The PanGenome Graph Builder

Andrea Guarracino[1]*, Simon Heumos[2]*, Flavia Villani[3], Emilio Rudbeck[4], Kaisa Thorell[4], Lorenzo Tattini[5], Christian Kubica[6], Sebastian Vorbrugg[6], Christian Fischer[7], Njagi Mwaniki[8], Sven Nahnsen[2], David Ashbrook[7], Robert Williams[7], Hao Chen[7], Vincenza Colonna[3], Pjotr Prins[7], Erik Garrison[7]

[1]University of Tor Vergata, Biology, Rome, Italy, [2]University of Tübingen, QBiC, Tübingen, Germany, [3]CNR, IGB, Naples, Italy, [4]University of Gothenburg, Infectious Diseases, Gothenburg, Sweden, [5]Université Côte d'Azur, CNRS, Nice, France, [6]MPI, Developmental Biology, Tübingen, Germany, [7]UTHSC, Genomics, Memphis, TN, USA, [8]KEMRI-Wellcome Trust, Training, Kilifi, Kenya

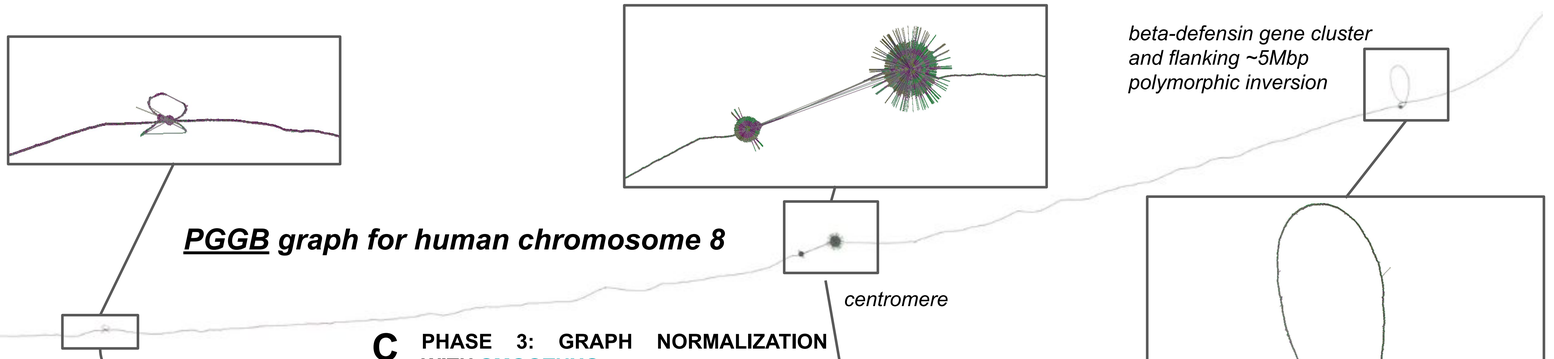
Contributed equally.

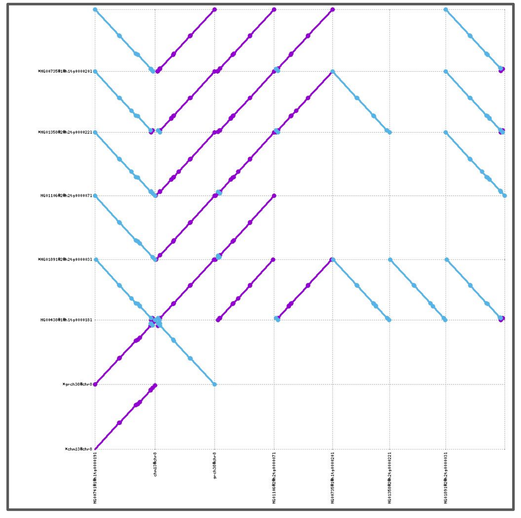
Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag; ZIM Zentrales Innovationsprogramm Mittelstand

Pangenome graphs[1] contain the full genomic information of a species. For the unbiased evaluation of a species' genomic variation space, all versus all comparisons are essential. We implement the PanGenome Graph Builder (PGGB), scaling efficiently to large collections of multi-gigabase genomes. The method does not require a reference. It consists of three phases. First, we generate all versus all alignments of the input sequences. Then, the sequences and alignments are used to induce a variation graph. Finally, the graph is normalized by sorting it with an unsupervised machine learning method and applying partial order alignment to blocks in the sorted order. We apply PGGB to sequence data of different species. The resulting graphs provide excellent targets for the mapping of short and long reads, and are a basis for comparative genomic applications.

To provide a visual explanation of PGGB, we apply it to human chromosome 8. As input, we use 2 reference assemblies (GRCh38 and CHM13) and 6 *de novo* assemblies from the Human Pangenome Reference Consortium' year 1 assemblies which span all or most of the chromosome.

PGGB has three distinct phases which require 20 minutes in total to obtain this graph: (**A**) all-to-all alignment with wfmash, (**B**) graph induction with seqwish, and (**C**) normalization with smoothxg, which produces the resulting graph shown. In (**D**) we display features that are visible in the structure of the graph[3]. The whole run takes around 20 minutes on a HPC compute node with an AMD EPYC 7402P 24-Core Processor and 128GB of RAM. In practice, we run with full human genomes by partitioning the input into chromosome-specific jobs, allowing turnaround of a full human pangenome from 91 haplotypes on a modest compute cluster in around a day.





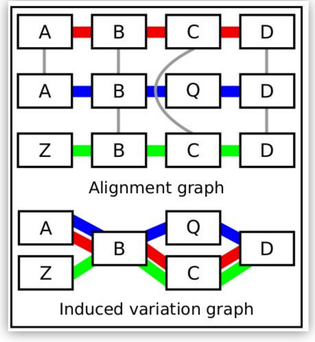
A PHASE 1: ALL-VS-ALL ALIGNMENTS WITH WFMASH

High-performance alignment of whole chromosomes is enabled by first applying MashMap2 with a 100kb segment length and 98% identity filter. A hierarchical implementation of the wavefront algorithm[2] allows us to obtain base-level global alignments for all mappings. Here, a dotplot shows the alignment relationships from which the above graph is built.

PHASE 2: GRAPH INDUCTION WITH SEQWISH

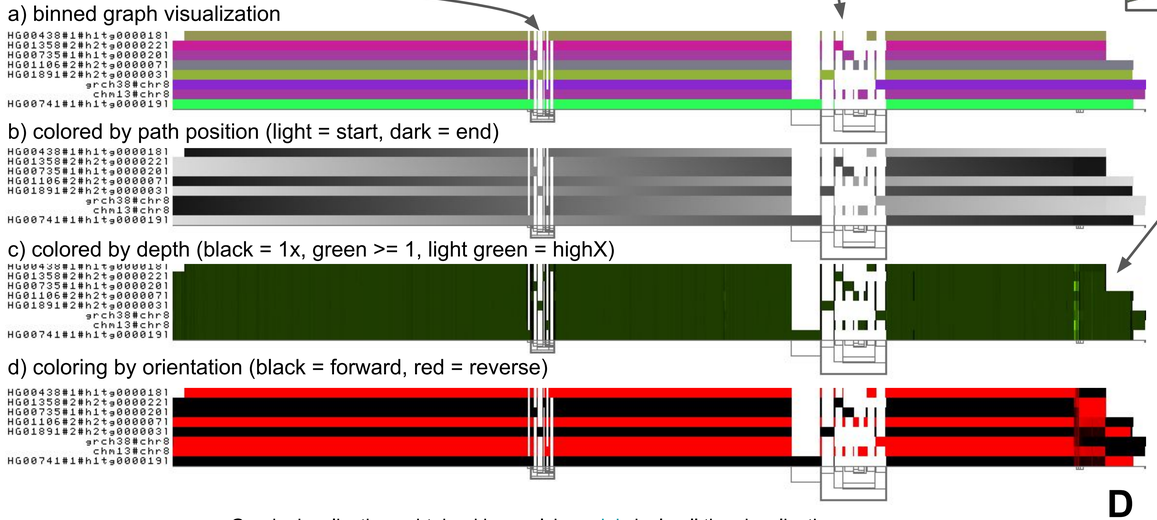
To build a graph, we render it as an alignment graph in which input sequences are nodes and special "alignment edges" connect sequences that have been aligned. To build a variation graph, we collapse the components of the graph connected by alignments and retrace the original paths through the graph.

B



C PHASE 3: GRAPH NORMALIZATION WITH SMOOTHXG

A partial order alignment is run for each part of the graph, resulting in a normalized graph with base-level resolution of all variant classes.



D

Here (c), the copy number variation in the beta-defensin cluster is apparent as a light green stripe near the beginning of the inversion.

d) The inversion can be seen in flips of orientation between the chromosome paths embedded in the graph and the graph's orientation.

Graph visualizations obtained by applying odgi viz. In all the visualizations

- The graph nodes' are arranged from left to right forming the pangenome sequence.
- The colored bars represent the linearized renderings of the embedded paths versus this pangenome sequence in a binary matrix
- The black lines under the paths, so called links, represent the topology of the graph.

References

1. Eizenga et al. (2020). Pangenome Graphs. *Annual Reviews of Genomics and Human Genetics*, 21, 1.
2. Marco-Sola et al. (2020). Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*.
3. Logsdon et al. (2021). The structure, function and evolution of a complete chromosome 8. *Nature.* 593. 101-107.

nf-core/pangenome

Acknowledgements

We thank Vincenza Colonna for organizing the Crusco Summer Hackathon and the Forentum Ritrovato museum for hosting it.
We thank the deNBI cloud for providing computational resources.
S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany.

de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE