Human Pangenome Workshop:

An Introduction to the Human Pangenome Tools and Workflows

Go here to follow along the workshop
tinyurl.com/PangenomeHugo24

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

Introduction to the HPRC

Karen Miga, PhD
Director of Human Pangenome Reference Production Center,
University of California, Santa Cruz

Introduction to the HPRC

Karen Miga, PhD
Director of Human Pangenome Reference Production Center,
University of California, Santa Cruz

Building and Analyzing Pangenome Graphs

Simon Heumos, PhD Candidate
Quantitative Biology Center (QBiC),
University of Tübingen, Tübingen, DE

Introduction to the HPRC

Karen Miga, PhD
Director of Human Pangenome Reference Production Center,
University of California, Santa Cruz

Building and Analyzing Pangenome Graphs

Simon Heumos, PhD Candidate
Quantitative Biology Center (QBiC),
University of Tübingen, Tübingen, DE

Introduction to the Draft Pangenome and or Giraffe

Xian Chang-Monlong, PhD Candidate
Computational Genomics Laboratory Research Group,
University of California, Santa Cruz

Introduction to the HPRC

Karen Miga, PhD
Director of Human Pangenome Reference Production Center,
University of California, Santa Cruz

Building and Analyzing Pangenome Graphs

Simon Heumos, PhD Candidate
Quantitative Biology Center (QBiC),
University of Tübingen, Tübingen, DE

Introduction to the Draft Pangenome and or Giraffe

Xian Chang-Monlong, PhD Candidate
Computational Genomics Laboratory Research Group,
University of California, Santa Cruz

Introduction to Short Read Mapping and Variant Calling to a Pangenome

Jean Monlong, Senior Scientist/Researcher
CRCN INSERM. IRSD, Toulouse, France

Workshop Logistics

9:00 **Introduction to the Human Pangenome Project**: *Karen Miga*

9:20 **Building and Analyzing Pangenome Graphs**: *Simon Heumos*

10:45 Break

11:05 **Introduction to the Draft Pangenome and Giraffe**: *Xian Chang*

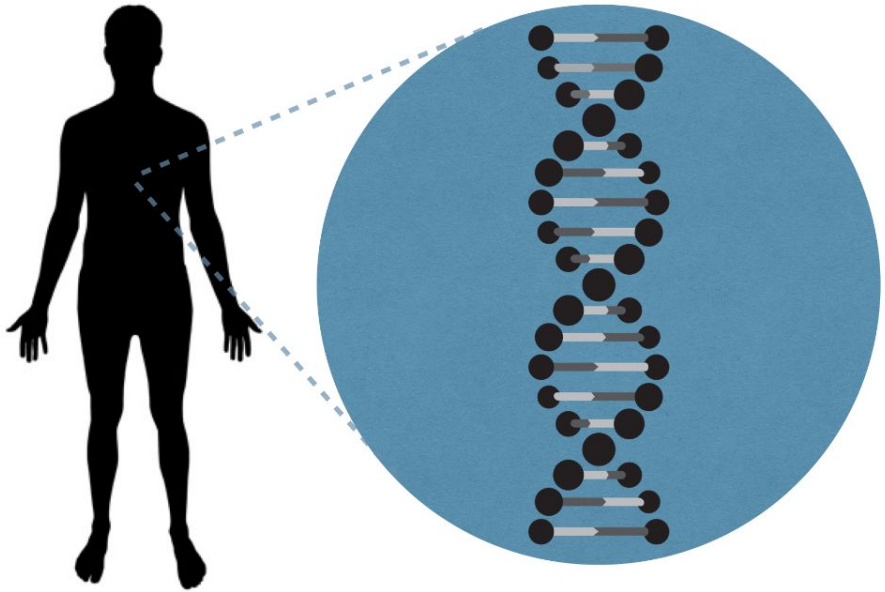
11:25 **Introduction to Short Read Mapping and Variant Calling to a Pangenome**: *Jean Monlong* and *Xian Chang*

12:30 End

Human Pangenome Reference

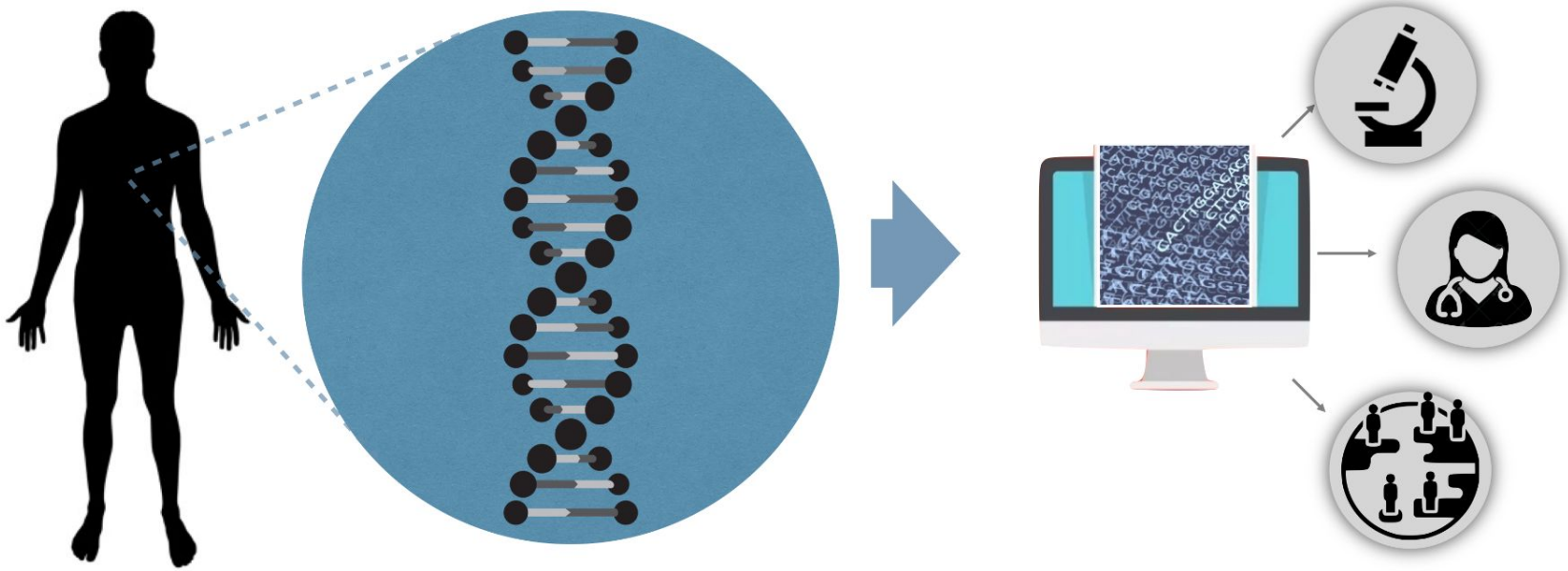
The Initial Human Genome Project

The Human Genome Project: **international, collaborative research program** whose goal was the complete mapping and understanding of all the genes of human beings.



The Initial Human Genome Project

Centralized coordinate system: critical for **sharing genomic data and data analysis standards** among international researchers.

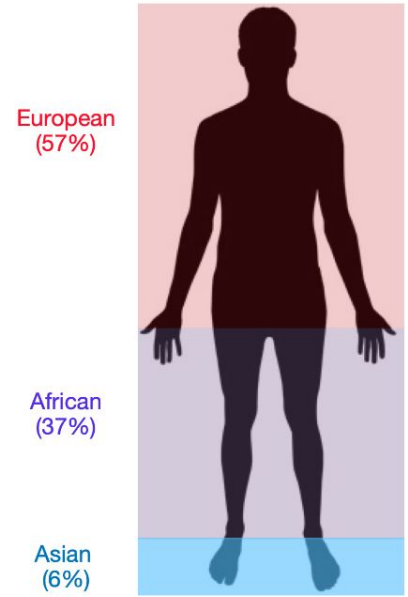


Challenge: Diversity In Genomic Research

Largely represents genomic information from a single individual

Current human genome reference does not adequately represent global genetic diversity

GRCh38

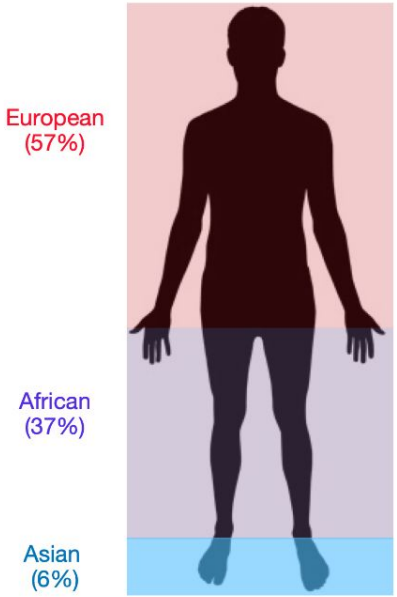


Green et al 2010

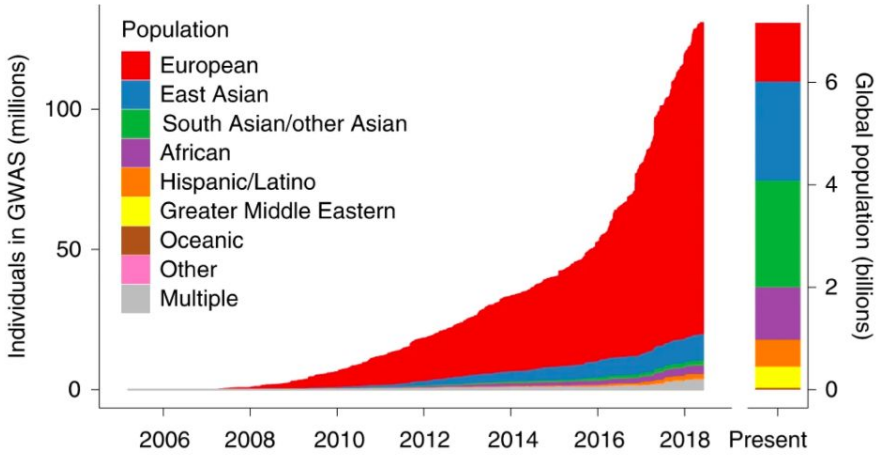
Challenge: Diversity In Genomic Research

Largely represents genomic information from a single individual

Current human genome reference does not adequately represent global genetic diversity



Green et al 2010



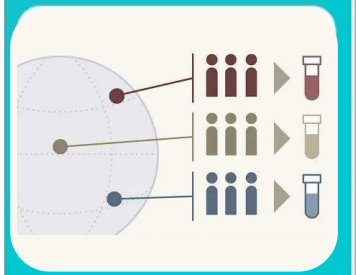
Martin et al 2019

Human Pangenome Reference Consortium

- Improve representation of **global genomic diversity** (>350 diverse diploid references)
- **Prioritizing quality**: we aim to release a complete (T2T) and comprehensive map of genome variation
- **Develop a new, non-linear reference data structure** and foster an innovative ecosystem of pangenomic tools
- Outreach, Education and Implementation

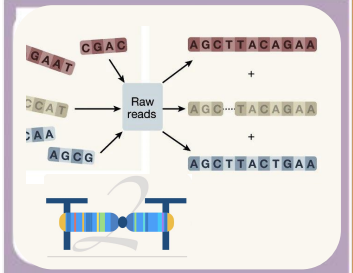
Highly collaborative, multidisciplinary, and cross-institutional working groups

Population sampling and representation

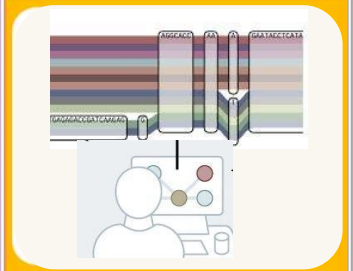


Technology Production

Phased/Finished T2T Assemblies



Pangenome and new workflows/tooling

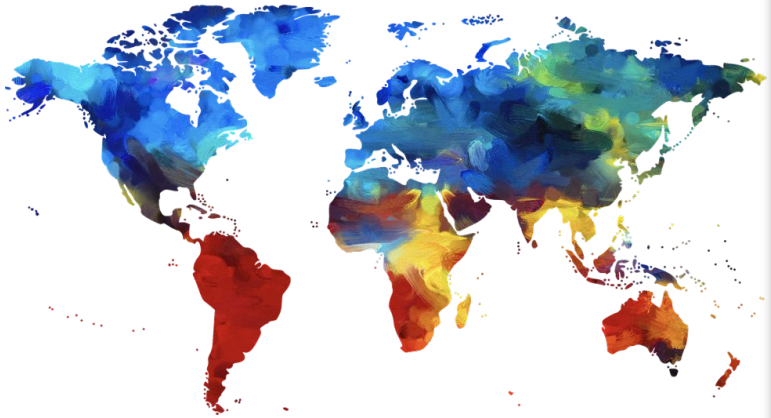


International Pangenome Project

Embedded Ethics and Policy:
Inter-disciplinary ethics working group/oversight committee

HPRC's Initial Sampling Efforts

Population Representation and Sampling



Initial Sampling Efforts:

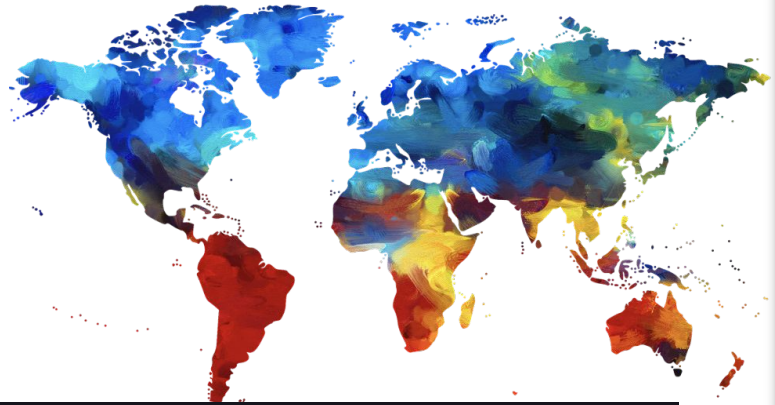
A THOUSAND GENOMES

African Ancestry	24%	Gambian, Western Division African Caribbean, Barbados Esan, Nigeria Mende, Sierra Leone
Americas	30%	Colombians, Medellin Peruvian, Lima Puerto Ricans, Puerto Rico
East Asia	18%	Southern Han Chinese Kinh Vietnamese
South Asia	28%	Bengali Punjabi Tamil

- ☑ Cover genetic and geographic diversity
- ☑ Availability of low passage cell lines
- ☑ Availability of trios/parental data.

HPRC's Initial Sampling Efforts

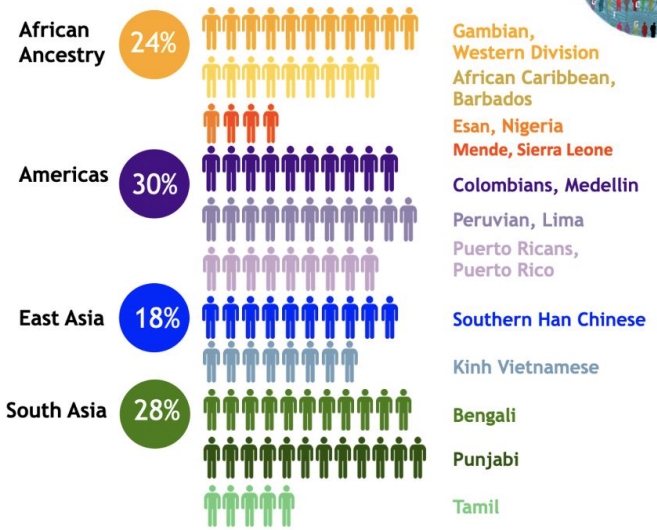
Population Representation and Sampling



1000 Genomes data alone is insufficient to fully represent genomic diversity within the human population

Initial Sampling Efforts:

A THOUSAND GENOMES



- ☑ Cover genetic and geographic diversity
- ☑ Availability of low passage cell lines
- ☑ Availability of trios/parental data.

HPRC's Data Flow

Dovetail GENOMICS

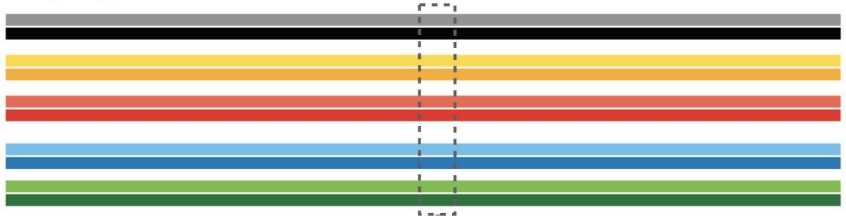
illumina

Oxford NANOPORE Technologies

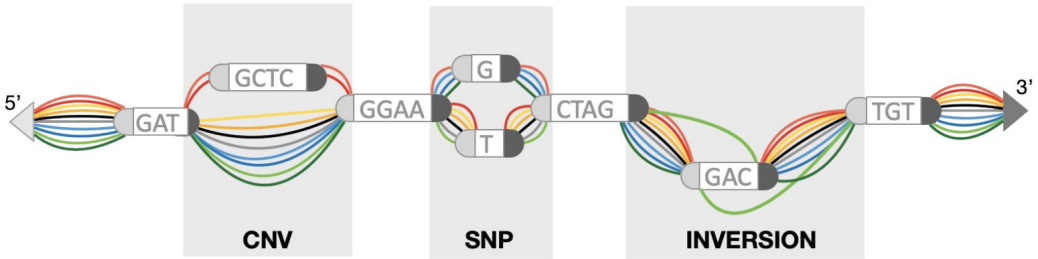
PACBIO®

Release → **Data/ Workflows**

Haplotype-Phased Assemblies:



Release → **Data/ Workflows**



Release → **Data/ Workflows**

First Data Release

sequencing data & QC metrics
for the first 30 samples*

30 HiFi (30x, 17-20kb)	30 ONT Ultra-Long (~6x 100 kb+)	30 Hi-C (Omni-C, ~60X)
30 Bionano Maps (N50>250kb, ~100X coverage)	60 Parental Datasets (30x, 150 bp PE)	10 Strand-Seq single-cell libraries

Open Data Sharing and Cloud-based Data Management

amazon web services
S3://human-pangenomics

AnVIL
AnVIL_HPRC

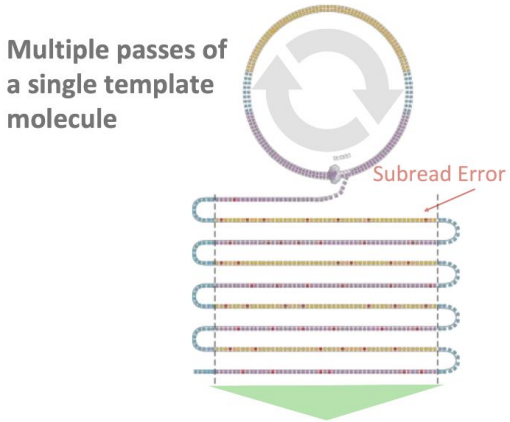
Dockstore
Human Pangenome

GitHub
Human-Pangenomics

Data Production Has An Emphasis On Long Reads

Advances in Long-Read Sequencing

PacBio High Fidelity (HiFi) Data

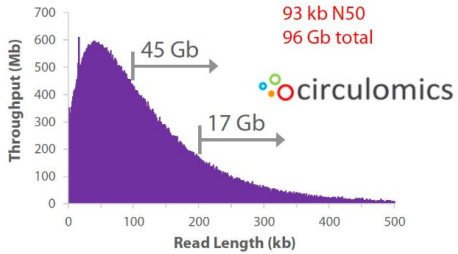


High-Quality Circular Consensus ("CCS") Read

99.9% Consensus Read Accuracy
(35-40x Coverage >Q20 HiFi Reads; 18-20kb)

Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Wenger et al. *Nature Biotechnology* (2019)

ONT Ultra-Long (UL) Data

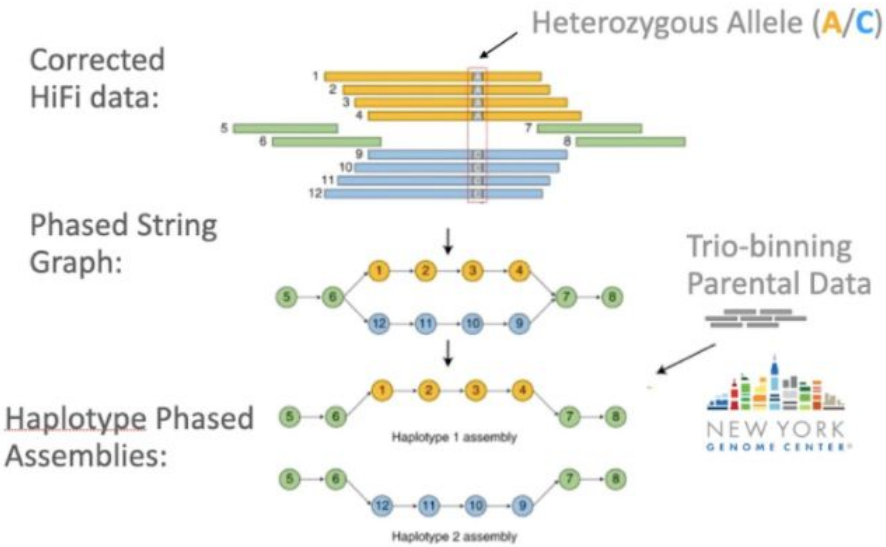


Nanopore sequencing and assembly of a human genome with ultra-long reads. Jain et al. *Nature Biotechnology* (2018)

Assembly Strategy

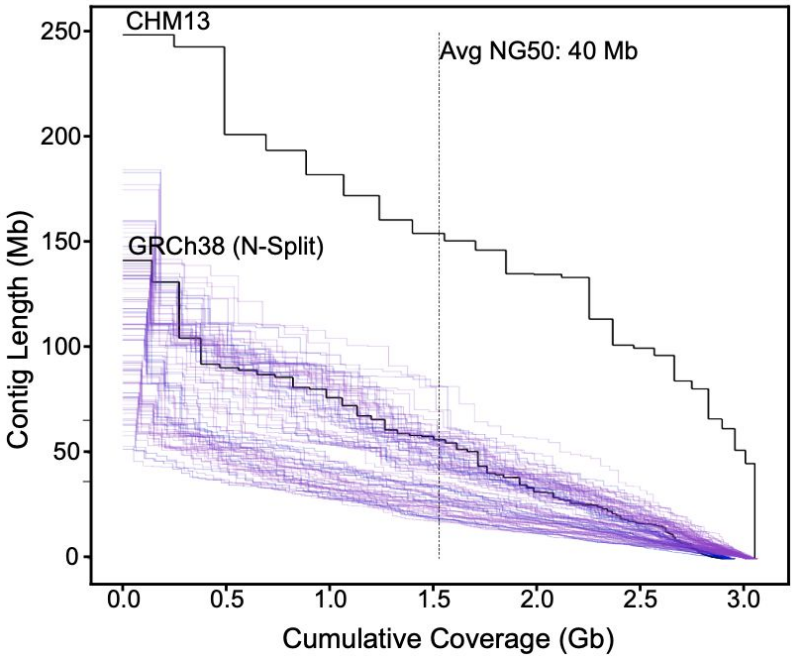
- **Needed a high-quality, automated workflow to create contig-level diploid assemblies**
- **Held a bake-off**
  - *Jarvis, et al. Nature 2022*
- **Developing methods for automated QC, scaffolding, T2T, & non-trio assembly**

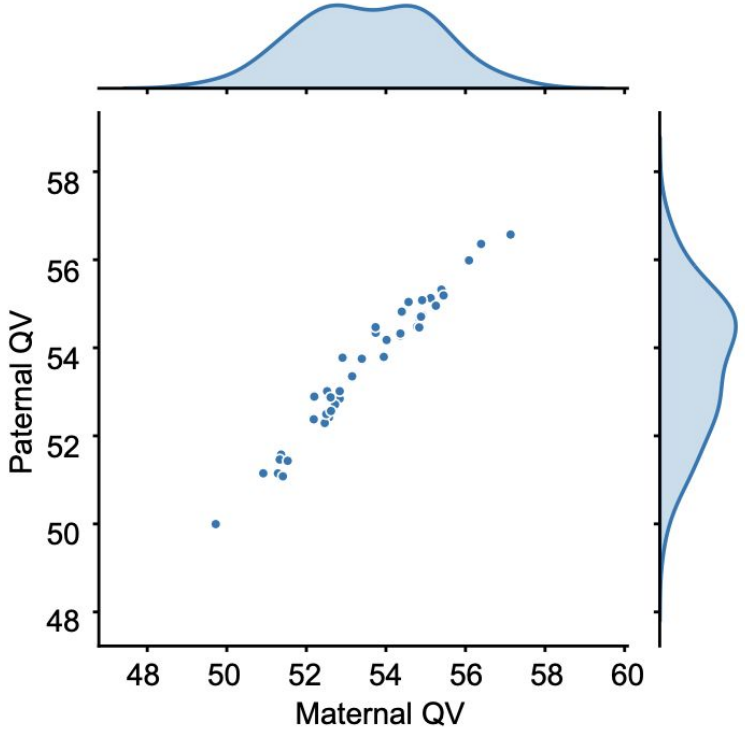
Hifiasm



Cheng, Haoyu, et al. *Nature Methods* (2021)

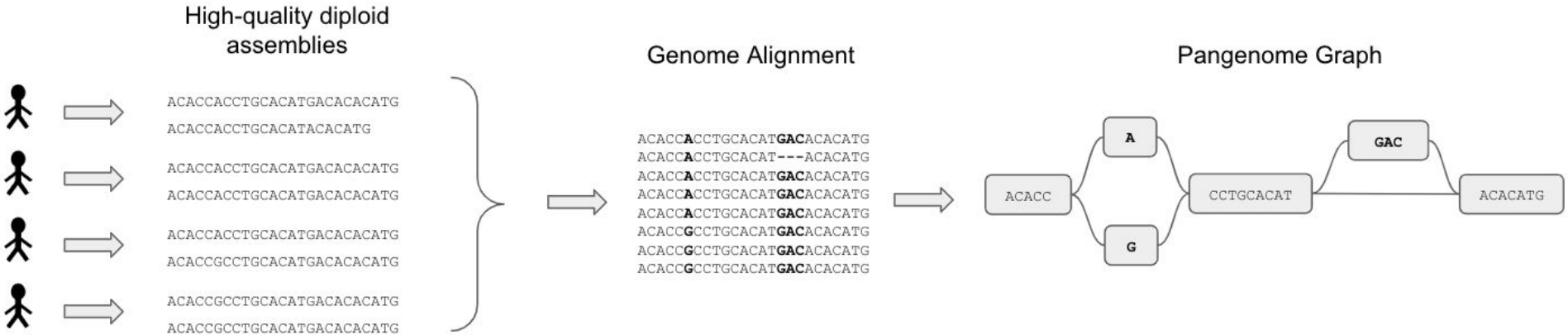
High-Quality Diploid Assemblies



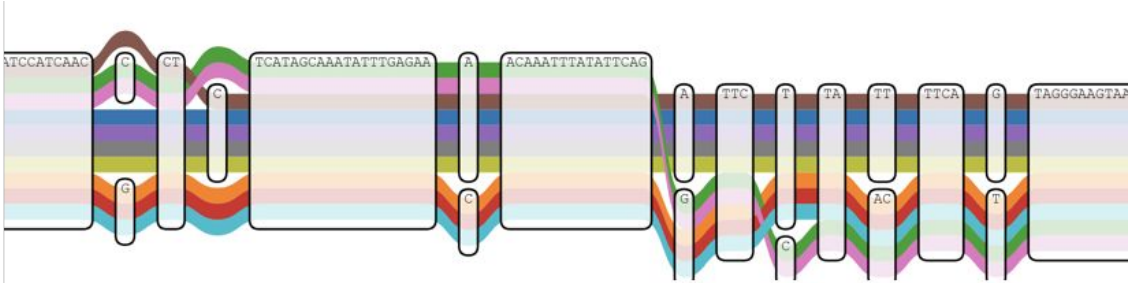


Liao, Asri, Ebler, et al. *BioArxiv* (2022)

Pangenome: Conceptual Approach To Creation



The First Release of a Human Pangenome



nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open Access | Published: 10 May 2023

A draft human pangenome reference

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, ... Benedict Paten

+ Show authors

Nature **617**, 312–324 (2023) | Cite this article

5 Citations | **2985** Altmetric | Metrics

Pangenome: Three Approaches

- Pangenome team currently has “Freeze1” pangenomes
  - The best pangenomes we can create now with current data and tools
  - 90 haplotypes were included (includes GRCh38 and CHM13)

	Minigraph	**Minigraph/CACTUS**	**PGGB**
Short Read Mapping	Untested	**Yes (fast)**	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: Three Approaches

●Generalization of minimap2
●Iterative construction
●SV only

	Minigraph	**Minigraph/CACTUS**	**PGGB**
Short Read Mapping	Untested	**Yes (fast)**	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: Three Approaches

- **Adds base-level alignments to minigraph**
- **Omits centromeric variation**

	Minigraph	**Minigraph/CACTUS**	**PGGB**
Short Read Mapping	Untested	**Yes (fast)**	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: Three Approaches

- **Constructed with all-to-all pairwise mapping**
- **Contains centromeric variation**

	Minigraph	**Minigraph/CACTUS**	**PGGB**
Short Read Mapping	Untested	**Yes (fast)**	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: How You Can Use A Pangenome

1

Calling SNV/InDels From Short Reads

VG-Giraffe

2

Short RNA Mapping

VG

3

Comparative Genomics

Create browser hub (MC)

Liftover with Comparative Analysis Toolkit

4

Calling Structural Variants From Short Reads

Pangenie

Liao, Asri, Ebler, et al. ***BioArxiv*** **(2022)**

Summary

A human pangenome is vital to ensuring all people are equally represented by the core reference structure that we all, as a community, use

The draft human pangenome adds:

- 120 megabases of new sequence to the human reference
- 282 billion bases of haplotype resolved assembly
- > 1500 gene duplications
- > 20 million small variants
- ~70 thousand structural variant sites

While adoption of the pangenome will take time, pangenome methods (mappers, downstream tools) are evolving fast and demonstrate promising applications right now

HPRC Data & Resources

HPRC Sequencing Data

The AnVIL

Public workspace (requires AnVIL account)

GitHub

Repository with data indexes for GCP and AWS files

Sequence Read Archive

BioProject with INSDC copies of raw sequencing data

HPRC Assemblies

The AnVIL

Public workspace (requires AnVIL account)

GitHub

Assembly repository with indexes for AWS & GCP files

Genbank

BioProject for HPRC assemblies

UCSC Genome Browser

Browser hub with year 1 assemblies & annotations

ENSEMBL

Project page containing assemblies & gene annotations

HPRC Pangenomes

The AnVIL

Public workspace (requires AnVIL account)

GitHub

Repository with data indexes for GCP and AWS files

European Nucleotide Archive

BioProject with pangenomes

This critical global resource must be developed by and provide benefit to the broader global community

HUMAN PANGENOME PROJECT

A Global Human Pangenome Resource

A Federated Alliance of Global Genomics Partners

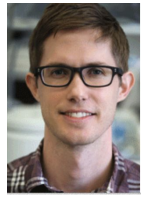
- Technical standards to ensure resource quality
- Ethical and policy standards
- Responsible and secure resource sharing + metadata
- Federated systems for data sharing
- CARE/FAIR principles

HPRC Associate Members

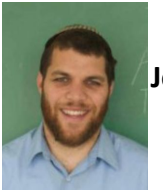
Hardip Patel
Australian National University

Simon Easteal
Australian National University

Steven Salzberg
Johns Hopkins University



Corey Watson
University of Louisville



Jeffrey Rosenfeld
Rutgers University

George Liu
USDA ARS

Obed Garcia
Stanford University

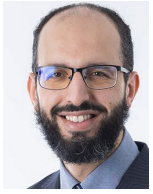
Kai Ye
Xi'an Jiaotong University

Shilpa Garg
University of Copenhagen

Ahmad Abou Tayoun
Al Jalila Children's

Guillaume Bourque
McGill University

Mile Sikic
Genome Institute of Singapore



Younes Mokrab
Sidra Medicine

Nathan Sheffield
University of Virginia

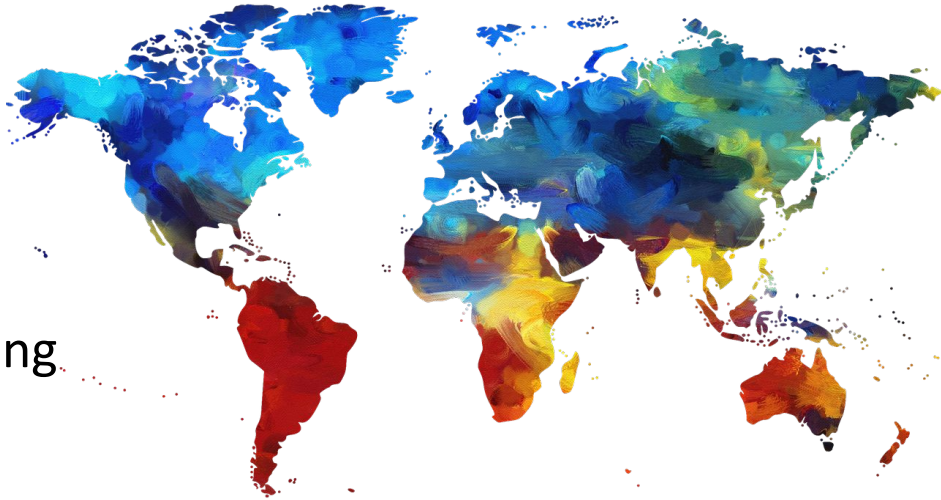
Yafei Mao
Shanghai Jiao Tong University

A Federated Alliance of Global Genomics Partners

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

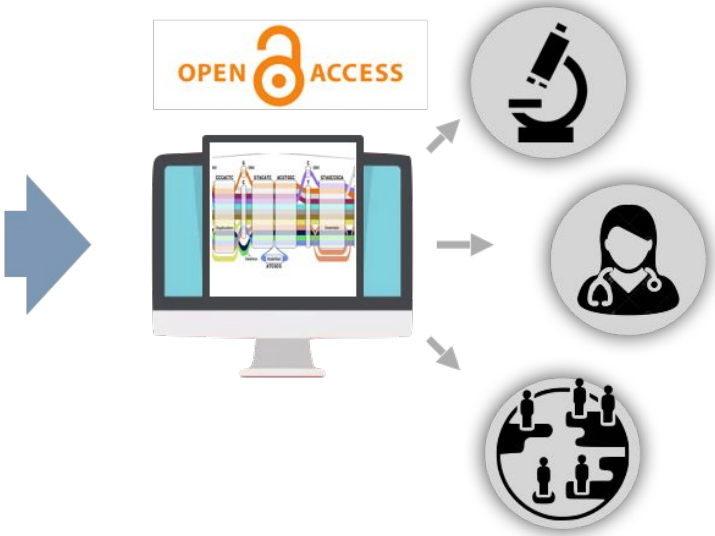
- HPRC is jumpstarting this effort
- Technical standards to ensure resource quality
  - Openly share workflows
- Ethical and policy standards
  - Acquisition and sharing
- Responsible and secure resource sharing



HUMAN PANGENOME PROJECT

HUMAN PANGENOME PROJECT

A Global Human Pangenome Resource



TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

THE GENOME INSTITUTE at Washington University

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Genome Sciences UNIVERSITY OF WASHINGTON

The Rockefeller University 1901

NCBI National Center for Biotechnology Information

CORIELL INSTITUTE FOR MEDICAL RESEARCH

DECODING THE GENOME

EMBL-EBI

HARVARD MEDICAL SCHOOL

Icahn School of Medicine at Mount Sinai

sanger institute

Yale University

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

We would like to acknowledge the National Genome Research Institute (NHGRI) for funding the following grants which are in support of creating the human pangenome reference: 1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973, 1U01HG010963, and the Human Pangenome Reference Consortium (https://humanpangenome.org/)

PacBio

illumina

Google Health

aws

NIH National Human Genome Research Institute

Cantata Bio

Global Alliance for Genomics & Health

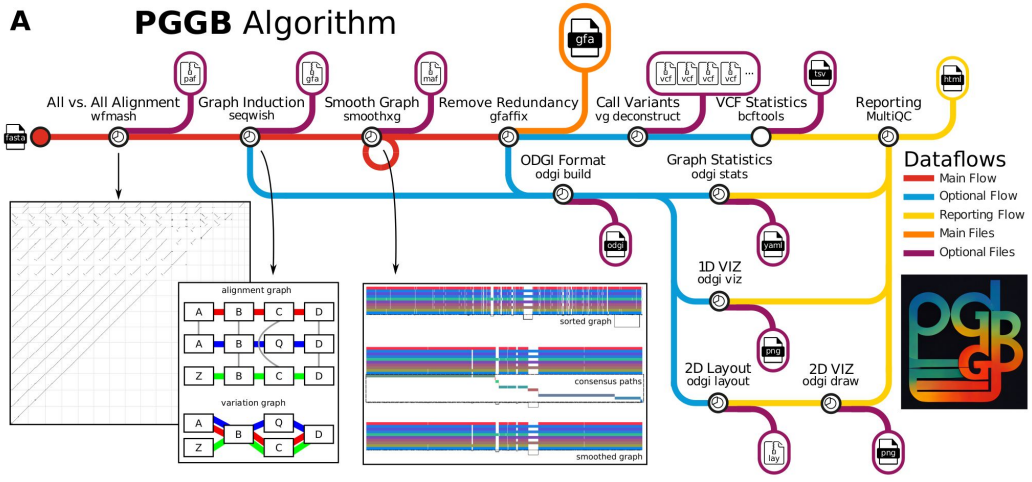
Collaborate. Innovate. Accelerate.

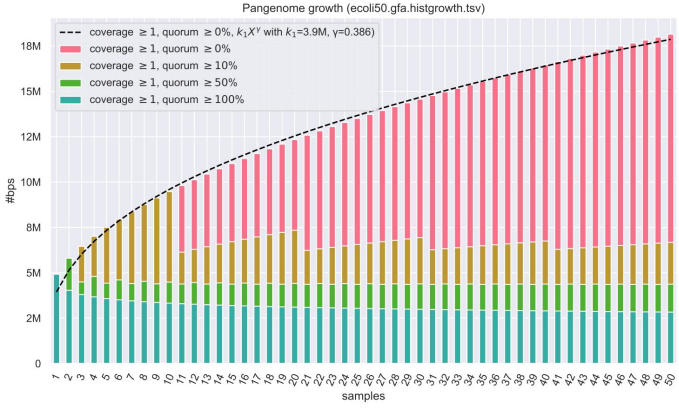
NIST

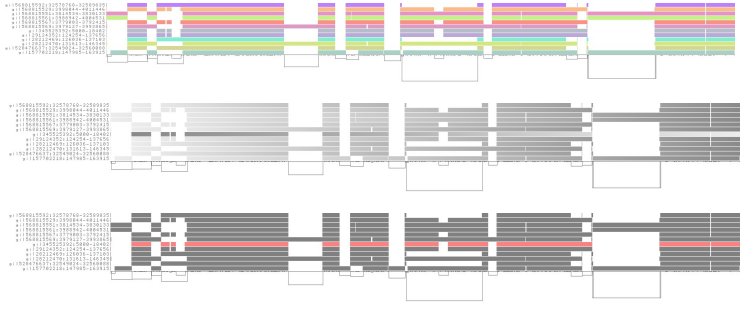
Building and Analyzing Pangenome Graphs

Simon Heumos

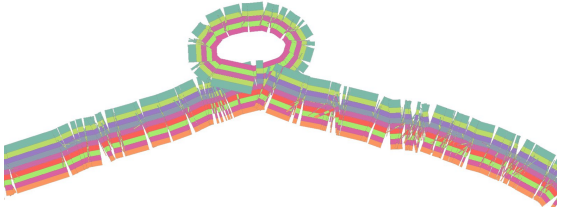
@HPRC HUGO24 Workshop, Italy, Rome
April 8, 2024

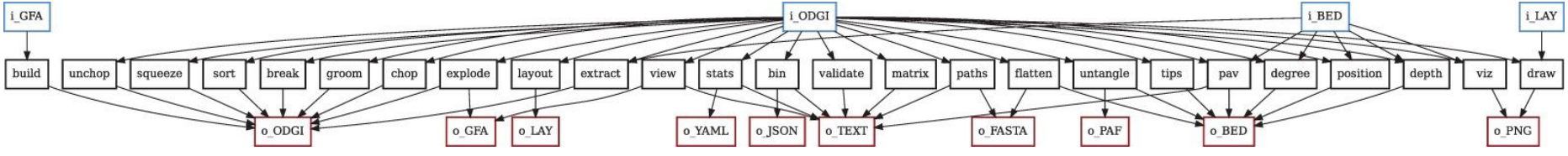






nf-core/pangenome

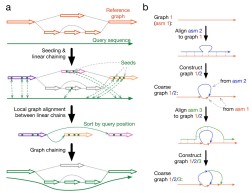




HPRC made 5 pangenome (reference) graphs

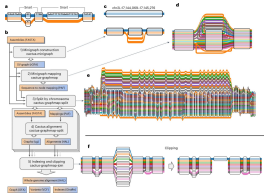
Minigraph

Proof-of-concept seq-to-graph mapper and graph generator



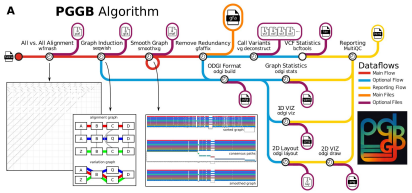
Minigraph-Cactus (MC)

Cactus is a reference-free whole-genome alignment program,
as well as a pangenome graph construction toolkit



PanGenome Graph Builder (PGGB)

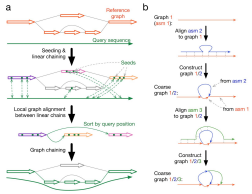
A reference-free pipeline for constructing unbiased pangenome graphs



HPRC made 5 pangenome (reference) graphs

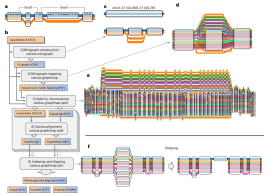
Minigraph

Proof-of-concept seq-to-graph mapper and graph generator



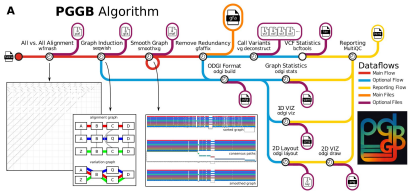
Minigraph-Cactus (MC)

Cactus is a reference-free whole-genome alignment program,
as well as a pangenome graph construction toolkit



PanGenome Graph Builder (PGGB)

A reference-free pipeline for constructing unbiased pangenome graphs



PanGenome Graph Builder (PGGB)

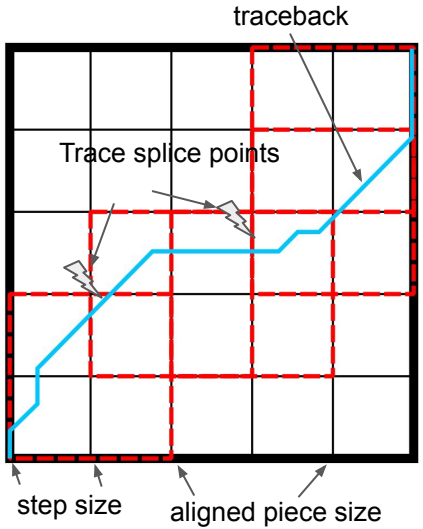

Erik Garrison


Andrea Guarracino

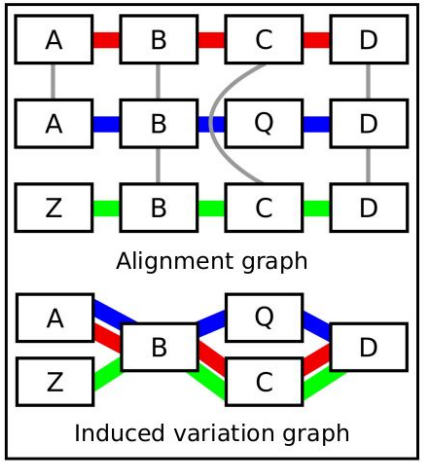

Simon Heumos

PGGB solves the whole genome alignment problem in 3 steps.

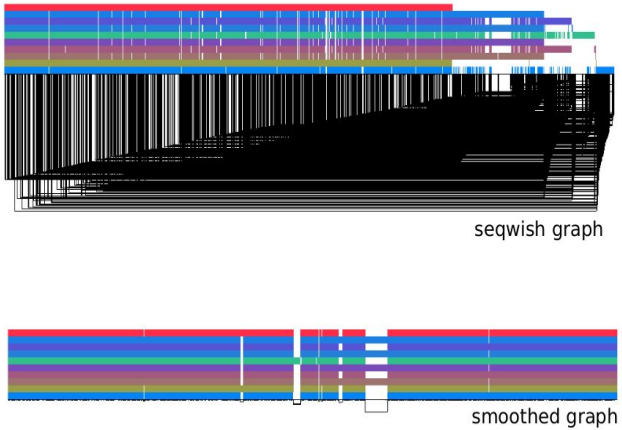
1) all-to-all alignment with **wfmash**



2) graph induction with **seqwish**



3) normalization with **smoothxg**



https://github.com/pangenome/pggb

https://doi.org/10.1101/2023.04.05.535718

PanGenome Graph Builder (PGGB)

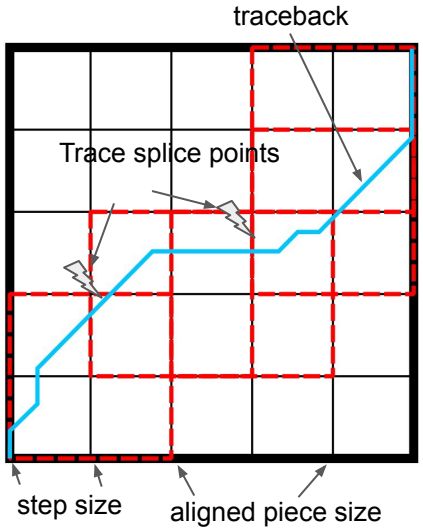

Erik Garrison


Andrea Guarracino

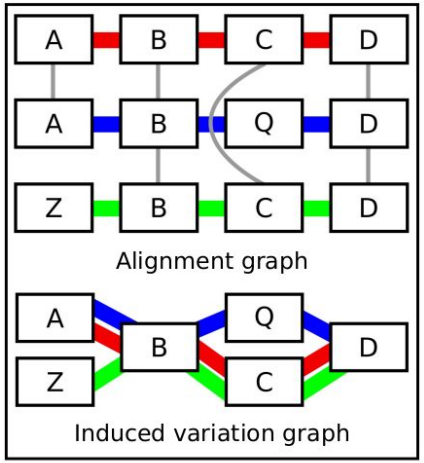

Simon Heumos

PGGB solves the whole genome alignment problem in 3 steps.

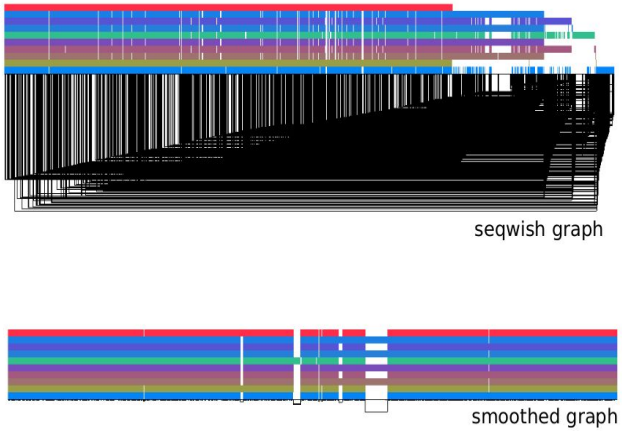
1) all-to-all alignment with **wfmash**



2) graph induction with **seqwish**



3) normalization with **smoothxg**



https://github.com/pangenome/pggb

https://doi.org/10.1101/2023.04.05.535718

wfmash - sparse homology mapping

We apply a heuristic homology mapping step (with MashMap) that efficiently finds regions of query and target sequences, *segments*, that are likely to be good alignments.

This is fast, but it has no facility to derive the precise base-level alignments.

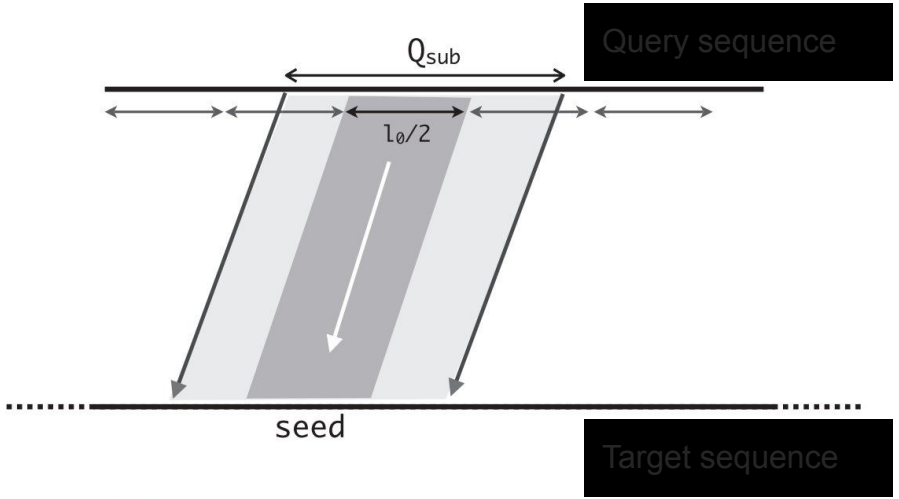
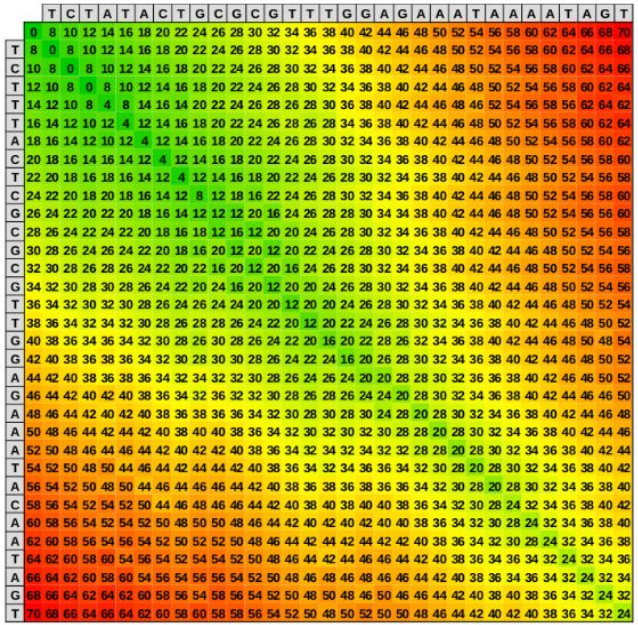


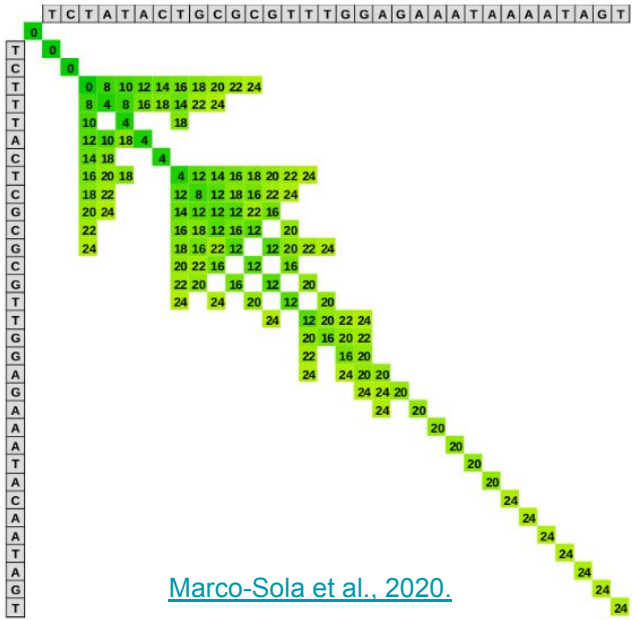
Fig. 1. A local alignment depicting the inclusion of a length $l_0/2$ fragment of the query sequence

Figure from Jain et al., 2018.

https://github.com/waveygang/wfmash

wfmash - WaveFront Alignment (WFA) algorithm





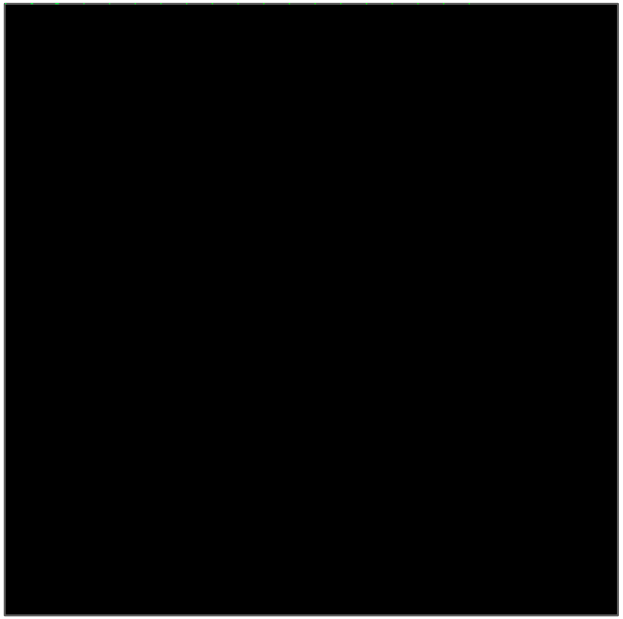
Marco-Sola et al., 2020.

https://github.com/waveygang/wfmash

wfmash - WaveFront Alignment (WFA) algorithm

We implement WFλ, wherein rather than comparing characters we call a function λ (v,h) on a particular set of matrix cells.

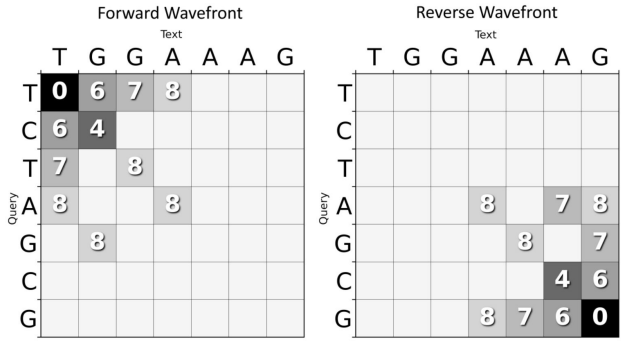
To align a query/target, we segment them into segments and use the WFA algorithm.

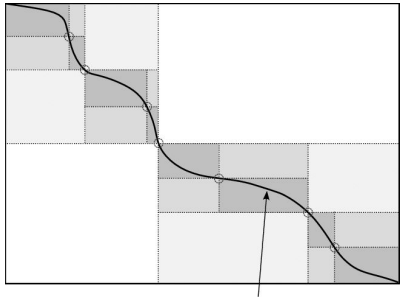


Animation from
https://github.com/RagnarGrootKoerkamp/astar-pairwise-aligner

All-to-all alignment with *bidirectional* wfmash

Bidirectional WFA (BiWFA)





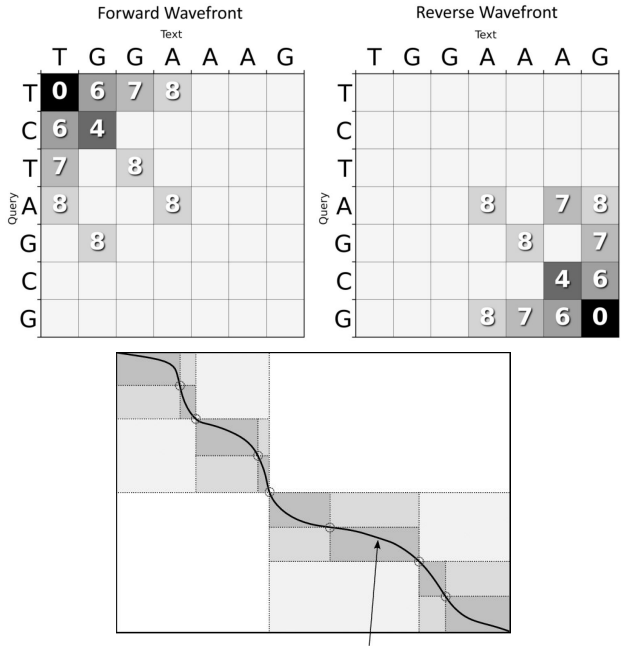
Optimal gap-affine alignment in $O(s)$ space

Santiago Marco-Sola [1,2*], Jordan M. Eizenga [3], Andrea Guarracino [4], Benedict Paten [3], Erik Garrison [5], Miquel Moreto [1,6]

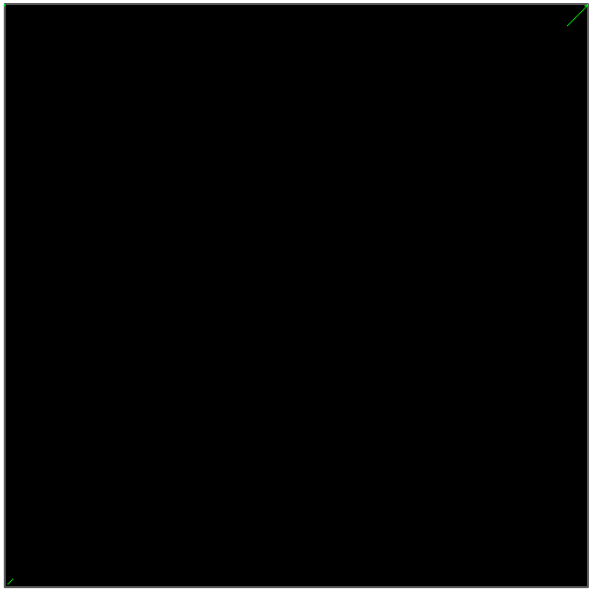
Marco-Sola et al. 2023

All-to-all alignment with *bidirectional* wfmash

Bidirectional WFA (BiWFA)



BiWFλ



Animation from
https://github.com/RagnarGrootKoerkamp/astar-pairwise-aligner

Optimal gap-affine alignment in $O(s)$ space

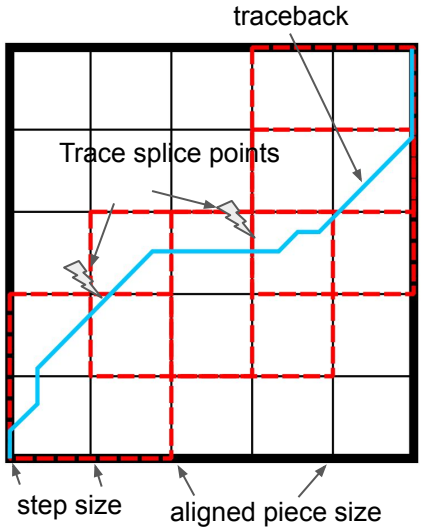
Santiago Marco-Sola [1,2*], Jordan M. Eizenga [3], Andrea Guarracino [4], Benedict Paten [3], Erik Garrison [5], Miquel Moreto [1,6]

Marco-Sola et al. 2023

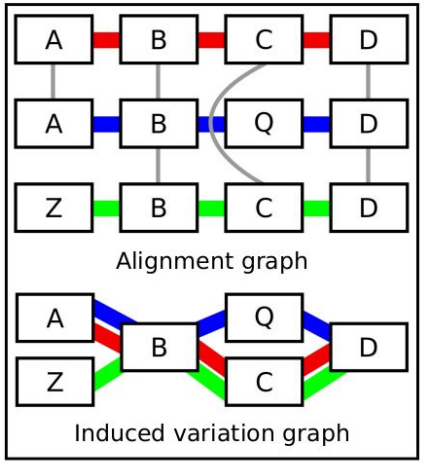
PanGenome Graph Builder (PGGB)

PGGB solves the whole genome alignment problem in 3 steps.

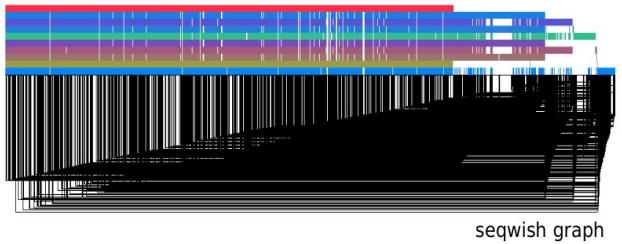
1) all-to-all alignment with **wfmash**

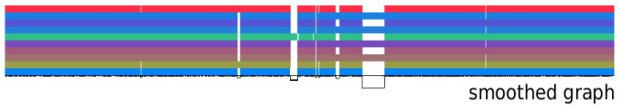


2) graph induction with **seqwish**



3) normalization with **smoothxg**

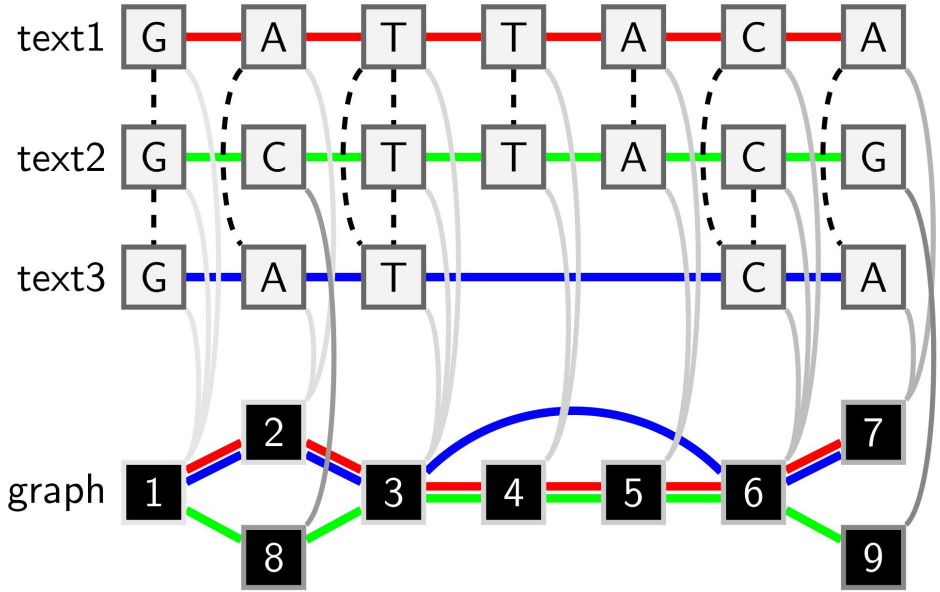




https://github.com/pangenome/pggb

https://doi.org/10.1101/2023.04.05.535718

Graph induction with seqwish



Unbiased pangenome graphs

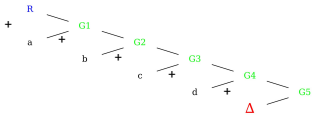
Erik Garrison [1]*, Andrea Guarracino [2]

Garrison et al., 2022, Bioinformatics

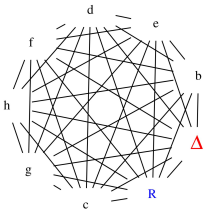
Unbiased pangenome graphs

Graph construction pipelines:

- minigraph (progressive)
- minimap2 + seqwish (all-vs-all)
- wfmash + seqwish (all-vs-all)

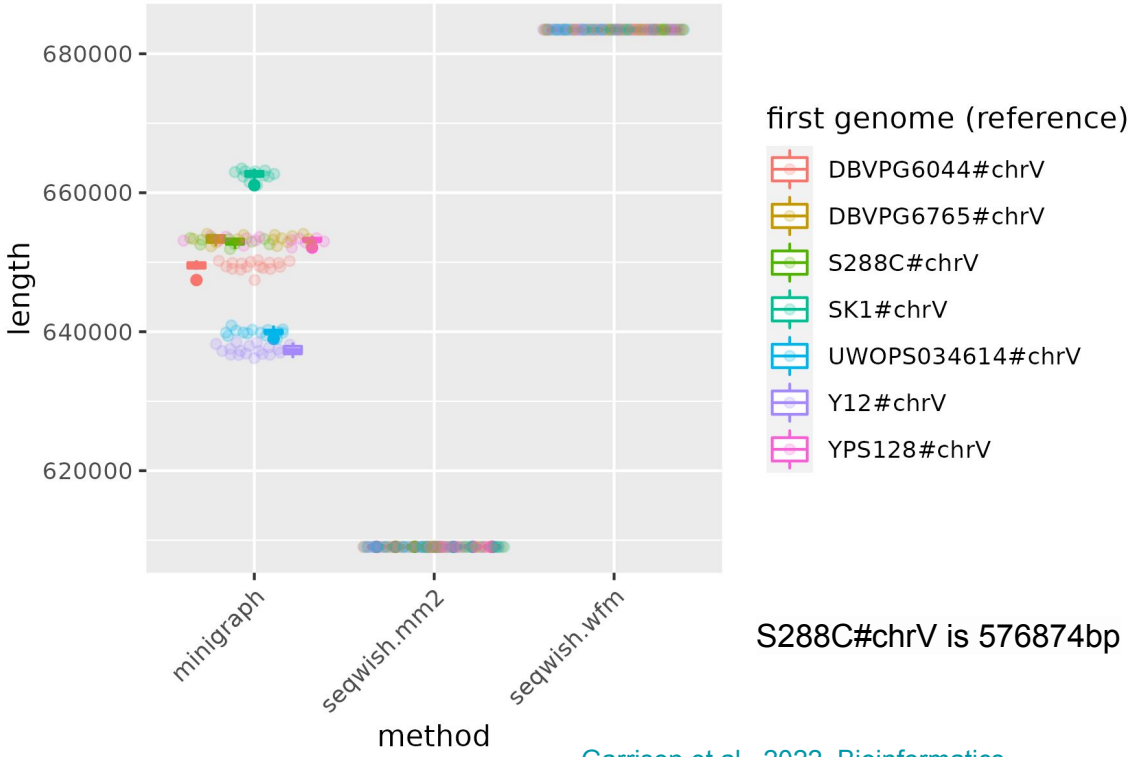


progressive



all-vs-all

Base-pair lengths of graphs built from 100 permutations of yeast chrV.



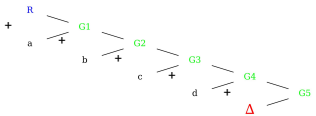
S288C#chrV is 576874bp

Garrison et al., 2022, Bioinformatics

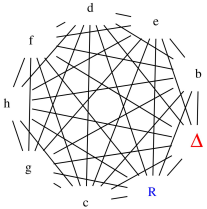
Unbiased pangenome graphs

Graph construction pipelines:

- minigraph (progressive)
- minimap2 + seqwish (all-vs-all)
- wfmash + seqwish (all-vs-all)



progressive



all-vs-all

Most of the variation in minigraph's graphs derives from the first genome which is picked.

Base-pair lengths of graphs built from 100 permutations of yeast chrV.

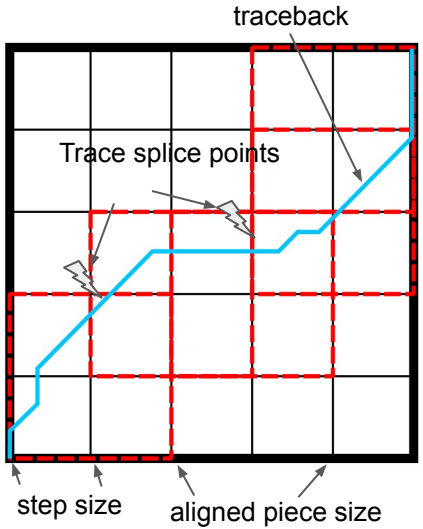
S288C#chrV is 576874bp

Garrison et al., 2022, Bioinformatics

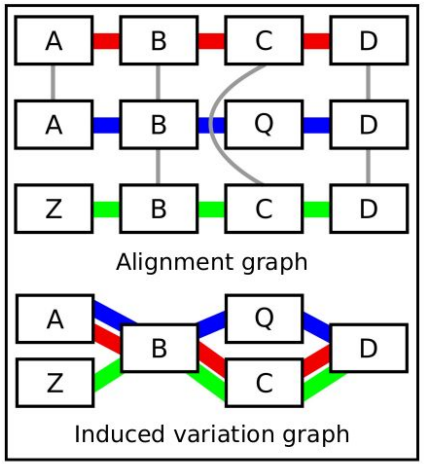
PanGenome Graph Builder (PGGB)

PGGB solves the whole genome alignment problem in 3 steps.

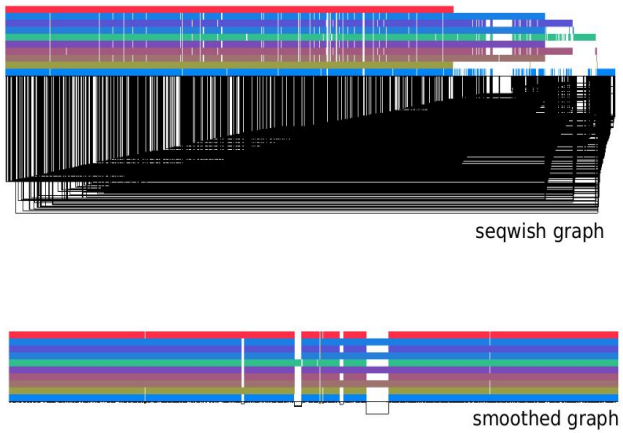
1) all-to-all alignment with **wfmash**



2) graph induction with **seqwish**



3) normalization with **smoothxg**

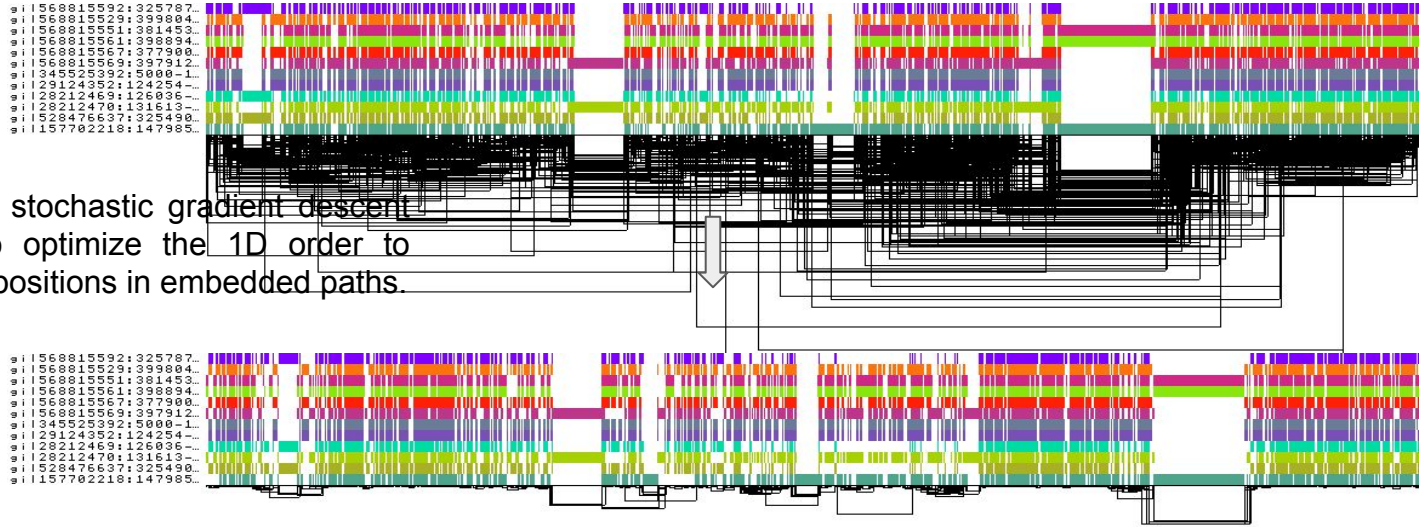


https://github.com/pangenome/pggb

https://doi.org/10.1101/2023.04.05.535718

Graph normalization with smoothxg

Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



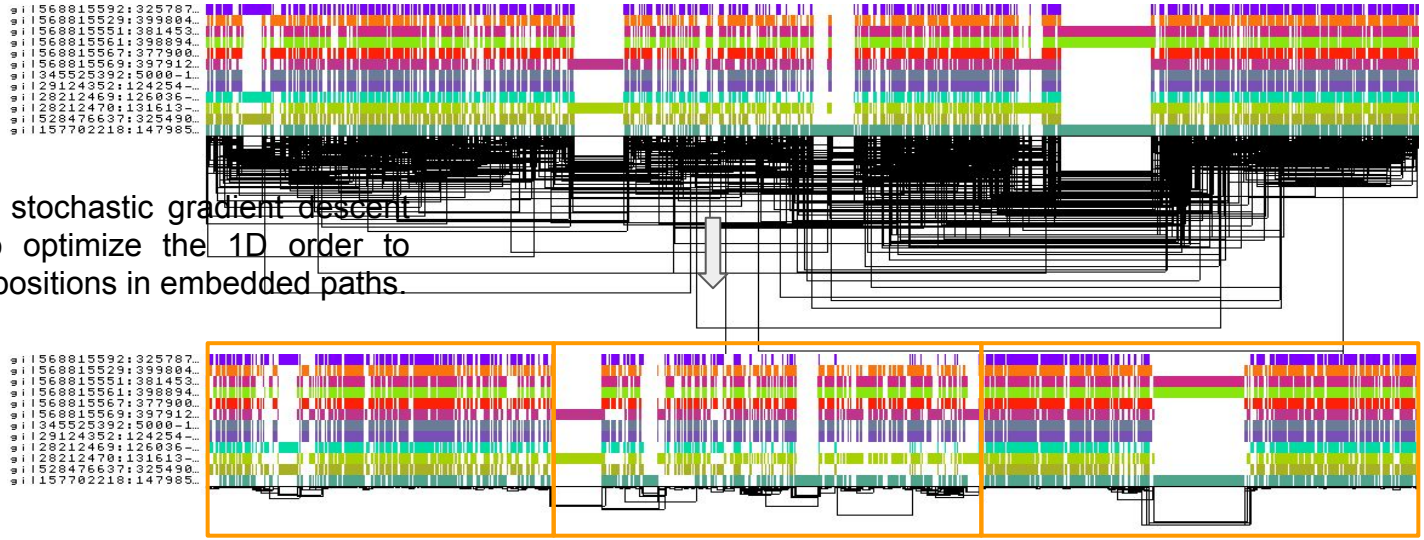
Path-guided stochastic gradient descent algorithm to optimize the 1D order to best-match positions in embedded paths.

https://github.com/pangenome/smoothxg

Heumos*, Guarracino* et al., 2023, bioRxiv
https://doi.org/10.1101/2023.04.05.535718

Graph normalization with smoothxg

Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Path-guided stochastic gradient descent algorithm to optimize the 1D order to best-match positions in embedded paths.

Multiple Sequence Alignment (MSA) over the ordered graph, locally

https://github.com/pangenome/smoothxg

Heumos*, Guarracino* et al., 2023, bioRxiv
https://doi.org/10.1101/2023.04.05.535718

PGGB key parameters

PGGB solves the whole genome alignment problem in 3 steps.

1) all-to-all alignment with **wfmash**

`--map-pct-id`: Percentage of sequence identity for mapping and alignment. Consult **mash**.

Default: 90.0.

`--segment-length`: Segment length for mapping.

Default: 5000.

2) graph induction with **seqwish**

`--min-match-length`: Filter exact matches below this length to prevent local spurious complexity.

Default: 23

3) normalization with **smoothxg**

`--poa-params`: Scoring parameters for the local MSAs in the form of *match,mismatch,gap1,ext1,gap2,ext2*

Default: 1,19,39,3,81,1

https://github.com/pangenome/pggb

https://doi.org/10.1101/2023.04.05.535718

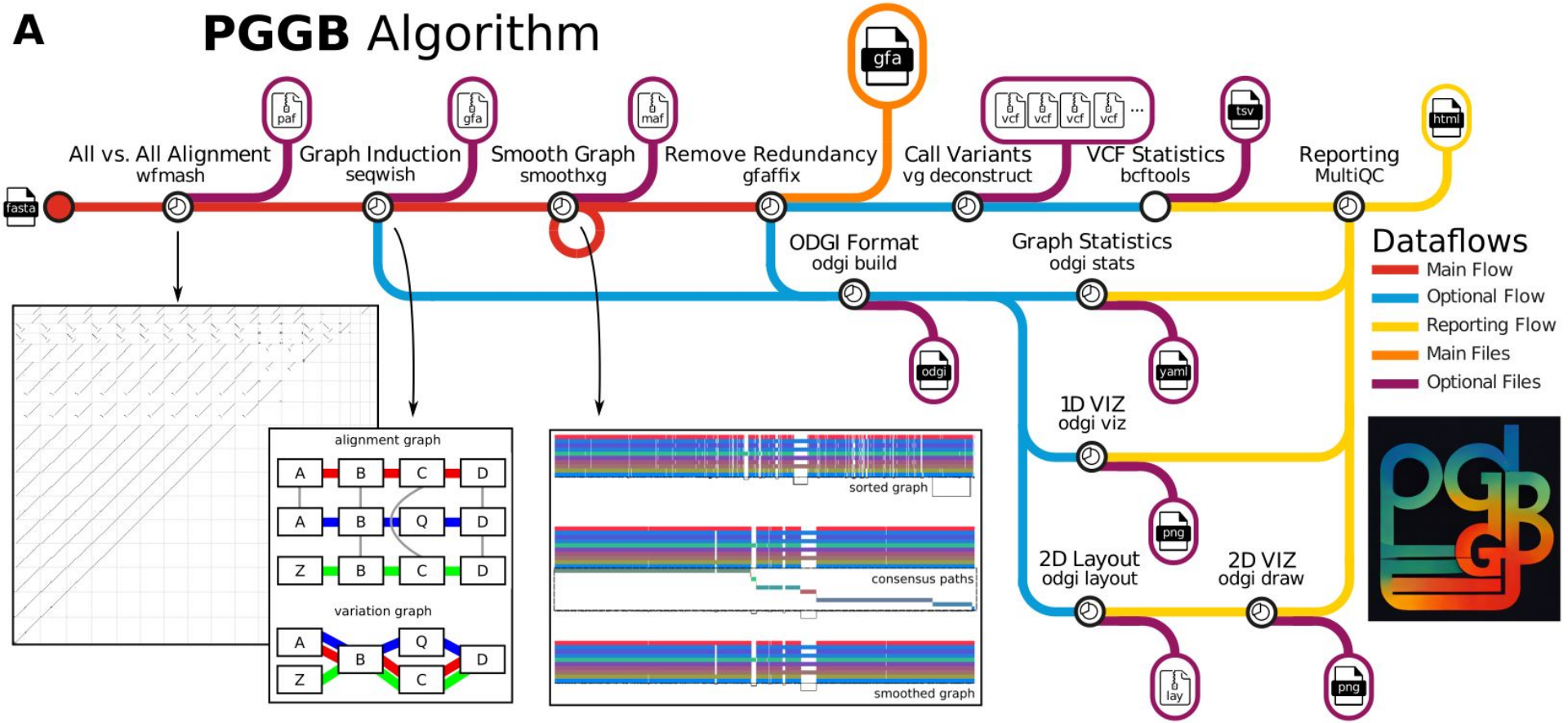
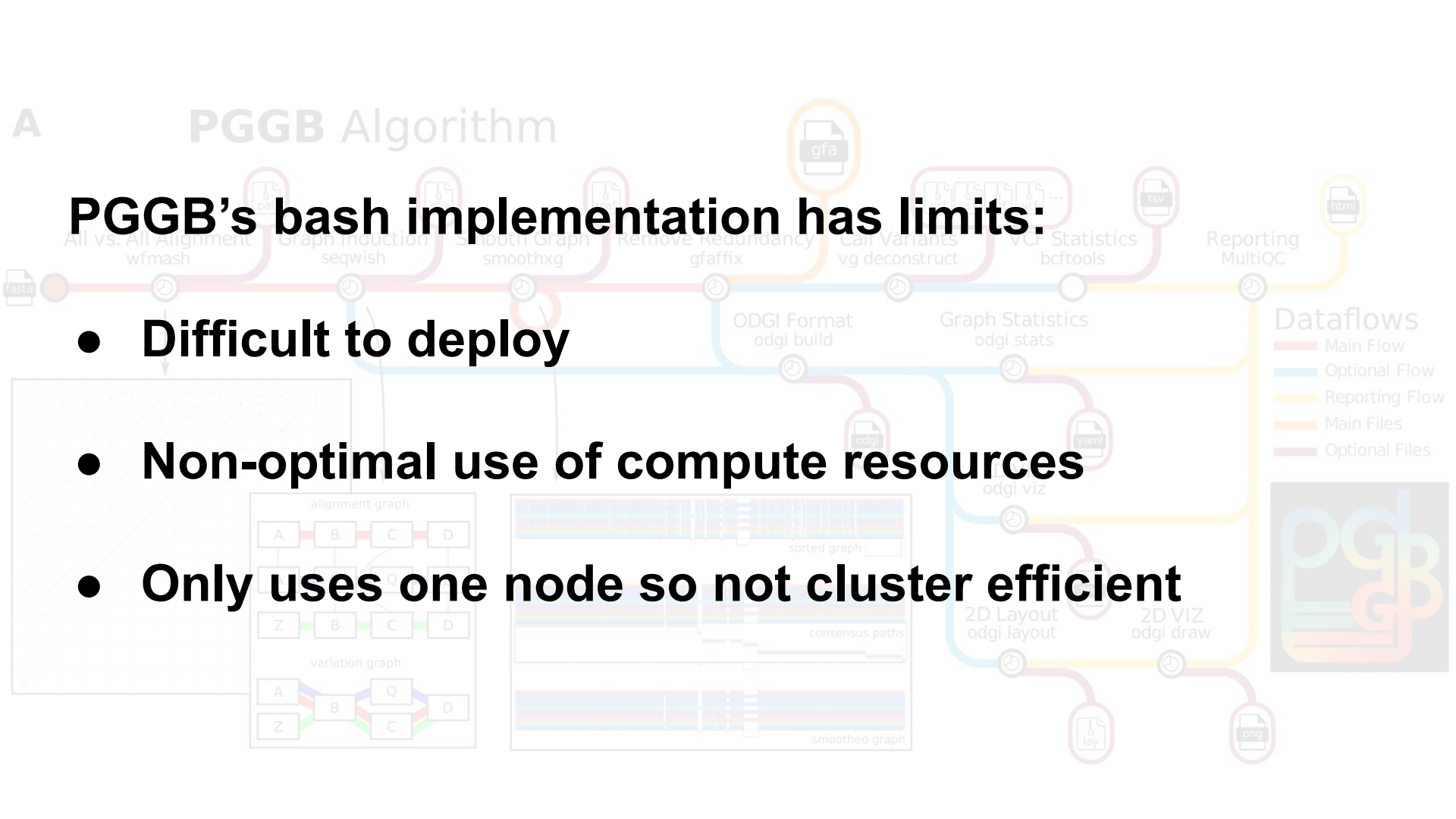
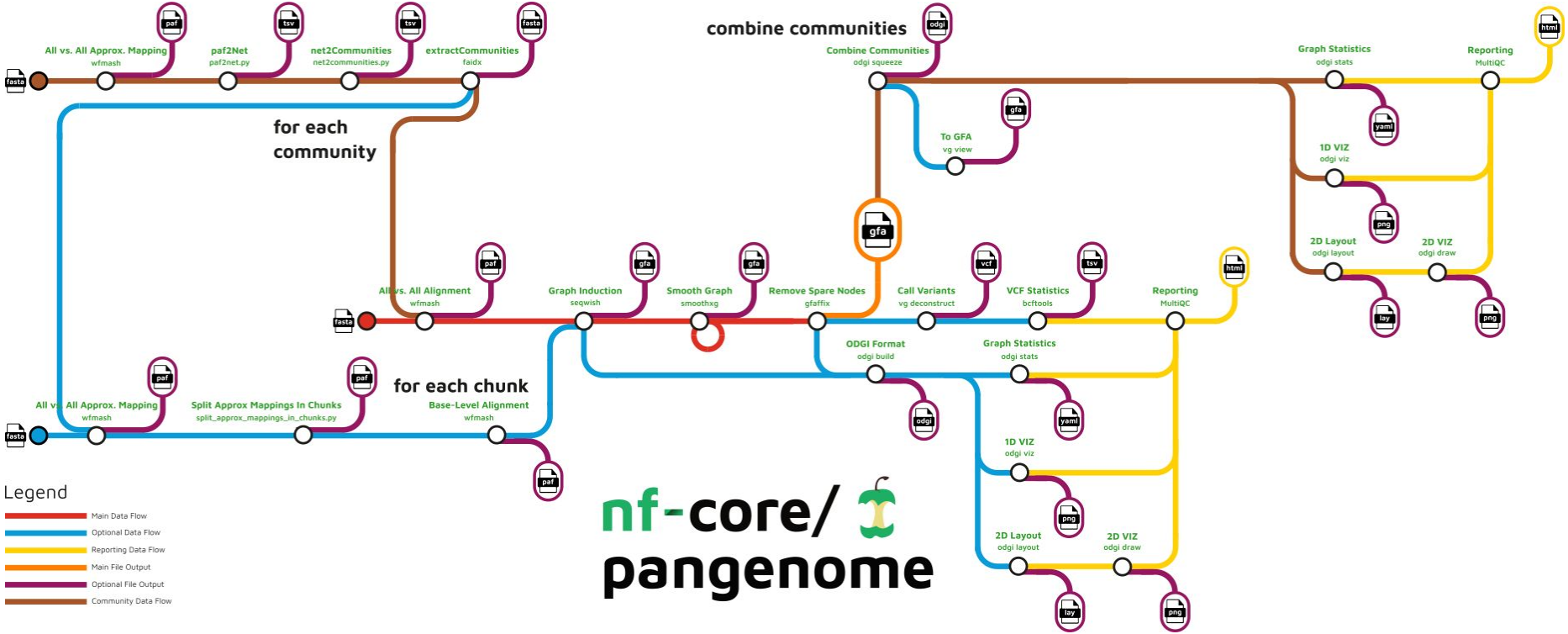


Figure from: [Garrison, Guarracino et al., 2023]

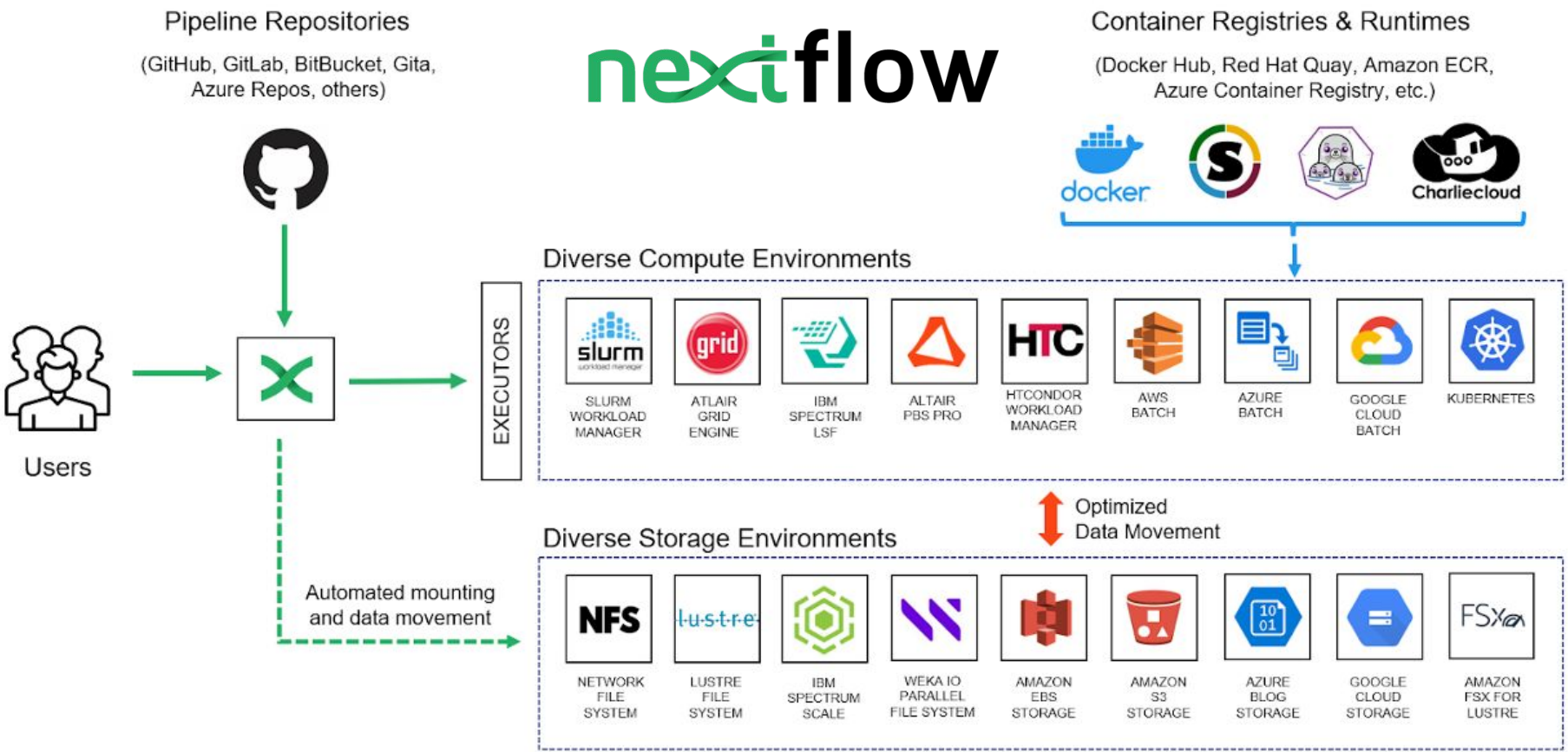


PGGB's bash implementation has limits:

- Difficult to deploy
- Non-optimal use of compute resources
- Only uses one node so not cluster efficient



Heumos et al., in preparation



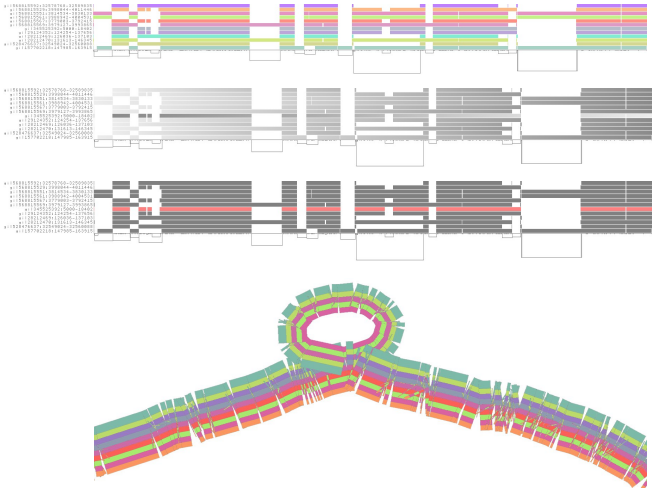
Di Tomasso P. , et al. (2017). Nature Biotechnology, 35, 316–319.

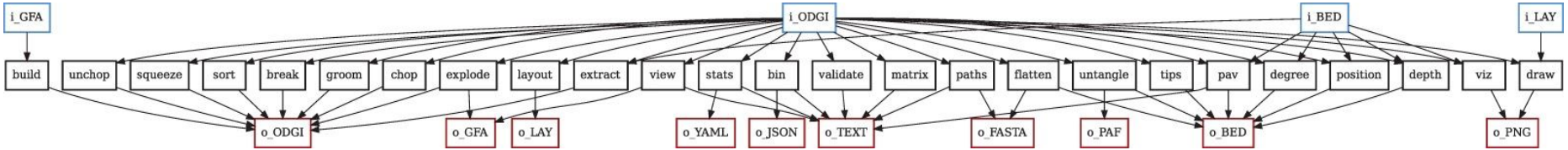
Understanding pangenome graphs with ODGI

ODGI is meant to be a basic toolkit for interacting with pangenome graphs.

It uses the embedded genomes as references.

ODGI offers more than 30 tools for graph interrogation, manipulation, and visualization.





Methods provided by ODGI (in black) and their supported input (in blue) and output (in red) data formats.

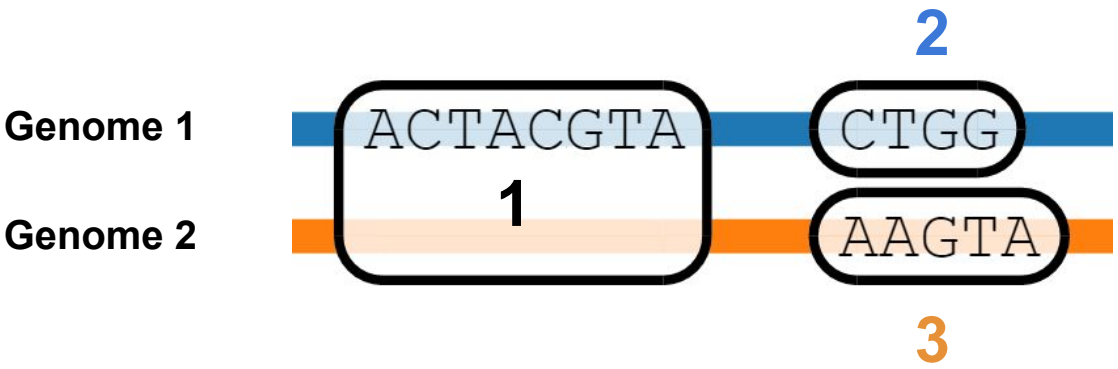
Guarracino*, Heumos* et al., 2022, Bioinformatics

ODGI works with *variation graphs*

Genome 1: **ACTACGTACTGG** Path: **1 2**

Genome 2: **ACTACGTAAAGTA** Path: **1 3**

Linear sequences are **paths** through nodes.



Graph topology is not directly shown.

The nodes represent DNA sequences.

Sketch made using SequenceTubeMap.

Paths can be contigs, haplotypes, reads, or whole chromosomes.

Towards a 1D visualization

Genome 1: **ACTACGTACTGG** Path: **1 2**

Genome 2: **ACTACGTAAAGTA** Path: **1 3**

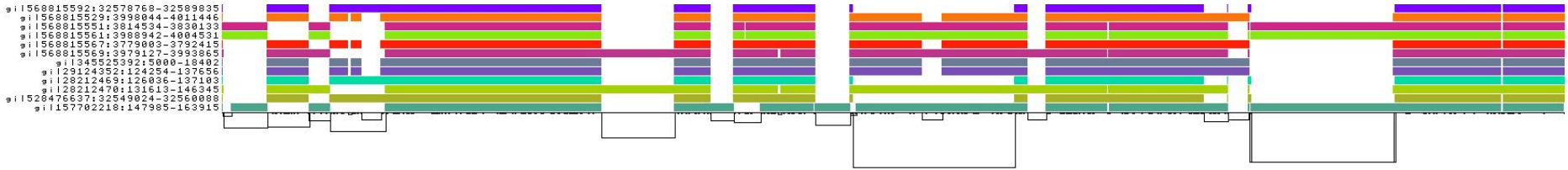
Concatenate nucleotides to a pangenome sequence.

Presence - absence matrix encodes actual genomic sequence.

Visualizations in 1D with odgi viz

By visualizing pangenome graphs we can gain insight into the mutual relationship between the embedded sequences and their variation.

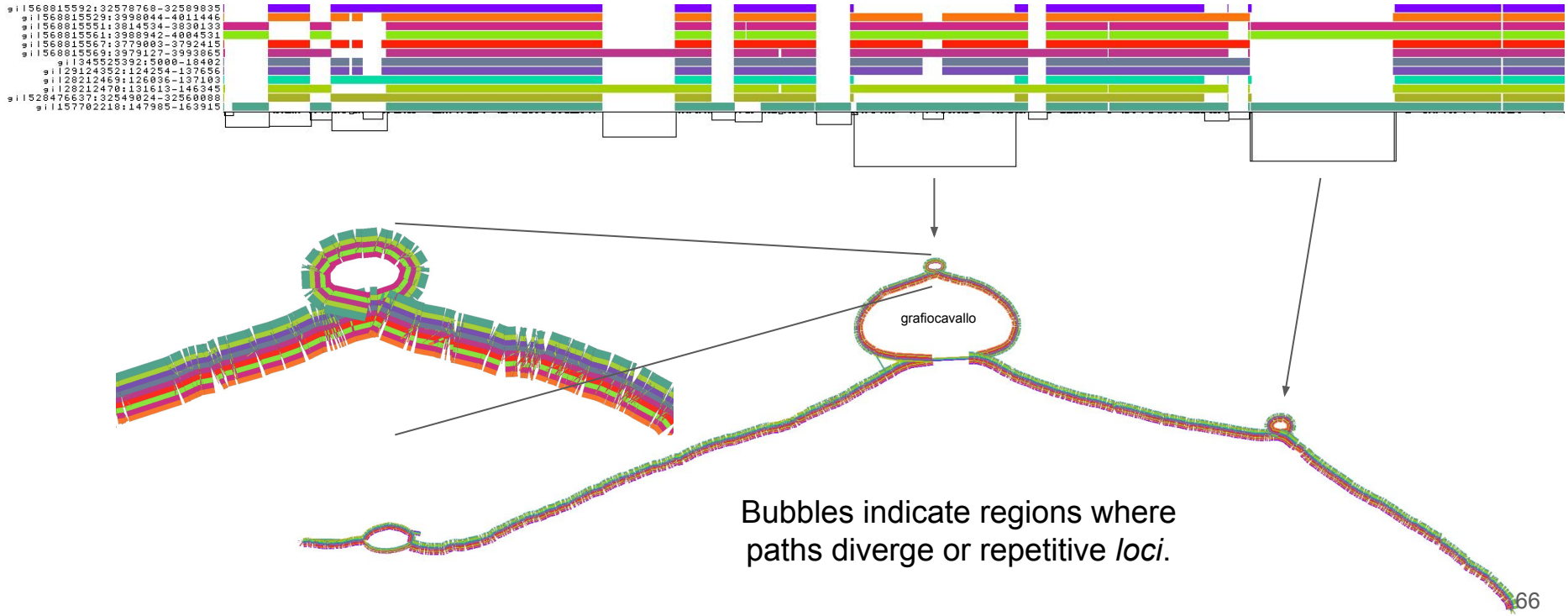
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



- The graph nodes are arranged from left to right, forming the pangenome sequence.
- The colored bars represent the paths versus the pangenome sequences in a binary matrix.
- The path names are visualized on the left.
- The black lines under the paths are the links, which represent the graph topology.

Visualizations in 1D and 2D with odgi viz and draw

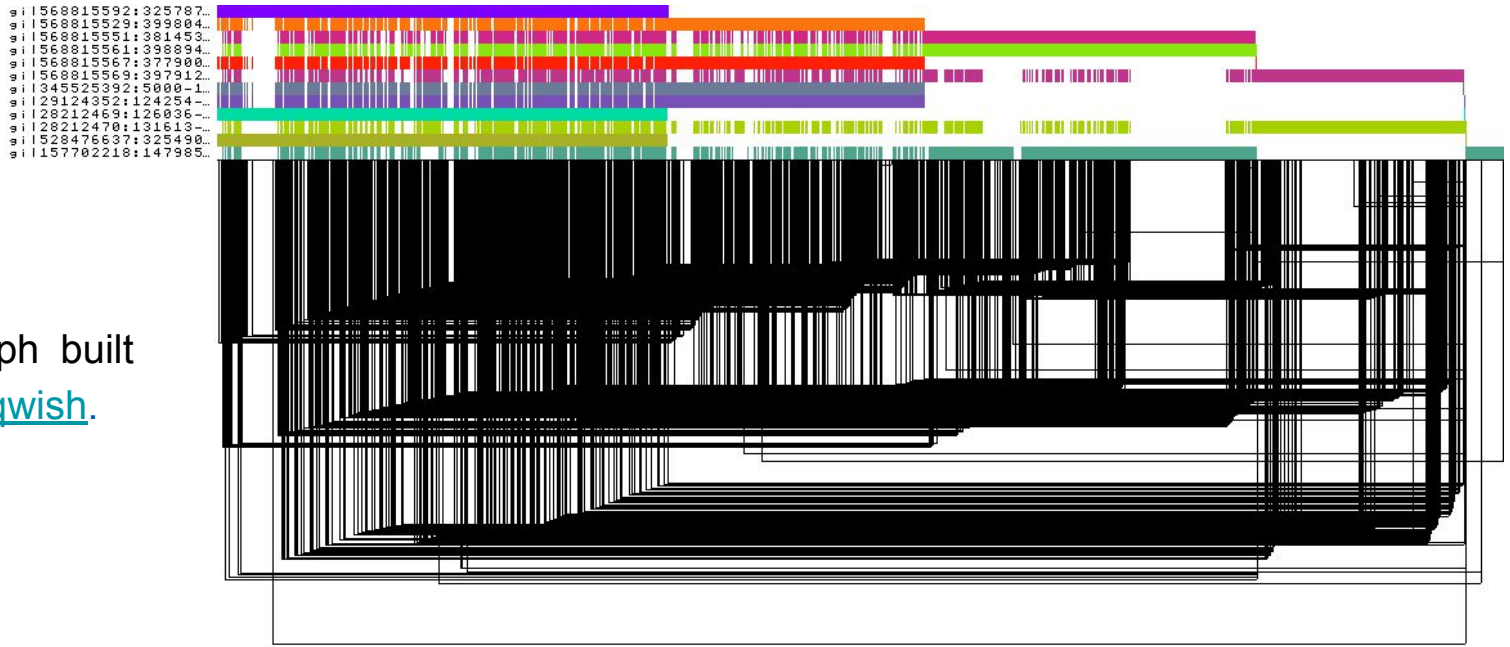
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Bubbles indicate regions where paths diverge or repetitive *loci*.

Finding latent structures with odgi sort

Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.

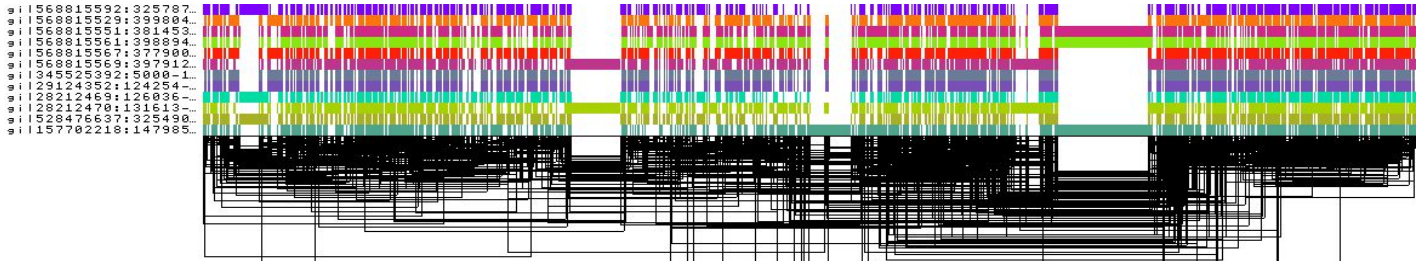


Raw graph built using seqwish.

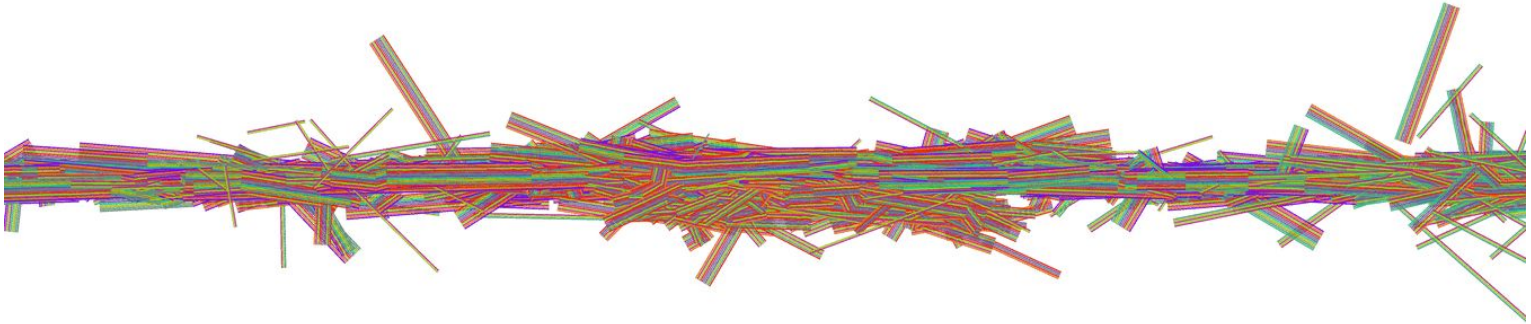
Finding latent structures in pangenome graphs with odgi sort and layout

Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.

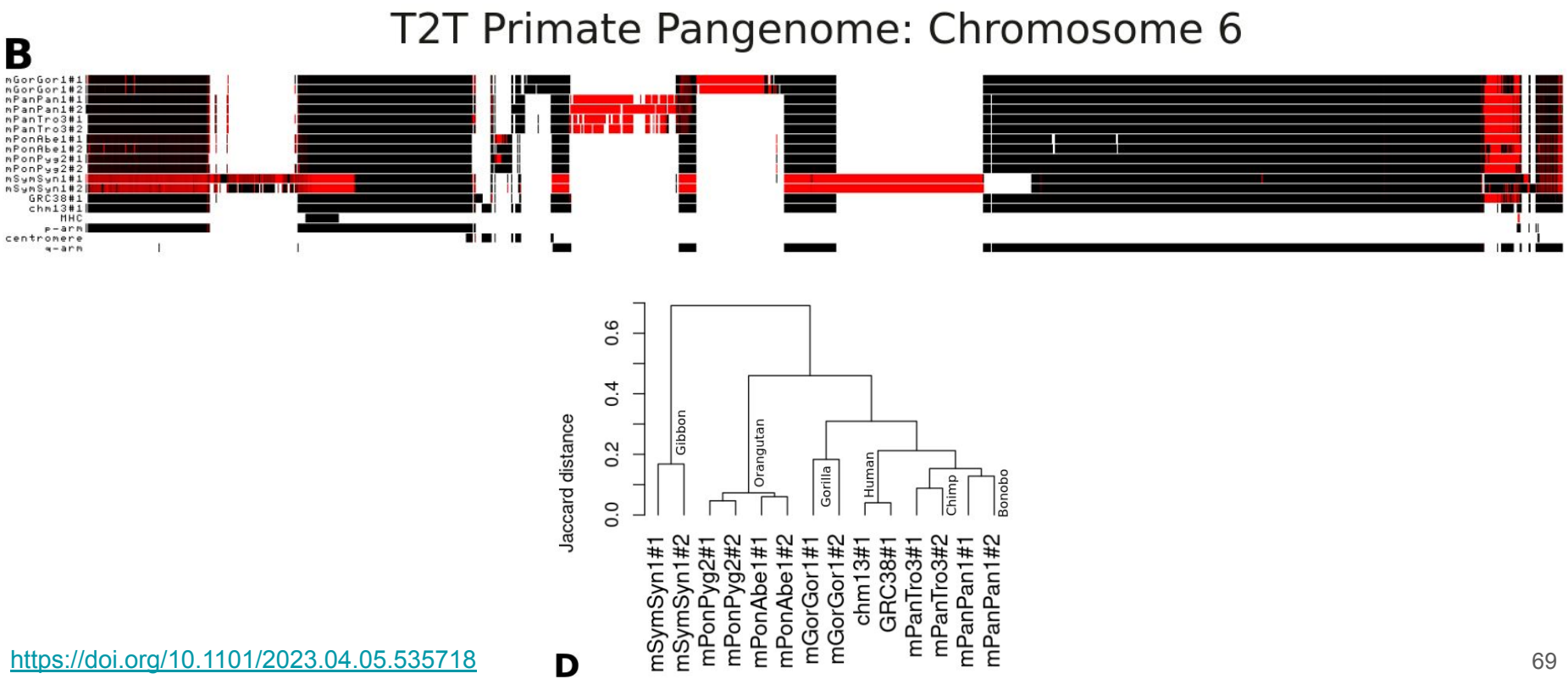
1D



2D



Phylogenetic trees with odgi similarity



https://doi.org/10.1101/2023.04.05.535718

Dissecting pangenome graphs with odgi extract

Downstream analyses may require focusing on specific *loci* in the pangenome.

Pangenome graph of the human chromosome 6 with 90 haplotypes (44 diploid *de novo* assemblies plus the GRCh38 and CHM13 reference genomes). A portion of the 2D layout is shown.

Dissecting pangenome graphs with odgi extract

Downstream analyses may require focusing on specific *loci* in the pangenome.

Pangenome graph of the human chromosome 6 with 90 haplotypes (44 diploid *de novo* assemblies plus the GRCh38 and CHM13 reference genomes). A portion of the 2D layout is shown.

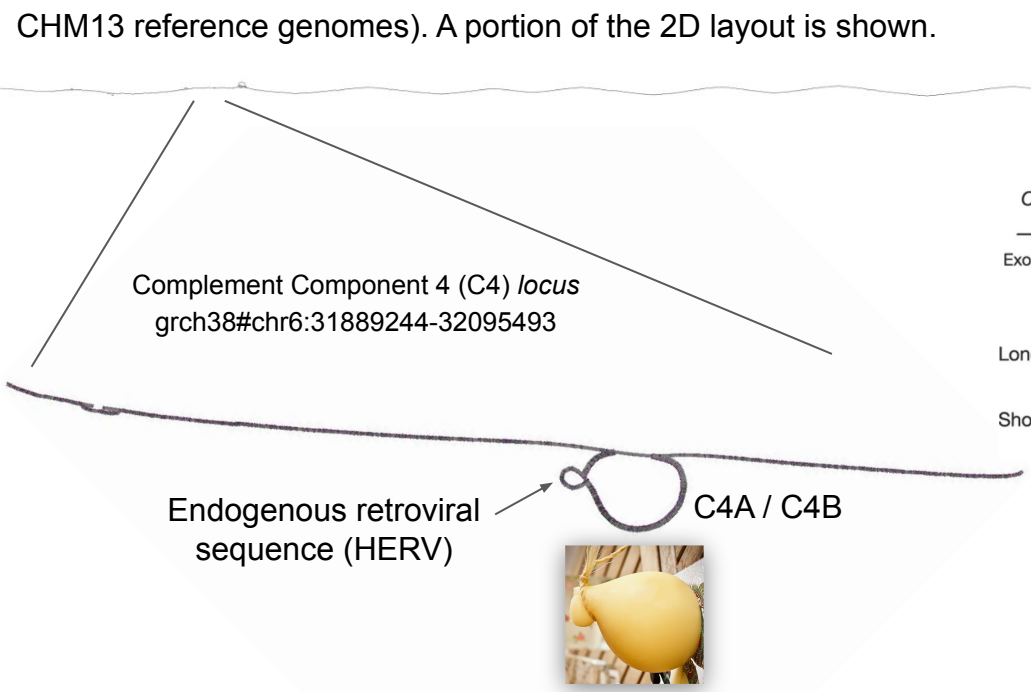
Centromere

Complement Component 4 (C4) *locus*
grch38#chr6:31889244-32095493

Dissecting pangenome graphs with odgi extract

Downstream analyses may require focusing on specific *loci* in the pangenome.

Pangenome graph of the human chromosome 6 with 90 haplotypes (44 diploid *de novo* assemblies plus the GRCh38 and CHM13 reference genomes). A portion of the 2D layout is shown.



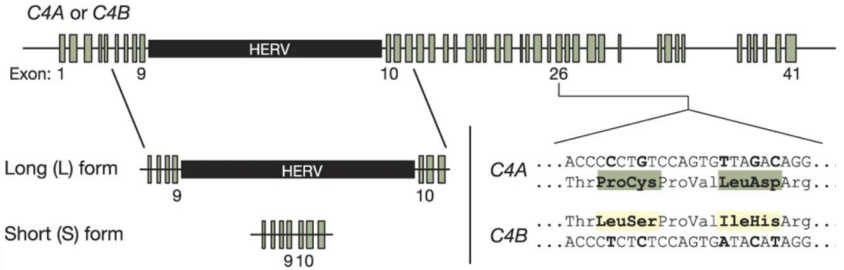


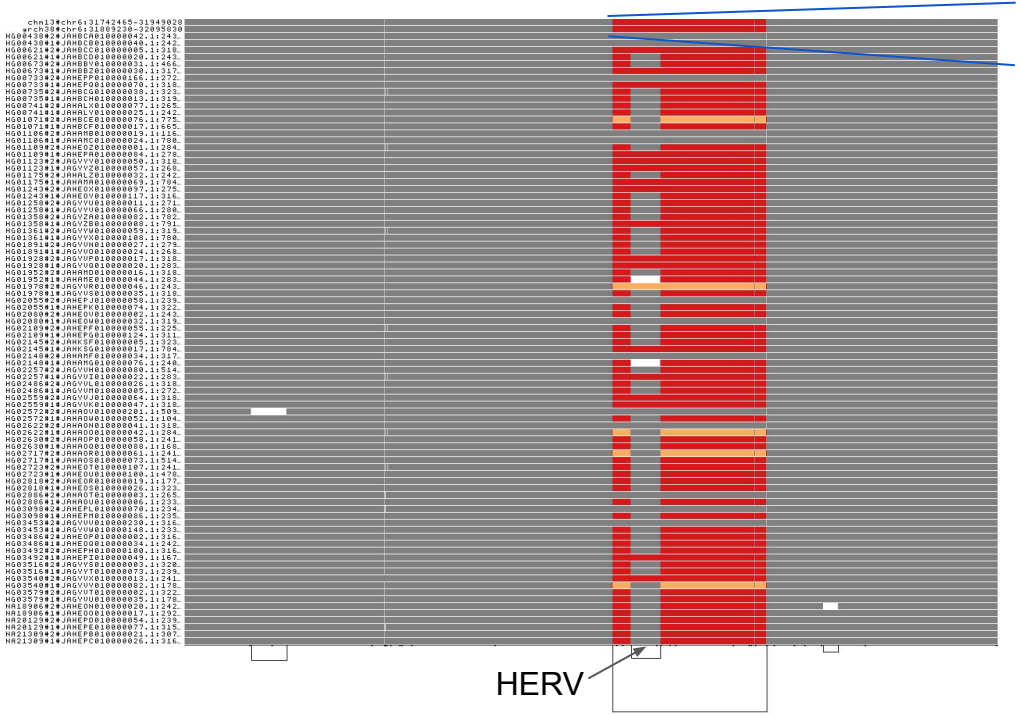
Figure from Sekar et al., 2016, Nature.

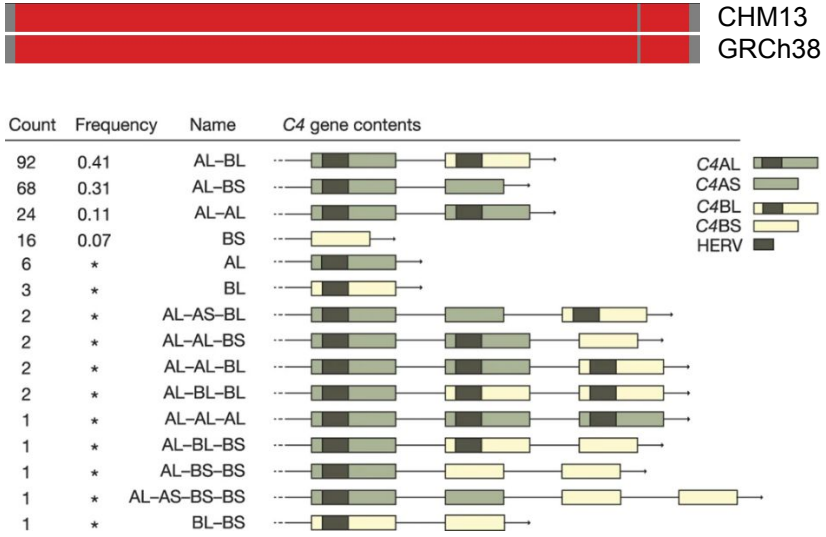
C4 *locus* complex variation is associated with Schizophrenia risk.

Dissecting pangenome graphs with odgi extract

Pangenome graph of the C4 *locus* with 90 haplotypes (44 diploid *de novo* assemblies plus the GRCh38 and CHM13 reference genomes).

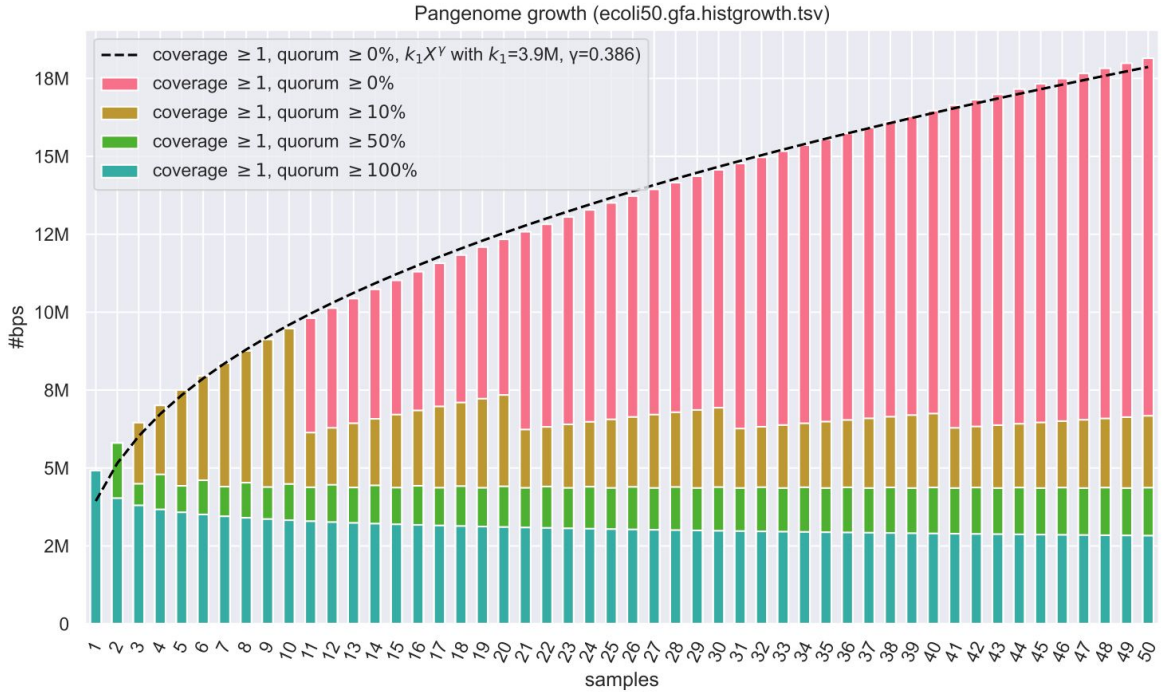
Colored by path depth (white = 0x, grey ~ 1x, red ~ 2x, yellow ~ 3x) -> Copy Number Status





222 chromosomes from 111 genetically unrelated individuals. Figure from Sekar et al., 2016, Nature.

Bonus: Pangenome growth curve with Panacus



quorum >= 0%: all seqs of all haps

quorum >= 10%: seqs traversed by at least 10% of haps - ***cloud pangenome***

quorum >= 50%: seqs traversed by at least 50% of haps - ***shell pangenome***

quorum >= 100%: seqs traversed by 100% of haps - ***core pangenome***

Activities

Learning objectives

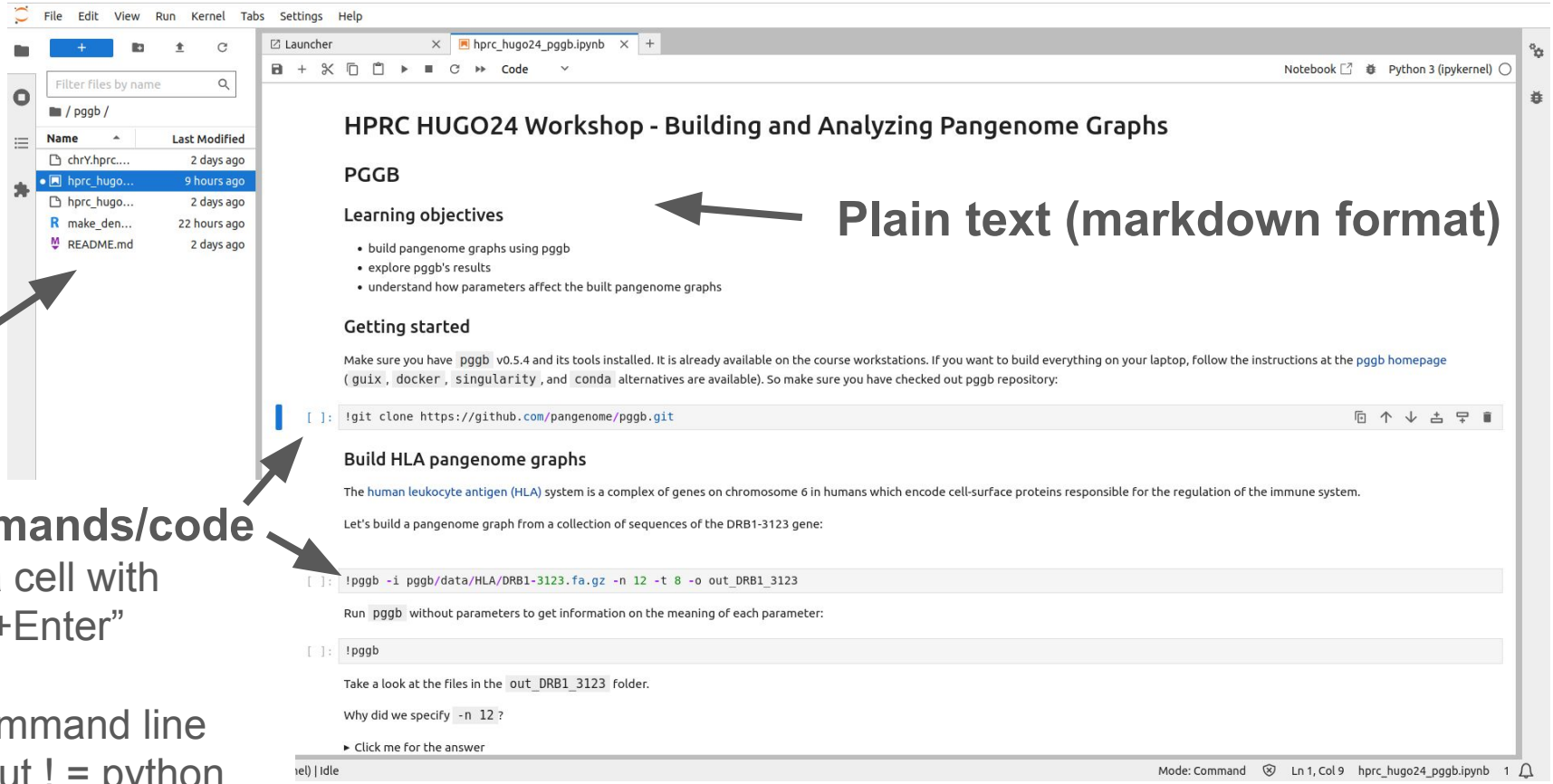
- Build pangenome graphs using `pggb`
- Explore `pggb`'s results
- Understand how different parameters affect the built pangenome graphs

- Cluster efficient pangenome graph construction with `nf-core/pangenome`

- Understanding pangenome graphs with `odgi`
  - Extract subgraphs representing *loci* of interest
  - Make phylogenetic trees
  - Identify variants in the graph

- Bonus: pangenome growth curves with `odgi heaps` and `panacus`

Find the link to the workshop's server at tinyurl.com/PangenomeHugo24



Find the link to the workshop's server at tinyurl.com/PangenomeHugo24

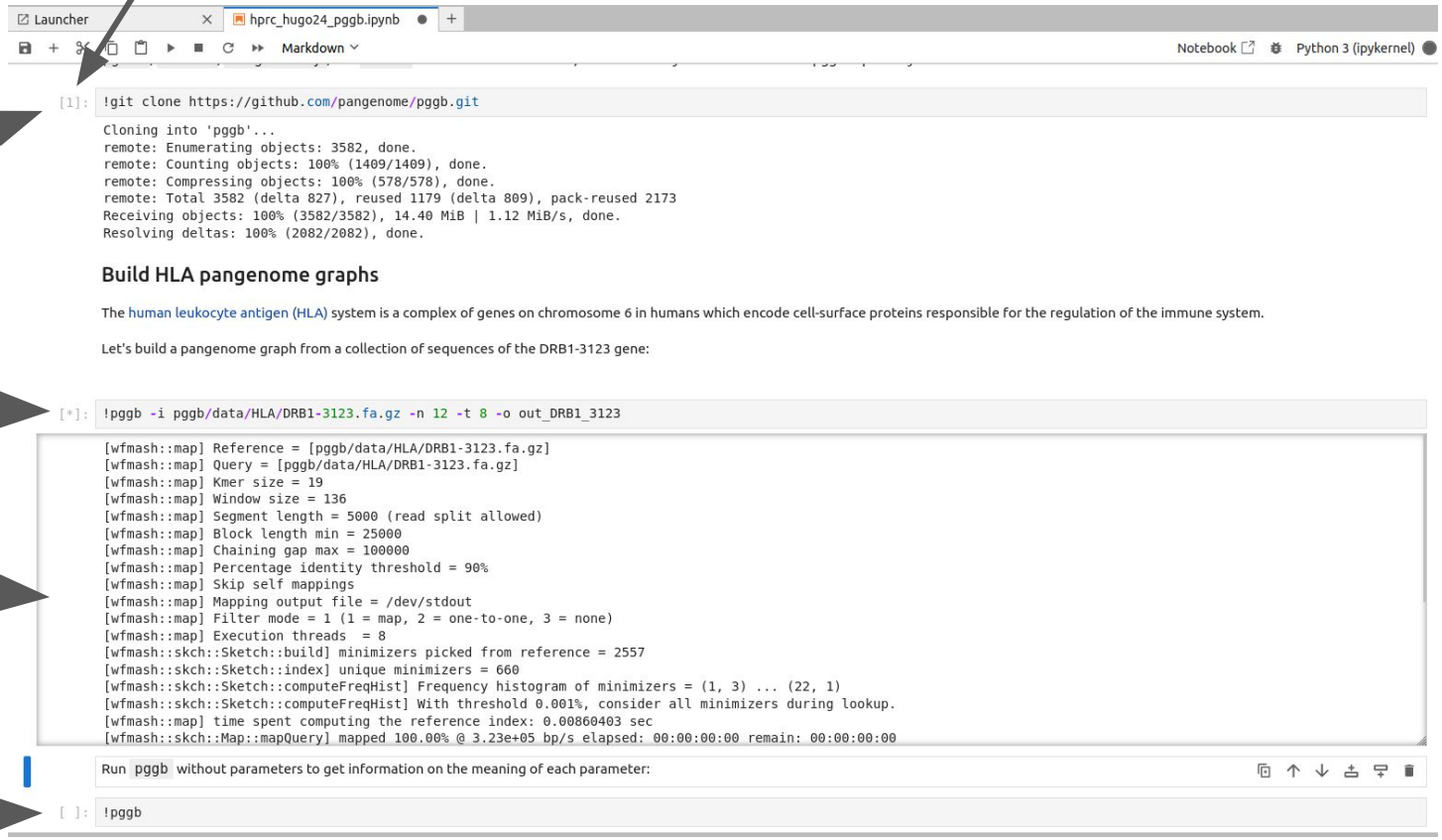
[#]
This cell has been run:
= execution order

[*]
This cell is currently being run

Output of the previous cell

[ ]
This cell has not yet been run

Cell status: [ ]



Variant calling on a pangenome with Giraffe-DeepVariant

Xian Chang & Jean Monlong

@HPRC HUGO24 Workshop, Italy, Rome
April 8, 2024

Standard variant calling pipeline

Sample DNA

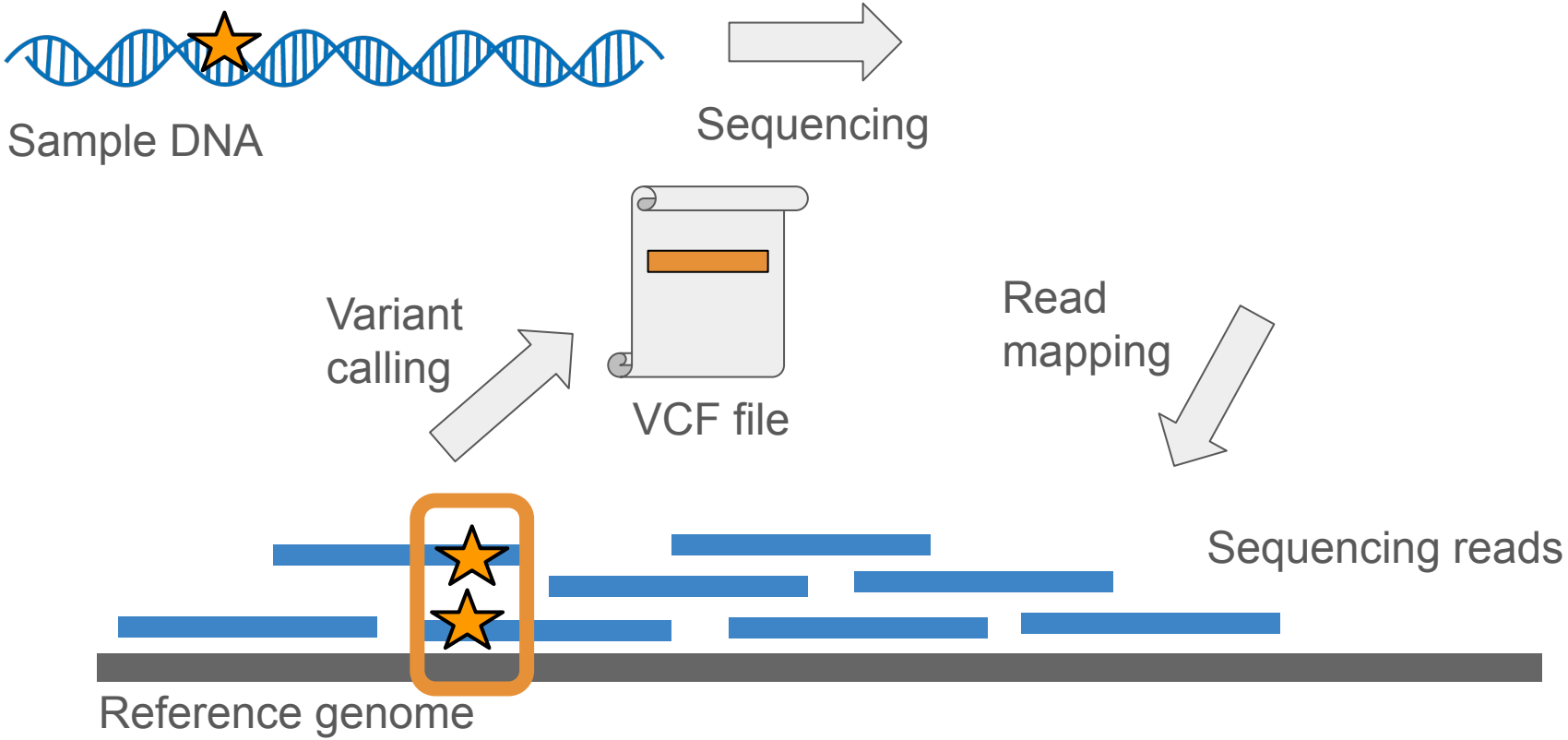
Sequencing

Sequencing reads

Read mapping

Reference genome

Standard variant calling pipeline



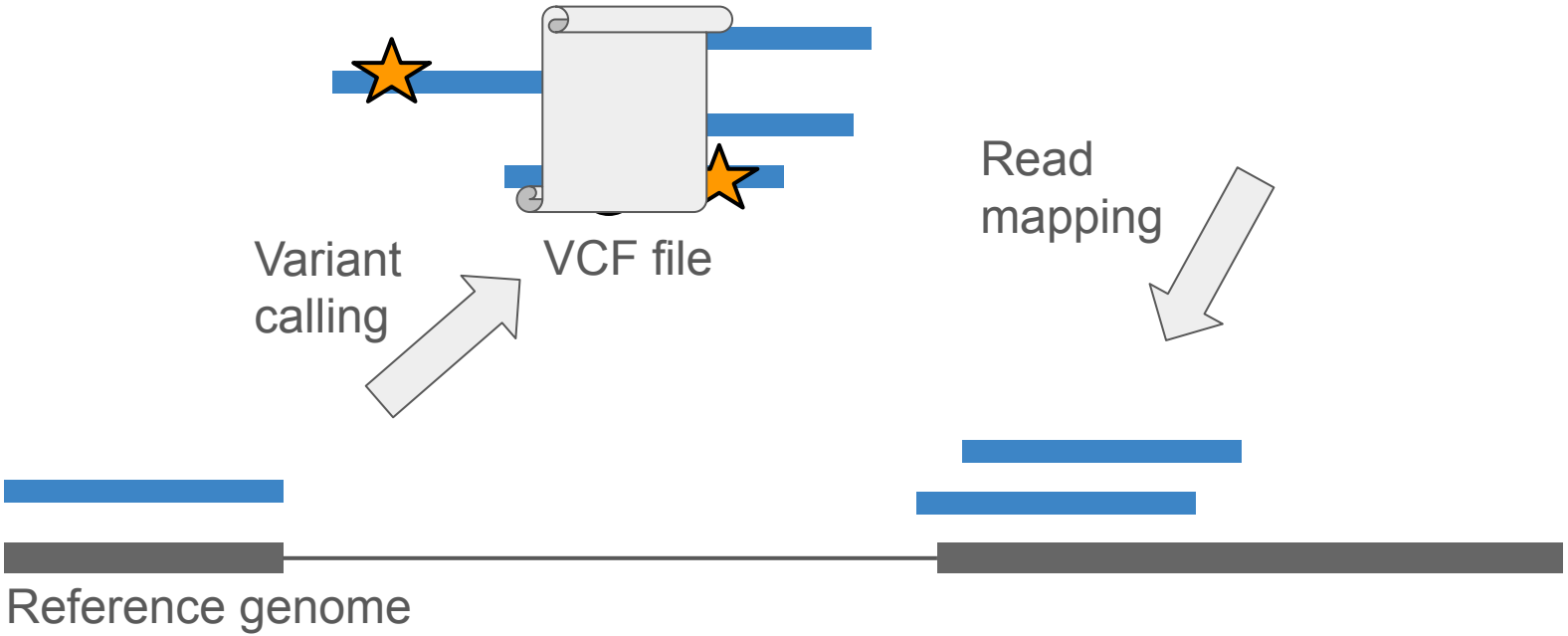
Shortcomings to standard variant calling

Sequencing reads

Read mapping

Reference genome

Shortcomings to standard variant calling



Variant calling on a pangenome

Sequencing reads

Read mapping

Reference pangenome

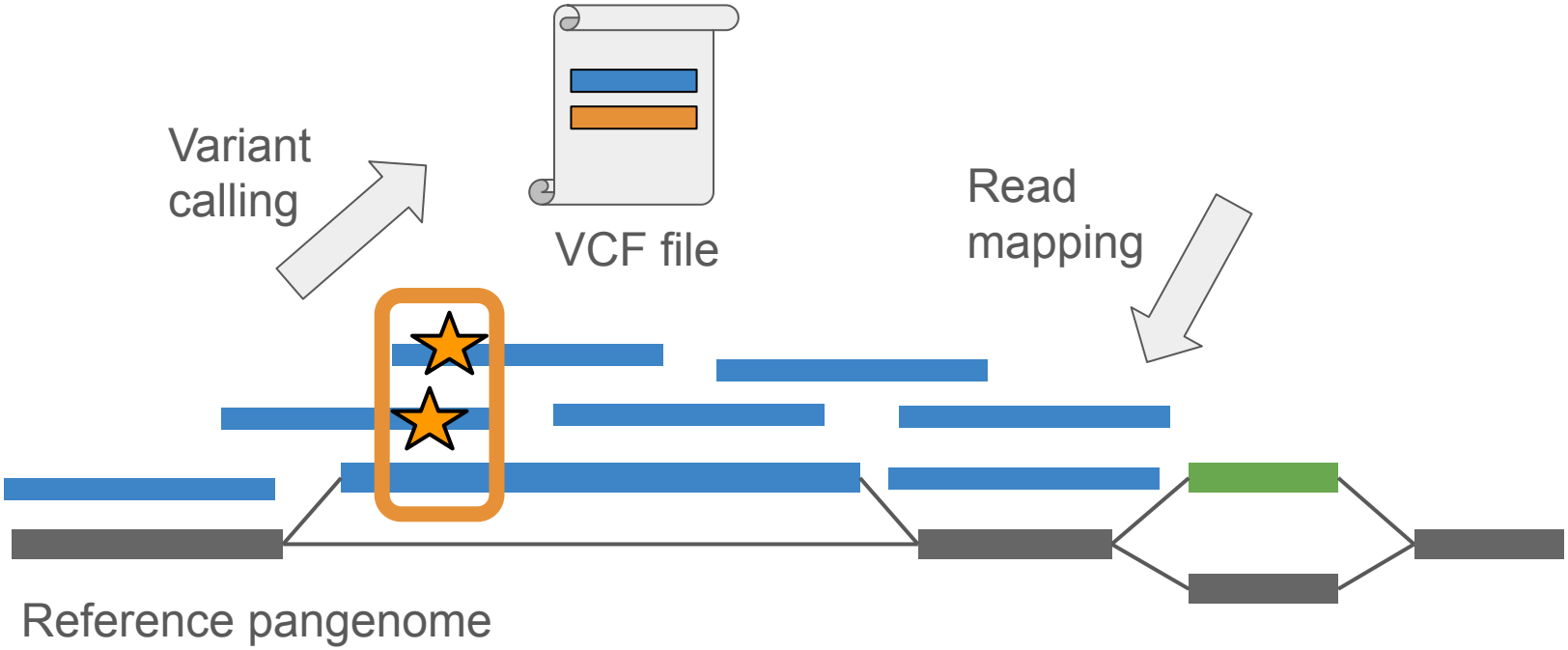
Variant calling on a pangenome

Sequencing reads

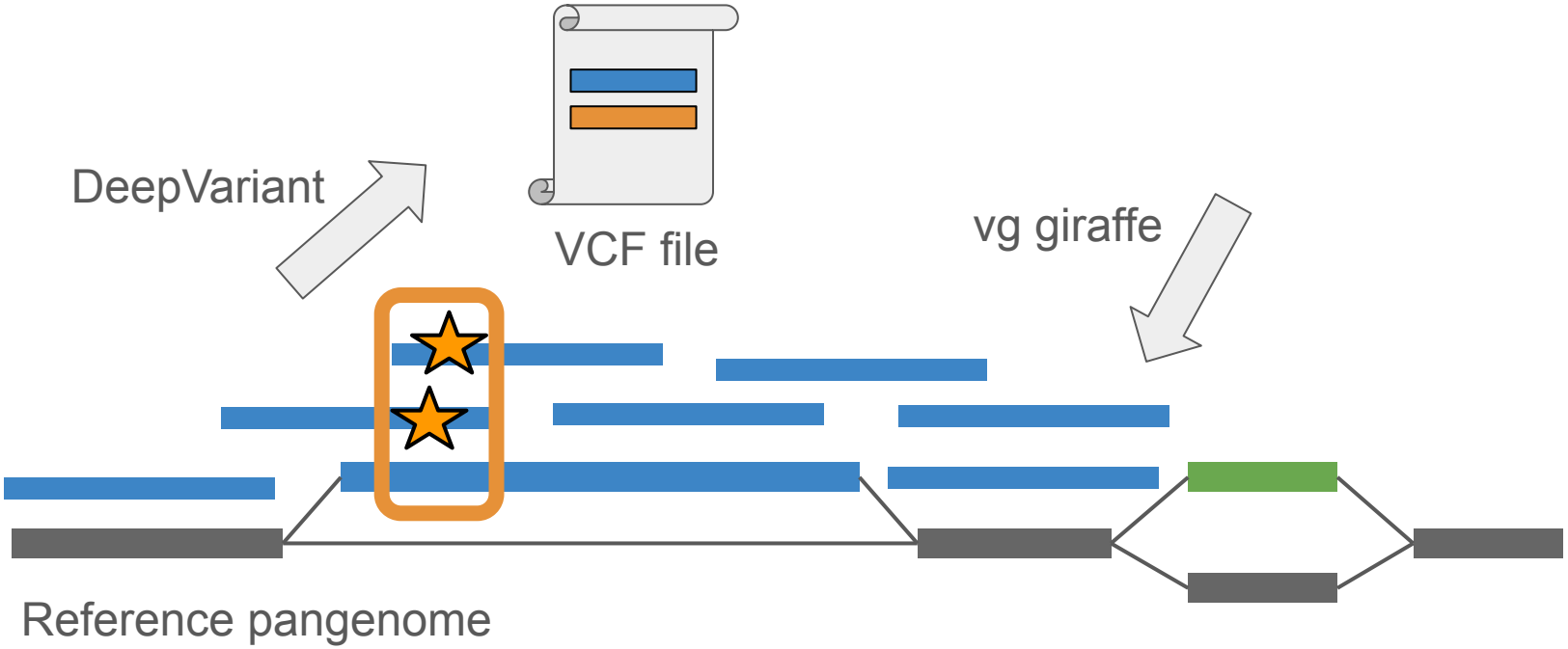
Read mapping

Reference pangenome

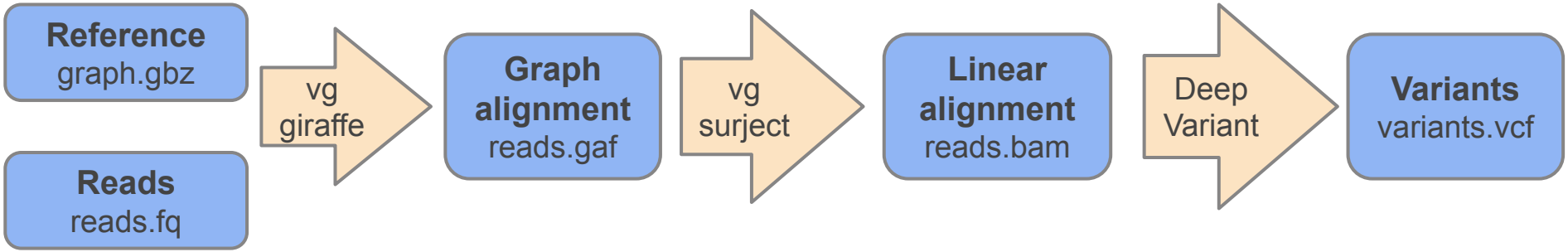
Variant calling on a pangenome



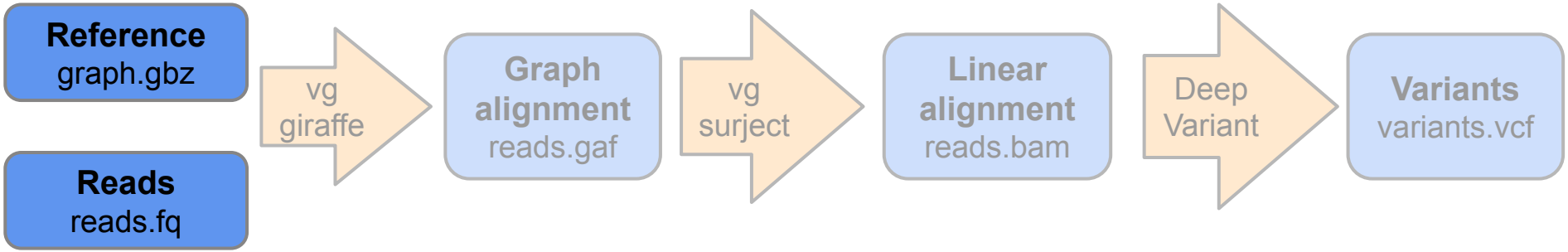
Variant calling on a pangenome



Overview of giraffe-DeepVariant pipeline



Overview of giraffe-DeepVariant pipeline

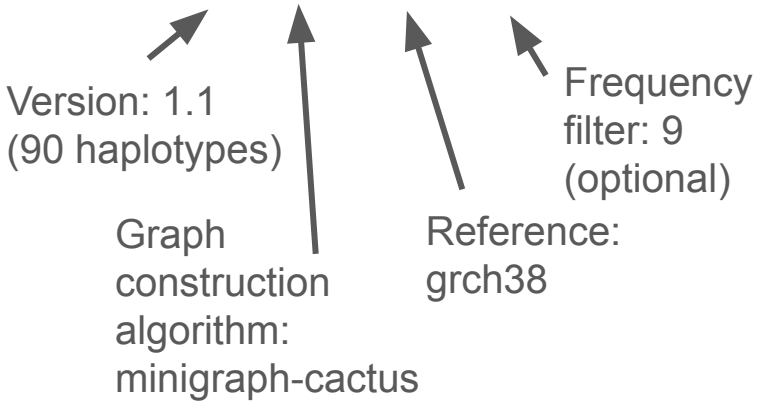


Reference: HPRC Pangenome

https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/freeze/freeze1/

General recommendation for read mapping and variant calling

hprc-v1.1-mc-grch38



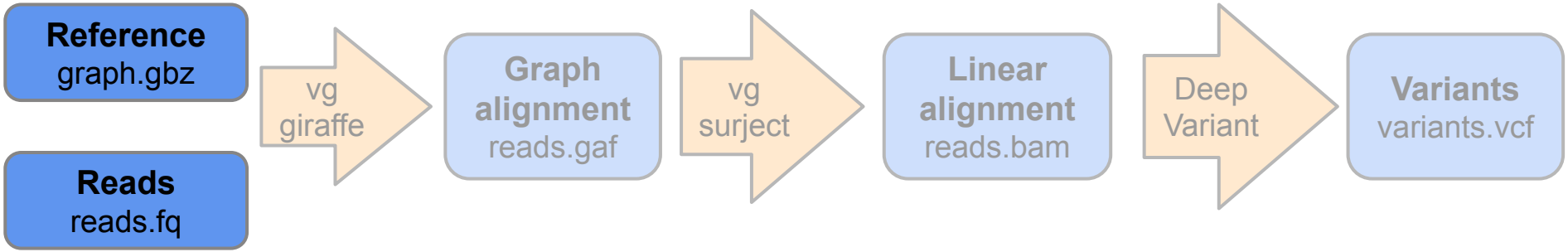
This workshop

hprc-v1.0-mc-grch38

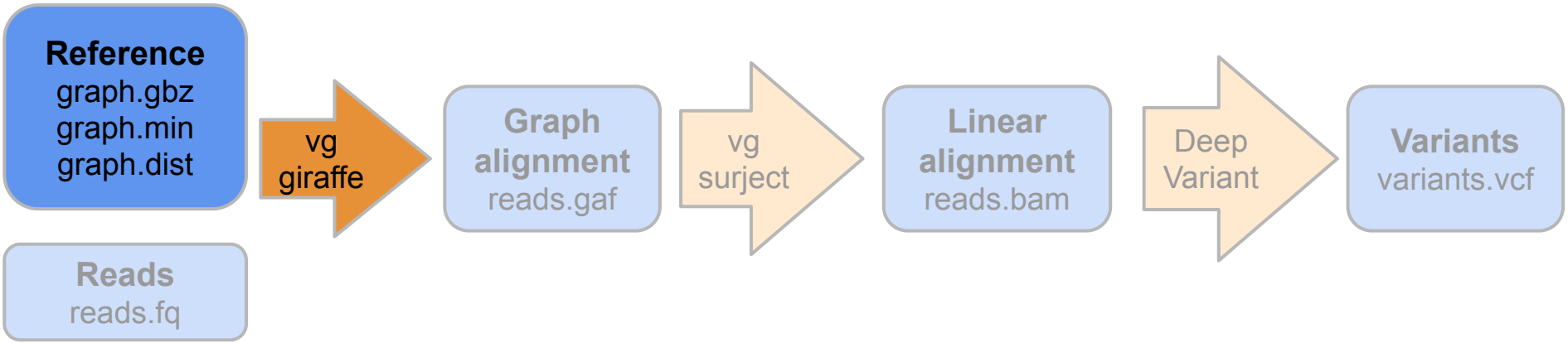
HG002 Illumina reads

https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/freeze/freeze1/minigraph-cactus/hprc-v1.1-mc-grch38/

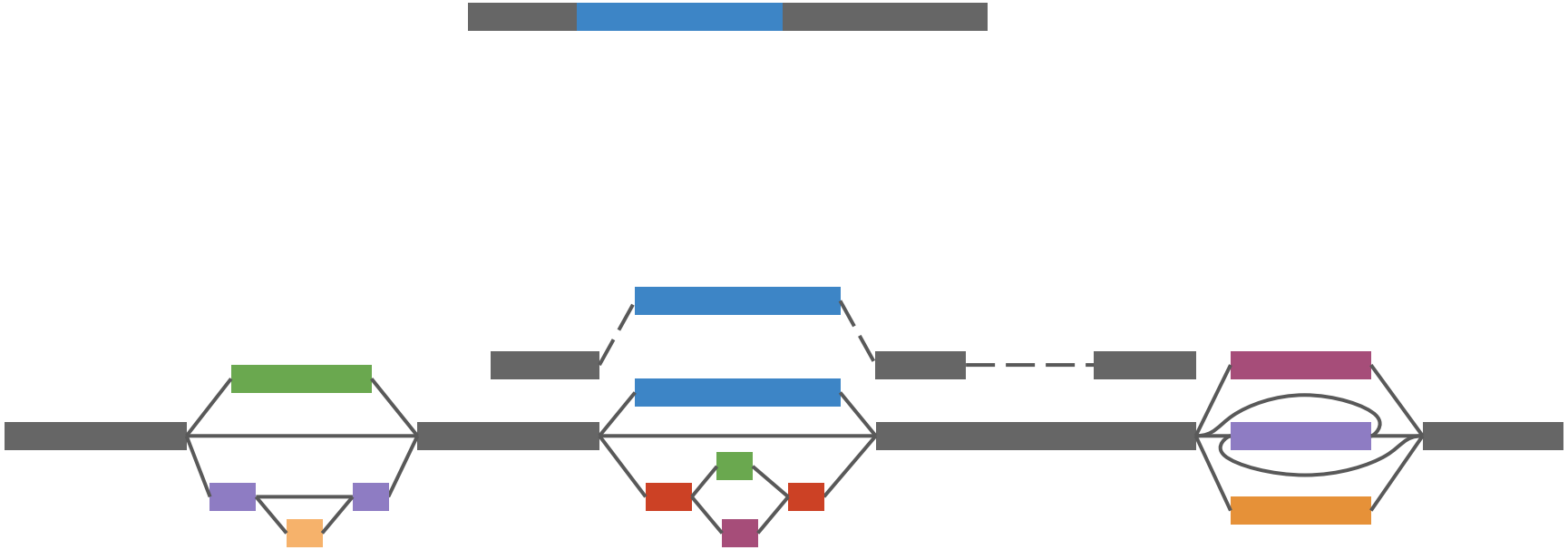
Overview of giraffe-DeepVariant pipeline



Overview of giraffe-DeepVariant pipeline

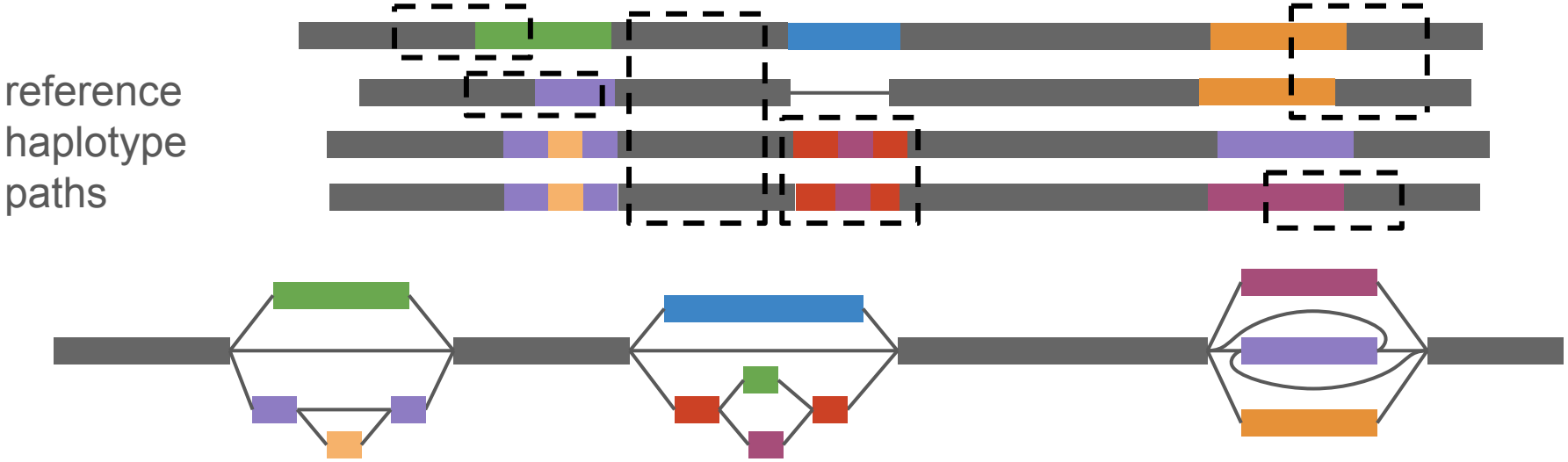


Read-to-graph alignment with giraffe



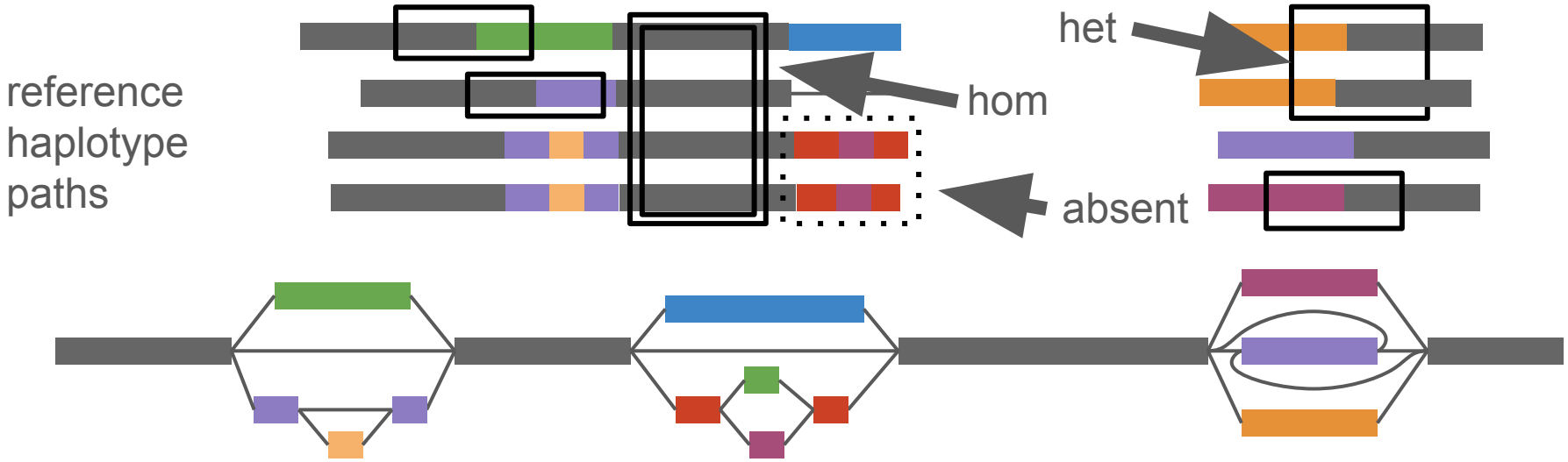
Graph preprocessing

1. Find graph-unique k-mers
2. Count k-mers in reads and classify



Graph preprocessing

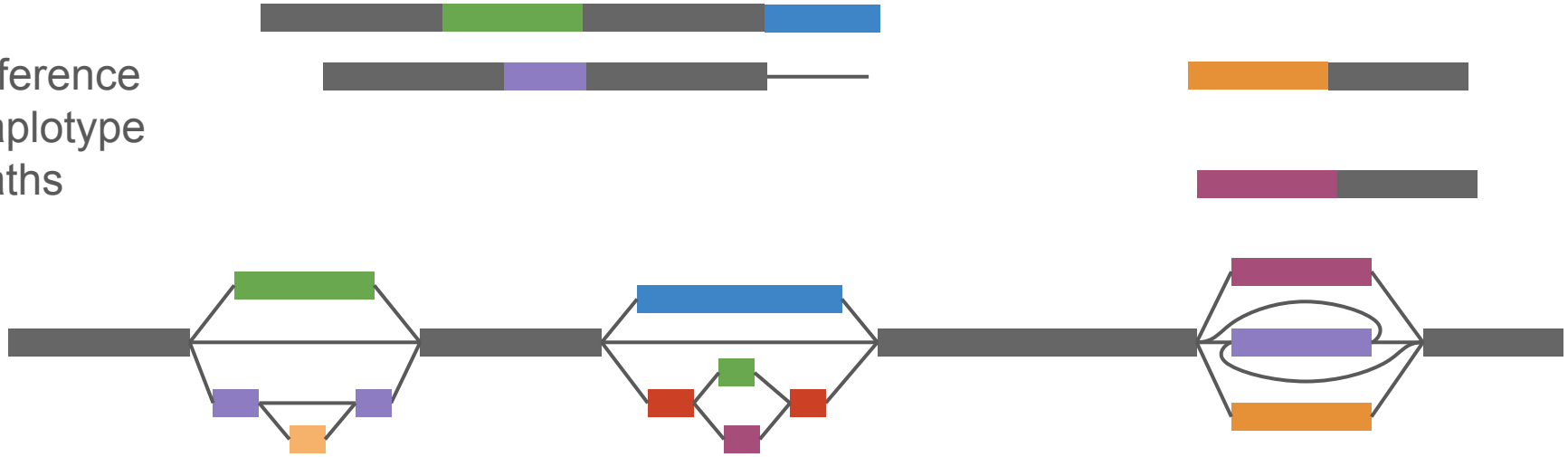
1. Find graph-unique k-mers
2. Count k-mers in reads and classify
3. Select haplotypes and subset graph



Graph preprocessing

1. Find graph-unique k-mers
2. Count k-mers in reads and classify
3. Select haplotypes and subset graph

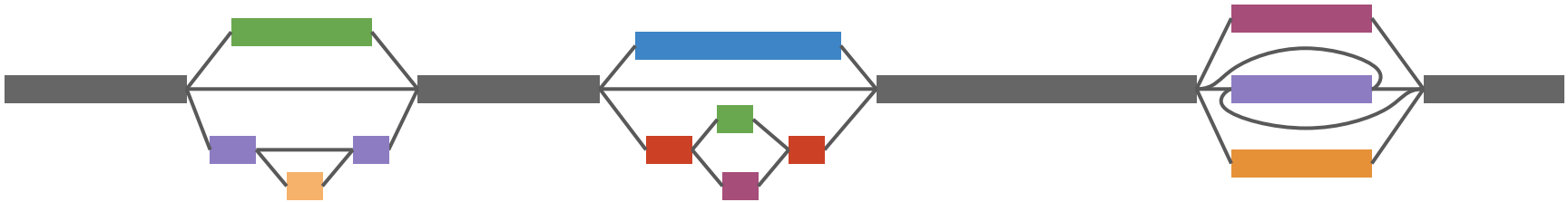
reference haplotype paths



Graph preprocessing

1. Find graph-unique k-mers
2. Count k-mers in reads and classify
3. Select haplotypes and subset graph

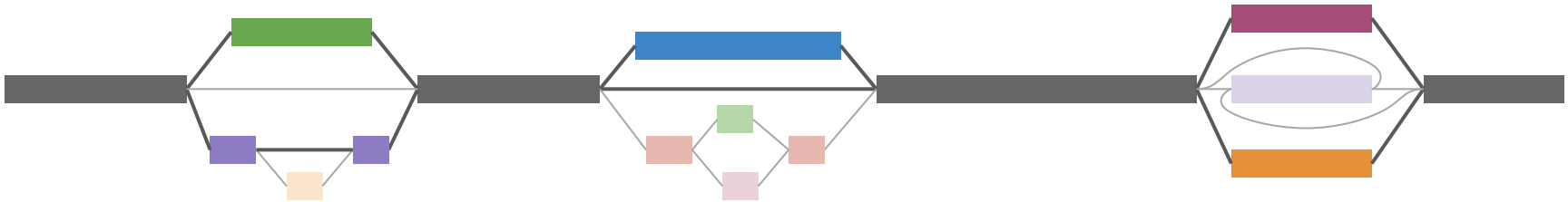
Sampled haplotypes



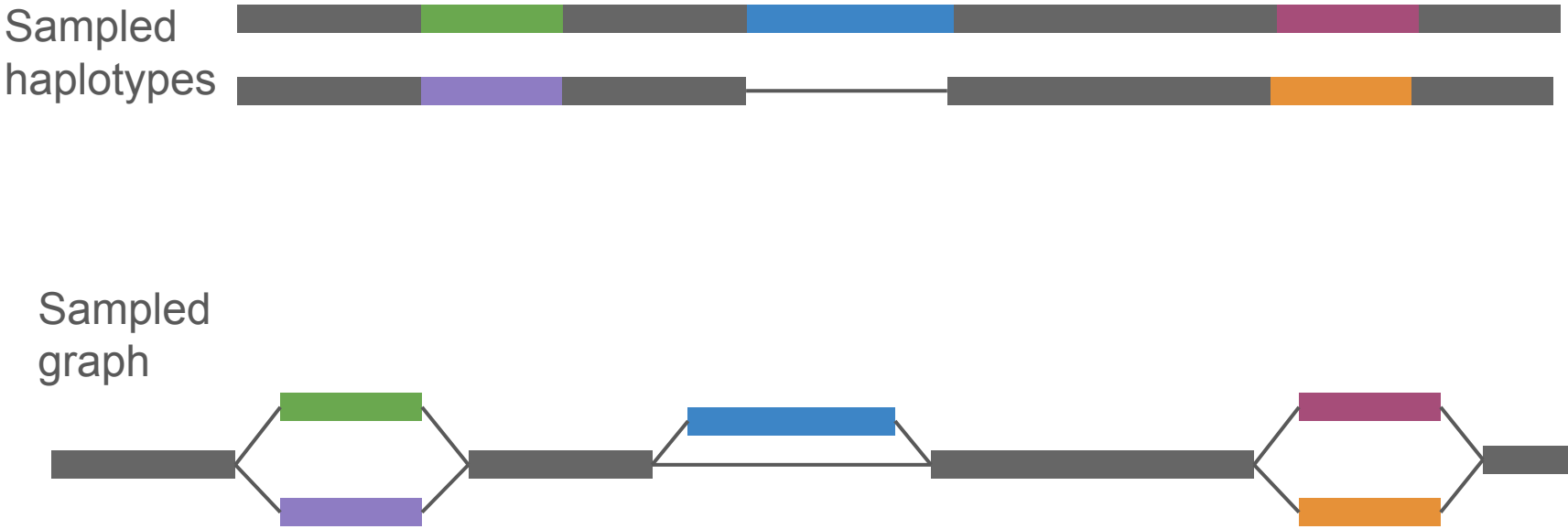
Graph preprocessing

1. Find graph-unique k-mers
2. Count k-mers in reads and classify
3. Select haplotypes and subset graph

Sampled haplotypes

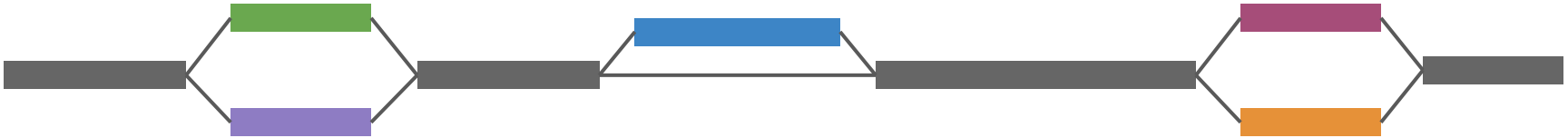


Graph preprocessing



Giraffe algorithm

Giraffe data structures

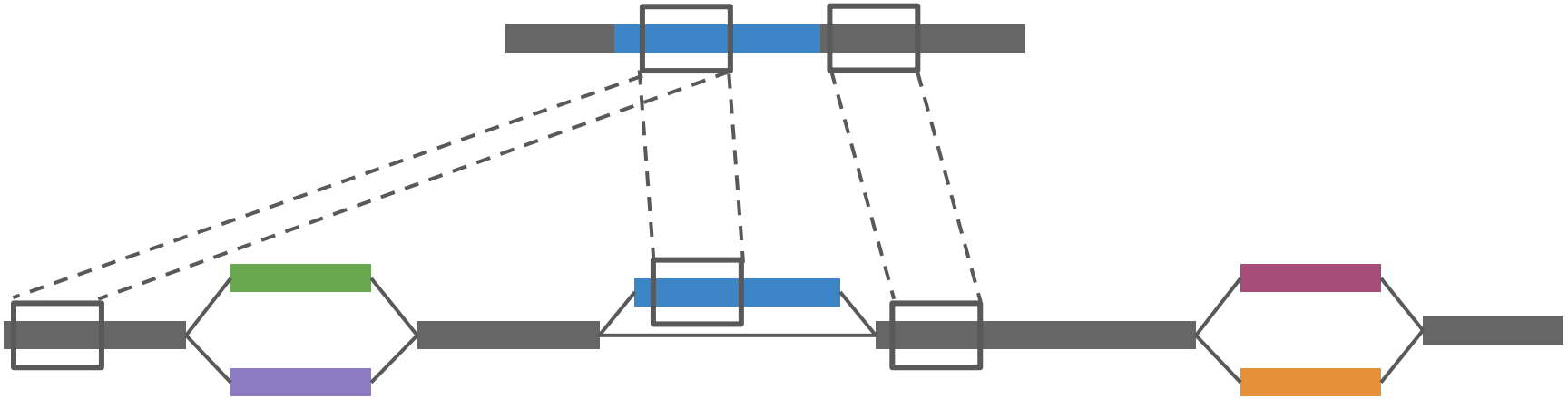


Giraffe algorithm

- Seed

Giraffe data structures

- Minimizer index (graph.min)

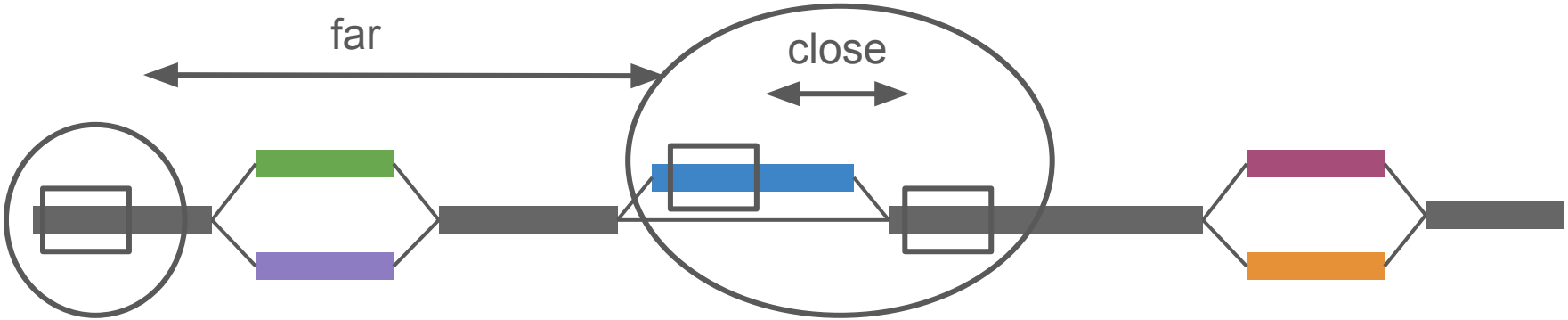


Giraffe algorithm

- Seed
- Cluster

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)

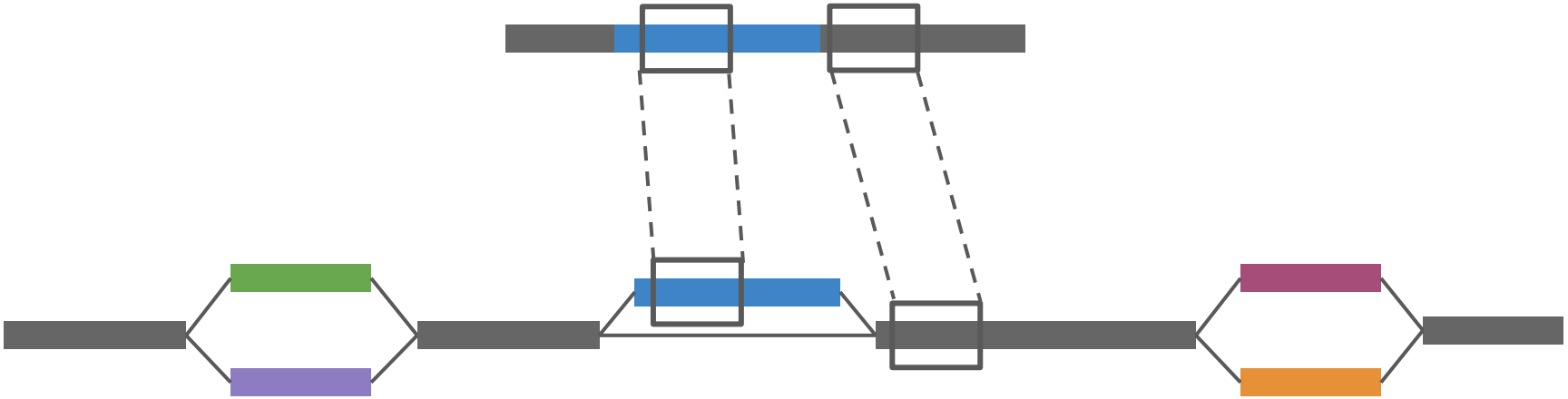


Giraffe algorithm

- Seed
- Cluster
- Gapless extension

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)

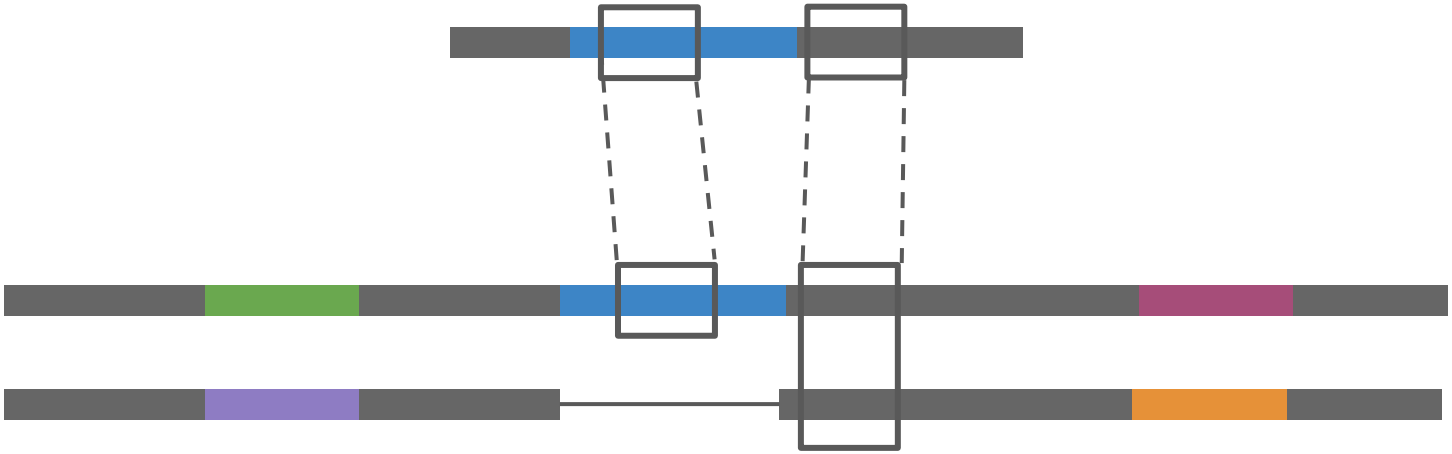


Giraffe algorithm

- Seed
- Cluster
- Gapless extension

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)

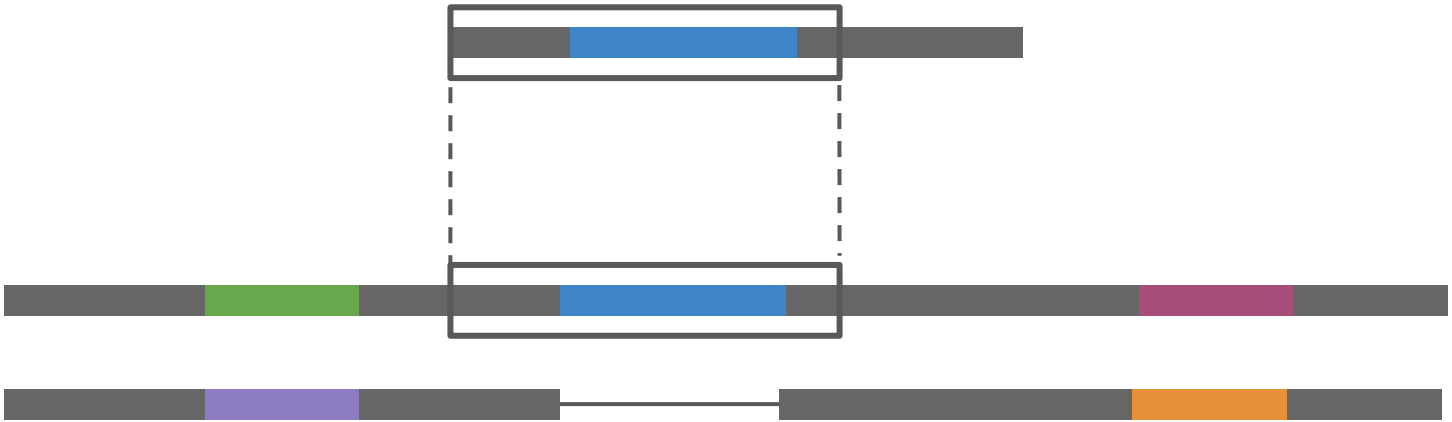


Giraffe algorithm

- Seed
- Cluster
- Gapless extension

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)

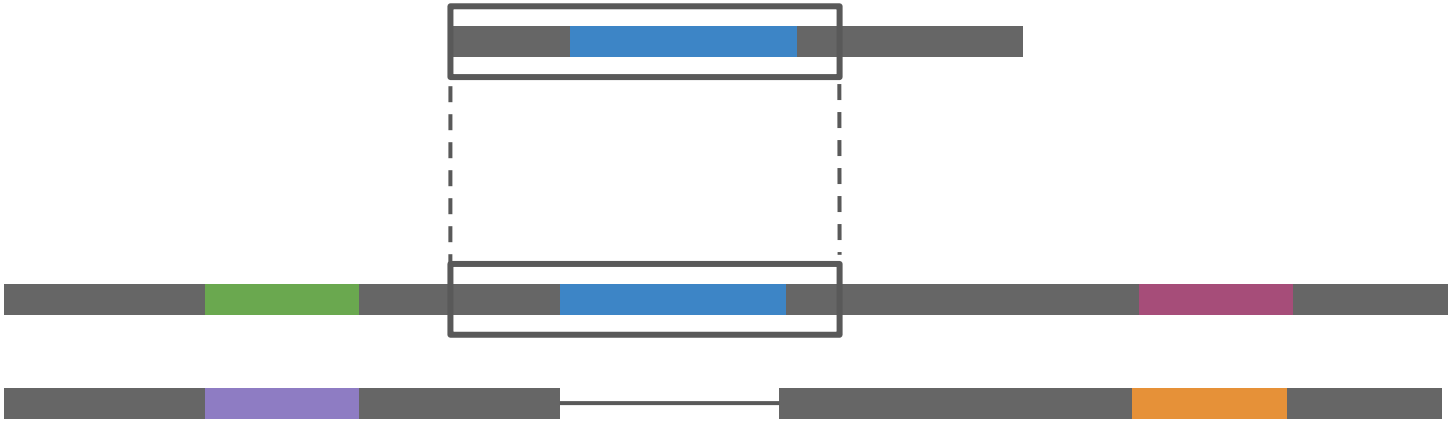


Giraffe algorithm

- Seed
- Cluster
- Gapless extension
- Gapped alignment

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)
- Graph (graph.gg/gbz/xg/hg)

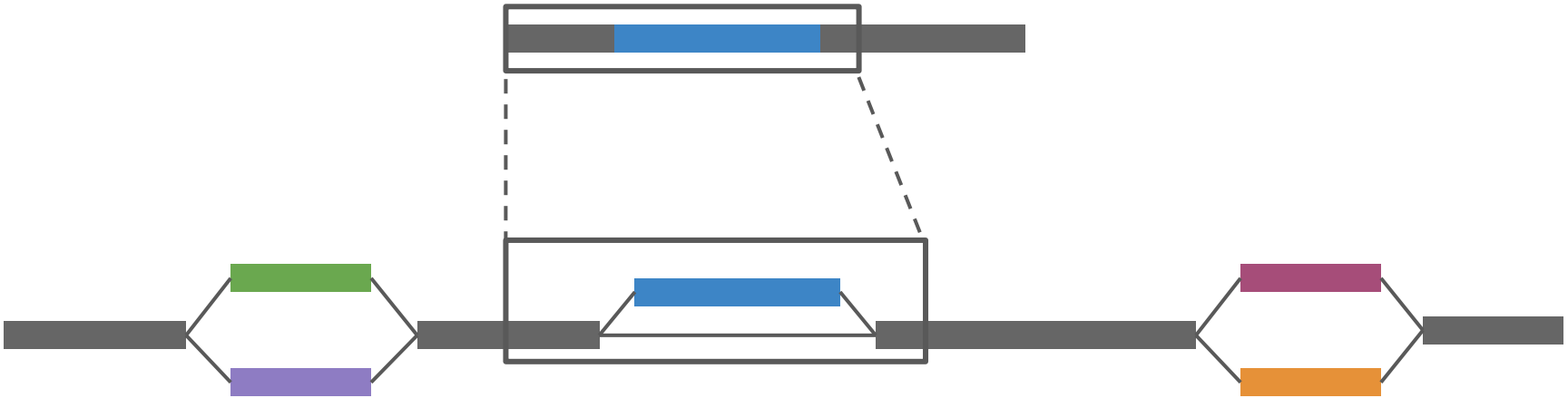


Giraffe algorithm

- Seed
- Cluster
- Gapless extension
- Gapped alignment

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)
- Graph (graph.gg/gbz/xg/hg)

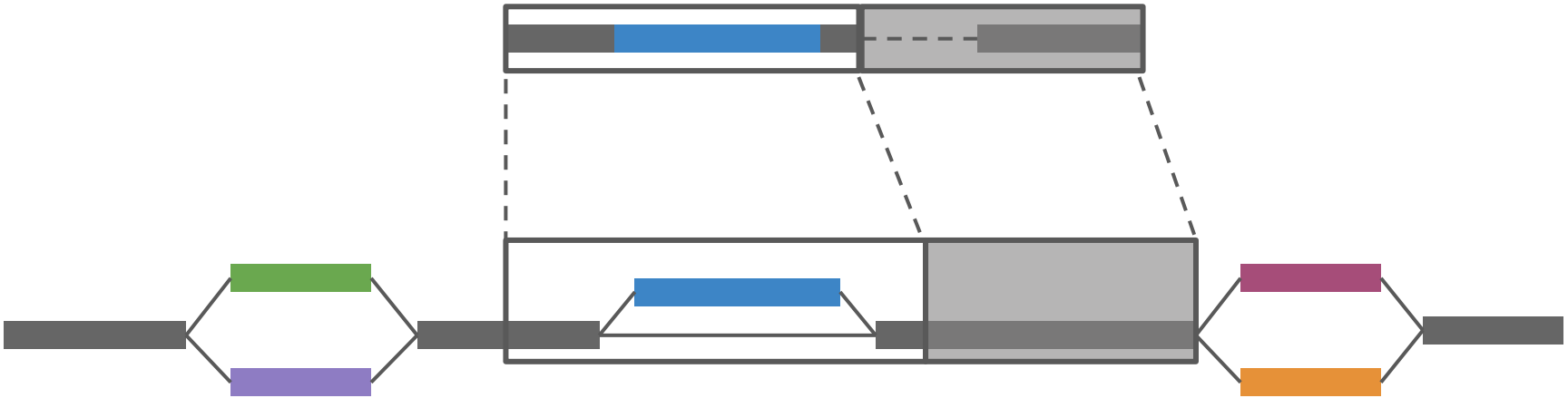


Giraffe algorithm

- Seed
- Cluster
- Gapless extension
- Gapped alignment

Giraffe data structures

- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)
- Graph (graph.gg/gbz/xg/hg)

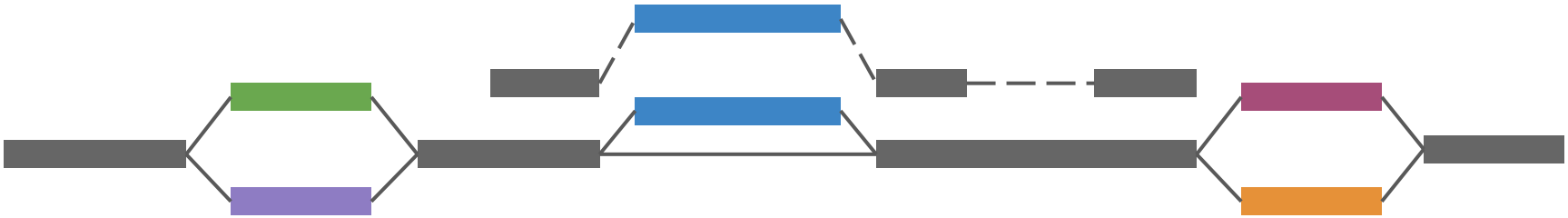


Giraffe algorithm

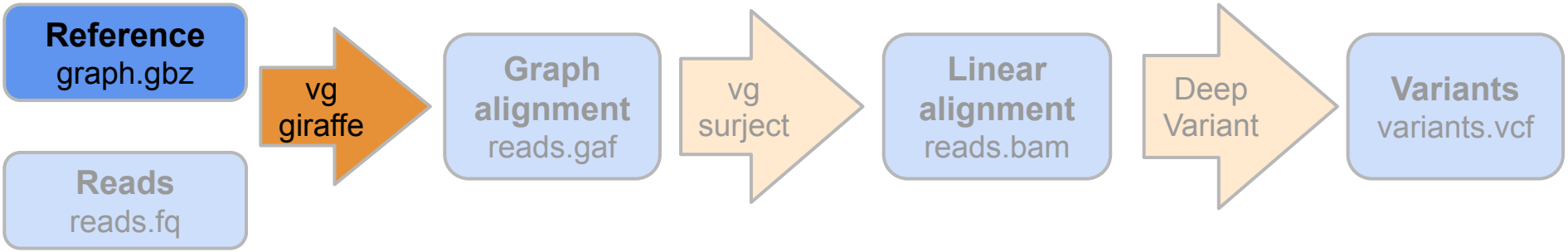
- Seed
- Cluster
- Gapless extension
- Gapped alignment

Giraffe data structures

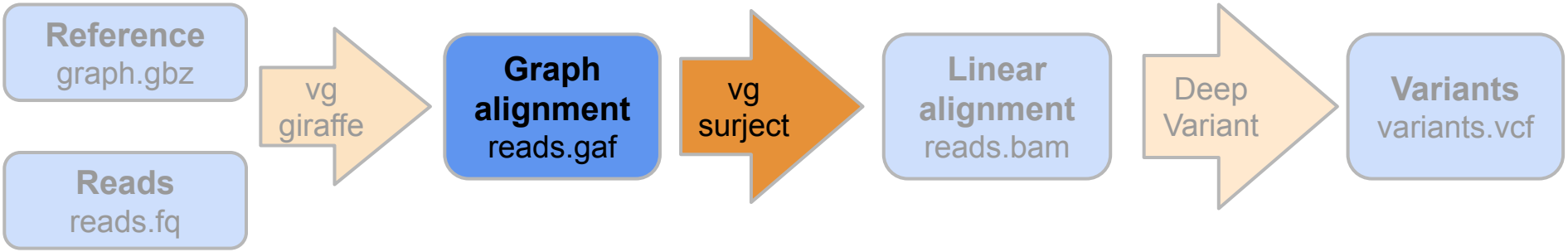
- Minimizer index (graph.min)
- Distance index (graph.dist)
- Haplotypes (graph.gbwt/gbz)
- Graph (graph.gg/gbz/xg/hg)



Overview of giraffe-DeepVariant pipeline

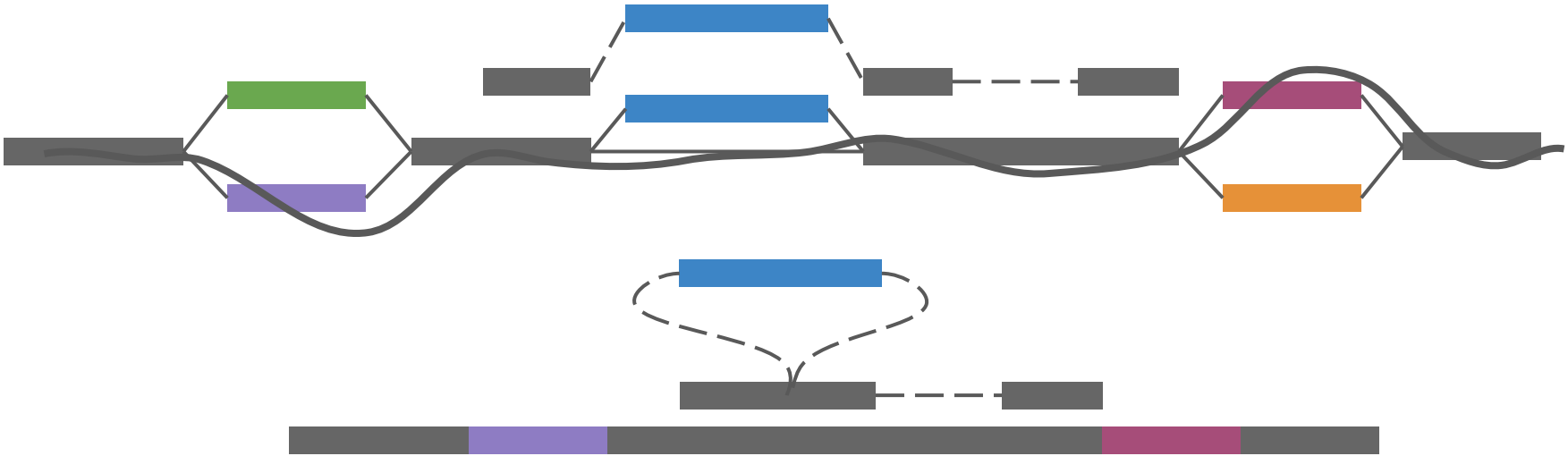


Overview of giraffe-DeepVariant pipeline



vg surject

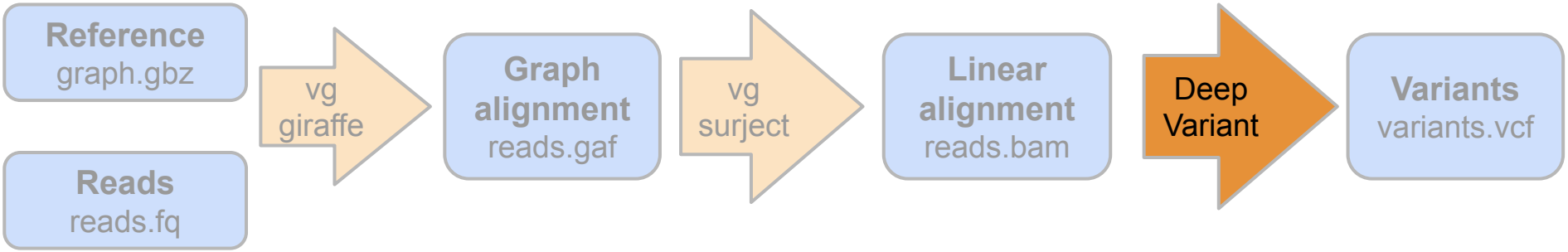
Project graph alignment onto reference genome



Reference genome

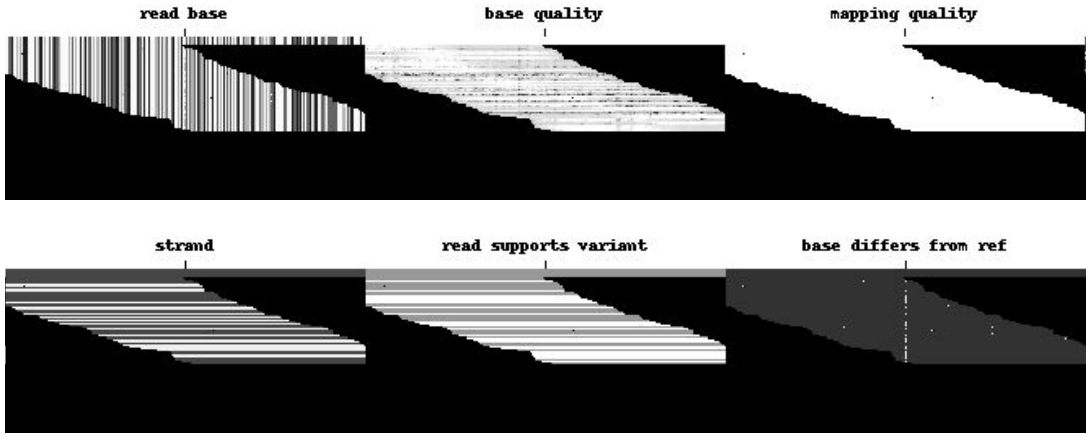
vg surject

Overview of giraffe-DeepVariant pipeline



DeepVariant

Machine learning approach to predict genotype at a candidate site from read pile-up images.



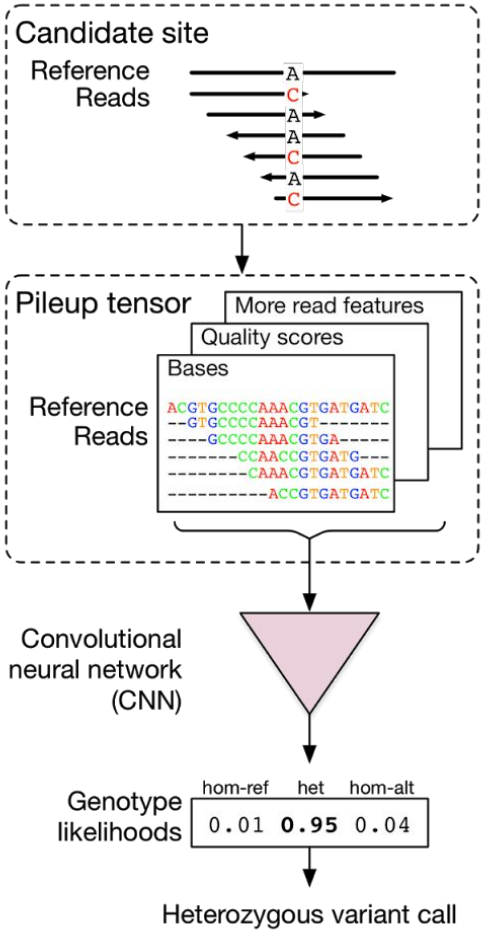
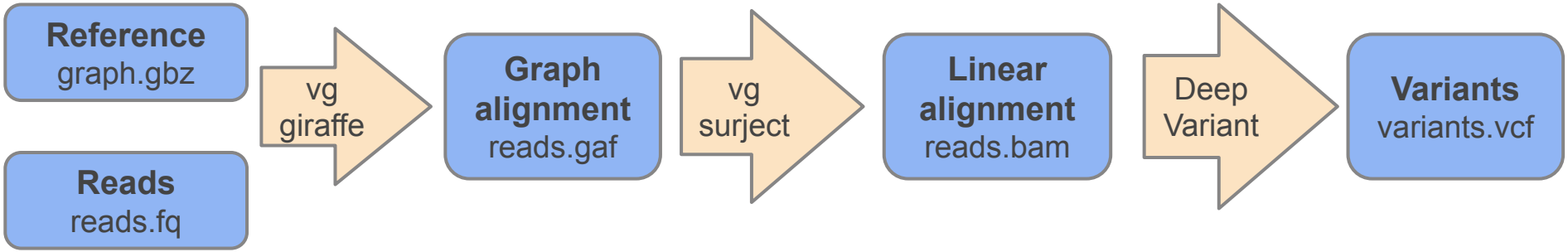


Image: https://google.github.io/deepvariant/

Overview of giraffe-DeepVariant pipeline



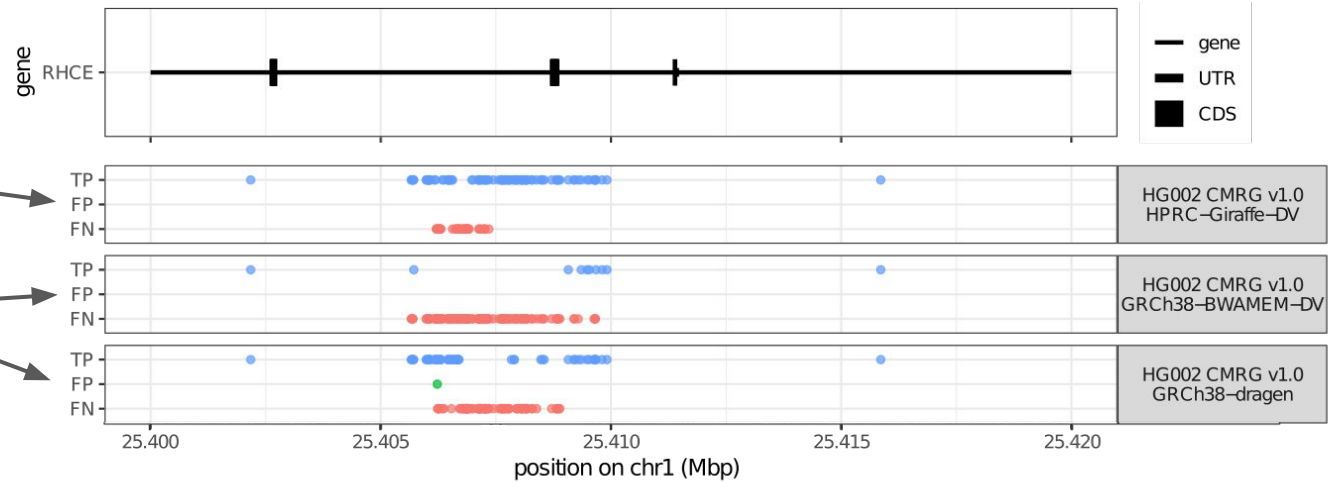
Demo: Hands-On Portion

Find the link to the workshop's server at tinyurl.com/PangenomeHugo24

RHCE: A Challenging, Medically-Relevant Gene

Mostly true positives with HPRC-Giraffe+DeepVariant

Many false negative with BWA+DeepVariant or Dragen



Each point is a variant called on HG002

Complex Variation Around RHCE

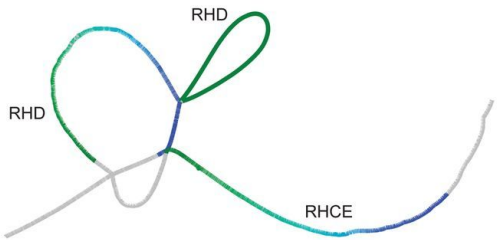
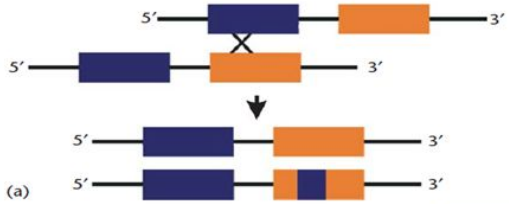


Image from Wikipedia

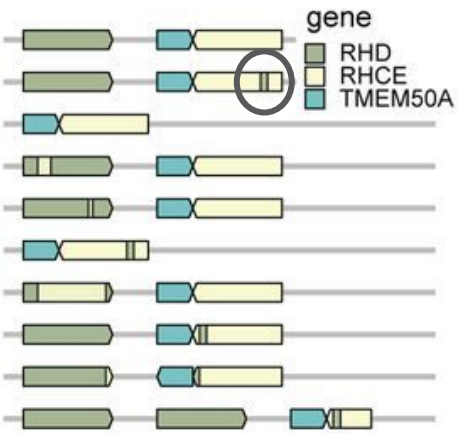


Gene conversion event

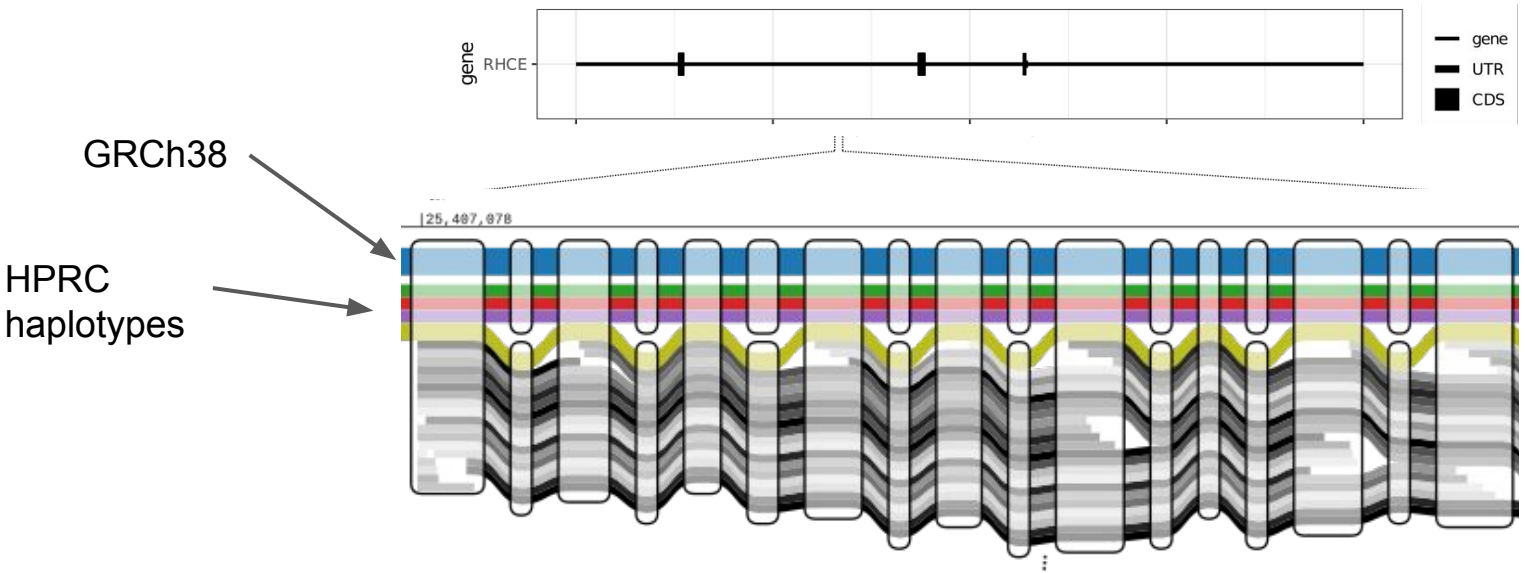
GRCh38

HG002

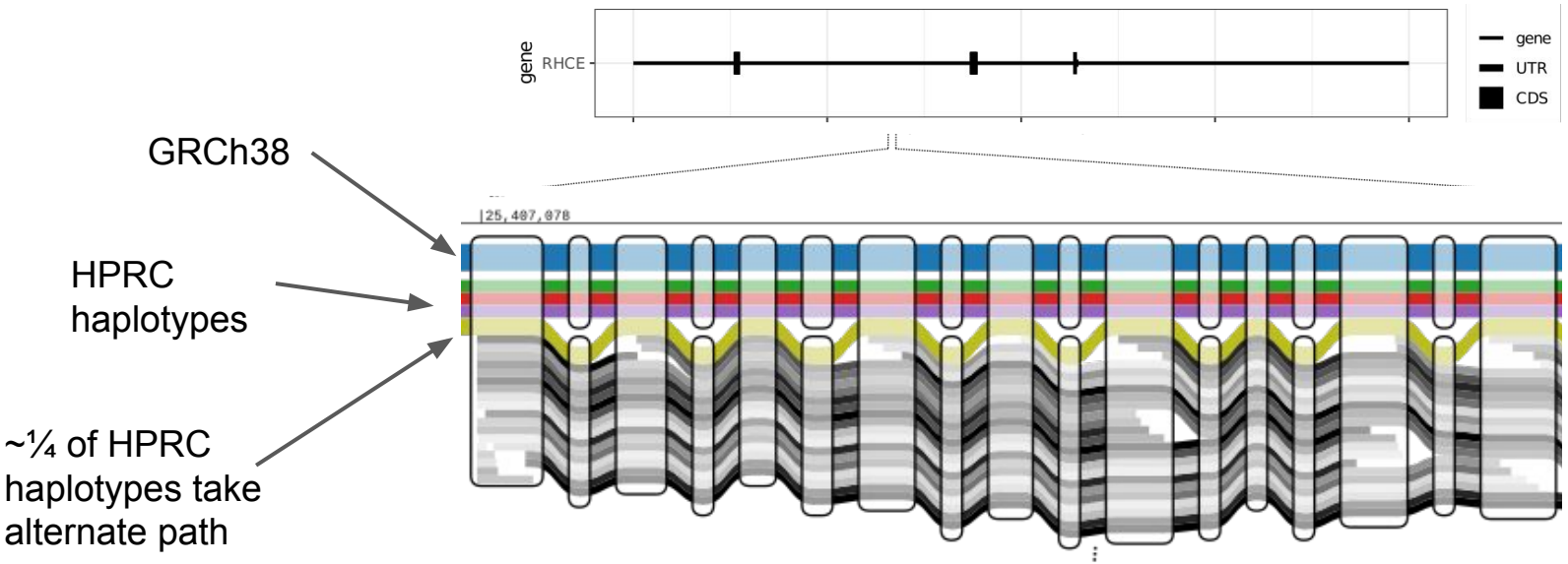
Count	Frequency	Haplotype name
43	0.48	RHD;RHCE
22	0.24	RHD;RHCE-RHD(2)-RHCE
15	0.17	RHCE
3	0.03	RHD-RHCE(2-3)-RHD;RHCE
2*	0.02	RHD-RHCE(8)-RHD;RHCE
1	0.01	RHCE-RHD(2)-RHCE
1*	0.01	RHD-RHCE(2-9)-RHD;RHCE
1*	0.01	RHD;RHCE-RHD(9)-RHCE
1*	0.01	RHD-RHCE(10);inv;RHCE-RHD(10)
1*	0.01	RHD;RHD;RHCE-RHD(9)-RHCE



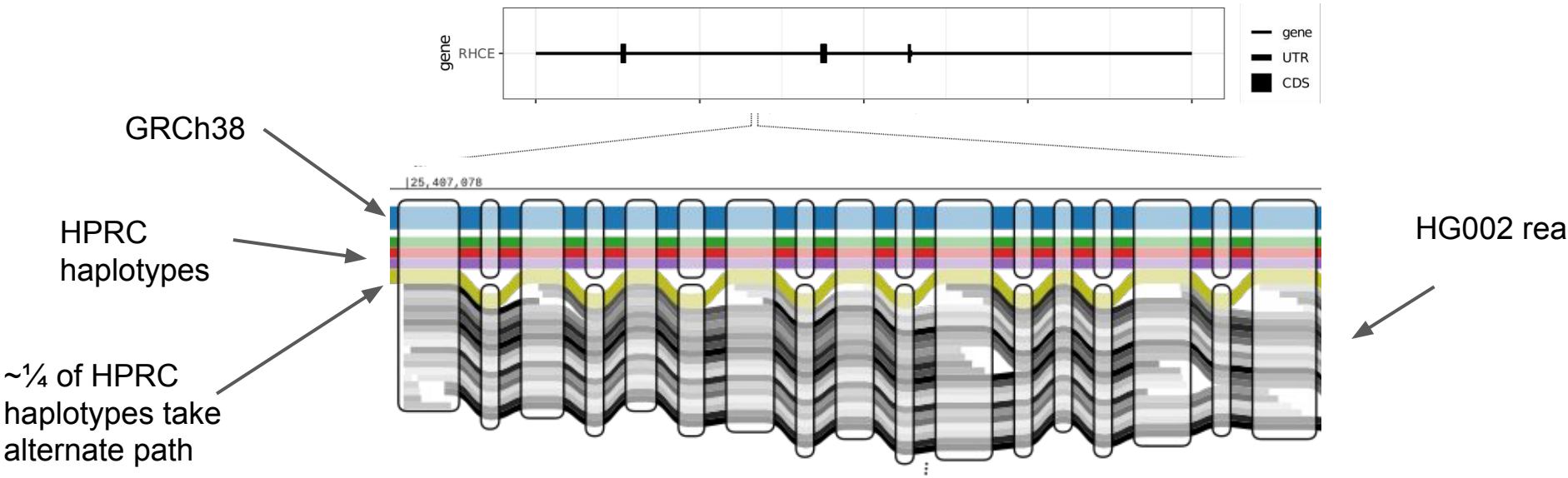
Small Reads Map To A Subset Of Haplotypes



Small Reads Map To A Subset Of Haplotypes

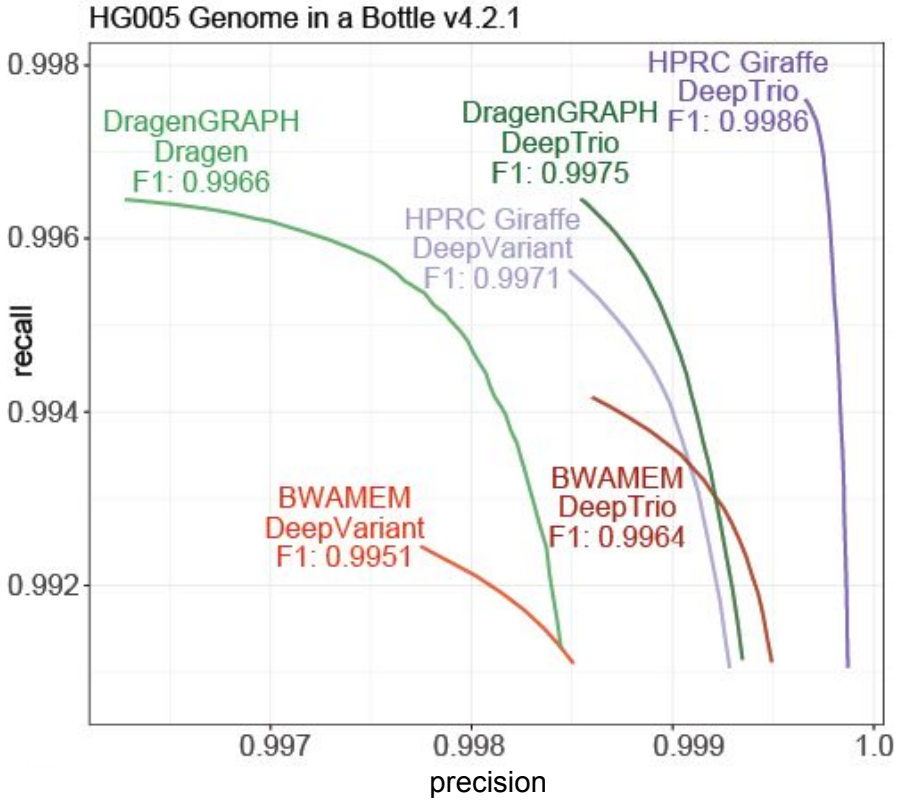


Small Reads Map To A Subset Of Haplotypes



Giraffe/DeepVariant Improves Variant Calling

On average, 34% less errors compared to the linear reference approach (GRCh38-BWA-DeepVariant)



END