A peer-reviewed version of this preprint was published in PeerJ on 22 November 2017.

View the peer-reviewed version (peerj.com/articles/4040), which is the preferred citable publication unless you specifically need to cite this preprint.

Ahmed Z, Ucar D. 2017. I-ATAC: interactive pipeline for the management and pre-processing of ATAC-seq samples. PeerJ 5:e4040 https://doi.org/10.7717/peerj.4040

A standalone software platform for the interactive management and pre-processing of ATAC-seq samples

Zeeshan Ahmed Corresp., [1] , **Duygu Ucar** Corresp. [2]

[1] University of Connecticut Health Center, Farmington, CT, United States

[2] The Jackson Laboratory For Genomic Medicine, Farmington, CT, United States

Corresponding Authors: Zeeshan Ahmed, Duygu Ucar
Email address: zahmed@uchc.edu, duygu.ucar@jax.org

Assay for Transposase Accessible Chromatin (ATAC-seq) is an open chromatin profiling assay that is adapted to interrogate chromatin accessibility from small cell numbers. ATAC-seq surmounted a major technical barrier and enabled epigenome profiling of clinical samples. With this advancement in technology we are now accumulating ATAC-seq samples from clinical samples at an unprecedented rate. These epigenomic profiles hold the key to uncover how transcriptional programs are established in diverse human cells and are disrupted by genetic or environmental factors. Thus, the barrier to deriving important clinical insights from clinical epigenomic samples is no longer one of data generation, but of data analysis. Specifically, we are still missing easy-to-use software tools that will enable non-computational scientists to analyze their own ATAC-seq samples. To facilitate systematic pre-processing and management of ATAC-seq samples, we developed an interactive, cross platform, user-friendly desktop application: interactive-ATAC (*I-ATAC*). *I-ATAC* integrates command-line data processing tools (e.g., *FASTQC* for quality checking) into an easy-to-use platform with user interface to automatically pre-process ATAC-seq samples with parallelized and customizable pipelines. Its performance has been tested using public ATAC-seq datasets in GM12878 and CD4+ T cells. *I-ATAC* is designed to empower non-computational scientists to process their own datasets and to break to exclusivity of data analyses to computational scientists.

A standalone software platform for the interactive management and pre-processing of ATAC-seq samples

Zeeshan Ahmed[2] and Duygu Ucar[1]

[1]The Jackson Laboratory for Genomic Medicine, 10 Discovery Dr., Farmington, CT, USA

[2]University of Connecticut Health Center, 195 Farmington Ave, Farmington, CT, USA

Corresponding author: **Duygu Ucar[1]**

Email address: duygu.ucar@jax.org

ABSTRACT

Assay for Transposase Accessible Chromatin (ATAC-seq) is an open chromatin profiling assay that is adapted to interrogate chromatin accessibility from small cell numbers. ATAC-seq surmounted a major technical barrier and enabled epigenome profiling of clinical samples. With this advancement in tecchnology we are now accumulating ATAC-seq samples from clinical samples at an unprecedented rate. These epigenomic profiles hold the key to uncover how transcriptional programs are established in diverse human cells and are disrupted by genetic or environmental factors. Thus, the barrier to deriving important clinical insights from clinical epigenomic samples is no longer one of data generation, but of data analysis. Specifically, we are still missing easy-to-use software tools that will enable non-computational scientists to analyze their own ATAC-seq samples. To facilitate systematic pre-processing and management of ATAC-seq samples, we developed an interactive, cross platform, user-friendly desktop application: interactive-ATAC (I-ATAC). I-ATAC integrates command-line data processing tools (e.g., FASTQC for quality checking) into an easy-to-use platform with user interface to automatically pre-process ATAC-seq samples with parallelized and customizable pipelines. Its performance has been tested using public ATAC-seq datasets in GM12878 and CD4+ T cells. I-ATAC is designed to empower non-computational scientists to process their own datasets and to break to exclusivity of data analyses to computational scientists.

INTRODUCTION

Assay for Transposase-Accessible Chromatin with high throughput Sequencing (ATAC-seq) is developed to profile chromatin accessibility from small cell numbers, making it uniquely suited to study epigenomic profiles of human clinical samples with a systems biology approach (Buenrostro *et al.*, 2013). ATAC-seq generates libraries via a simple two-step protocol using hyperactive Tn5 transposase, which inserts itself to open chromatin sites and generates double-strand breaks. ATAC-seq is attracting a growing interest in genomics applications due to its simple protocol, high sensitivity, and low expectations for starting material amounts (500–50,000 cells) (Tsompana and Buck, 2014). Therefore, data processing and management of samples generated by this new assay is becoming an important first step to study the open chromatin sites in diverse human cells.

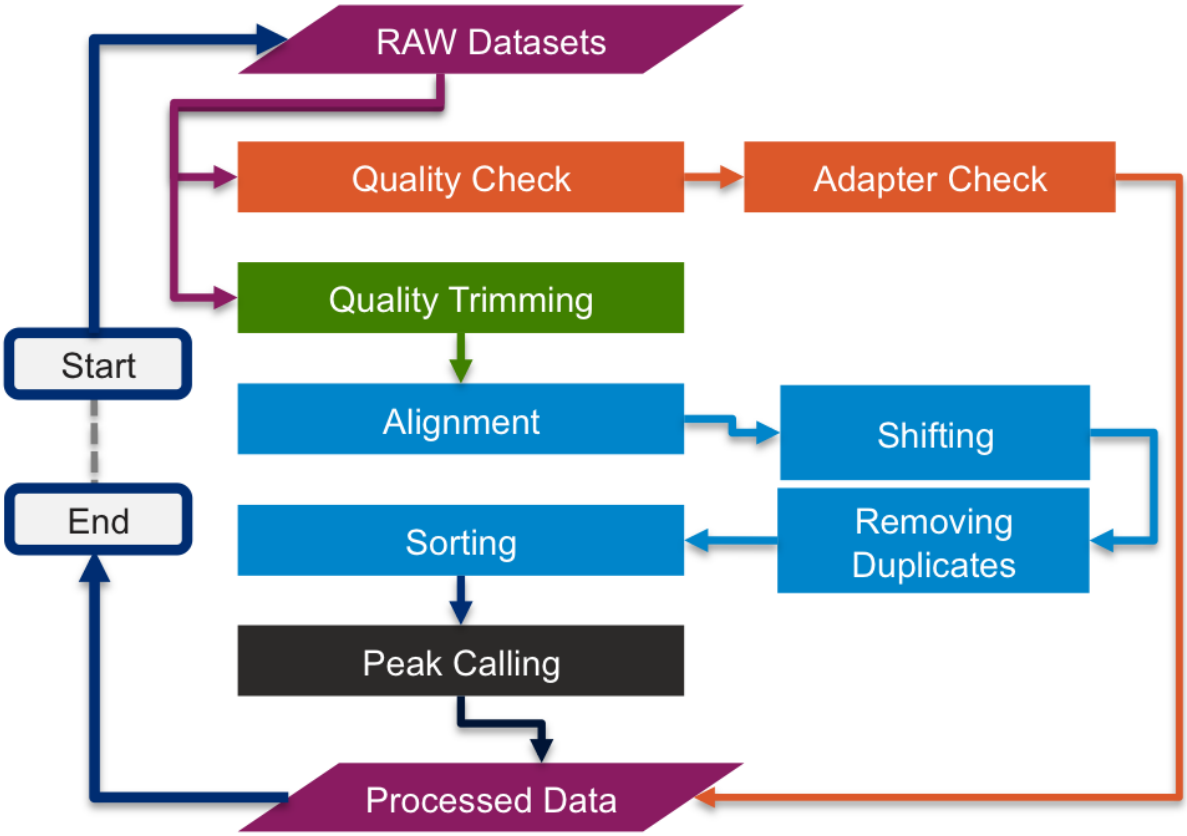

Figure 1. ATAC-seq data pre-processing pipeline's workflow.

Traditional way of next generation sequencing (NGS) data pre-processing is based on running a series of command-line applications, which requires good programming skills and ability to work in the UNIX environment (Fig. 1). Several integrated platforms exist to help in managing and building pipelines for NGS data pre-processing e.g. *Galaxy* (Scholtalbers *et al.*, 2013) (Giardine *et al.*, 2005), *SMITH* (Venco *et al.*, 2013), *SeqBench* (Dander *et al.*, 2014), *Wasp* (McLellan *et al.*, 2012), *NG6* (Mariette *et al.*, 2012), *openBIS* (Bauch *et al.*, 2011), etc. However, there is no open source software that is standalone, interactive, and easy-to-use, which enables biologists with no programming experience to analyze their ATAC-seq samples. To facilitate data analysis by the scientists who generate the data, we developed interactive and cross-platform software for the processing of ATAC-seq samples, namely Interactive-ATAC (I-ATAC) (Fig. 2).

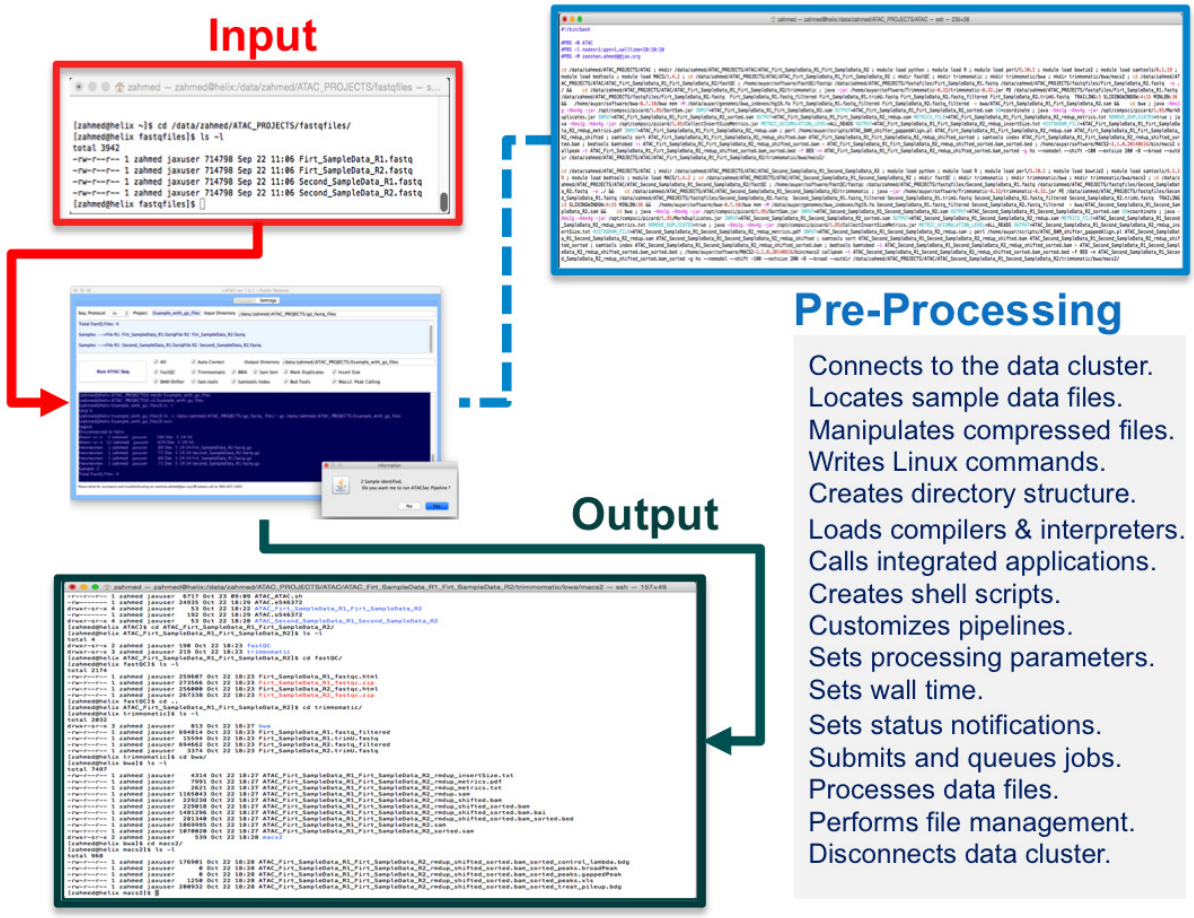

Figure 2. I-ATAC: Sample data input, data pre-processing & management, and output.

METHODS AND IMPLEMENTATION

I-ATAC is based on an I/O redirection framework (*FASTQ, FASTQ.gz, txt, sam, bam, bed, bdg, broadPeak, gappedPeak, xls, pdf and html*) that integrates several publicly available command-line tools within this framework for data quality control, adapter filtering, trimming, alignment, shifting, duplicate read filtering and peak calling. Within I-ATAC, we utilized *FASTQC* for computing the quality statistics; *Trimmomatic* (Bolger *et al.*, 2014) for the identification and trimming of the adapter and bad quality sequences; Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009) for aligning ATAC-seq reads to a reference genome; Sequence Alignment/Map (SAM) tools (Li, *et al.*, 2009) and *Picard* for generating, processing and viewing "sam" and "bam" files; Browser Extensible Data (BED) tools (Quinlan and Hall, 2010) for generating and processing "bed" files, and Model-based Analysis of ChIP-Seq (MACS) (Zhang, *et al.*, 2008) for identifying regions of the genome enriched in ATAC-seq reads (i.e., peaks) that are the putative open chromatin sites (Fig. 3).

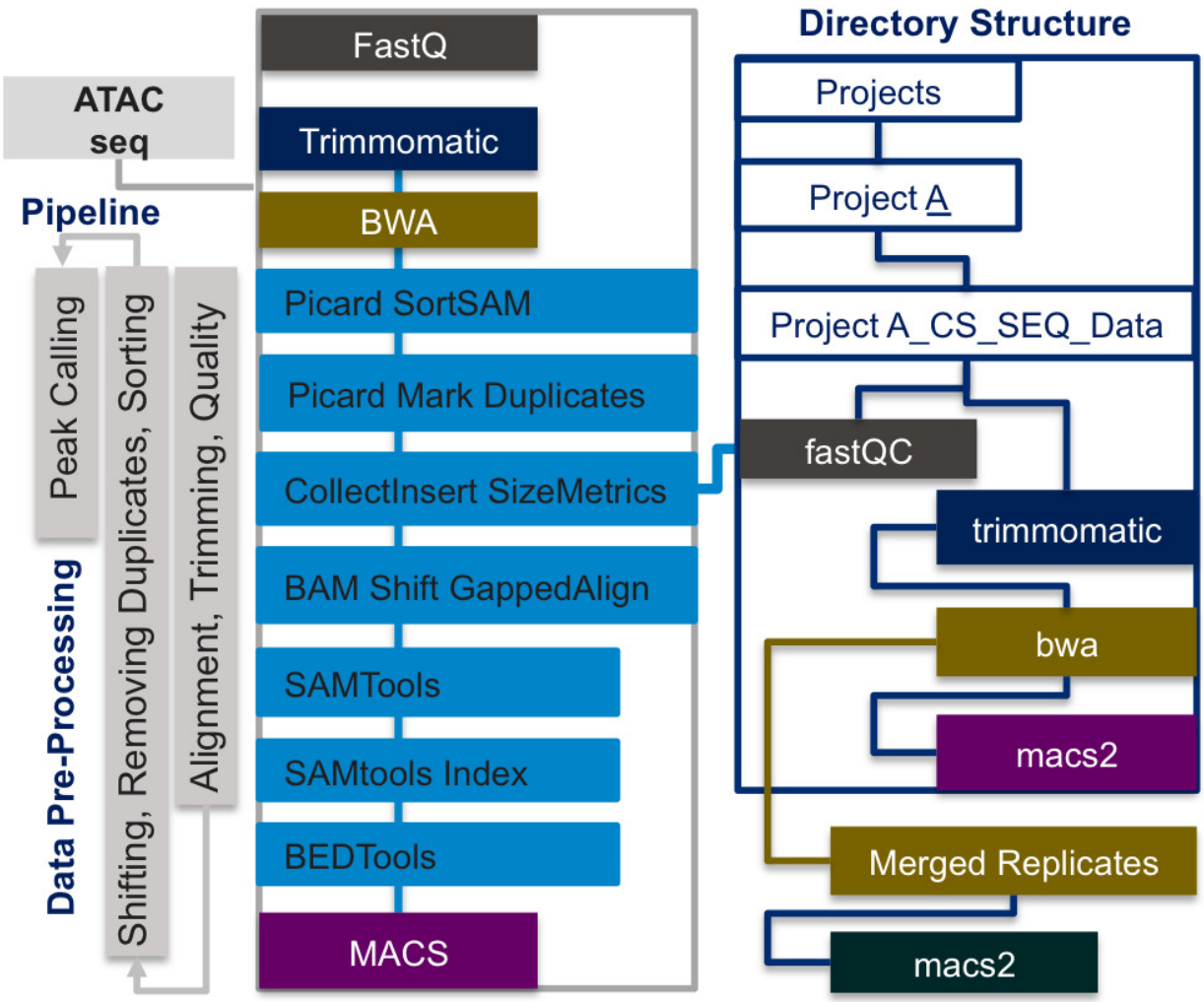

Figure 3. Integrated applications, data pre-processing steps and project directory structure.

I-ATAC is a platform designed by following software engineering principles for the sustainable bioinformatics software implementation (Ahmed *et al.*, 2014). It is a Java based desktop application, which requires Java Runtime Environment and all integrated applications to be installed in data cluster (or local computer) as well as the reference genomes that will be used for the alignment.

Operational Workflow of I-ATAC

The basic work flow of I-ATAC is very simple, as it requires only login information, project name and path to the samples files as input, however, pipeline operations can be customized (Fig. 4) by choosing the applications between *FASTQC, Trimmomatic, BWA, Sam Sort, Mark Duplicates, Insert Size, BAM Shifter, SAM tools, SAM tools index* and *BED tools*. To avoid exceptions, system will not let the user select any application without selecting its pre-requisites.

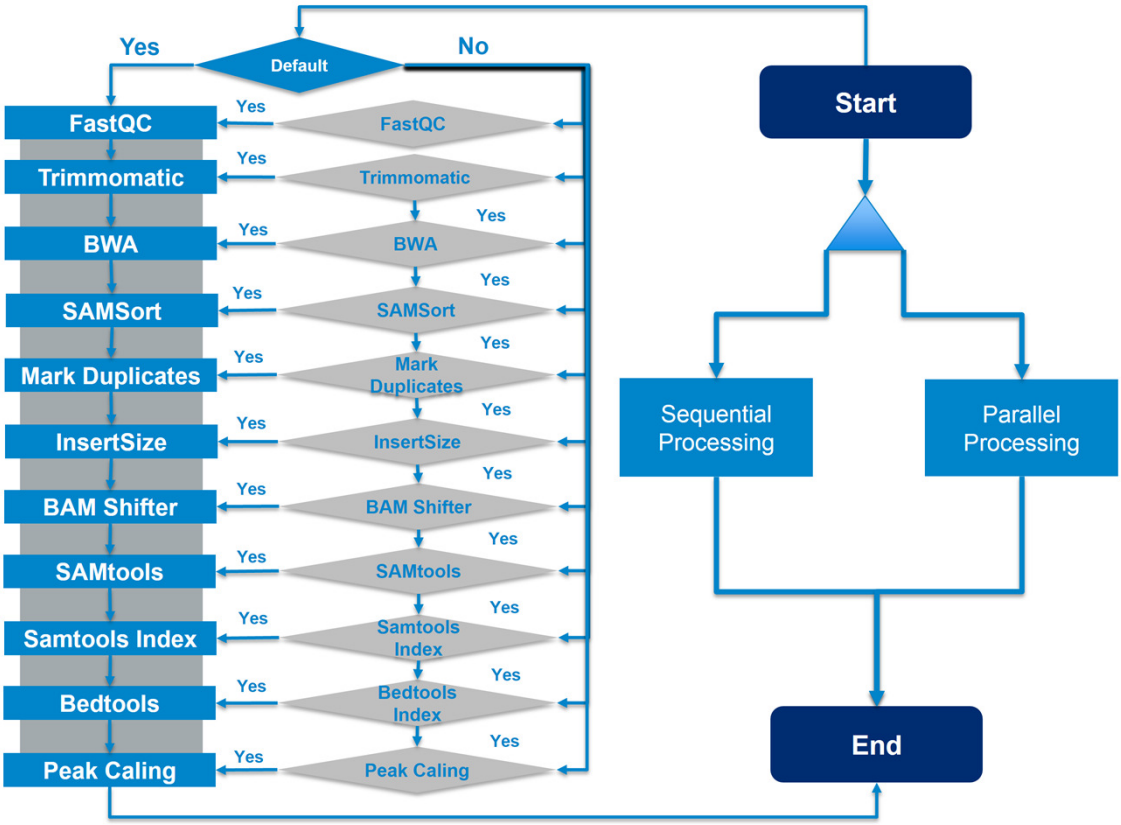

Figure 4. I-ATAC: Customization of ATAC-seq data pre-processing pipeline with sequential (multiple jobs in one script) and parallel (multiple jobs in multiple scripts, one of each) processing.

User can remotely handle sample data files for processing by either keeping them in the same parent directory and putting only pre-processed results in the main project and sub-project directories or by first copying compressed files into the project directory, unzips them and then process them. Additionally, I-ATAC allows user to automatically create and submit one sequential job (Unix based Secure Shell Scripts) for multiple samples, as well as, creating and submitting multiples parallel jobs for multiple samples (one for each).

DATA PROCESSING AND RESULTS

We have applied I-ATAC to several publicly available ATAC-seq datasets in GM12878 and CD4+ T cells to process these samples with the help of an easy to use software platform. Along with the recommended settings for tools that we use for pre-processing, the only input to the I-ATAC is the path to the directory where ATAC-seq samples (single or paired end) can be found (FASTQ files). With just one click operation (by pressing “Run ATAC-Seq” button), it automatically connects and interacts with the data cluster to locate sample data files, writes command line instructions, manipulates (copy, paste, unzip) compressed input files, loads compilers & interpreters, calls applications, creates shell scripts, generates multiple, parallel, sequential and customized data analysis pipelines, submits and queues jobs, creates output directory structure, processes data files, places output data files in relevant directories, sets notifications and disconnects to the connected data cluster (Fig. 5).

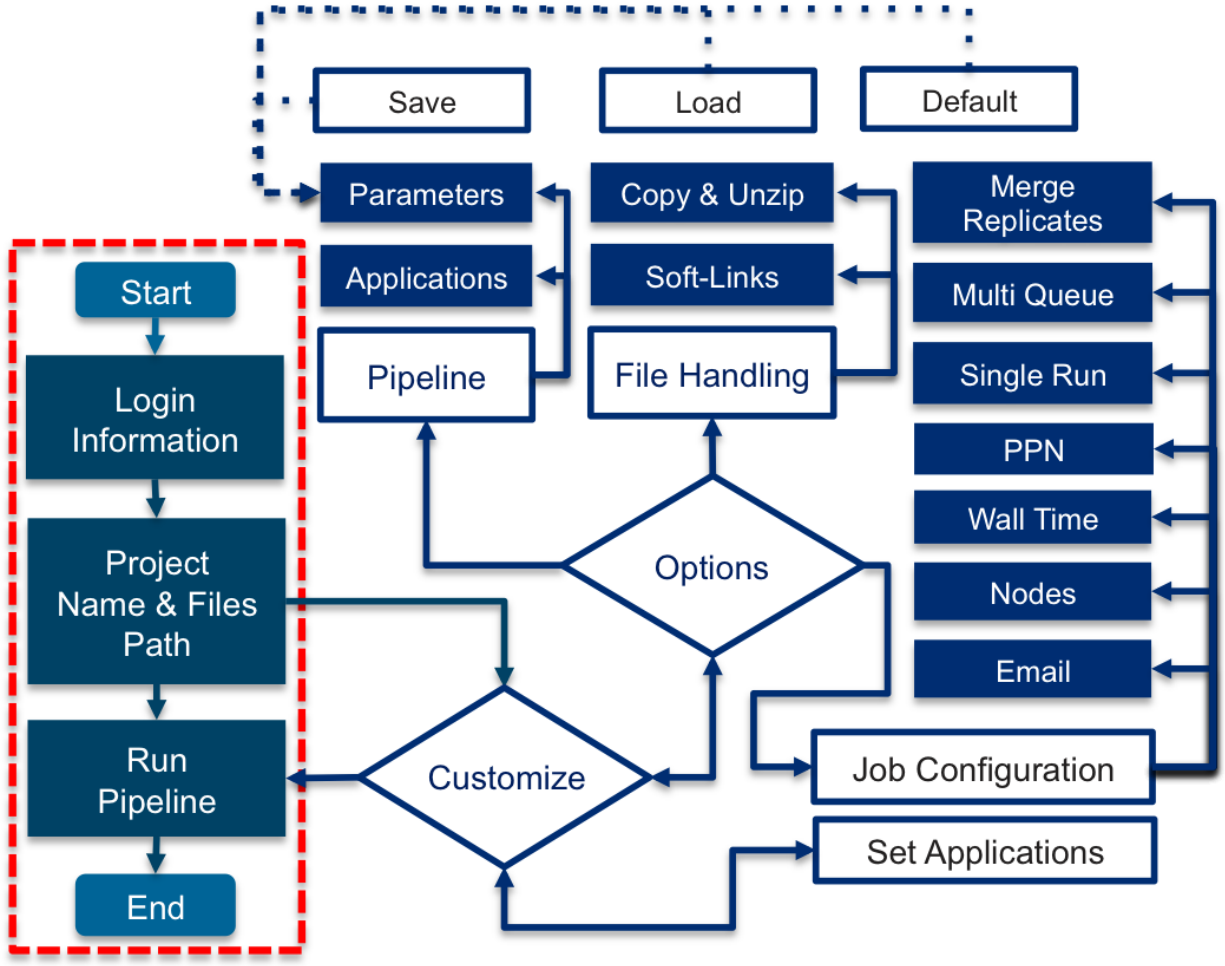

Figure 5. Direct and customized components workflow of I-ATAC.

I-ATAC also enables users to customize parameters (Fig. 6) used for data pre-processing steps (Fig. 7) by letting the user to choose between applications as well as by setting different parameters, which enables customizing this pipeline for the analyses of other data types, such as ChIP-seq data. As the output, I-ATAC produces data quality reports that can be visualized within the platform. It also outputs ATAC-seq reads that are filtered, trimmed and aligned as well as peak calls from these reads. These peaks can be visualized using frequently used genome browsers (e.g. *USCS*, *IGV*) and can be further processed for annotation and for differential open chromatin detection. Features of I-ATAC platform are explained in detail with two example case studies in the attached supplementary material.

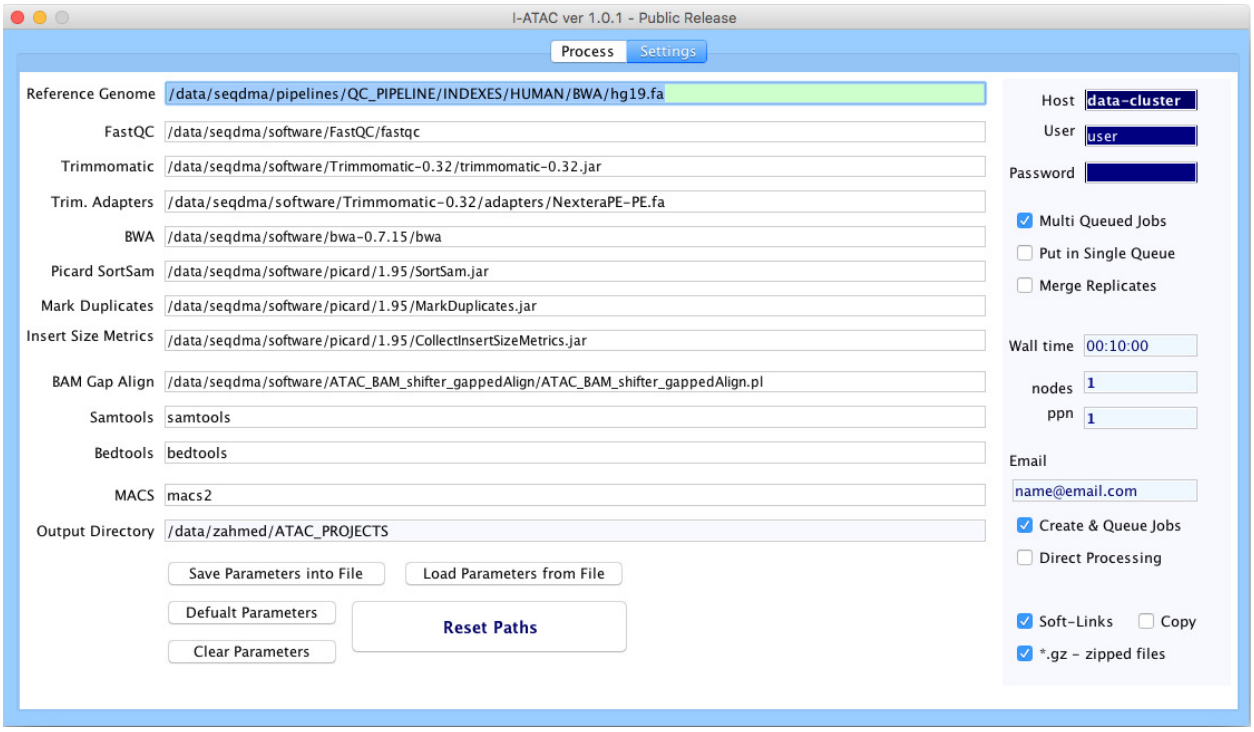

Figure 6. Graphical User Interface of I-ATAC: Set parameters and user credentials.

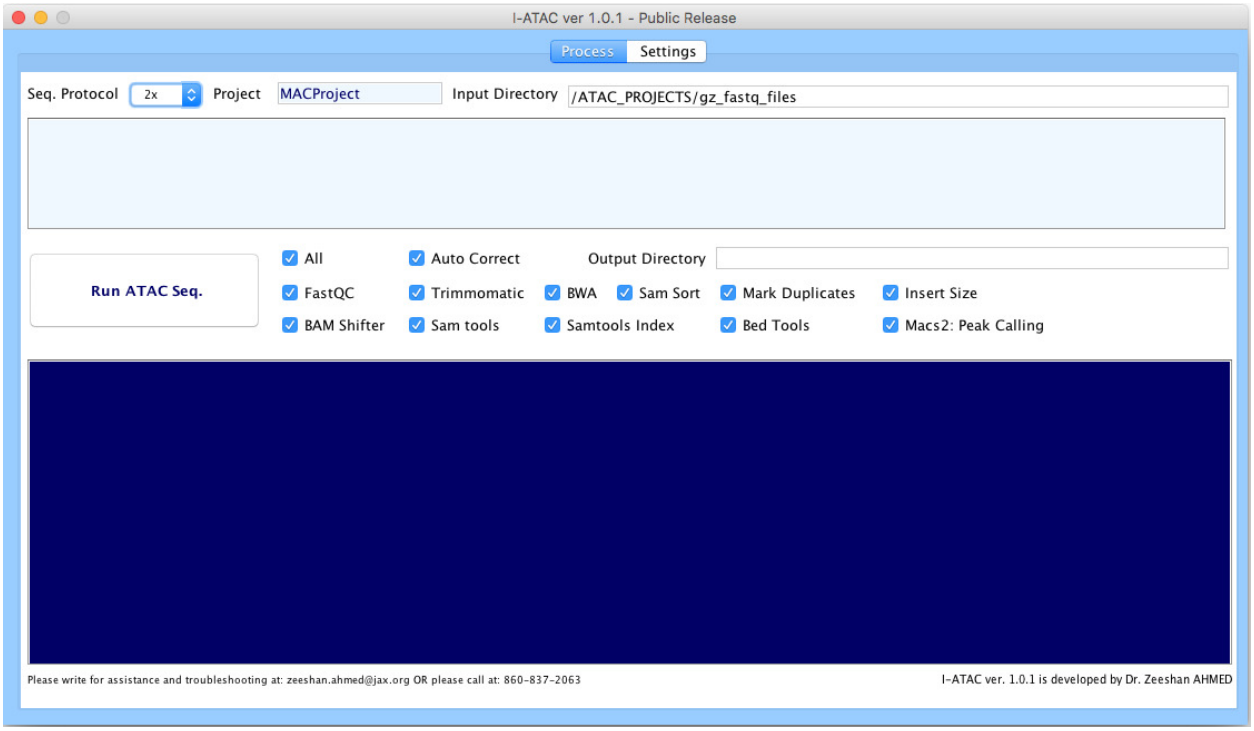

Figure 7. Graphical User Interface of I-ATAC: Create and run data processing jobs.

CONCLUSION

One of the major requirements for the downstream analysis of any kind of genomics data (e.g. RNA-seq, ChIP-seq, ATAC-seq etc.) is to first demultiplex and then pre-process FASTQ files using respective data pre-processing pipelines. The focus of this study is to develop an interactive, cross platform software for ATAC-seq data pre-processing. Many bioinformatics tools are open source and publicly available, which are helpful in compiling pipelines for ATAC-seq data pre-processing e.g. *ENCODE's pipelines, Galaxy Biostar* etc. However, these pipelines assume that the data analyst can operate with command line tools, which is not always the case. Moreover, none of the available pipelines have a cross platform graphical user interface, which can be helpful in supporting non-computational scientist in tracking FASTQ files, loading default/customized settings, creating automatic directory structure, automatically generating shell scripts and submitting jobs to the attached data clusters, regardless of the number and kind (single or paired end) of input FASTQ files. To overcome these limitations in current command line pipelines, we have developed a novel platform i.e. I-ATAC, that facilitates processing of ATAC-seq samples by non-computational scientists (Fig. 4). We have successfully tested I-ATAC on ATAC-seq single and paired end data (in-house and publicly available) at The Jackson Laboratory for Genomics Medicine, USA. I-ATAC enables easy generation and tracking of output files including FASTQ files with high quality reads (trimmed out sequence adapter and low quality reads), sorted SAM, BAM and BED files. While I-ATAC have been implemented and well tested with ATAC-seq data but it can also be applied to perform quality checking and pre-preprocessing of WGS (whole genome sequencing) and ChIP-seq data.

ACKNOWLEDGEMENT

The Jackson Laboratory (JAX) supports and owns this project. Special thanks to Ucar and Banchereau labs as well as the Genome Technologies and Computational Sciences cores at JAX, who provided insight and expertise that greatly, assisted the research and development of this platform.

FUNDING STATEMENT

This work was supported by The Jackson Laboratory, USA.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

SUPPLEMENTARY MATERIAL

I-ATAC tutorial is provided.

ADDITIONAL REQUIREMENTS

For additional information, please refer to the project webpage: https://www.jax.org/research-and-faculty/tools/i-atac. Source code, JAR files for MAC OS X and Windows, and complete source code package for Eclipse IDE is available at https://github.com/UcarLab/I-ATAC. Example dataset is available at: https://zenodo.org/record/46079#.WAe3l5MrK7Y. Supporting software and dependencies are available at: https://zenodo.org/record/162023#.WAe3dJMrK7Y

REFERENCES

Ahmed, Z., Zeeshan, S., Dandekar, T. (2014) Developing sustainable software solutions for bioinformatics by the "Butterfly" paradigm. *F1000Res*., **3**, 71.

Bauch, A. (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinf*., **12**, 468.

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

Bolger, A.M., Lohse, M., Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*., **30**, 2114-20.

Dander, A., *et al.* (2014) SeqBench: integrated solution for the management and analysis of exome sequencing data. *BMC Res Notes*., **7**, 43.

Giardine, B. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. **15**, 1451-1455.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*., **25**, 1754-60.

Li, H., *et al*. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics.,* **25**, 2078-2079.

Mariette, J., *et al.* (2012) NG6: Integrated next generation sequencing storage and processing environment. *BMC Genomics*. **13**, 462.

McLellan, A. S., *et al.* (2012) The Wasp System: an open source environment for managing and analyzing genomic data. *Genomics*., **100**, 345-51.

Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*., **26**, 841-842.

Scholtalbers, J. (2013) Galaxy LIMS for next-generation sequencing. *Bioinformatics*. **29**,1233-1234.

Tsompana, M and Buck, M. J. (2014) Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*., **7**, 33.

Venco, F., *et al.* (2014) SMITH: a LIMS for handling next-generation sequencing workflows. *BMC Bioinf*., **15**, (Suppl 14):S3.

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