BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 17 2013, pages 2195–2196
doi:10.1093/bioinformatics/btt354

Genome analysis

Advance Access publication June 19, 2013

GPViz: dynamic visualization of genomic regions and variants affecting protein domains

Rene Snajder[1,2], Zlatko Trajanoski[1] and Hubert Hackl[1,*]

[1]Biocenter, Division of Bioinformatics, Innsbruck Medical University and [2]Oncotyrol Center for Personalized Cancer Medicine GmbH, 6020 Innsbruck, Austria

Associate Editor: Burkhard Rost

ABSTRACT

Summary: GPViz is a versatile Java-based software for dynamic gene-centered visualization of genomic regions and/or variants. User-defined data can be loaded in common formats as resulting from analysis workflows used in sequencing applications and studied in the context of the gene, the corresponding transcript isoforms, proteins and their domains or other protein features. Both the genomic regions and variants can be also defined interactively. Various gene filter options are provided to enable an intersection of variants, genomic regions and affected protein features. Finally, by using GPViz, we identified differentially expressed exons, which could indicate alternative splicing events, and found somatic variants in different cancer types affecting metabolic proteins. GPViz is freely available at http://icbi.at/gpviz (released under GNU general public license), is based on Java 7 and can be used as a stand-alone or Web Start application.

Availability: http://icbi.at/gpviz

Contact: hubert.hackl@i-med.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 1, 2013; revised on May 24, 2013; accepted on June 17, 2013

1 INTRODUCTION

Recently emerging high-throughput sequencing technologies enable the detection of mutations and genomic variants (Single Nucleotide Variants) from exome or whole genome sequencing data and the identification of differentially expressed exons or transcripts from RNAseq data. Several tools are available for mapping features to the genome, including subsequent visualization [e.g. Integrative Genomics Viewer (Thorvaldsdóttir *et al.*, 2012)]. To functionally interpret the data, it is necessary to analyze and visualize variants or other genomic features also in the context of proteins and protein domains. Algorithms like Sorting Intolerant From Tolerant (SIFT) or PolyPhen-2 and other functional impact metrics allow prediction of the effect of variants (i.e. nsSNPs) and amino acid substitution on protein functions [see recent review (Pabinger *et al.*, 2013)]. MuSiC (Dees *et al.*, 2012) can be applied to perform a proximity analysis of mutations. Another application allows domain mapping of disease mutations (Peterson *et al.*, 2010), and within the Ensembl protein summary view (Flicek *et al.*, 2013) reported mutations can be visualized in the context of protein domains. Alternative splicing and exons can be analyzed and visualized using AltAnalyze and domain graph (Emig *et al.*, 2010). The web application that covers a number of intended issues to visualize genes, transcripts and protein domains or other features is FancyGene (Rambaldi and Ciccarelli, 2009). Finally, there are a number of further possibilities for visualization of multidimensional genomic and sequencing data (Schroeder *et al.*, 2013).

To the best of our knowledge, an interactive tool for visualization of differentially expressed exons (or defined genomic regions) and nucleotide variants in the context of gene structure, transcript variants and encoded proteins is still missing. To close this gap, we developed GPViz, a versatile Java-based software, which enables this visualization based on the gene and transcript structure (genomic coordinates of exons), gene, protein and domain annotations.

2 FEATURES

2.1 Input and output data formats

GPViz supports common data formats, as used within and resulting from standard sequencing workflows and genome browsers. The following four types of information can be optionally loaded:

(1) *Gene structures* can be loaded as General Transfer Format files. These data are essential for usability.

(2) *Genomic regions* can be added in Browser Extensible Data (BED) format (tab-delimited text file, including information about chromosome, start and end position within chromosome and strand).

(3) *Nucleotide variants and mutations* can be loaded in standardized formats (mutation annotation format, mutation format, variant call format and SNP file format).

(4) *Protein domains and features* can be loaded in formats as retrieved from protein domain databases (i.e. Conserved Domain Database or Pfam) or as customized tab-delimited text file including domain information (protein ID, type, from, to, ID and name).

Additionally, to map additional gene and protein information (e.g. UniProt ID) annotation files are loaded automatically in the background. This feature can be configured in the option dialog.

Publication-ready images can be saved in different sizes and resolutions either individually (genewise), as a batch of several different files (in many common file formats including scalable

*To whom correspondence should be addressed.

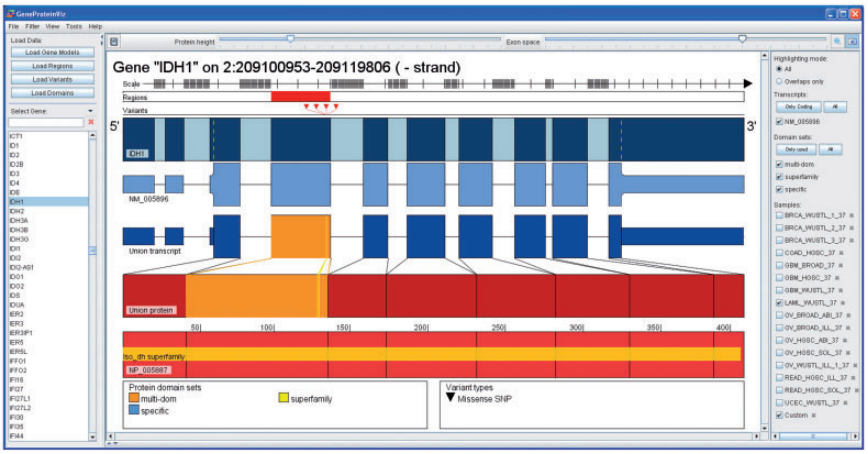

Fig. 1. Somatic mutations in acute myeloid leukemia (data provided by The Cancer Genome Atlas) are affecting a protein domain of IDH1 (amino acid position 132) visualized using GPViz

vector graphic), or as multiple images into one multipage document (Portable Document Format or Tagged Image File Format).

2.2 Display and filter options

GPViz provides various display options including:

- Interactively adding and removing regions and/or variants that are displayed as individual tracks (different types of mutations are shown with different symbols),
- Showing and hiding of individual samples (variants or regions), transcripts or protein domain sets,
- Highlighting of regions and variants in a union transcript and union protein and
- Changing color schemes for all features, zooming view and varying intron–exon space ratio.

The advanced filter options enable filtering of genes that show variants, regions mapping to the gene, mapping to protein domains or mapping and intersecting defined regions and variants.

3 RESULTS

All data can be customized by the user because GPViz is developed as a visualization tool to be used for any application and organism. We additionally provide test datasets considering the latest genome assembly of human and mouse and annotation from Refseq and Ensembl. To demonstrate how GPViz can be used, two case studies were analyzed step-by-step (see Supplementary Material). Briefly, RNAseq analyses using applications like DEXSeq (Anders *et al*., 2012) or exon array analyses result in a list of differentially expressed exons. GPViz analysis enabled the selection of candidate exons to distinguish between transcript isoforms (indicating alternative splicing events) as well as filtering and visualization of exons affecting protein domains. In the second case study, we could show that somatic mutations in different cancer types might affect metabolic enzymes and, therefore, alter cancer metabolism. A representative example of a somatic mutation in the gene isocitrate dehydrogenase 1 (IDH1) (amino acid position 132) occurring in acute myeloid leukemia (TCGA, 2013) is shown in Figure 1. Using an additional sample showed that this mutation occurs also in other cancer types (i.e. rectal adenocarcinoma).

Funding: Tiroler Standortagentur (Bioinformatics Tyrol) and the FFG project Oncotyrol.

Conflict of Interest: none declared.

REFERENCES

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res*., **22**, 2008–2017.

Dees,N.D. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res*., **22**, 1589–1598.

Emig,D. *et al.* (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res*., **38**, W755–W762.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res*., **41**, D48–D55.

Pabinger,S. *et al.* (2013) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. [Epub ahead of print, doi: 10.1093/bib/bbs086, January 21, 2013].

Peterson,T.A. *et al.* (2010) DMDM: domain mapping of disease mutations. *Bioinformatics*, **26**, 2458–2459.

Rambaldi,D. and Ciccarelli,F.D. (2009) FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics*, **25**, 2281–2282.

Schroeder,M.P. *et al.* (2013) Visualizing multidimensional cancer genomics data. *Genome Med*., **5**, 9.

TCGA. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med*. **368**, 2059–2074.

Thorvaldsdóttir,H. *et al.* (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*., **14**, 178–192.