Bioinformatics, 33(7), 2017, 1081–1082
doi: 10.1093/bioinformatics/btw761
Advance Access Publication Date: 31 December 2016
Applications Note

OXFORD

Genome analysis

PanViz: interactive visualization of the structure of functionally annotated pangenomes

Thomas Lin Pedersen[1,2,*], Intawat Nookaew[3,4], David Wayne Ussery[3,4] and Maria Månsson[2]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, DK-2800 Lyngby, Denmark, [2]Assays, Culture and Enzymes Division, DK-2970 Hørsholm, Denmark, [3]Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA and [4]Department Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 4, 2016; revised on November 23, 2016; editorial decision on November 24, 2016; accepted on November 29, 2016

Abstract

Summary: PanViz is a novel, interactive, visualization tool for pangenome analysis. PanViz allows visualization of changes in gene group (groups of similar genes across genomes) classification as different subsets of pangenomes are selected, as well as comparisons of individual genomes to pangenomes with gene ontology based navigation of gene groups. Furthermore it allows for rich and complex visual querying of gene groups in the pangenome. PanViz visualizations require no external programs and are easily sharable, allowing for rapid pangenome analyses.
Availability and Implementation: PanViz is written entirely in JavaScript and is available on https://github.com/thomasp85/PanViz. A companion R package that facilitates the creation of PanViz visualizations from a range of data formats is released through Bioconductor and is available at https://bioconductor.org/packages/PanVizGenerator.
Contact: thomasp85@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Visualization plays an integral part in modern biology research, as the size and complexity of biological data has increased (Land *et al.*, 2015). Many new tools for analysis of pangenomes have recently been published (Grant *et al.*, 2012; Hallin *et al.*, 2008; Lechat *et al.*, 2012; Rokicki *et al.*, 2014). However, several of these do not scale well. A general tendency for pangenome visualizations is to use the chromosome of a reference genome as an axis and plot synteny between genomes along that. Visualization based on a reference genome fails to take into account novel pan-genes not in the reference. As more genomes are added, the reference genome becomes less representative of the full dataset. Some attempts have been made to create reference free, scalable pangenome visualization of different types. GenoSets (Cain *et al.*, 2012) is a visualization that uses parallel sets to facilitate gene group selections based on presence-absence in pangenome subsets and GenomeRing (Herbig *et al.*, 2012) tries to overcome the reference bias by merging all chromosomes into a superchromosome that can be used as a backbone for visualization. Here we present a new interactive visualization, PanViz, aimed at letting users explore the structure of functionally annotated pangenomes and pangenome subsets, while performing visual queries to search for gene groups. PanViz is based purely on GO annotation and the presence/absence pattern of gene groups, and is thus not dependent on a single reference genome

2 Implementation

PanViz is written entirely in JavaScript using D3 (Bostock *et al.*, 2011). It is completely self-contained, embedded in a single HTML file, and does not require any connection to external sources. A companion R package, PanVizGenerator, has been released on Bioconductor (Gentleman *et al.*, 2003; Huber *et al.*, 2015) that

facilitates the creation of new PanViz visualizations. The input data needed by PanVizGenerator is a pangenome matrix giving the presence/absence pattern of each gene group across the included genomes, as well as a Gene Ontology (GO) (Ashburner *et al.*, 2000) based functional annotation of each gene group. The latter can be derived by analyzing a representative sequence for each gene group using e.g. InterProScan (Jones *et al.*, 2014; Zdobnov and Apweiler, 2001) or Blast2GO (Conesa *et al.*, 2005). A gene group in this context is a group of similar genes across genomes. The simple nature of the input data means that PanViz works with any pangenome tool, though the method of achieving a pangenome matrix might vary.

3 Overview

PanViz consists of four main areas (see Supplementary Fig. S1 in sup plementary material). The left part is reserved for (pan)genome navigation (Supplementary Fig. S1A), the center part for pangenome visualization (Supplementary Fig. S1B), while the right part is for legends and additional lookup information (Supplementary Fig. S1C). In the bottom a list of all the currently selected gene groups is available, as well as tools to modify the selection (Supplementary Fig. S1D).

3.1 Genome navigation

The genomes in the pangenome are represented by a dual linked view with a principal component analysis/multidimensional scaling based scatterplot on top, and a zoomable hierarchical clustering in the bottom (Supplementary Fig. S1A) both based on the pangenome matrix. Both views support selecting single genomes in order to transition into the genome-pangenome comparison state, and the dendrogram allows for selection of subsets of the pangenome by selecting the branch points of the dendrogram. The overview plots are also linked to the gene group table (Supplementary Fig. S1D) so that all genomes containing the gene group currently hovered over will be highlighted.

3.2 Pangenome overview

The main view of the visualization is a radial representation of the 3 presence-based gene group groupings in the pangenome: Core, Accessory and Singleton gene groups (Supplementary Fig. S1B1). Each of these is furthermore divided based on the distribution of top level biological process GO terms. As different sub-pangenomes are selected, the changes in the pangenome are animated by moving sections of each GO term arc around. After the animation ends the dynamics can furthermore be shown as chords when hovering over a specific GO arc.

3.3 Genome-pangenome comparison

When one or two genomes are selected the main view transitions into a stacked bar chart showing the pangenome in the middle (Supplementary Fig. S1B2). The genomes are represented by their GO term composition and weighted bezier curves connectes the genes in the genomes to their location in the pangenome (if present). If two genomes are selected the proportion of each GO term they share with each other and the pangenome is visible as a darker shaded bezier curve.

3.4 Gene ontology naivgation

To gain insight into the distribution of lower level GO terms a treemap weighted by the number of gene groups in the current pangenome having a specific term is available upon selecting a top level GO term bar from the pangenome (Supplementary Fig. S1B3). The treemap is zoomable and features descriptions of each included GO term.

3.5 Visual querying

As each visual element represents of a set of gene groups it makes sense to build a querying mechanism based on set arithmetic (union, intersection, complement, etc.). An icon on top of the gene group table indicates the different set operations available (Supplementary Fig. S1D). The operations will be performed between the current content of the table and the gene groups contained in the visual element selected. Based on the six different operations it is possible to intuitively create very complex gene group queries guided by the insights gained from the visualization.

4 Conclusion

PanViz offers a novel and unbiased approach to visualizing the structure of pangenome data. Interactions and animations are utilized to invite users to investigate the data, and the reliance of a single self-contained HTML file makes it easy to share with fellow researchers. The main visualization is fully scalable to thousands of gene groups and genomes as it relies on summaries, but larger pangenomes will require faster hardware due to the dynamic nature of the visualization. PanViz helps researchers in understanding how different pangenomes differ on a functional level rather than simply in terms of shared gene groups. Future work will focus on implementing state saving within the URL to facilitate sharing of different states of a PanViz visualization, as well as performance improvement to the implementation.

Funding

This work was supported by The Danish Agency for Science, Technology and Innovation.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bostock,M. *et al.* (2011) D^3: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.*, **17**, 2301–2309.

Cain,A.A. *et al.* (2012) GenoSets: visual analytic methods for comparative genomics. *PLoS ONE*, **7**, e46401.

Conesa,A. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Gentleman,R.C. *et al.* (2003) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80–R80.

Grant,J.R. *et al.* (2012) Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics*, **13**, 202.

Hallin,P.F. *et al.* (2008) The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology. *Mol. bioSyst.*, **4**, 363–371.

Herbig,A. *et al.* (2012) GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, **28**, i7–15.

Huber,W. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.

Jones,P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Land,M. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.

Lechat,P. *et al.* (2012) SynTView – an interactive multi-view genome browser for next-generation comparative microorganism genomics. *BMC Bioinformatics*, **14**, 277–277.

Rokicki,J. *et al.* (2014) CodaChrome: a tool for the visualization of proteome conservation across all fully sequenced bacterial genomes. *BMC Genomics*, **15**, 65.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.