Bioinformatics, 33(5), 2017, 740–742
doi: 10.1093/bioinformatics/btw688
Advance Access Publication Date: 5 December 2016
Applications Note

Genome analysis

ALTRE: workflow for defining ALTered Regulatory Elements using chromatin accessibility data

Elizabeth Baskin†, Rick Farouni† and Ewy A. Mathé*

Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on July 14, 2016; revised on October 23, 2016; editorial decision on October 25, 2016; accepted on October 26, 2016

Abstract

Summary: Regulatory elements regulate gene transcription, and their location and accessibility is cell-type specific, particularly for enhancers. Mapping and comparing chromatin accessibility between different cell types may identify mechanisms involved in cellular development and disease progression. To streamline and simplify differential analysis of regulatory elements genome-wide using chromatin accessibility data, such as DNase-seq, ATAC-seq, we developed ALTRE (ALTered Regulatory Elements), an R package and associated R Shiny web app. ALTRE makes such analysis accessible to a wide range of users—from novice to practiced computational biologists.
Availability and Implementation:https://github.com/Mathelab/ALTRE
Contact: ewy.mathe@osumc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Assays that measure chromatin accessibility genome-wide, such as FAIRE-seq (Giresi *et al.*, 2007), DNase-seq (Crawford *et al.*, 2006; John *et al.*, 2013; Thurman *et al.*, 2012), and ATAC-seq (Buenrostro *et al.*, 2013), enable global mapping of regulatory elements (REs), including promoters and enhancers. Organization of these REs, particularly enhancers, is cell-type specific (Kieffer-Kwon *et al.*, 2013; Rendeiro *et al.*, 2016; Stergachis *et al.*, 2013) and is a strong determinant of disease mutational landscapes, including those of cancer (Polak *et al.*, 2015). Thus, identifying REs that differ in accessibility between cell types, such as cancerous and non-cancerous cell lines and tissues, holds promise for pinpointing mechanisms involved in disease progression. Furthermore, REs that control disease-related genes and pathways can be investigated as putative therapeutic targets, or may even be such targets themselves (Heinz *et al.*, 2015; Lam *et al.*, 2013).

To the best of our knowledge, no comprehensive and user-friendly workflow for downstream analysis of chromatin accessibility data is available. Downstream analysis includes guiding chromatin accessibility alignment and peak data to interpretable results of REs and pathways of interest. However, there are no standardized approaches or guidelines. Typically, individual data analyses pipelines must be created from scratch in-house, thereby making reproducible, shareable data-analysis difficult. ALTRE provides a workflow so users can identify altered REs between two different cell types or conditions, and includes a Shiny (RStudio shiny: Easy web applications in R. 2014) web interface for those not as fluent in the R statistical language.

2 Implementation

2.1 Data preparation and set-up

Typical of high-throughput sequencing data, chromatin accessibility data are delivered in FASTQ files. Quality control, alignment and peak calling of the FASTQ file reads, described in detail elsewhere (Baek *et al.*, 2012; Boyle *et al.*, 2008; Jalili *et al.*, 2016; Rashid *et al.*, 2011; Zhang *et al.*, 2008), must be performed before using ALTRE. To start the ALTRE workflow, users need to generate a comma-separated-values CSV file with 4 columns for each sample to be analyzed: (1) name of alignment (BAM) files; (2) name of peak (BED) files; (3) sample name; (4) replicate number. All files should be placed in the same folder and the software will detect the location of the files when reading in the CSV. A minimum of 2 replicates per sample is required to run the workflow. To get started with ALTRE, users need to have R (≥3.2.0) installed.

2.2 General aspects and design

ALTRE was designed to be user-friendly and to streamline differential analysis of REs genome-wide. The steps of the workflow analysis are delineated in Figure 1 and include loading data, defining consensus peaks (found in multiple replicates), annotating (e.g. Transcription Start Site (TSS)-distal and TSS-proximal) and optionally merging peaks, identifying significantly altered REs based on quantitative data using DESeq2 (Love *et al.*, 2014), creating tracks for visualizing categorized REs in a genome browser, comparing altered REs with those defined based on binary (peak present/absent) data only, and finally, defining pathways that are enriched in cell- or condition-type specific or shared REs using GREAT (Gu, Z. rGREAT: Client for GREAT Analysis. R package version 1.4.2. 2016; McLean *et al.*, 2010).

ALTRE's embedded Shiny app takes alignment files (BAM format) and hotspot/peak files (BED format) as input. The workflow guides users through the steps described above and delineated in Figure 1. At each step, users can define thresholds, such as number of replicate samples required to define a peak as consensus, and fold changes and p-value cutoffs for definition of cell type specific or shared REs. Users can then quickly retrieve summary statistics and visualization plots (heatmaps, barplots) to ensure the appropriateness of their parameters. For ease of use, default options are provided at each step for guidance. Of note, while tools for differential binding and annotation of sequencing data exist (Bailey *et al.*, 2013; Chabbert *et al.*, 2016; Ross-Innes *et al.*, 2012; Yu *et al.*, 2015; Zhu, 2013; Zhu *et al.*, 2010; Stark and Brown, 'DiffBind: differential binding analysis of ChIP-Seq peak data' 2011), ALTRE supports peak merging and annotation, differential analysis and pathway enrichment analysis in one streamlined tool.

3 Results and discussion

Users can install ALTRE with the function install_github() from the devtools R package (Wickham H and Chang, W. 2016. devtools: Tools to Make Developing R Packages Easier). Full installation instructions are found at https://github.com/Mathelab/ALTRE. Users can then run the workflow either in the R console or by launching the embedded web application by typing 'runShinyApp()' in the R console. A detailed vignette (https://mathelab.github.io/ALTRE/vignette.html) walks users through an example workflow analysis step-by-step.

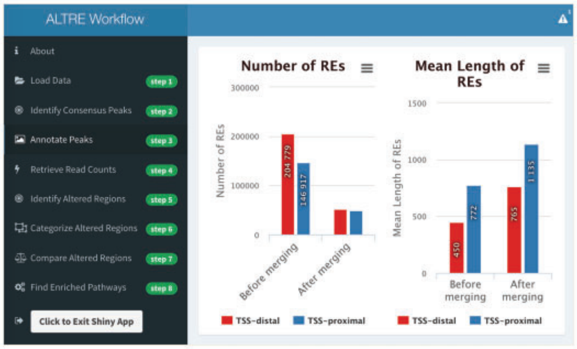

Fig. 1. Snapshot of ALTRE Shiny web application showing workflow steps

A sample dataset is provided on GitHub and can be accessed at https://mathelab.github.io/ALTREsampledata/. This sample dataset includes ENCODE data for cancerous and associated non-cancer lung cell lines, A549 and SAEC, respectively. On a machine with 16 GB memory and a 2.5 GHz Intel Core i7 processor, the workflow takes ∼334 s to complete for the example dataset using all chromosomes.

For real-time analysis of results, the ALTRE Shiny app enables users to change their parameters and directly visualize the effect of those changes through summary statistics tables and plots. For example, users can readily visualize the number of REs that are sample-type specific or shared based on their input fold change and adjusted *P*-value thresholds through a volcano plot and an associated statistics table. In addition, processed data can be saved after key steps in the analysis and all plots can be modified (e.g. colors) and saved as high resolution images.

With the increasing interest in researching REs to better understand transcriptional regulation and diseases, and improvements in techniques to assess these regions (Buenrostro *et al.*, 2013), chromatin accessibility assays are being increasingly generated. With this in mind, ALTRE provides a user-friendly workflow that guides the analysis and interpretation of these data.

Funding

This work was supported by The Ohio State University and the Translational Data Analytics Initiative.

Conflict of Interest: none declared.

References

Baek,S. *et al.* (2012) Quantitative analysis of genome-wide chromatin remodeling. *Methods Mol. Biol.*, **833**, 433–441.

Bailey,T. *et al.* (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.

Boyle,A.P. *et al.* (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.

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.

Chabbert,C.D. *et al.* (2016) DChIPRep, an R/Bioconductor package for differential enrichment analysis in chromatin studies. *PeerJ*, **4**, e1981.

Crawford,G.E. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**, 123–131.

Giresi,P.G. *et al.* (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.

Heinz,S. *et al.* (2015) The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell. Biol.*, **16**, 144–154.

Jalili,V. *et al.* (2016) MuSERA: multiple sample enriched region assessment. *Brief Bioinform*, DOI: 10.1093/bib/bbw029.

John,S. *et al.* (2013) Genome-scale mapping of DNase I hypersensitivity. *Curr. Protoc. Mol. Biol.*, Chapter 27:Unit 21 27.

Kieffer-Kwon,K.R. *et al.* (2013) Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, **155**, 1507–1520.

Lam,M.T. *et al.* (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature*, **498**, 511–515.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Polak,P. *et al*. (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, **518**, 360–364.

Rashid,N.U. *et al*. (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol*., **12**, R67.

Rendeiro,A.F. *et al*. (2016) Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat. Commun*., **7**, 11938.

Ross-Innes,C.S. *et al*. (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

Stergachis,A.B. *et al*. (2013) Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, **154**, 888–903.

Thurman,R.E. *et al*. (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

Yu,G. *et al*. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.

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

Zhu,L.J. (2013) Integrative analysis of ChIP-chip and ChIP-seq dataset. *Methods Mol. Biol*., **1067**, 105–124.

Zhu,L.J. *et al*. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinf*., **11**, 237.