# Accurate short variant calling from sequencing data with pangenomes and DeepVariant

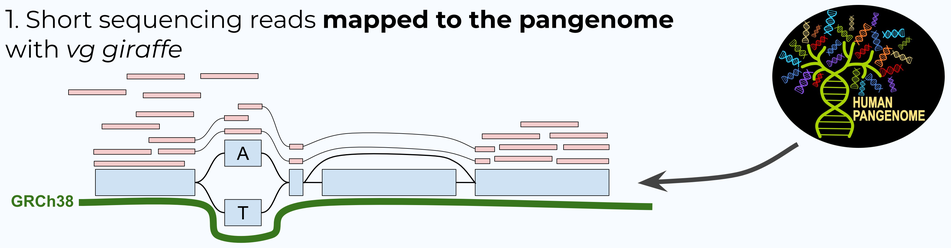
Jean Monlong, Adam M. Novak, Charles Markello, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, Benedict Paten, Human Pangenome Reference Consortium

## Background

- Accurate identification of genomic variants is critical for genetic studies and for clinical genetic testing
- Some genomics regions and variants are still challenging to detect with the standard techniques.
- The Human Pangenome Reference Consortium (HPRC) aims at producing **high-quality phased de novo assemblies** for more than 300 diverse individuals, and building a **comprehensive and representative human pangenome**

## Methods

1. Short sequencing reads **mapped to the pangenome** with *vg giraffe*



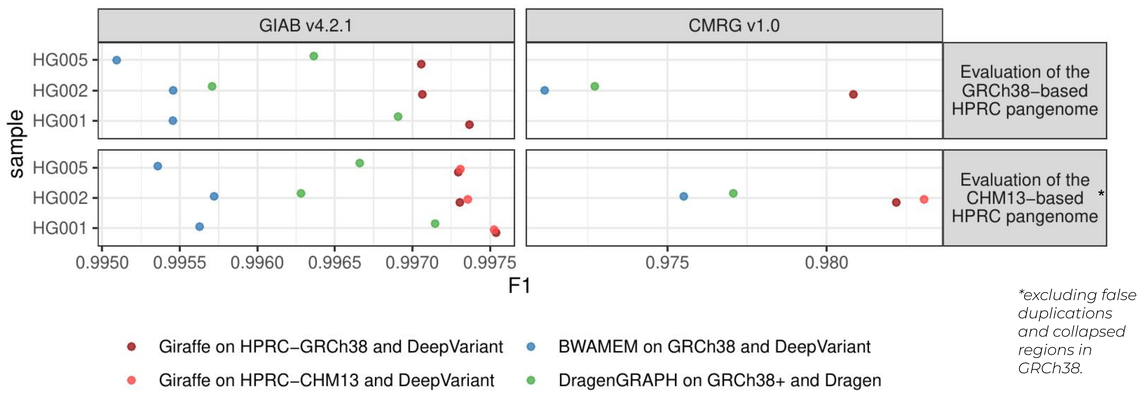

HPRC draft pangenome v1.0 (90 haplotypes)

2. Aligned reads **projected to GRCh38** with *vg surject*


3. **Variants called with DeepVariant** trained on surjected alignments, after indel realignment.

## The HPRC pangenome improves the mapping of short reads and, as a result, the identification of short variants.

Evaluation against the Genome in a Bottle (GIAB) and Challenging Medically-Relevant Genes (CMRG) truthsets.

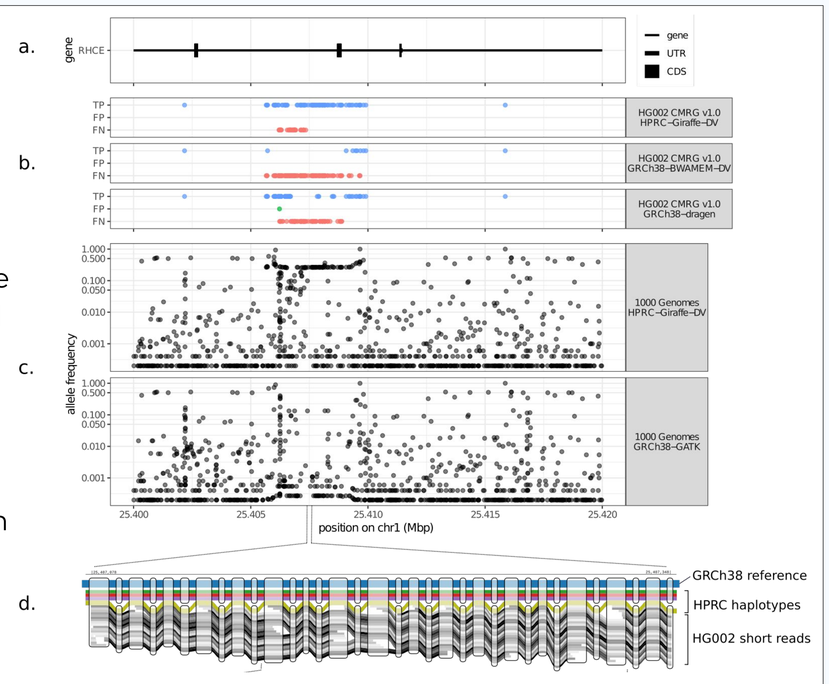



**excluding false duplications and collapsed regions in GRCh38.*

In RHCE (a), our approach can correctly call more variants in a region with gene conversion (b).

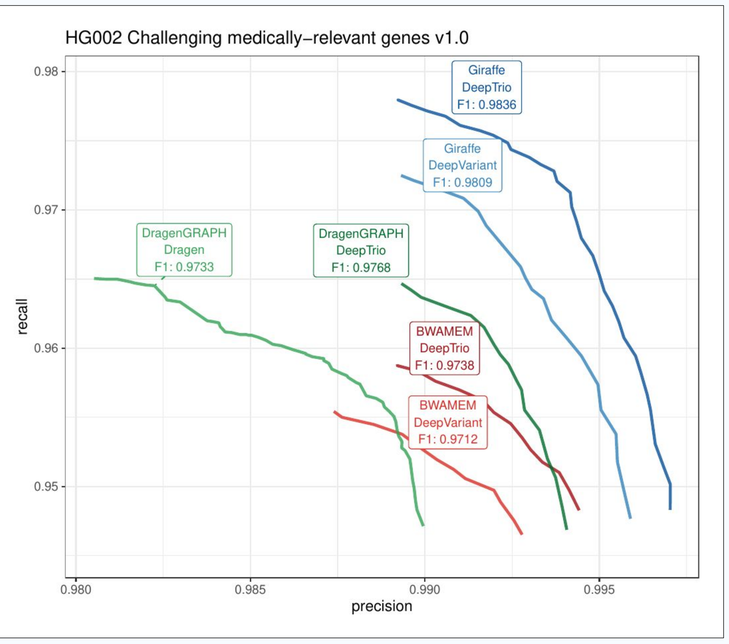
We could derive better allele frequencies from short read datasets in this region (c).

That's because the pangenome contains the gene-converted allele which ensure that reads are mapped to the correct location (d).




Best performance with DeepTrio trio-based calling

In challenging regions, single-sample pangenomic calling performed better than trio-based standard approaches




## Conclusion

- Our pangenomic approach results in about 34% less errors, and calls short variants in a larger fraction of the genome, including in more challenging regions.
- Resources:
  - *A Draft Human Pangenome Reference*. bioRxiv 2022 DOI:10.1101/2022.07.09.499321
  - HPRC pangenomes: https://github.com/human-pangenomics/hpp_pangenome_resources
  - Workflows (WDL) DOI: 10.5281/zenodo.6655968

This work was funded by National Human Genome Research Institute of the National Institutes of Health under Award Numbers U01HG010961, U41HG010972, R01HG010485, U24HG011853.