diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data

Citation	Lareau, Caleb A., and Martin J Aryee. 2017. “diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data.” Bioinformatics 34 (4): 672-674. doi:10.1093/bioinformatics/btx623. http://dx.doi.org/10.1093/bioinformatics/btx623.
Published Version	doi:10.1093/bioinformatics/btx623
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:35982229
Terms of Use	This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Bioinformatics, 34(4), 2018, 672–674
doi: 10.1093/bioinformatics/btx623
Advance Access Publication Date: 29 September 2017
Applications Note

Genome analysis

diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data

Caleb A. Lareau[1,2,3] **and Martin J. Aryee**[1,2,3,4,*]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [3]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA and [4]Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, MA 02129, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on March 5, 2017; revised on September 23, 2017; editorial decision on September 26, 2017; accepted on September 27, 2017

Abstract

Summary: The 3D architecture of DNA within the nucleus is a key determinant of interactions between genes, regulatory elements, and transcriptional machinery. As a result, differences in DNA looping structure are associated with variation in gene expression and cell state. To systematically assess changes in DNA looping architecture between samples, we introduce diffloop, an R/Bioconductor package that provides a suite of functions for the quality control, statistical testing, annotation, and visualization of DNA loops. We demonstrate this functionality by detecting differences between ENCODE ChIA-PET samples and relate looping to variability in epigenetic state.

Availability and implementation: Diffloop is implemented as an R/Bioconductor package available at https://bioconductor.org/packages/release/bioc/html/diffloop.html

Contact: aryee.martin@mgh.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The organization of DNA within the nucleus into hierarchical 3D structures plays a key role in regulating gene expression by determining the accessibility of genes to the transcriptional machinery as well as the proximity of genes to their distal regulatory elements. Differences in 3D architecture, such as the presence or absence of loops between enhancers and their target genes, are associated with transcriptional variation in both normal and disease states. Intriguingly, several recent studies have implicated alterations in genome topology with a diverse set of diseases (Flavahan *et al.*, 2016; Hnisz *et al.*, 2016).

Experimental techniques that couple chromatin conformation capture (3C; Dekker *et al.*, 2002) with high-throughput sequencing have made the genome-wide identification of 3D interactions feasible. For example, the high-throughput chromosome conformation capture (Hi-C) assay, which can theoretically yield a near-complete map of chromatin interactions, has been used to map the 3D genome at a 1-kb resolution (Rao *et al.*, 2014). As Hi-C requires billions of reads to achieve this resolution, methods such as Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) (Tang *et al.*, 2015), HiChIP (Mumbach *et al.*, 2016), or promoter-capture Hi-C (Mifsud *et al.*, 2015) use capture techniques to enrich for specific subsets of interactions such as structural loops or enhancer–promoter interactions, allowing lower sequencing depths. These assays when coupled with appropriate preprocessing tools (Cairns *et al.*, 2016; Phanstiel *et al.*, 2015) produce interaction frequencies between pairs of genomic loci.

In order to fully explore the role that 3D genome organization plays in determining normal and pathogenic cell states, statistical tools are needed to identify differences in DNA loops in a similar manner to which differential expression analysis is applied to transcriptional data. Additionally, the systematic integration of biological prior knowledge, such as the location of active enhancer regions, into topology analyses can provide annotation and insight into the regulatory role of a loop. To address these needs, we have developed diffloop, an R/Bioconductor package that implements statistical testing for differential DNA looping between samples from ChIA-PET, HiChIP, and related 3C assays. While existing tools such as DiffBind (Stark and Brown, 2015) and diffHiC (Lun and Smyth, 2015) provide functionality for identifying differential features from ChIP-seq and Hi-C experiments, diffloop provides a suite of functions tailored to chromatin loop data. Here, we briefly demonstrate some of the utility of the diffloop package by comparing chromatin interactions inferred by ChIA-PET replicates between the MCF7 and K562 cell lines (ENCODE Project Consortium, 2012).

2 Materials and methods

Following import of raw loop read counts diffloop combines counts across samples and assigns statistical significance to each putative loop using the method developed by Phanstiel *et al.* (2015). The calculation uses a model that takes into account the signal intensity at each of the anchors and the expected background chromatin interaction frequency for the given anchor separation distance.

To identify differential loops, diffloop by default applies the statistical test in edgeR (Robinson *et al.*, 2010) where counts are modeled using the negative binomial distribution and an empirical Bayes procedure is used to moderate the degree of overdispersion. The counts matrix, rather than representing reads mapped to genes or transcripts as is typical in a differential expression analysis, instead contains PETs (i.e. paired-end reads). A scaling size factor is calculated for each sample to account for variations in read depth.

The diffloop provides functionality for annotation of loops and loop anchors and to facilitate interpretation of the functional relevance of the significantly differential loops identified. A typical use case involves annotating anchors with chromatin mark data and, promoter overlap and gene expression levels. Loops may be categorized based on these annotations into categories such as CTCF-CTCF or enhancer-promoter and can be visualized with ease using novel diffloop functions.

3 Results

POL2 ChIA-PET data from two MCF7 and two K562 samples were individually preprocessed from raw reads to loops using the Mango preprocessing pipeline (Phanstiel *et al.*, 2015). Across the union of the four samples considered for our analyses, we observed a total of 87 456 anchor pairs involving 24 576 autosomal loci (anchors). After filtering out loci biased by copy number variation, loops only detected in a single sample, and anchor pairs with interaction frequencies within the range of the background signal (Phanstiel *et al.*, 2015), we retained 9320 loops for differential testing (see Supplementary Material).

At an FDR of 1%, we identified 2633 differential loops between the cell lines, including 1974 loops that were annotated as enhancer-promoter loops. Supplementary Table S3 summarizes the top five differential enhancer–promoter loops specific to each cell line. Figure 1 provides a sample visualization of one of these differential loop regions where mutliple loops near the *MTHFR* gene were more prevalent in the K562 cell line than the MCF7 cell line.

To characterize the structural differences globally, we identified nine pathways enriched for genes involved in differential enhancer–promoter looping (see Supplementary Material). Genes related to estrogen response such as GREB1 and XBP1, for example, are linked by several strong loops to unique enhancers in the MCF-7 breast cancer cell line. Conversely, targets associated with c-MYC transcription factor, which plays a well-documented role in leukemia and hematopoiesis were enriched in K562. These results suggest that differential topology analyses can systematically uncover known and novel regulatory loops related to disease and other phenotypes of interest. Thus, we suggest that cell type-specific chromatin loops such as those identified here by diffloop can serve as a valuable epigenetic feature for characterizing cell identity.

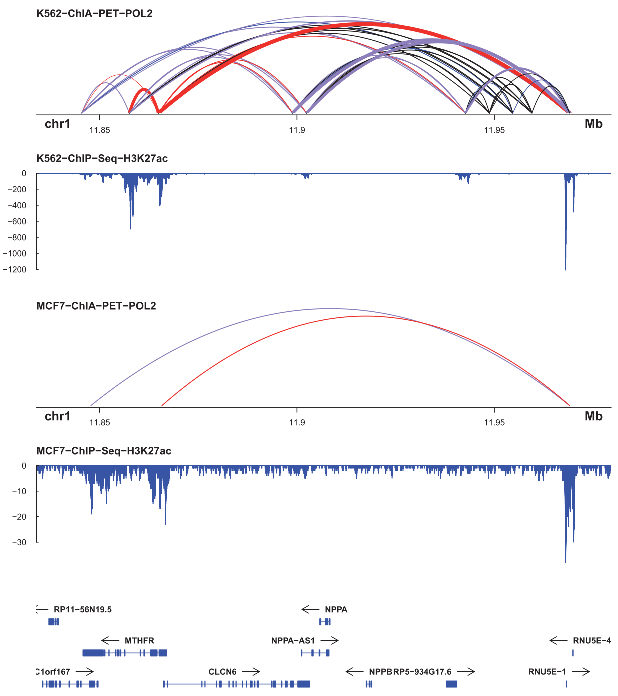

Fig. 1. Sample visualizations of differential looping. The figure shows the combined POL2 ChIA-PET replicates for the K562 and MCF-7 cell lines as well as the cell type-specific H3K27ac ChIP-Seq track. Line widths are indicative of the number of PETs supporting a loop while colors represent biological annotation (red, enhancer-promoter; purple, enhancer-enhancer; black: no special annotation). The region highlighted contains the MTHFR gene, which has previously been implicated as an up-regulated feature of human leukemias such as the K562 cell line

Acknowledgements

We thank D. Day, D. Hnisz and members of the R. Young Lab for their useful discussion and insight.

Funding

This work has been supported by the NSF (Graduate Fellowship no. DGE1144152) and an MGH Department of Pathology Startup Fund.

Conflict of Interest: none declared.

References

Cairns,J. *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.*, **17**, 127.

Dekker,J. *et al.* (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Flavahan,W.A. *et al.* (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110–114.

Hnisz,D. *et al*. (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.

Lun,A.T.L. and Smyth,G.K. (2015) diffHiC: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, **16**, 258.

Mifsud,B. *et al*. (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.

Mumbach,M.R. *et al*. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.

Phanstiel,D.H. *et al*. (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**, 3092–3098.

Rao,S.S.P. *et al*. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Stark,R. and Brown,G. (2015) DiffBind: Differential Binding Analysis of ChIP-Seq peak data. R package version 1.10.2.2014. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.

Tang,Z. *et al*. (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.