Annotations in the pangenome with indexed GAF files

Jean Monlong[1,2], Adam M. Novak[1], Dickson Chung[1], Glenn Hickey[1], Toshiyuki T. Yokoyama[3], Erik Garrison[4], Benedict Paten[1]

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA
[3]Department of Computational Biology and Medical Sciences, The University of Tokyo, Chiba, Japan
[2]Institut de Recherche en Santé Digestive, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France
[4]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

A genomic range can be represented as a path in the pangenome. 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. To explore this approach within the vg toolkit, two subcommands were updated: `vg gamsort` to **sort and index bgzipped GAF files**; `vg annotate` to **project annotation on the latest HPRC pangenomes**.

GAF sorting and indexing

Methods

Sort paths by minimum, then maximum node ID.

Tweak HTSlib to index bgzipped GAF files
- Tab-separated file, like VCF or BED
- Instead of indexing on CHR:START-END, index on MIN_NODE-MAX_NODE

Query a node ID range or path range, like we would a genomic range.

Benchmark

Sort 30x Illumina short-read dataset (~300M paired-end reads)

- GAM:
  - File size: 69G sorted
  - Sorting: ~9h and ~2.5Gb mem
- GAF:
  - File size: **24 Gb** sorted
  - Sorting: **~6h** and ~3.5 Gb mem
- GAM to GAF conversion: ~2h

Sorted GAF file with **minimum** and maximum node IDs highlithed

```
read_name_6  100  0  100  +  <394<393<392<391          128  25  124  100  100  0
read_name_7  100  0  100  +  <2075<4074<2073<2072      128  7   106  100  100  0
read_name_2  100  0  100  +  <2300<2299<2298<2297      128  7   106  100  100  0
read_name_9  100  0  100  +  <3222<3221<3220<3219      128  18  117  100  100  0
read_name_5  100  0  100  +  <13108<13107<13106<13105  128  7   106  100  100  0
read_name_1  100  0  100  +  <15216<15215<15214<15213  128  28  127  100  100  0
read_name_3  100  0  100  +  >18612>18613>18614>18615  128  15  114  100  100  0
read_name_8  100  0  100  +  >19602>19603>19604>19605  128  11  110  100  100  0
read_name_4  100  0  100  +  <19770<19769<19768<19767  128  8   107  100  100  0
```

Commands

```
> vg gamsort -G reads.gaf.gz | bgzip > reads.sorted.gaf.gz
> tabix -p gaf reads.sorted.gaf.gz
> vg find -F reads.sorted.gaf.gz -o 200-300
> vg chunk -x graph.vg -F -a reads.sorted.gaf.gz -p chr22:1000-1200 -c 10
```

Annotation with GAF files

Methods

Look for path in pangenome graph and create an "alignment" record. If path is broken in pieces, break the region in multiple alignments

Record the path name from BED 4th column or GFF Name.

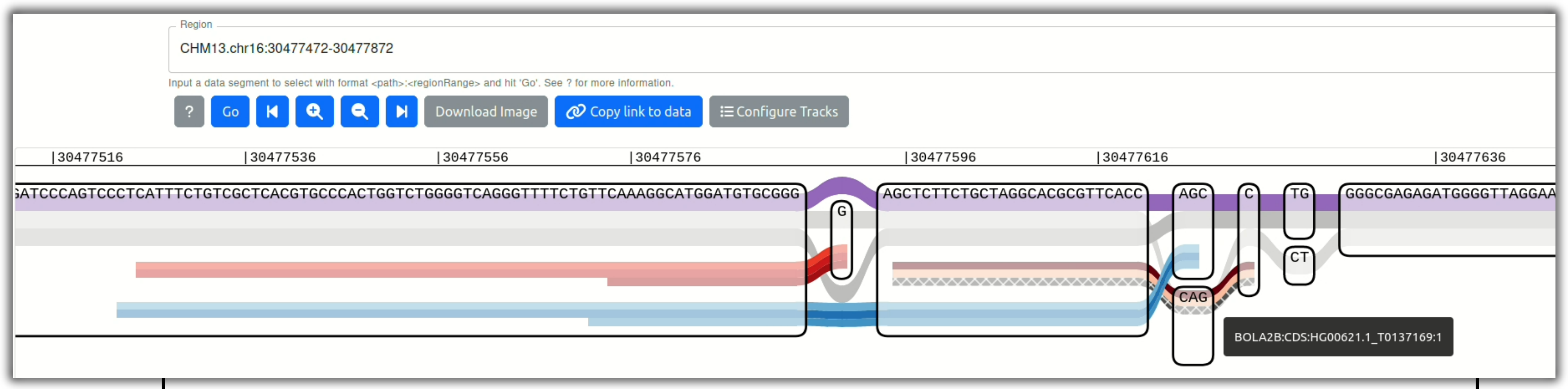
Commands

```
> vg annotate -x graph.gbz -b anno.bed | \
     vg convert -G - graph.gbz | gzip > anno.gaf.gz
> vg annotate -x graph.gbz -f anno.gff | \
     vg convert -G - graph.gbz | gzip > anno.gaf.gz
> vg gamsort -t 1 -pG anno.gaf.gz | bgzip > anno.sorted.gaf.gz
> tabix -p gaf anno.sorted.gaf.gz
```

HPRC pangenome annotation

Annotate genes, segmental duplications, tandem repeats and repeats annotations from the HPRC freeze 1 into the CHM13-based Minigraph-Cactus pangenome.

On average per haplotype, `vg annotate` ~4M gene annotations in ~16 mins, and ~5.5M repeats from RepeatMasker in ~9 mins.

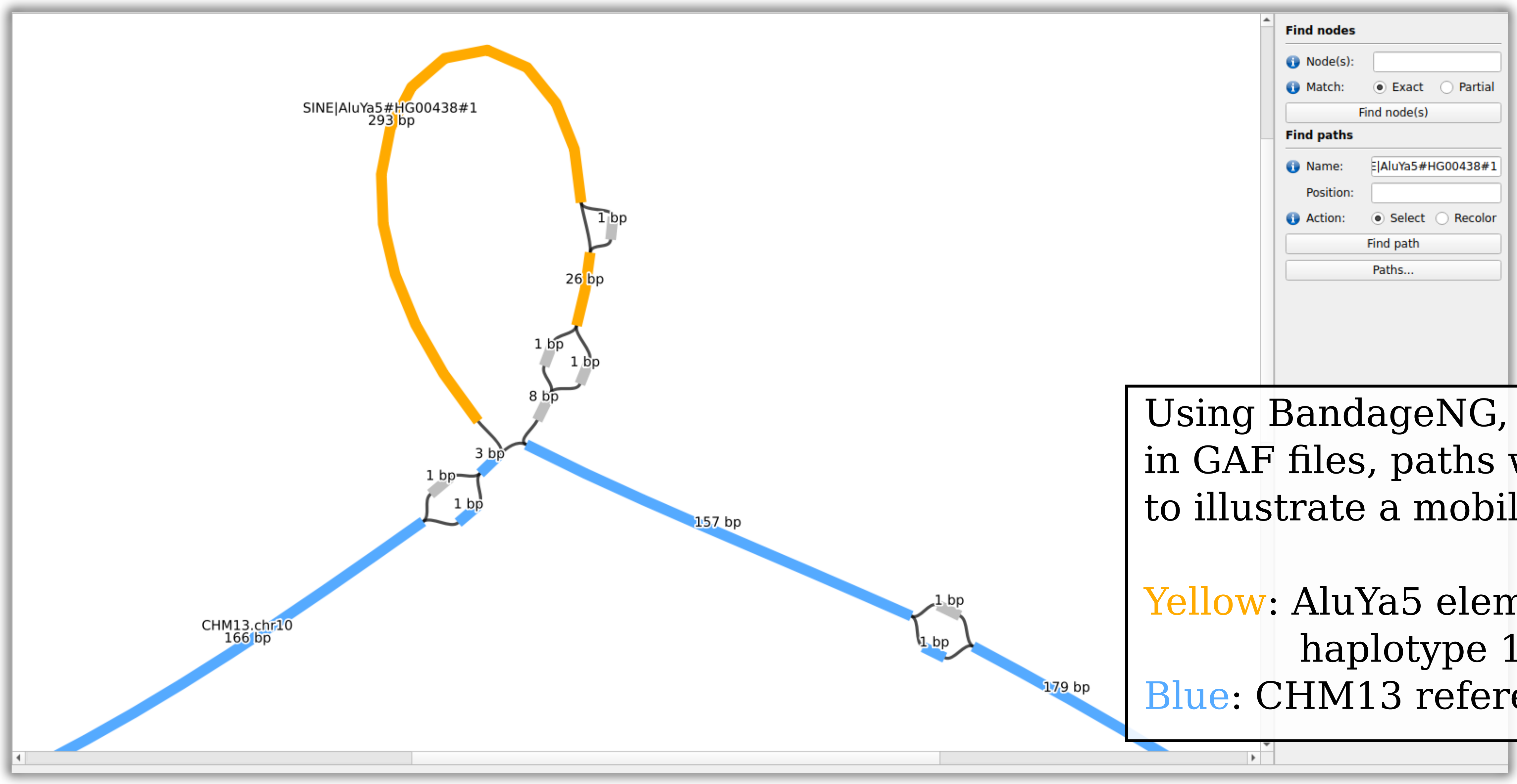


Region: CHM13.chr16:30477472-30477872

Input a data segment to select with format <path>:<regionRange> and hit 'Go'. See ? for more information.

Using sequenceTubeMap, haplotypes, read alignments and paths can be visualized interactively. Hovering on a path displays its name, here the ID of a coding region of the BOLA2B gene.

Haplotypes: CHM13 (purple), HG00621 (greys). Annotated CDS for HG00621 hap 1 (reds) and 2 (blues).



Using BandageNG, a fork that can import paths in GAF files, paths were searched and colored to illustrate a mobile element insertion.

Yellow: AluYa5 element annotated in the haplotype 1 of HG00438.
Blue: CHM13 reference path.

Limitations

- No metadata recorded, all in one path name.
- Simplistic handling of clipped paths.
- Optimized for short paths/ranges.
- Requires ordered integer node IDs for best performance

Links

Code and tutorial: `https://github.com/jmonlong/HPRC2023-gaf-annotation`

vg `https://github.com/vgteam/vg` ("gafidx" branch)
docker: `quay.io/jmonlong/vg:gafidx`

sequenceTubeMap `https://github.com/vgteam/sequenceTubemap`
docker: `quay.io/jmonlong/sequencetubemap:gaf`

Bandage NG v2022.09
`https://github.com/asl/BandageNG`

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute — Inserm La science pour la santé From science to health — irsd

Contact: `jean.monlong@inserm.fr`