epiTAD: a web application for visualizing high throughput chromosome conformation capture data in the context of genetic epidemiology

Jordan H. Creed, Alvaro N. Monteiro, and Travis A. Gerke*

Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA

January 5, 2018

Abstract

The increasing availability of public data resources coupled with advancements in genomic technology has created vast opportunities for researchers to examine the genome on a large and complex scale. To meet the need for integrative genome wide exploration, we present epiTAD. This web-based tool enables researchers to compare genomic structures and annotations across multiple databases and platforms in an interactive manner in order to facilitate *in silico* discovery. epiTAD can be accessed at https://gerkelab.shinyapps.io/epiTAD/.

1 Introduction

Hi-C is a high-throughput sequencing technique that enables estimation of 3-D genomic structure. Uptake of the Hi-C assay has driven recent genome-wide mapping of topologically associated domains (TADs) at varying resolutions. These units of genomic organization are actively being studied for their role in biological function. Various web tools and stand-alone programs enable the visualization of a researcher's Hi-C data, or utilize competing algorithms to call TADs; however, they do not adequately show the relationship between these measures and other units of genomic organization. Examples of useful complementary information include linkage disequilibrium, eQTLs, and functional annotation. Widespread integration of this data into research pipelines may be limited by a dearth of user-friendly software applications that link TAD data to other genomic databases. To meet this need, we present epiTAD, a user-friendly web application.

The 3D structure of the genome is known to impact its functional aspects. An intrinsic genome organization hierarchy exists in which double stranded DNA is wrapped around histone octamers forming nucleosomes, which are then folded into chromatin. The chromatin is arranged into loops and is in turn condensed into chromosomes. Chromatin loops serve to bring promotors and enhancers, that linearly would be far apart, into close contact[1]. TADs are regions of chromatin that interact extensively, and can be visualized as triangles on a heatmap of Hi-C data. These domains are largely conserved across both cell lines and species[2].

The Hi-C Browser[3] and Juicebox[4] are other free web tools for visualizing Hi-C data. These tools and epiTAD produce heatmaps, but differ in additional annotations and data sources provided. Juicebox is strongest when used as either part of Aiden Lab's Hi-C pipeline or as a stand alone program for visualizing published experimental data in a single image, while Hi-C Browser contains a unique track for DNase I hypersensitive sites (DHSs). Both tools also allow users to either select specific cell lines or use their own data. While these tools are highly valuable, they may be of limited use to researchers with an interest in variant level genome structure or population genetics, as neither tool delivers such annotations. Furthermore, high-resolution figures are not downloadable. epiTAD serves to fill this role for researchers.

*Corresponding author: travis.gerke@moffitt.org

2 Materials and Methods

2.1 Data Collection

The Hi-C reads provided in epiTAD are from human fibroblast IMR90 cell profiles as described by Dixon et al[2]. TAD boundaries were also taken from the Dixon experiments and were determined on the basis of a "bi-directionality index" of chromatin interactions[2]. Linkage disequilibrium (LD) measures are queried from HaploReg, with data originating from the 1000 Genomes project[5]. eQTLs are obtained through GTEx Version 7, which determines eQTLs from 635 samples over 53 tissues[6]. Coordinates and HGNC symbols for genes are taken from ENSEMBL, while data from various sources are scraped from the annotation aggregator Oncotator. All data is queried in real time from publicly available resources.

2.2 Implementation

The web application is hosted on R Shiny servers through Shiny IO. The user interface is built on a combination of R and Java, while all data querying, table production and visualization is processed through R and performed in real time. Full source code for the application is available at https://github.com/tgerke/epiTAD.

3 Access and Display

epiTAD is freely available online from https://gerkelab.shinyapps.io/epiTAD/, and is compatible with all major web browsers (Chrome, Internet Explorer and Safari). No login or user information is required for use. The display consists of two pages: the primary page, which contains all tables, visuals, and user inputs; and an informational page, accessible from a tab on the left-hand side of the page. The tool is broken down into five panels: inputs, output options, variant annotations, gene annotations, and visualization.

At a minimum, users need to input at least one SNP, in the form of a dbSNP rs id (e.g. rs10486567). Multiple SNPs can be uploaded and queried at once either as a comma separated list or as a text file, however they must all reside on the same chromosome.

Variant annotations from the HaploReg and RegulomeDB databases are available under separate tabs. The amount of information shown can be customized by the user from the "Select Output" panel under the tab of the same name. The TADs tab informs the user if the SNP(s) are located within a known TAD and if so, provides the coordinates and also contains a link to the Hi-C Browser for additional visualization options.

A query region is created that contains all SNPs within the maximum of the LD or TAD boundaries. If there are no SNPs above the selected LD threshold and the variants of interest are not located within a TAD, then a region of 53,500 bp is added to either side of the SNP to create the query region. This query region is then used for all gene level annotations and for the initial coordinates for the visualization. The gene names and coordinates can be downloaded from the site as a CSV file. Oncotator queries are broken down by source and can pull information from the Cancer Gene Census, HUGO Gene Nomenclature Committee, and UniProt. If eQTLs are available from GTEx the user can select tissue(s) of interest to subset results.

Live links to additional resources and tools, including ClinVar, the UCSC Genome Browser, Juicebox, Hi-C Browser, and GTEx, are made available to the user for the region/SNPs.

A figure is automatically rendered that spans the query region and contains four tracks. The first track displays the Hi-C contact matrix, the second contains gene coordinates and names within the region, the third shows the query SNP(s) (color) and SNPs in LD (grey), and the final track shows the TAD boundaries. The plot can be updated with user-defined coordinates, as long as the coordinates are at least 20,000 bp apart (Hi-C is binned in 10,000 bp), and there is a button to return to the original coordinates. The user also has the option to choose from 5 color schemes. The figure can also be downloaded as a high-resolution PDF.

Figure 1 demonstrates the output for rs10486567. This SNP was previously found to significantly associate with *HIBADH* expression in prostate tumor tissue and with *TAX1BP1* expression in normal prostate tissue[7]. This finding may be, in part, explained by observing that both of these genes as well as the query SNP are located within the same TAD, with the Hi-C data showing a large amount of contact across this area.

4 Discussion

In light of the ongoing and rapid growth in publicly available genomic data, efficient systems to compare genomic databases become increasingly vital to researchers. epiTAD provides a single application to integrate genomic annotations and measurements across multiple public databases covering a uniform region. The application allows researchers to access and plot large amounts of data related to major genome organization structures without programming knowledge. The resulting web app may prove a broadly useful component of *in silico* functional genomics discovery.

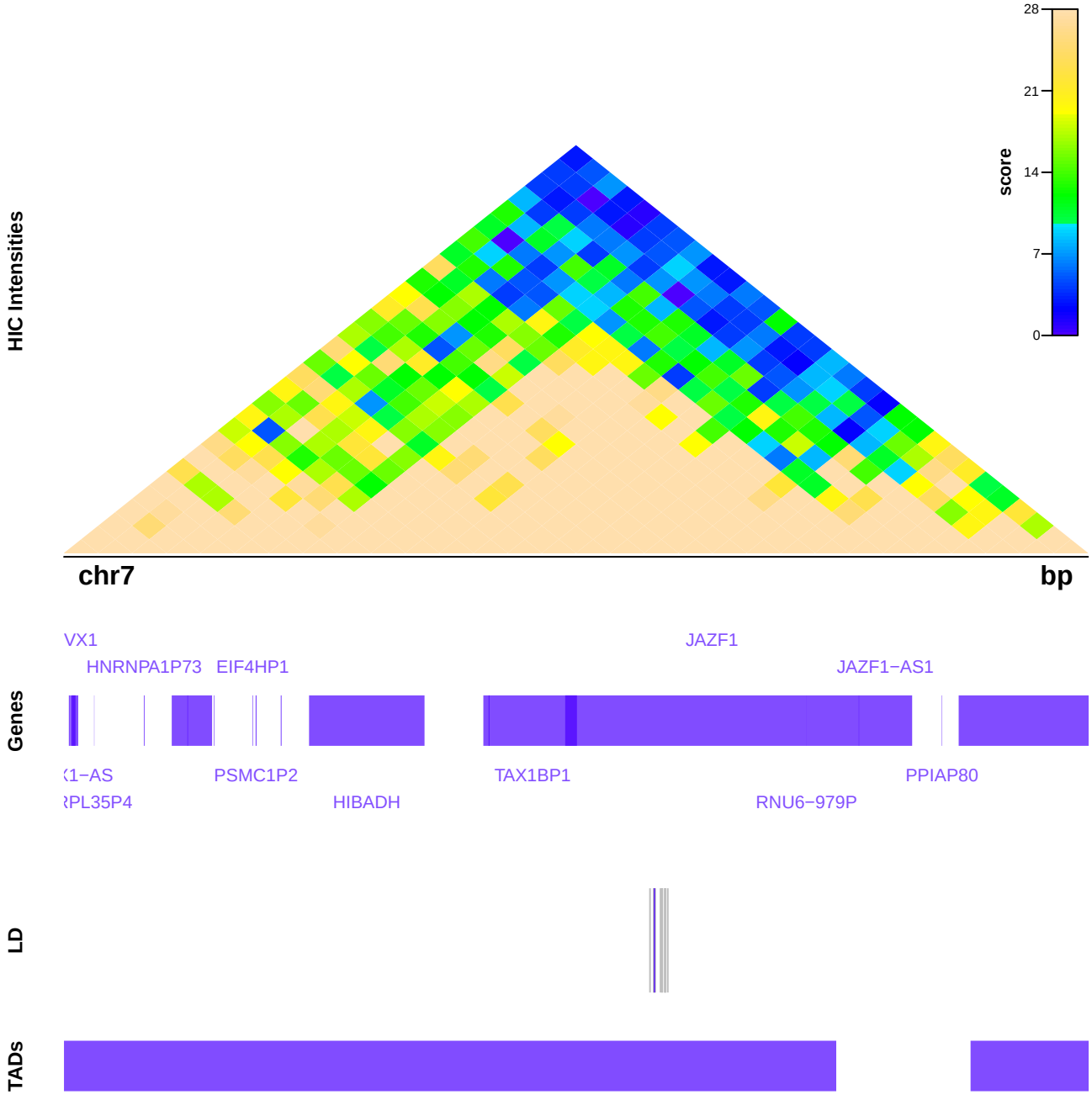

Figure 1: An example of the epiTAD figure output for rs10486567.

References

[1] B. V. Iyer, M. Kenward, and G. Arya. Hierarchies in eukaryotic genome organization: Insights from polymer theory and simulations. *BMC Biophys*, 4:8, Apr 2011.

[2] Dixon JR, Selvaraj S, Yue F, and et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, Apr 2012.

[3] Yanli Wang, Bo Zhang, Lijun Zhang, Lin An, Jie Xu, Daofeng Li, Mayank NK Choudhary, Yun Li, Ming Hu, Ross Hardison, Ting Wang, and Feng Yue. The 3d genome browser: a web-based browser for visualizing 3d genome organization and long-range chromatin interactions. *bioRxiv*, 2017.

[4] N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*, 3(1):99–101, Jul 2016.

[5] Ward LD and Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40:D930–D934, Jan 2012.

[6] GTEx Consortium. *GTEx Portal*, 2017 (accessed January 5, 2018). https://www.gtexportal.org/home/.

[7] K. L. Penney, J. A. Sinnott, S. Tyekucheva, T. Gerke, I. M. Shui, P. Kraft, H. D. Sesso, M. L. Freedman, M. Loda, L. A. Mucci, and M. J. Stampfer. Association of prostate cancer risk variants with gene expression in normal and tumor tissue. *Cancer Epidemiol. Biomarkers Prev.*, 24(1):255–260, Jan 2015.