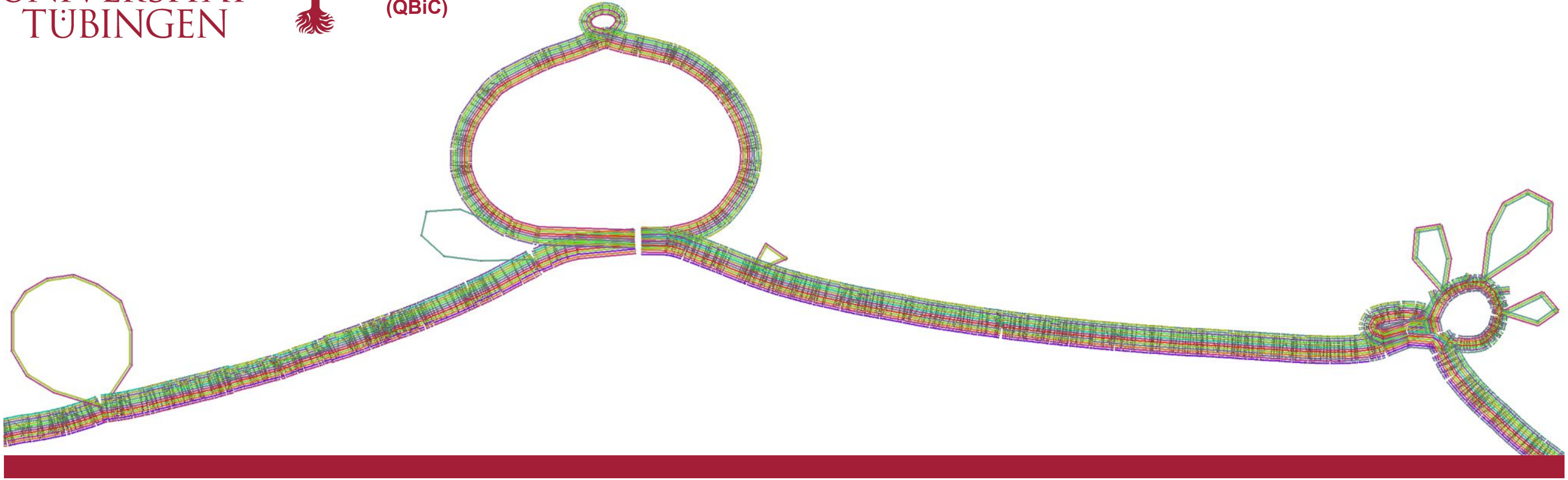
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Quantitative Biology Center
(QBiC)

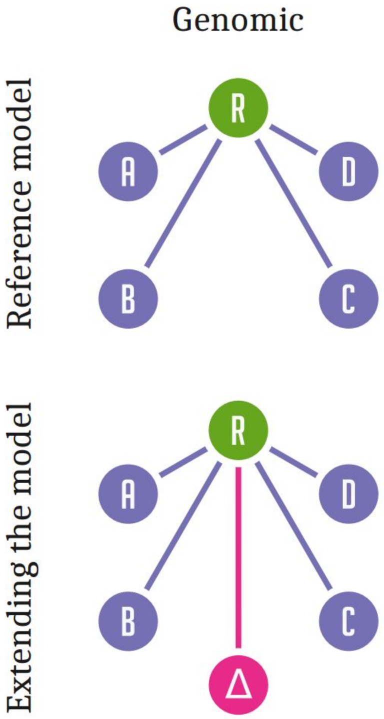


Cluster efficient pangenome graph construction with

nf-core/ pangenome

Simon Heumos - M3 Workshop 22/03/2024

De novo assembly and a pangenomic model

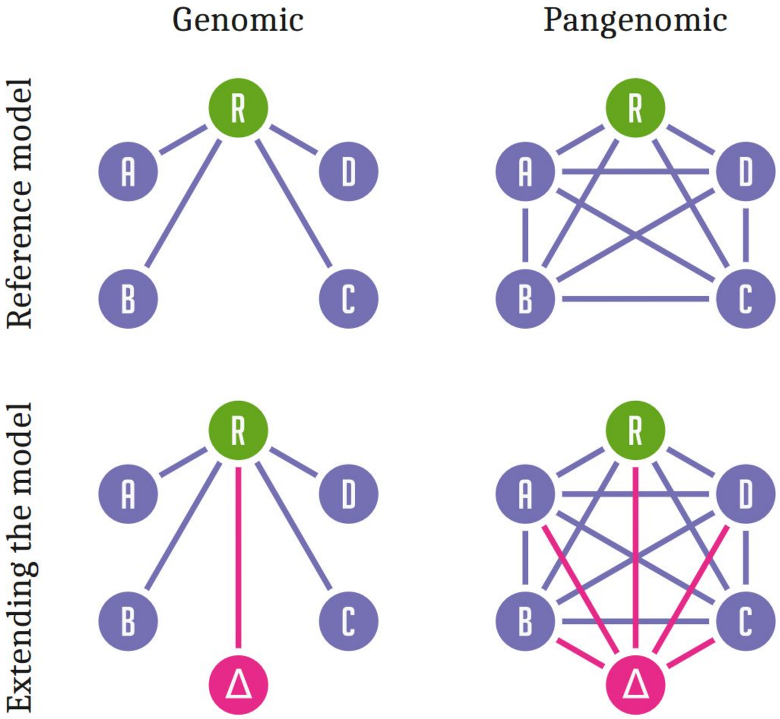


Δ: new genome; R: reference genome.

Figure from Eizenga et al., 2020.

Thanks to advances in sequencing technology, new **telomere-to-telomere** quality genome assemblies are produced at a high rate.

De novo assembly and a pangenomic model



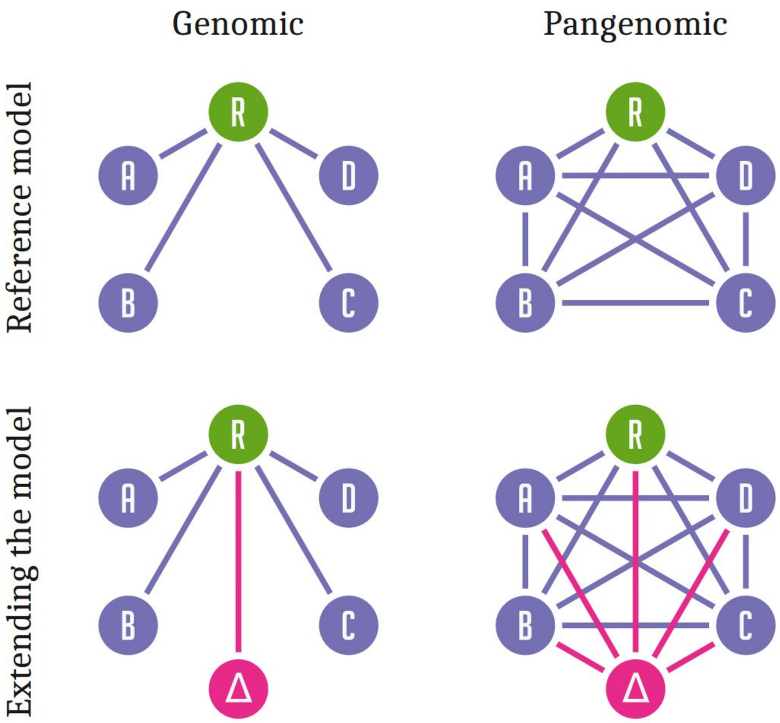
Δ: new genome; R: reference genome.

Figure from Eizenga et al., 2020.

Thanks to advances in sequencing technology, new **telomere-to-telomere** quality genome assemblies are produced at a high rate.

Pangenomes can **model** the full set of genomic elements in a given species or clade, reducing the **reference-bias**.

A pangenome encoded as a graph



Δ: new genome; R: reference genome.

Figure from Eizenga et al., 2020.

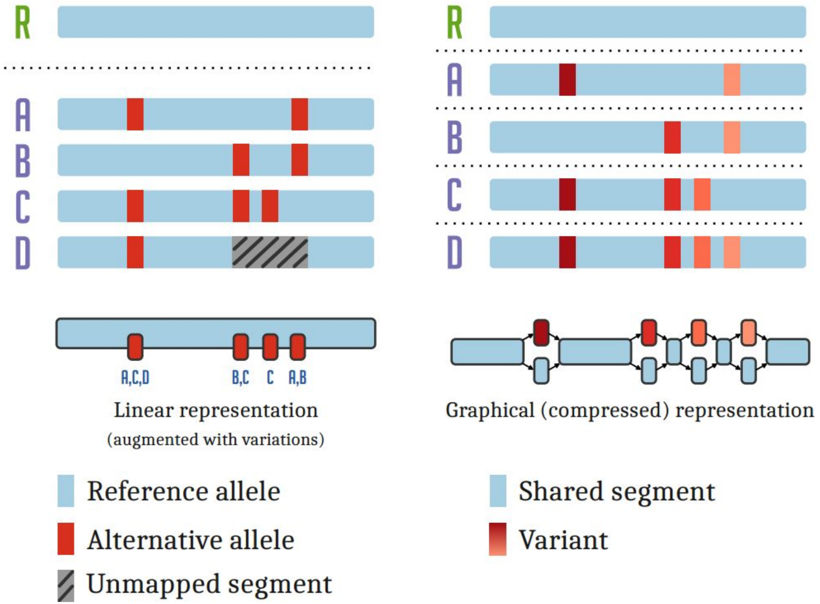


Figure from Eizenga et al., 2020.

Pangenome graphs - representation

Pangenomes can take many forms, including **graph-based** data structures.

Pangenome graphs compress redundant sequences into a smaller data structure that is still representative of the full set.

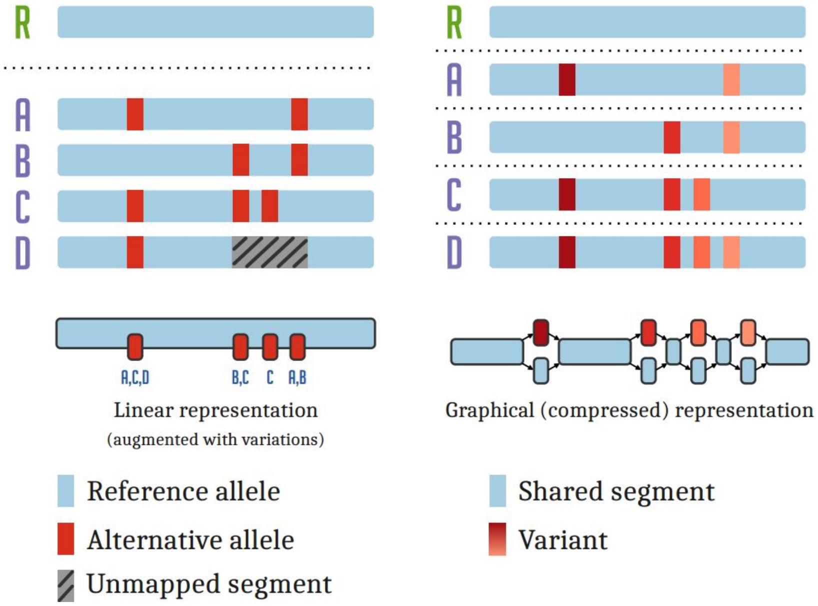


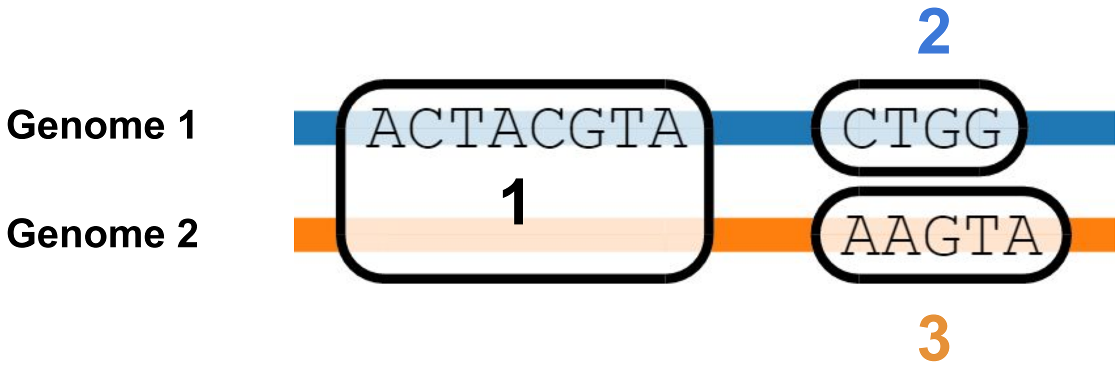
Figure from Eizenga et al., 2020.

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.

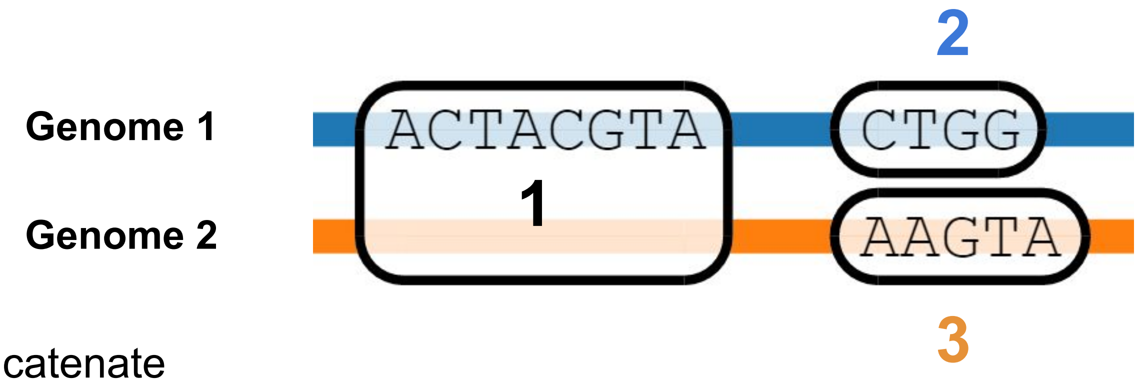
Sketch made using SequenceTubeMap.

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

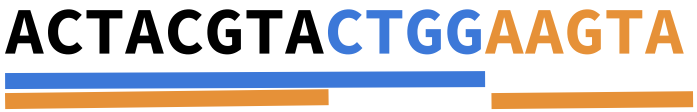
Towards a 1D visualization

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

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



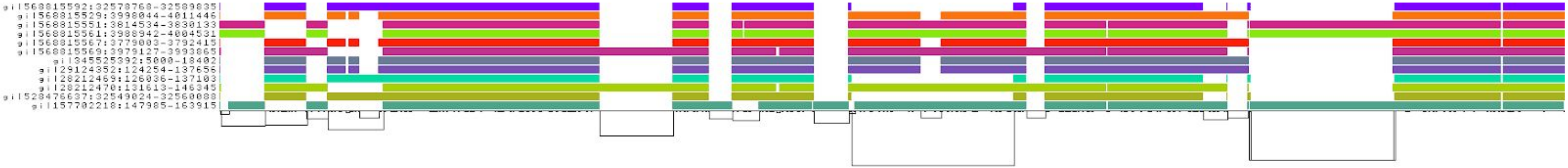
Concatenate nucleotides to a pangenome sequence.


ACTACGTACTGGAAGTA

Presence - absence matrix encodes actual genomic sequence.

1D Graph visualization explained

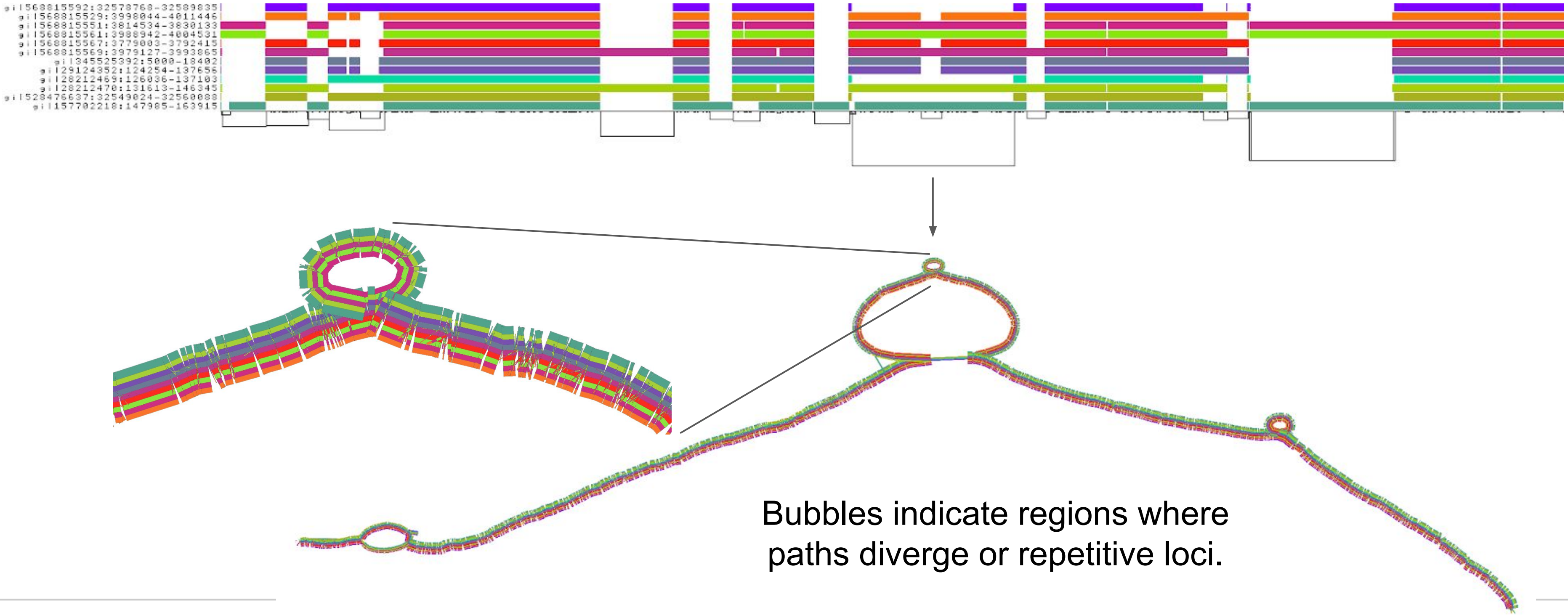
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.


gi|568815592:32578768-32589835
gi|568815529:3998044-4011446
gi|568815551:3814534-3830133
gi|568815561:3988942-4004531
gi|568815567:3779003-3792415
gi|568815569:3979127-3993865
gi|345525392:5000-18402
gi|29124352:124254-137656
gi|28212469:126036-137103
gi|28212470:131613-146345
gi|528476637:32549024-32560088
gi|157702218:147985-163915

- Graph nodes are arranged from left to right forming the pangenome sequence
- Colored bars are the paths versus the pangenome sequences in a binary matrix
- Path names are left
- The black lines under the paths are the links representing the graph topology

2D Graph visualization explained

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


gi|568815592:32578768-32589835
gi|568815529:3998044-4011446
gi|568815551:3814534-3830133
gi|568815561:3988942-4004531
gi|568815567:3779003-3792415
gi|568815569:3979127-3993865
gi|345525392:5000-18402
gi|29124352:124254-137656
gi|28212469:126036-137103
gi|28212470:131613-146345
gi|528476637:32549024-32560088
gi|157702218:147985-163915

Bubbles indicate regions where paths diverge or repetitive loci.

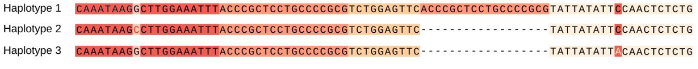
Pangenome research timeline

2000-2010s: Counting genes from MSAs

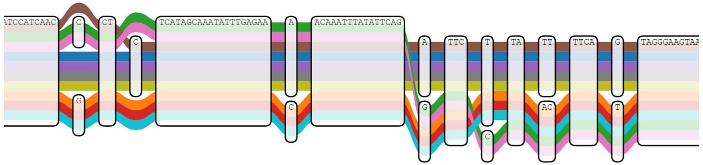
~2015: Sequence level genome graphs

2020s: HPRC pangenomes graphs

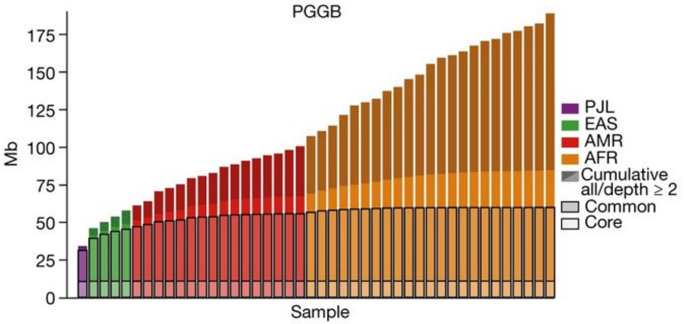
https://doi.org/10.1093/bib/bbw089


Haplotype 1 CAAATAAGGCTTGGAAATTTACCCGCTCCTGCCCCGCGTCTGGAGTTCACCCGCTCCTGCCCCGCGTATTATATTCCAACTCTCTG
Haplotype 2 CAAATAAGCCTTGGAAATTTACCCGCTCCTGCCCCGCGTCTGGAGTTC------------------TATTATATTCCAACTCTCTG
Haplotype 3 CAAATAAGGCTTGGAAATTTACCCGCTCCTGCCCCGCGTCTGGAGTTC------------------TATTATATTACAACTCTCTG

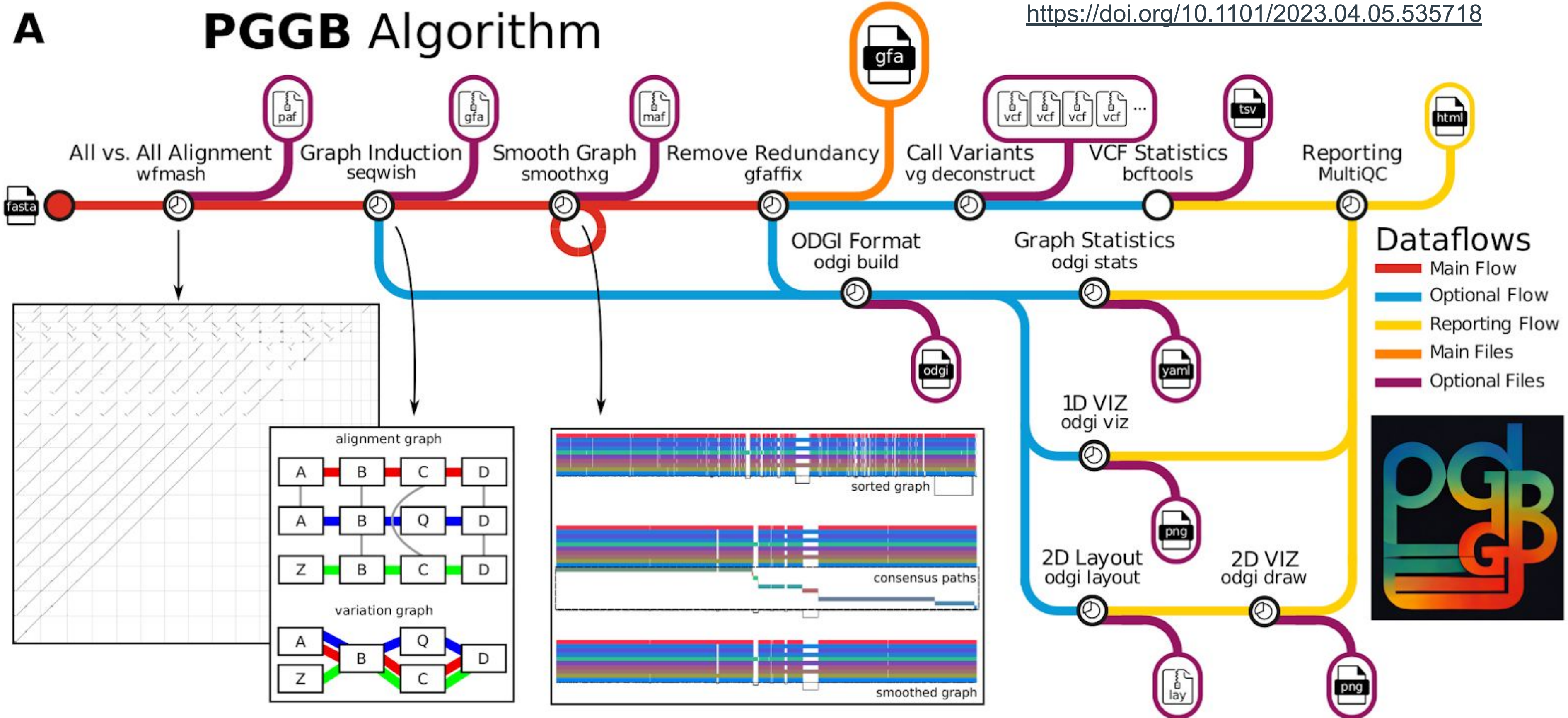
Multiple sequence alignment



https://doi.org/10.1038/s41586-023-05896-x



https://doi.org/10.1101/2023.04.05.535718



All vs. All = quadratic!

Alignment dot plot on the left: E*xascale* matrix of chr6 in all great apes.

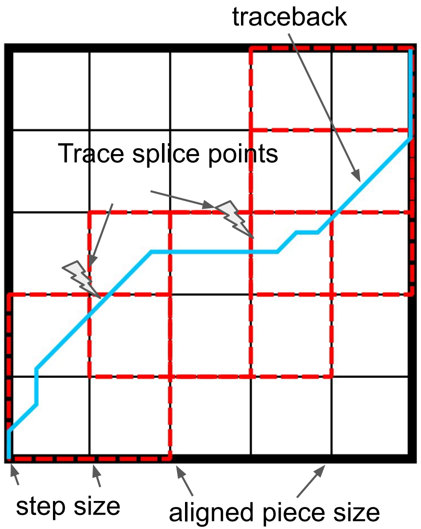
Erik Garrison

Andrea Guarracino

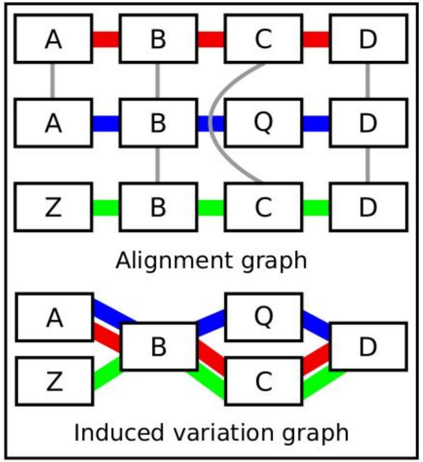
PanGenome Graph Builder - PGGB

PGGB solves the whole genome alignment problem in 3 steps.

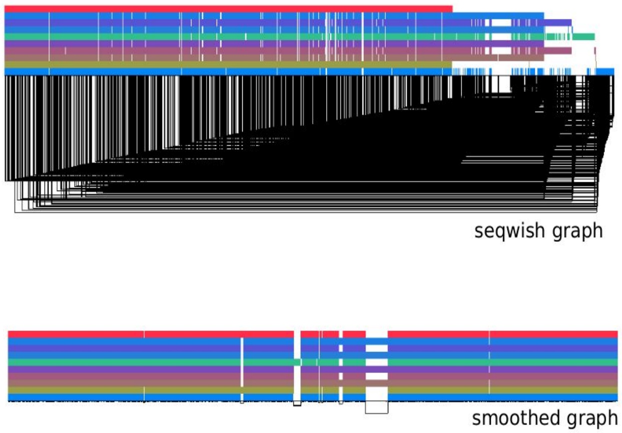
1) all-to-all alignment with **WFMASH**



2) graph induction with **SEQWISH**



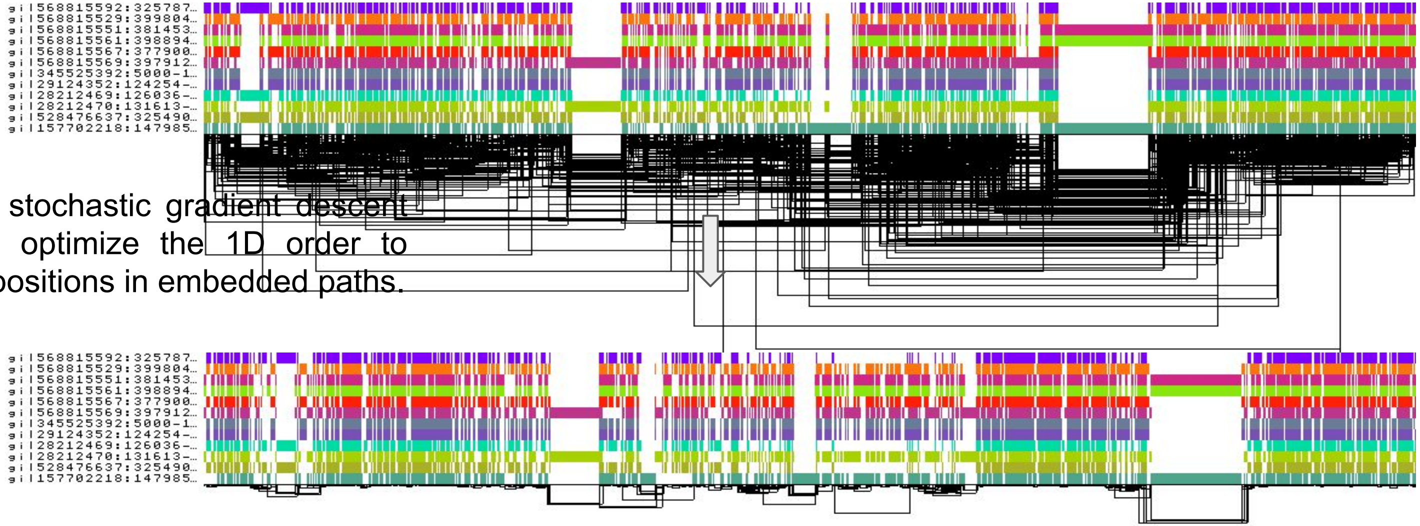
3) normalization with **SMOOTHXG**



https://github.com/pangenome/pggb https://doi.org/10.1101/2023.04.05.535718

Graph normalization

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



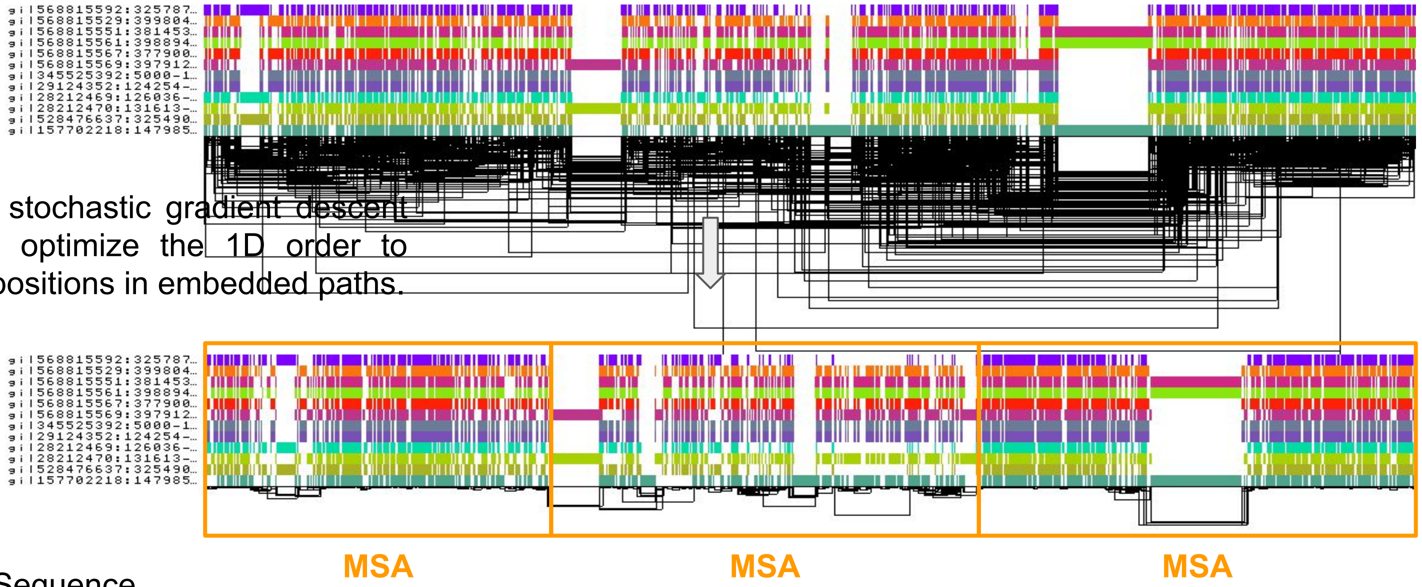
Path-guided stochastic gradient descent algorithm to optimize the 1D order to best-match positions in embedded paths.

https://github.com/pangenome/smoothxg

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

Graph normalization

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



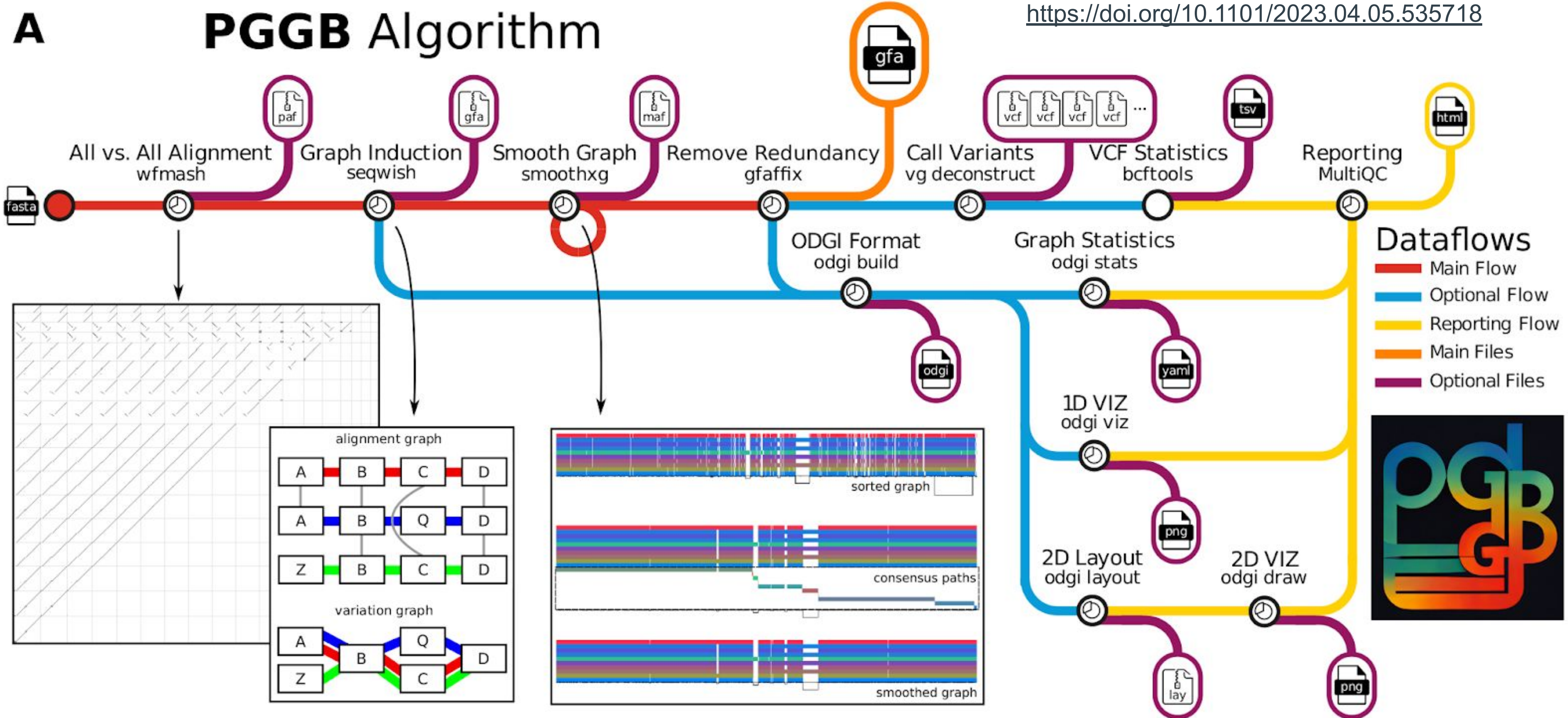
Path-guided stochastic gradient descent algorithm to optimize the 1D order to best-match positions in embedded paths.

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

https://doi.org/10.1101/2023.04.05.535718



All vs. All = quadratic!

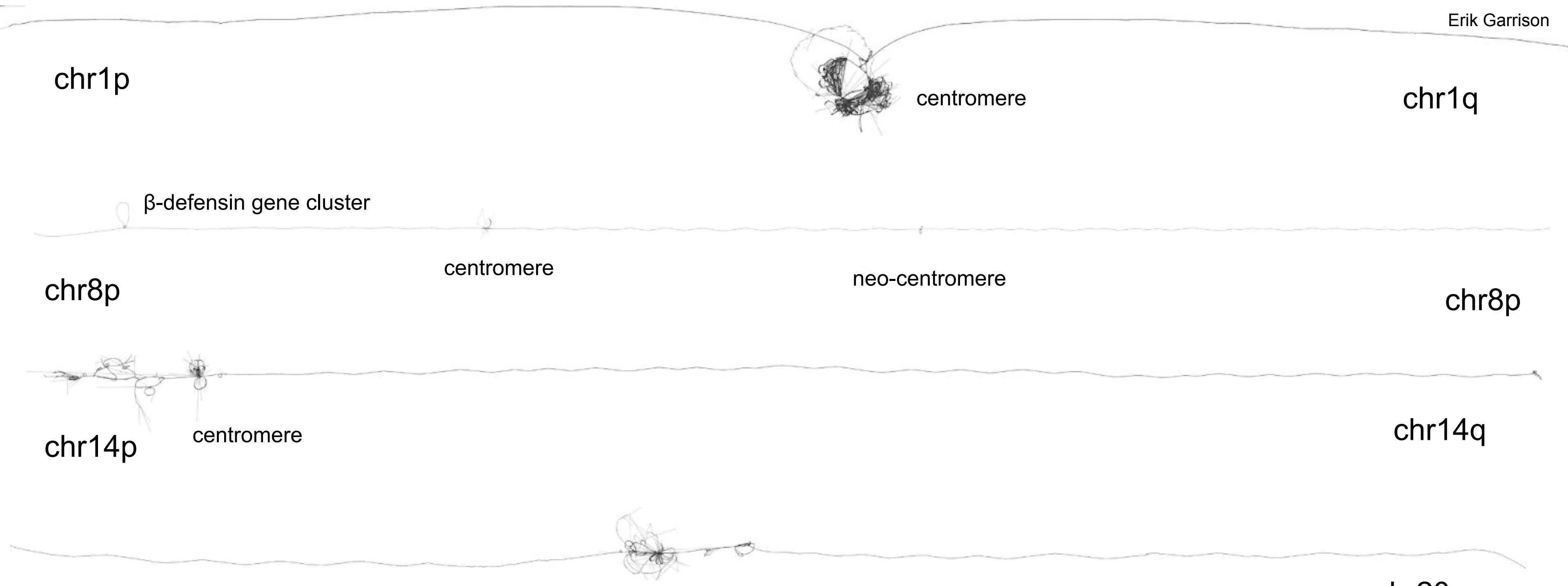
Alignment dot plot on the left: E*xascale* matrix of chr6 in all great apes.

Erik Garrison

Andrea Guarracino

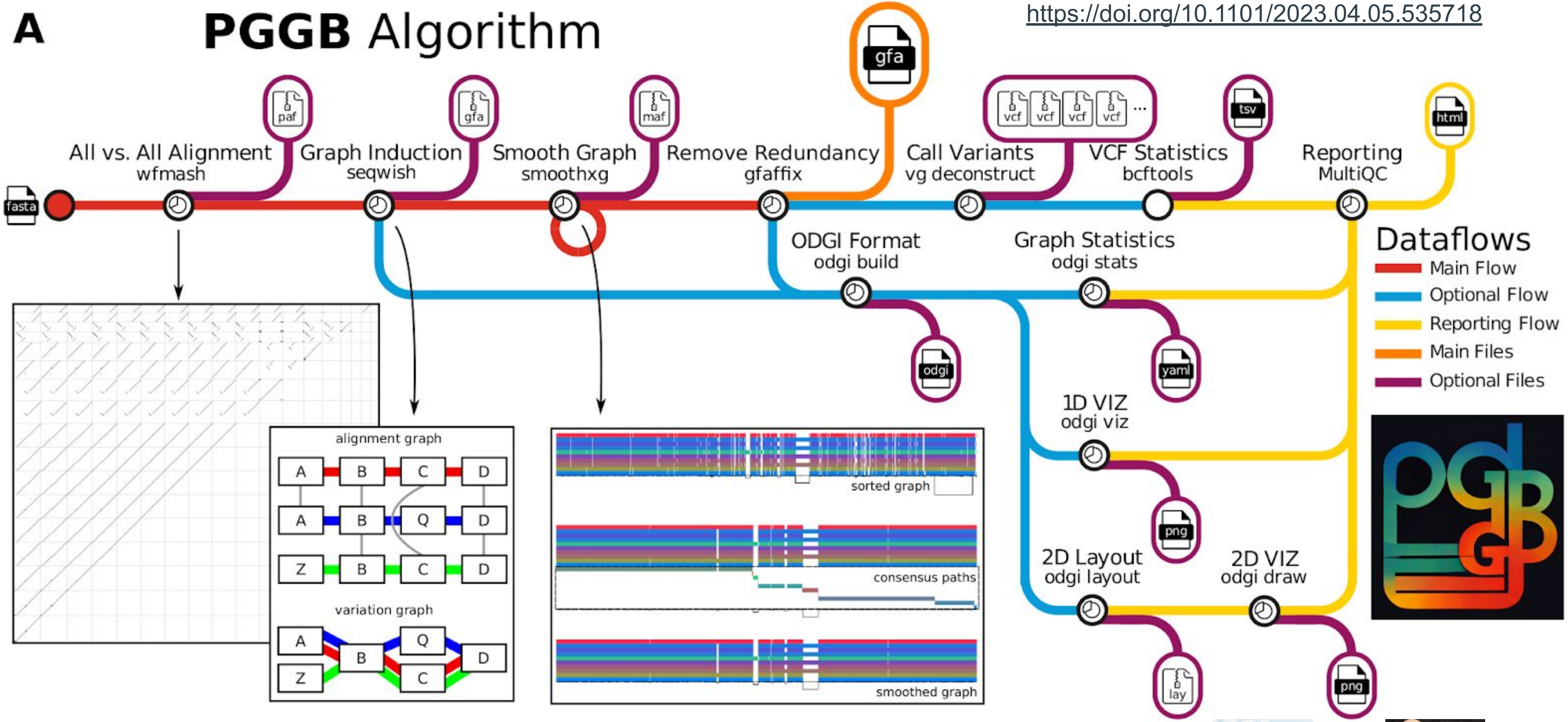
Some Human Pangenome Reference Consortium graphs

Erik Garrison



https://doi.org/10.1101/2023.09.22.558964

https://doi.org/10.1101/2023.04.05.535718



All vs. All = quadratic!

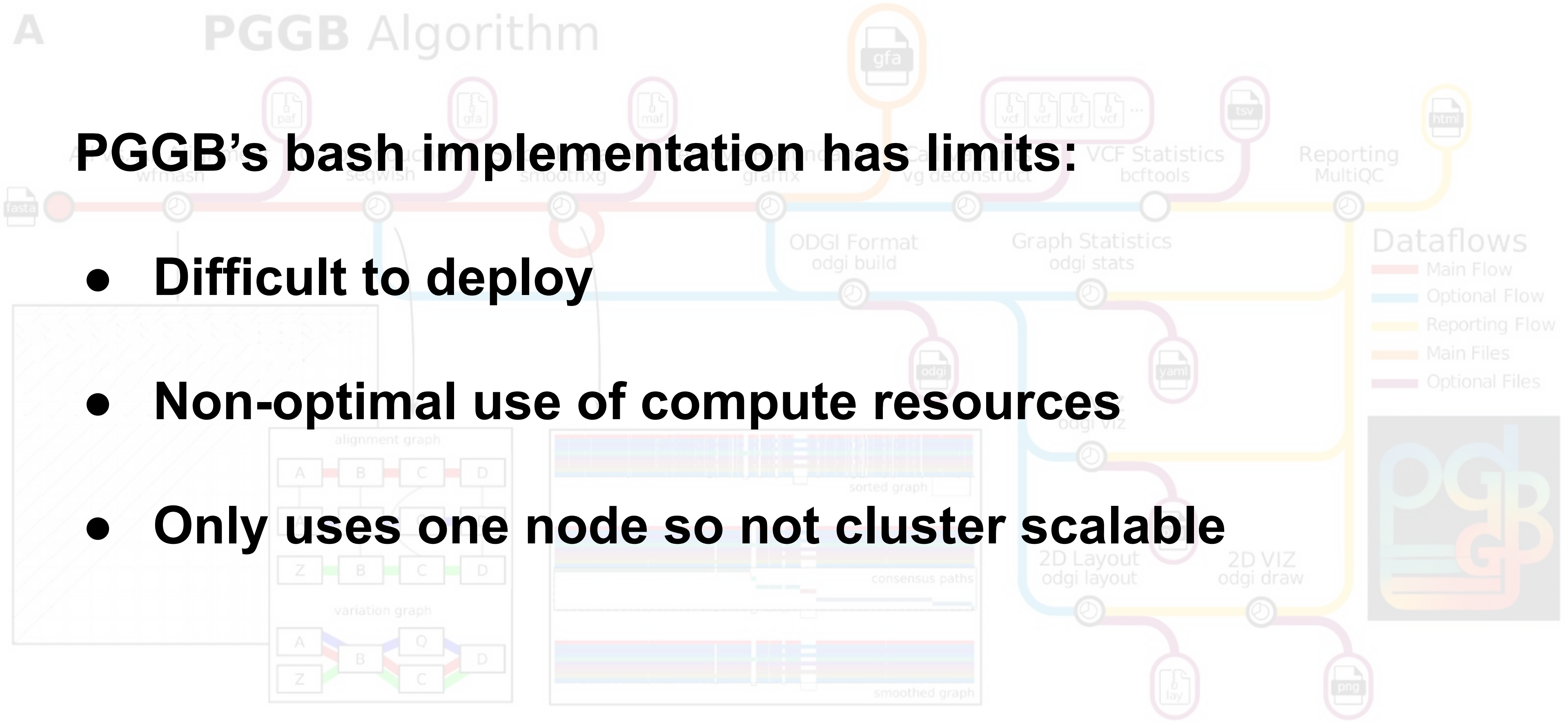
Alignment dot plot on the left: E*xascale* matrix of chr6 in all great apes.

Erik Garrison

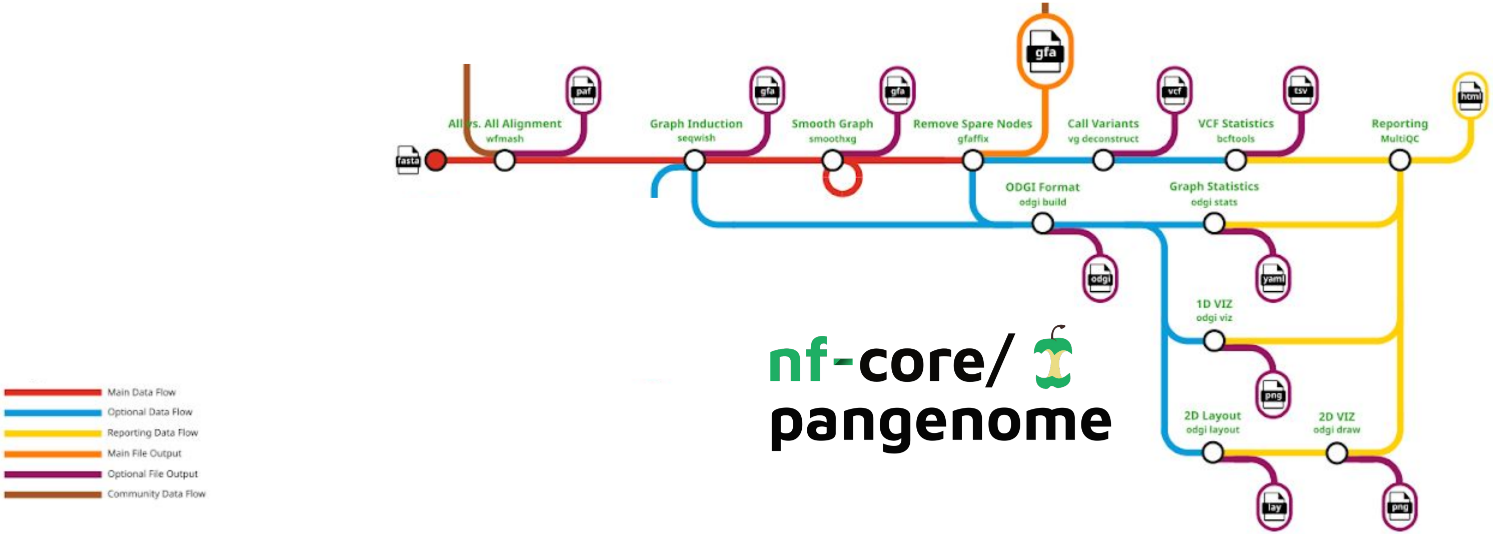
Andrea Guarracino

PGGB’s bash implementation has limits:

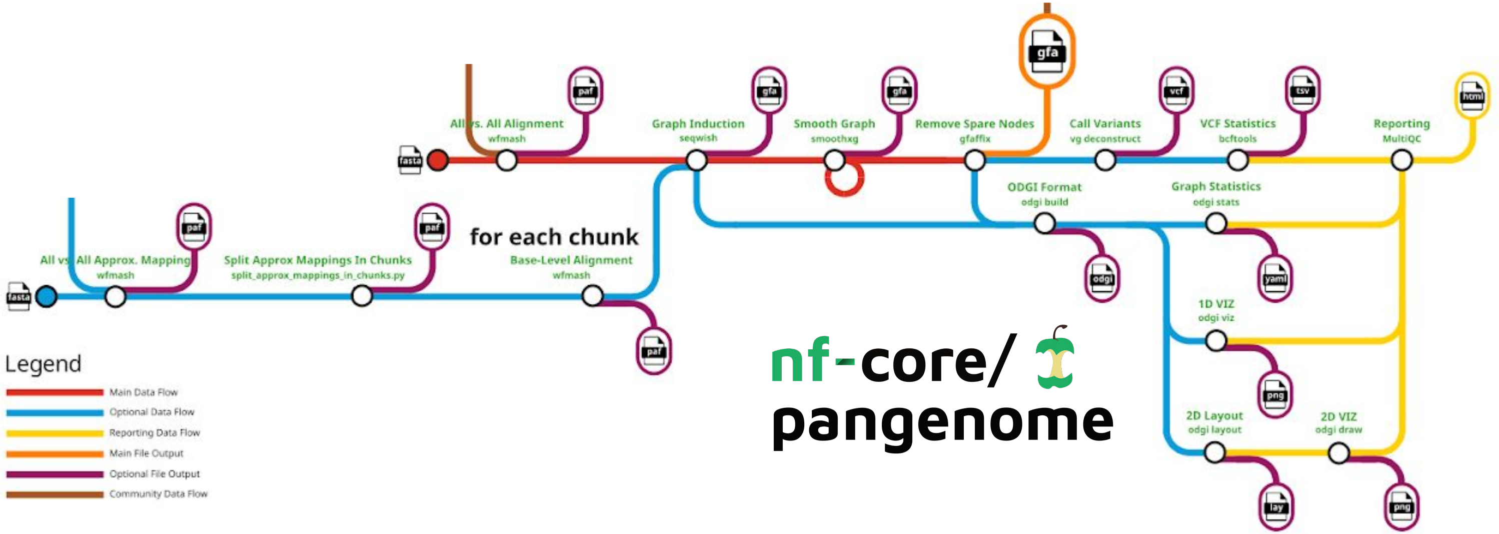
- Difficult to deploy
- Non-optimal use of compute resources
- Only uses one node so not cluster scalable



nf-core/pangenome

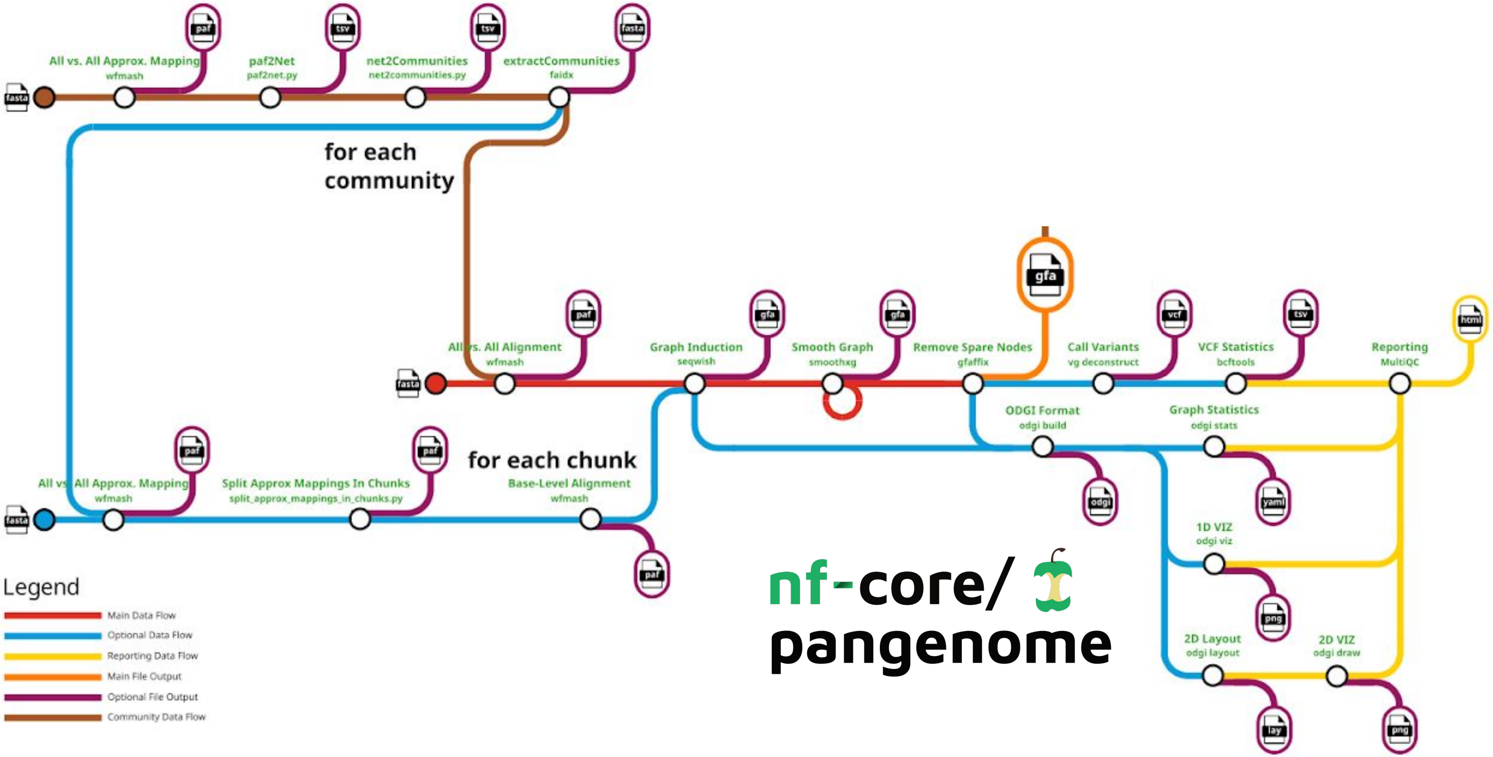


Core workflow taken over from PGGB: Garrison, Guarracino et al., 2023.



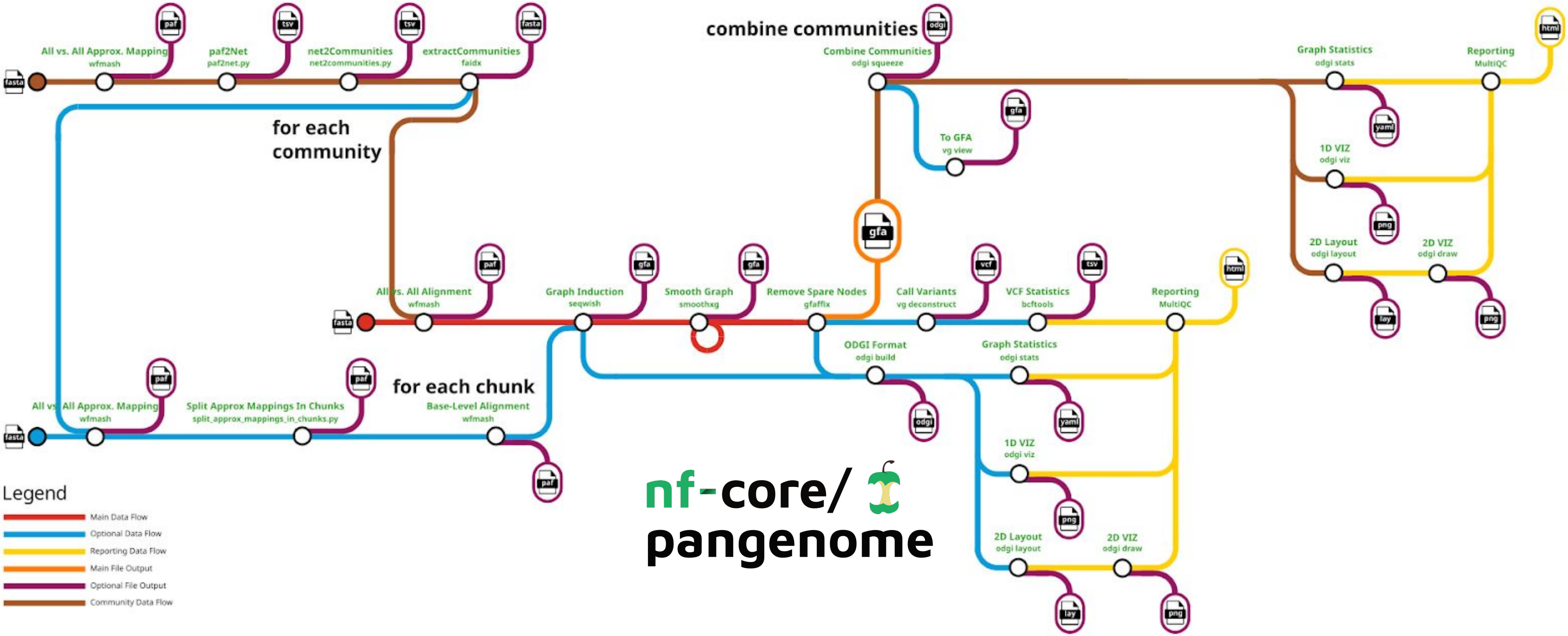
nf-core/

pangenome



Clustering with the Leiden algorithm: Edge weight is `mapped_length * mapped_identity`

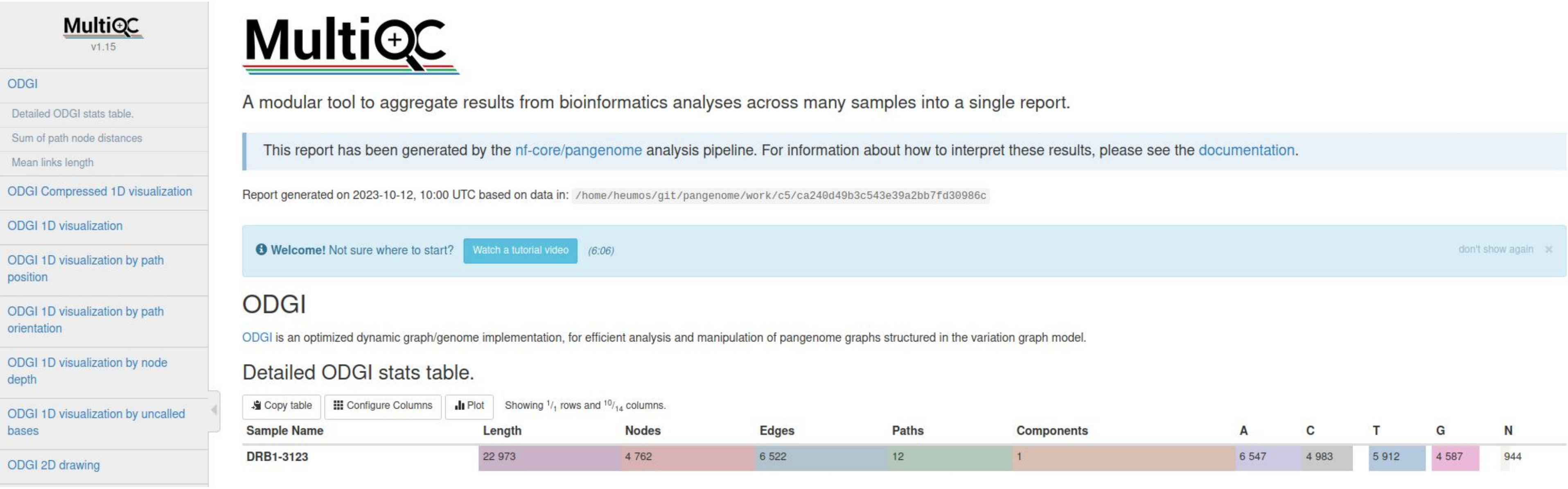
Leiden was applied in Guarracino et al., 2023.



Clustering with the Leiden algorithm: Edge weight is `mapped_length * mapped_identity`

Leiden was applied in Guarracino et al., 2023.

MultiQC Report


MultiQC
v1.15
ODGI
Detailed ODGI stats table.
Sum of path node distances
Mean links length
ODGI Compressed 1D visualization
ODGI 1D visualization
ODGI 1D visualization by path position
ODGI 1D visualization by path orientation
ODGI 1D visualization by node depth
ODGI 1D visualization by uncalled bases
ODGI 2D drawing
MultiQC
Watch a tutorial video
don't show again
Copy table
Configure Columns
Plot

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

This report has been generated by the nf-core/pangenome analysis pipeline. For information about how to interpret these results, please see the documentation.

Report generated on 2023-10-12, 10:00 UTC based on data in: `/home/heumos/git/pangenome/work/c5/ca240d49b3c543e39a2bb7fd30986c`

Welcome! Not sure where to start? *(6:06)*

ODGI

ODGI is an optimized dynamic graph/genome implementation, for efficient analysis and manipulation of pangenome graphs structured in the variation graph model.

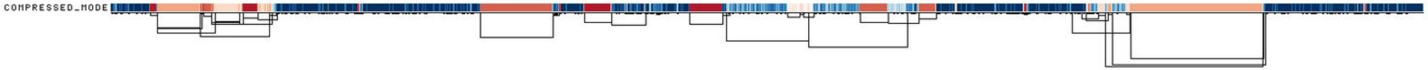
Detailed ODGI stats table.

Showing $^1/_1$ rows and $^{10}/_{14}$ columns.

Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N
DRB1-3123	22 973	4 762	6 522	12	1	6 547	4 983	5 912	4 587	944

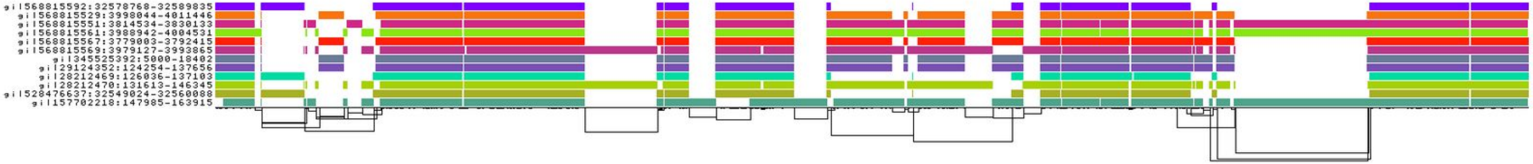
ODGI Compressed 1D visualization

This image shows a 1D rendering of the built pangenome graph. The graph nodes are arranged from left to right, forming the pangenome sequence. Summarization of path coverage across all paths. A heatmap color-coding from https://colorbrewer2.org/#type=diverging&scheme=RdBu&n=11 is used. Dark blue means highest coverage. Dark red means lowest coverage. The path names are placed on the left. The black lines under the paths are the links, which represent the graph topology.



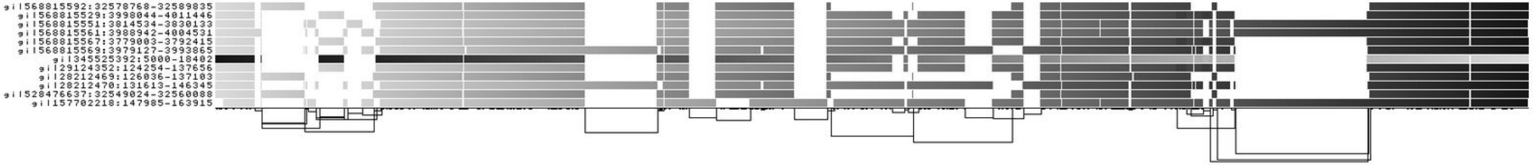
ODGI 1D visualization

This image shows a 1D rendering of the built pangenome graph. The graph nodes are arranged from left to right, forming the pangenome sequence. The colored bars represent the paths versus the pangenome sequence in a binary matrix. The path names are placed on the left. The black lines under the paths are the links, which represent the graph topology.



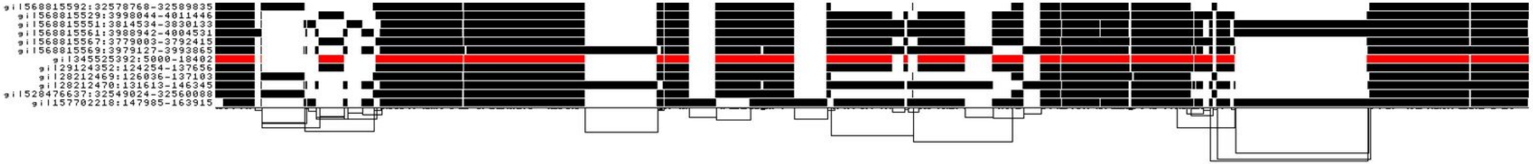
ODGI 1D visualization by path position

This shows a 1D rendering of the built pangenome graph where the paths are colored according to their nucleotide position. Light grey means a low path position, black is the highest path position.



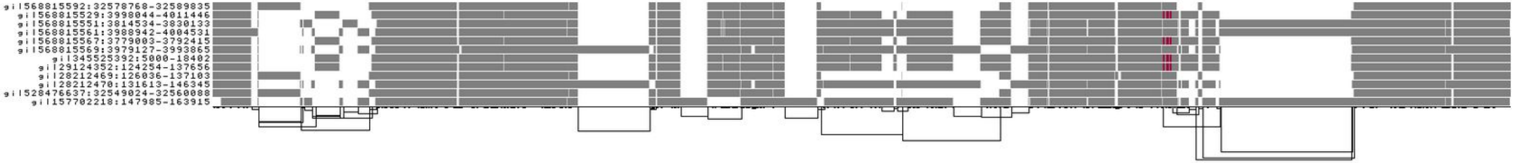
ODGI 1D visualization by path orientation

This image shows a 1D rendering of the built pangenome graph where the paths are colored by orientation. Forward is black, reverse is red.



ODGI 1D visualization by node depth

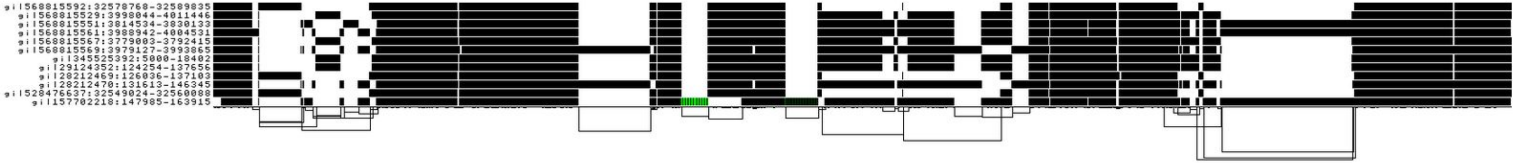
This shows a 1D rendering of the built pangenome graph where the paths are colored according to path depth. Using the Spectra color palette with 4 levels of path depths, white indicates no depth, while grey, red, and yellow indicate depth 1, 2, and greater than or equal to 3, respectively.



gi|568815592:32578768-32589835
gi|568815529:3998044-4011446
gi|568815551:3814534-3830133
gi|568815561:3988942-4004531
gi|568815567:3779003-3792415
gi|568815569:3979127-3993865
gi|345525392:5000-18402
gi|29124352:124254-137656
gi|28212469:126036-137103
gi|28212470:131613-146345
gi|528476637:32549024-32560088
gi|157702218:147985-163915

ODGI 1D visualization by uncalled bases

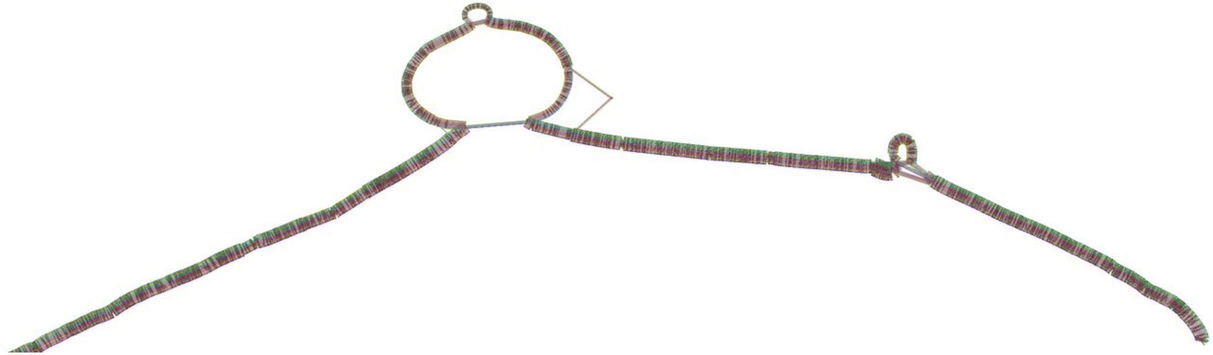
This shows a 1D rendering of the built pangenome graph where the paths are colored according to the coverage of uncalled bases. The lighter the green, the higher the 'N' content of a node is.



gi|568815592:32578768-32589835
gi|568815529:3998044-4011446
gi|568815551:3814534-3830133
gi|568815561:3988942-4004531
gi|568815567:3779003-3792415
gi|568815569:3979127-3993865
gi|345525392:5000-18402
gi|29124352:124254-137656
gi|28212469:126036-137103
gi|28212470:131613-146345
gi|528476637:32549024-32560088
gi|157702218:147985-163915

ODGI 2D drawing

This image shows a 2D rendering of the built pangenome graph.

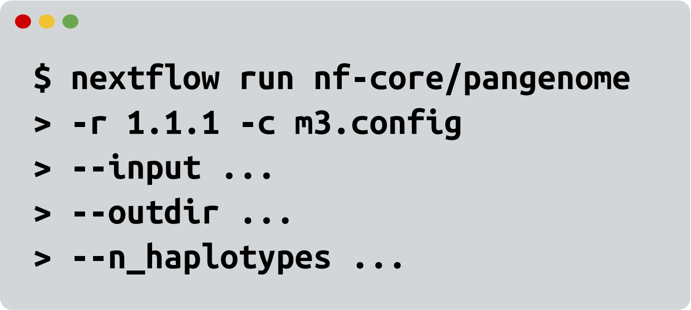


Run nf-core/pangenome

1. Put all input sequences in one FASTA
2. Bonus: Sequence names respect PanSN-spec
3. `bgzip` FASTA
4. `samtools faidx` FASTA.gz
5. Select parameters
6. (Advanced use:) Fine tune parameters
7. Launch pipeline

nf-core/pangenome

Renders a collection of sequences into a pangenome graph.



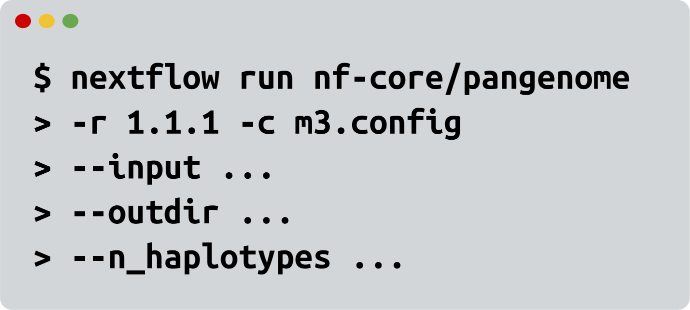
```
$ nextflow run nf-core/pangenome
> -r 1.1.1 -c m3.config
> --input ...
> --outdir ...
> --n_haplotypes ...
```

Run nf-core/pangenome

1. Put all input sequences in one FASTA
2. Bonus: Sequence names respect PanSN-spec
3. `bgzip` FASTA
4. `samtools faidx` FASTA.gz
5. **Select parameters**
6. (Advanced use:) Fine tune parameters
7. Launch pipeline

nf-core/pangenome

Renders a collection of sequences into a pangenome graph.



```
$ nextflow run nf-core/pangenome
> -r 1.1.1 -c m3.config
> --input ...
> --outdir ...
> --n_haplotypes ...
```

nf-core/pangenome key parameters

nf-core/pangenome solves the whole genome alignment problem in 3 steps.

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

`--wfmash_map_pct_id`: Percentage of sequence identity for mapping and alignment. Consult **mash**.

Default: 90.0.

`--wfmash_segment_length`: Segment length for mapping.

Default: 5000.

2) graph induction with **SEQWISH**

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

Default: 23

3) normalization with **SMOOTHXG**

`--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

M3 cluster demo:

Building a Lipoprotein(a) (LPA) pangenome graph from 14 haplotypes

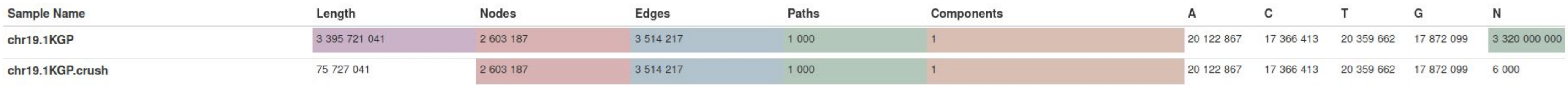
LPA is a risk factor for:

- Atherosclerosis
- Coronary heart disease
- Stroke

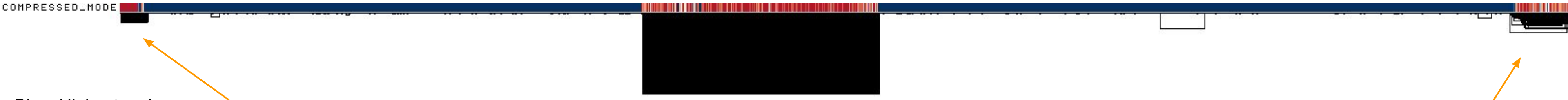
Building a human 1KG chromosome 19 pangenome graph

- 1000 sequences of chr19 of the 1000 genomes project
- Chr19 length: ~59Mb
- Takes ~4 days on our Core Facility Cluster (CFC)

Building a human 1KG chromosome 19 pangenome graph



Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N
chr19.1KGP	3 395 721 041	2 603 187	3 514 217	1 000	1	20 122 867	17 366 413	20 359 662	17 872 099	3 320 000 000
chr19.1KGP.crush	75 727 041	2 603 187	3 514 217	1 000	1	20 122 867	17 366 413	20 359 662	17 872 099	6 000



Blue: Highest node coverage

Red: Lowest node coverage

Complex telomeric region which can't be resolved with short read data

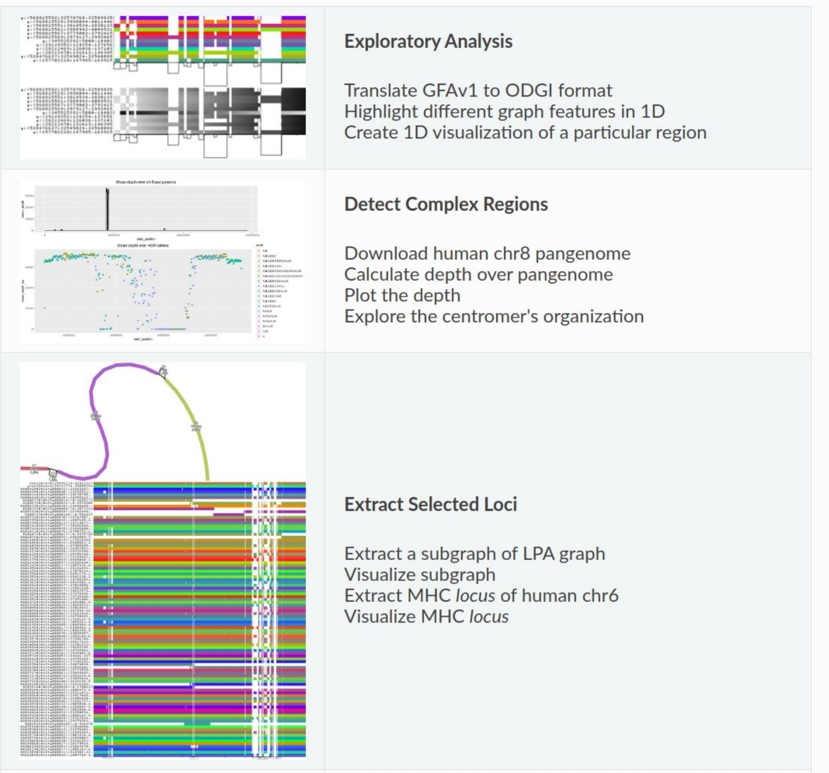
Centromere region which can't be resolved with short read data

Complex telomeric region which can't be resolved with short read data

Building a 2146 sequences *E. coli* pangenome graph

- 2146 sequences from GeneBank including 133 plasmids
- *E. coli* length: ~5Mb
- Our CFC is barely sufficient for this task: ~10 days

Downstream analyses with ODGI



Exploratory Analysis

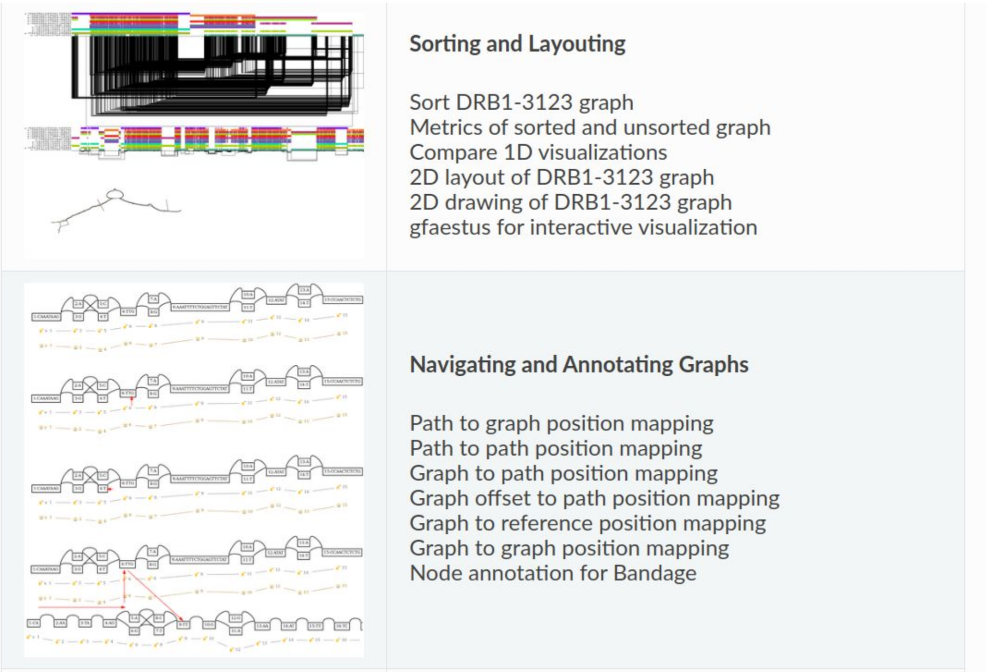
Translate GFAv1 to ODGI format
Highlight different graph features in 1D
Create 1D visualization of a particular region

Detect Complex Regions

Download human chr8 pangenome
Calculate depth over pangenome
Plot the depth
Explore the centromer's organization

Extract Selected Loci

Extract a subgraph of LPA graph
Visualize subgraph
Extract MHC *locus* of human chr6
Visualize MHC *locus*



Sorting and Layouting

Sort DRB1-3123 graph
Metrics of sorted and unsorted graph
Compare 1D visualizations
2D layout of DRB1-3123 graph
2D drawing of DRB1-3123 graph
gfaestus for interactive visualization

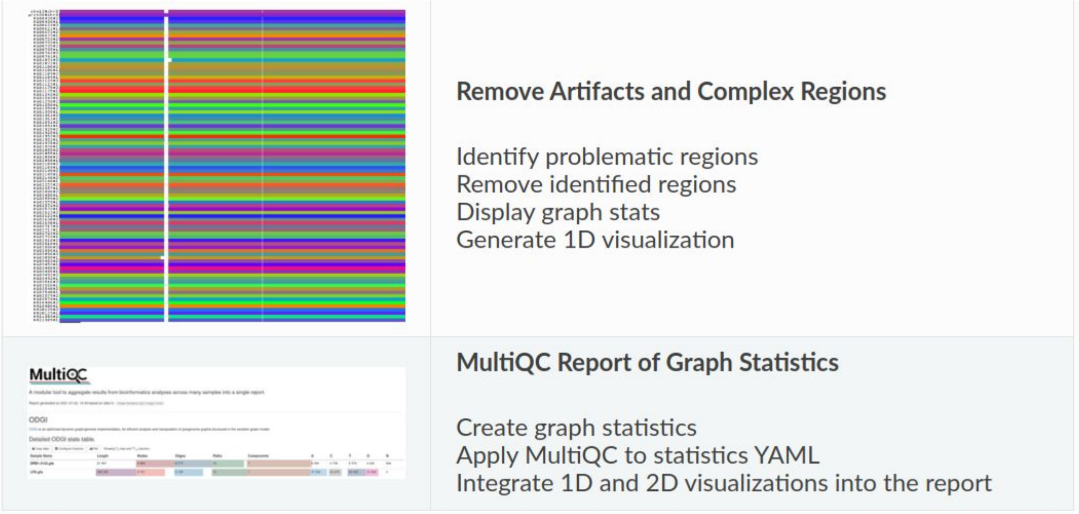
Navigating and Annotating Graphs

Path to graph position mapping
Path to path position mapping
Graph to path position mapping
Graph offset to path position mapping
Graph to reference position mapping
Graph to graph position mapping
Node annotation for Bandage

Guarracino*, Heumos* et al., 2023

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Downstream analyses with ODGI



Remove Artifacts and Complex Regions

Identify problematic regions
Remove identified regions
Display graph stats
Generate 1D visualization

MultiQC

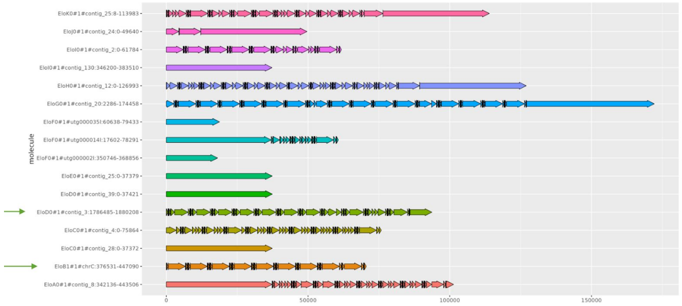
A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

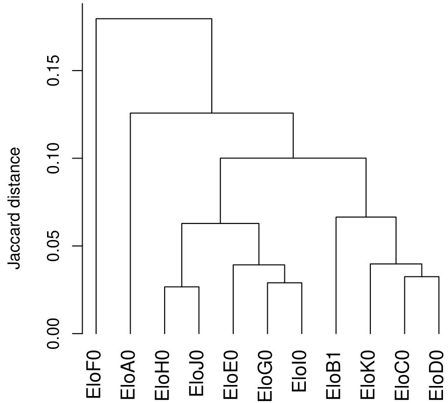
ODGI

Detailed ODGI stats table.

MultiQC Report of Graph Statistics

Create graph statistics
Apply MultiQC to statistics YAML
Integrate 1D and 2D visualizations into the report





Guarracino*, Heumos* et al., 2023

Tutorial materials

- Memphis Pangenome Course 2023:

  https://pangenome.github.io/MemPanG23/

- ODGI Tutorials:

  https://odgi.readthedocs.io/en/latest/rst/tutorials.html

- Future Memphis Pangenome Course 2024:

  https://pangenome.github.io/MemPanG24/

Acknowledgements

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

ZIM
Zentrales
Innovationsprogramm
Mittelstand

Sven Nahnsen

Oliver Kohlbacher

Michael Krone

Gisela Gabernet

Friederike Hanssen

Antonia Schuster

Júlia Mir Pedrol

Susanne Jodoin

Erik Garrison

Andrea Guarracino

Pjotr Prins

Vincenza Colonna

Flavia Villani

David G. Ashbrook

Robert W.Williams

Christian Fischer

Christian Kubica

Sebastian Vorbrugg

Jörg Hagmann

Jerven Bollemann

Toshiyuki T. Yokoyama

Torsten Pook

Franziska Huth

HUMAN PANGENOME

Lukas Heumos

Philipp Ehmele

Daniel Dörr

Tomas Vinar

Brona Brejova

Jozeph Nosek

LodElo Consortium

nf-core