JOURNAL.PONE.0011471.
35. Kuhn RM, Haussler D, James KW. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144–61. https://doi.org/10.1093/bib/bbs038.
36. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92. https://doi.org/10.1093/bib/bbs017.
37. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153:320–34. https://doi.org/10.1016/j.cell.2013.03.036.
38. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19. https://doi.org/10.1016/j.cell.2013.03.035.
39. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39-49.
40. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22:1813–31. https://doi.org/10.1101/gr.136184.111.
41. McLeod C, Gout AM, Zhou X, Thrasher A, Rahbarinia D, Brady SW, et al. St. Jude cloud: a pediatric cancer genomic data-sharing ecosystem. Cancer Discov. 2021;11:1082–99. https://doi.org/10.1158/2159-8290.cd-20-1230.
42. Tao T, Shi H, Mariani L, Abraham BJ, Durbin AD, Zimmerman MW, et al. LIN28B regulates transcription and potentiates MYCN-induced neuroblastoma through binding to ZNF143 at target gene promotors. Proc Natl Acad Sci U S A. 2020;117:16516–26. https://doi.org/10.1073/pnas.1922692117.
43. Botcheva K, McCorkle SR, McCombie WR, Dunn JJ, Anderson CW. Distinct p53 genomic binding patterns in normal and cancer-derived human cells. Cell Cycle. 2011;10:4237–49. https://doi.org/10.4161/cc.10.24.18383.

Publisher's Note

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