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

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

1. Department of Orthopedic of the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310029, China,
2. Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, International Campus, Zhejiang University, 718 East Haizhou Road, Haining 314400, China,
3. Department of Biochemistry, McGill University, Montreal, QC H3G 1Y6, Canada.

*Corresponding author. Email: wanluliu@intl.zju.edu.cn

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 neglected, largely due to the lack of reliable tools to link the distal 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. RAD maps the up- and down-regulated genes associated with any genomic regions of interest (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.

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, a new generation tool incorporates transcriptome and ChIP-seq data to infer direct target genes (Wang, et al., 2013). Combination of multi-level omics data may provide new insights into the regulation of transcription by genome-wide epigenetic landscape.

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 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. 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-Bartolome, 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.

RAD functions

RAD is an open-access, user-friendly web application (Supplemental Figure 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.

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 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. The title and color palette of the output bar plot can be customized according to user's preference.

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 (Figure 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. To calculate the enrichment of up- or down-regulated genes within different distance bins, observed over expected ratio is calculated as indicated in Figure 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 hypergeometric test to calculate the *p-value*. 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 Figure 1, suggesting the potential proximal and distal transcriptional promotion role of those naïve specific ATAC-seq peak (Data from Pastor, et al., 2018).

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.

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 and Dr. Li, J. Jessica from UCLA for helpful discussion. This work is supported by Zhejiang Provincial Natural Science Foundation of China, [LQ20C060004 to W.L.]; and the Fundamental Research Funds for the Central Universities[K20200099]. 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.

Competing interests. The authors declare no competing interest.

References

Barski, A., et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129(4):823-837.

Boyle, A.P., et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132(2):311-322.

Buenrostro, J.D., et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;10(12):1213-1218.

Cokus, S.J., et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008;452(7184):215-219.

Gallego-Bartolome, J., et al. Co-targeting RNA polymerases IV and V promotes efficient de novo DNA methylation in Arabidopsis. Cell 2019;176(5):1068-1082. e1019.

Harris, C.J., et al. A DNA methylation reader complex that enhances gene transcription. Science 2018;362(6419):1182-1186.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.

McLean, C.Y., et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 2010;28(5):495-501.

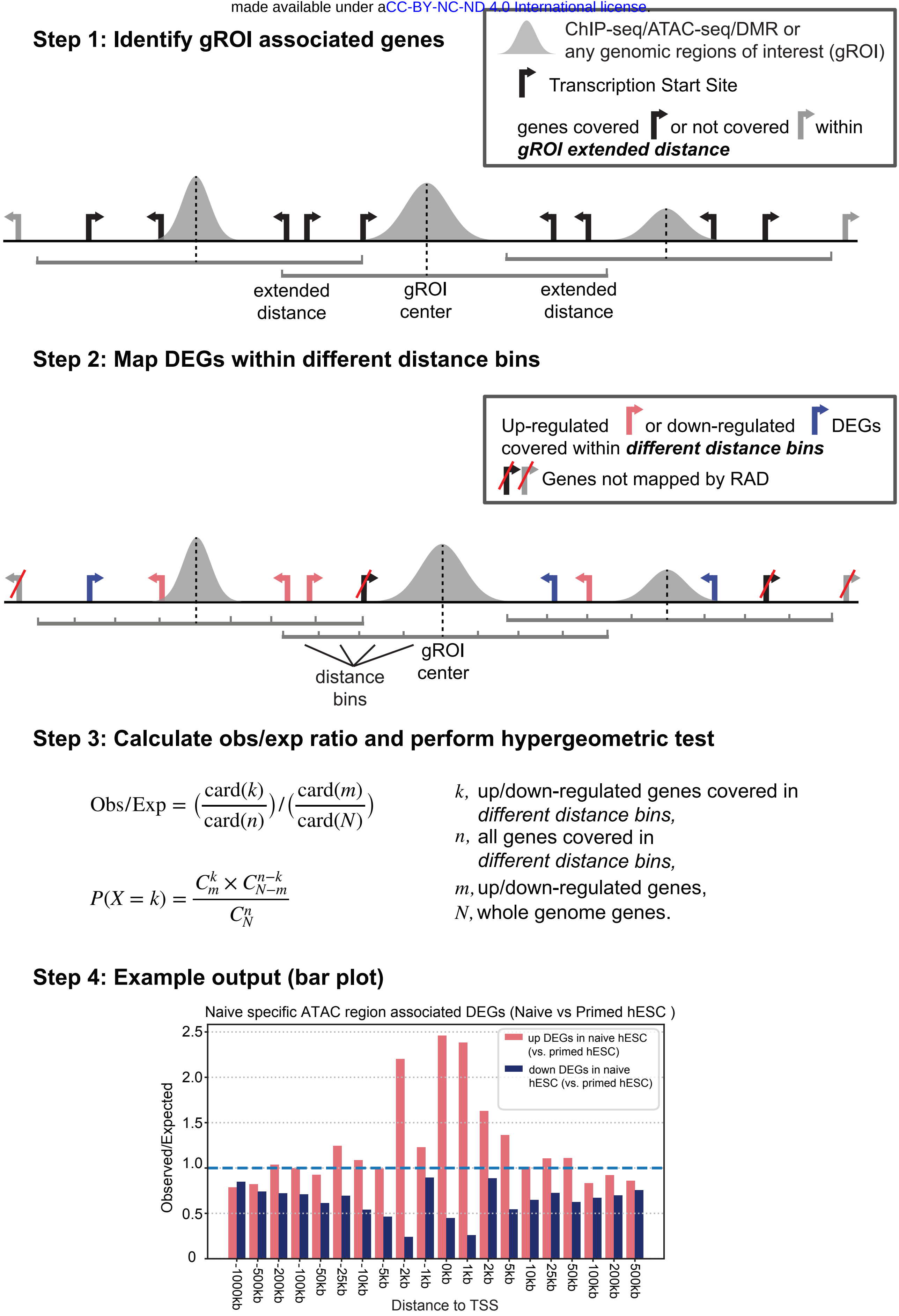
Pastor, W.A., et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. Nature cell biology 2018;20(5):553-564.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841-842.

Schones, D.E., et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132(5):887-898.

Shlyueva, D., Stampfel, G. and Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 2014;15(4):272-286.

Wang, S., et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc 2013;8(12):2502-2515.


Step 1: Identify gROI associated genes
ChIP-seq/ATAC-seq/DMR or any genomic regions of interest (gROI)
Transcription Start Site
genes covered or not covered within gROI extended distance
extended distance
gROI 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
distance bins
gROI center
Step 3: Calculate obs/exp ratio and perform hypergeometric test
k, up/down-regulated genes covered in different distance bins,
n, all genes covered in different distance bins,
m, up/down-regulated genes,
N, whole genome genes.
Step 4: Example output (bar plot)
Naive specific ATAC region associated DEGs (Naive vs Primed hESC)
up DEGs in naive hESC (vs. primed hESC)
down DEGs in naive hESC (vs. primed hESC)
Observed/Expected
0
0.5
1.0
1.5
2.0
2.5
-1000kb
-500kb
-200kb
-100kb
-50kb
-25kb
-10kb
-5kb
-2kb
-1kb
0kb
1kb
2kb
5kb
10kb
25kb
50kb
100kb
200kb
500kb
Distance to TSS

$$\mathrm{Obs/Exp} = \left(\frac{\mathrm{card}(k)}{\mathrm{card}(n)}\right) / \left(\frac{\mathrm{card}(m)}{\mathrm{card}(N)}\right)$$

$$P(X = k) = \frac{C_m^k \times C_{N-m}^{n-k}}{C_N^n}$$

Figure 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 test; Step 4) example output bar plot (Data from *Pastor, et al., 2018*).