BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–2

DNAism: Exploring genomic datasets on the web with Horizon Charts.

David Rio Deiros [1,*], Richard A. Gibbs [1,2] and Jeffrey Rogers [1,2]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

Received on XXXXXXX ; revised on September YYYYYYY; accepted on ZZZZZZZZ

Associate Editor: Name Here.

ABSTRACT

Summary:

Computational biologists daily face the need to explore massive amounts of genomic data. New visualization techniques help researchers navigate and understand these big data. Horizon Charts are a relatively new visualization method that, under the right circumstances, maximizes data density without losing graphical perception. This visualization technique has been successfully applied to understand multi-metric time series data. We have adapted an existing Javascript library (Cubism[1]) that implements Horizon Charts for the time series domain so that it works effectively with genomic datasets. We call this new library DNAism.

Availability and implementation: Source code, documentation and usage examples can be found at: http://dnaism.github.com.

Contact: deiros@bcm.edu

1 INTRODUCTION

Sharing and communicating about large and intricate datasets produced by Next-Gen sequencing can be a challenging task. Visual channels are an effective way to explore data. However, the accelerating increase in data quantity is pushing the limits of current approaches for representing these datasets visually without sacrificing accuracy or graphical perception. Data volume is increasing vertically as throughput per sample increases, and horizontally as studies involving large numbers of subjects become more feasible. Thus, more effective visualization techniques are needed to understand the most challenging Next-Gen sequencing datasets.

Horizon Charts (Few, 2008) have proved to be an effective (Heer et al., 2009) visualization approach when working with multi-metric time series encoded data. On the other hand, BED[2] files are the gold standard for capturing genomic metrics in the Next-Gen sequencing domain. In time series, metrics are monitored over time, however, BED files use genomic coordinates. We have adapted a time series Javascript library to the genomic domain. We call our new library DNAism.

2 IMPLEMENTATION

Contrary to time series data, in genomic datasets, the variable under study is associated with chromosomal coordinates instead of timestamps. We have modified an existing time series data visualization library (based on D3 (Bostock et al., 2011)) called Cubism to support genome coordinate data. This makes DNAism a flexible and effective tool to explore multi-sample genomic datasets using Horizon charts.

To visualize genomic datasets, we have modified most of the components of the original Cubism library. The two major ones being 'context' and 'source'. The 'context' component performs several functions, most importantly, it defines the region of the genome we want to explore. This component also specifies, in pixels, how much vertical space we have available for the visualization. The 'source' component parses the genomic raw data and generates the data points necessary for visualization. Our library provides two sources: 'bedfile' and 'bedserver'. Once the sources are created we can use the metric component to instantiate metrics pointing to specific samples. Finally, the horizon component encapsulates the functionality necessary to create the visual elements. Figure 1 illustrates visualization of sample data.

One of the crucial features of DNAism is the ability to efficiently parse and load the genomic data for visualization. We have provided two alternatives via the bedfile and the bedserver sources. A bedfile is a simple solution that loads all the genomic data in memory and returns the relevant data when queried. However, this approach is not adequate for larger datasets, especially those involving multi-sample data. To handle such cases, the bedserver source can be used. A bedserver is a dedicated server that implements a RESTful API interface. The client's code running in the browser can send queries to this server to obtain the data of interest. The server uses pre-indexed (Li, 2011) data to speed up random access and returns only the necessary information for the visualization back to the client. Hence, this approach becomes much more scalable even with large sized genomic data sets. We have implemented bedserver as a Python package.

Our source code has a decoupled interface that facilitates the extension of this library to new data sources. DNAism is data agnostic. As a result, users can create new sources to capture their specific backend peculiarities.

*to whom correspondence should be addressed

[1] https://github.com/square/cubism.

[2] http://genome.ucsc.edu/FAQ/FAQformat.html

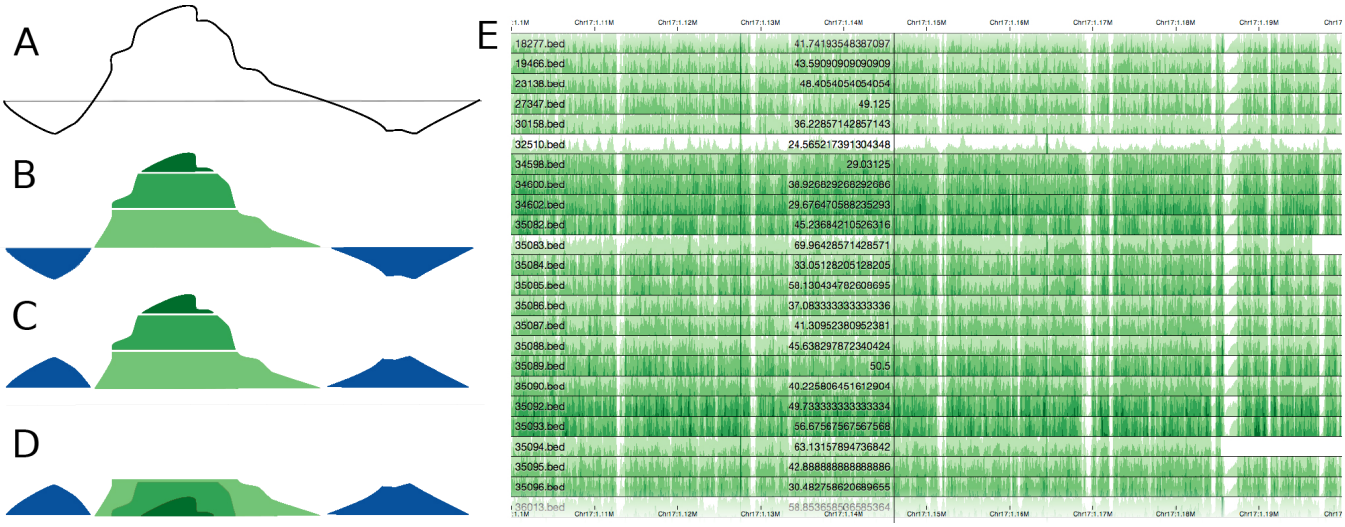

Fig. 1. Horizon Charts emerge from applying a set of changes to traditional line charts **(A)**. We start by coloring the underlying area of the line chart, using different hues for positive and negative values. Next, we divide the graph in bands and apply a gradient of color that increases along with the quantitative value of the variable we are investigating **(B)**. In the next step, negative values are flipped over the baseline **(C)**, effectively reducing the vertical space by two fold. In a final step, bands are collapsed making all of them start at the baseline and providing another level of space reduction **(D)**. We used this technique to rapidly identify problematic samples when performing quality control on large scale sequencing. You can see the read depth across whole genome sequences from 24 rhesus macaque samples (30x coverage) for genomic region Chr17:1.1M-1.2M **(E)**. There are regions consistently underrepresented across all the samples and sample 32510 has low coverage across the whole genomic region. Note that the variable we are exploring in this example, read depth, does not contain negative values. Therefore, only green hues appear in **(E)**.

3 RESULTS

We introduce the genomics community to a powerful visualization technique previously used in the time series data domain. This method facilitates identification of abnormal patterns across multi-sample datasets. In addition, this approach helps to explore and visualize high density datasets more effectively, thereby, helping the researchers to understand the data easily.

Our library keeps the effective and elegant interface of the original, while allowing users to leverage its power for genomic data. By providing a library, we maintain flexibility regarding how to use these resources. Users can build full applications or use the library within their existing ones.

The companion lightweight server will facilitate the exploration of large genomic datasets without affecting user experience by using indexed datasets. Alternatively, users can create their own data sources to reflect the details of their own environments.

Funding: NIH(NCRR) R24OD011173 and NIH grant 2U54HG003273.

REFERENCES

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309.

Few, S. (2008). Time on the horizon. *Visual Business Intelligence Newsletter*.

Heer, J., Kong, N., and Agrawala, M. (2009). Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *CHI*, pages 1303–1312.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, 27(5):718–719.