epiRomics: a multi-omics R package to identify and visualize enhancers

Alex M. Mawla[1], Mark O. Huising[1,2*]

[1]Department of Neurobiology, Physiology & Behavior, College of Biological Sciences, University of California, Davis, CA 95616, USA

[2]Department of Physiology and Membrane Biology, School of Medicine, University of California, Davis, CA 95616, USA

* Corresponding author:

Mark O. Huising, Ph.D.

Associate Professor

University of California, Davis

Department of Neurobiology, Physiology & Behavior, College of Biological Sciences

Department of Physiology and Membrane Biology, School of Medicine

One Shields Avenue,

Davis, CA, 95616

Ph: (530) 752-4670

Email: mhuising@ucdavis.edu

Abstract

Summary: epiRomics is an R package designed to integrate multi-omics data in order to identify and visualize enhancer regions alongside gene expression and other epigenomic modifications. Regulatory network analysis can be done using combinatory approaches to infer regions of significance such as enhancers, when combining ChIP and histone data. Downstream analysis can identify co-occurrence of these regions of interest with other user-supplied data, such as chromatin availability or gene expression. Finally, this package allows for results to be visualized at high resolution in a stand-alone browser.

Availability and Implementation: epiRomics is released under Artistic-2.0 License. The source code and documents are freely available through Github (https://github.com/Huising-Lab/epiRomics).

Contact: ammawla@ucdavis.edu or mhuising@ucdavis.edu

Supplementary information: Supplementary data, and methods are available online on biorxiv or *Github*.

Introduction

The evaluation of the transcriptional landscape between cell types grants the scientific community a deeper understanding of cellular identity, and helps paint the underlying mechanisms that drive phenotype and function (Capobianco, 2014). Bulk RNA sequencing has been a gold standard in the field, followed more recently with the advent of single-omics approaches (Chen, et al., 2019; Conesa, et al., 2016; Kolodziejczyk, et al., 2015; Kulkarni, et al., 2019). However, gene expression represents only a single aspect of what is a sophisticated and interlaced network of genetic and epigenomic regulators that drive and determine cell identity, with perturbations leading to dysfunction and, sometimes, disease (Karczewski and Snyder, 2018).

Chromatin remodeling is a dynamic process that represents one of the epigenetic layers of cell fate maintenance and identity (Andrey and Mundlos, 2017; Muller and Leutz, 2001). Approaches such as DNase I hypersensitive site sequencing (DNASE-Seq) (Boyle, et al., 2008) and Assay for transposase-accessible chromatin using sequencing (ATAC-Seq) (Buenrostro, et al., 2013; Buenrostro, et al., 2015), are commonly used to compare chromatin accessibility between cell types and states. Chromatin immunoprecipitation Sequencing (ChIP-Seq) is another approach used to assess the epigenomic regulators driven by specific transcription factors acting as either activators or suppressors on the genic region, or at distal-intergenic regions, associated with enhancer activity (Daugherty, et al., 2017; de la Torre-Ubieta, et al., 2018; Neph, et al., 2012; Pastor, et al., 2014; Starks, et al., 2019). Alternatively, ChIP-Seq is used in conjunction with antibodies that pull down specific histone modifications associated with regions of chromatin, whose presence can be used to infer whether the region is active, poised or repressed (Calo and Wysocka, 2013; Mellor, 2005). The interrogation of transcription factor binding is not

enough to infer transcriptional behavior, as whether or not chromatin is accessible between cell types, and whether the region is active, poised, or repressed, demarked by co-occurrences of histone marks, must be considered in order to fully evaluate the biology (Bemer, 2018; Calo and Wysocka, 2013). Lastly, methylation of DNA, quantifiable through whole-genome bisulfite sequencing analysis (BS-Methyl Seq) (Adusumalli, et al., 2015), can help determine whether accessible chromatin that has recruited the correct histone marks is even available for transcription factor binding and recruitment of co-modulators to drive or suppress gene transcription within cell types (Arand, et al., 2012; He, et al., 2011).

While tools exist to compare these different layers in a pairwise manner, a gap exists to integrate multiple omics layers quickly, and easily to generate high resolution visuals in order to derive more biological meaning behind results. We developed a novel 'epigenomics in R', epiRomics, package to solve this issue. We designed epiRomics to accept either browser extensible data (BED) (Kent, et al., 2002) or bigwig (Kent, et al., 2010) files as input for any of the aforementioned types of data. Inclusion of functional annotations, *i.e.* FANTOM (Lizio, et al., 2017), single nucleotide polymorphism (SNP) data from GWAS (Wang, et al., 2010; Zheng, et al., 2015), or Ultra Conserved Non Coding Elements (UCNEs) (Dimitrieva and Bucher, 2012) is also possible – in order to more fully integrate many slices of the genetic and epigenetic pie.

Functions

epiRomics takes in a user-submitted comma-separated values (csv) file containing hard paths to all BED or bigwig formatted files, optional hexadecimal (hex) color code associations for each file and a user-defined label to group each input data set (*e.g.*, ChIP, ATAC, RNA, functional, etc.). The *epiRomics_build_dB* command quickly generates a comprehensive and

easily accessible variable of the class "epiRomics-class" containing a GenomicRanges (*GRanges)* object (Lawrence, et al., 2013) that tracks each of these submitted data, along with all other data related to the species, pulled automatically from the UCSC genome database (Kent, et al., 2002). The epiRomics-class variable can easily be integrated with other packages, and the user can also save these data in a csv format for further manual exploration in excel, or other comparable third-party tools.

Putative enhancers can efficiently be called, and then categorized separately – active, poised, repressed, etc., through the *epiRomics_putative_enhancers* function, which will consider user provided histone data. For example, the histone marks H3K27ac and H3K4me1 are commonly used to demark active enhancer regions (Calo and Wysocka, 2013; Creyghton, et al., 2010; Spicuglia and Vanhille, 2012; Spitz and Furlong, 2012) outputting an epiRomics-class variable for downstream use within the package, or outside. This variable can be used further to identify key enhanceosome regions with evidence of co-binding of multiple user-selected transcription factors by implementing the *epiRomics_putative_enhanceosome* command. These data can also be filtered against functional data annotations, such as methylation calls, FANTOM, SNP regions, or UCNEs, through the use of *epiRomics_putative_enhancers_filtered.* A side function is provided within the package making use of decision trees (Kingsford and Salzberg, 2008) in order to classify which transcription factors were most meaningfully associated with different enhancer types, through the use of *epiRomics_predictors*.

For visualization of these differently classified regions, and integration with bigwig data such as gene expression or chromatin availability between cell types, the tool *epiRomics_track_layer* can be used. This makes use of the package *GViz (Hahne and Ivanek, 2016)* to generate resolution, publication-quality encapsulated postscript (eps) files. Specific calls

for enhancer regions provided by *epiRomics_putative_enhanceosome* can be visualized. Conversely, if users have specific regions or genes of interest they wish to evaluate, they can do so using *epiRomics_region_of_interest.*

These tools were designed to allow biological relevance to be determined from the integrated multi-omics data that is available for a particular tissue or cell type. For example, a common enhancer region may be present between cell types of a common progenitor, with chromatin accessible across all cell types, methylation may block activity in one cell type, but not the other. Drug treatment, or healthy versus diseased comparisons can quickly be made, and the multitude of SNPs amassed via GWAS can be seamlessly connected to narrow in on deleterious variants that may contribute to disease.

Results

epiRomics is developed as an R package to be made available through Bioconductor (Gentleman, et al., 2004), and is available under Artistic-2.0 License. epiRomics is designed to integrate a multitude of -omics data – in either BED or bigwig format – in order to identify regions of regulatory interest, such as enhancers, and provide sophisticated, high quality resolution visuals in EPS format for use in publications. Users with little programming experience can use epiRomics to encode colors for individual tracks, cross-reference diverse types of -omics data – such as ATAC- and RNA- Seq, and produce strong candidate lists for putative enhancers common or unique to cell types. Finally, epiRomics is easy to use, with a full walkthrough with sample data available through its companion vignette.

Funding

This work was supported by grants from the National Institutes of Health (NIDDK110276), the Juvenile Diabetes Research Foundation (2-SRA-2021-1054-M-N) and

American Diabetes Association (1-19-IBS-078) to M.O.H. A.M.M. was supported by the Stephen F. and Bettina A. Sims Immunology Fellowship and the Amazon AWS Machine Learning Research Fellowship.

Conflict of interest

The authors have no conflicts of interests to declare.

References

Adusumalli, S., *et al.* (2015) Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief Bioinform* 16(3):369-379.

Andrey, G. and Mundlos, S. (2017) The three-dimensional genome: regulating gene expression during pluripotency and development. *Development* 144(20):3646-3658.

Arand, J., *et al.* (2012) In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet* 8(6):e1002750.

Bemer, M. (2018) Unraveling the Complex Epigenetic Mechanisms that Regulate Gene Activity. 1675.

Boyle, A.P., *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132(2):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(12):1213-1218.

Buenrostro, J.D., *et al.* (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* 109:21 29 21-21 29 29.

Calo, E. and Wysocka, J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol Cell* 49(5):825-837.

Capobianco, E. (2014) RNA-Seq Data: A Complexity Journey. *Comput Struct Biotechnol J* 11(19):123-130.

Chen, G., Ning, B. and Shi, T. (2019) Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet* 10:317.

Conesa, A., *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol* 17:13.

Creyghton, M.P., *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* 107(50):21931-21936.

Daugherty, A.C., *et al.* (2017) Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res* 27(12):2096-2107.

de la Torre-Ubieta, L., *et al.* (2018) The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* 172(1-2):289-304 e218.

Dimitrieva, S. and Bucher, P. (2012) Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics* 28(18):i395-i401.

Gentleman, R.C., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80.

Hahne, F. and Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. In: Mathé, E. and Davis, S., editors, *Statistical Genomics: Methods and Protocols*. New York, NY: Springer New York; 2016. p. 335-351.

He, X.J., Chen, T. and Zhu, J.K. (2011) Regulation and function of DNA methylation in plants and animals. *Cell Res* 21(3):442-465.

Karczewski, K.J. and Snyder, M.P. (2018) Integrative omics for health and disease. *Nature reviews. Genetics* 19(5):299-310.
Kent, W.J., *et al.* (2002) The human genome browser at UCSC. *Genome Res* 12(6):996-1006.
Kent, W.J., *et al.* (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26(17):2204-2207.
Kingsford, C. and Salzberg, S.L. (2008) What are decision trees? *Nat Biotechnol* 26(9):1011-1013.
Kolodziejczyk, A.A., *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol Cell* 58(4):610-620.
Kulkarni, A., *et al.* (2019) Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr Opin Biotechnol* 58:129-136.
Lawrence, M., *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol* 9(8):e1003118.
Lizio, M., *et al.* (2017) Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. *Nucleic acids research* 45(D1):D737-D743.
Mellor, J. (2005) The dynamics of chromatin remodeling at promoters. *Mol Cell* 19(2):147-157.
Muller, C. and Leutz, A. (2001) Chromatin remodeling in development and differentiation. *Curr Opin Genet Dev* 11(2):167-174.
Neph, S., *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150(6):1274-1286.
Pastor, W.A., *et al.* (2014) MORC1 represses transposable elements in the mouse male germline. *Nat Commun* 5:5795.
Spicuglia, S. and Vanhille, L. (2012) Chromatin signatures of active enhancers. *Nucleus* 3(2):126-131.
Spitz, F. and Furlong, E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* 13(9):613-626.
Starks, R.R., *et al.* (2019) Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin* 12(1):16.
Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 38(16):e164.
Zheng, H.F., *et al.* (2015) Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PloS one* 10(1):e0116487.