BIOINFORMATICS

Genome Analysis

FIREcaller: an R package for detecting frequently interacting regions from Hi-C data

Cheynna Crowley[1], Yuchen Yang[1], Yunjiang Qiu[2,3], Benxia Hu[1], Hyejung Won[1], Bing Ren[2,4,5], Ming Hu[6,*] and Yun Li[1,7,8,*]

[1]Department of Genetics, [7]Biostatistics, [8]Computer Science, University of North Carolina, Chapel Hill, North Carolina, USA. [2]Ludwig Institute for Cancer Research, La Jolla, California, USA. [3]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, California, USA. [4]Department of Cellular and Molecular Medicine, University of California San Diego. La Jolla, California USA. [5]Institute of Genomic Medicine and Moores Cancer Center, University of California San Diego. La Jolla, California, USA. [6]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, Ohio, 44195, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Hi-C experiments have been widely adopted to study chromatin spatial organization, which plays an important role in genome function. Well-established Hi-C readouts include A/B compartments, topologically associating domains (TADs) and chromatin loops. We have recently proposed another readout: frequently interacting regions (FIREs) and discovered them to be informative about tissue-specific gene expression. However, computational tools for detecting FIREs from Hi-C data are still lacking.

Results: In this work, we have developed FIREcaller, a stand-alone, user-friendly R package for detecting FIREs from Hi-C data. FIREcaller takes raw Hi-C contact matrix as input, performs within-sample and cross-sample normalization via HiCNormCis and quantile normalization respectively, and outputs continuous FIRE scores, dichotomous FIREs and super-FIREs.

Availability and implementation: The FIREcaller package is implemented in R, freely available at https://yunliweb.its.unc.edu/FIREcaller.

Contact: yunli@med.unc.edu or hum@ccf.org

INTRODUCTION

Chromatin folding in three-dimensional (3D) space is closely related to genome function (Dekker, et al., 2013). In particular, transcription regulation is orchestrated by a collection of *cis*-regulatory elements, including promoters, enhancers, insulators and silencers. For example, enhancers, which can be hundreds of kilobase (Kb) or even over a megabase (Mb) away from their target gene(s), are brought into close spatial proximity of gene promoter, through looping between the corresponding chromatin segments. Alteration of chromatin spatial organization in the human genome can lead to gene dysregulation and consequently complex diseases including developmental disorders and cancers (Krijger and de Laat, 2016; Li, et al., 2018).

High-throughput chromatin conformation capture (Hi-C) has been widely used to measure genome-wide chromatin spatial organization since first introduced in 2009 (Lieberman-Aiden, et al., 2009; Tjong, et al., 2016; Yardimci, et al., 2019). Analyzing Hi-C data has led to the discovery of structural readouts at a cascade of resolutions, including: (1) A/B compartments (Lieberman-Aiden, et al., 2009), which are multiple megabases (Mb) in size and largely correspond to active or inactive chromatin; (2) Topologically associating domains (TADs) (Dixon, et al., 2012), which are on average approximately 1Mb in size and serve as the basic structural and functional unit of the genome to constrain long-range chromatin interactions largely within TADs; (3) chromatin loops (Rao, et al., 2014) at Kb resolution, which are nested within TADs and anchored by a pair of convergent CTCF motifs; and (4) chromatin 3D contacts (Ay, et al., 2014; Ma, et al., 2015; Xu, et al., 2016; Xu, et al., 2016), again at Kb resolution, which are pairs of chromatin segments brought in physical proximity more often than expected by random chromatin looping or collision. Among these Hi-C readouts, TADs and chromatin loops are largely conserved across cell types (Dixon, et al., 2012; Rao, et al., 2014), while A/B compartments and 3D contacts exhibit rather moderate levels of cell type specificity (Dixon, et al., 2015). In an attempt to identify Hi-C readouts that are better indicative of cell type or tissue-specific chromatin spatial organizations, we discovered thousands of frequently interacting regions (FIREs) by studying a compendium of Hi-C data across 14 human primary tissues and 7 cell types (Schmitt, et al., 2016). We defined FIREs as 40Kb genomic regions which have substantially more frequent contacts with its neighboring regions. FIREs are enriched for tissue-specific enhancers and nearby tissue-specifically expressed genes, suggesting their potential relevance to tissue specific transcription regulatory programs. In our original study, we relied on an in-house pipeline to identify FIREs, limiting the general application of FIRE analysis and the full exploration of tissue-specific chromatin interaction features from Hi-C data. In this work, we developed FIREcaller, a stand-alone, user-friendly R package for detecting FIREs from Hi-C data.

IMPLEMENTATION

The current implementation encompasses three components. First, FIREcaller takes the raw Hi-C contact matrix as input and calculates the total number of local (<200Kb) interactions for each genomic locus (40Kb bin by default). Next, FIREcaller performs within-sample normalization, and cross-sample normalization when the input data contain multiple samples. Here, we use one sample to denote one Hi-C dataset, and multiple samples correspond to multiple Hi-C datasets, for instance, from different tissues or cell types, or from different biological replicates. Specifically, we leverage the HiCNormCis method for within-sample normalization, which adopts a Poisson regression approach to adjust for systematic biases from restriction enzyme cutting, GC content and mappability (Hu, et al., 2012). The residuals from Poisson regression represent the normalized total local interactions and are termed as FIRE scores. When analyzing multiple samples, FIREcaller first applies HiCNormCis to each sample separately for within-sample normalization, and then uses R function "*normalize.quantiles*" in the "*preprocessCore*" package to perform quantile normalization across samples. Subsequently, FIREcaller converts the FIRE scores into Z scores, calculates the one-sided *p*-values based on the standard normal distribution, and defines genomic loci with *p*-value < 0.05 as FIREs. In addition, similar to the observation that many typical enhancers are clustered to form stitched enhancers or super-enhancers (Whyte, et al., 2013), we noticed that many FIREs are close to each other. FIREcaller has a function to merge adjacent FIREs, following the ROSE algorithm (Whyte, et al., 2013), to identify clustered FIREs, termed as super-FIREs. FIREcaller outputs FIRE scores for each genomic locus, a list of FIREs and super-FIREs, for each sample.

APPLICATION EXAMPLES

We used the Hi-C data from human hippocampus tissue released in our previous study (Schmitt, et al., 2016) to showcase the usage of FIREcaller. **Figure 1** shows an illustrative example of a 400Kb super-FIRE (red horizontal bar), which overlaps with two hippocampus super-enhancers (two light blue horizontal bars). Notably, this super-FIRE contains a schizophrenia-associated GWAS SNP rs9960767 (black vertical line) (Stefansson, et al., 2009), and largely overlaps with gene *TCF4* (pink horizontal bar depicted at the top), a gene that plays an important role in neurodevelopment (Forrest, et al., 2014). Since rs9960767 resides within a super-FIRE with highly frequent local chromatin interactions, we hypothesize that chromatin spatial organization may play an important role in gene regulation in this region, elucidating potential mechanisms regarding how rs9960767 affects schizophrenia risk.

In addition, we applied FIREcaller to cortical tissue samples across two developmental epochs, fetal (Won, et al., 2016) and adult (Wang, et al., 2018). We identified 3,925 and 3,926 FIREs from fetal and adult brains, respectively. Among them, 4,815 FIREs are differentially regulated between fetal and adult, where 2,407 FIREs are fetal brain-specific and 2,408 adult brain-specific. The massive changes in FIREs recapitulate recently reported extensive chromatin rewiring during brain development (Wang, et al., 2018), exemplifying the tissue-specific nature of FIREs. We then overlapped FIREs with gene promoters and found that FIREs dynamic across brain developmental stages are closely associated with developmental gene regulation. Specifically, integrative analysis of FIREs and brain gene expression data (Li, et al., 2018) have shown that genes with fetal brain-specific FIREs overlapping their promoters are significantly up-regulated in the fetal brain (p-value<2.2e-16), while genes with adult brain-specific FIREs overlapping their promoters are significantly up-regulated in the adult brain (p-value <2.2e-16) (**Supplementary Information**).

CONCLUSION

In sum, we developed FIREcaller, a user-friendly R package to identify FIREs from Hi-C data. We demonstrated its utilities through applications to two real datasets. We believe that FIREcaller will become a useful tool in studying tissue-specific chromatin spatial organization features.

FUNDING

This research was supported by the National Institute of Health grant R01HL129132 (awarded to YL), U54DK107977 (awarded to BR and MH), and R00-MH113823 (awarded to HW). *Conflict of interest*: none declared.

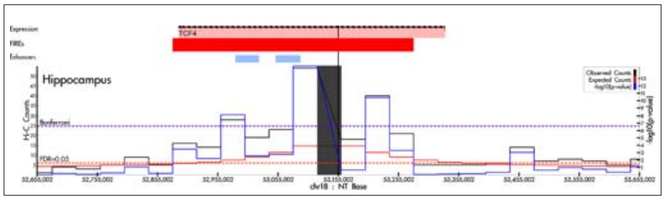

Fig 1. An example of super-FIRE in human hippocampus tissue. Virtual 4C plot of a 1Mb region (chr18:52,665,002-53,665,002) anchored at schizophrenia-associated GWAS SNP rs9960767 (black vertical line), visualized by HUGIn (Martin, et al., 2017). The solid black, red and blue lines represent the observed contact frequency, expected contact frequency, and –log10(p-value) from Fit-Hi-C (Ay, et al., 2014), respectively. The dashed purple and reds line represent significant thresholds corresponding to Bonferroni correction and 5% FDR, respectively. The red horizontal bar depicts the 400Kb super-FIRE region. The two blue horizontal bars mark two hippocampus super-enhancers in the region.

REFERENCES

Ay, F., Bailey, T.L. and Noble, W.S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research* 2014;24(6):999-1011.

Dekker, J., Marti-Renom, M.A. and Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics* 2013;14(6):390-403.

Dixon, J.R., *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;518(7539):331-336.

Dixon, J.R., *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485(7398):376-380.

Forrest, M.P., *et al.* The emerging roles of TCF4 in disease and development. *Trends in molecular medicine* 2014;20(6):322-331.

Hu, M., *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics (Oxford, England)* 2012;28(23):3131-3133.

Krijger, P.H. and de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* 2016;17(12):771-782.

Li, M., *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science (New York, N.Y.)* 2018;362(6420).

Li, Y., Hu, M. and Shen, Y. Gene regulation in the 3D genome. *Human molecular genetics* 2018;27(R2):R228-r233.

Lieberman-Aiden, E., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326(5950):289-293.

Ma, W., *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nature methods* 2015;12(1):71-78.

Martin, J.S., *et al.* HUGIn: Hi-C Unifying Genomic Interrogator. *Bioinformatics* 2017.

Rao, Suhas S.P., *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 2014;159(7):1665-1680.

Schmitt, A.D., *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell reports* 2016;17(8):2042-2059.

Schmitt, A.D., Hu, M. and Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* 2016;17(12):743-755.

Stefansson, H., *et al.* Common variants conferring risk of schizophrenia. *Nature* 2009;460(7256):744-747.

Tjong, H., *et al.* Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences of the United States of America* 2016;113(12):E1663-1672.

Wang, D., *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science (New York, N.Y.)* 2018;362(6420).

Whyte, W.A., *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;153(2):307-319.

Won, H., *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 2016;538(7626):523-527.

Xu, Z., *et al.* A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics* 2016;32(5):650-656.

Xu, Z., *et al.* FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics (Oxford, England)* 2016.

Yardimci, G.G., *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome biology* 2019;20(1):57.