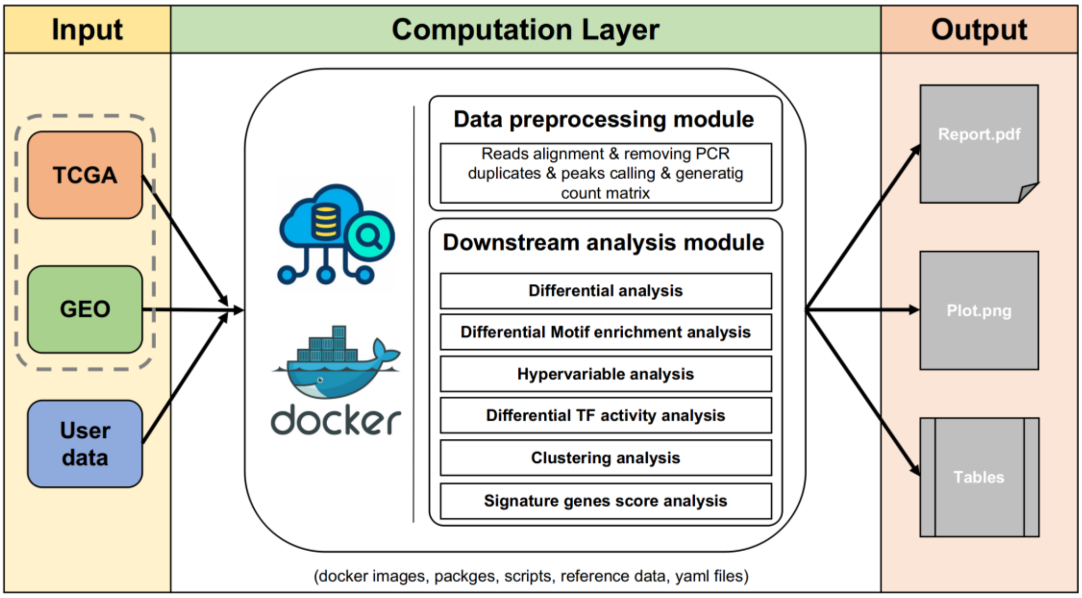
Introduction

Epigenome profiling is critical for gene expression regulation studies for individual development and disease progression. Epigenome sequencing, such as ChIP-seq and ATAC-seq have enabled us to dissect the epigenetic heterogeneity of development cells and disease cohorts by identifying epigenomic alterations and the potential associated transcription factors. A tremendous amount of ChIP/ATAC-seq data have been generated by large consortium projects such as ENCODE and TCGA, providing opportunities for data mining and broader understanding of gene expression regulation. Additionally, there is no systematic epigenomic analysis tools to explore the ChIP/ATAC-seq data from various studies deposited in NCBI GEO. Despite the availability of large set of computational tools and methods for ChIP/ATAC-seq data analysis, it is still challenging for experimental biologist to deploy them and integrate these tools into workable pipeline. Especially in heterogeneous cohort study (e.g. cancer cohort), conventional analysis tools are not applicable. To fill these gaps, we have developed EAP (Epigenomic Analysis Platform, https://sdap.biosino.org/epigenetics/), a customizable and interactive ChIP/ATAC-seq data analysis web tool. EAP uses state-of-the-art statistical algorithms to transform ChIP/ATAC-seq data from heterogeneous samples into biologically meaningful and interpretable results. At present, EAP provides data preprocessing, differential analysis, differential TF motifs enrichment analysis, hypervariable analysis, differential TF activity analysis, clustering analysis and signature genes score analysis, all these analysis tools are specially developed for heterogeneous disease cohort study. The comprehensive epigenomic analysis through EAP can greatly facilitate data mining in different research areas, such as cancer subtyping and therapeutic target discovery.

EAP leverages cloud computing technology for large scale data set analysis and docker container technology to enable reproducible results. These greatly facilitate the end-to-end bioinformatics analysis. EAP data preprocessing pipeline is implemented as a workflow task, or interchangeably referred to as Data preprocessing module. This module uses a standardized analysis pipeline to quality control, read alignment, peak calling and read counting, and also creates a PDF report including quality control plots and summary statistics to facilitate filtering poor quality samples. A comprehensive collection of ChIP/ATAC-seq data analysis tools are encapsulated in the Down-stream analysis module and provides a simple interface for users to customize their analysis parameters (such as adjusted p-value cutoff, number of clusters, the number of principal components used for hierarchical clustering and variable of interest used for differential analysis). Altogether, EAP provides a reproducible, customizable and interactive ChIP/ATAC-seq data analysis platform capable of exploring the epigenetic heterogeneity in development cells and disease cohorts.


Input
Computation Layer
Output
TCGA
GEO
User data
Data preprocessing module
Reads alignment & removing PCR duplicates & peaks calling & generatig count matrix
Downstream analysis module
Differential analysis
Differential Motif enrichment analysis
Hypervariable analysis
Differential TF activity analysis
Clustering analysis
Signature genes score analysis
docker
Report.pdf
Plot.png
Tables
(docker images, packges, scripts, reference data, yaml files)

Overview of EAP architecture and the various analysis modules