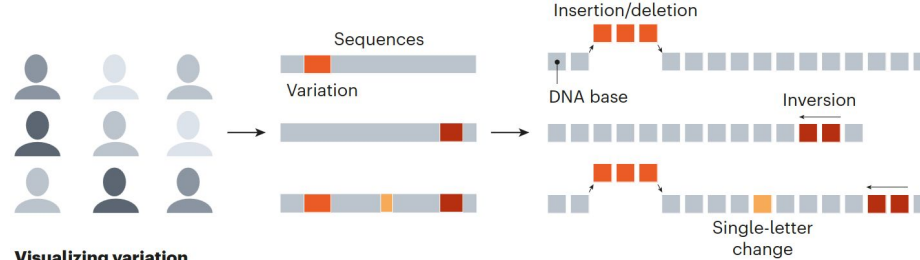


# Cluster scalable pangenome graph construction with

## **nf-core/** **pangenome**

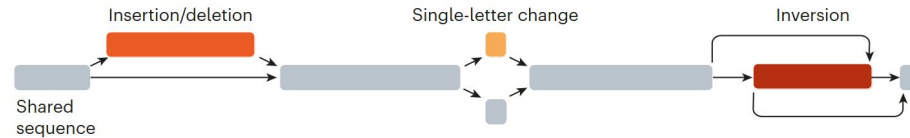


# Every base everywhere all at once - Pangenomes



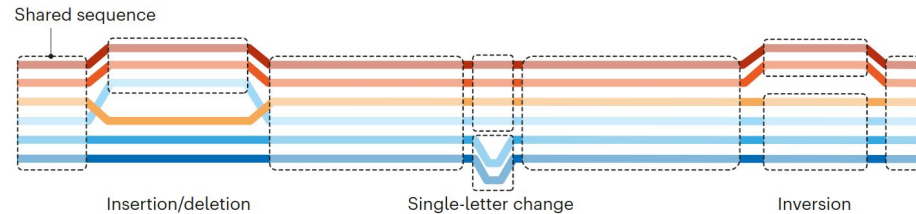
## Visualizing variation

Graphical models can present variation data in a way that doesn't assume a standard, or default reference genome.



## Exploring the pangenome

Representations that look like subway maps allow researchers to compare the variations in a population at a sequence level.



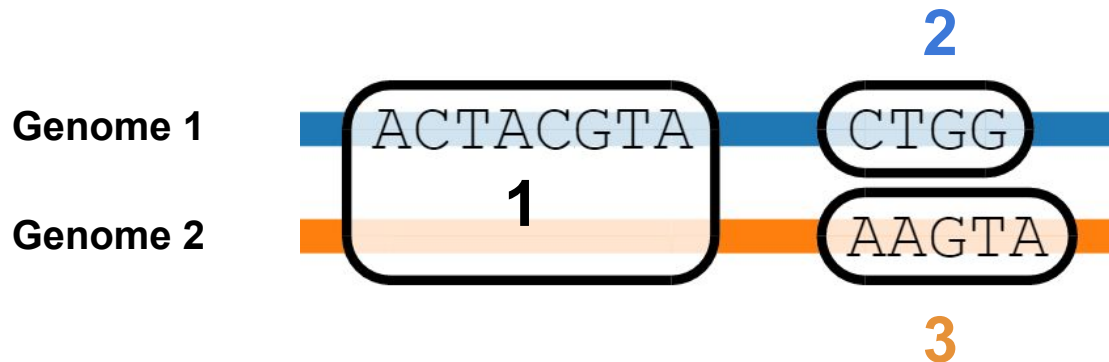


# Variation Graphs

— Genome 1: ACTACGTA**CTGG** Path: 1 **2**

— Genome 2: ACTACGTA**AAGTA** Path: 1 **3**

Linear sequences are **paths** through nodes.



Graph topology is not directly shown.

The nodes represent DNA sequences.

Sketch made using  
[SequenceTubeMap](#).

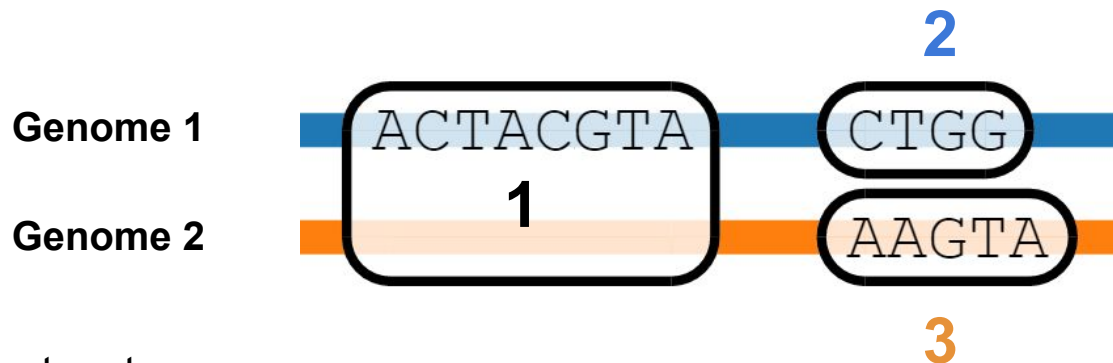
**Paths** can be contigs, haplotypes, reads, or whole chromosomes.



# Towards a 1D visualization

Genome 1: **ACTACGTA**CTGG Path: 1 2

Genome 2: **ACTACGTA**AAGTA Path: 1 3



Concatenate  
nucleotides to a  
pangenome  
sequence.

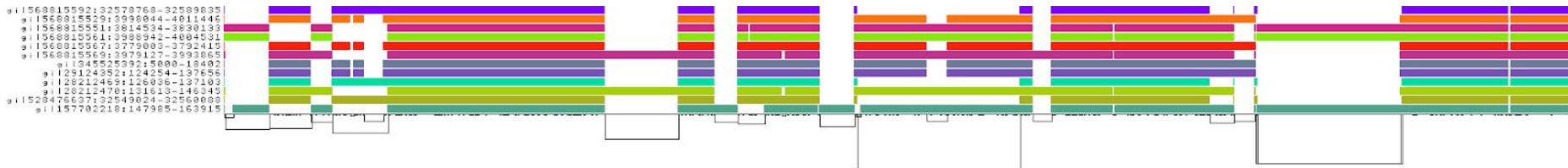
**ACTACGTA**CTGGAAGTA

Presence - absence  
matrix encodes actual  
genomic sequence.



# 1D visualization explained

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

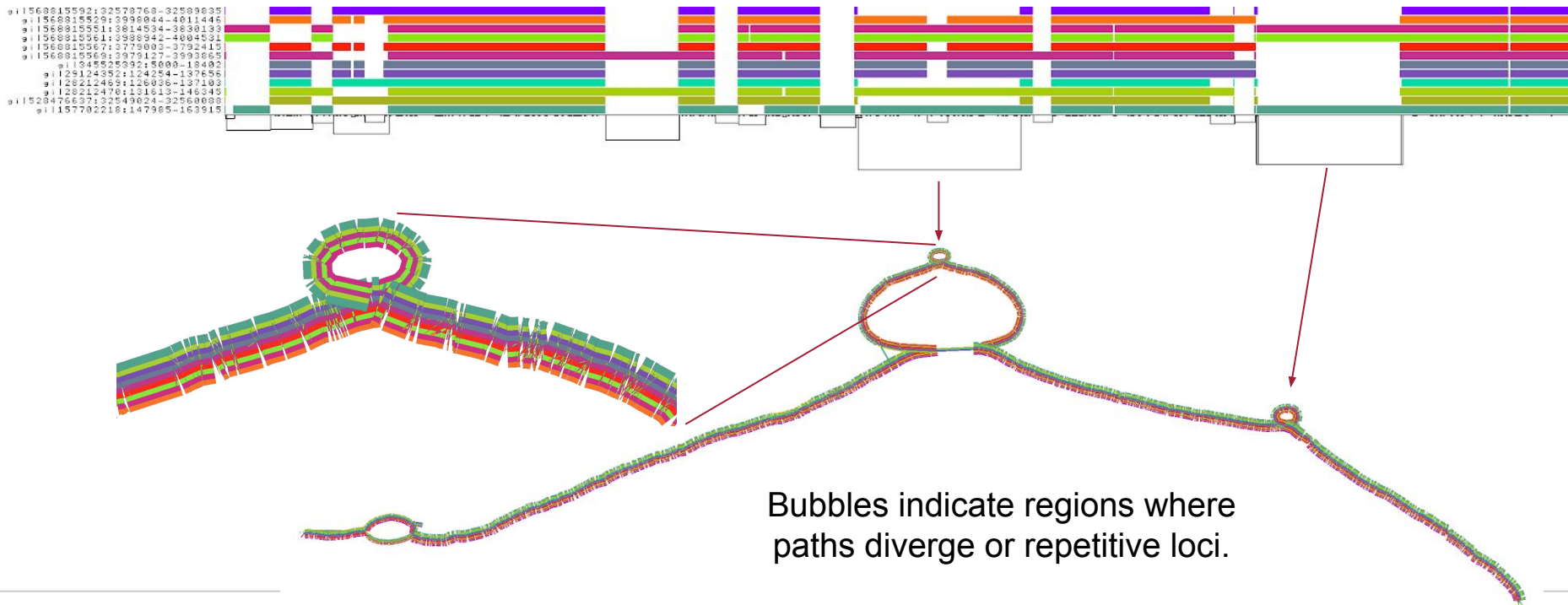


- Graph nodes are arranged from left to right forming the pangenome sequence
- Colored bars are the paths versus the pangenome sequences in a binary matrix
- Path names are left
- The black lines under the paths are the links representing the graph topology



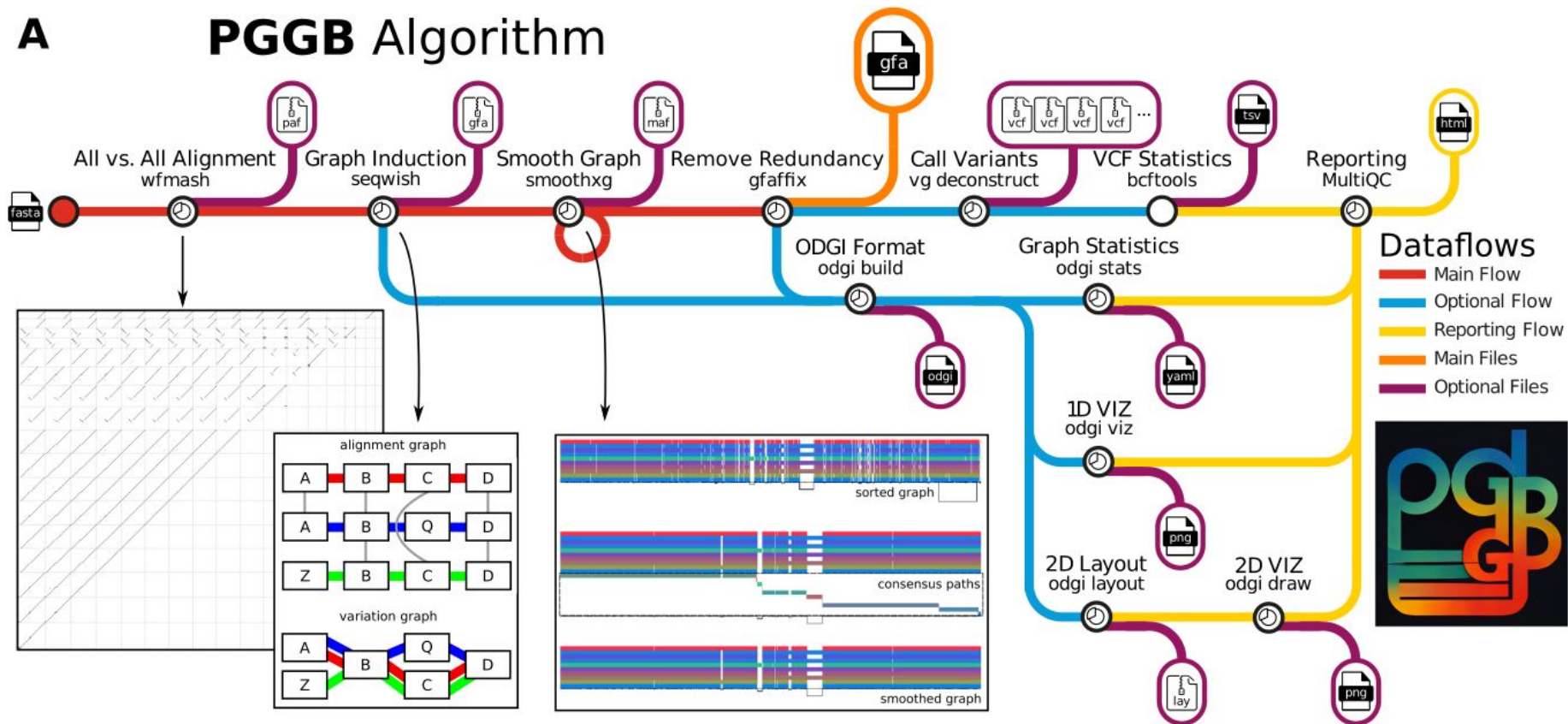
# 2D visualization explained

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





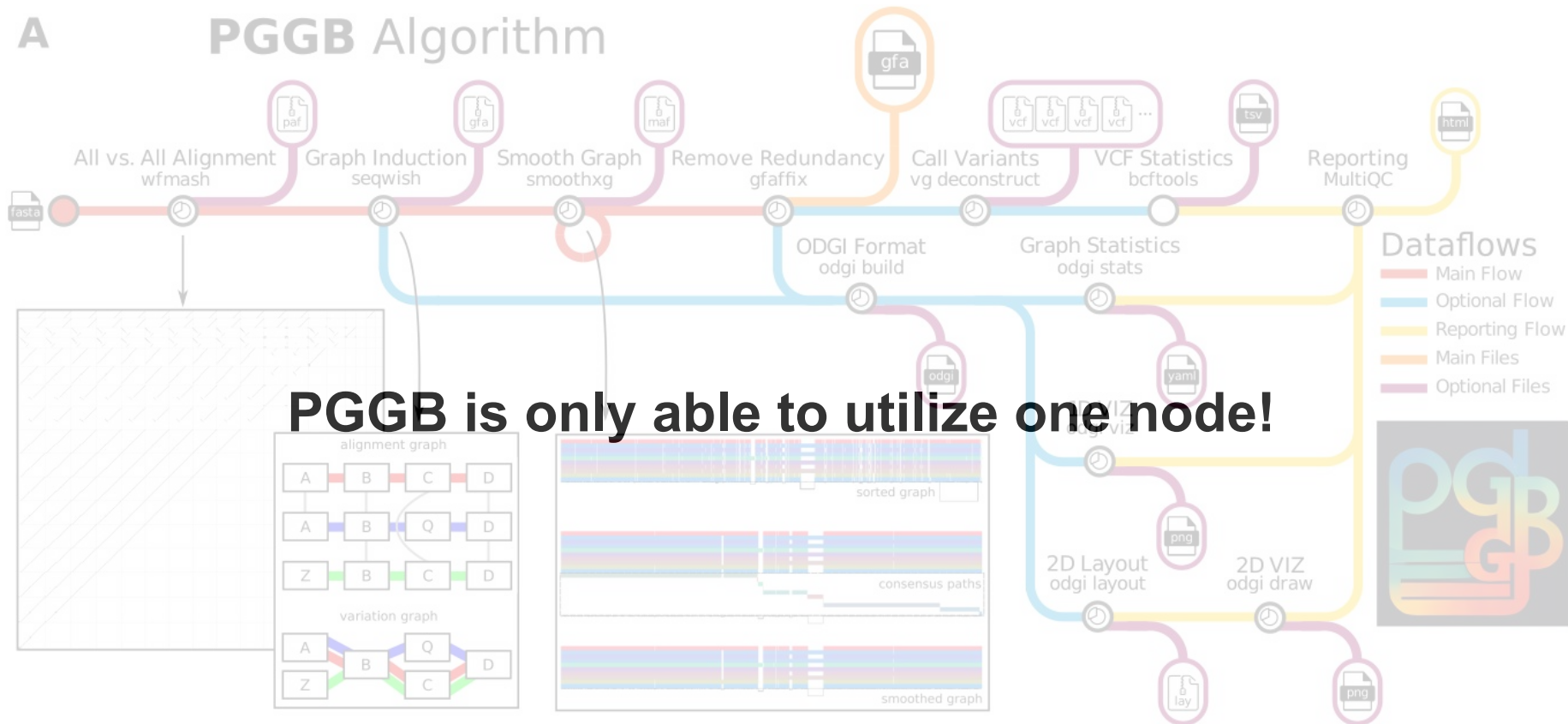
# A PGGB Algorithm



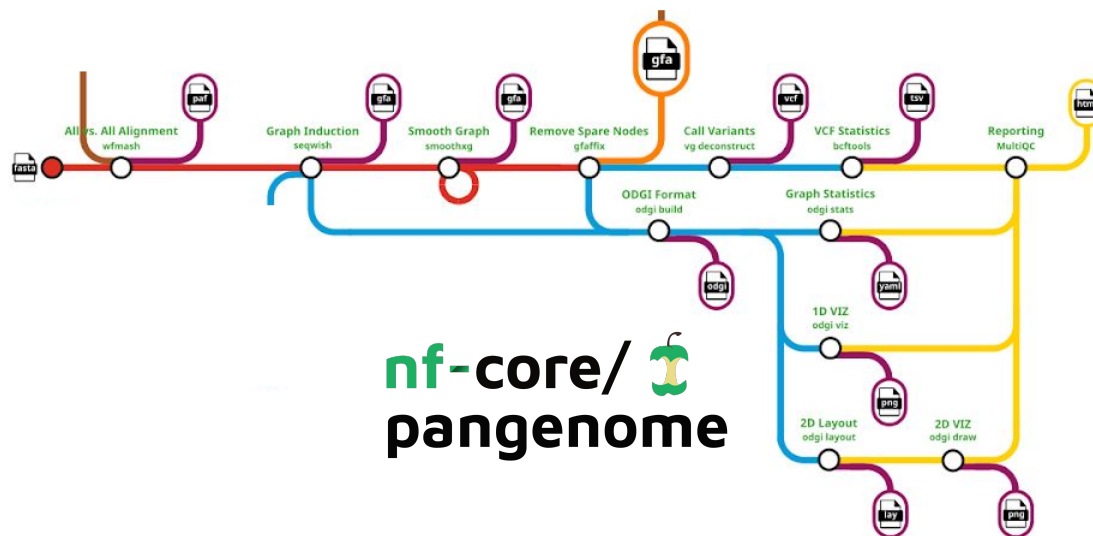


A

# PGGB Algorithm







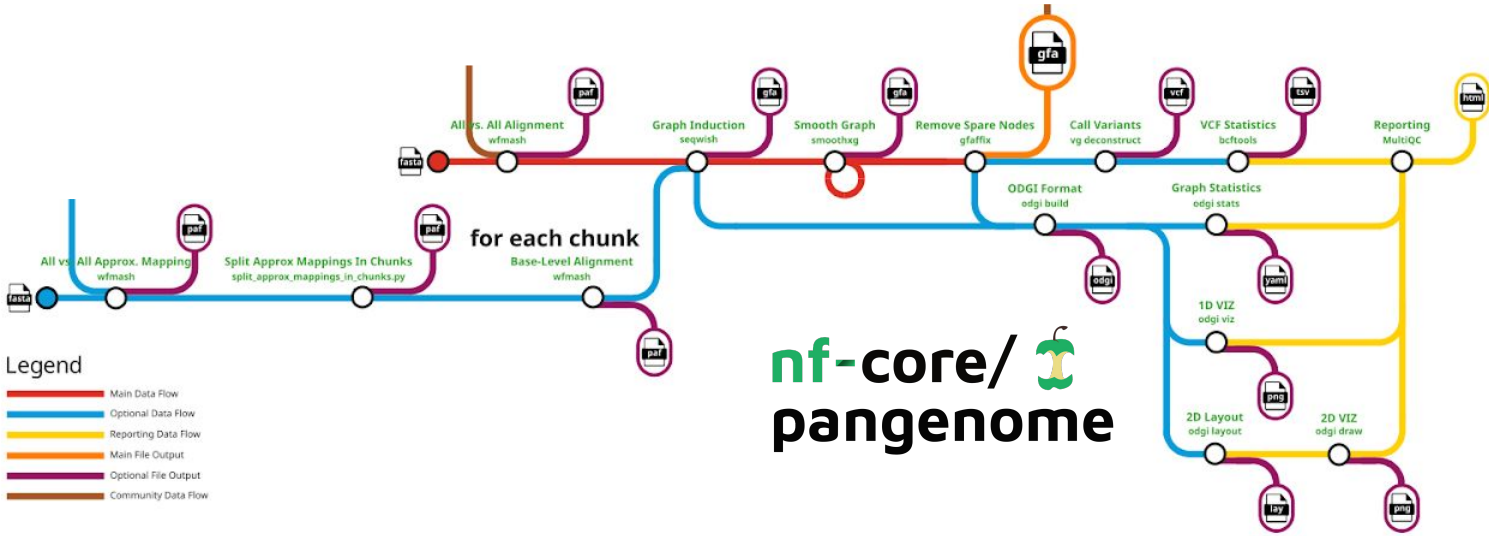
Core workflow taken over from PGGB: [Garrison, Guarracino et al., 2023](#).

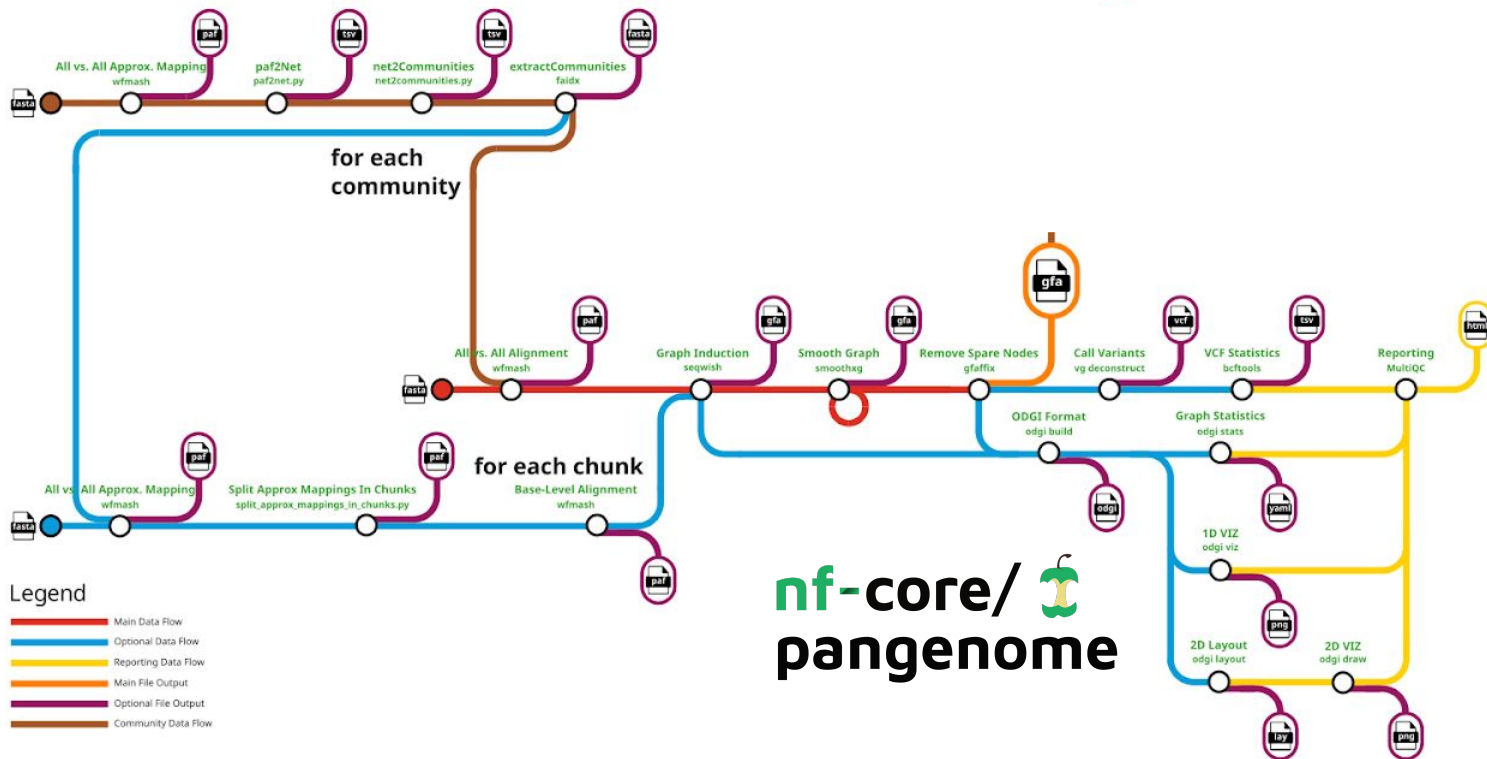


Workflow

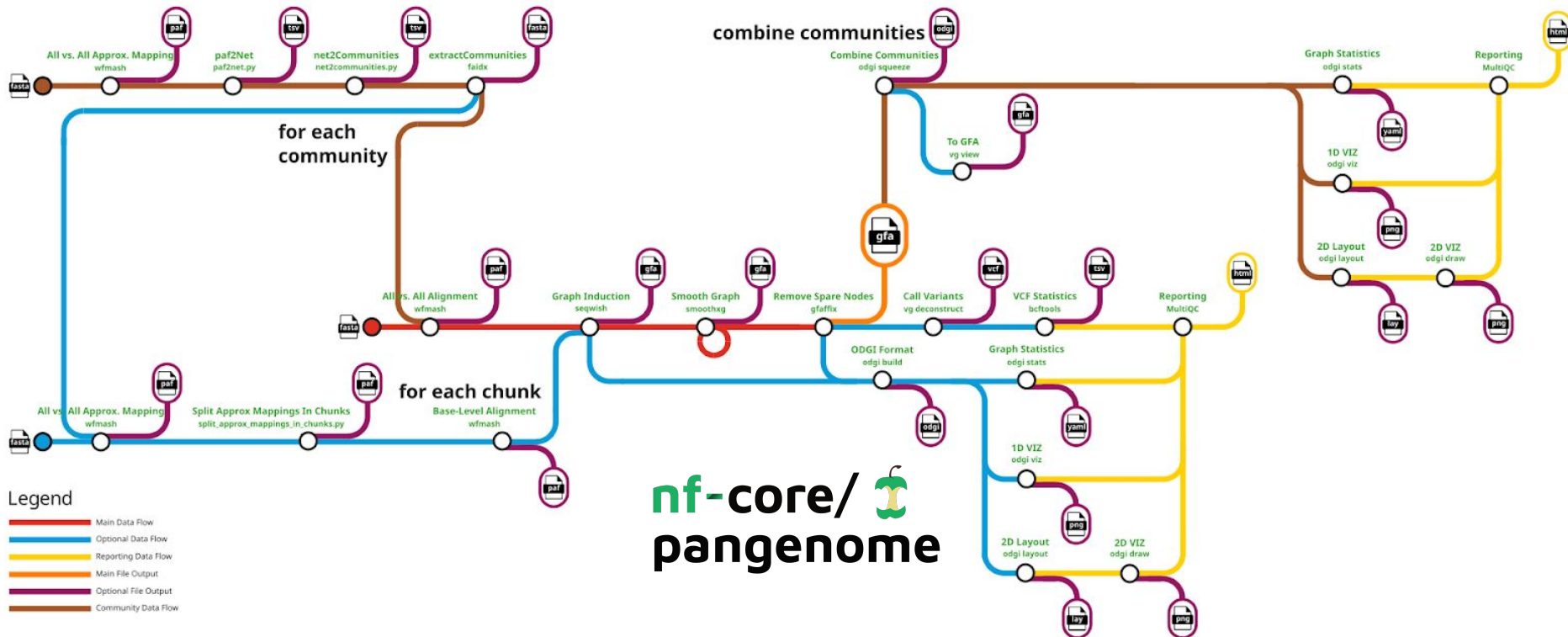
Tools

Outputs





Clustering with the [Leiden](#) algorithm: Edge weight is mapped\_length \* mapped\_identity



Clustering with the [Leiden](#) algorithm:  $\text{Edge weight} = \text{mapped\_length} * \text{mapped\_identity}$



# 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**

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? [Watch a tutorial video](#) (6:06)

don't show again ✕

## ODGI

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

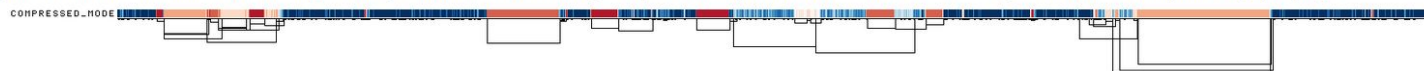
### Detailed ODGI stats table.

Copy table ☐ ☐ Configure Columns ☐ Plot 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

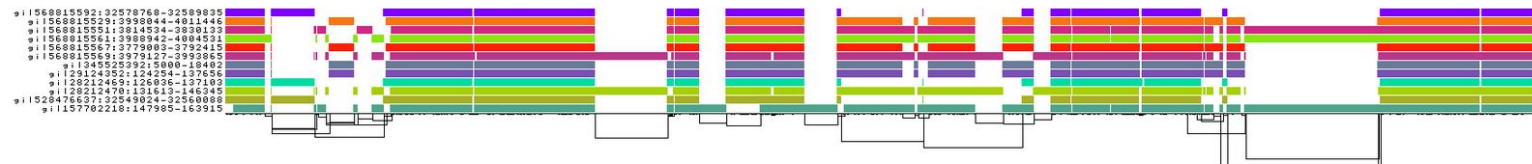
## ODGI Compressed 1D visualization

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.



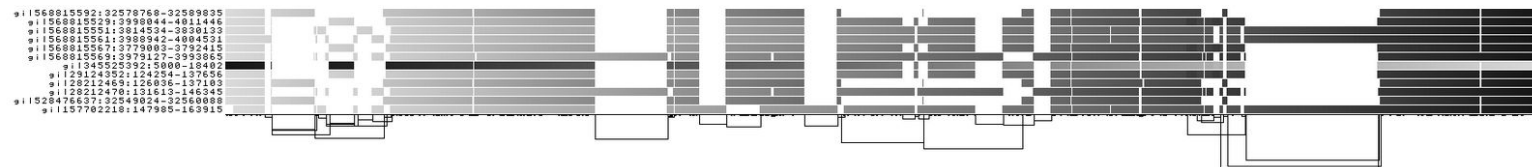
## ODGI 1D visualization

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.

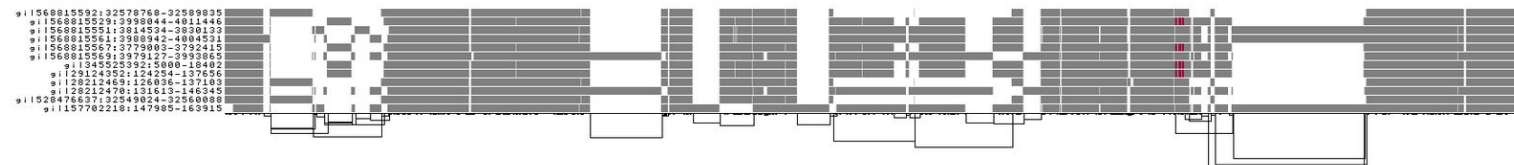


Layouts generated with PG-SGD: [Heumos, Guarracino et al., 2023.](#)

PNGs generated with ODGI: [Guarracino, Heumos et al., 2022.](#)

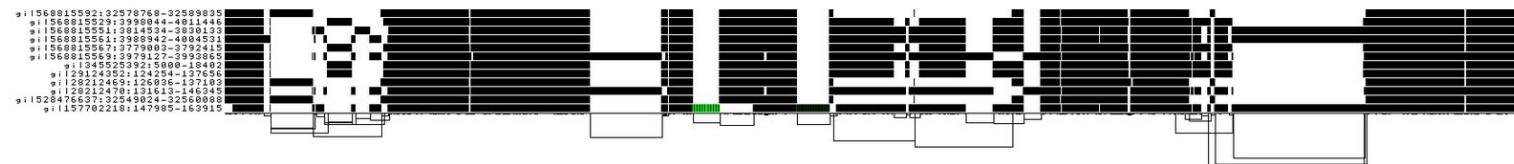
## ODGI 1D visualization by node depth

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.



## 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.



## ODGI 2D drawing

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



Layouts generated with PG-SGD: [Heumos, Guarracino et al., 2023.](#)

PNGs generated with ODGI: [Guarracino, Heumos et al., 2022.](#)





# Building a *Lodderomyces elongisporus* pangenome graph

- A yeast fungi
- An underestimated pathogen!
- Genome length: ~15Mb
- 8 chromosomes + mtDNA
- ALPACA / PANGAIA Winter Wet Lab School 2023:
  - 11 assemblies from Nanopore and Illumina data

[New Microbes New Infect.](#) 2018 Nov; 26: 20–24.

Published online 2018 Jul 18. doi: [10.1016/j.nmni.2018.07.004](https://doi.org/10.1016/j.nmni.2018.07.004)

PMCID: PMC6141678

PMID: [30245829](https://pubmed.ncbi.nlm.nih.gov/30245829/)

*Lodderomyces elongisporus*: a bloodstream pathogen of greater clinical significance


[K. Al-Obaid](#),<sup>1</sup> [S. Ahmad](#),<sup>2</sup> [L. Joseph](#),<sup>2</sup> and [Z. Khan](#)<sup>2,\*</sup>

 | Mycology | Research Article | 27 April 2023



Genomic Analyses of a Fungemia Outbreak Caused by  
*Lodderomyces elongisporus* in a Neonatal Intensive Care Unit in  
Delhi, India

Authors: [Anamika Yadav](#), [Peeyush Jain](#), [Kusum Jain](#), [Yue Wang](#), [Aditi Singh](#), [Ashutosh Singh](#), [Jianping Xu](#)    
[Anuradha Chowdhary](#)   | [AUTHORS INFO & AFFILIATIONS](#)

DOI: <https://doi.org/10.1128/mbio.00636-23> •  Check for updates





chrA



chrG



chrH



chrC



chrD



chrF



chrB



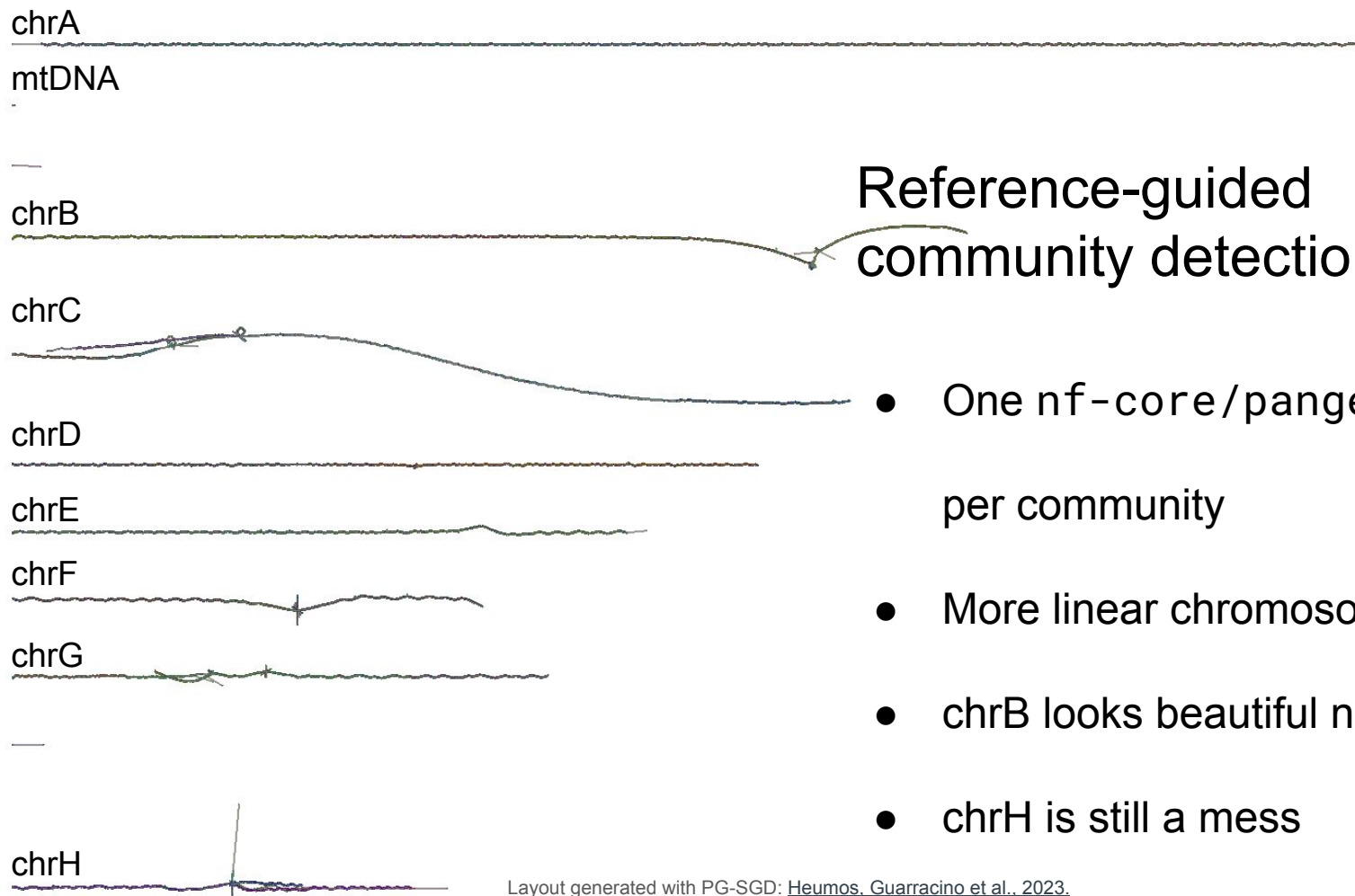
chrE



mtDNA

## nf-core/pangenome in community mode

- **9 communities:** Each community contains one chromosome!
- Most chromosomes are linear
- Thin tails → unmapped sequence
- chrB and chrH are a mess!

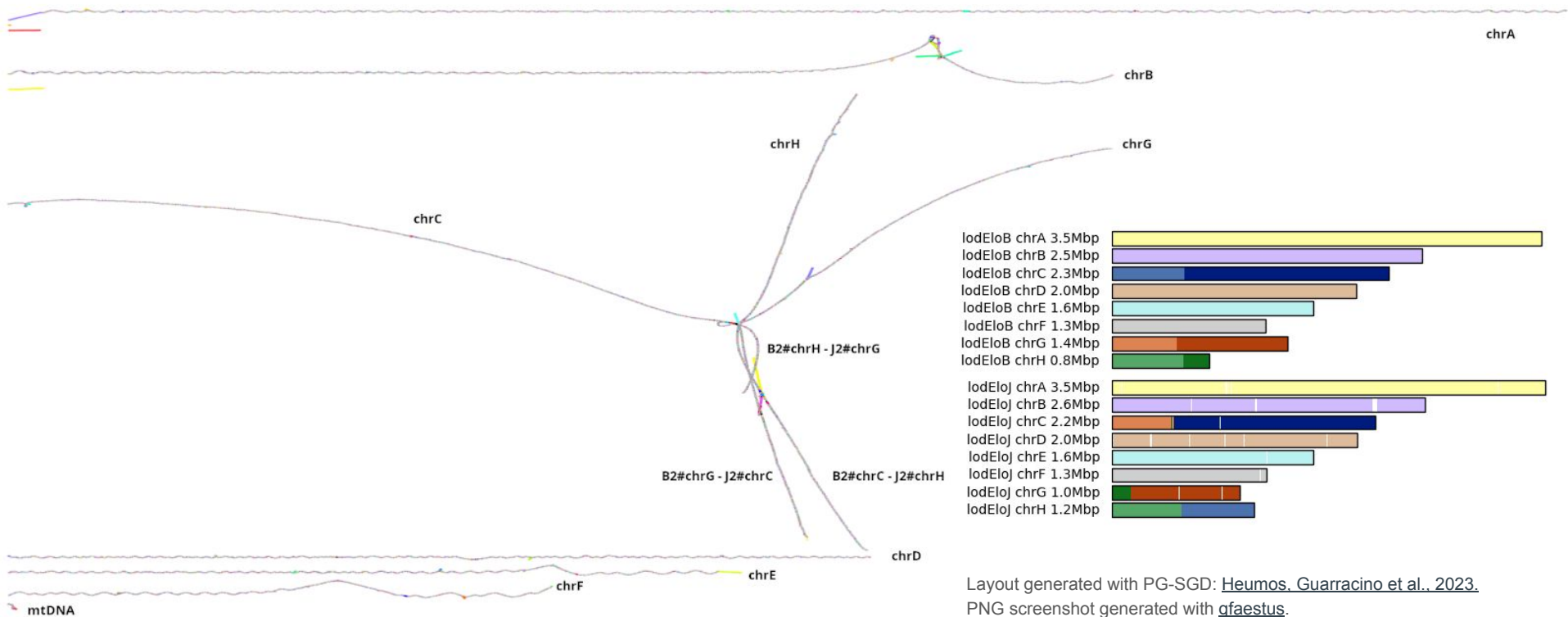


## Reference-guided community detection

- One nf-core/pangenome run per community
- More linear chromosomes
- chrB looks beautiful now
- chrH is still a mess



# Reference and assembly guided communities



Layout generated with PG-SGD: [Heumos, Guarracino et al., 2023](#).  
PNG screenshot generated with [gfaestus](#).

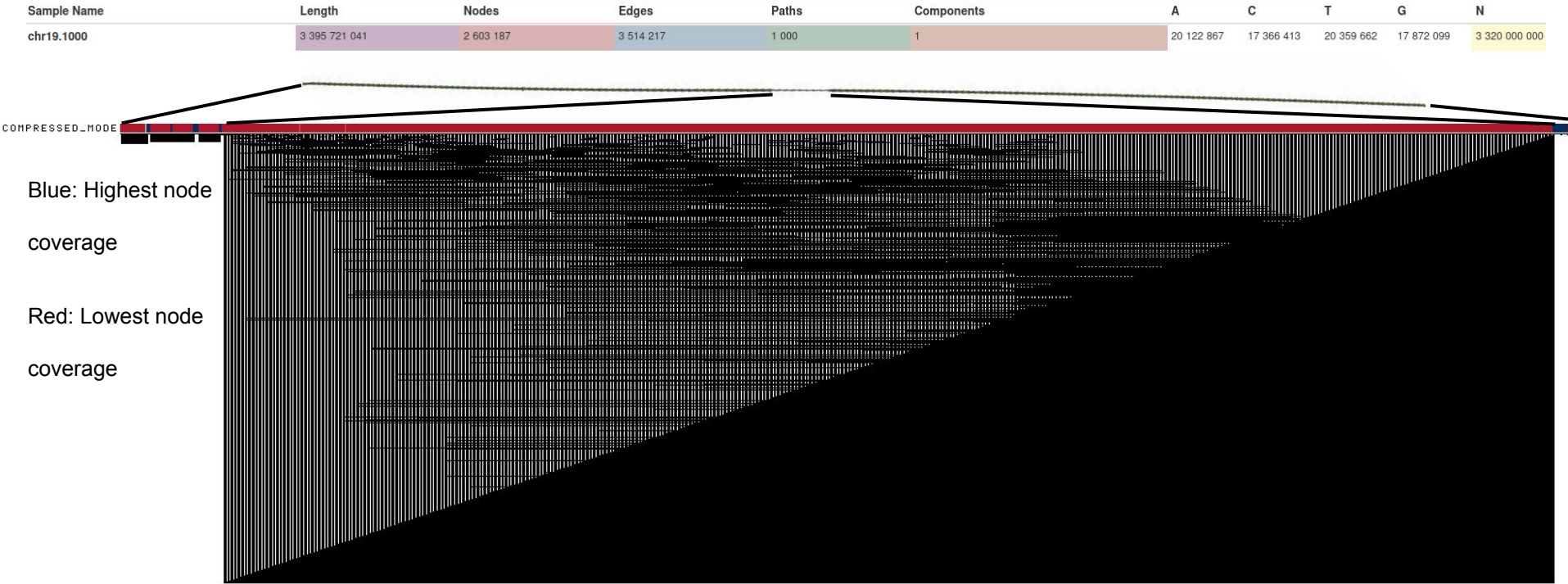


# Building a human 1KG chromosome 19 pangenome graph

- 1000 sequences of chr19 of the 1000 genomes project
- Chr19 length: ~59Mb
- Timings:
  - wfmash map: 9h
  - wfmash align: 100 \* 2h
  - seqwish: 1d 13h
  - smoothxg: 15h



# Building a human 1KG chromosome 19 pangenome graph





## Building a 2146 sequences *E. coli* pangenome graph

- 2146 sequences from GeneBank including 133 plasmids
- *E. coli* length: ~5Mb
- Quadratic all versus all alignment problem!
  - wfmash map: 1h 30min
  - wfmash align: 1000 x 20min: ~666GB of PAF files!
  - seqwish: 2TB of scratch space did not suffice!

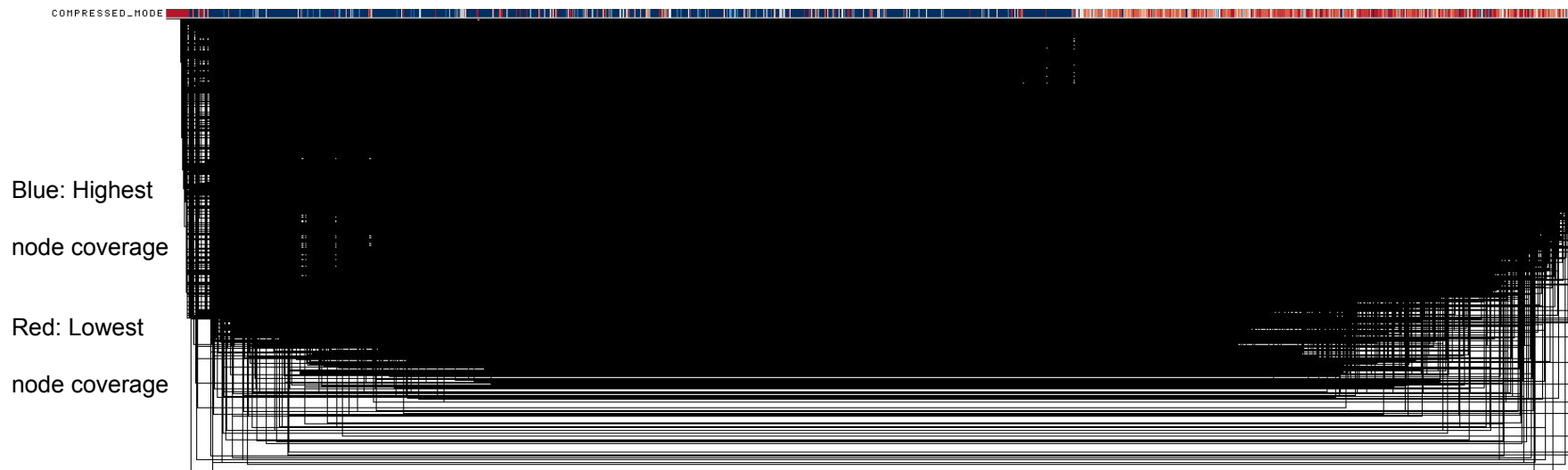


## Building a 2146 sequences *E. coli* pangenome graph

- Network storage I/O was too slow!
- wfmash sparse map factor: only retain ~0.03% of all mappings
- wfmash align: 100 x 5min - previously: 1000 x 20min
- 500GB RAM not sufficient for seqwish
- seqwish transclose batch decreased by 2 orders of magnitude: 5 hours
- smoothxg: only one round of smoothing: 62 hours

# Building a 2146 sequences *E. coli* pangenome graph

Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N
ecoli_2146	358 911 871	6 393 864	9 558 161	2 146	40	78 846 987	70 759 278	78 416 199	71 085 690	59 803 717







# Acknowledgements



Sven Nahnsen

Erik Garrison

Christian Kubica

Lukas Heumos

Oliver Kohlbacher

Andrea Guarracino

Sebastian Vorbrugg

Philipp Ehmele

Michael Krone

Pjotr Prins

Jörg Hagmann

Daniel Dörr

Gisela Gabernet

Vincenza Colonna

Jerven Bollemann

Tomas Vinar

Friederike Hanssen

Flavia Villani

Toshiyuki T. Yokoyama

Brona Brejova

Antonia Schuster

David G. Ashbrook

Torsten Pook

Jozeph Nosek

Júlia Mir Pedrol

Robert W. Williams

Franziska Huth

LodElo Consortium

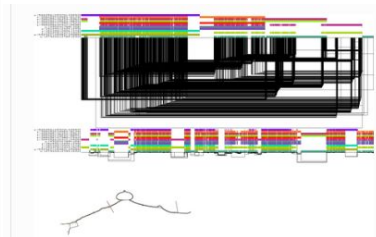
Susanne Jodoin

Christian Fischer



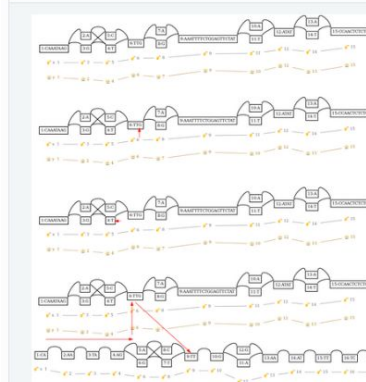
# Downstream analyses with ODGI

	<p><b>Exploratory Analysis</b></p> <p>Translate GFAv1 to ODGI format Highlight different graph features in 1D Create 1D visualization of a particular region</p>
	<p><b>Detect Complex Regions</b></p> <p>Download human chr8 pangenome Calculate depth over pangenome Plot the depth Explore the centromer's organization</p>
	<p><b>Extract Selected Loci</b></p> <p>Extract a subgraph of LPA graph Visualize subgraph Extract MHC locus of human chr6 Visualize MHC locus</p>



## 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



# 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.

Read more: <https://multiqc.info/>

#### ODGI

ODGI is an open-source graph visualization tool that allows users to visualize and interact with graph data in a web browser.

Read more: <https://odgi.io/>

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

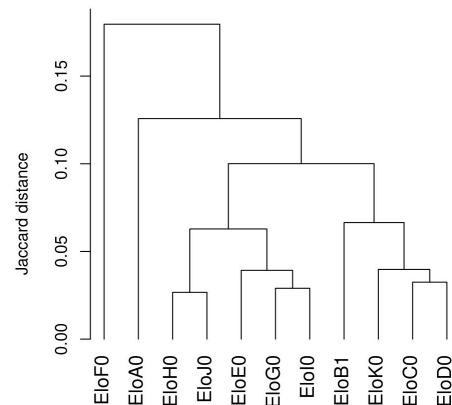
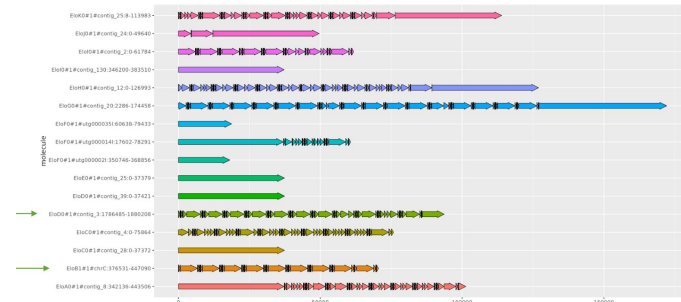
ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

ODGI is a web-based tool for visualizing and interacting with graph data.

## MultiQC Report of Graph Statistics

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





## Core Facility Cluster - hardware specifications

- 28 nodes
- Parallel BeeGFS Filesystem (SFS9) with a total capacity of 400TB
- 24 Regular nodes each:
  - 32Cores/64Threads (2\* AMD EPYC 7343), 512GB RAM, 2TB NVME
- 4 HighMem nodes each:
  - 64Cores/128Threads (2\* AMD EPYC 7513), 2048GB RAM, 4TB NVME
- I can occupy at most 1228 CPU Threads ~19 nodes.

ID=LODELOB2\_p59160;Name=cox1-GENE;  
ID=LODELOB2\_p59170;Name=atp8-GENE;  
ID=LODELOB2\_p59180;Name=atp6-GENE;  
ID=LODELOB2\_p59190;Name=cox3-GENE;  
ID=LODELOB2\_p59200;Name=nad5-GENE;  
ID=LODELOB2\_p59210;Name=nad4L-GENE;



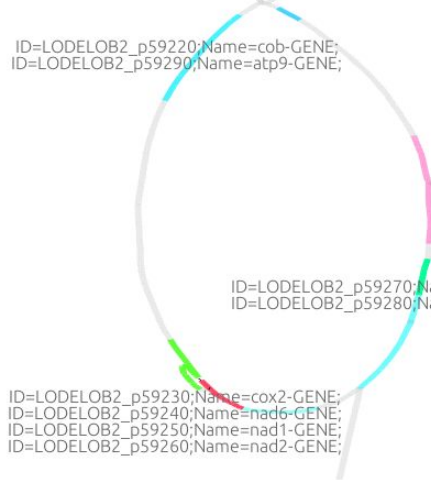
This is the annotated mtDNA graph.

Nodes are colored by annotation.

As will be the case for the following 2D plots.

The black links are the edges in the graph.

ID=LODELOB2\_p59220;Name=cob-GENE;  
ID=LODELOB2\_p59290;Name=atp9-GENE;



ID=LODELOB2\_p59270;Name=nad3-GENE;  
ID=LODELOB2\_p59280;Name=nad4-GENE;

ID=LODELOB2\_p59230;Name=cox2-GENE;  
ID=LODELOB2\_p59240;Name=nad6-GENE;  
ID=LODELOB2\_p59250;Name=nad1-GENE;  
ID=LODELOB2\_p59260;Name=nad2-GENE;

Shown is the complex region in the middle, where all chromosomes cross each other.

