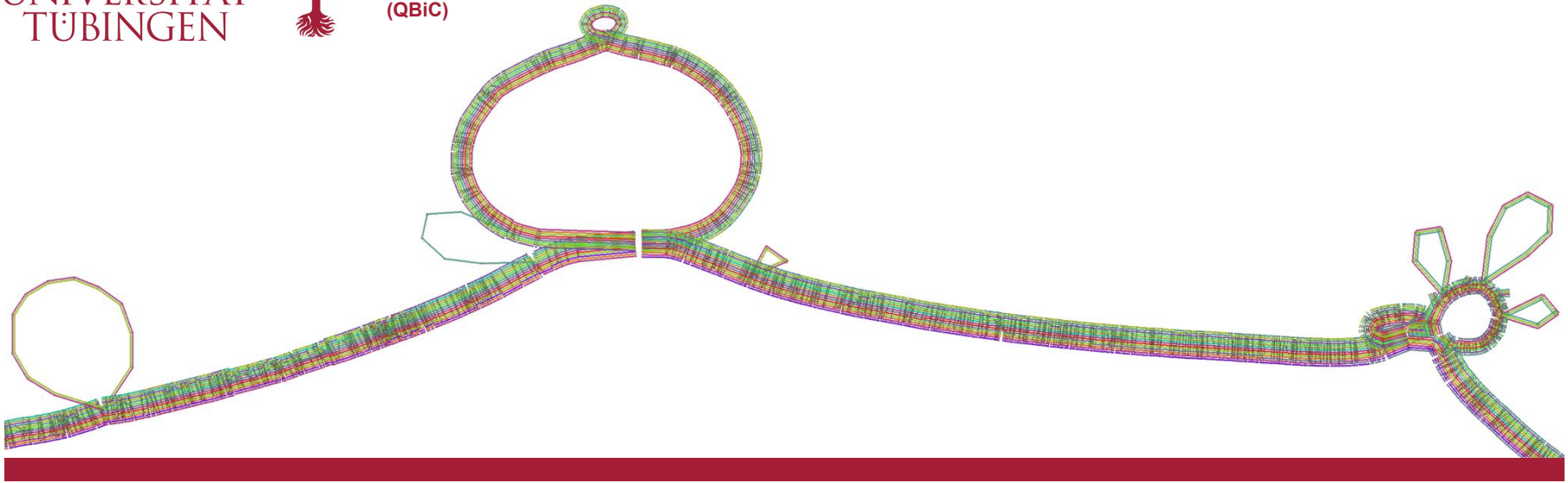
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Quantitative Biology Center
(QBiC)



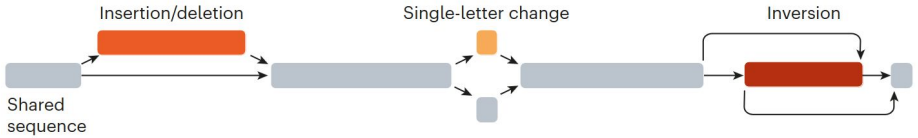
Cluster scalable pangenome graph construction with

nf-core/pangenome

Simon Heumos - Nextflow Summit 2023 Barcelona

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.

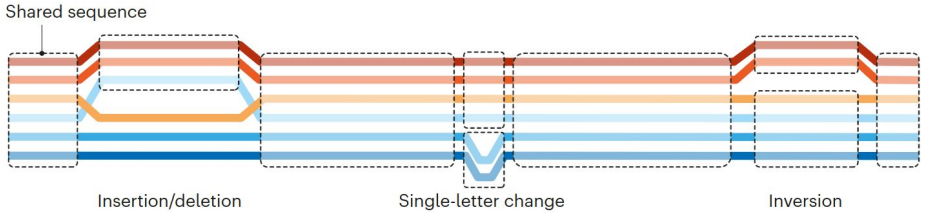
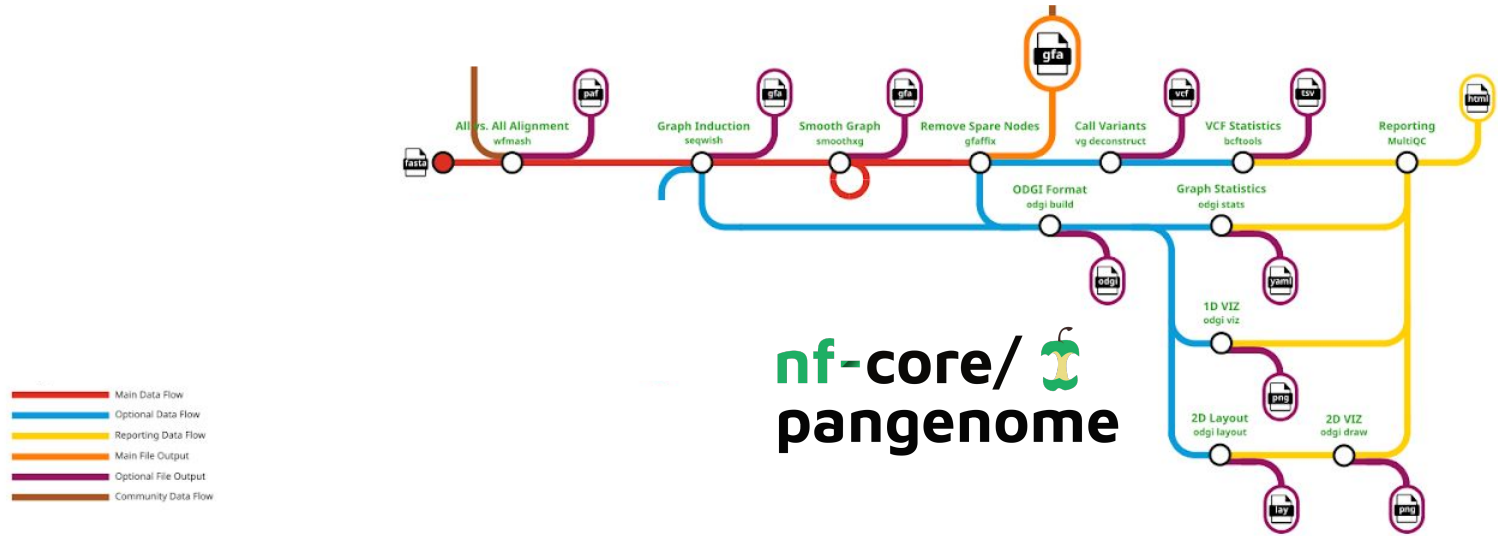
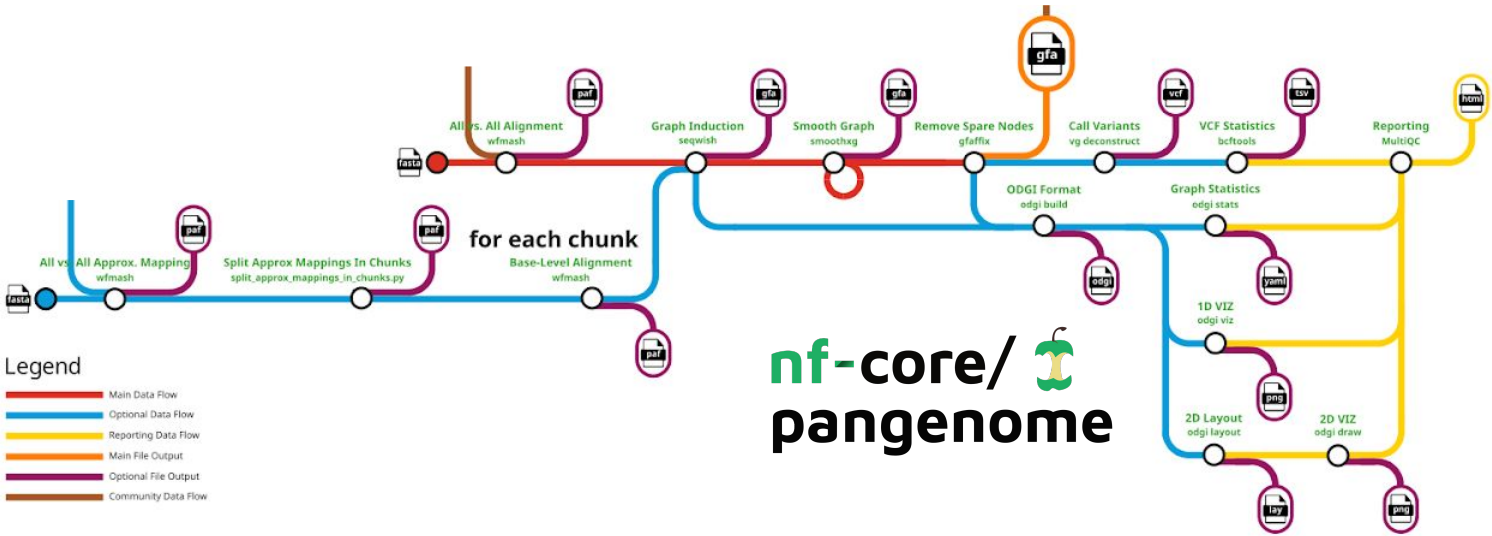
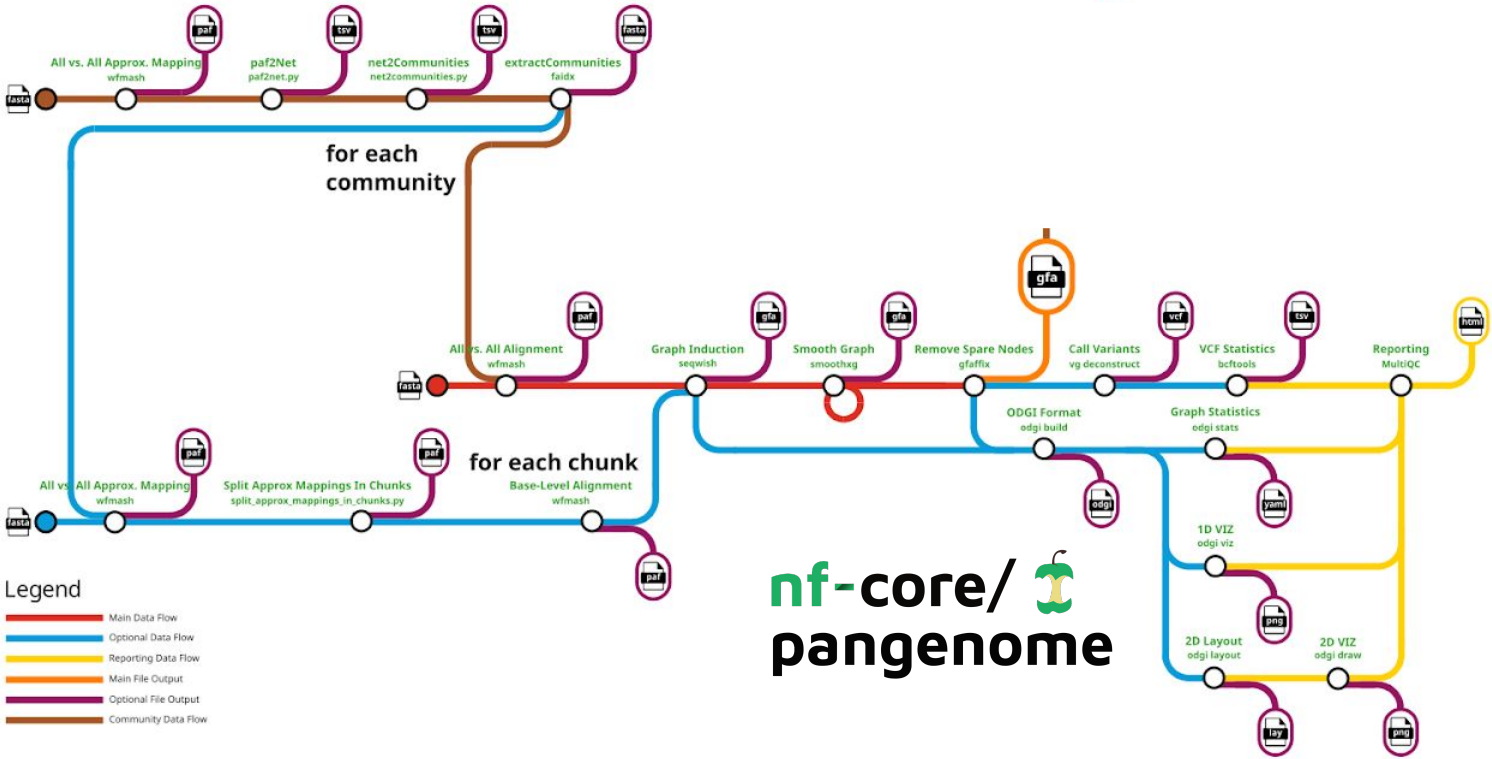


Figure from Eisenstein 2023.



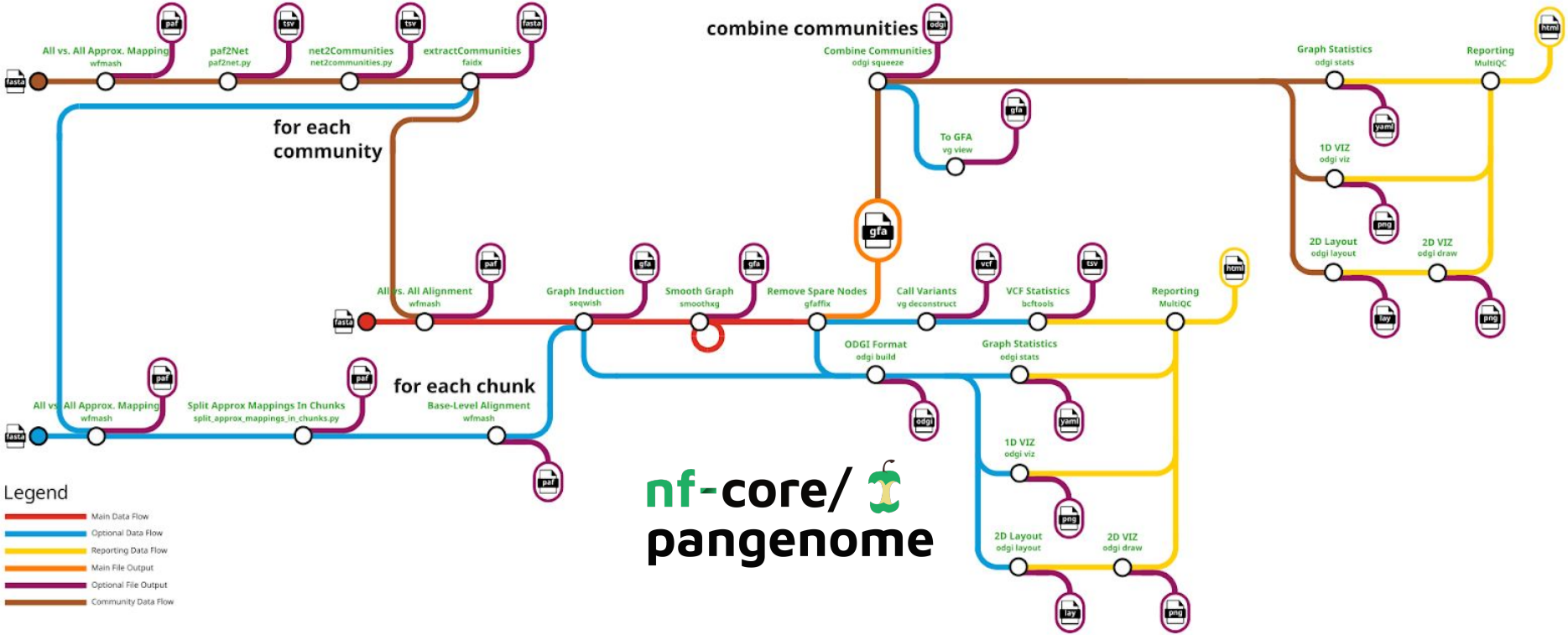
Core workflow taken over from PGGB: Garrison, Guarracino et al., 2023.


All vs. All Alignment
wfmash
Graph Induction
seqwish
Smooth Graph
smoothxg
Remove Spare Nodes
gfaffix
Call Variants
vg deconstruct
VCF Statistics
bcftools
Reporting
MultiQC
ODGI Format
odgi build
Graph Statistics
odgi stats
1D VIZ
odgi viz
2D Layout
odgi layout
2D VIZ
odgi draw
All vs. All Approx. Mapping
wfmash
Split Approx Mappings In Chunks
split_approx_mappings_in_chunks.py
for each chunk
Base-Level Alignment
wfmash
Legend
Main Data Flow
Optional Data Flow
Reporting Data Flow
Main File Output
Optional File Output
Community Data Flow
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.

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
PMCID: PMC6141678
PMID: 30245829

Lodderomyces elongisporus: a bloodstream pathogen of greater clinical significance

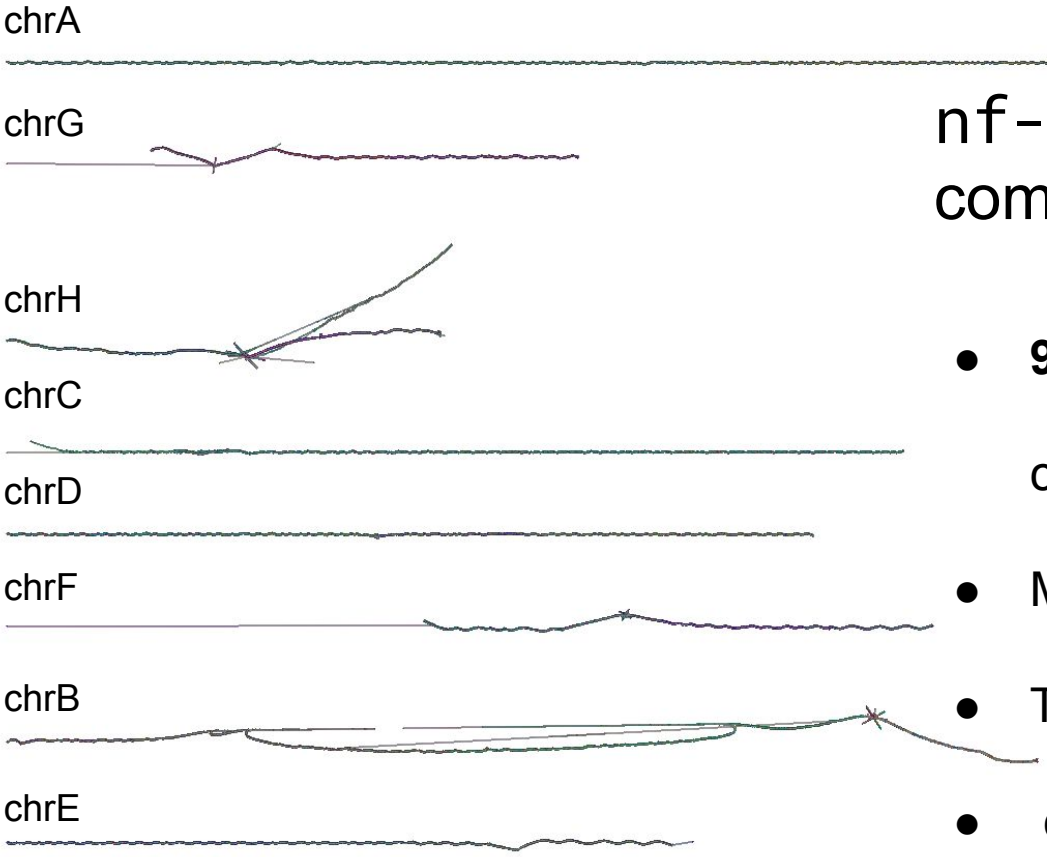
K. Al-Obaid,[1] S. Ahmad,[2] L. Joseph,[2] and Z. Khan[2,*]

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

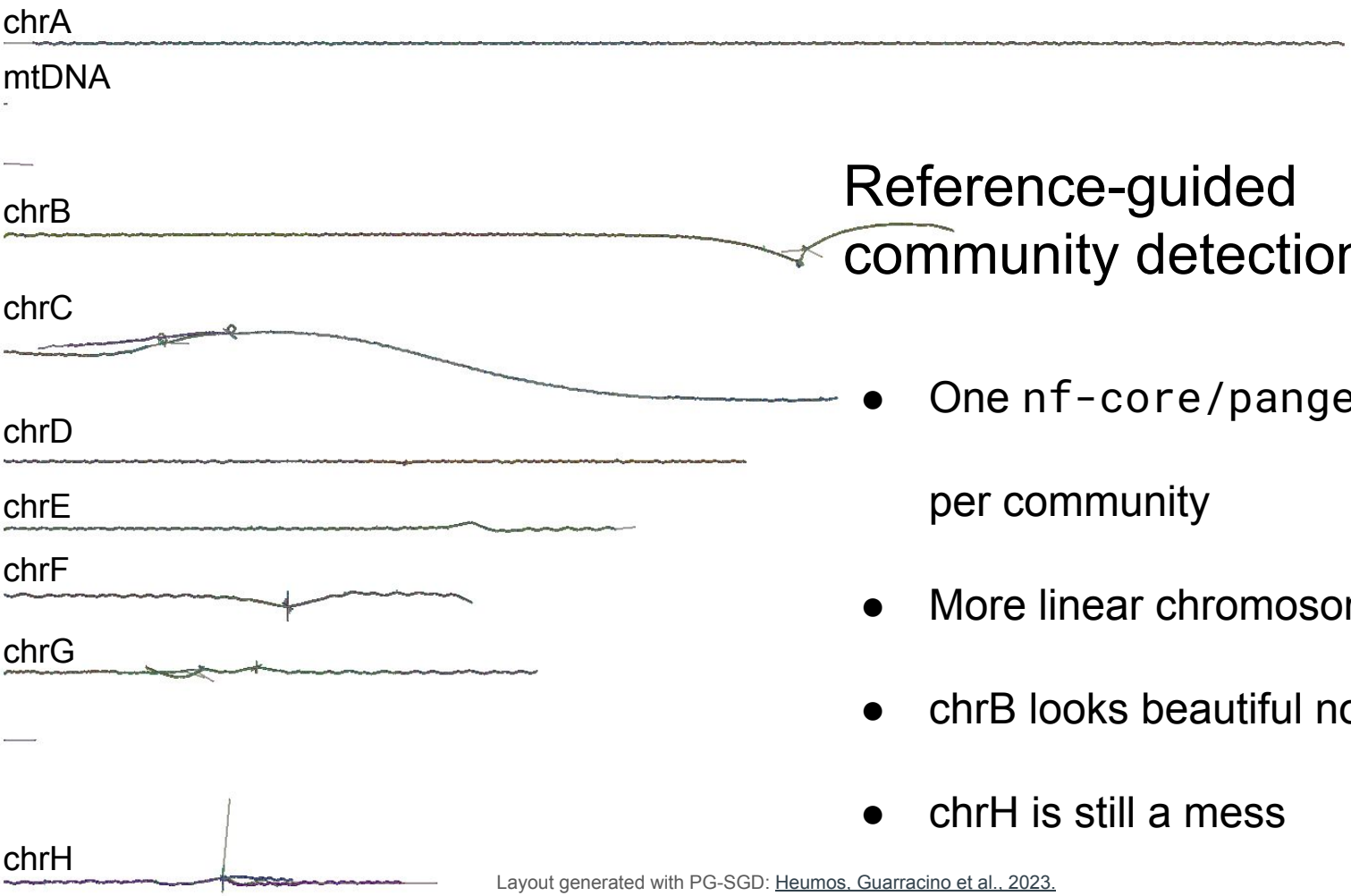
DOI: https://doi.org/10.1128/mbio.00636-23 · Check for updates



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!

Layout generated with PG-SGD: Heumos, Guarracino et al., 2023.
PNG generated with ODGI: Guarracino, Heumos et al., 2022.

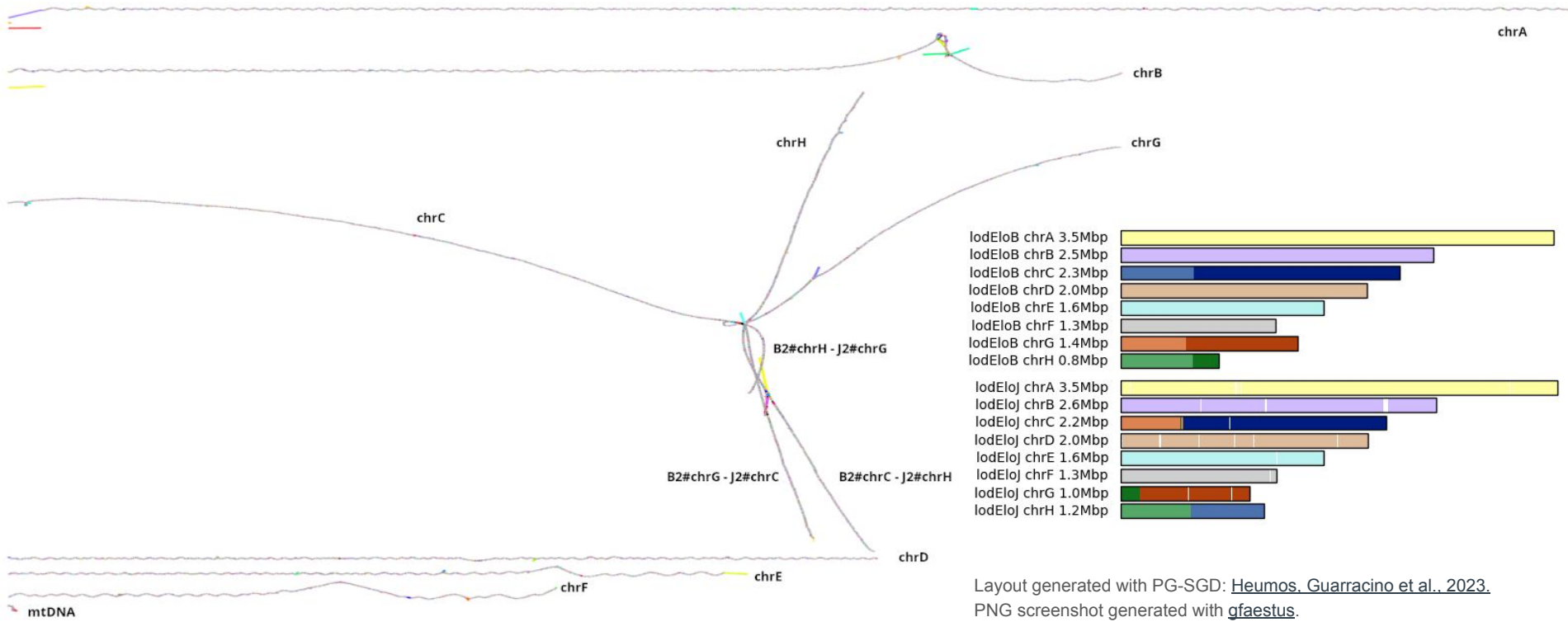


Reference-guided community detection

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

Layout generated with PG-SGD: Heumos, Guarracino et al., 2023.
PNG generated with ODGI: Guarracino, Heumos et al., 2022.

Reference-guided and assembly-guided communities



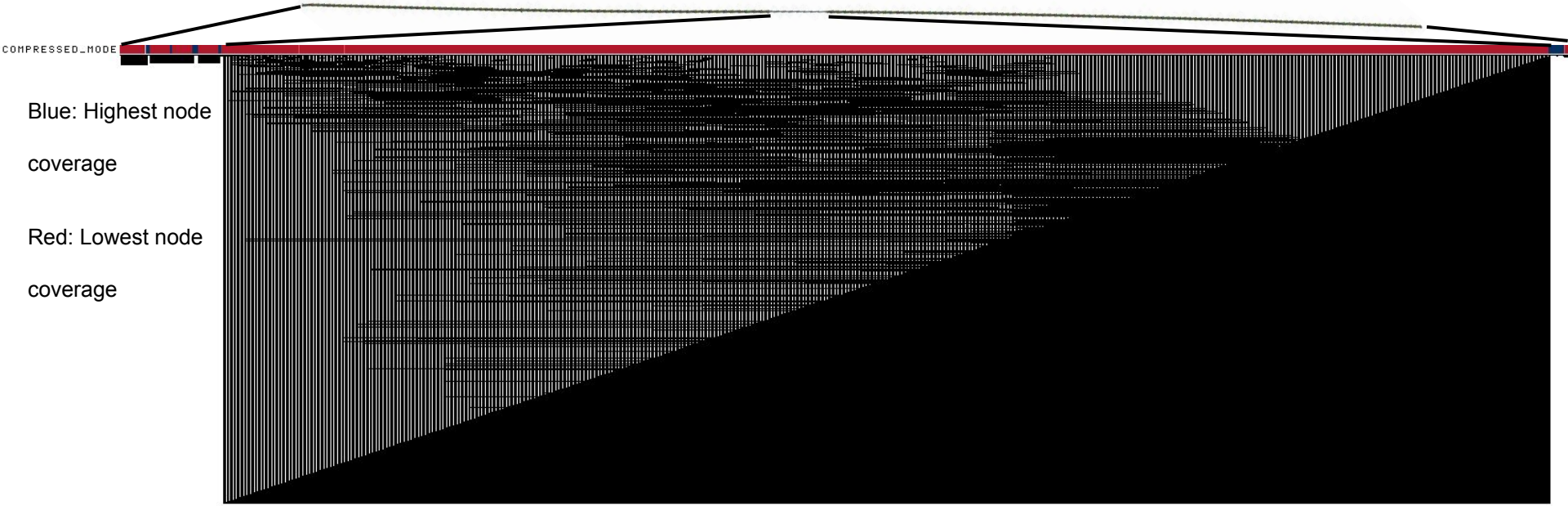
Layout generated with PG-SGD: Heumos, Guarracino et al., 2023.
PNG screenshot generated with gfaestus.

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

Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N
chr19.1000	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



Blue: Highest node coverage

Red: Lowest node coverage

Building a 2146 sequences *E. coli* pangenome graph

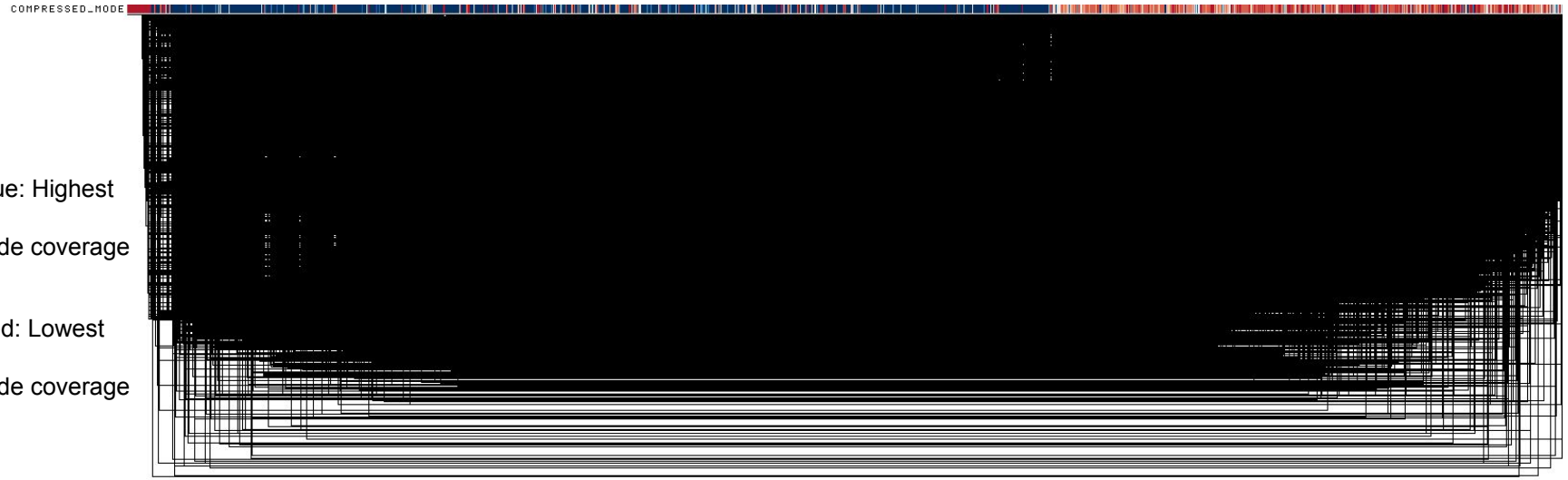
- 2146 sequences from GeneBank including 133 plasmids
- *E. coli* length: ~5Mb
- Quadratic all versus all alignment problem!
  - `wfmash map`: 1h 30min
  - `wfmash align`: 1000 x 20min: ~666GB of PAF files!
  - `seqwish`: 2TB of scratch space did not suffice!

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



Blue: Highest node coverage

Red: Lowest node coverage

Acknowledgements

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

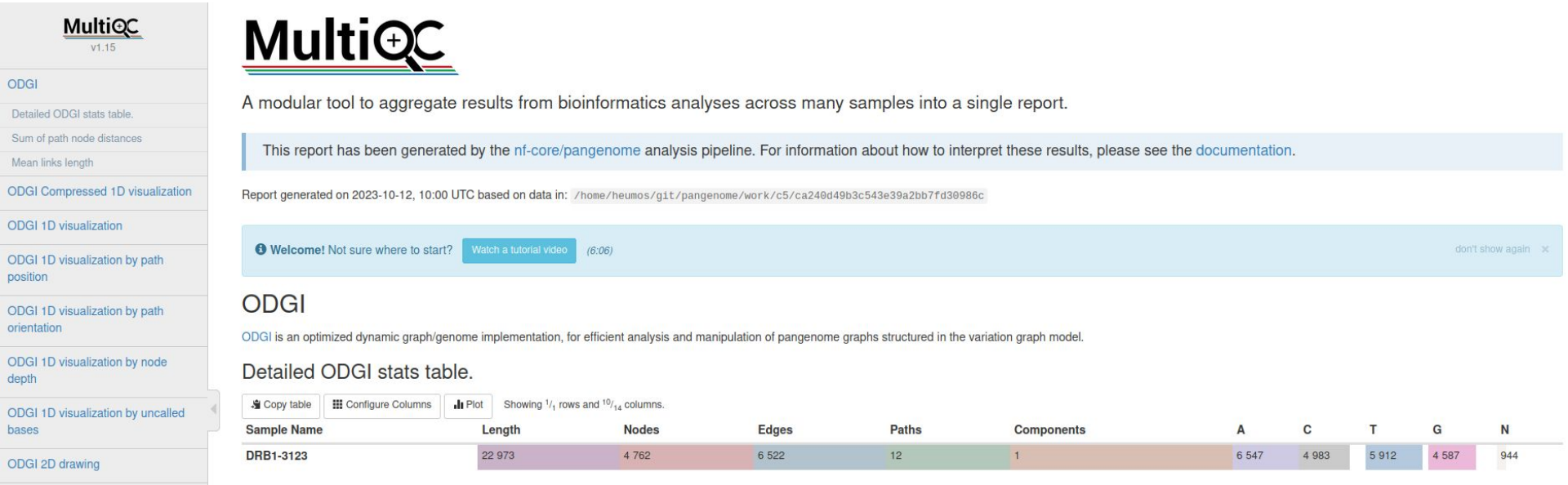
ZIM
Zentrales Innovationsprogramm Mittelstand

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

HUMAN PANGENOME

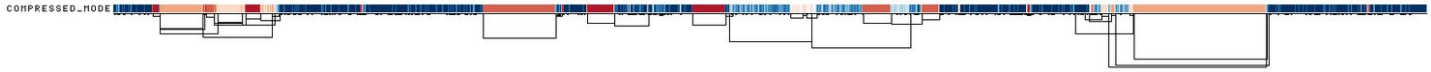
nf-core

MultiQC Report



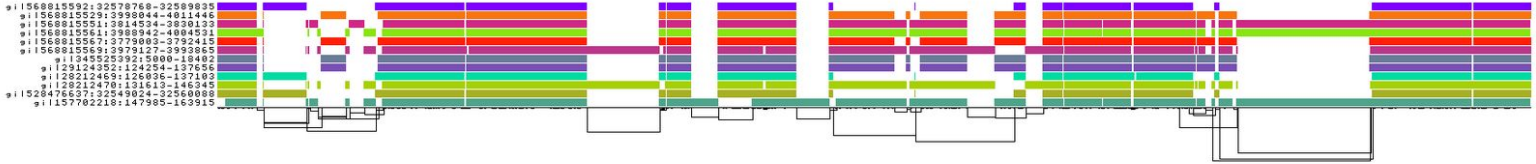
ODGI Compressed 1D visualization

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

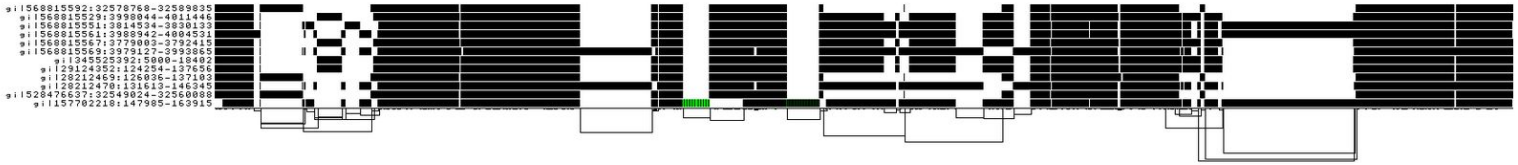
This image shows a 1D rendering of the built pangenome graph where the paths are colored by orientation. Forward is black, reverse is red.

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.

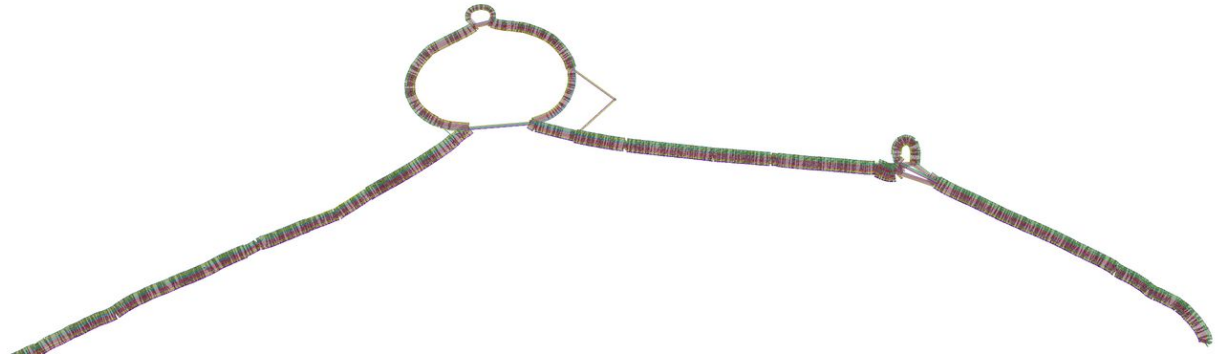
ODGI 1D visualization by uncalled bases

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.

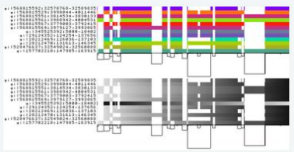


ODGI 2D drawing

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

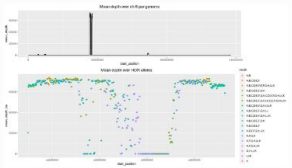


Downstream analyses with ODGI



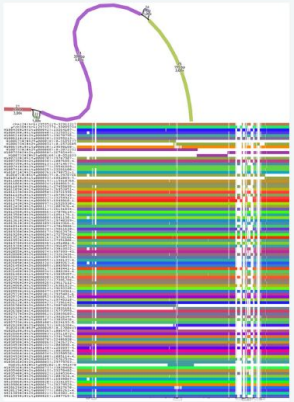
Exploratory Analysis

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



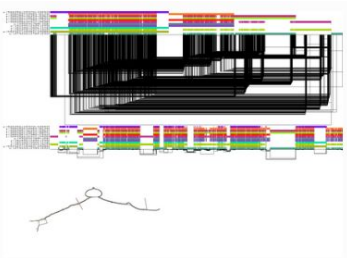
Detect Complex Regions

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



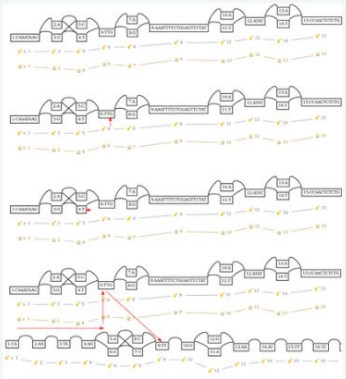
Extract Selected Loci

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



Sorting and Layouting

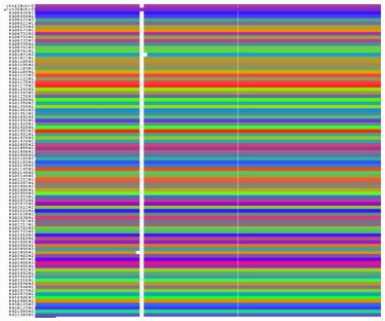
Sort DRB1-3123 graph
Metrics of sorted and unsorted graph
Compare 1D visualizations
2D layout of DRB1-3123 graph
2D drawing of DRB1-3123 graph
gfaestus for interactive visualization



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

Downstream analyses with ODGI

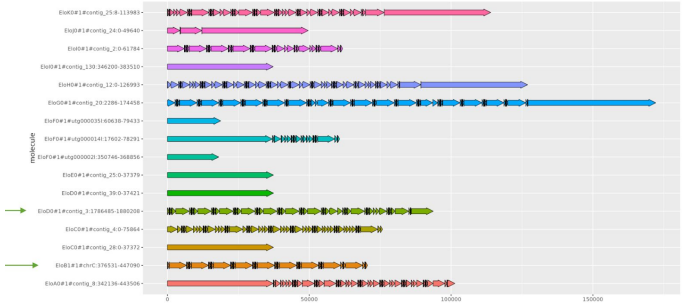


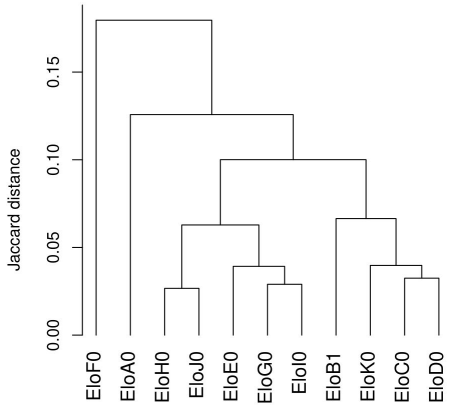
Remove Artifacts and Complex Regions

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

MultiQC Report of Graph Statistics

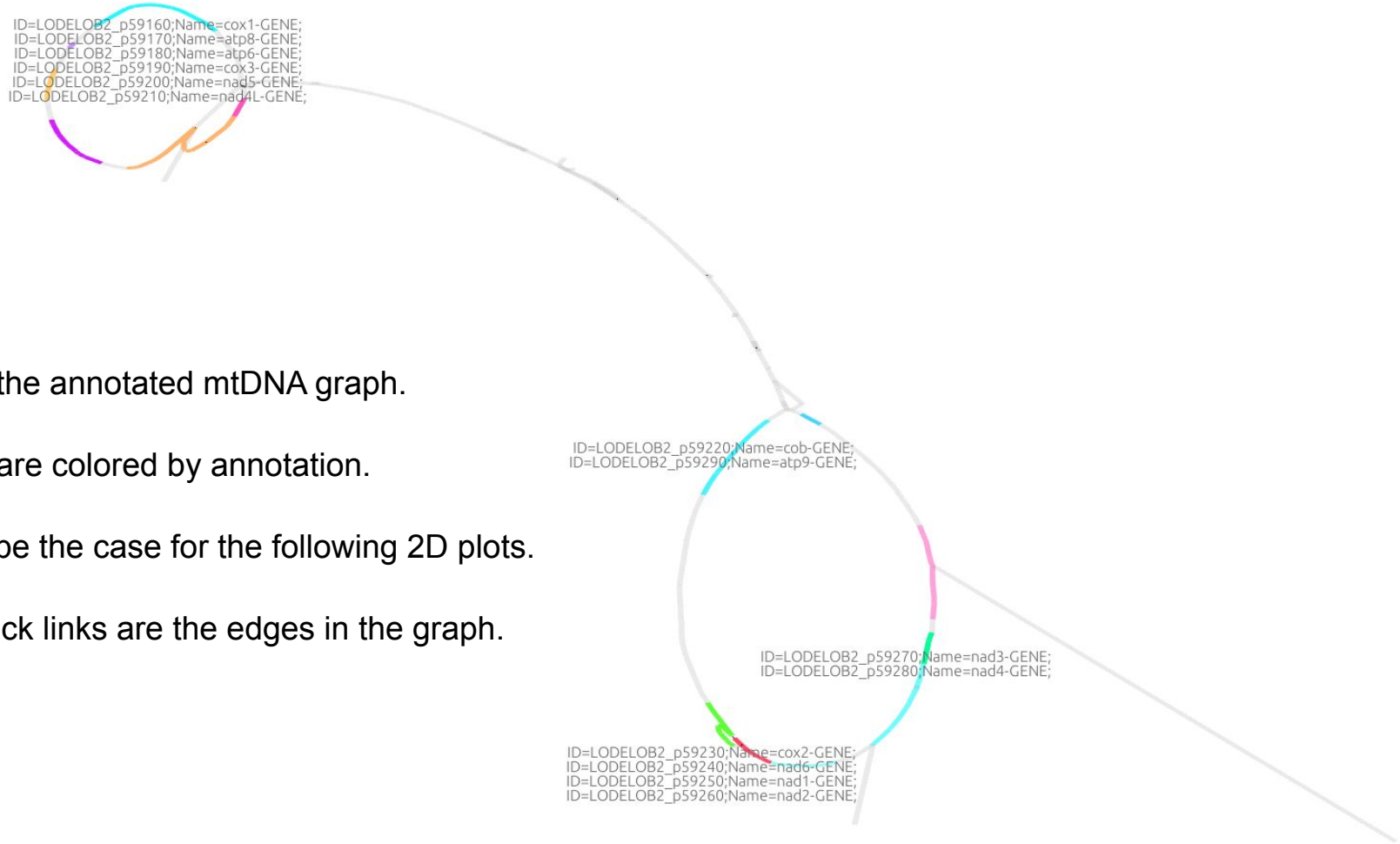
Create graph statistics
Apply MultiQC to statistics YAML
Integrate 1D and 2D visualizations into the report





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.



This is the annotated mtDNA graph.

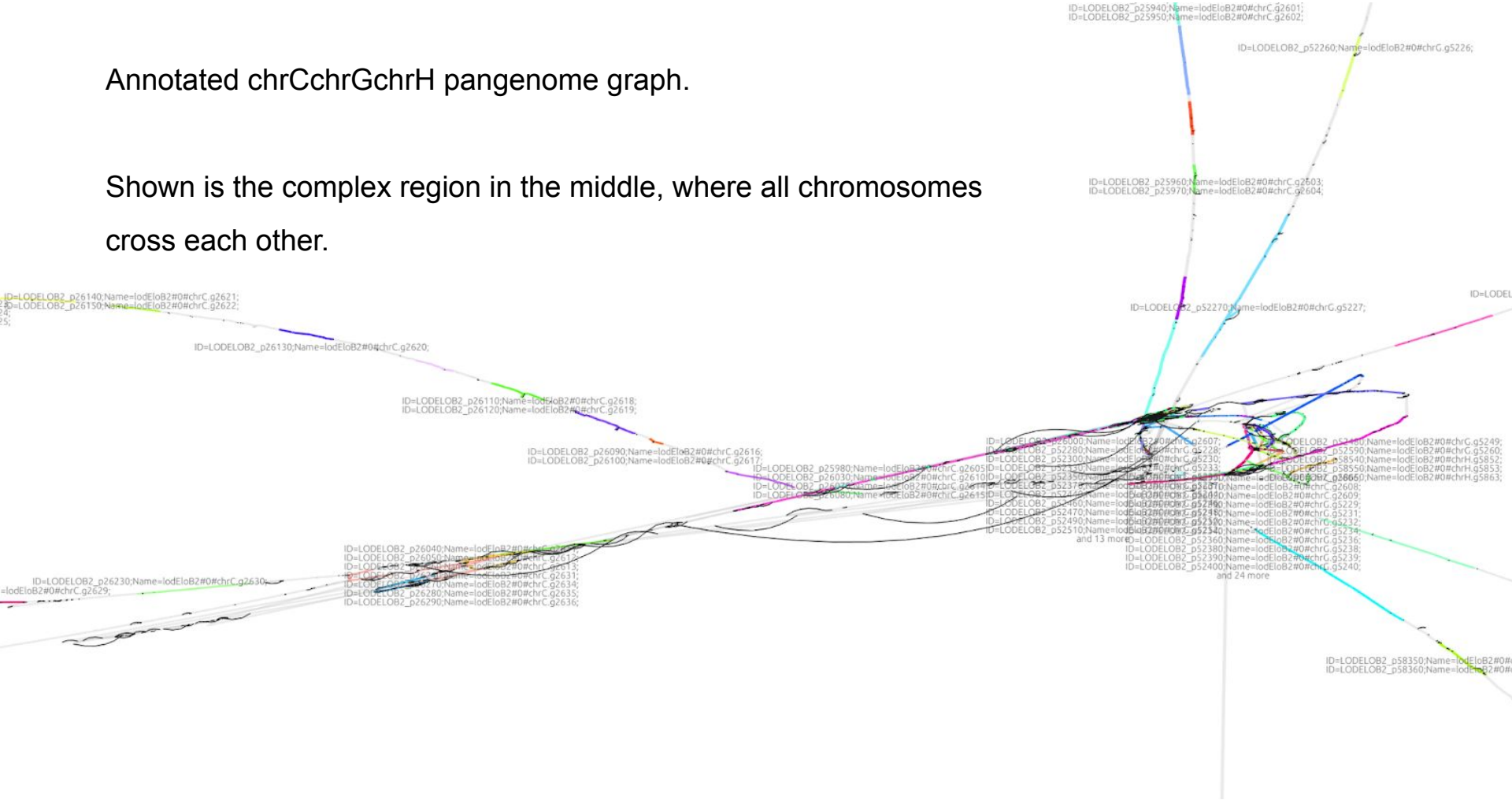
Nodes are colored by annotation.

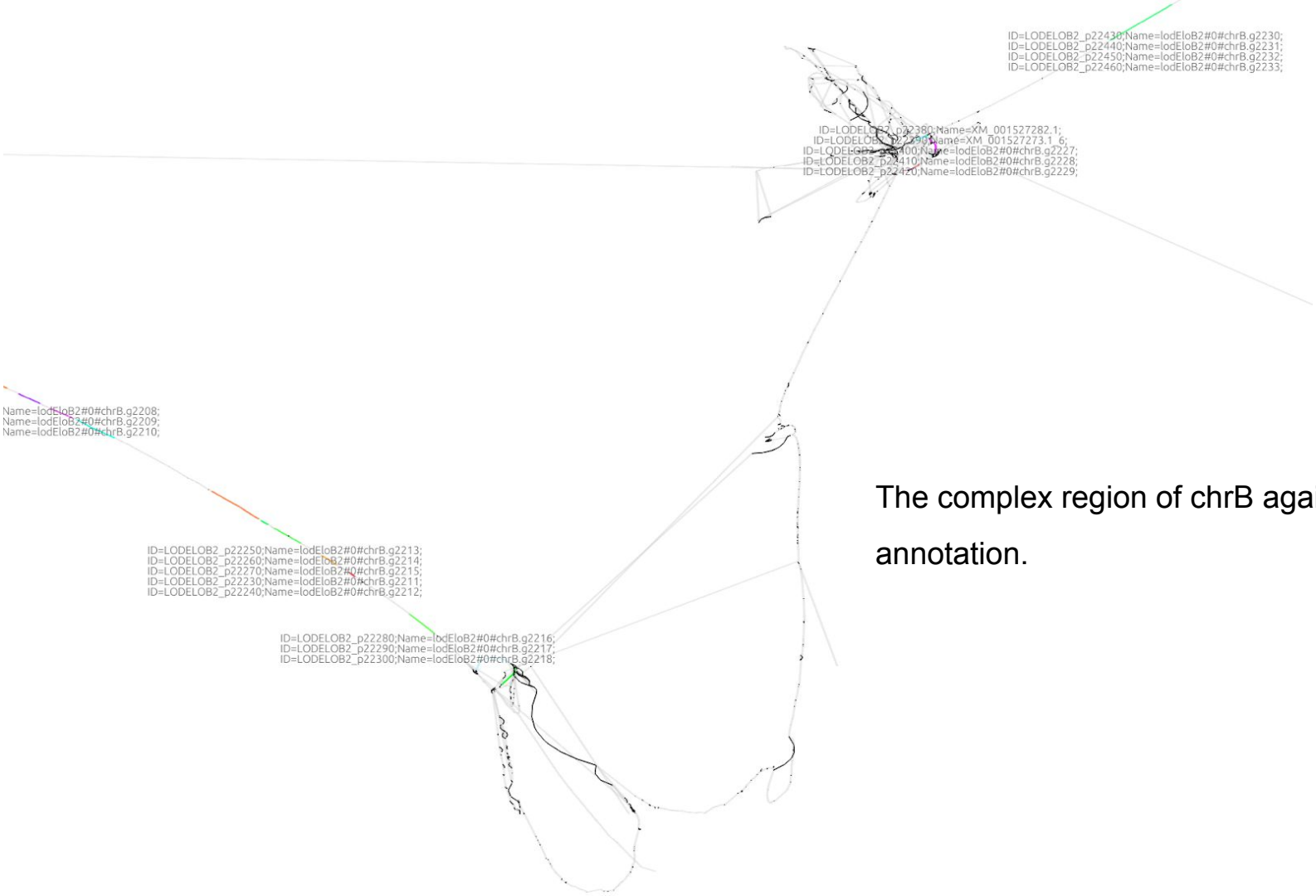
As will be the case for the following 2D plots.

The black links are the edges in the graph.

Annotated chrCchrGchrH pangenome graph.

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





The complex region of chrB again with annotation.