Genome analysis

Sushi.R: Flexible, quantitative, and integrative genomic visualizations for publication-quality multi-panel figures

Douglas H Phanstiel[1], Alan P Boyle[1], Carlos L Araya[1], and Michael P Snyder[1,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

Associate Editor: Dr. Inanc Birol

ABSTRACT

Motivation: Interpretation and communication of genomic data require flexible and quantitative tools to analyze and visualize diverse data types, yet a comprehensive tool to display all common genomic data types in publication quality figures does not exist to date. To address this shortcoming, we present Sushi.R, an R/Bioconductor package that allows flexible integration of genomic visualizations into highly-customizable, publication-ready, multi-panel figures from common genomic data formats including BED, bedGraph, and BEDPE. Sushi.R is open source and made publicly available through GitHub (https://github.com/dphansti/Sushi) and Bioconductor (http://bioconductor.org/packages/release/bioc/html/Sushi.html).

1 INTRODUCTION

Genomic science is a rich, data-intensive field in which diverse data types are combined to uncover and explore characteristics of sequence elements on a large scale. However, despite a growing set of mature, standard visualization techniques and file formats, no comprehensive tools exist to facilitate multi-panel visualization across a broad range of standard genomic data types. To address this deficiency we developed Sushi.R, a flexible R library that leverages standard visualization techniques and file formats to produce highly customizable, publication-quality figures of genomic data within the widespread analysis environment, R(R Core Team, 2013).

2 METHODS

Sushi.R is written exclusively in the R software environment. The Sushi.R package includes 13 example data sets and a vignette detailing the usage of each (Sanyal *et al.*, 2012; Li *et al.*, 2012; Consortium, 2012; Neph *et al.*, 2012; Ehret *et al.*, 2011; Dixon *et al.*, 2012; Rhee and Pugh, 2011). Data sets that were mapped to hg19 were converted to hg18 using the liftOver tool. Sushi is compatible with all organisms and genome builds. Large data sets were filtered to include only regions shown in figure 1. ChIA-PET interactions were additionally filtered to remove interactions between regions <= 1000 base pairs apart. To facilitate use, Sushi.R is open source and is distributed through both Bioconductor for one-step installation and GitHub for version control, issue management, and third party development(Gentleman *et al.*, 2004).

3 FEATURES

Quantitative and qualitative genomic information can typically be broken down into 3 data types: features, signals, and interactions. Sushi.R provides flexible methods to plot each data type, allowing users to represent virtually any type of genomic data in an aesthetically pleasing, coherent, and integrative fashion. A Sushi plot made entirely within R (without any modifications in image editing software) displaying multiple data types is shown in **Figure 1**. The code and data to make **Figure 1** are included as part of the Sushi.R package.

Feature data describes genomic regions characterized by a unique combination of chromosome, start, and stop coordinates. Often stored in BED format, feature data can be used to represent sites of transcription factor binding, gene structures, transcript structures, sequence read alignments, GWAS hits and data from an array of other sources. The Sushi functions *plotBed*, *plotGenes*, and *plotManhattan* facilitate the visualization of feature data in a host of different formats ranging from heatmaps of feature density to feature pile-ups (**Figure 1a, g-k, n**).

Signal data representing quantitative values across genomic coordinates are commonly stored in bedGraph format and can be used to represent diverse forms of data including sequence conservation, transcription, transcription factor binding, chromatin accessibility, and nascent transcription rates, among others. The Sushi function *plotBedgraph* provides flexible methods to plot, overlay, and compare signal track data to appropriately represent data from each one of these disparate sources (**Figure 1e, f, l, m**).

Finally, interaction data can be used to describe interactions between distal genomic elements in both a qualitative or quantitative fashion. Interaction data describing for example three-dimensional chromatin structure are commonly stored in BEDPE format or in interaction matrices. Sushi functions *plotHiC* and *plotBedpe* are used to plot interactions data as either trapezoidal heatmaps, arched lines, or box and line structures, and support quantitative mapping of interaction signals on *y*-axis values, color scales, and line-widths (**Figure 1b-d**).

Sushi plots can easily be combined and augmented *via* a number of annotation functions including *zoomsregion*, *zoombox*, *maptocolor*, and *addlegend* allowing customizable scaling of colors, line types, and line widths for flexible, quantitative presentation. Zoom inset features facilitate visualization at multiple scales and diverse genomic contexts. Images can be written to all formats supported by R including EPS, PDF, and PNG.

4 DISCUSSION

The rapid proliferation and complexity of genomics experiments –fueled by high-throughput sequencing– has concomitantly driven demand for analysis and visualization tools that facilitate interpretation and communication of rich and diverse genomic data types. Sushi fills a critical void among currently available visualization tools by providing a means to easily pro-
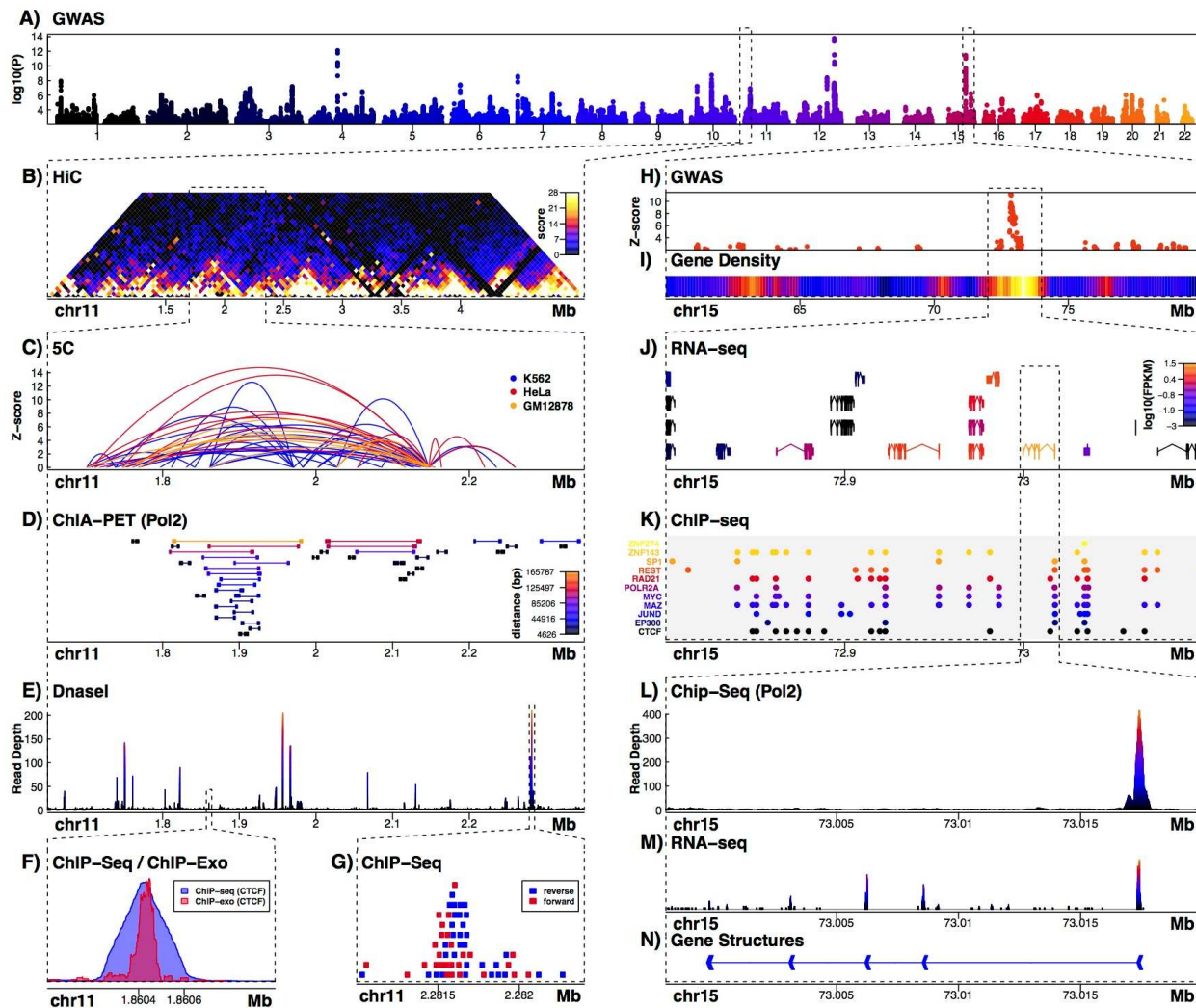
*To whom correspondence should be addressed.

Fig. 1. Multi-panel Sushi plot made without modification by external image editing software. The Sushi functions used to create the plot include (**a**) *plotManhattan*, (**b**) *plotHic*, (**c**) *plotBedpe*, (**d**) *plotBedpe*, (**e**) *plotBedgraph*, (**f**) *plotBedgraph*, (**g**) *plotBed*, (**h**) *plotManhattan*, (**i**) *plotBed*, (**j**) *plot-Genes*, (**k**) *plotBed*, (**l**) *plotBedgraph*, (**m**) *plotBedgraph*, and (**n**) *plotGenes*. The code and data to make this figure are included as part of the Sushi.R package.

duce sophisticated, customizable, genomic visualizations. Sushi.R will be of great use to the genomic community as it accelerates our ability to uncover, document, and communicate important scientific findings derived from increasingly abundant, and complex, genomic data.

REFERENCES

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

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

Ehret,G.B. *et al.* (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**, 103–109.

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

Li,G. *et al.* (2012) Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*, **148**, 84–98.

Neph,S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.

R Core Team (2013) R: A Language and Environment for Statistical Computing

Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.

Sanyal,A. *et al.* (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.