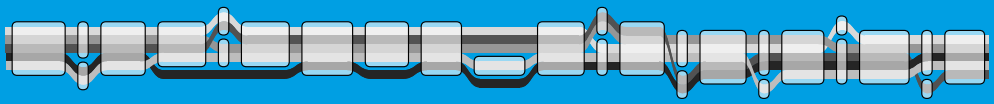


Pangenome graph construction from genome alignments with Minigraph-Cactus

Jean Monlong

JOBIM HIGHLIGHT
26/06/2024

irsd
Institut de recherche en santé digestive

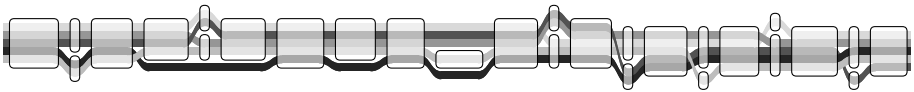
Inserm
La science pour la santé
From science to health

Outline

Introduction - Genomes and pangenomes

Methods - Building a pangenome from assemblies

Results - Human pangenome reference analysis

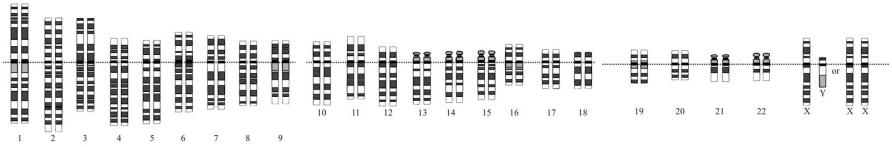


Highlight of *Hickey, Monlong et al. Nature Biotech 2023*

Introduction - Genomes and pangenomes

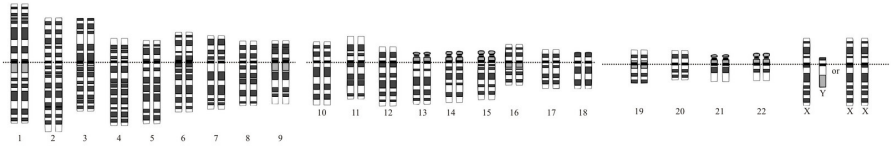
The human reference genome

One copy of the human genome produced by a worldwide effort that took more than 10 years (2.7 billion dollars).



The human reference genome

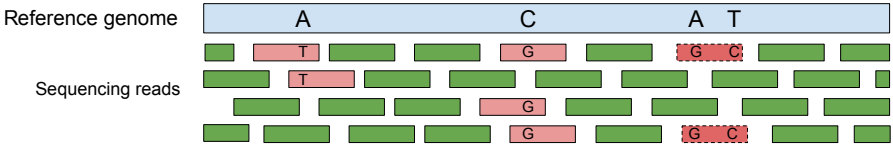
One copy of the human genome produced by a worldwide effort that took more than 10 years (2.7 billion dollars).



Limitations

1. Not a real haplotype. Composite from several different anonymous individuals.
2. Not complete. Mega-bases of gaps, i.e. missing sequence.
3. Just one copy. Not representative of the human genetic diversity.

Aligning sequencing reads to a reference genome

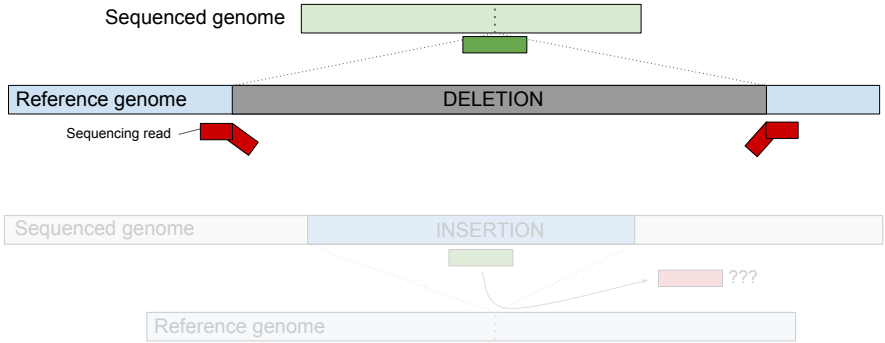


Assuming the reads are correctly placed, we can

- find genomic variants
- quantify gene/transcript expression
- identify functional regions

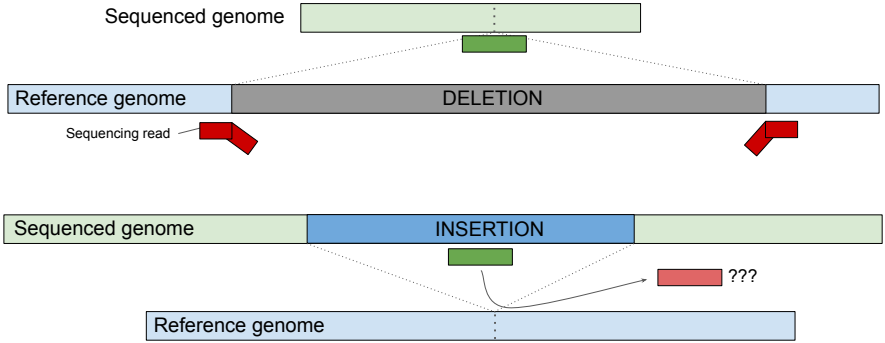
The challenges of structural variant detection

Structural variants (SVs) involve 50 bp or more.



The challenges of structural variant detection

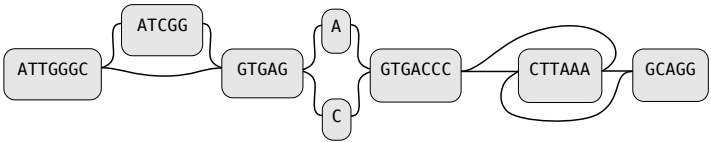
Structural variants (SVs) involve 50 bp or more.



Pangenomes represent genetic diversity succinctly

A pangenome represents a **collection of genomes** and the genetic variants among them.

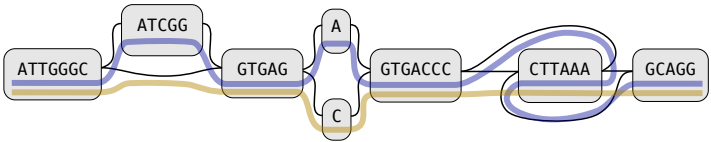
ATTGGGC**ATCGG**GTGAG**A**GTGACCC**TTTAAG**GCAGG
ATTGGGC**-----**GTGAG**C**GTGACCC**CTTAAA**GCAGG



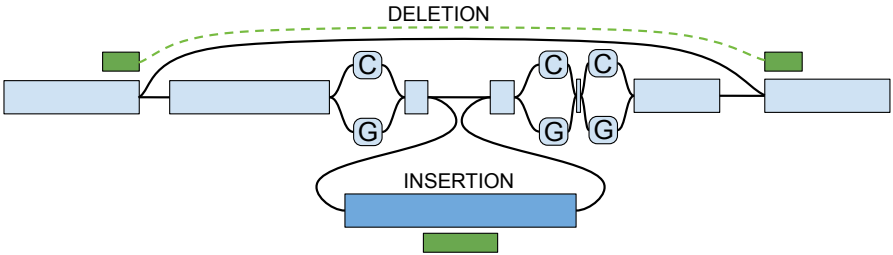
Pangenomes represent genetic diversity succinctly

A pangenome represents a **collection of genomes** and the genetic variants among them.

ATTGGGC**ATCGG**GTGAG**A**GTGACCC**TTTAAG**GCAGG
ATTGGGC**-----**GTGAG**C**GTGACCC**CTTAAA**GCAGG



Aligning reads to a reference pangenome



https://github.com/vgteam/vg


Garrison, et al. Nature Biotech 2018

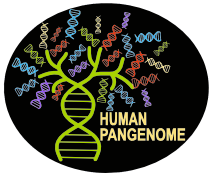
Grytten, Rand, et al. PLoS Comput Biol 2019

Hickey, Heller, Monlong, et al. Genome Biology 2020

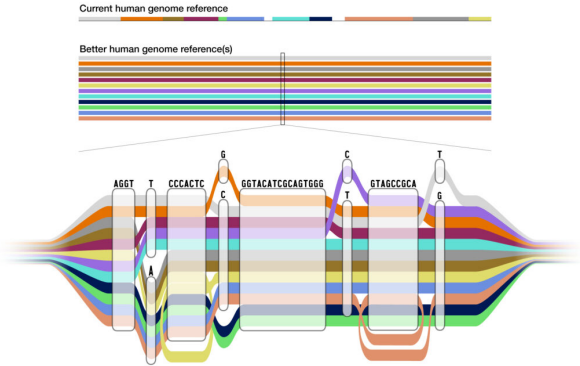

Siren, Monlong, Chang, Novak, Eizenga, et al. Science 2021

Sibbesen, Eizenga, et al. Nature Methods 2023

Building a Human pangenome reference



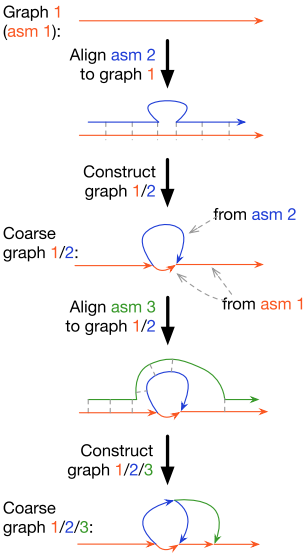
- Human Pangenome Reference Consortium (HPRC)
- Latest sequencing technologies for 350 diverse individuals
- Pangenome containing a comprehensive catalog of variants



Liao, Asri, Ebler, et al. Nature 2023

Methods - Building a pangenome from assemblies

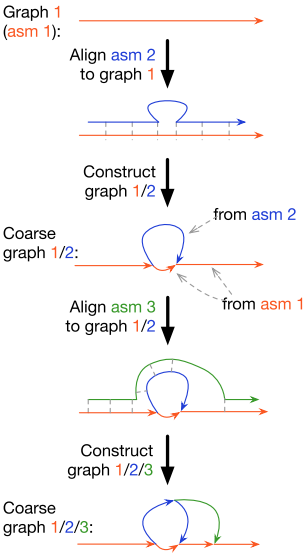
Minigraph efficiently builds a SV graph



Methods

Add each genome/assembly iteratively but only structural variation (50 bp or more).

Minigraph efficiently builds a SV graph



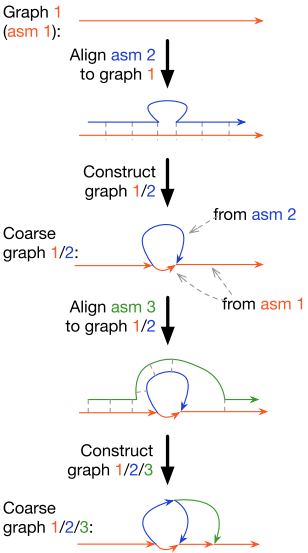
Methods

Add each genome/assembly iteratively but only structural variation (50 bp or more).

Limitation

- Dependent on the order of input genomes.
- Small variation absent.

Minigraph efficiently builds a SV graph



Methods

Add each genome/assembly iteratively but only structural variation (50 bp or more).

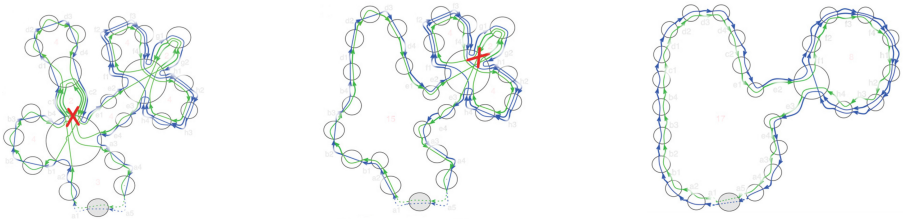
Limitation

- Dependent on the order of input genomes.
- Small variation absent.

→ Cactus to add back small variation and genomes losslessly embedded as paths.

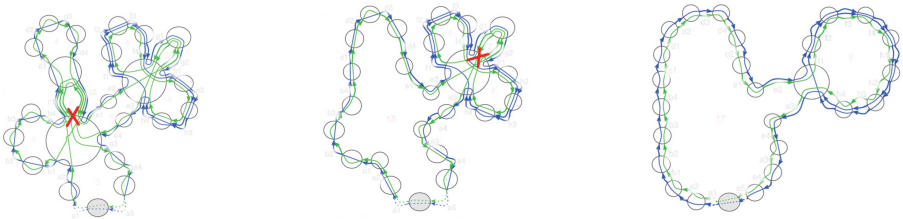
Li et al. Genome Biology 2020

Cactus organizes an alignment graph in independent blocks



Paten et al. Genome Research 2011

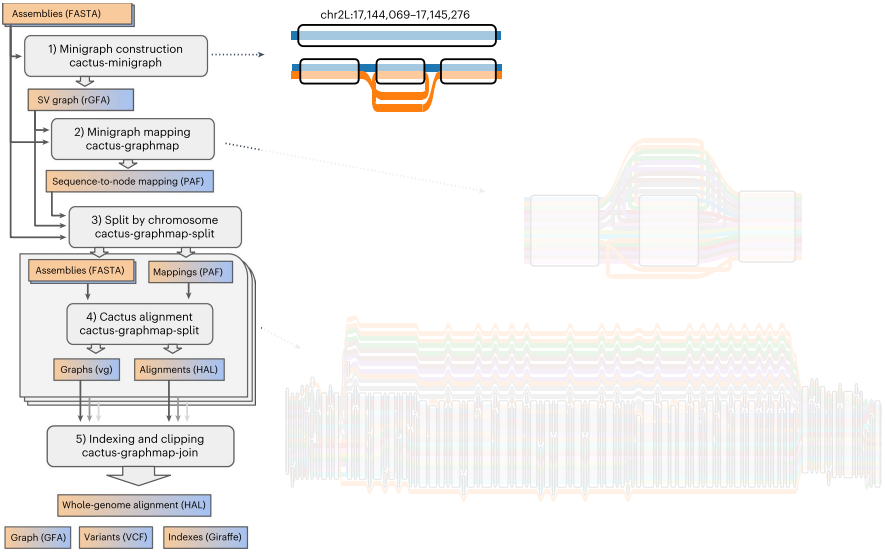
Cactus organizes an alignment graph in independent blocks



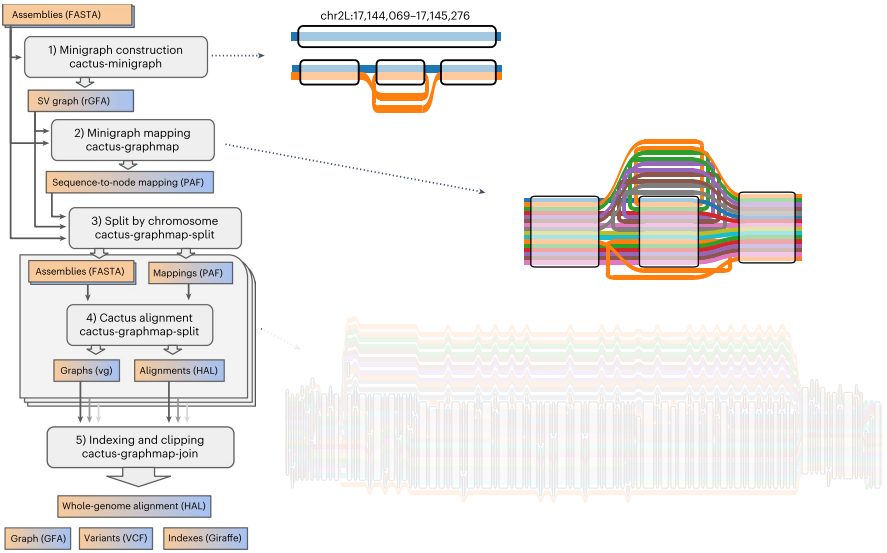
Paten et al. Genome Research 2011

- **Cactus Alignment Filter (CAF)** iteratively extends larger chains.
- **Base-level alignment refinement (BAR)** re-aligns unaligned sequences.

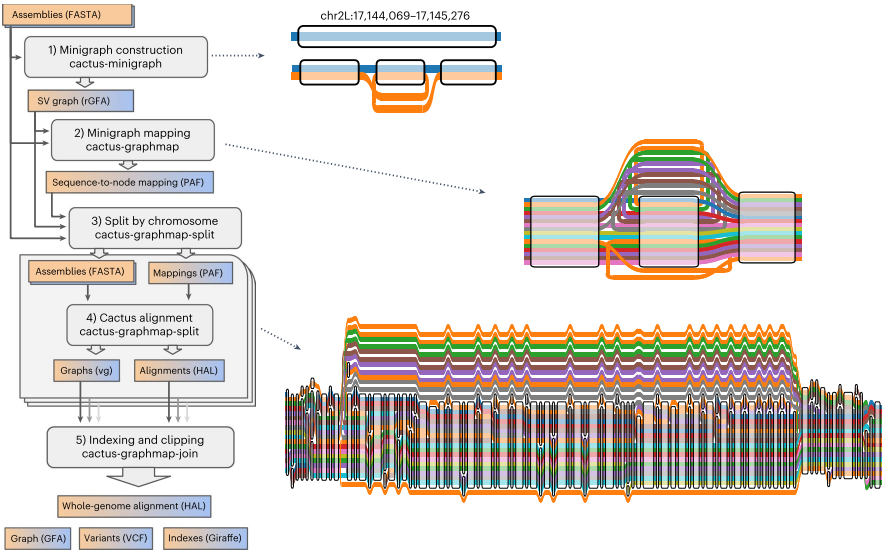
The Minigraph-Cactus pipeline



The Minigraph-Cactus pipeline



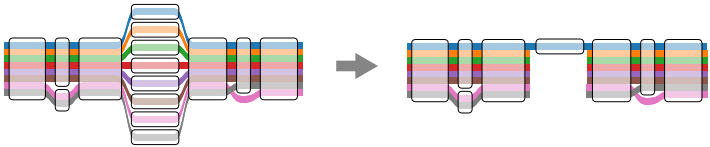
The Minigraph-Cactus pipeline



Post-alignment refinements

Clip out regions $\geq$ 10 Kbp that don't align to the SV graph (e.g. centromeres, satellites arrays).

Clipping of unaligned sequences



Removal of redundant sequences with GFAffix

https://github.com/marschall-lab/GFAffix

Results - Human pangenome reference analysis

Four Minigraph-Cactus human pangenomes

Pangenomes built from 90 human haplotypes from HPRC freeze 1.

- **Two "backbones"**: GRCh38 or CHM13 (recent complete human genome).

Four Minigraph-Cactus human pangenomes

Pangenomes built from 90 human haplotypes from HPRC freeze 1.

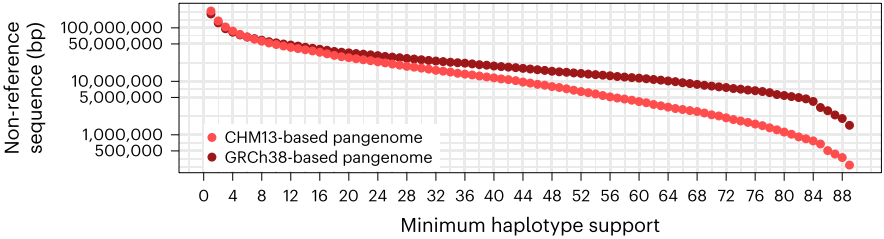
- **Two "backbones"**: GRCh38 or CHM13 (recent complete human genome).
- **Two sets of filters**
  - full pangenome
  - frequency filtered (keep $\geq$ 10% frequency)

Four Minigraph-Cactus human pangenomes

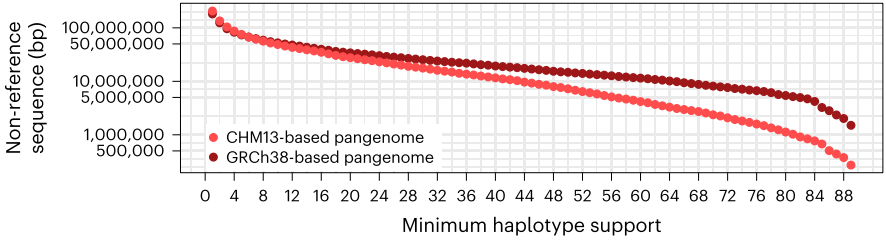
Pangenomes built from 90 human haplotypes from HPRC freeze 1.

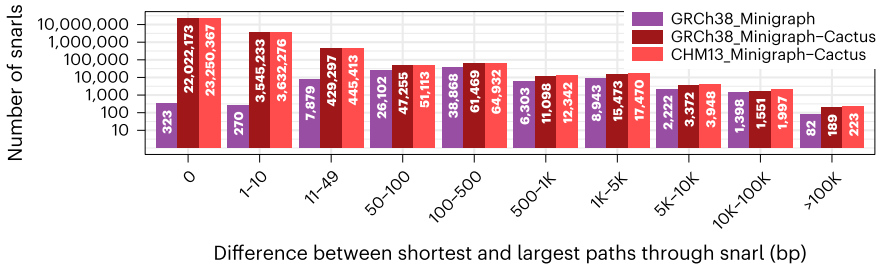
- **Two "backbones"**: GRCh38 or CHM13 (recent complete human genome).
- **Two sets of filters**
  - full pangenome
  - frequency filtered (keep $\geq$ 10% frequency)
- Most of our mapping/variants calling analysis used the **frequency-filtered GRCh38-based pangenome**.

Amount of non-reference sequence and variation sites

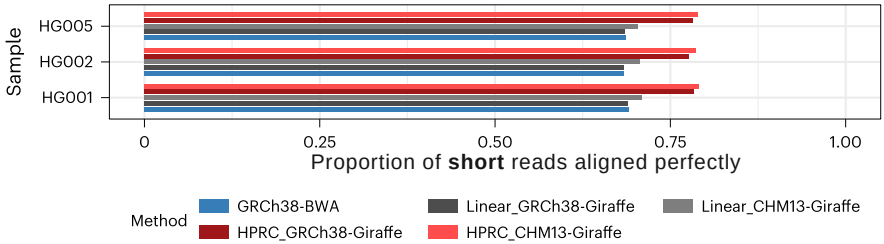


Amount of non-reference sequence and variation sites

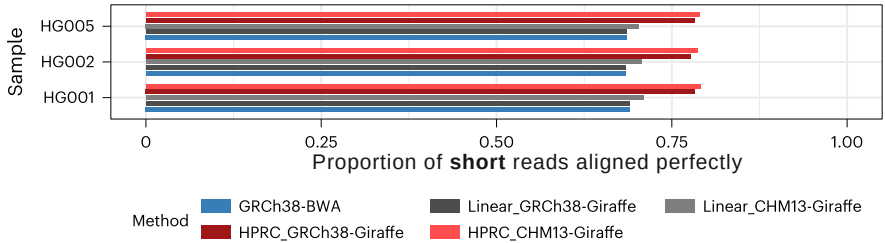


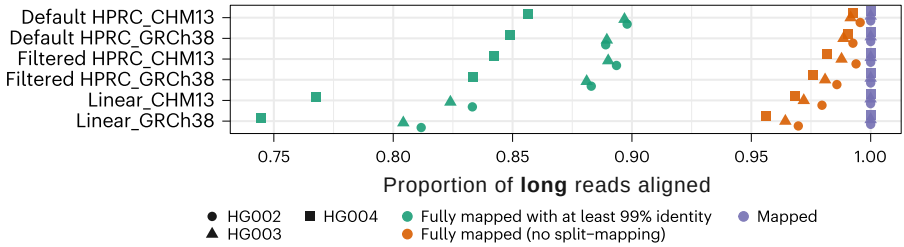


Both short and long sequencing reads align better

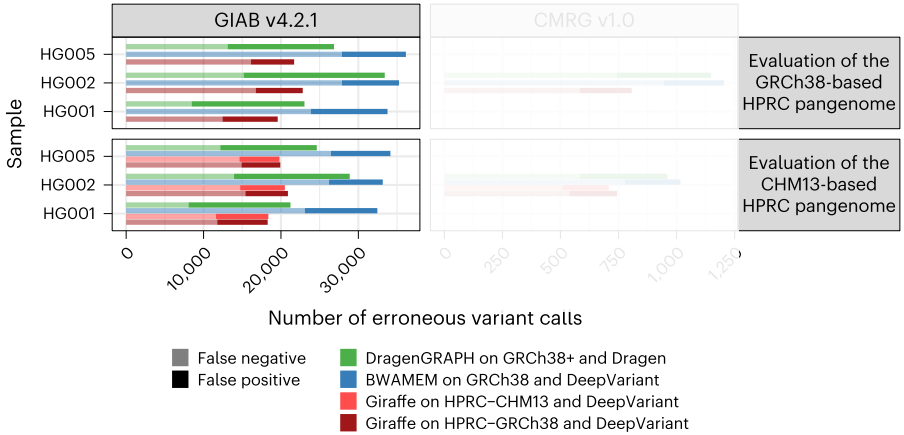


Both short and long sequencing reads align better



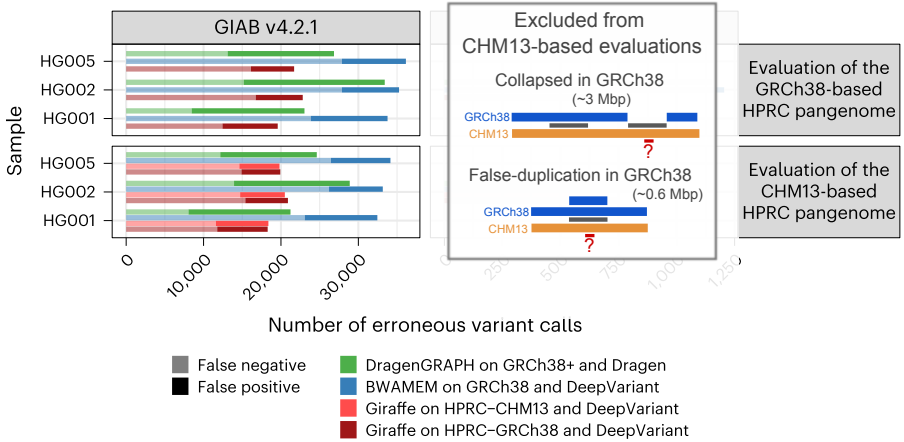


Less errors when calling small variants (SNPs/indels)



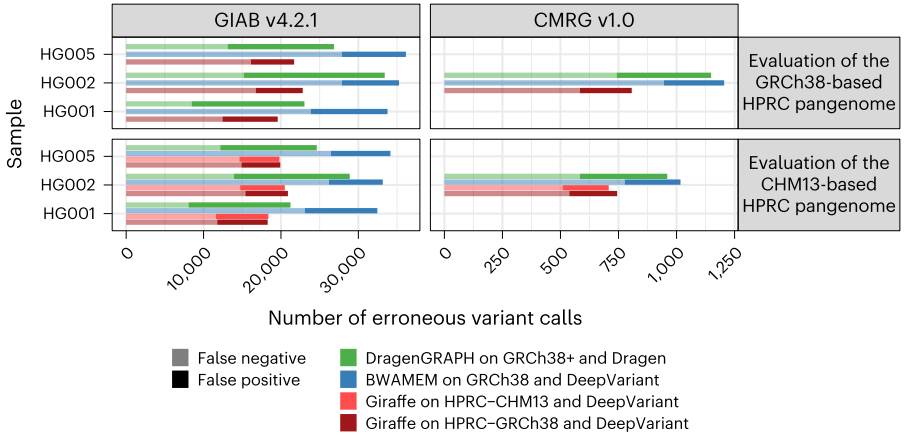
CMRG: Challenging Medically-Relevant Genes truthset from the Genome in a Bottle (GIAB).

Less errors when calling small variants (SNPs/indels)



CMRG: Challenging Medically-Relevant Genes truthset from the Genome in a Bottle (GIAB).

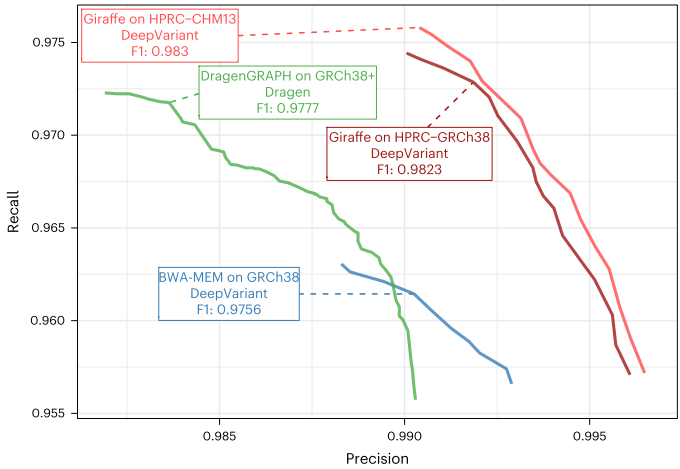
Less errors when calling small variants (SNPs/indels)



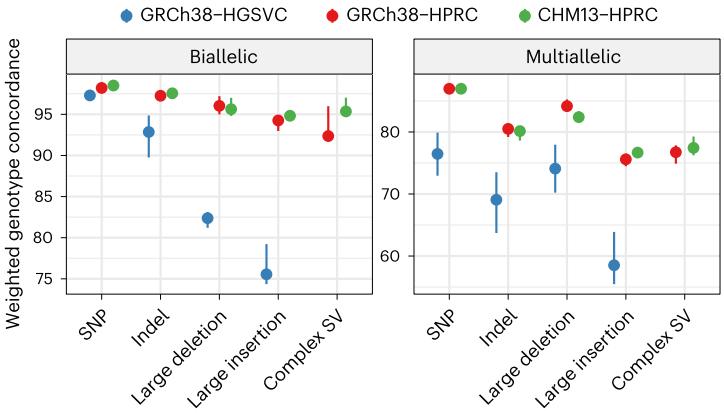
CMRG: Challenging Medically-Relevant Genes truthset from the Genome in a Bottle (GIAB).

CHM13-based pangenome superior in challenging regions

SNPs/indels calling in Challenging Medically-Relevant Genes (CMRG) truthset.



Better variant genotyping with PanGenie

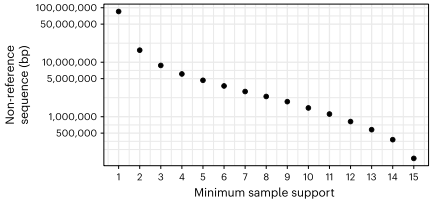


The *GRCh38-HGSVC* pangenome was built by combining variant calls (VCFs).

`https://github.com/eblerjana/PanGenie` Ebler, et al. Nature Genetics 2022

Application to Drosophila

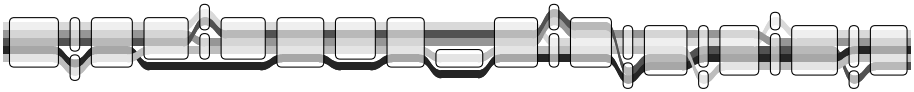
- Pangenome from 16 drosophila assemblies.
- Experiment with short reads across 100 individuals.



Conclusion

Minigraph-Cactus builds pangenomes that **fully represent the input genomes** and serves as a **better reference** for sequencing data analysis.

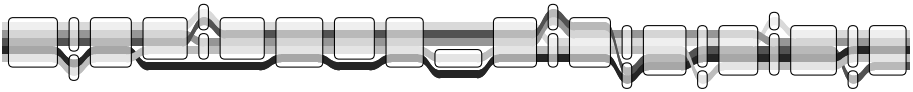
`https://github.com/ComparativeGenomicsToolkit/cactus`



Conclusion

Minigraph-Cactus builds pangenomes that **fully represent the input genomes** and serves as a **better reference** for sequencing data analysis.

`https://github.com/ComparativeGenomicsToolkit/cactus`



Limitations

- Centromeres and clipped regions
  - ⇒ *Method for centromere alignment in preparation at UCSC.*
- Crude frequency-filtering.
  - ⇒ *Personalized Pangenome References. Sirén et al. bioRxiv 2023*
- Unsupported interchromosomal events (e.g. cancer).

Acknowledgments

University of California, Santa Cruz

- **Glenn Hickey**
- **Benedict Paten**
- Adam Novak
- Jordan Eizenga
- Xian Chang
- Jouni Sirén

IMBM Düsseldorf

- Jana Ebler
- Tobias Marschall

Dana-Farber Cancer Institute

- Heng Li

UC SANTA CRUZ
Genomics Institute

HUMAN PANGENOME

vg