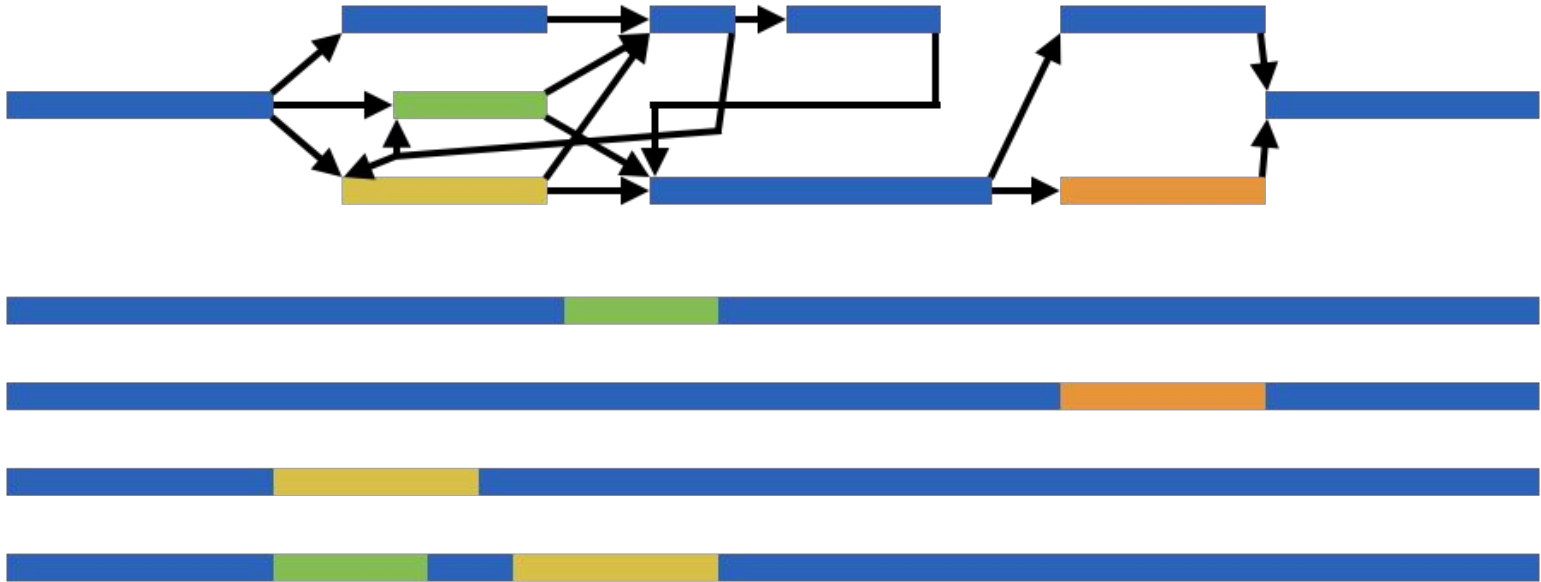


GBWTGraph

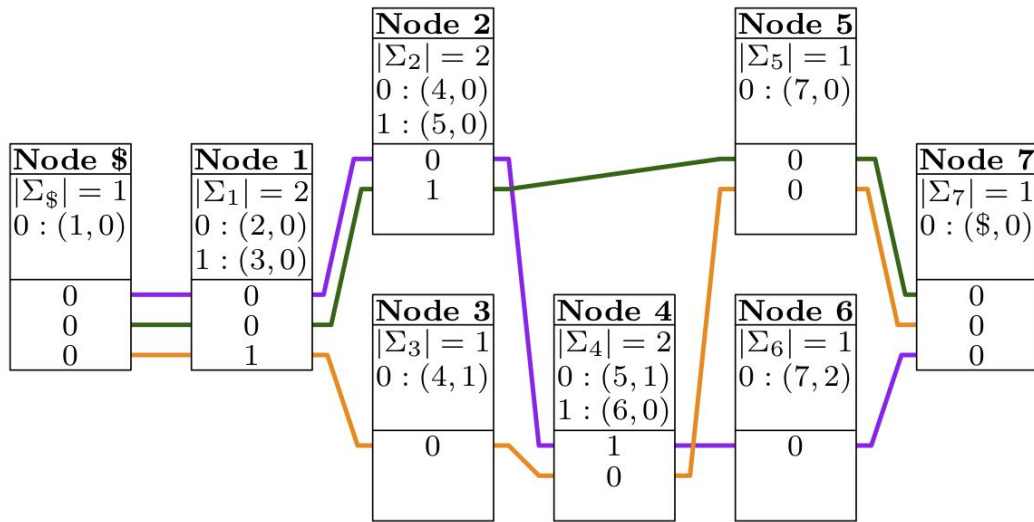
Jouni Sirén

UCSC Genomics Institute



We have:

- A **graph** that tells which **positions** in the haplotypes are **equivalent**.
- **Haplotypes** that tell which **paths** in the graph **make sense**.



Sirén, Garrison, Novak, Paten, and Durbin:

Haplotype-aware graph indexes.

Accepted to Bioinformatics, 2019.

<https://doi.org/10.1093/bioinformatics/btz575>

<https://github.com/vgteam/vg>

<https://github.com/jltsiren/gbwt>

- We developed the GBWT for storing haplotypes as paths over a VG graph.
- The GBWT is an FM-index storing sequences of node identifiers.
- The index contains the topology of the graph induced by the haplotypes.
- If we add the sequences stored in the nodes, the GBWT becomes a fully functional graph representation.
- When we traverse a path, we can easily tell how many haplotypes support the path.

Giraffe = minimizer index + distance index + GBWTGraph

- We are developing a fast haplotype-aware short read to graph aligner.
- Most Illumina sequencing errors are substitutions. The haplotypes already contain most real indels. Hence we can align most reads without expensive dynamic programming.

Future goals

- Develop a standalone GBWTGraph implementation that depends on libhandlegraph, GBWT, and SDSL.
- Explore various improvements to GBWTGraph.
- Applications?