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

Genomic studies have resulted in catalogs of genetic variants in humans. Studying the pattern of damaging and non-damaging variants can not only help understand evolution but can also potentially improve human health by identifying the key driver elements.

We present BioJS-HGV Viewer, a BioJS component to represent and visualize genetic variants pooled from various sources. The component presents information at different levels allowing the end user to study the pattern of variations in detail in a user friendly 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 large numbers. One of the chief outcomes of such projects has been catalog of genetic variants such as dbSNP[1] and COSMIC[2]. These catalogs contain publicly accessible sets of genetic variants found in humans which can be utilized to study evolutionary relationships and disease specific variations. COSMIC database is a curated set of somatic mutations as observed in cancer samples. The number of such variations are huge. dbSNP 129 had reportedly more than 14 million unique variants [3]. The availability of data at such a large scale makes the analysis challenging.

Any exploratory attempt at making sense of 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. BioJS-HGV Viewer is a BioJS [4] component developed to visualize genetic variants in a comprehensive manner.BioJS is an open source project providing various components to visualize biological data. These components use javascript for rendering visualization. The visualizations are web based and hence are absolutely platform independent.

II. Methods

The functionality provided by BioJS-HGV Viewer has two parts:

- 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

*skchoudh@usc.edu

available through EBI. This service has an indexed database of protein variants as reported in the COSMIC and UniProt[5] database and is made available as a JSON[6] file. The support for standard data formats such as VCF[7] is under process.

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 demo at `http://saketkc.github.io/biojs` loads the protein *J3KP33*. 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`

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*.

I. Overview Mode

In the default mode the viewer presents variant information in a condensed format focusing on the number and *type of mutations* at each site. 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[8] and SIFT[9]. Polyphen generate scores 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 with pre-generated scores for mutations across all proteins 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.

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 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

BioJS-HGV Viewer can be used to study the pattern of mutations. A visual inspection can help discover sites of conserved regions or regions which are frequently mutated. Being entirely web based, it can be accessed from any platform.

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

[9] 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.

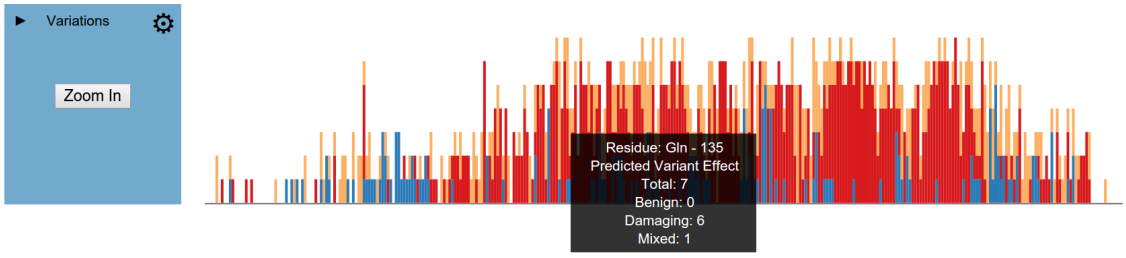

Figure 1: *'Overview' of genetic variants as shown in by HGV viewer. Tooltips are used to display the number of mutations in benign, damaging and mixed categories.*

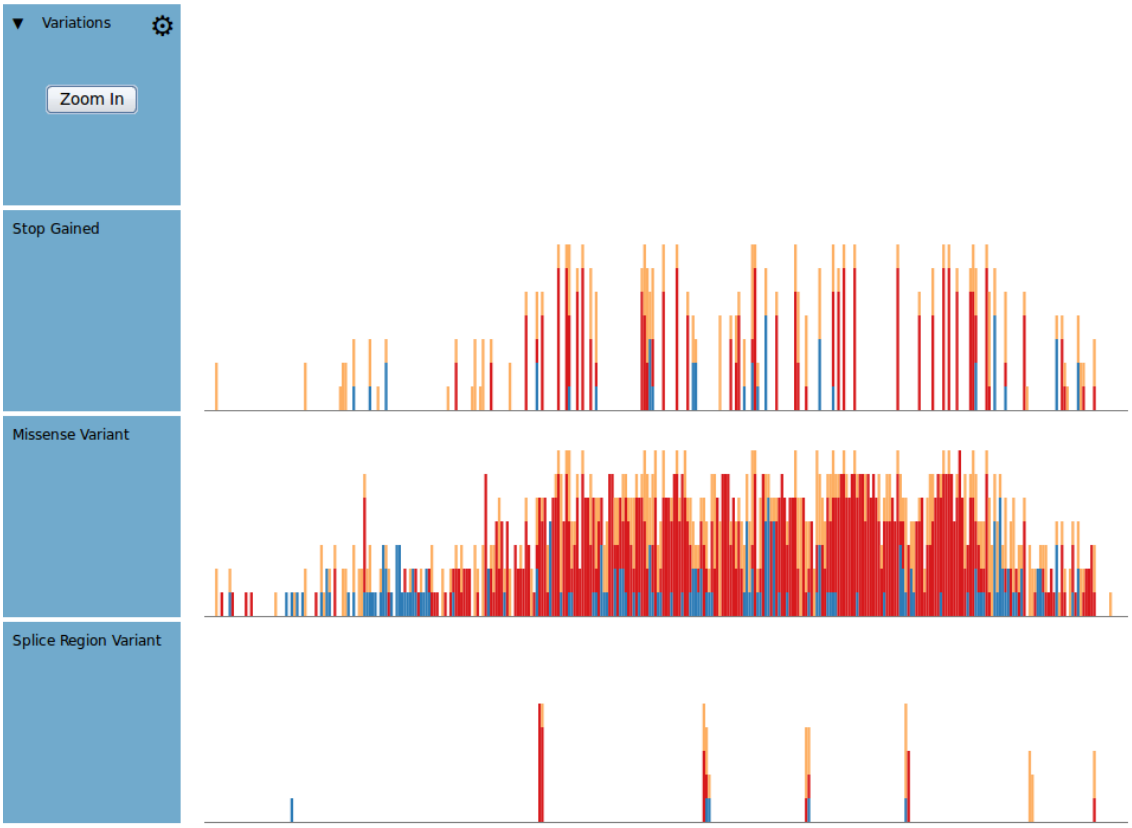

Figure 2: *Overview with open view ON*

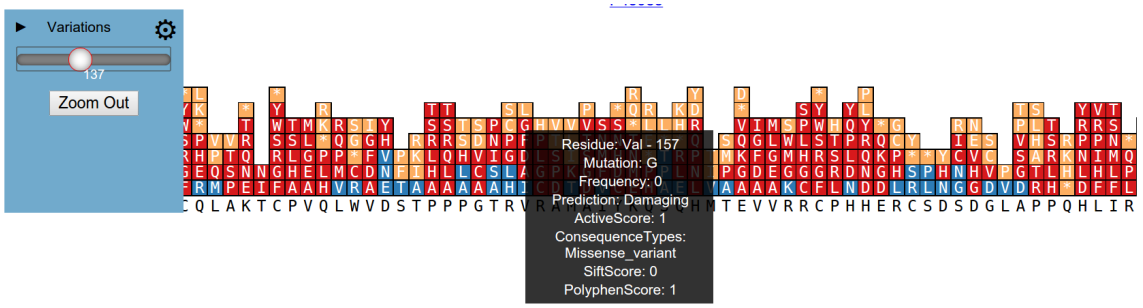

Figure 3: *'Detailed view' of genetic variants. The SIFT/Polyphen scores and associated information with the mutations is rendered using tooltips*

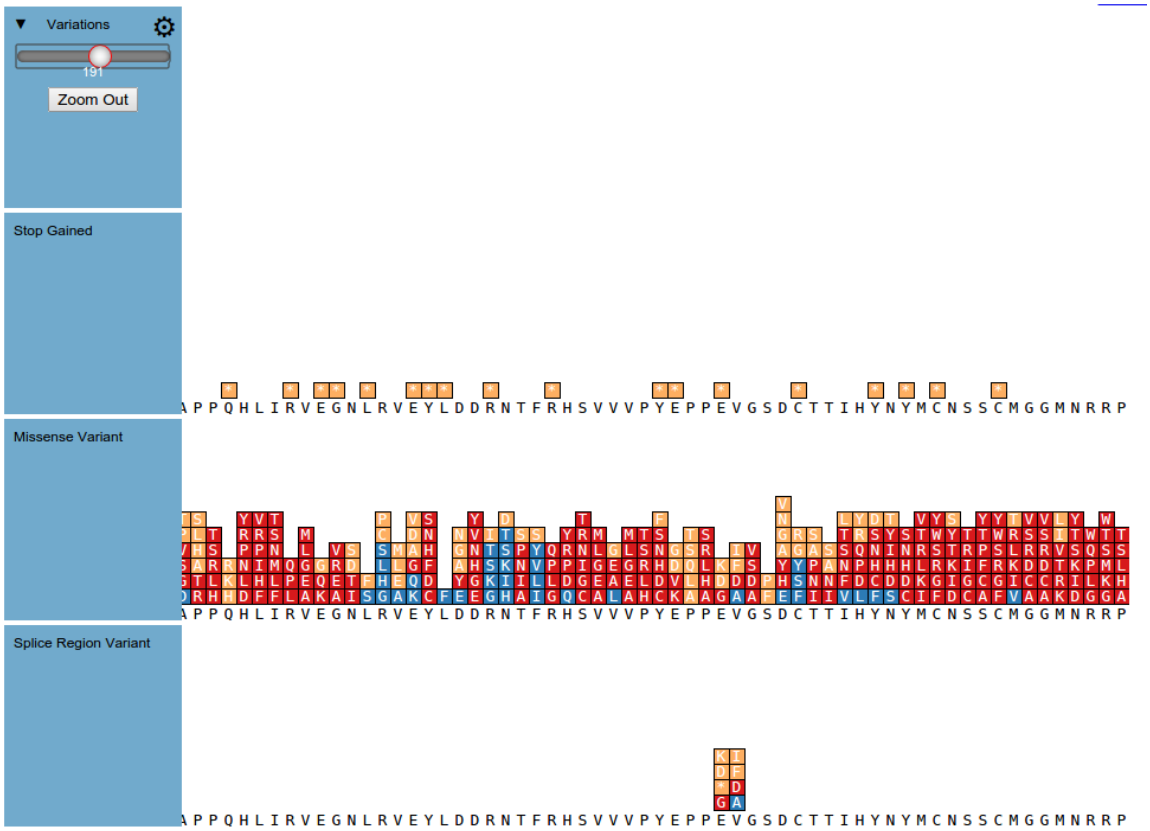

Figure 4: *Zoomed with open view*