Genome variation graphs with the vg toolkit

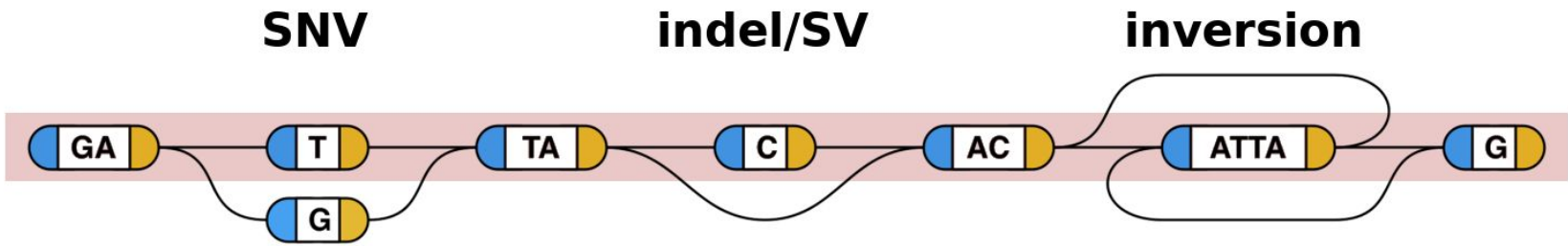
Jean Monlong

Updates from the GRC & GIAB
Oct 15, 2019

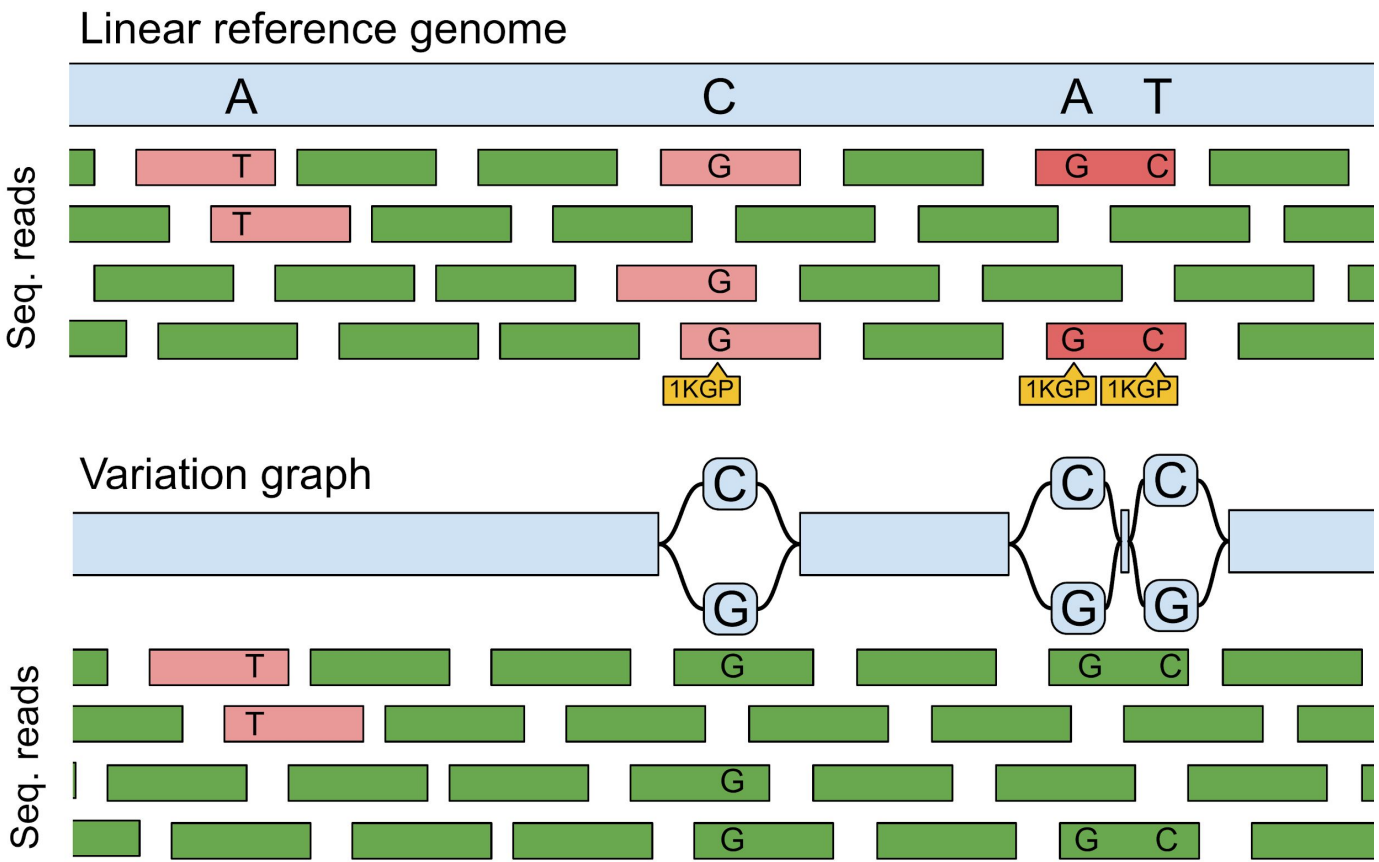
UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

Variation Graphs

An approach to incorporating information on human diversity into the genomic reference.

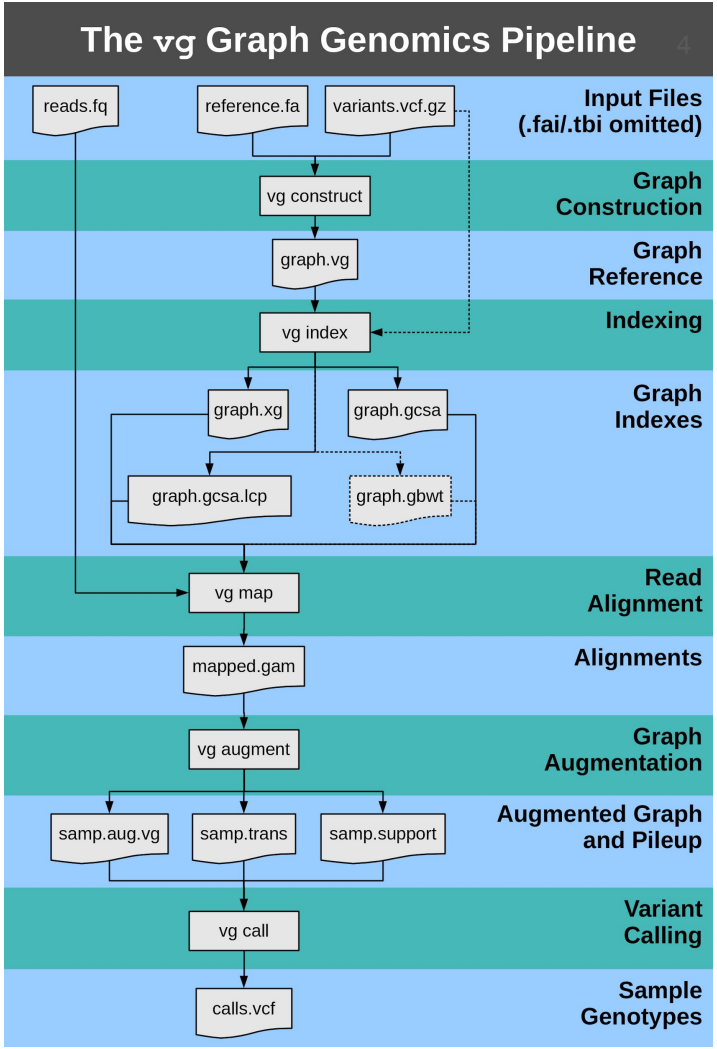


Sequencing reads map better on variation graphs

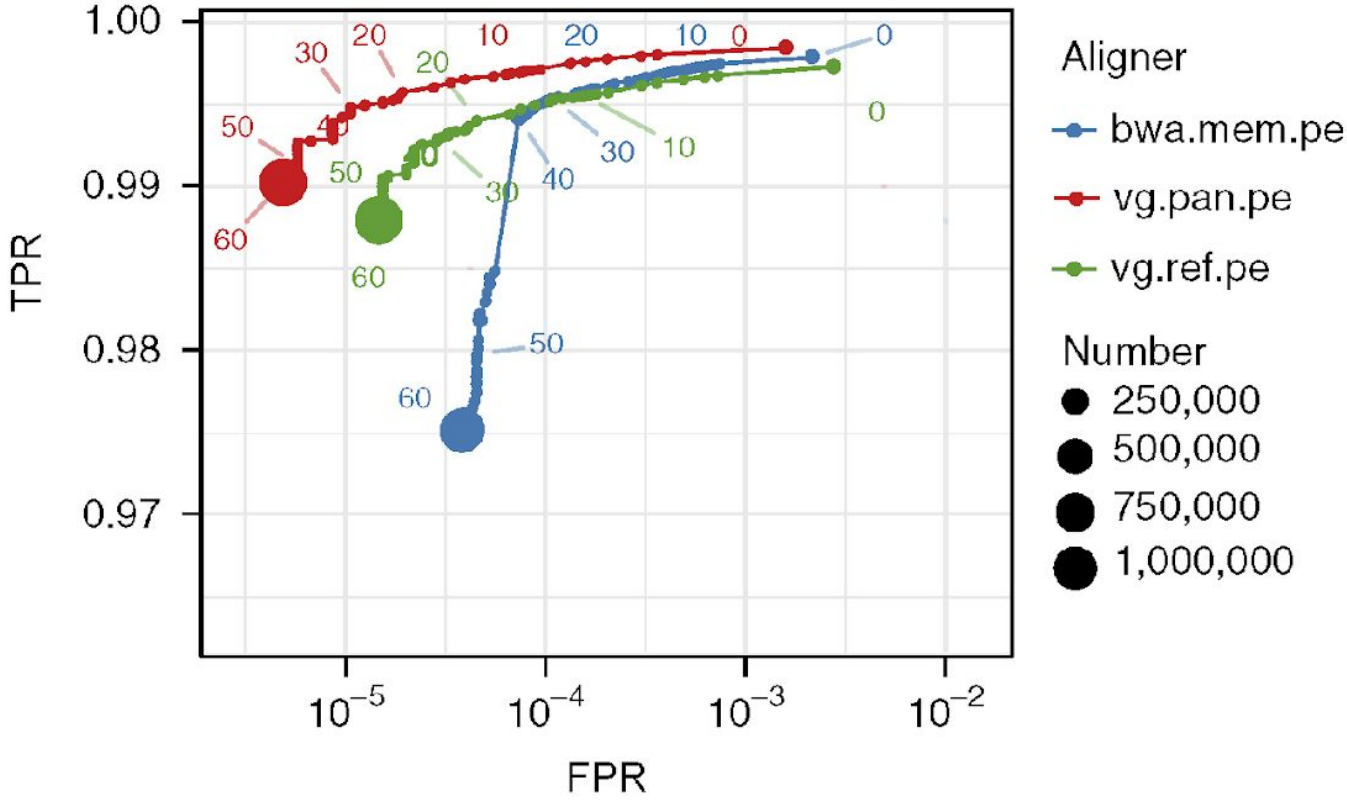


is a complete,
open source solution
for graph construction,
read mapping,
and variant calling.

https://github.com/vgteam/vg

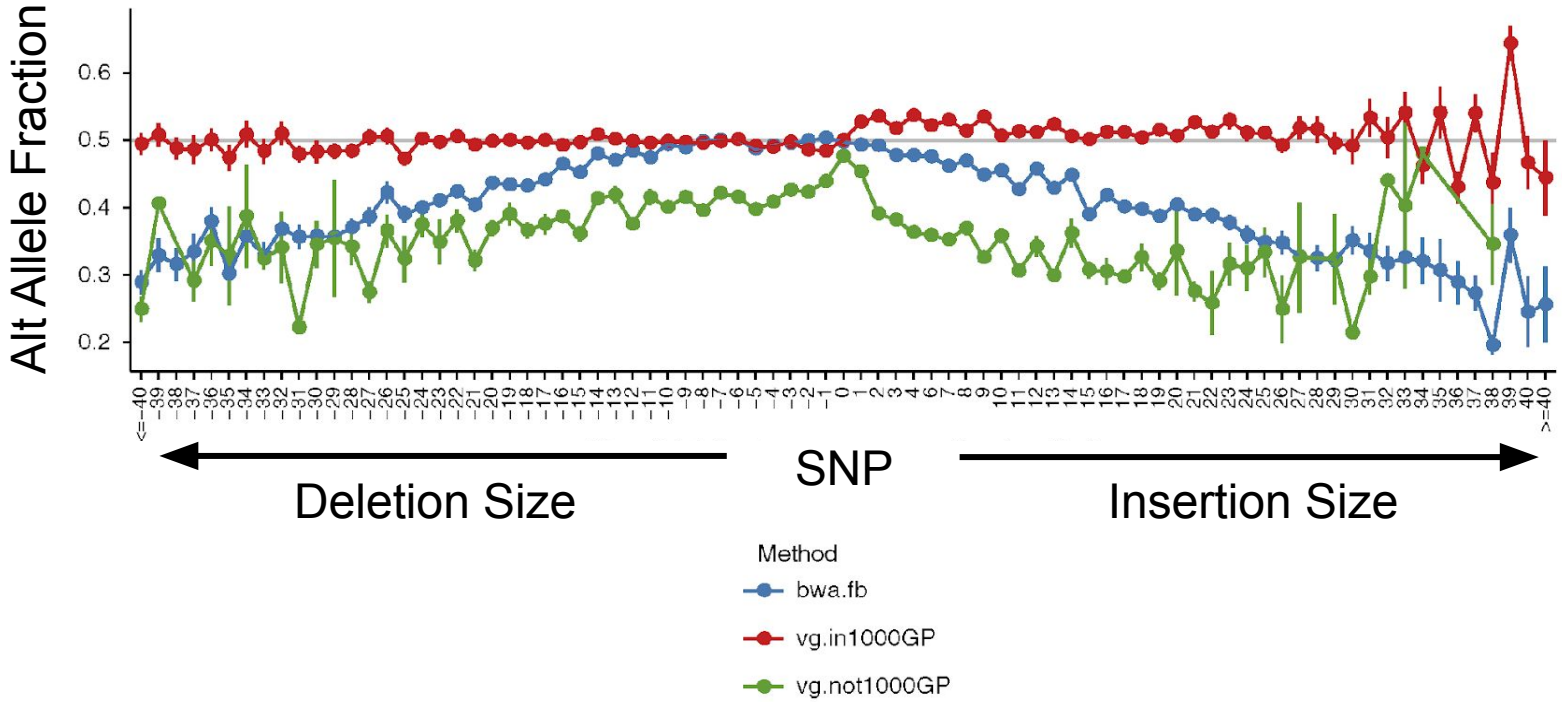


Better read mapping in regions with variants



Garrison et al (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology.

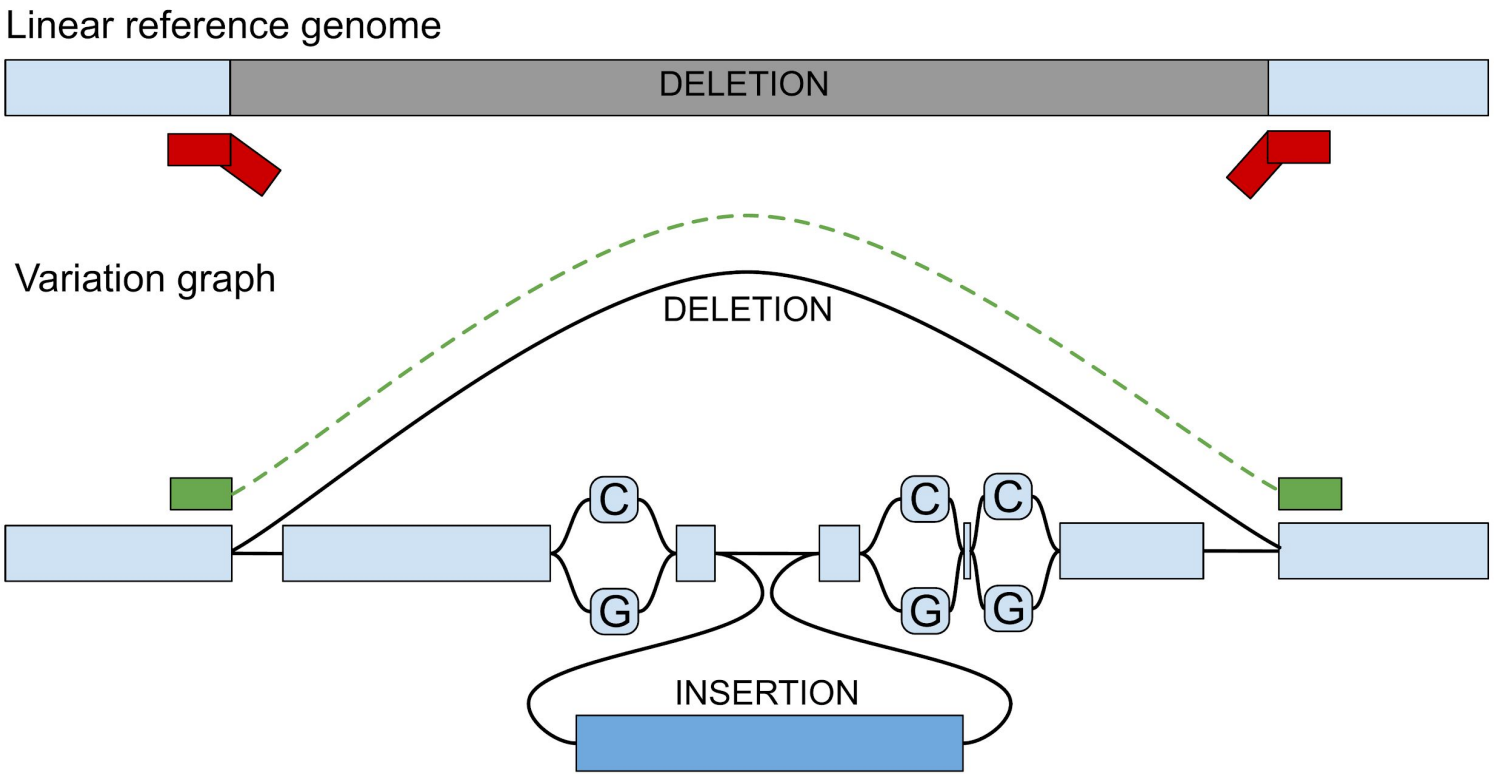
Better allele balances at heterozygous sites



Garrison et al (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology.

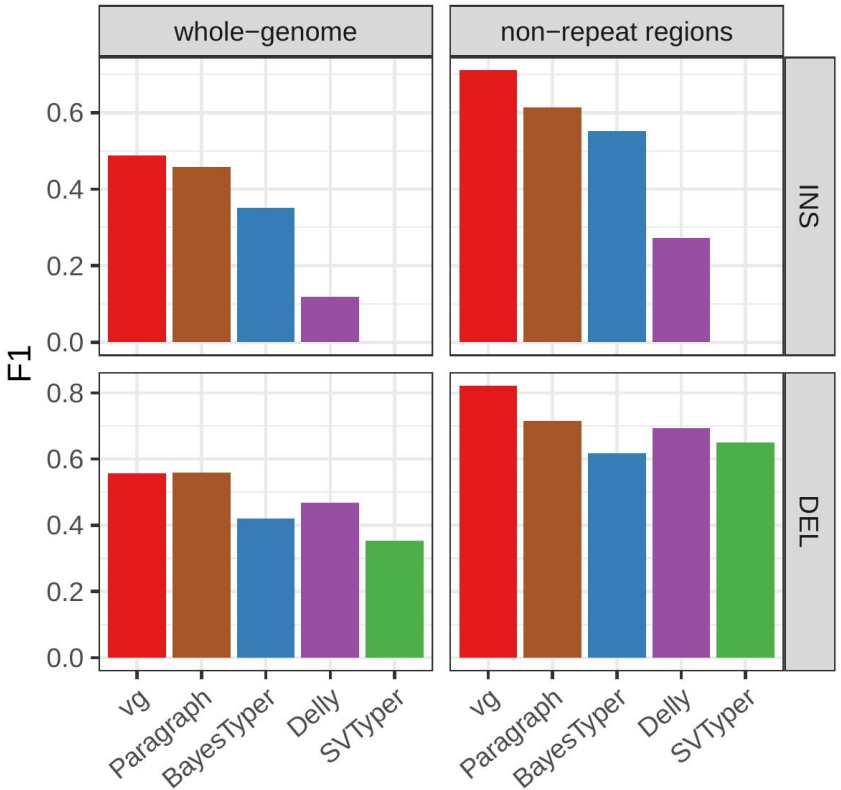
Structural Variants (SVs)

Genomic variants >50bp. E.g. insertions, deletions, inversions.



Genotyping SVs from short-read data

High-quality SV catalogs from long-read sequencing studies
(HGSVC 2019, GIAB 2019, SVPOP 2019).



Graph-based SV genotypers:
vg, Paragraph, BayesTyper

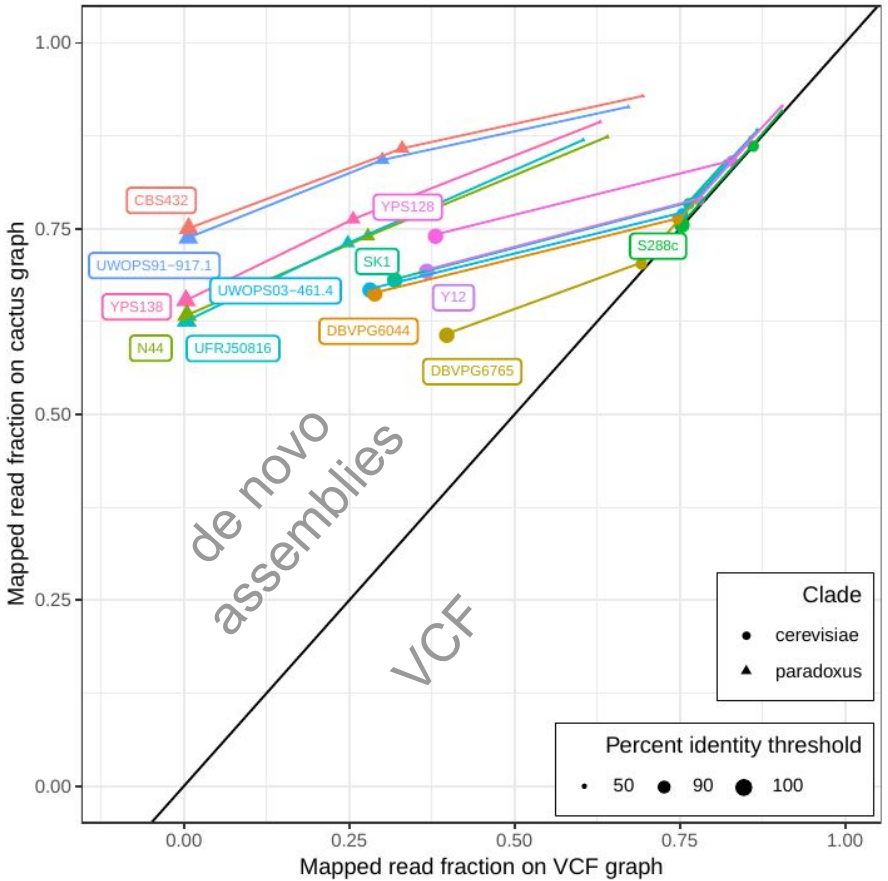
Traditional SV genotypers:
Delly, SVTyper

Hickey et al (2019). Genotyping structural variants in pangenome graphs using the vg toolkit. bioRxiv.

Graph from de novo assemblies

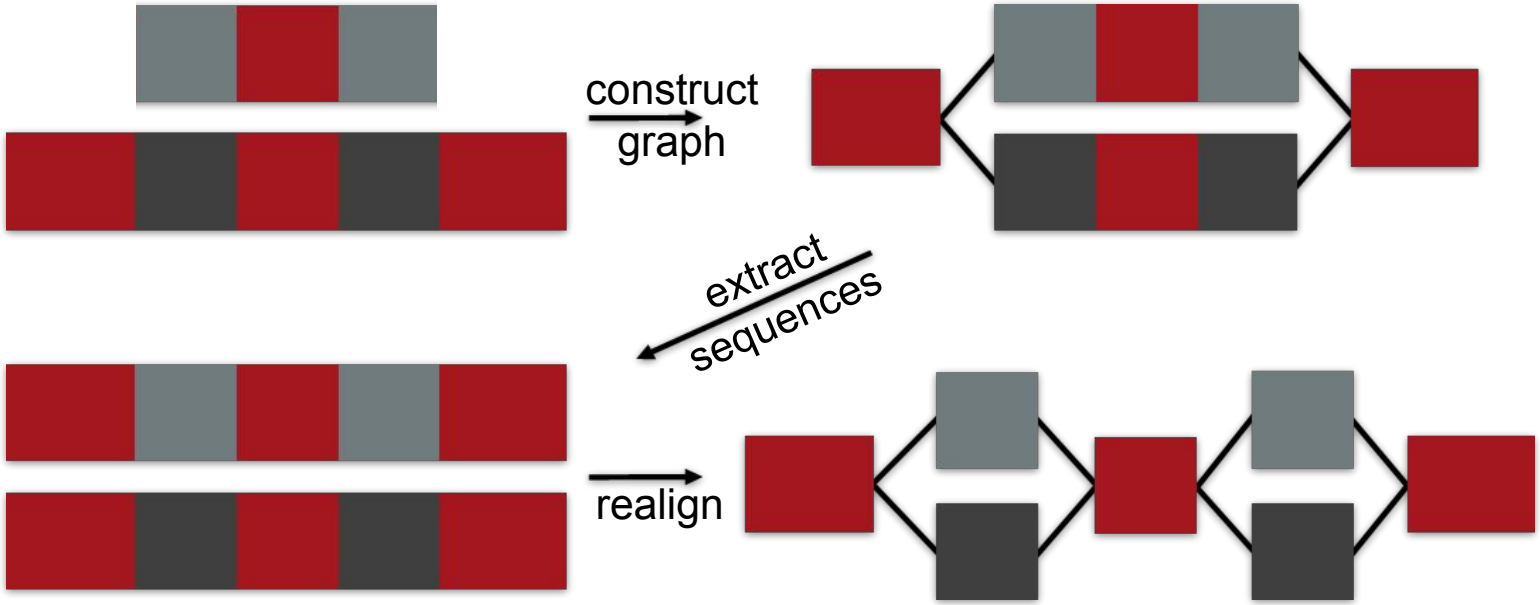
Experiment with 12 yeast strains.

- better read mapping.
- SV better supported by reads.



Hickey et al (2019). Genotyping structural variants in pangenome graphs using the vg toolkit. bioRxiv.

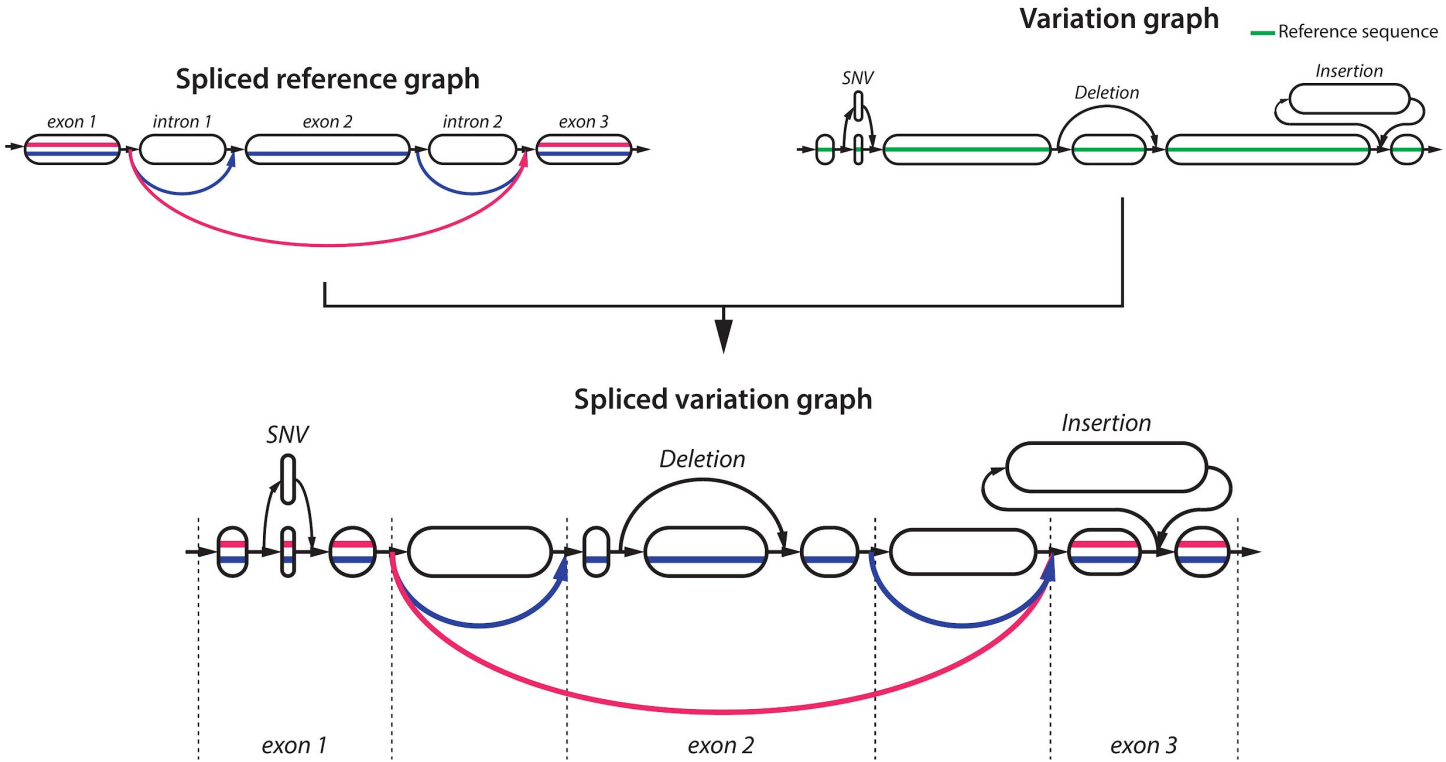
Variant normalization



Robin Rounthwaite

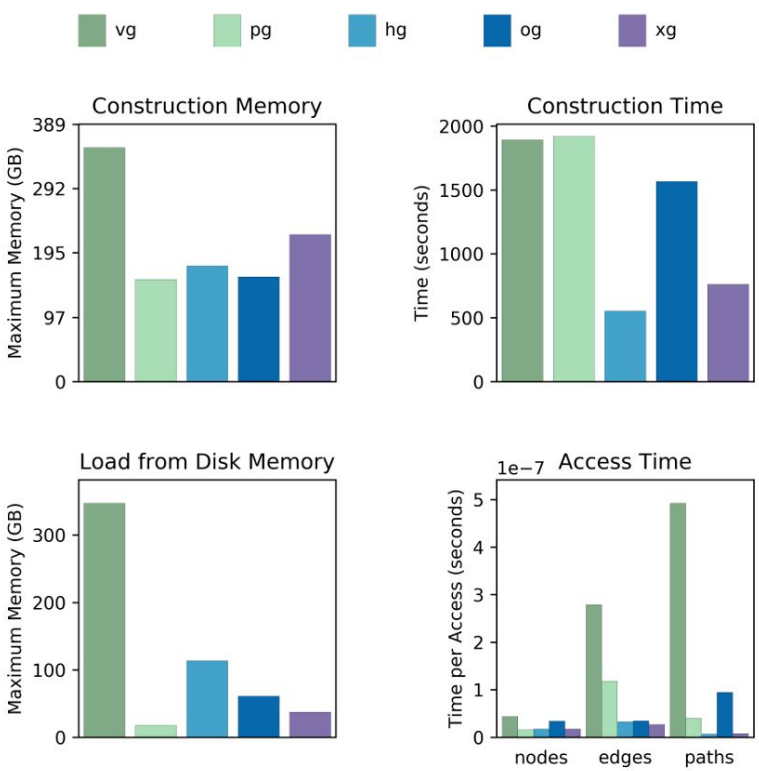
Transcriptome analysis with variation graphs

https://github.com/jonassibbesen/rpvg



New graph formats optimized for memory and speed

https://github.com/vgteam/libbdsg



VG: Legacy

PG (PackedGraph): Small

HG (HashGraph): Fast

OG (Optimized Dynamic Graph Index): Balanced

XG: Small and fast, but static

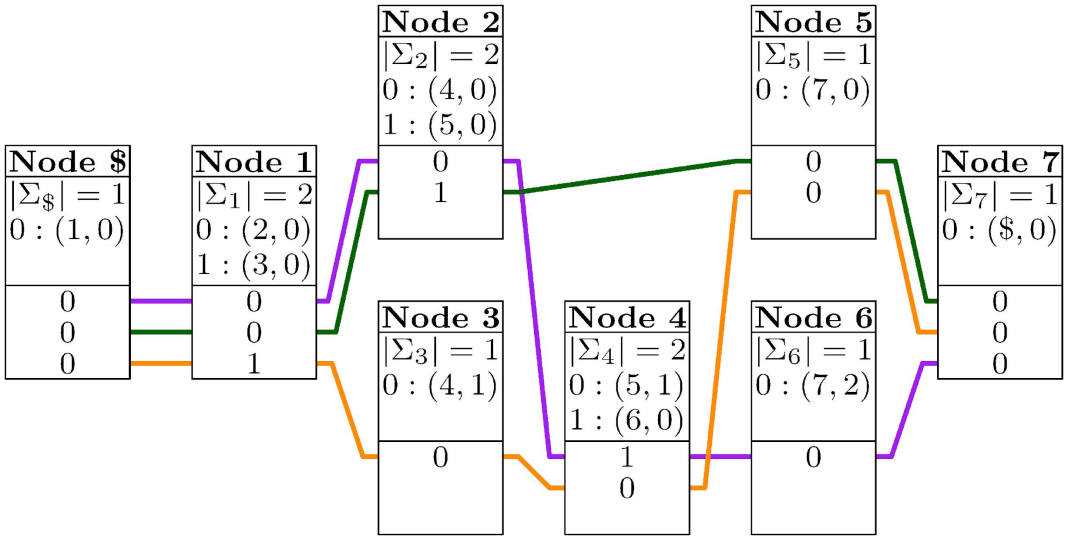
Jordan Eizenga, Emily Fujimoto, Erik Garrison

Graph Burrows-Wheeler Transform

https://github.com/jltsiren/gbwtgraph

Stores 2,504 haplotypes from 1K Genomes Project in 14.6 GiB

Can scale past tens of thousands of haplotypes

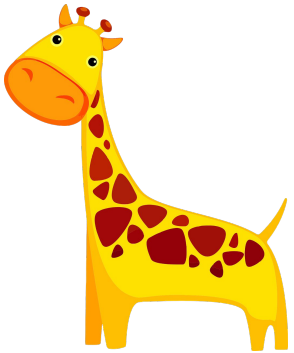


Sirén et al. (2018). Haplotype-aware graph indexes. In 18th International Workshop on Algorithms in Bioinformatics (WABI)

Faster short read mapping with giraffe

Assuming that most indels are in the graph and a large haplotype database.

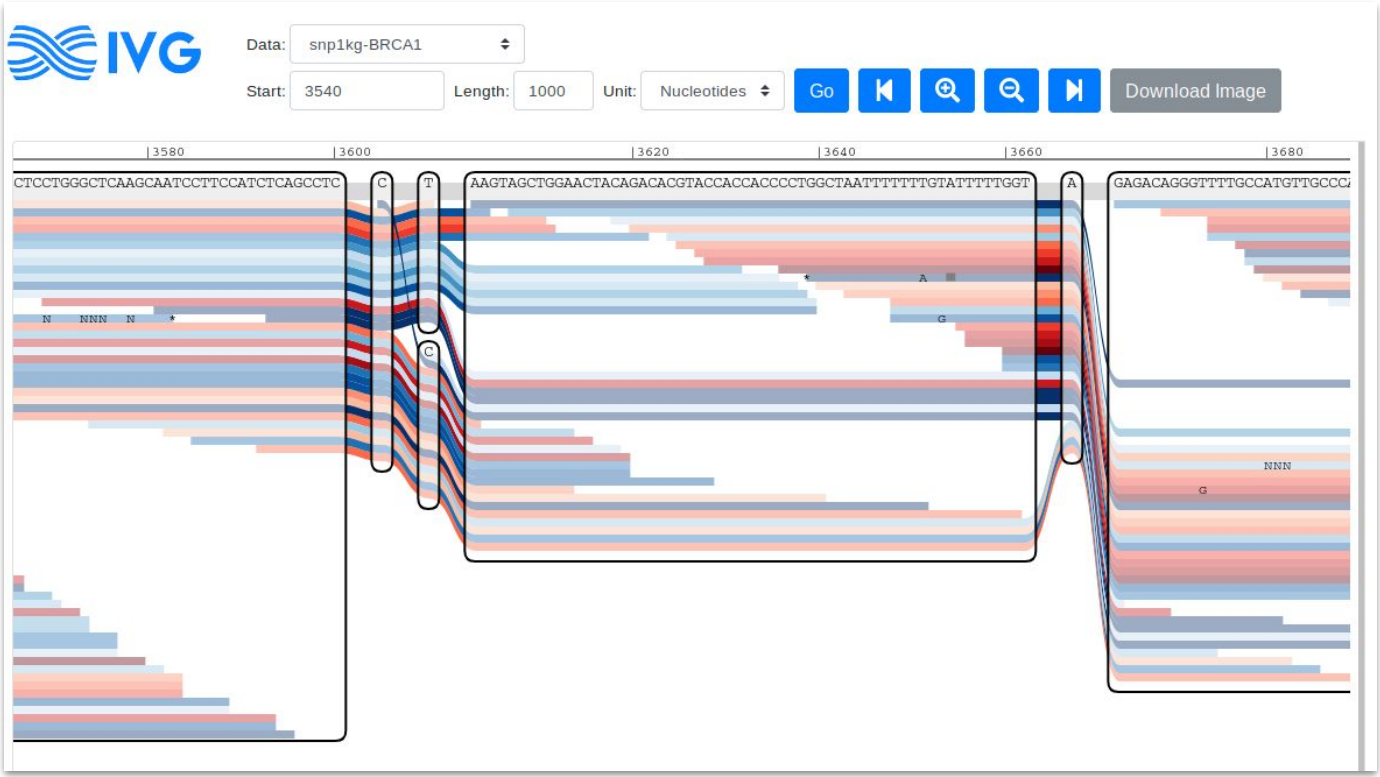
- Minimizer-based indexing.
- Restrict to known haplotypes.
- Gapless extension.
- Faster clustering using a snarl tree.



Xian Chang, Jouni Siren, Adam Novak

Visualization

https://github.com/vgteam/sequenceTubeMap



Beyer et al. (2019). Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics (Oxford, England).

Pipelines in Toil and WDL

https://github.com/vgteam/toil-vg

https://github.com/vgteam/vg_wdl

	toil-vg	WDL
Graph construction	x	
Read mapping	x	x
Variant calling(SNV/indels & SVs)	x	x

Acknowledgements

Benedict Paten
Adam Novak
Erik Garrison
Jordan Eizenga
Glenn Hickey
Jouni Siren
David Heller

Jonas Sibbesen
Xian Chang
Charles Markello
Yohei Rosen
Robin Rounthwaite
Susanna Morin
Emily Fujimoto

https://github.com/vgteam/vg

Eric Dawson
Mike Lin
Wolfgang Beyer

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Richard Durbin
Daniel Zerbino

Check out Charles Markello' Platform Talk (PgmNr 15) on applying vg for rare variant discovery!