BioJS-HGV Viewer: Genetic Variation Visualizer

Saket Choudhary*[1], Leyla Garcia[2], and Andrew Nightingale[2]

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, USA
[2]European Bioinformatics Institute EMBL-EBI, Cambridge, England

Abstract

Studying the pattern of genetic variants can help us understand a individuals within a population, identify key 'driver variants' that affect disease states and evolution. Catalogs of genetic variants are vast but lack an interactive exploratory interface.

Here we present BioJS-HGV Viewer, a BioJavascript component to represent and visualize genetic variants pooled from various sources. The tool displays sequences and variants at different levels facilitating representation of variant sites and annotations in a user friendly and interactive manner.

The code for BioJS-HGV Viewer is available at: `https://github.com/saketkc/biojs-genetic-variation-viewer`.

A demo is available at: `http://saketkc.github.io/biojs`

I. Introduction

With the advent of next-generation sequencing technologies, it has been possible to profile genomes in larger numbers. One of the chief outcomes of such projects has been the cataloguing of genetic variants in database resources such as dbSNP[1] and Catalogue Of Somatic Mutations In Cancer (COSMIC)[2]. These catalogs contain publicly accessible sets of genetic variants found in species such as mouse, rat, zebrafish and human which can be utilized to study evolutionary relationships, population diversity and disease specific variations. The COSMIC database is a curated set of somatic mutations as observed in cancer samples. Where the current release (v70) contains 1,564,699 unique variants taken from 1,029,547 samples covering 28,735 genes. dbSNP build 142 contains two orders of magnitude more human variants than COSMIC with 112 million refSNPs[3]. The availability of data at such a large scale makes the analysis and interpretation challenging.

Any exploratory attempt at analyzing the variation data would involve visualizing the variants across the genome to determine specific sites, if any, where the mutations are more frequent or are absent completely. Thus, visualization is critical from the point of interpretation of the vast catalogs of variants. *You don't want to piss off COSMIC they will most probably use you tool* There are variant visualization available; for example the COSMIC Genome Browser[2] offers a view with limited flexibility to customize the view and annotations for each variant.

BioJS-HGV Viewer is a BioJS [4] component developed to visualize genetic variants in a comprehensive manner. BioJS is an open source javascript library providing various components to visualize biological data. The visualizations are web based and hence are absolutely platform independent.

*skchoudh@usc.edu

II. Methods

The functionality provided by BioJS-HGV Viewer has two mode views:

- Overview
- Detailed or Zoomed View

The architecture of this component is designed to handle both DNA and protein variants. The current implementation makes use of protein variants. These variant sites have been generated by an un-published webservice made available through EBI. This service has an indexed database of protein variants as reported in the COSMIC and UniProtKB[5] database and is made available as a JSON[6] file. With support for standard data formats, such as VCF,[7] being implemented.

The demo at `http://saketkc.github.io/biojs` loads the variants for protein *J3KP33*, by default. The component however allows loading other proteins by passing an additional argument to the url. For example: `http://saketkc.github.io/biojs/src/test/javascript/TestHGVViewer.html?q=P00533`.

By default, SIFT and Polyphen scores are averaged and the type of mutations are then decided based on this average score. The component however allows user to choose from either or all of the scores.

The user can also choose to hide a particular category of mutations. Both the overview and detailed mode have another *'open view'*2 where these mutations can further be separately visualized as *Stop Gained*, *Missense* and *Splice Region*.

All the visualizations are rendered as scalable vector graphics(SVG) using the *d3js*[8] javascript library

I. Overview Mode

In the default mode the viewer presents variant information in a condensed format using a stacked bar char displaying the number and *type of mutations* at each site. The detailed annotations are displayed on hovering over the rectangle as a *tooltip*. The *type of mutations* are classified as:

- **Benign**
- **Damaging**
- **Mixed**

The 'Mixed' category represents an **intermediate** state between damaging and benign.

The classification currently uses the predictions scores of Polyphen[9] and SIFT[10]. Polyphen scores are on a scale of $[0, 1]$ with 1 indicating that the mutation is damaging and 0 indicating the mutation being benign. SIFT scores also operate on the scale of $[0, 1]$ however 0 indicates a damaging mutation. The webservice has a database of all mutations across various proteins with pre-generated scores which can be retrieved as a JSON file.

The data thus received is parsed for calculating the number of mutations in each category. Each category is defined by threshold levels. For example a Polyphen score between 0.75 and 1.0 can be considered to reflect a damaging mutation. These threshold levels can be modified by the user. The height of each rectangular box depicting the mutation is dynamically adjusted based on the maximum number of variants at any site.

II. Detailed View

In the detailed view3 each individual amino acid on the protein is displayed as a rectangular box with all variants at that site, the height of the rectangle being proportional to the reported frequency. The box for variants is colored based on its type. On a *mouse over* action at the variant box, the tooltip shows detailed information about that particular mutation.

III. Discussion

There has been a lock of interactive tools to visualize catalog of mutations comprehensively. – What do you mean here carefully read!! BioJS-HGV Viewer is an open source BioJS component that can be used as tool to visualize variants in a flexible manner. Thus, BioJS-HGV Viewer can be a powerful tool for visual interpretation of mutation data. Being entirely web based, it is

platform independent.

IV. Acknowledgments

We would like to thank the BioJS community for insightful discussions. This project was funded by Google Summer of Code 2014.

References

[1] E. M. Smigielski, K. Sirotkin, M. Ward, and S. T. Sherry, "dbSNP: a database of single nucleotide polymorphisms.," *Nucleic acids research*, vol. 28, pp. 352–5, Jan. 2000.

[2] S. a. Forbes *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.," *Nucleic acids research*, vol. 39, pp. D945–50, Jan. 2011.

[3] "dbsnp data statistics." `http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm`. Accessed: 2014-10-04.

[4] M. Corpas, "The BioJS article collection of open source components for biological data visualisation," *F1000Research*, vol. 56, pp. 6–8, Feb. 2014.

[5] C. H. Wu, R. Apweiler, *et al.*, "The Universal Protein Resource (UniProt): an expanding universe of protein information.," *Nucleic acids research*, vol. 34, pp. D187–91, Jan. 2006.

[6] "Javascript object notation." `http://www.json.org/`. Accessed: 2014-10-04.

[7] "Vcf (variant call format) version 4.1." `http://www.1000genomes.org/wiki/analysis/variant%20call%20format/vcf-variant-call-format-version-41`. Accessed: 2014-10-04.

[8] "D3.js data driven documents." `http://d3js.org`. Accessed: 2014-10-04.

[9] V. Ramensky, P. Bork, and S. Sunyaev, "Human non-synonymous SNPs: server and survey.," *Nucleic acids research*, vol. 30, pp. 3894–900, Sept. 2002.

[10] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.," *Nature protocols*, vol. 4, pp. 1073–81, Jan. 2009.

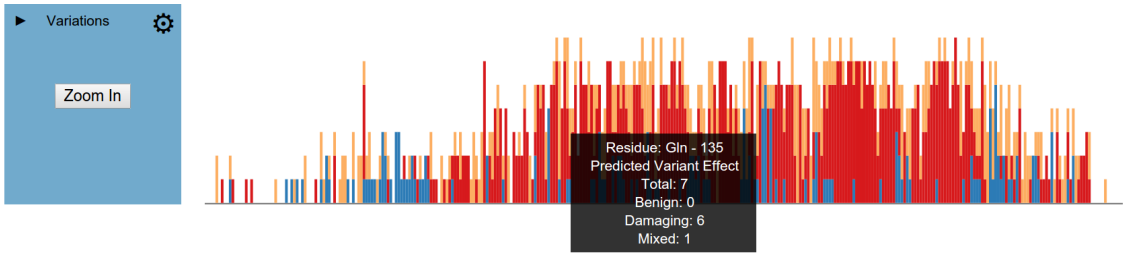

Figure 1: *'Overview' of genetic variants as shown in by HGV viewer. Tooltips are used to display the number of mutations in benign, damaging and mixed categories.*

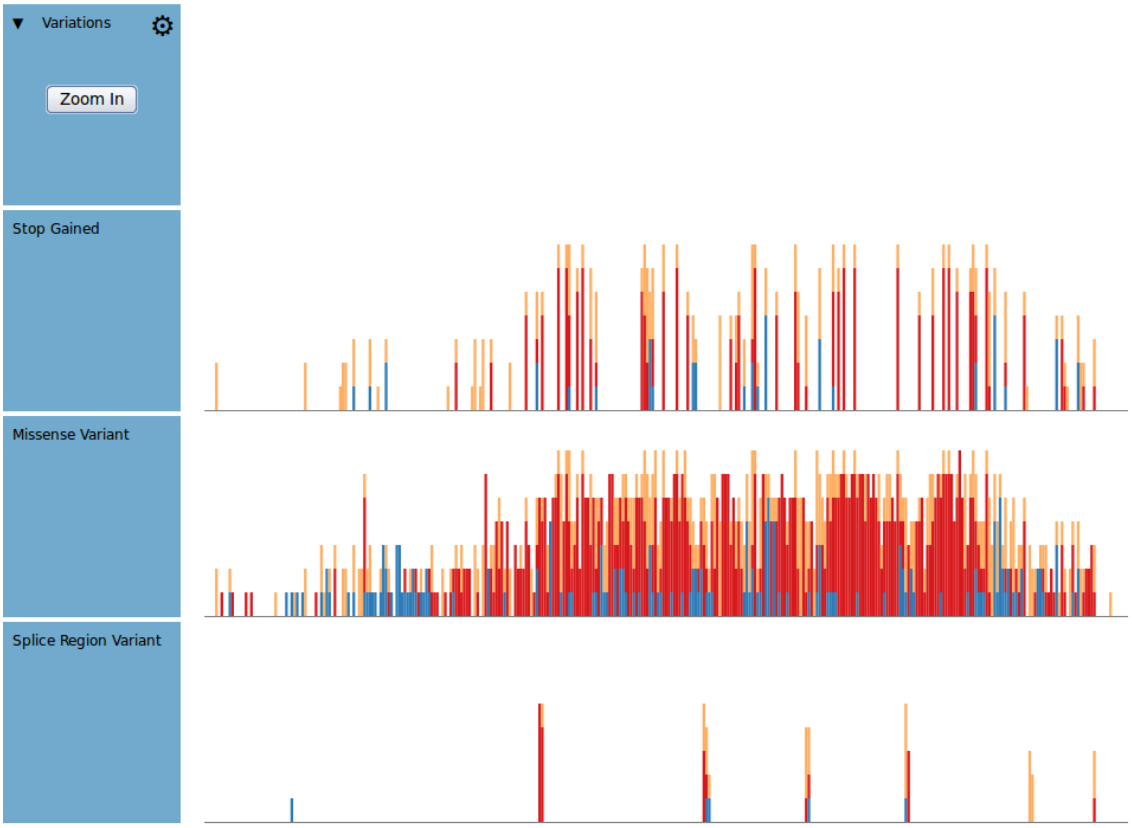

Figure 2: *Overview with open view ON*

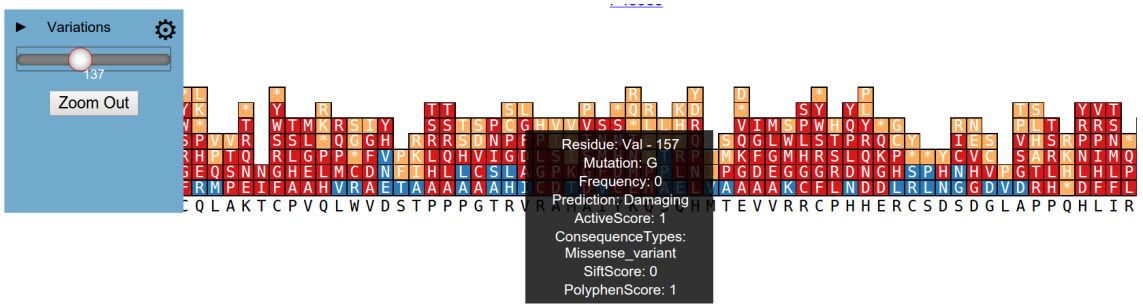

Figure 3: *'Detailed view' of genetic variants. The SIFT/Polyphen scores and associated information with the mutations is rendered using tooltips*

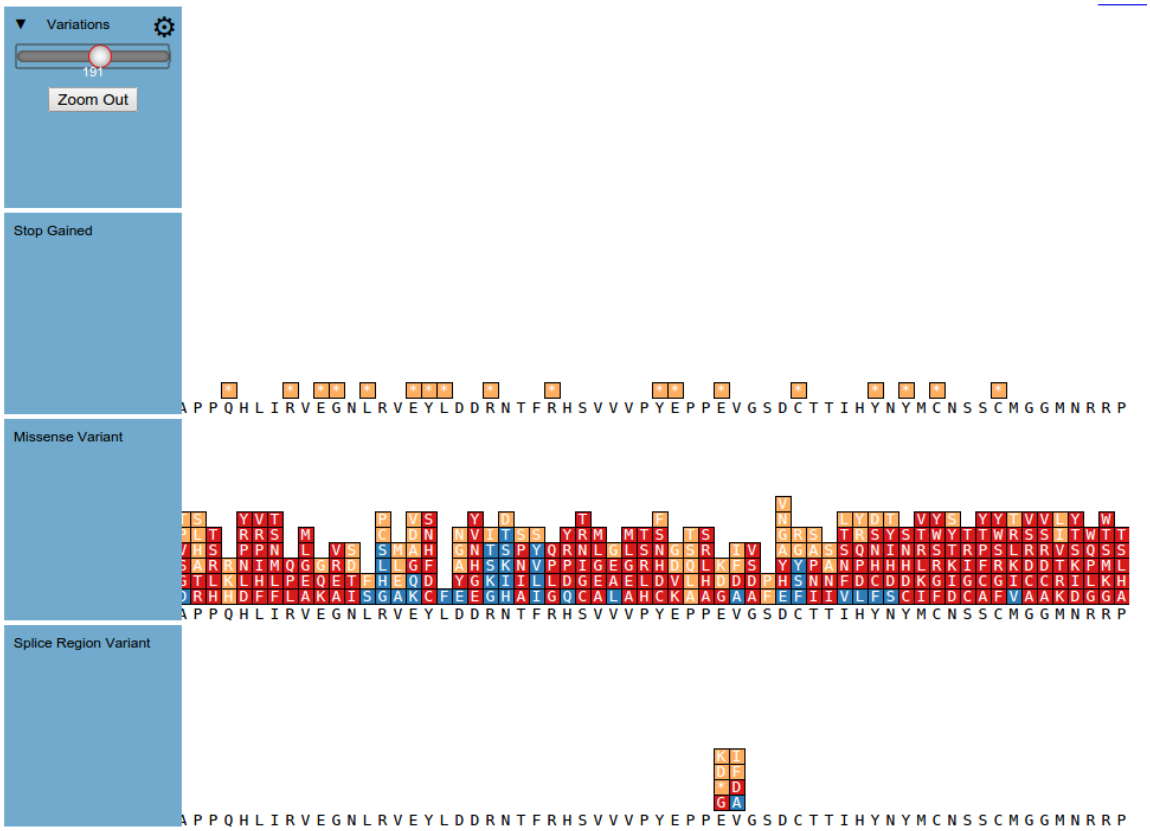

Figure 4: *Zoomed with open view*