Pantograph: Scalable Interactive Graph Genome Visualization

Josiah Seaman, Simon Heumos, Andrea Guarracino, Artem Tarasov, Bonface Munyoki, Christine Seaman, Dmytro Trybushnyi, Eloi Durant, Hannah Sewell, Jack Tierney, Jacob Windsor, Jerven Bolleman, Jörg Hagmann, Katherine Innamorati, Njagi Mwaniki, Robert Fornof, Mark Seaman, Michael R. Crusoe, Stacie Seaman, Thomas Townsley, Torsten Pook, Toshiyuki T. Yokoyama, Travis Clark, Erik Garrison

ABSTRACT

Introduction: Traditionally, an individual's genetic variation is inferred by comparing to one single reference. This reference biased view is supported by a toolchain including genome browsers. Pangenome graphs provide extensive benefits over reference-based approaches but so far lack scalable visualization solutions. Here we present Pantograph, a scalable, interactive graphical pangenome browser.

Background: Pantograph is an open source browsable pangenome visualization for graph genomes. It allows researchers to see the full genetic diversity in large populations. Graph genomes naturally express genome rearrangements, SNPs, and indels. Visualized variation graphs can be expressed as knowledge graphs. These can further be enriched with annotations, geographical locations, and patient outcomes. This makes Pantograph an ideal tool for tracking viral strains of SARS-CoV-2. Without the constraints of a reference genome, viral strains can be smoothly integrated as they are sequenced.

Description: Pantograph is a data and visualization solution that scales to thousands of individuals while preserving all types of sequence variation. Other tools either are not scalable or discard all genome rearrangements. Pantograph achieves scalability by identifying syntenic blocks and interconnecting them with nonlinear variants. Pantograph's application to the COVID-19 pandemic is driven by the unique evolutionary scenario we are facing. The infected population size is a multiplier for the number of mutations available for selection. Vaccinating the population during a pandemic is a selection sweep for resistant viral mutations. This is the biggest viral selection sweep in human history. Pantograph can help predict vaccine effectiveness in different regions of the world by integrating our total knowledge of genetic diversity.

METHODS

To achieve scalability, Pantograph implements a multi-stage data analysis pipeline written in C++, Python, R, and JavaScript. Graphical pangenomes built by vg (Garrison et al. 2018)[1] or seqwish[2] are stored in GFAv1[3]. Odgi (Eizenga et al. 2020) can sort them into a roughly linear ordering. Linear ordering allows for easy browsability unlike a pure graph method such as Bandage (Wick et al. 2015) or GfaViz (Gonnella et al. 2018). MSAs are easy to understand but lack markup for repeated sequences and translocations (Fig. 1). Pantograph identifies co-linear syntenic regions and encapsulates them in *Components*. They are then connected in a graph using *Links*, which are edges representing inversions or translocations. This visualization technique allows Pantograph to treat any individuals which share a rearrangement as a single "variant" called a *Link Column* which contain colored cells indicating presence or absence in an individual (Fig. 2).

```
TCCAA---CTCTCTGTGGTTCCGGTTGCAT*
TCC-AGGTCTCTCTGTGGTTCC----GCATGGT*
TCCAA---CTCTCTGGGGTTCCGG-TGTATGGT*
TCC-AGGTCTCTCTGAGGTTCC----GTAT*
TCC-AGGTCTCTCTGTGGTTCCGG-TGTAT*
```

Fig. 1: *Five example sequences in a multiple sequence alignment. Repeated sequence fragments have been colored by hand to indicate rearrangements not shown in a multiple sequence alignment.*

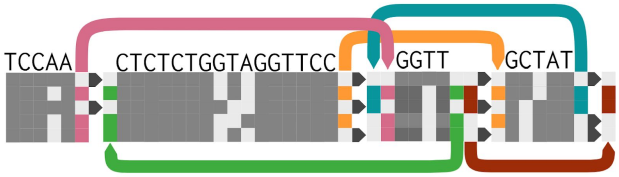

Fig. 2: *A Preview of a Pangenome Schematic showing the same variation as Fig. 1 but including markup for the transpositions and duplications, SNPs, duplications, and rearrangements in five individuals. Every component reads from left to right; only follow each Link once. Each row is an individual. For example, the bottom row reads TCCA then follows the pink Link to GGT then follows the green link to CTCTCGTGGTTCC. The black connector adds a second copy of GGT, we ignore the Green Link this time and proceed to GTAT at the end.*

[1] https://pangenome.github.io/

[2] https://github.com/ekg/seqwish

[3] https://github.com/GFA-spec/GFA-spec/blob/master/GFA1.md

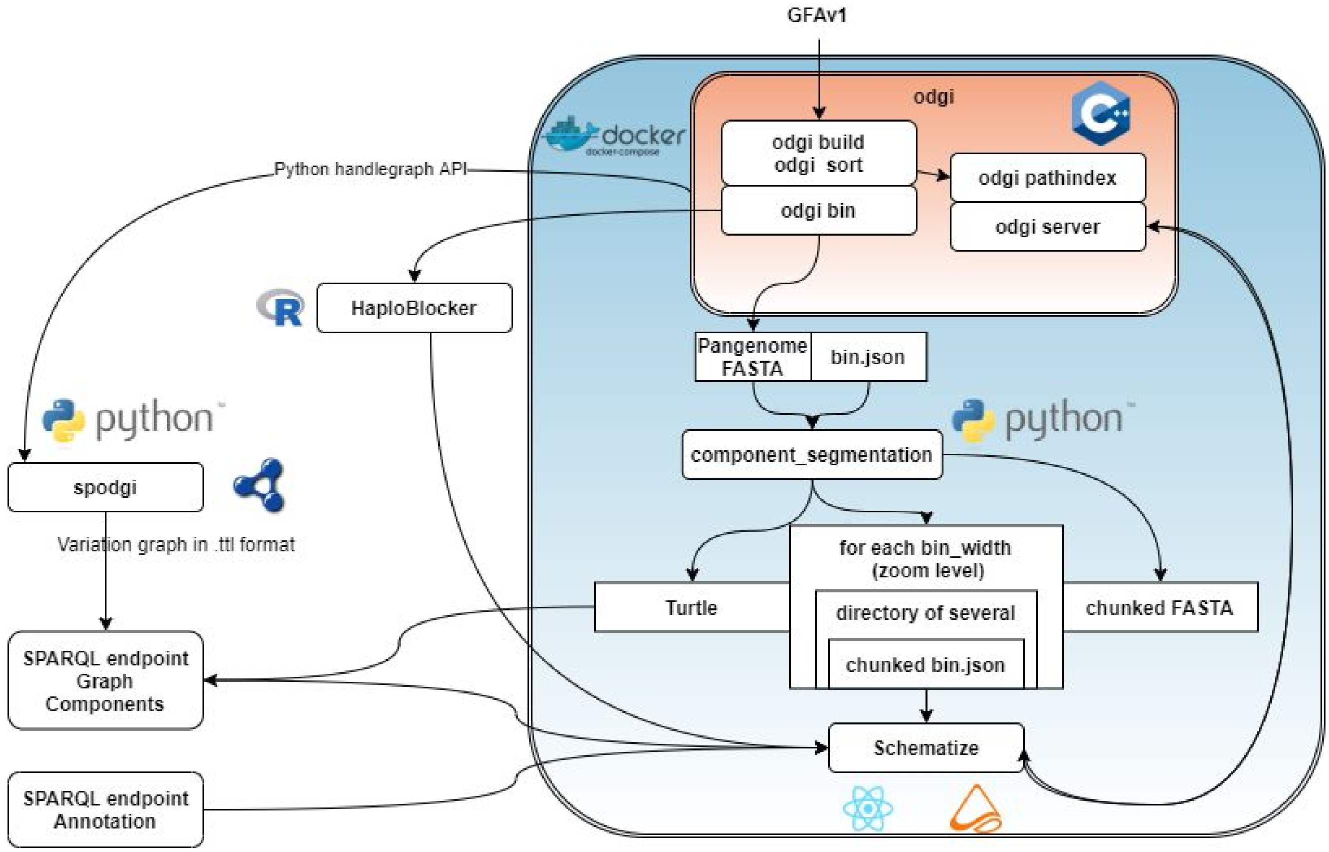

***Fig. 3:** Pantograph pipeline diagram*

On the frontend, the React JavaScript library is used to render variations within a *Component* as an MSA. *Links* in the graph behave like hyperlinks to navigate to connected loci. Pantograph can visualize the complete genome of thousands of individuals down to the nucleotide level and enable visual inspection datasets on a scale that has been previously impossible.

RESULTS

Pangenome browser development was part of a larger COVID-19 Virtual Biohackathon 2020[4] involving over 300 people. We have generated a SARS-CoV-2 pangenome from 850 individual *de novo* genome assemblies. A public server provides an instance of Pantograph, giving researchers the ability to browse a live-updating view of our complete knowledge of SARS-CoV-2 genetic diversity. See graphgenome.org for more details. Source code is available at https://github.com/graph-genome.

Future development will integrate annotations, sample metadata, and knowledge graph resources via SPARQL queries.

REFERENCES

Eizenga,J.M. et al. (2020) Succinct dynamic variation graphs. bioRxiv,2020.04.23.056317, 1–6.

Garrison,E. et al. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol., 36, 875–879.

Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualisation of de novo genome assemblies. Bioinformatics, 31(20), 3350-3352.

Giorgio Gonnella, Niklas Niehus, Stefan Kurtz. GfaViz: Flexible and interactive visualization of GFA sequence graphs. Bioinformatics, bty1046 (2018). DOI:10.1093/bioinformatics/bty104

[4] https://github.com/virtual-biohackathons/covid-19-bh20