Current options to index, represent, and visualize annotations in a pangenome with the vg toolkit

This manuscript (permalink) was automatically generated from jmonlong/manu-vggafannot@8989ca0 on August 22, 2024.

Authors

- **Jean Monlong** ✉
  0000-0002-9737-5516 · jmonlong
  IRSD - Digestive Health Research Institute, University of Toulouse, INSERM, INRAE, ENVT, UPS, Toulouse, France

✉ — Correspondence possible via email to Jean Monlong <jean.monlong@inserm.fr>.

Abstract

The current reference genome is the backbone of diverse and rich annotations. To enable similar enrichment of a pangenome reference, there is a dire need for tools and formats for pangenomic annotation. Simple text formats, like VCF or BED, have been widely adopted and helped this critical exchange of genomic information. The Graph Alignment Format (GAF) text format, which was proposed to represent alignments, could be used to represent any type of annotation in a pangenome graph. Here I review how some features of the `vg` ecosystem can already provide indexing, querying, and visualization capabilities for annotations represented as paths.

We developed efficient sorting, indexing and querying for GAF files. This approach can for example extract annotations overlapping a subgraph quickly. Alignments are currently sorted based on the covered node IDs, similar to the approach for sorting read alignments in the GAM format, a binary format used previously by the `vg` toolkit. To index the bgzipped GAF file, we extended HTSlib/tabix to work with the GAF format. Second, `vg annotate` was recently updated to better produce graph annotations as paths, starting from annotation files relative to linear references. More precisely, it can to take annotations in BED or GFF3 files, written relative to reference paths or haplotypes, and produce GAF files representing the equivalent paths through the pangenome.

To showcase these commands, we projected annotations for all haplotypes in the latest draft human pangenome (HPRC v1.1 GRCh38-based Minigraph-Cactus pangenome). This included genes, segmental duplications, tandem repeats and repeats annotations. `vg annotate` can annotate ~4M gene annotations in ~16 mins, and ~5.5M repeats from RepeatMasker in ~9 mins on a single-threaded machine. Finally, these rich annotations can then be quickly queried with `vg` and visualized using existing tools like the sequenceTubeMap or Bandage.

References