Genome Analysis

RAD: a web application to identify region associated differentially expressed genes

Yixin Guo[1], Ziwei Xue[1], Ruihong Yuan[1], Jingyi Jessica Li[2], William A. Pastor[3] and Wanlu Liu[1,4,5*]

[1]Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, Haining 314400, China., [2]Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA., [3]Department of Biochemistry, McGill University, Montreal, QC H3G 1Y6, Canada., [4]Department of Orthopedic, the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310029, China., [5]Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Zhejiang University.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on July 10, 2020; revised on December 30, 2021; accepted on January 22, 2021

Abstract

With the advance of genomic sequencing techniques, chromatin accessible regions, transcription factor binding sites and epigenetic modifications can be identified at genome-wide scale. Conventional analyses focus on the gene regulation at proximal regions; however, distal regions are usually less focused, largely due to the lack of reliable tools to link these regions to coding genes. In this study, we introduce RAD (Region Associated Differentially expressed genes), a user-friendly web tool to identify both proximal and distal region associated differentially expressed genes (DEGs). With DEGs and genomic regions of interest (gROI) as input, RAD maps the up- and down-regulated genes associated with any gROI and helps researchers to infer the regulatory function of these regions based on the distance of gROI to differentially expressed genes. RAD includes visualization of the results and statistical inference for significance.

Availability: RAD is implemented with Python 3.7 and run on a Nginx server. RAD is freely available at http://labw.org/rad as online web service.

Contact: wanluliu@intl.zju.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Data-rich methods such as micrococcal nuclease sequencing (MNase-seq, Schones *et al.*, 2008), DNase I sequencing (DNase-seq, Boyle *et al.*, 2008), chromatin immunoprecipitation sequencing (ChIP-seq, Barski *et al.*, 2007) assay for transposase-accessible chromatin sequencing (ATAC-seq, Buenrostro *et al.*, 2013) and whole genome bisulfite sequencing (WGBS, Cokus *et al.*, 2008) that analyze genome-wide epigenetic landscape have been widely used to provide information on the binding of transcription factors (TFs) and chromatin accessibility of cis-regulatory elements (CREs) including promoters and enhancers. Through peak or DMR (differential methylated regions) calling, one can identify genomic regions of interest (gROI) in genomic data, which provides the basis for further analysis. Integration of these methods with RNA sequencing (RNA-seq) data allows researchers to determine whether differentially expressed genes (DEGs) are regulated by TF binding, chromatin accessibility, or other epigenetic modifications such as DNA methylation.

Methods for the integrative analysis of multi-level omics data have been introduced in recent years, such as GREAT, which incorporates ChIP-seq data and gene ontologies to highlight the association between CREs and gene function (McLean *et al.*, 2010). BETA is another new generation tool incorporating transcriptome and ChIP-seq data to infer direct target genes (Wang *et al.*, 2013). More recently, LISA is introduced to predict transcriptional regulators based on chromatin models constructed by user-defined gene sets, histone mark ChIP-seq data, and chromatin accessibility profiles (Qin *et al.*, 2020). BART is a recently developed software package for predicting functional transcription factor using gene sets or ChIP-seq datasets as input (Wang *et al.*, 2018). Combining multi-omics data may provide new insights into the epigenetic regulation of transcription.

Conventional analyses that focus on proximal regulatory events often omit information distal to gROI. Unlike promoters that are adjacent to transcription start site (TSS) (⩽ 1kb), enhancers may activate their target promoters and regulate the expression of target genes from long distance (Shlyueva *et al.*, 2014). In this study, we introduce a user-friendly web application, Region Associated DEGs (RAD), to intuitively measure both proximal and distal association between TF binding, chromatin accessibility, epigenetic modification or any other gROI and the transcriptional changes of surrounding genes. Using a hypergeometric or binomial statistical test, we can potentially infer whether nearby genes are up-regulated or down-regulated by differential TF binding, chromatin accessibility or epigenetics changes, and whether this regulation is mediated via proximal and/or distal interaction. RAD has several differences comparing to existing computational tools. For example, GREAT emphasizes the pathway analysis with a list of regions, while RAD infers the transcriptional regulatory function of those gROI based on nearby DEGs.

The algorithm used in RAD has been successfully implemented in recent publications to investigate the association of DEGs with chromatin accessibility changes (Pastor *et al.*, 2018), TF binding (Harris *et al.*, 2018) and DMRs (Gallego-Bartolomé *et al.*, 2019). The web application thus allows users to infer potential regulatory effects of transcription factors, epigenetic modifications or any gROI in genomic data.

2 Usage

2.1 RAD Functions

RAD is an open-access, user-friendly web application (Supplemental Fig. 1-4) for studying the relationship between gROI and DEGs. Visualization of DEGs surrounding gROI are implemented in RAD to help researchers to infer the potential regulatory function of transcription factors or chromatin features.

2.2 RAD Input

The input files for RAD include three files: 1-2) line-break text file containing up- or down-DEGs (*upregulated_genes.txt*, *downregulated_genes.txt*); 3) file containing gROI information in browser extensible data format (*gROI_file.bed*). Up- or down-regulated genes can be calculated with DESeq2 (Love *et al.*, 2014) or other methods that identify differentially expressed genes. Genes in the up- or down-DEGs files should be separated by line breaks (i.e. each line should only contain one gene symbol or Ensembl ID). gROI file provides the genomic region information including chromosome number, start and end position of the region. Instead of uploading the data, the user can also directly paste a list of gene names and genomic regions into the text-input area on the website. Required options include user-defined reference genome and gROI extended distance. Reference genome corresponding to the data should be specified by the user and we support several widely used mammalian (*Homo sapiens*, *Mus musculus*) and plant (*Arabidopsis thaliana*) reference genomes, including GRCh38, GRCh37, GRCm38, GRCm37, and TAIR10 (www.ensembl.org). gROI extended distance can be chosen from 1kb, 10kb, 25kb, 50kb, 100kb, 500kb, and 1000kb with 1000kb as default. Any other extended distance of interest could also be customized according to the user's preference. The user can also define the title and color palette for the output bar plot.

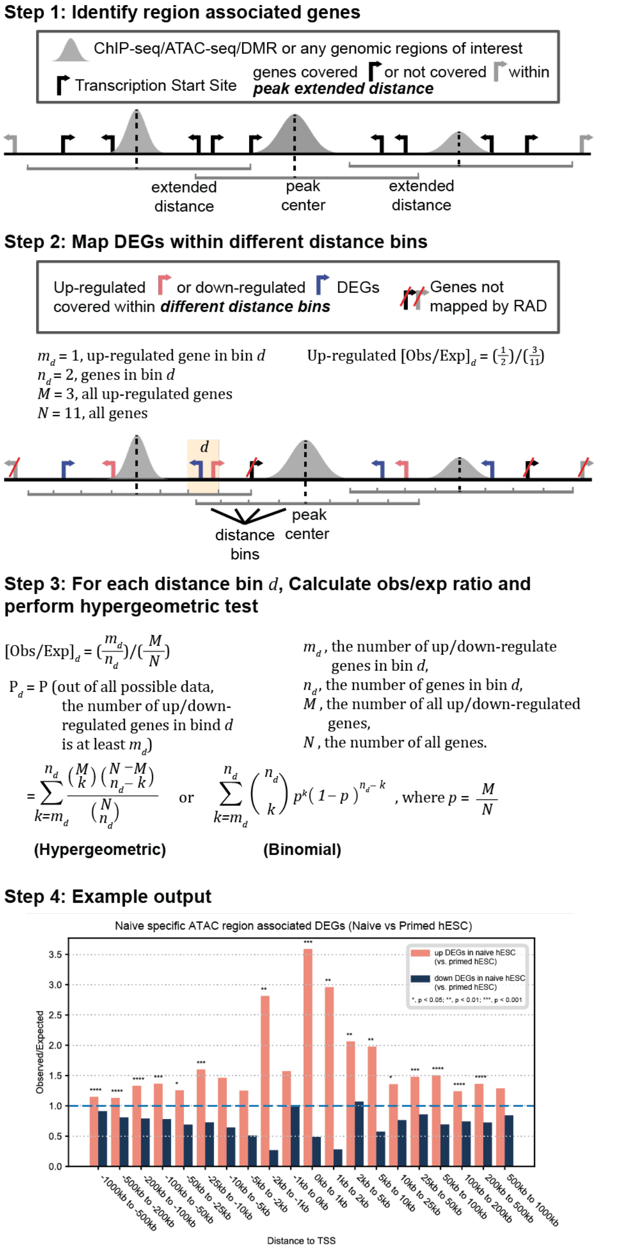
Step 1: Identify region associated genes
ChIP-seq/ATAC-seq/DMR or any genomic regions of interest
Transcription Start Site
genes covered or not covered within
peak extended distance
extended distance
peak center
extended distance
Step 2: Map DEGs within different distance bins
Up-regulated or down-regulated DEGs covered within different distance bins
Genes not mapped by RAD
m_d = 1, up-regulated gene in bin d
n_d = 2, genes in bin d
M = 3, all up-regulated genes
N = 11, all genes
Up-regulated $[\mathrm{Obs/Exp}]_d = (\frac{1}{2})/(\frac{3}{11})$
d
distance bins
peak center
Step 3: For each distance bin d, Calculate obs/exp ratio and perform hypergeometric test
$[\mathrm{Obs/Exp}]_d = (\frac{m_d}{n_d})/(\frac{M}{N})$
P_d = P (out of all possible data, the number of up/down-regulated genes in bind d is at least m_d)
$= \sum_{k=m_d}^{n_d} \frac{\binom{M}{k}\binom{N-M}{n_d-k}}{\binom{N}{n_d}}$ or $\sum_{k=m_d}^{n_d} \binom{n_d}{k} p^k(1-p)^{n_d-k}$, where $p = \frac{M}{N}$
(Hypergeometric)
(Binomial)
m_d, the number of up/down-regulate genes in bin d,
n_d, the number of genes in bin d,
M, the number of all up/down-regulated genes,
N, the number of all genes.
Step 4: Example output
Naive specific ATAC region associated DEGs (Naive vs Primed hESC)
Observed/Expected
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
up DEGs in naive hESC (vs. primed hESC)
down DEGs in naive hESC (vs. primed hESC)
*, p < 0.05; **, p < 0.01; ***, p < 0.001
-1000kb to -500kb
-500kb to -200kb
-200kb to -100kb
-100kb to -50kb
-50kb to -25kb
-25kb to -10kb
-10kb to -5kb
-5kb to -2kb
-2kb to -1kb
-1kb to 0kb
0kb to 1kb
1kb to 2kb
2kb to 5kb
5kb to 10kb
10kb to 25kb
25kb to 50kb
50kb to 100kb
100kb to 200kb
200kb to 500kb
500kb to 1000kb
Distance to TSS

Fig. 1. Major steps of the algorithm implemented in RAD. The algorithm includes Step 1) identify of gROI associated genes; Step 2) map DEGs within different distance bins; Step 3) calculate observed over expected ratio and perform hypergeometric or binomial test; Step 4) example output bar plot using hypergeometric test as default (Data from Pastor *et al.*, 2018).

2.3 RAD Workflow and Implementation

RAD web application was implemented with Python (version 3.7) programming language, on a Nginx server with Centos 7.06 operating system. The website was developed using AngularJS and Flask framework. The algorithm can be divided into four steps (Fig. 1). The first step is to identify gROI associated genes within user defined gROI extended distance through *awk* and *bedtools* (Quinlan and Hall, 2010). Then gROI extended distance will be split into different distance bins. Up- or down-regulated genes are then mapped into different distance bins. Some pre-defined extended distance bins with varying-length are provided. However, if the user prefers extended distance bins with same length, customized extended distance could be defined. Based on our simulated data, using bins with varying-length or same length are both able to capture distal or proximal regulatory events (Supplemental Fig. 5, 6). To calculate the enrichment of up- or down-regulated genes within different distance bins, observed over expected ratio is calculated as indicated in Fig. 1. Genes that are outside of the gROI extended distance (too far from gROI), or not differentially regulated are excluded. The third step is to perform the hypergeometric or the binomial statistical test to calculate the *p-value*. The formulae for calculating the *p*-value are shown in Fig. 1. Finally, DEGs covered by gROI extended distance, count of up- or down-regulated DEGs in each distance bins and the calculated *p-value* will be reported in text files. The observed over expected ratio will be displayed on the website as bar plot. An example output bar plot comparing naïve human embryonic stem cells (hESC) specific ATAC-seq peaks and naïve hESC up- or down-DEGs is displayed in Fig. 1, suggesting the potential proximal and distal transcriptional promotion role of those naïve specific ATAC-seq peak (Data from Pastor *et al.*, 2018).

2.4 RAD Output

RAD output contains three files: 1) DEGs covered by extended gROI are stored in a text file named *RAD_genename_distance.txt*; 2) The count of up- or down-regulated DEGs in each distance bins, total genes count genome-wide as well as the calculated *p-value* will be reported in a text file named *RAD_genecount_pvalue.txt*; 3) The bar plot of observed over expected ratio in each distance bins can be downloaded as png, SVG or pdf format.

3 Conclusion

We developed a web application RAD to identify gROI associated DEGs and provide a graphic output as well as gROI associated DEGs list for further analysis. Downstream analysis such as gene ontology (GO) enrichment analysis for DEGs in certain distance bins could be performed to help biologists to further infer potential functions of gROI.

Acknowledgements

We thank Dr. Chen, Di, Dr. Wang, Chaochen from Zhejiang University for helpful discussion. W.L., W.P. developed the algorithm. Y.G., Z.X. developed the web application. R.Y. implemented the algorithm in python. Y.G., Z.X., W.L. wrote manuscript and W.L. coordinated research.

Funding

This work is supported by Zhejiang Provincial Natural Science Foundation of China [LQ20C060004 to W.L.]; the Fundamental Research Funds for the Central Universities [K20200099] and Alibaba Cloud.

Conflict of Interest: none declared.

References

Barski,A. et al. (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. Cell, 129, 823–837.

Boyle,A.P. et al. (2008) High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell, 132, 311–322.

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.

Cokus,S.J. et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature, 452, 215–219.

Gallego-Bartolomé,J. et al. (2019) Co-targeting RNA Polymerases IV and V Promotes Efficient De Novo DNA Methylation in Arabidopsis. Cell, 176, 1068-1082.e19.

Harris,C.J. et al. (2018) A DNA methylation reader complex that enhances gene transcription. Science, 362, 1182–1186.

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.

Pastor,W.A. et al. (2018) TFAP2C regulates transcription in human naive pluripotency by opening enhancers. Nat. Cell Biol., 20, 553–564.

Qin,Q. et al. (2020) Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. Genome Biol., 21, 32.

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

Schones,D.E. et al. (2008) Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell, 132, 887–898.

Shlyueva,D. et al. (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet., 15, 272–286.

Wang,S. et al. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat. Protoc., 8, 2502–2515.

Wang,Z. et al. (2018) BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. Bioinformatics, 34, 2867–2869.