Check for updates

SOFTWARE TOOL ARTICLE

REVISED SeqAcademy: an educational pipeline for RNA-Seq and ChIP-Seq analysis [version 3; peer review: 1 approved, 1 approved with reservations, 1 not approved]

Syed Hussain Ather [ID][1], Olaitan Igbagbo Awe [ID][2], Thomas J. Butler[3], Tamiru Denka[4], Stephen Andrew Semick [ID][5], Wanhu Tang[6], Ben Busby[4]

[1]National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, 20892, USA
[2]National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA
[3]National Institute on Aging , National Institutes of Health, Baltimore , MD, 21224, USA
[4]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, 20894, USA
[5]Lieber Institute for Brain Development , Baltimore , MD, 21205, USA
[6]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, 20892, USA

Abstract

Quantification of gene expression and characterization of gene transcript structures are central problems in molecular biology. RNA sequencing (RNA-Seq) and chromatin immunoprecipitation sequencing (ChIP-Seq) are important methods, but can be cumbersome and difficult for beginners to learn. To teach interested students and scientists how to analyze RNA-Seq and ChIP-Seq data, we present a start-to-finish tutorial for analyzing RNA-Seq and ChIP-Seq data: SeqAcademy (*source code:* https://github.com/NCBI-Hackathons/seqacademy, *webpage:* http://www.seqacademy.org/). This user-friendly pipeline, fully written in markdown language, emphasizes the use of publicly available RNA-Seq and ChIP-Seq data and strings together popular tools that bridge that gap between raw sequencing reads and biological insight. We demonstrate practical and conceptual considerations for various RNA-Seq and ChIP-Seq analysis steps with a biological use case - a previously published yeast experiment. This work complements existing sophisticated RNA-Seq and ChIP-Seq pipelines designed for advanced users by gently introducing the critical components of RNA-Seq and ChIP-Seq analysis to the novice bioinformatician. In conclusion, this well-documented pipeline will introduce state-of-the-art RNA-Seq and ChIP-Seq analysis tools to beginning bioinformaticians and help facilitate the analysis of the burgeoning amounts of public RNA-Seq and ChIP-Seq data.

Open Peer Review

Reviewer Status ✗ ✓ ?

	Invited Reviewers		
	1	**2**	**3**
version 4 (revision) 22 Sep 2020			
version 3 (revision) 09 May 2019			? report
version 2 (revision) 30 Nov 2018	✗ report ⬆		
version 1 22 May 2018	✗ report	✓ report	

1. **Philip Ewels** [ID], Department of Biochemistry and Biophysics, Stockholm, Sweden

2. **Norann A. Zaghloul**, University of Maryland School of Medicine, Baltimore, USA

This article is included in the International Society for Computational Biology Community Journal gateway.

3. **Xi Chen** , Simons Foundation, New York, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Syed Hussain Ather (shussainather@gmail.com)

> **REVISED** **Amendments from Version 2**
>
> The steps of the pipeline should work. I've removed the Jupyter notebook completely for simplification so that the tutorials may attract a more general audience.
>
> **See referee reports**

Introduction

RNA sequencing (RNA-Seq) is a rapidly expanding technique used to answer broad questions in the life sciences, ranging from mitochondrial function (Mercer *et al.*, 2011) to the pathogenesis of breast cancer (Li *et al.*, 2017). Chromatin immunoprecipitation sequencing (ChIP-Seq) is a genome-wide technique for profiling histone modifications, protein interactions, and transcription factor binding sites (Barski *et al.*, 2007). Using this technique to analyze protein interactions involves very large data sets for computational analysis. The computational steps can identify the locations of features such as DNA-binding enzymes, modified histones, chaperones, nucleosomes, and transcription factors (TFs) (Bailey *et al.*, 2013).

The expanding importance of RNA-Seq and ChIP-Seq data is reflected by its explosive growth in terabytes in the primary public repository for storing this data - the Sequence Read Archive (SRA) (Wheeler *et al.*, 2008). This incredible increase in the amount of public data has not been met with an equal increase in the number of scientists who can skillfully and thoughtfully analyze this important resource. Given the fundamental role that RNA-Seq and ChIP-Seq data, among other next-generation sequencing data types, are likely to play in the coming decades, there is a critical need to teach RNA-Seq and ChIP-Seq analysis to life scientists with diverse interests and backgrounds.

The goal of analyzing RNA-Seq data is often to identify and characterize quantitative differences in gene expression between biological samples from two or more groups. For ChIP-Seq, the goal is to characterize DNA-protein interactions. Biological samples may originate from several different study designs including: different tissue types from the same individual (e.g. cancerous tissue vs. non-cancerous tissue), the same strain of cells under different environmental conditions, or the same tissue under a time-course experiment.

There are major barriers to the novice bioinformatician who is interested in learning how to analyze RNA-Seq and ChIP-Seq data. RNA-Seq and ChIP-Seq data are costly to generate (>$1,000/sample) and cumbersome to store; with data from a single sample often occupying several gigabytes of storage space. However, recent advances, including a greater impetus to deposit sequencing data in SRA ("Principles and Guidelines for Reporting Preclinical Research," 2015) and the innovative alignment of streamed sequencing data (Kim *et al.*, 2015), offer new opportunities to overcome these long-standing problems. The second barrier to entry is inherent to RNA-Seq and ChIP-Seq data. These datasets are large and complex: there are over 20,000 known genes in the human genome (Naidoo *et al.*, 2011) and the transcriptional diversity of the human genome is not yet fully characterized (Yamashita *et al.*, 2011).

Furthermore, RNA-Seq data is susceptible to "batch effects" and other confounders that can occlude real biological effects or, worse, mislead the un-skeptical researcher. Thus, appropriate analysis of these data requires advanced algorithms and sophisticated statistical methods, coupled with traditional scientific skepticism, to uncover biological insight buried in the data.

These difficulties dissuade many from attempting RNA-Seq and ChIP-Seq analysis, particularly those lacking previous data analysis experience, but the genomics community needs more scientists who can adeptly analyze RNA-Seq and ChIP-Seq data. Moreover, shared understanding of RNA-Seq and ChIP-Seq analysis will produce higher quality discourse between the biologists who are responsible for conducting RNA-Seq and ChIP-Seq experiments and the bioinformaticians who are experts at analyzing the resulting data produced from these experiments. Several well-developed pipelines currently exist for processing RNA-Seq and ChIP-seq data from start to finish (Djebali *et al.*, 2017; Park *et al.*, 2017; Torres-García *et al.*, 2014; Yalamanchili *et al.*, 2017); however, these pipelines are generally designed for advanced bioinformaticians who often have existing practical experience in analyzing high-throughput data. A pipeline designed to teach those with little experience how to analyze high-throughput sequencing data is therefore needed. Thus, we developed a proof-of-concept, well-documented "tutorial pipeline" over the course of a three-day NCBI-sponsored hackathon intended to teach RNA-seq and ChIP-seq analysis to beginners. This tutorial pipeline, "SeqAcademy," incorporates state-of-the-art RNA-Seq and ChIP-seq analysis tools into a simple, easy to use workflow tutorial and we demonstrate its use with publicly available data.

Methods

Implementation

SeqAcademy uses self-contained tutorials, which runs Python, R, and Bash scripts among others, all from the document itself. It requires about 16 GB of memory storage. The tutorial files facilitate open science and reproducible code by mixing code chunks with notes and markup. This format, known as "literate programming," is particularly amenable to teaching bioinformatics because it allows learners to follow along in the document while running each code step directly within the notebook.

Operation

The tutorial begins with an explanation of how to install necessary dependencies and select interesting data from the BioProjects browser. Alignment while streaming the data is done with HISAT2 version 0.1.6 and subsequent quality control with MultiQC version 1.5. The tutorial then splits into two separate protocols: one for RNA-seq, the other for ChIP-seq analysis.

The workflow involved setup, alignment, quality control, analysis, and visualization steps for publicly available RNA-Seq and ChIP-seq data sets. There are many appropriate tools available for each step of RNA-seq and ChIP-seq analysis. Our goal is to present an easy to use and understandable pipeline

rather than an exhaustive list of analysis tools. For each step below, we will explain the role of the bioinformatic tool, as well as our rationale for including it in this tutorial pipeline (Figure 1). Here, we present an overview of the steps; further details for each subsection can be found on the project's Github page.

Setup

The setup step uses the Bioconda channel (Grüning *et al*., 2017) for the conda package manager to install all of the programmatic dependencies for the entire pipeline. The data sets were selected by searching NCBI BioProject web browser (Barrett *et al*., 2011). For our use case, we searched for publically available RNA-Seq and ChIP-Seq datasets that were relatively small and thus could be easily downloaded and processed, and would be relatively straightforward to interpret biologically. We therefore selected RNA-Seq and ChIP-Seq data from yeast (*Saccharomyces cerevisiae)* samples (Mulla *et al*., 2017; Rawal *et al*., 2018).

The RNA-Seq data demonstrates the differences in genetic expression between aneuploid and euploid yeast (Mulla *et al*., 2017). The ChIP-Seq data demonstrates the effects of 3-Amino-1,2,4-triazole (3-AT) on chromatin accessibility (Rawal *et al*., 2018). We downloaded the reference sequence for *Saccharomyces cerevisiae* from Ensembl version 84 (RNA-Seq SRA study number: SRP106028 ChIP-Seq SRA study number: SRP132584). We note that the *SraRunTables* file can be adjusted to specific user data, different from the RNA-Seq or ChIP-seq data sets used in this project. Thus, this lightweight, portable educational pipeline can be adapted to meet the usage needs and interests of a broad base of bioinformatics beginners and teachers.

Alignment

HISAT2 is a software program used for the alignment of raw sequence data, consisting of FASTQ files (Kim *et al*., 2015. We chose to use HISAT2 because it allows users to stream raw sequence data rather than downloading it to the local machine, reducing disk space and time requirements for users of the SeqAcademy educational tool - an exemplary use of "edge-computing" in bioinformatics. One disadvantage of this approach is that it requires a stable internet connection, as the aligned raw sequence files are downloaded as SAM (sequence alignment mapping) files along with the log files. Nevertheless, by choosing to use HISAT2 for alignment, we reduced required disk space and broadened the potential user base of this pipeline.

RNA-Seq

After alignment and quality control, users convert the SAM files to BAM files with the samtools package version 1.8 (Li *et al*., 2009). Then, gene expression is quantified with HTSeq version 0.9.1 (Anders *et al*., 2015). Afterwards, we demonstrate how to extract biological significance from these various analyses, by showing students how to visualize gene expression patterns and undertake exploratory data analysis with principal component analysis (PCA). Finally, we show how to undertake differential expression analysis using DESeq2 version 1.21.0 (Love *et al*., 2014) and how to visualize these differences with volcano plots and experiment-specific visualizations in the
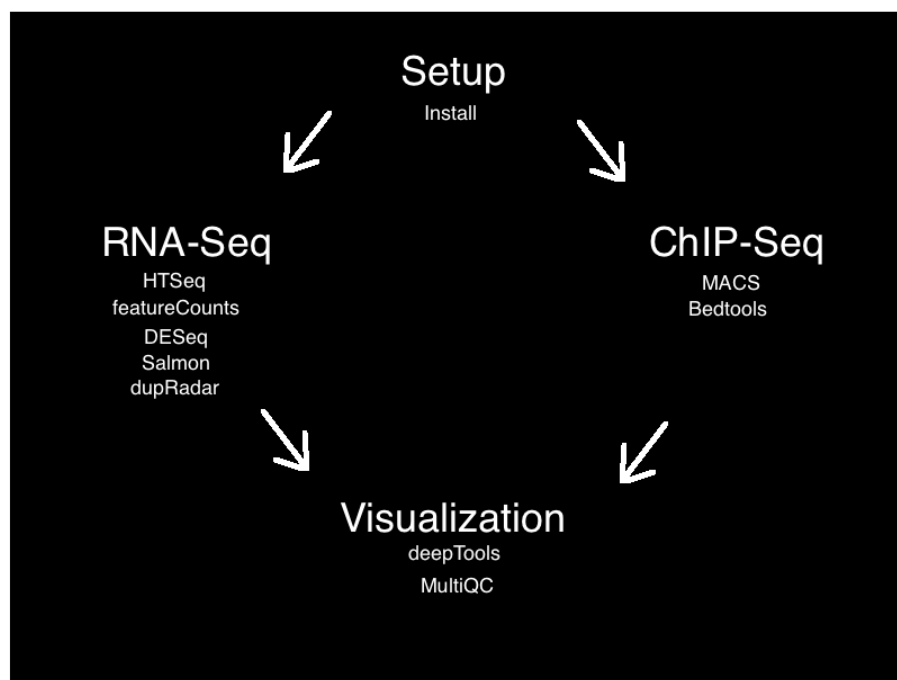
Figure 1. Flowchart of the SeqAcademy tutorial.

R package ggplot2 version 2.2.1 (Wickham, 2009). Thus, students can learn how to quantify gene expression, answer biologically relevant questions through differential gene expression analysis, and visualize gene expression patterns.

ChIP-Seq

After alignment, we perform peak-calling to determine protein-binding locations in the ChIP-seq data. The peak-calling step of ChIP-Seq involves finding differentially binding sites between the two ChIP-Seq signals (input and immunoprecipitate). Numerous peak callers exist to distinguish biologically relevant signal peak from technical noise for the Chip-Seq experiments. Here, we used the peak-calling algorithm MACS (Model-based Analysis for ChIP-Seq) version 1.4.2 (Zhang *et al.*, 2008). MACS is a commonly used peak-caller and has been shown to have more accurate results than competing peak-callers (Hocking *et al.*, 2017). After calling peaks, the results are sorted and analyzed for intersections using bedtools version 2.27.0, a set of tools for analyzing genomic data (Quinlan & Hall, 2010). Lastly, bedtools output is visualized with Integrative Genomics Viewer (IGV) version 2.4, a genomic data set viewer that allows for visualization of genomic features (Robinson *et al.*, 2011).

Quality control

To generate a quality control report about the success of the alignment, we used MultiQC (Ewels *et al.*, 2016). MultiQC reports the number of reads mapped to one unique location, reads mapped to multiple unique locations, and reads not mapped to any location in the reference genome. MultiQC can provide reports for both RNA-Seq and ChIP-seq data. Reads mapped to one unique location have a higher confidence level of being correctly mapped, as reads mapped to multiple unique locations cannot be localized to the reference with a high degree of probability. While MultiQC is not strictly necessary for this pipeline--the plots and statistics it produces are based off of the HISAT2 alignment summary files - we chose to include it to introduce users to a useful tool that is built for quality control.

Use cases
Target audience

This educational pipeline is designed for students without previous programming experience who are looking for an introduction to the acquisition, processing, analysis, and visualization of either RNA-Seq or ChIP-seq data. Students of next-generation sequencing analysis may range the academic spectrum, from undergraduates to professors, all of whom share an interest in learning to analyze sequencing data. SeqAcademy also

offers a useful introduction to the core steps of RNA/ChIP-Seq analysis for use by bioinformatics educators who are teaching a class or mentoring students. Motivated individual learners, for instance a graduate student who is attempting RNA-Seq analysis, may also benefit by working through SeqAcademy. The tutorial completely self-contained, so users do not need to manage additional input files or tools beyond what is provided directly in the notebook document—every line of code to be run has already been written and tested. Thus, this flexible tutorial may be a suitable introduction to RNA-seq and ChIP-seq analysis for workshops, graduate school classes, or motivated individual learners. We also hope that fellow bioinformatics educators will build off of SeqAcademy to teach intermediate and advanced bioinformatics concepts and skills. The pipeline is simple and modular, so it can easily be adapted to analyze different datasets and customized to meet different user needs.

Learning objectives

The learning objectives of SeqAcademy are two-fold. The first and most immediate or practical objective is for a student to learn how to conduct the core steps of an RNA/ChIP-seq analysis, beginning with a search for publicly available sequencing data and ending with biologically meaningful results. The second objective is to foster a greater understanding of the concepts behind each step. This includes biological reasons behind why certain experiments such as ChIP-Seq and RNA-Seq are run, and the logic behind alignment, differential gene expression, and peak-calling. The tutorial pipeline is purposefully simple, as this will introduce important component of next generation sequencing more gently, and will encourage students to build off of it to create more advanced pipelines that will meet the unique goals of the student.

Table 1 and Table 2 illustrate the sample input yeast data for RNA-Seq and ChIP-Seq, respectively. The RNA-Seq data examines aneuploidy while the ChIP-Seq data shows induction by 3-Amino-1,2,4-triazole (3-AT). Results of the principal component analysis, an unsupervised data reduction technique, of the RNA-Seq data are shown in Figure 2a. The slight clustering of the data into two different groups, euploid and aneuploid can be observed. A volcano plot is used to visualize significant differentially expressed genes between two groups, in this case euploid and aneuploid (Figure 2b). Figure 2c displays the enrichment of chromosome X for differentially expressed genes, consistent with the aneuploid sample having an extra X chromosome. Figure 3 shows an IGV screenshot of how peaks of protein-enrichment are distributed across the yeast genome.

Table 1. Example RNA-Seq input. This data presents the RNA-Seq data used in this tutorial. This tutorial observes RNA-Seq data of aneuploidy in yeast.

BioSample	Experiment	MBases	MBytes	Run	SRA_Study
SSAMN06859 211	SRX2775581	1632	575	SRR5494627	SRP106028
SAMN06859 210	SRX2775582	940	331	SRR5494628	SRP106028
SAMN06859 209	SRX2775583	1195	421	SRR5494629	SRP106028
SAMN06859 208	SRX2775584	815	288	SRR5494630	SRP106028
SAMN06859 207	SRX2775585	946	333	SRR5494631	SRP106028
SAMN06859 206	SRX2775586	1152	407	SRR5494632	SRP106028

Table 2. Example ChIP-Seq input. This data presents the ChIP-Seq data used in this tutorial. This tutorial observes ChIP-Seq data of induction by 3-AT in yeast.

BioSample	Experiment	MBases	MBytes	Run	SRA_Study
SAMN08513506	SRX3677830	8816	3690	SRR6703656	SRP132584
SAMN08513513	SRX3677835	9614	4022	SRR6703661	SRP132584
SAMN08513512	SRX3677836	6049	2749	SRR6703662	SRP132584
SAMN08513511	SRX3677837	6918	3140	SRR6703663	SRP132584
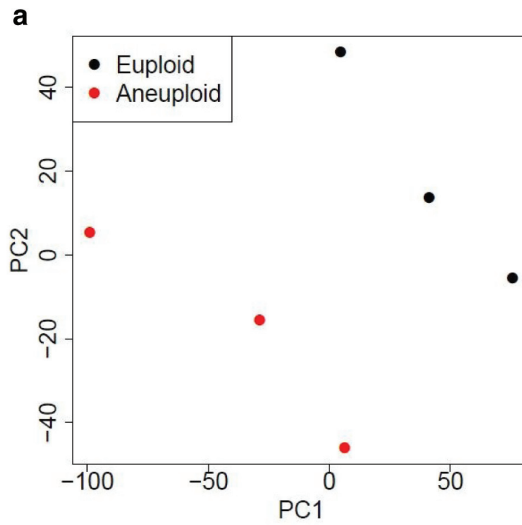
Figure 2a. Principal component analysis (PCA) of yeast. PCA suggests gene expression for euploid yeast samples (haploid) clusters distinctly from that of the aneuploid yeast samples (diploid chromosome X). The first two Principal Components account for ~70% of the variance in expressed genes). Data provided by Mulla *et al.*, 2017.
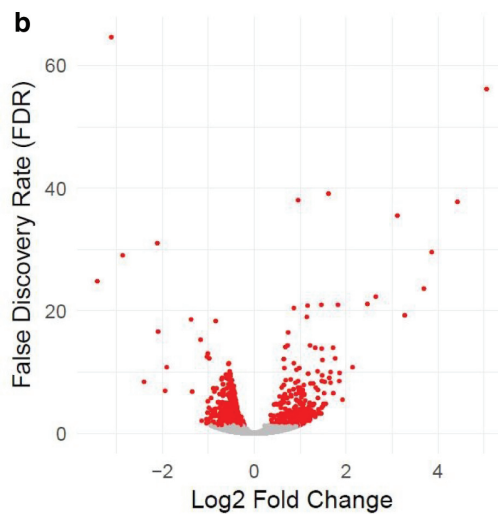
Figure 2b. Volcano plot of differentially expressed genes between euploid yeast colonies versus aneuploid yeast colonies. The x-axis represents the difference in gene expression between the conditions. False discovery rate (FDR), a method for controlling for multiple testing, is along the y-axis. Each point represents a tested gene (N=3,926). Red points are those reaching genome-wide significance (at FDR<0.05, N=663), whereas grey points are genes not reaching statistical significance (FDR>0.05, N=3,263). Data provided by Mulla *et al.*, 2017.
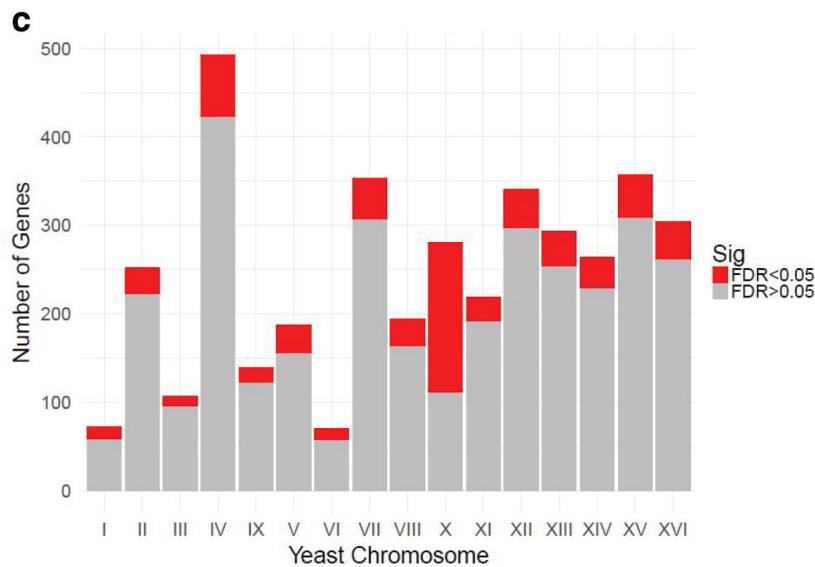
Figure 2c. Relative enrichment of chromosome X for differentially expressed genes. The relative enrichment of chrX for differentially expressed genes suggests the downstream results of this processing pipeline are consistent with biological expectations. The RNA-seq experiment was performed on yeast colonies with an extra chromosome X. Data provided by Mulla *et al.*, 2017.
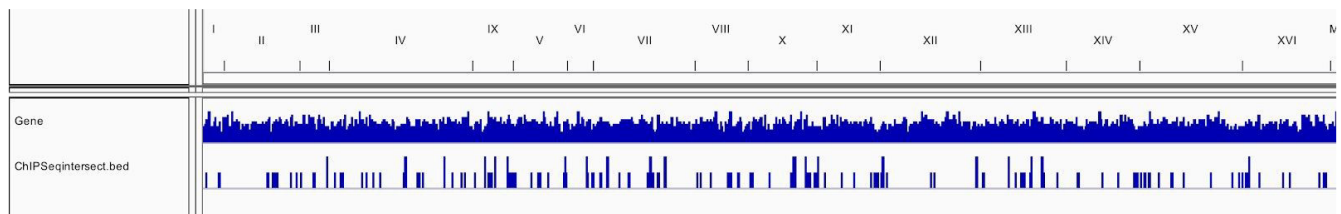
Figure 3. Distribution of intersected peaks across the yeast genome. This IGV screenshot shows in the bottom row the intersected peaks between the two treatment conditions of the yeast samples. The matching genes with each intersected peak can be analyzed. Data provided by Rawal *et al.*, 2018.

The corresponding genes can be examined to determine proteins involved in 3-AT induction.

Conclusion and next steps
Limitations and future directions
There are several limitations to take into account with this tutorial and future directions for further work. In this tutorial, we focused on using RNA-seq on "bulk" or homogenate tissue samples, as opposed to single-cell RNA-seq, which has distinct analytical considerations. Our pipeline is currently limited to only two of the various next generation sequencing analyses, and we would like to broaden the scope to also include DNA sequencing and other epigenetic sequencing protocols, such as whole-genome bisulfite sequencing. Our platform can also be developed further to incorporate more advanced features, such user interfaces for performing bioinformatics analyses from the web browser, login systems for users to keep track of their own progress, and forums and messaging systems for community feedback. We would also like to translate the pipeline into other languages to broaden its scope. In subsequent improvements, we plan to make the pipeline easily individualized for a user's own data sources by adjusting SraRunTables. Future hackathons may offer a useful setting to further improve this developing resource. Despite these limitations, SeqAcademy provides a solid starting foundation for beginners to learn the fundamentals.

Summary
We have presented a novel, standalone educational tool for two types of next generation sequencing data: RNA-Seq and ChIP-Seq data. This project offers a simple guidebook to an introductory analysis pipeline used in RNA-Seq and ChIP-Seq data. We introduced a cutting-edge bioinformatics tools frequently used for the acquisition, alignment, processing, analysis, and visualization of large-scale sequencing data and referenced further resources for continued learning. SeqAcademy meets the need for an educational analysis pipeline which can be used to teach undergraduate and graduate students with limited bioinformatics

experience how to analyze publically available sequencing data.

Data availability

Use case data is available for the NCBI Sequence Read Archive Run Selector under accession numbers – SRP132584 and SRP106028

Software availability

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.2662541 (Ather *et al.*, 2018)

The code for this project is deposited under an MIT License on GitHub: https://github.com/NCBI-Hackathons/seqacademy

References

Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics.* 2015; **31**(2): 166–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Ather SH, Awe OI, Butler TJ, *et al.*: **SeqAcademy: an educational pipeline for RNA-Seq and ChIP-Seq analysis.** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.2662541

Bailey T, Krajewski P, Ladunga I, *et al.*: **Practical guidelines for the comprehensive analysis of ChIP-seq data.** *PLoS Comput Biol.* 2013; **9**(11): e1003326.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Barrett T, Clark K, Gevorgyan R, *et al.*: **BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata.** *Nucleic Acids Res.* 2011; **40**(Database issue): D57–D63.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Barski A, Cuddapah S, Cui K, *et al.*: **High-resolution profiling of histone methylations in the human genome.** *Cell.* 2007; **129**(4): 823–37.
PubMed Abstract | **Publisher Full Text**

Djebali S, Wucher V, Foissac S, *et al.*: **Bioinformatics Pipeline for Transcriptome Sequencing Analysis.** *Methods Mol Biol.* 2017; **1468**: 201–219.
PubMed Abstract | **Publisher Full Text**

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: A sustainable and comprehensive software distribution for the life sciences.** *bioRxiv.* 2017.
Publisher Full Text

Hocking TD, Goerner-Potvin P, Morin A, *et al.*: **Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning.** *Bioinformatics.* 2017; **33**(4): 491–499.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods.* 2015; **12**(4): 357–60.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Li Y, Liu X, Tang H, *et al.*: **RNA Sequencing Uncovers Molecular Mechanisms Underlying Pathological Complete Response to Chemotherapy in Patients with Operable Breast Cancer.** *Med Sci Monit.* 2017; **23**: 4321–4327.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 550.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mercer TR, Neph S, Dinger ME, *et al.*: **The human mitochondrial transcriptome.** *Cell.* 2011; **146**(4): 645–658.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mulla WA, Seidel CW, Zhu J, *et al.*: **Aneuploidy as a cause of impaired chromatin silencing and mating-type specification in budding yeast.** *eLife.* 2017; **6**: pii: e27991.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Naidoo N, Pawitan Y, Soong R, *et al.*: **Human genetics and genomics a decade after the release of the draft sequence of the human genome.** *Hum Genomics.* 2011; **5**(6): 577–622.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Park SJ, Kim JH, Yoon BH, *et al.*: **A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages.** *Genomics Inform.* 2017; **15**(1): 11–18.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Principles and Guidelines for Reporting Preclinical Research. Retrieved April 18, 2018, 2015.
Reference Source

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rawal Y, Chereji RV, Valabhoju V, *et al.*: **Gcn4 Binding in Coding Regions Can Activate Internal and Canonical 5' Promoters in Yeast.** *Mol Cell.* 2018; **70**(2): 297–311.e4.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–26.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Torres-García W, Zheng S, Sivachenko A, *et al.*: **PRADA: pipeline for RNA sequencing data analysis.** *Bioinformatics.* 2014; **30**(15): 2224–2226.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wheeler DL, Barrett T, Benson DA, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2008; **36**(Database issue): D13–D21.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** Springer-Verlag. 2009.
Publisher Full Text

Yalamanchili HK, Wan YW, Liu Z: **Data Analysis Pipeline for RNA-seq Experiments: From Differential Expression to Cryptic Splicing.** *Curr Protoc Bioinformatics.* 2017; **59**: 11.15.1–11.15.21.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Yamashita R, Sathira NP, Kanai A, *et al.*: **Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis.** *Genome Res.* 2011; **21**(5): 775–789.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Zhang Y, Liu T, Meyer CA, *et al.*: **Model-based Analysis of ChIP-Seq (MACS).** *Genome Biol.* 2008; **9**(9): R137.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✗ ✓ ?

Version 3

Reviewer Report 13 June 2019

https://doi.org/10.5256/f1000research.20598.r49787

? **Xi Chen** (ID)

Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY, USA

I appreciate the efforts that the authors have made to make the RNA/ChIP-seq data processing easy by developing this pipeline. I can foresee the educational value of this paper. However, after reading this paper, I have several comments, especially about explaining 'why' each processing step is needed to users.

It is not clear to me what's the meaning of protein interactions. Are you talking about protein-DNA interactions or protein-protein interactions? ChIP-seq is mainly used to identify protein-DNA interactions. If multiple ChIP-seq profiles are jointly analyzed, co-binding relationship between multiple proteins can be studied. However, this co-binding relationship is different from protein-protein interaction. The author should say "protein-DNA interactions" instead of "protein interactions" as the latter usually refer to protein-protein interactions.

After reading the abstract and introduction, it is not clear to me why RNA-seq and ChIP-seq data are jointly discussed in this paper. Most of the reference papers are just talking about one data type. Please clarify why you choose these two. I know people can integrate both data types to identify regulatory networks (doi: 10.1093/bioinformatics/btx827)[1]. Is this your motivation? But this is not mentioned in this paper. As this manuscript is prepared to guide beginners lacking enough experience in this filed, it is necessary to tell them why these two data types are jointly discussed here. Although integrating these two data types is not the main focus of this paper, users can preprocess the raw data using this pipeline and do further downstream analysis by using other tools. Definitely, current description of each data type is necessary to let readers know each data type provides unique information.

Quality Control should be placed right after alignment as it is a general approach to check read quality and alignment rate. It is very misleading to readers that they need to get peaks or gene expression and then do QC check on BAM files.

The following part of RNA-seq data analysis should be improved "Afterwards, we demonstrate how to extract biological significance from these various analyses, by showing students how to

visualize gene expression patterns and undertake exploratory data analysis with principal component analysis (PCA). Finally, we show how to undertake differential expression analysis using DESeq2 version 1.21.0 (Love et al., 2014) and how to visualize these differences with volcano plots and experiment-specific visualizations in the R package ggplot2 version 2.2.1 (Wickham, 2009)." Users may have their own data and it can be: (1) one sample, (2) multiple sample under one condition, (3) multiple sample under two conditions. It would be better to discuss processing steps in different scenarios. If only one RNA-seq sample is given, there is no need to do PCA or DESeq2. For (2), PCA can tell if there is batch effect. For (3), PCA can tell how different (at high level) those samples are between two conditions and further DESeq2 can tell which genes are different. In this paragraph, the authors say a lot of 'how', but to beginners, they need to know 'why' each step is needed.

The same issue to ChIP-seq data. Why do users need the intersection step using bedtools? If they get multiple replicates, the intersection step will return reliable peak loci. If data of multiple factors is obtained, they can use the interaction step to extract co-binding events. And gene annotation is usually needed after getting those peaks, which is missing in this pipeline.

References
1. Chen X, Gu J, Wang X, Jung J, et al.: CRNET: an efficient sampling approach to infer functional regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data. *Bioinformatics*. 2018; **34** (10): 1733-1740 Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: NGS data analysis, computational tool development, bulk or single cell data integration, epigenetics, genetics

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 2

Reviewer Report 11 March 2019

https://doi.org/10.5256/f1000research.18855.r41334

✗ **Philip Ewels** (iD)
Stockholm University, Department of Biochemistry and Biophysics, Stockholm, Sweden

The authors have made significant improvements to the tool in this revision, addressing a number of points raised in my initial review. However, sadly I was still unable to get the tutorial code to work myself. A number of the steps generated syntax and execution errors making progress impossible. The notebook has some statuses from the authors' previous execution steps with similar error messages, so I don't think that this is just my system (downloading the genome & building the index onwards).

There seems to be some active development work ongoing in the repository - both a `devel` workbook (better to use git branches instead of separate files), but also some duplicate code blocks in the main notebook. There also seems to be a discrepancy between the introduction text on the main website and on the repository.

In summary - I still think that the intention and design of the tool is excellent, and the manuscript text good. However, whilst the tutorial itself seems to be broken and not usable by others, I cannot approve this article. If these issues can be fixed so that it's possible for others to easily run through the tutorial, then I will be happy to approve.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Version 1

Reviewer Report 12 June 2018

https://doi.org/10.5256/f1000research.16196.r34336

✔ **Norann A. Zaghloul**
Department of Medicine, Division of Endocrinology, Diabetes and Nutrition, Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

This article presents a streamlined tool aimed toward novice bioinformaticians who wish to analyze RNA-Seq or ChIP-Seq data in one simple tool. The premise for this tool is excellent. Investigators who are new to these approaches often find the analysis of data daunting and are unsure of the tools to use for each step. SeqAcademy provides a single tool for all the steps in an easy to download package. Given that the target for this tool is investigators who do not have extensive expertise in the tools and analyses available, the article would benefit from more detailed descriptions of some individual components. Specifically, the following:
1. Alignment: The rationale for use of HISAT2 is clearly described. The time needed for this step should be discussed and, if possible, it would be helpful to provide an alternate local tool that could be used if users are unable to maintain consistent internet connectivity over the time needed for alignment completion.
2. RNA-Seq: A critical piece of RNA-Seq analyses are the statistical tools used to define differential expression. Some brief discussion of important cut-offs would be helpful.
3. Most tools incorporated into the package are described in 1-2 sentences. This should be consistent throughout, including for HTSeq, DESeq2.

The tool is really quite useful and I believe will be very valuable to novice investigators and students learning the pipeline of RNA-Seq or ChIP-Seq. Given that, it would be very helpful to

provide more, not less, explanation of some of the basic tools such that the educational goal of this package will be served.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 05 June 2018

https://doi.org/10.5256/f1000research.16196.r34335

✗   **Philip Ewels** (iD)

Stockholm University, Department of Biochemistry and Biophysics, Stockholm, Sweden

Bioinformatics training is of critical importance for the field of genomics, especially given the current shortage of experienced bioinformaticians and the utility of cross-training lab scientists in data analysis. In this manuscript, the authors describe a training tool developed to help teach people new to bioinformatics how to handle data from RNA and ChIP high throughput sequencing experiments.

Such a tool is valuable to the genomics community - it is a resource for interested parties to use in their own time and could also be very useful for those designing bioinformatics courses.

Manuscript

The manuscript does a good job of describing the resource, as well as the reason it is helpful. In my opinion, it is longer than needed - the Introduction could be more concise and the *Use cases* section would feel more natural in the introduction. The figures could also easily be condensed into one and the table be a supplementary figure. Making the manuscript shorter would make it more accessible.

Tutorial

The authors take an excellent approach to the design of their tutorial, using a Jupyter notebook and software installations from bioconda. These are appropriate, user friendly and provide an excellent starting point for well documented and reproducible analyses. Some dependencies are required, but these are well described on the tool's homepage and commonly used in bioinformatics. I attempted to run the tutorial on my laptop, but I ran into a few issues that required some debugging.

Broken steps
I came across a couple of problems in the notebook: I couldn't get the MultiQC commands to work - the mix of Python variables in the bash commands failed in my Jupyter notebook. I changed the command to use the following command:

!multiqc test/{"* test/".join(RNASeqoutrun)}* --quiet --outdir test/multiqc_rnaseq --force

Next, the Bedtools steps didn't work - the intersect command refers to a sample SRR6703663, which does not seem to be included anywhere in the tutorial before this point. I gave up at this point, so I'm not sure about the final steps. If the authors think that I did something wrong, I'd be happy to have another go.

I recommend trying to get some colleagues to run this tutorial from scratch to get more feedback.

External scripts
My main concern with the tutorial is that much of the downstream processing is done within R scripts outside of the Jupyter notebook. For me, this sort of misses the point - it would be preferable to have this code mixed in with the tutorial without any external scripts. It should be possible to have both Python (bash) and R in the same notebook, and this would make those steps part of the exercise rather than an abstract *"runDeseq.R"* script that requires the user to go and find and read through.

Data subsampling
Although the authors have endeavoured to use a small dataset (Yeast, not Human), the data used is still large. The authors note in the tutorial that the alignment *"will most likely take several hours."*, which limits the utility of the tutorial in any teaching scenario.

It should be quite easy to use the -u/--qupto flag in HiSAT2 to limit the number of alignments performed. I imagine that most of the downstream analysis will still work with subsampled data, and will run much faster. I tried doing this with 1,000,000 reads and the alignment step completed in a couple of minutes.

Tutorial - Minor points

Software requirements
I had trouble getting the bioconda installation command to first install and then activate inside the the notebook. I would instead recommend creating the conda environment on the terminal first (including jupyter as a package to install), activating it and then running jupyter. This is what I ended up doing to get the tutorial to work.

Streaming Input Data
The approach of streaming data from the SRA for the alignment step is a little non-standard. The reasoning is described in the paper as helping to reduce the disk space footprint required for the analysis. Whilst its cool that this step works, I would argue that it lessens the value of the tutorial, as most bioinformatics projects can not use such a method. Starting with regular static FastQ files would be more typical and useful for future analyses. This would also allow the additional step of FastQC for raw data analysis, which is pretty typical for such pipelines. The downside of this is that it invalidates my above subsampling idea! But SRA data could still be streamed to FastQ files with a similar subsampling approach.

Two-in-one analysis
The analysis presented uses common alignment and quality control steps for both sets of data, before splitting into bespoke RNA and ChIP-seq analysis. I see the reasoning for this but I think this complicates issues a little for users. Personally, I would prefer to see two separate Jupyter notebooks and separate analyses for the two different data types to avoid confusing the two. This would simplify future expansion of the resource across additional data types, which may not be able to share such analysis steps.

MultiQC reporting
The Quality Control step using MultiQC is placed after the HiSAT2 alignment. It would be better placed at the end of the analysis, as MultiQC also supports outputs from MACS, HTSeq and deepTools. A more complete report would be generated if it were run with all of these outputs.

Conclusions

I think that the aim of this paper is worthwhile and the approach is very nice. The tutorial itself feels a little unpolished and could do with a few trial runs on willing volunteers to iron out the issues I encountered. With a few tweaks and the changes above, I think that this tutorial could be the beginning of an excellent resource for budding bioinformaticians and educators.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research