Exploring pangenome graphs and possible applications

Institute for Medical Biometry and Bioinformatics
21 October 2021

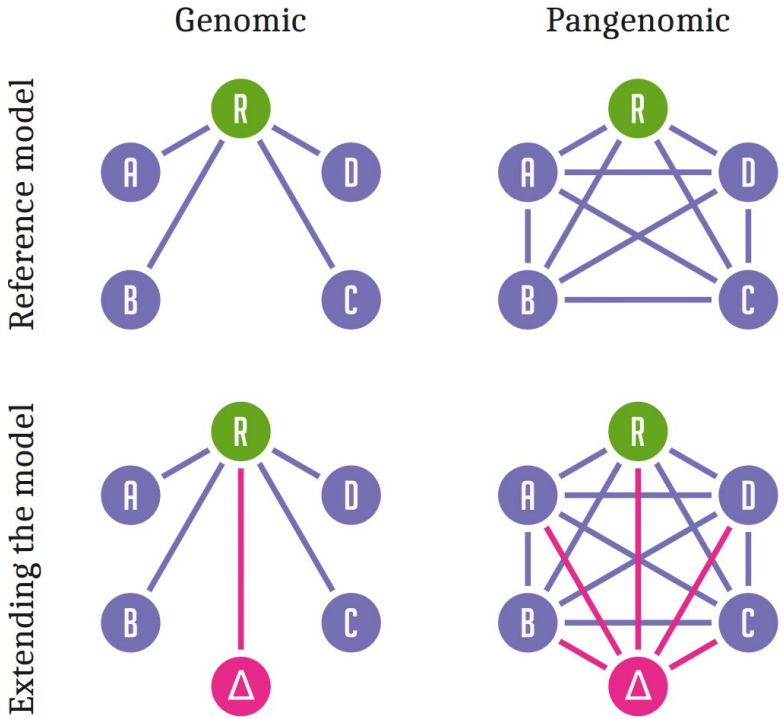
Simon Heumos*, Andrea Guarracino,
Jerven Bolleman, Toshiyuki T. Yokoyama, Pjotr Prins,
Sven Nahnsen, Erik Garrison

*Quantitative Biology Center (QBiC), Tübingen

QBiC

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

De novo assembly and a pangenomic model

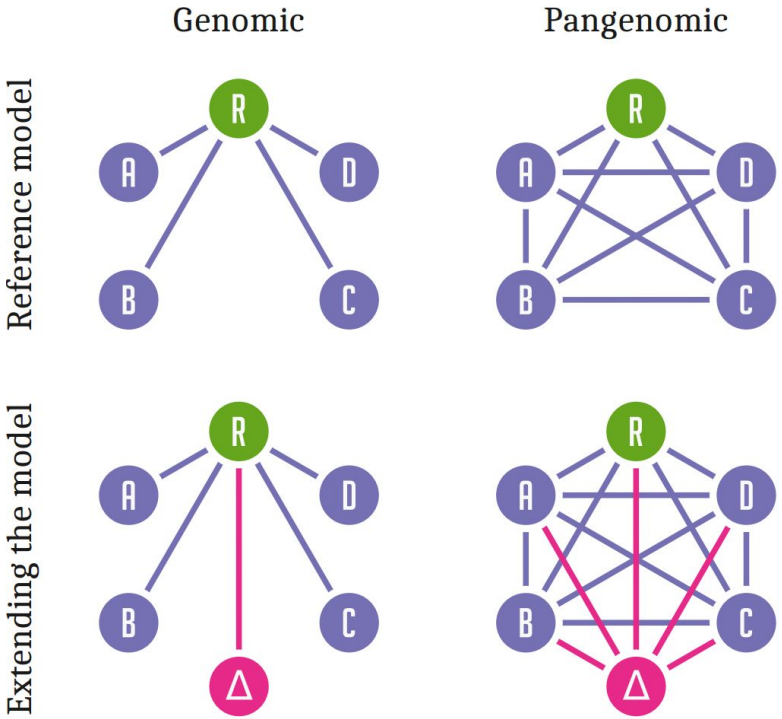


Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

Thanks to advances in sequencing technology, new **telomere-to-telomere** 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.

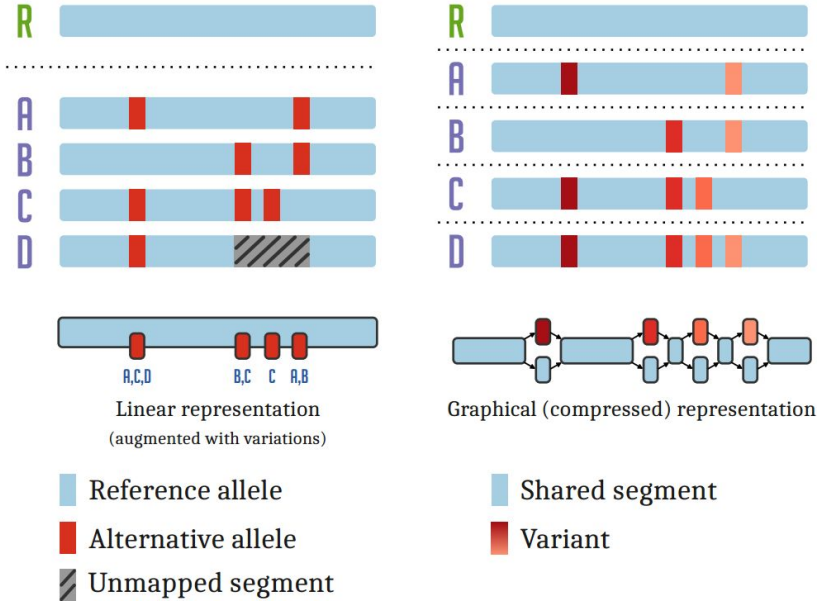


Figure from Eizenga et al., 2020.

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.

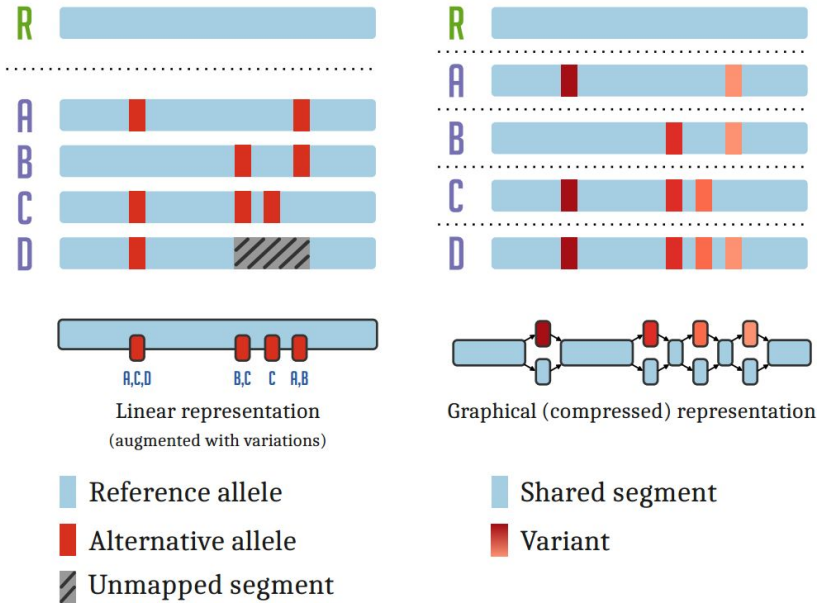


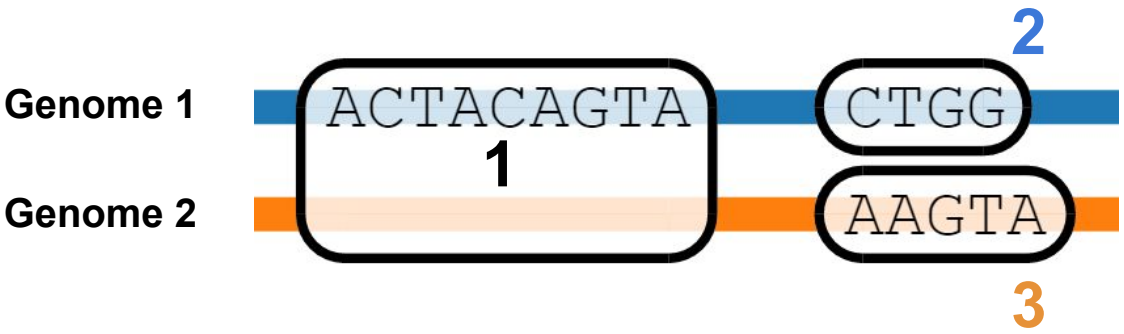
Figure from Eizenga et al., 2020.

Variation graphs

Genome 1: **ACTACAGTACTGG** Path: **1 2**

Genome 2: **ACTACAGTAAAGTA** Path: **1 3**

Linear sequences are **paths** through nodes.



Graph topology is not directly shown.

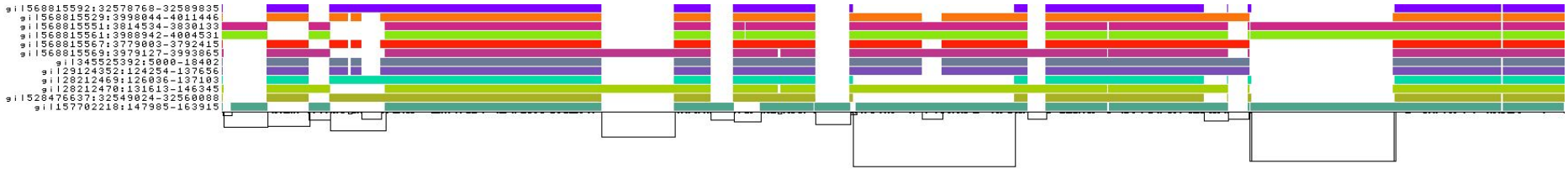
The nodes represent DNA sequences.

Sketch made using SequenceTubeMap.

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

1D Graph visualization explained

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

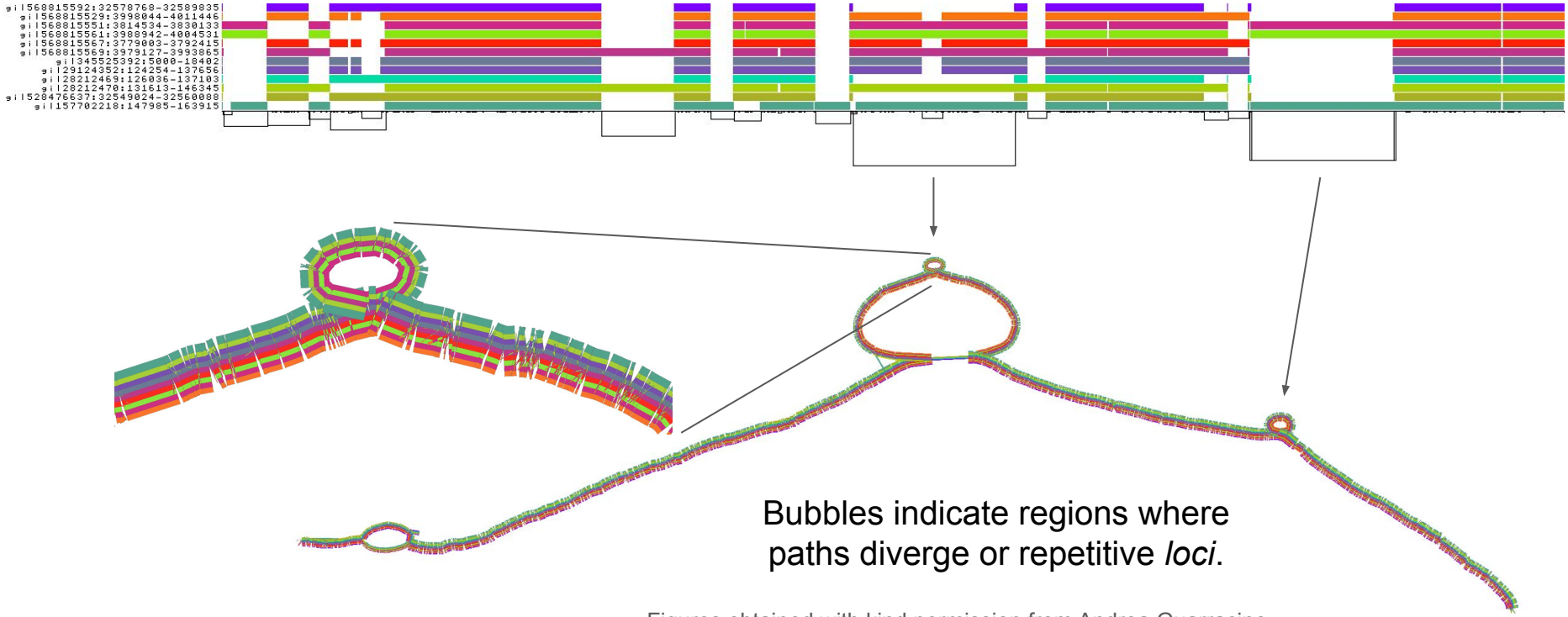


- The graph nodes are arranged from left to right, forming the pangenome sequence.
- The colored bars represent the paths versus the pangenome sequences in a binary matrix.
- The path names are placed on the left.
- The black lines under the paths are the links, which represent the graph topology.

Figure obtained with kind permission from Andrea Guarracino

2D Graph visualization explained

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



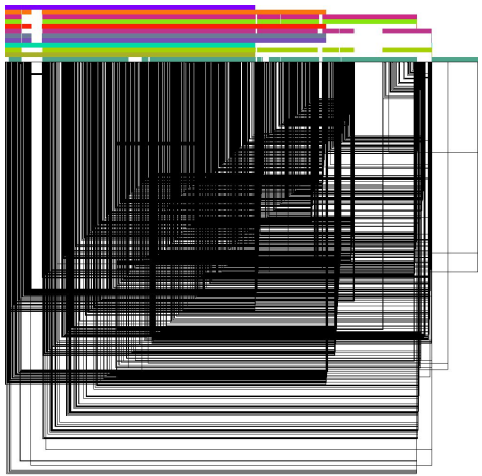
Figures obtained with kind permission from Andrea Guarracino

1D Graph Sorting and 2D Layouting by Path-Guided Stochastic Gradient Descent (PG-SGD)

A pangenome graph induced from raw alignments can be very complex and hard to analyse downstream.

Solution Make the graph more linear by reordering of nodes.

- Visualization
- Comparative genomics
- Mapping
- Interpretation



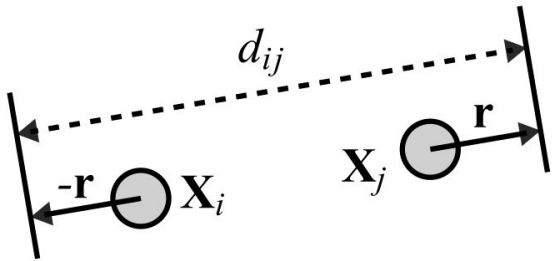
Raw graph built using seqwish.

Erik tried more than 10 sorting algorithms, but none of them did the job. The most promising one was Graph drawing by Stochastic Gradient Descent

1D Graph Sorting by P-SGD - The Algorithm Explained

Objective: Move a single pair of nodes at a time.

Optimizing the disparity between the layout distance of a node pair and the actual nucleotide distance of a path traversing these nodes.



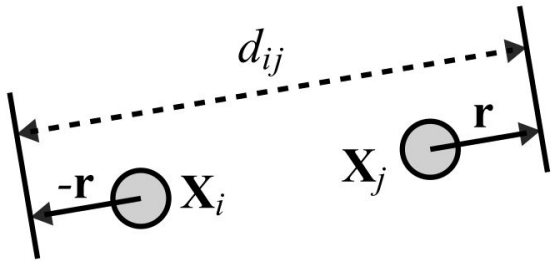
- The first node X_i of a pair is a uniform path step pick from all nodes.

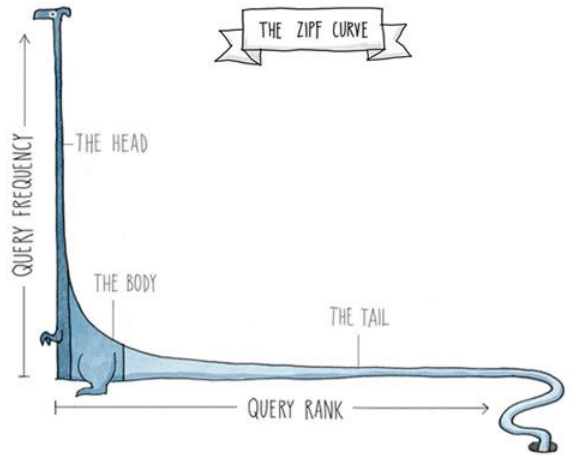
1D Graph Sorting by P-SGD - The Algorithm Explained

Objective: Move a single pair of nodes at a time.

Optimizing the disparity between the layout distance of a node pair and the actual nucleotide distance of a path traversing these nodes.

- The first node X_i of a pair is a uniform path step pick from all nodes.
- The second node X_j of a pair is sampled from the same path following a Zipfian distribution.

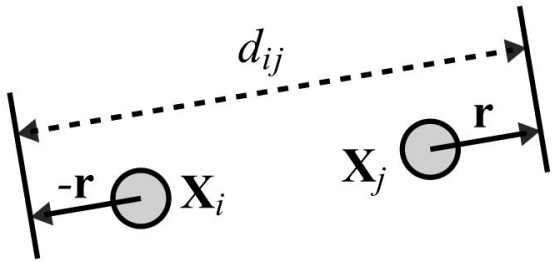




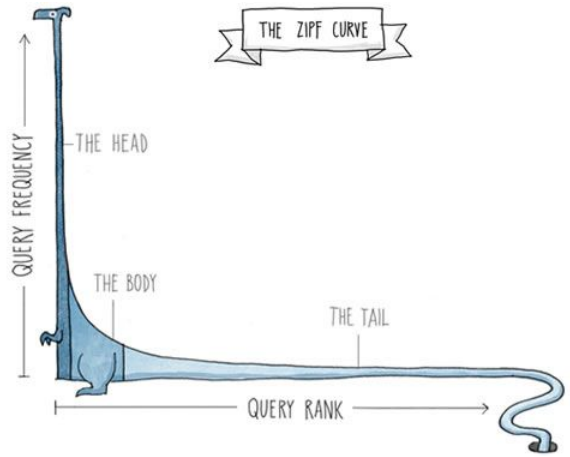
1D Graph Sorting by P-SGD - The Algorithm Explained

Objective: Move a single pair of nodes at a time.

Optimizing the disparity between the layout distance of a node pair and the actual nucleotide distance of a path traversing these nodes.

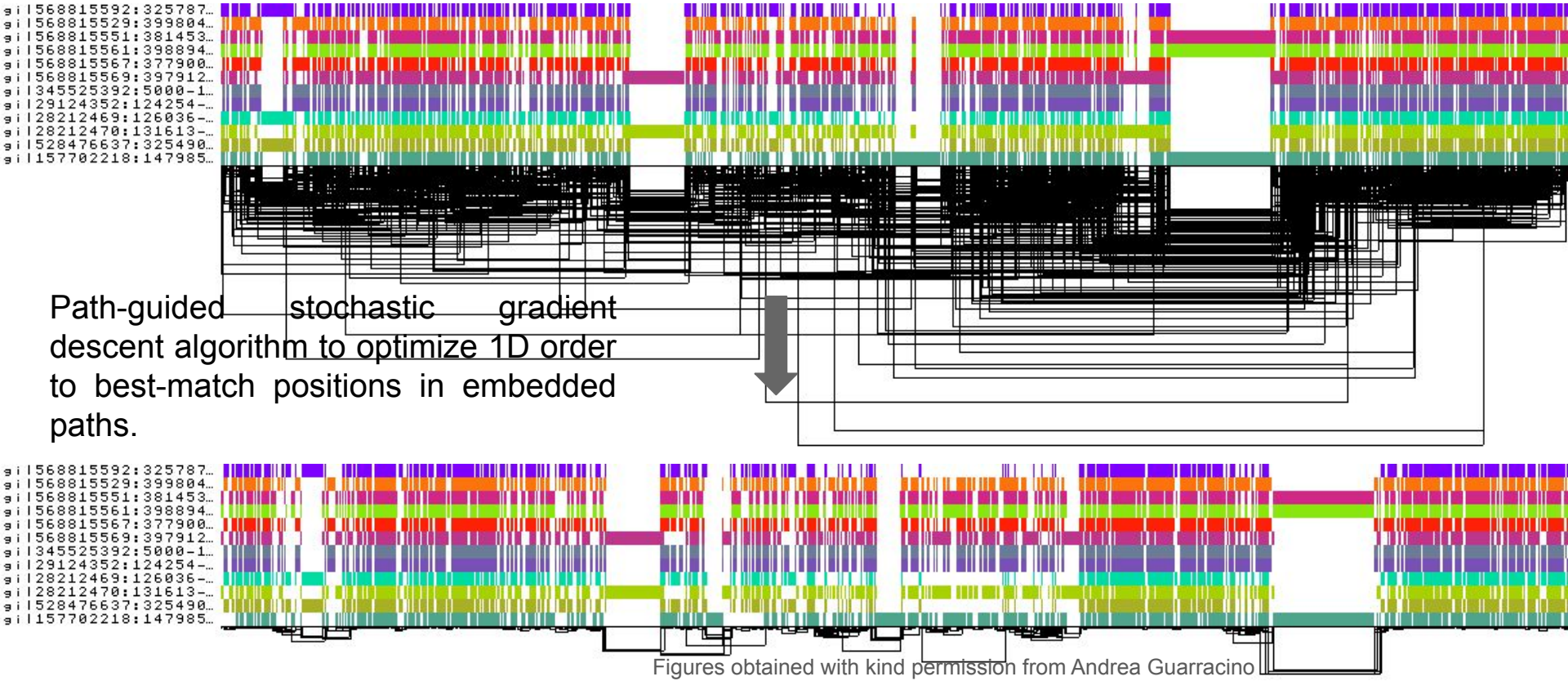


- The first node X_i of a pair is a uniform path step pick from all nodes.
- The second node X_j of a pair is sampled from the same path following a Zipfian distribution.
- The path nucleotide distance of the nodes in the pair guides the actual layout distance d_{ij} update of these nodes. The magnitude r of the update depends on the current learning rate of the SGD.



1D Graph sorting by PG-SGD - Hogwild!

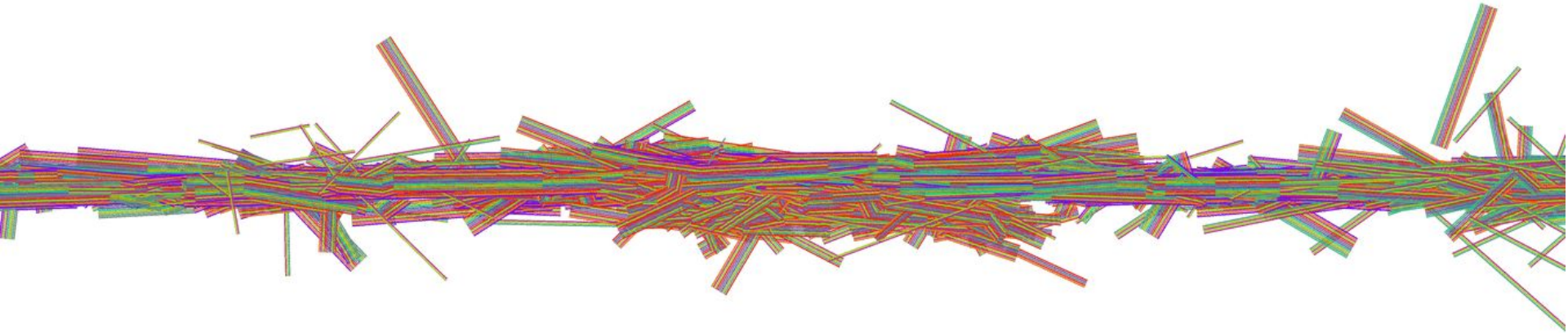
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Figures obtained with kind permission from Andrea Guarracino

Bonus: 2D Graph layout by PG-SGD - Also Hogwild!

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



Path-guided stochastic gradient descent algorithm to optimize 2D layout. Path-labeled rendering with `odgi draw`.

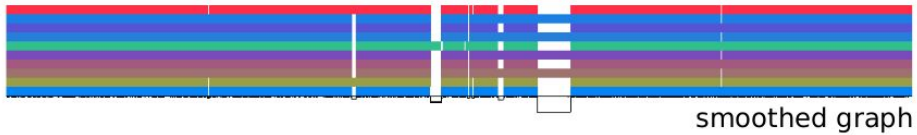
The layout can be plugged into gfaestus for interactive visualization.

Figure obtained with kind permission from Andrea Guarracino

1D PG-SGD implementation is **the key step** in pangenome graph simplification pipeline smoothxg

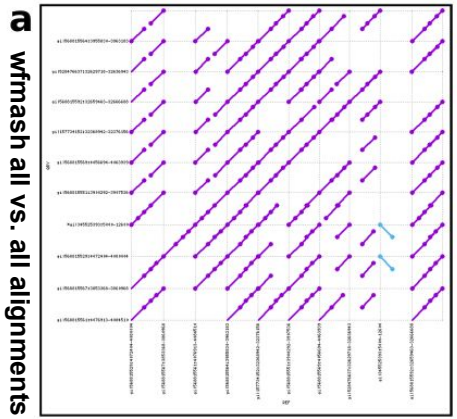
- smoothxg runs Partial Order Alignment (POA) for each block of paths that are collinear within a seqwish induced variation graph.
- A prerequisite is that the graph nodes are sorted according to their occurrence in the graph's embedded paths
- Our 1D path-guided SGD algorithm is designed to provide this kind of sort.

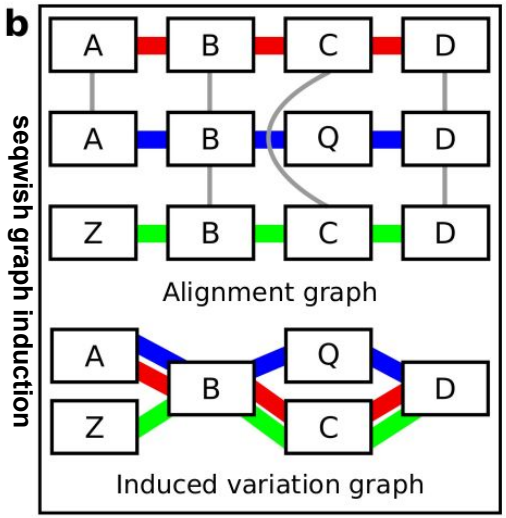
PanGenome Graph Builder (PGGB)

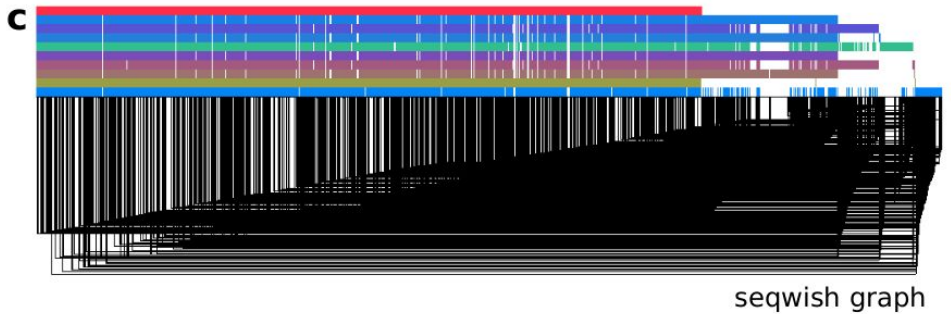


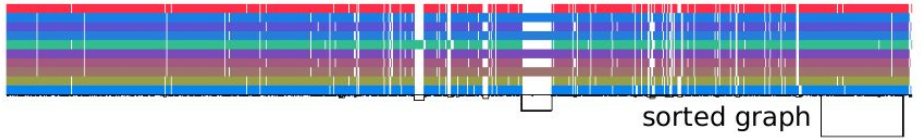
Figures obtained with kind permission from Erik Garrison

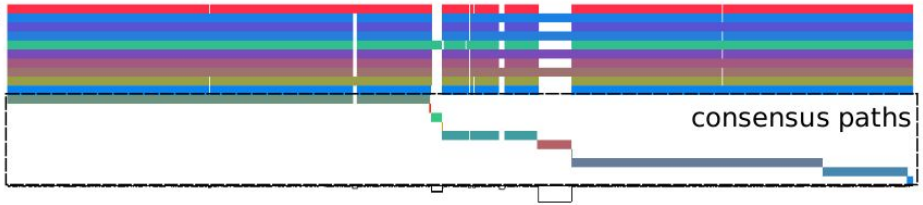
PanGenome Graph Builder (PGGB)

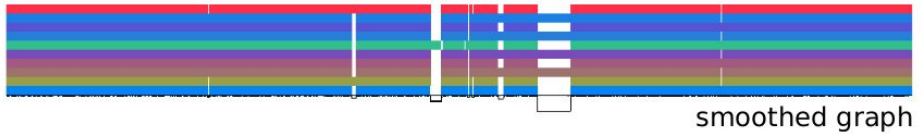












Figures obtained with kind permission from Erik Garrison

nf-core/pangenome

```
[heumos@savasana pangenome]$ pggb -i data/HLA/DRB1-3123.fa.gz -N -w 50000 -s 500
0 -I 0 -p 80 -n 10 -k 8 -t 16 -v -L -o out
```

```
[heumos@savasana pangenome]$ nextflow run main.nf -profile docker -with-docker h
eumos/pangenome:latest --input ~/git/pggb/data/HLA/DRB1-3123.fa.gz --alignment_n
o_splits --smoothxg_max_block_weight 50000 --alignment_segment_length 5000 --smo
othxg_block_id_min 0 --alignment_map_pct_id 80 --alignment_n_secondary 10 --seqw
ish_min_match_length 8 --do_viz --outdir potato_out/ --wfmash --max-cpus 16 --ma
x-memory 24G --file_name_prefix "pggb" -c savasana.config
```

```
[heumos@savasana pangenome]$ ls potato_out/
alignment     multiqc_report.html  odgi_stats  pipeline_info  smoothxg
multiqc_data  odgi_build           odgi_viz    seqwish
```

Parameters

- Input/output options
- Alignment options
- Seqwish options
- Smoothxg options
  - `--smoothxg_max_block_weight`
  - `--smoothxg_max_path_jump`
  - `--smoothxg_max_edge_jump`
  - `--smoothxg_max_poa_length`
  - `--smoothxg_consensus_spec`
  - `--smoothxg_block_id_min`
  - `--smoothxg_ratio_contain`
  - `--smoothxg_poa_params`
- Visualization options

nf-core/ pangenome

command

```
nextflow run nf-core/pangenome -r
```

clones in last 11 months

3016

stars	watchers
12	2
last release	last updated
	2 months ago
open issues	pull requests
12	42

collaborators

nf-core/pangenome - MultiQC

MultiQC

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

Report generated on 2021-10-18, 11:56 based on data in: `/home/heumos/pggb/DRB1`

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 6/6 rows and 11/14 columns.

Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N	% GC
cons@10000__y_0_1000000	13 180	1	0	1	1	3 944	3 073	3 539	2 624	0	43.2%
cons@1000__y_0_1000000	13 180	1	0	1	1	3 944	3 073	3 539	2 624	0	43.2%
cons@100__y_0_1000000	14 125	14	18	4	1	3 944	3 073	3 540	2 624	944	40.3%
cons@10__y_0_1000000	14 260	44	60	10	1	3 980	3 115	3 566	2 655	944	40.5%
seqwish	37 415	1 192	1 596	12	1	11 031	8 124	9 769	7 547	944	41.9%
smooth	22 517	4 860	6 649	13	1	6 422	4 932	5 659	4 560	944	42.2%

nf-core/pangenome - Future scaling up

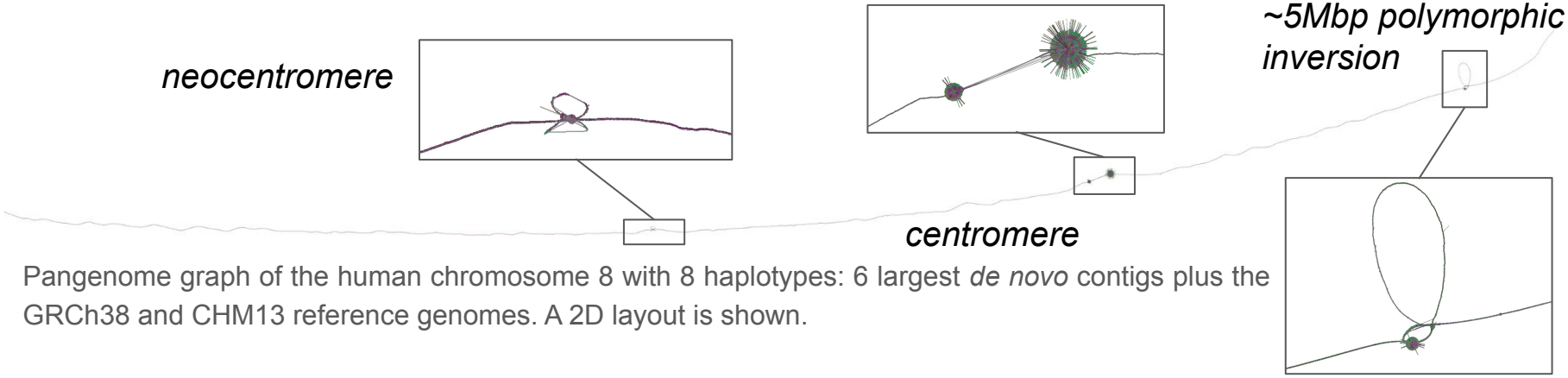
wfmash: Split all versus all run into one versus all wfmash runs so each becomes one execution job.

seqwish: Generate seqwish runs by steering the number of bp to use for transitive closure batch and pack each batch into one execution job.

smoothxg: Each abPOA alignment becomes one execution job.

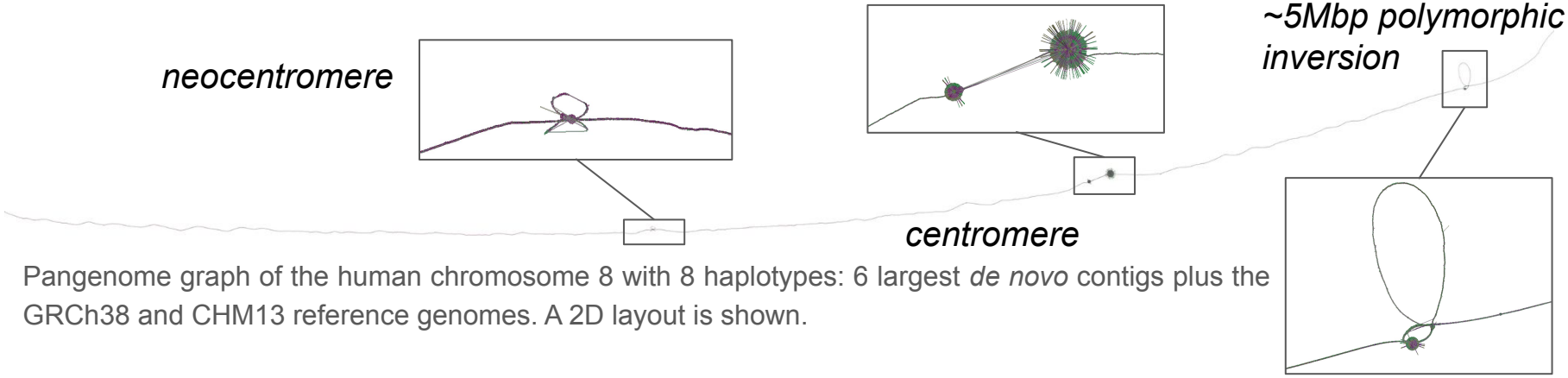
Improvement: Pangenome building of one chromosome can scale across a whole cluster and not just one node.

A PGGB human chr8 pangenome graph



Pangenome graph of the human chromosome 8 with 8 haplotypes: 6 largest *de novo* contigs plus the GRCh38 and CHM13 reference genomes. A 2D layout is shown.

A PGGB human chr8 pangenome graph



Pangenome graph of the human chromosome 8 with 8 haplotypes: 6 largest *de novo* contigs plus the GRCh38 and CHM13 reference genomes. A 2D layout is shown.

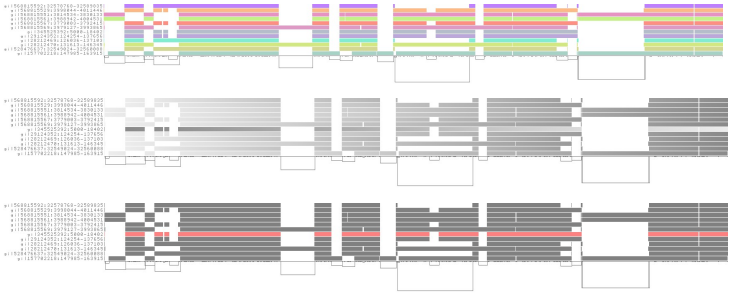
PGGB and nf-core/pangenome:

- Building a pangenome graph from all versus all alignments.
- No reference bias!
- All versus all relationship in one, compressed data structure.
- We can reconstruct the centromeres and **ultra deep complex regions**!

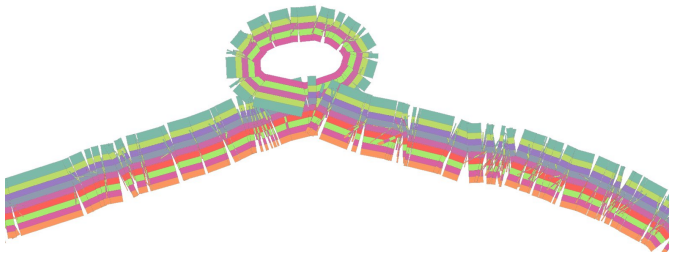
Figure obtained with kind permission from Erik Garrison

Our solution: a new suite of tools for pangenome graphs

To overcome these problems, we have developed an **Optimized Dynamic Genome/Graph Implementation** (ODGI), a new suite of tools to work with pangenome graphs structured in the variation graph model.



- ODGI supports GFA version 1 (GFAv1)
- The majority of ODGI's tools are index-free
- Path manipulation in parallel



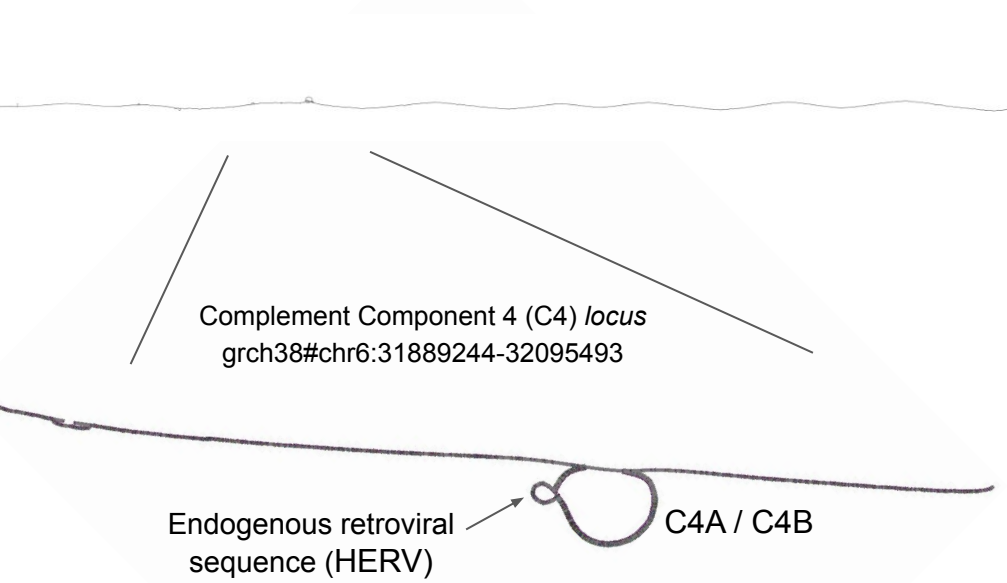
ODGI offers more than 30 tools for graph interrogation, manipulation, and visualization.

Figures obtained with kind permission from Andrea Guarracino

Dissecting pangenome graphs - `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.



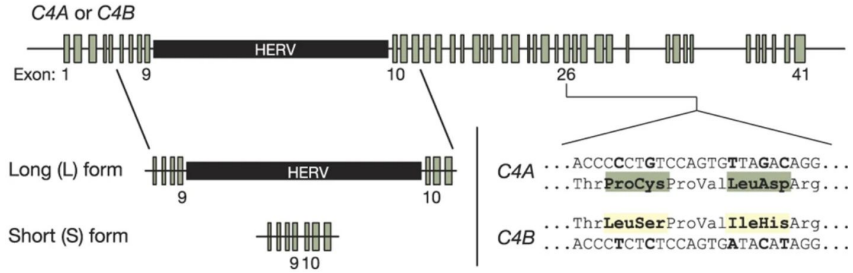


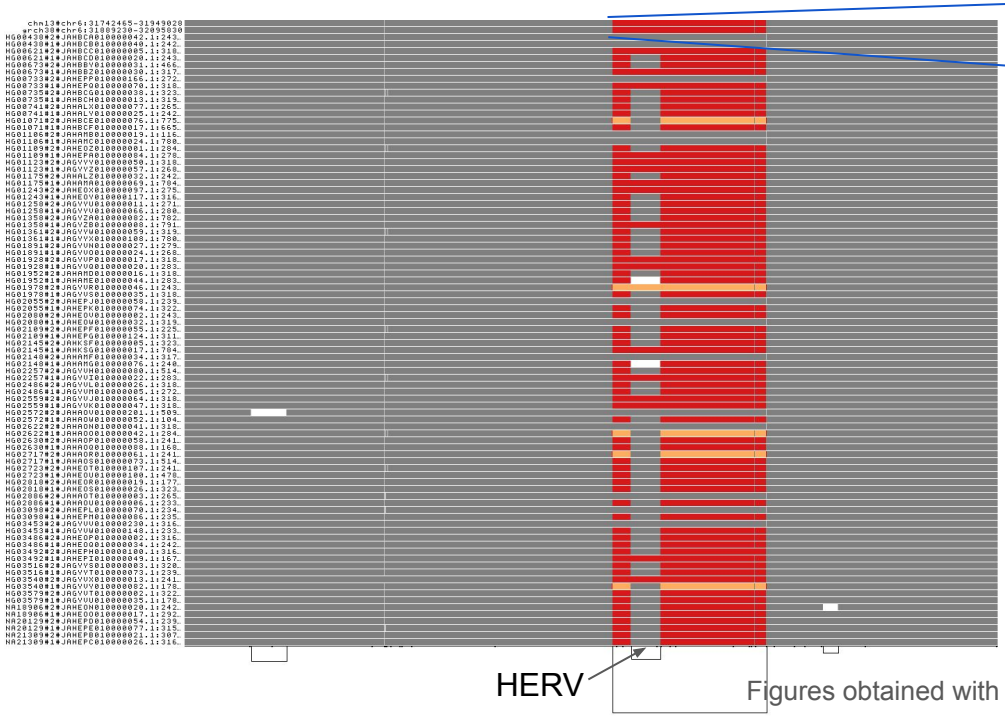
Figure from Sekar et al., 2016.

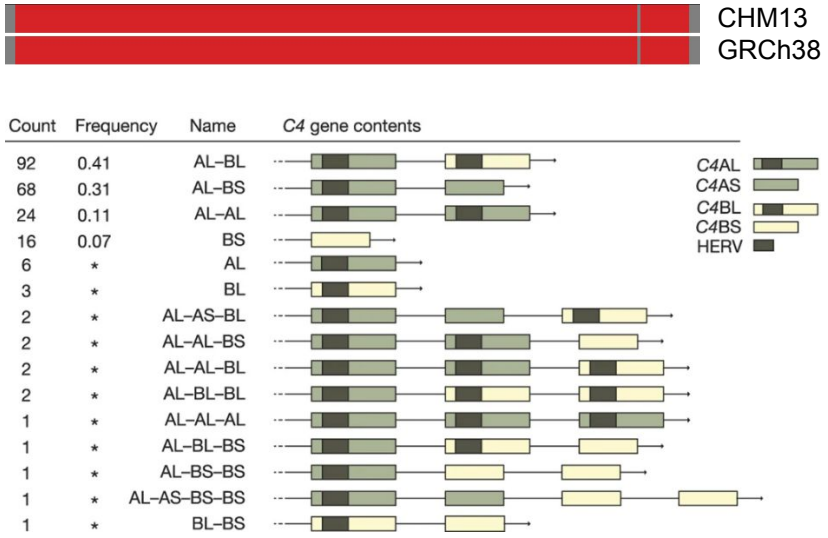
Figures obtained with kind permission from Froggy

Dissecting pangenome graphs - `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)



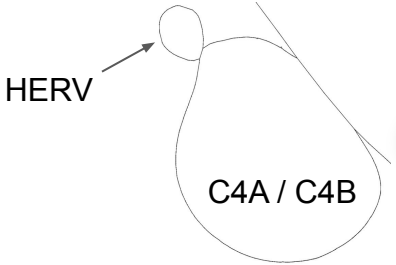


222 chromosomes from 111 genetically unrelated individuals. Figure from Sekar et al., 2016.

Figures obtained with kind permission from Andrea Guarracino

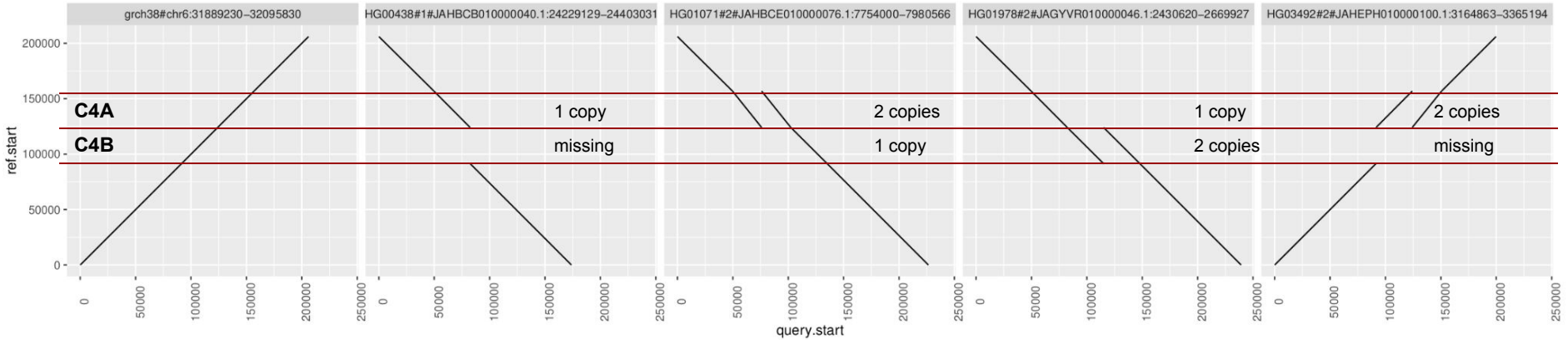
Untangling pangenome graphs - `odgi untangle`

Repetitive sequences produce collapsed repeats in the pangenome graphs.



Paired BED output

query.name	query.start	query.end	ref.name	ref.start	ref.end	score	inv	self.cov	nth.best
HG03492#2#JAHEPH010000100.1:3164863-3365194	0	91469	chm13#chr6:31742465-31949028	0	91441	0.998121	+	1	1
HG03492#2#JAHEPH010000100.1:3164863-3365194	91469	124145	chm13#chr6:31742465-31949028	124001	156853	0.99149	+	1.80362	1
HG03492#2#JAHEPH010000100.1:3164863-3365194	124145	150633	chm13#chr6:31742465-31949028	124001	156853	0.802825	+	1.99185	1
HG03492#2#JAHEPH010000100.1:3164863-3365194	150633	199828	chm13#chr6:31742465-31949028	156853	206060	0.997848	+	1.00026	1



Haplotypes representing the most frequent configurations found at the C4 *locus* in the HPRC dataset.

Figures obtained with kind permission from Andrea Guarracino

Annotating pangenome graphs - `odgi position`

Input: Chr8 human **consensus graph** originating from 44 *de novo* assemblies from the HPRC. Both references CHM13 and GRCh38 are fully preserved in the graph.

Tool: `odgi position` - Annotation lift over from an annotated reference path in the graph to the nodes in the graph via e.g. **BED**

Output: TSV file with the gene annotation per node for e.g. visualization.

Chr8 consensus graph - cytobands annotation

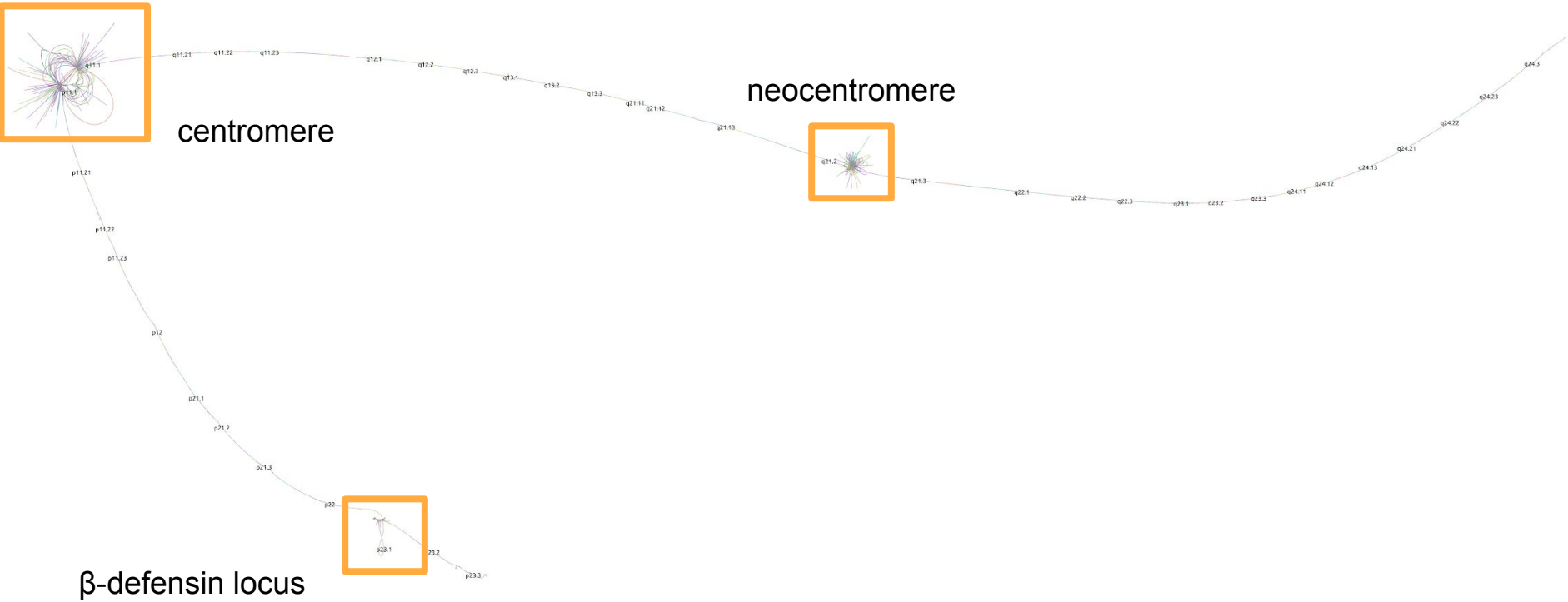
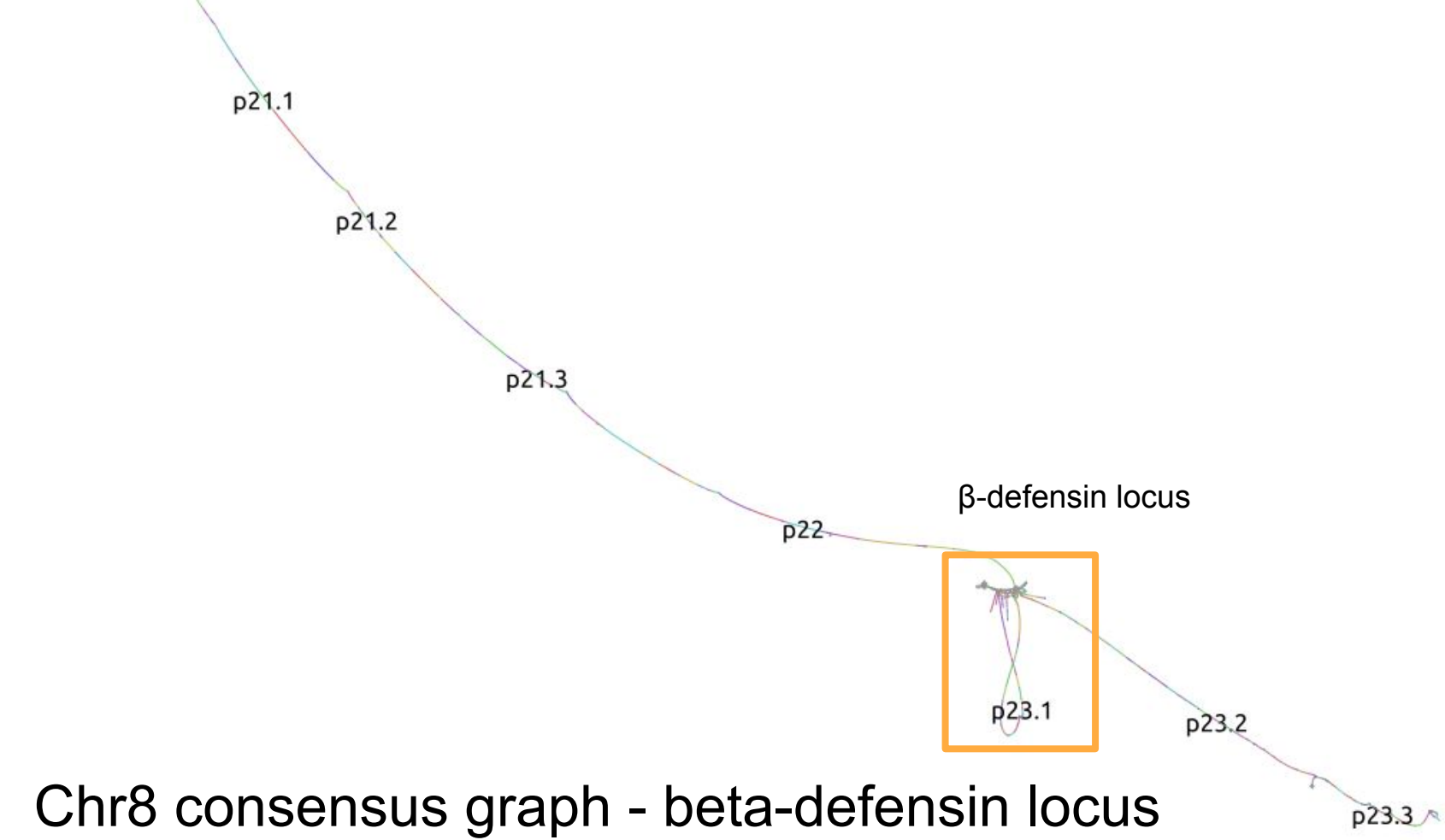


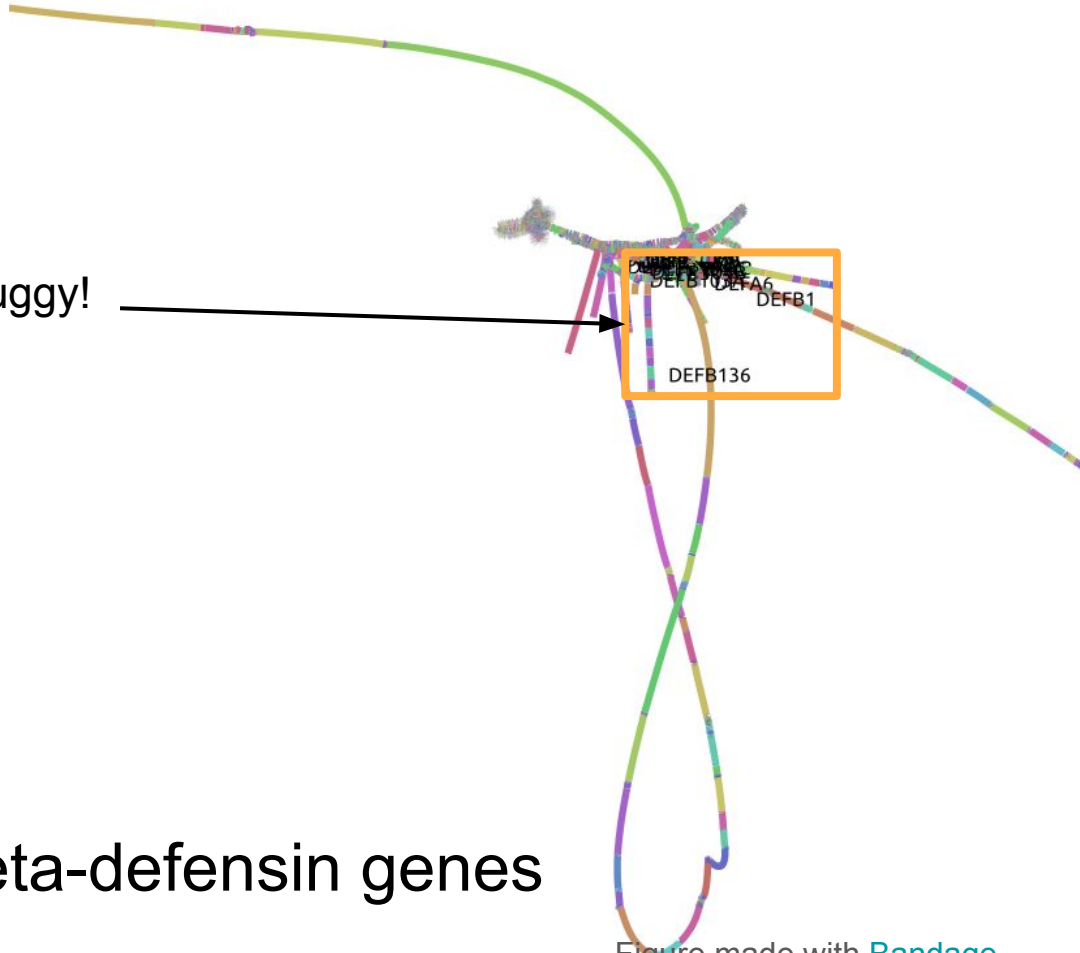
Figure made with Bandage.



Chr8 consensus graph - beta-defensin locus

Figure made with Bandage.

This graph must be buggy!



Chr8 consensus graph - beta-defensin genes

Figure made with Bandage.

Identifying assembly breakpoints relative to the references

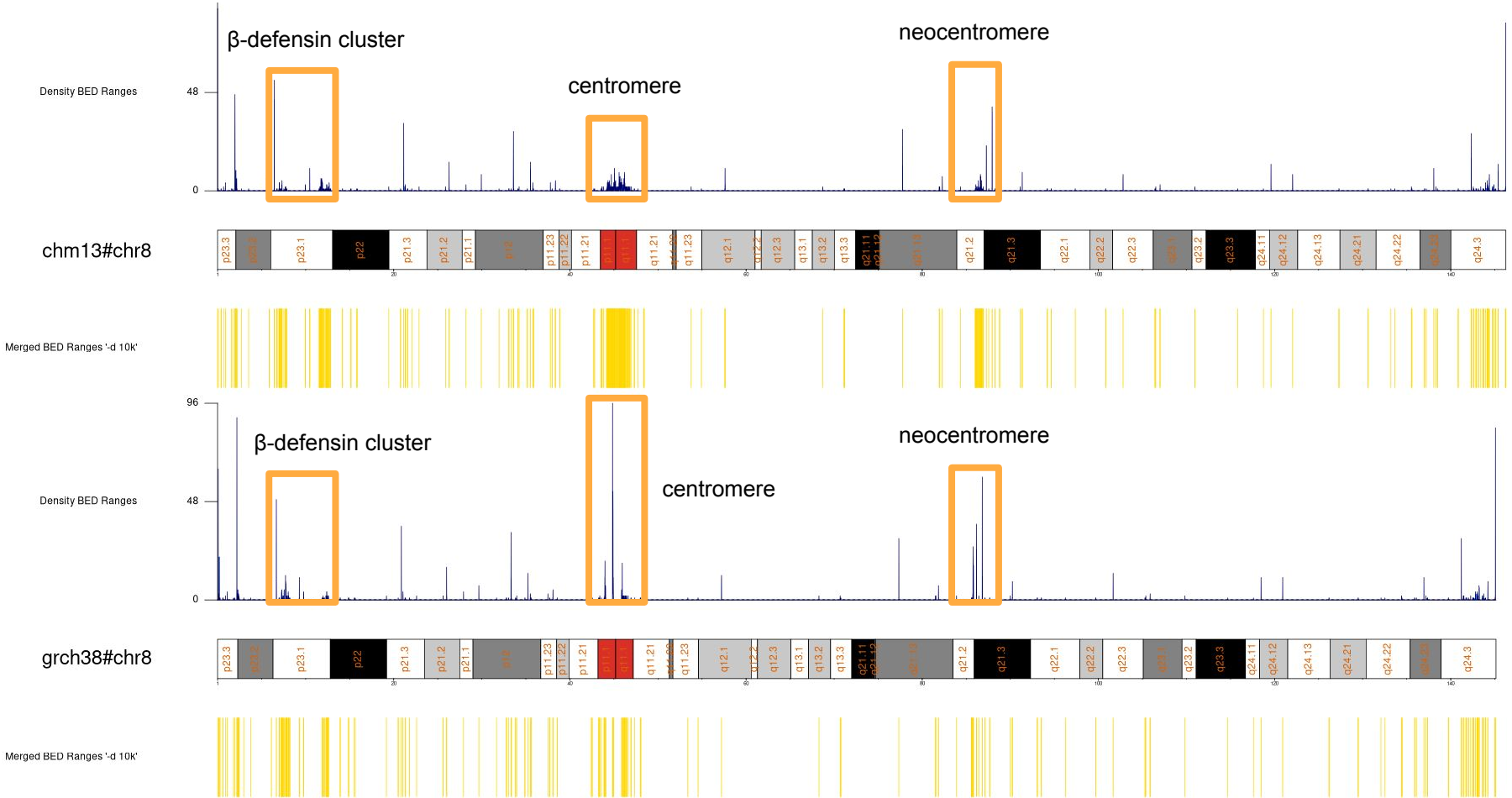
Where do our contigs' ends match the references? - Detecting regions that are difficult to assemble.

Input: Chr8 human pangenome graph made with contigs from 44 *de novo* assemblies from the HPRC adding CHM13 and GRCh38 - 90 haplotypes.

Tool: `odgi tips` - Walking from the ends of a contig until a reference node is found. For each contig range (e.g. a `tip`) we look at each possible reference window and find the most-similar one.

Output: BED file with the best reference hit and position for each contigs' ends.

Assembly Breakpoint Ranges of the Contigs in chr8 in the HPRC PGGB RC1 Graph relative to CHM13 and GRCh38



Discussion

ODGI repository

ODGI documentation

ODGI: State-of-the-art tool box to transform, analyse, simplify, validate, and visualize pangenome graphs at large scale.

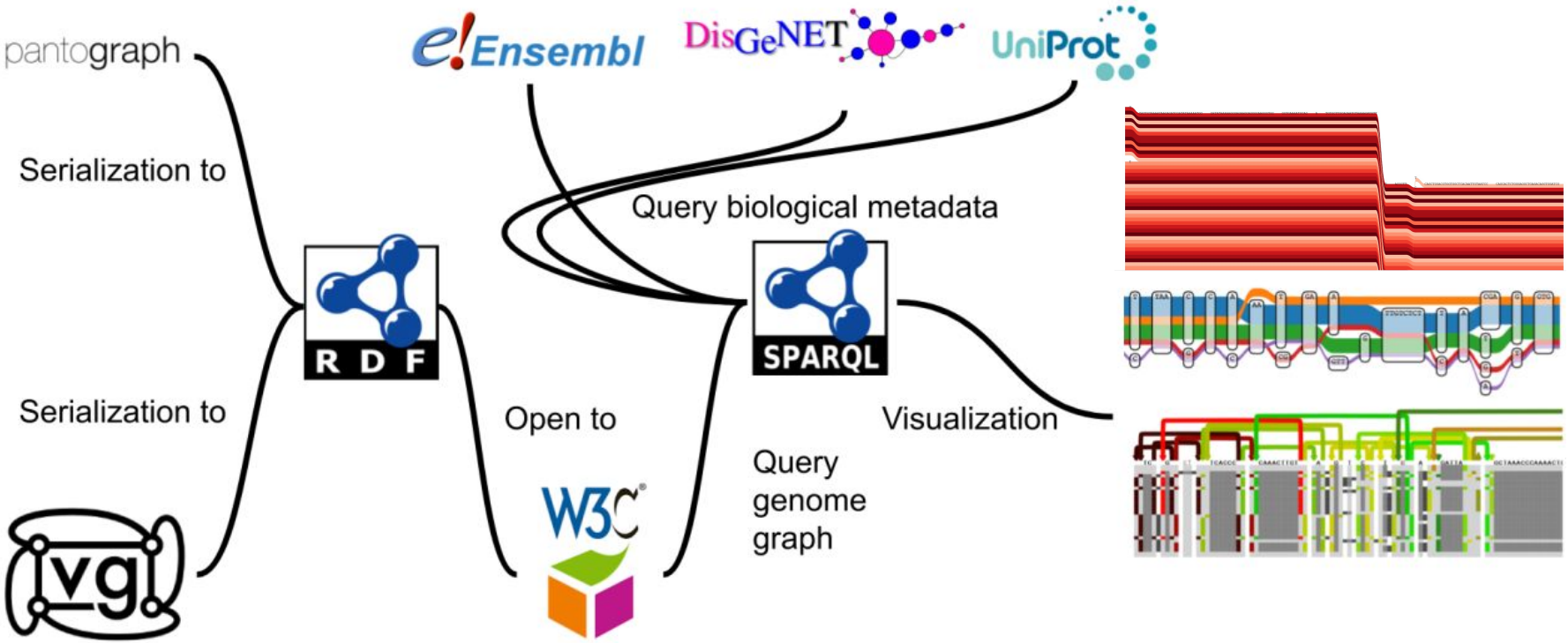
Bridge between linear reference genome analysis and pangenome graphs: Subgraph extraction, lifting over annotations, linearizing nested graph structures

Discover the underlying biology of pangenome graphs:
Detect complex regions, identify assembly breakpoints

The tools already are the backbone of pipelines such as the Pangenome Graph Builder (PGGB) or nf-core/pangenome.

Future work: RNA and protein support, expand metadata capabilities

The overview of Pangenome ontologies



http://biohackathon.org/resource/vg

Resource Description Framewok (RDF)

- **IRI**
  Synonym: URI is often an URL

- **Literal**
  “Strings”, 1, 1.001

- **Bnode**
  Placeholder identifiers: Something exists, but you don’t know its identity

- A **Triple** is a statement
  Subject → Predicate → Object

Resource Description Framework (RDF)

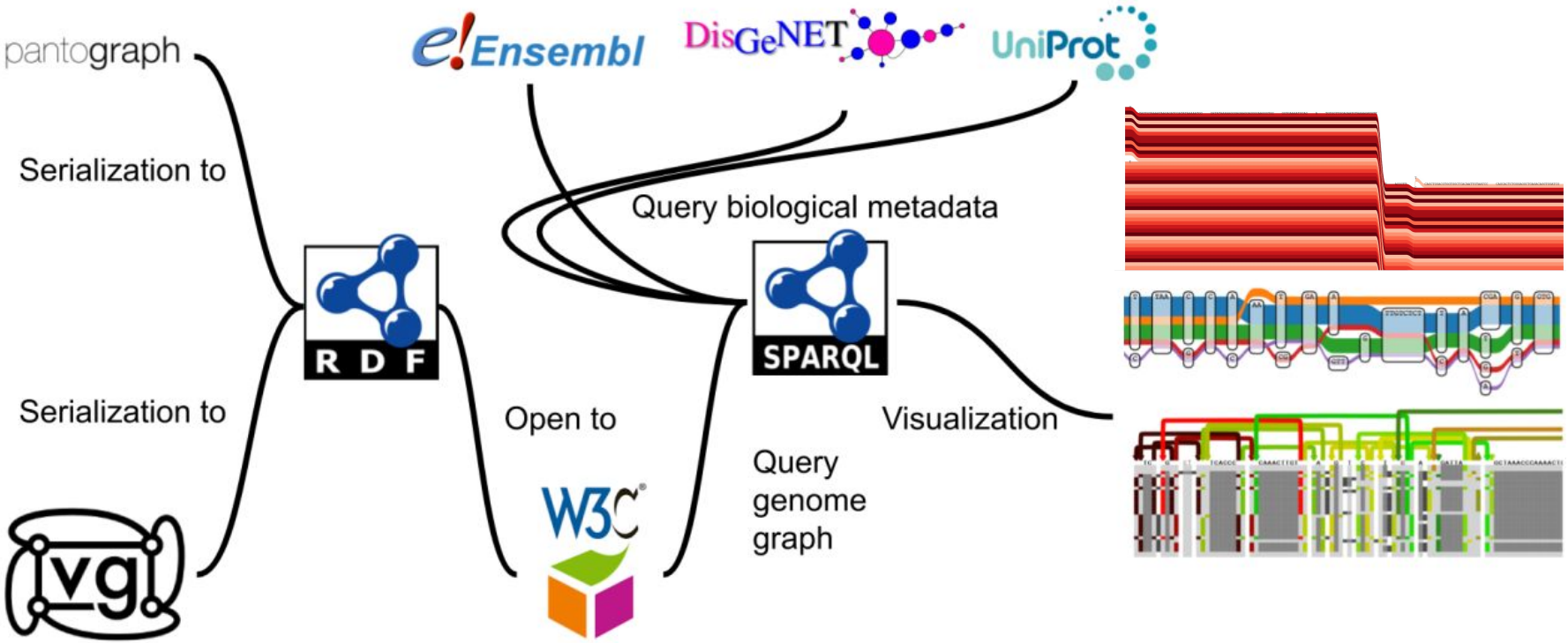
- A **Triple** is a statement
  Subject → Predicate → Object

  I → love → my cat

<https://uni-tuebingen.de/forschung/forschungsinfrastruktur/zentrum-fuer-quantitative-biologie-qbic/team/simon-heumos/> ↷

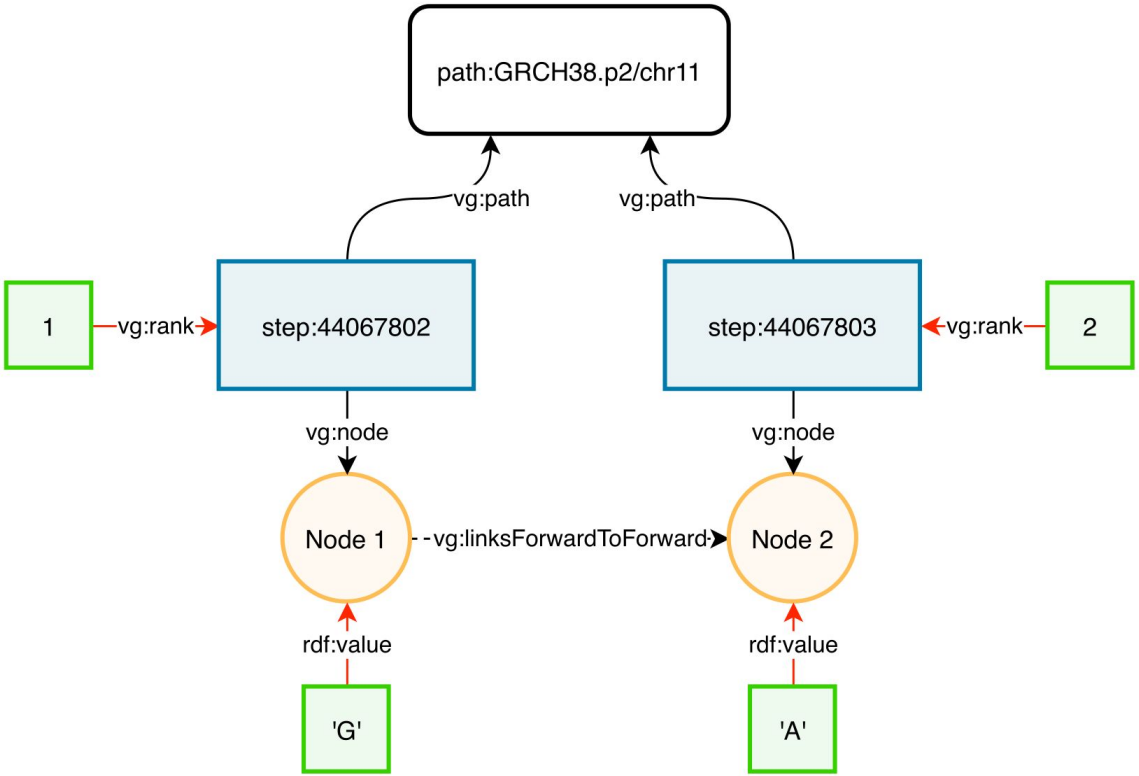
<https://www.dictionary.com/browse/love>↷

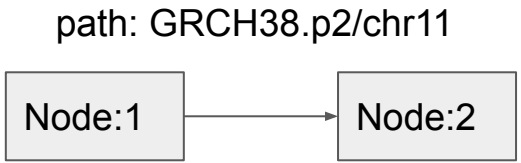
The overview of Pangenome ontologies



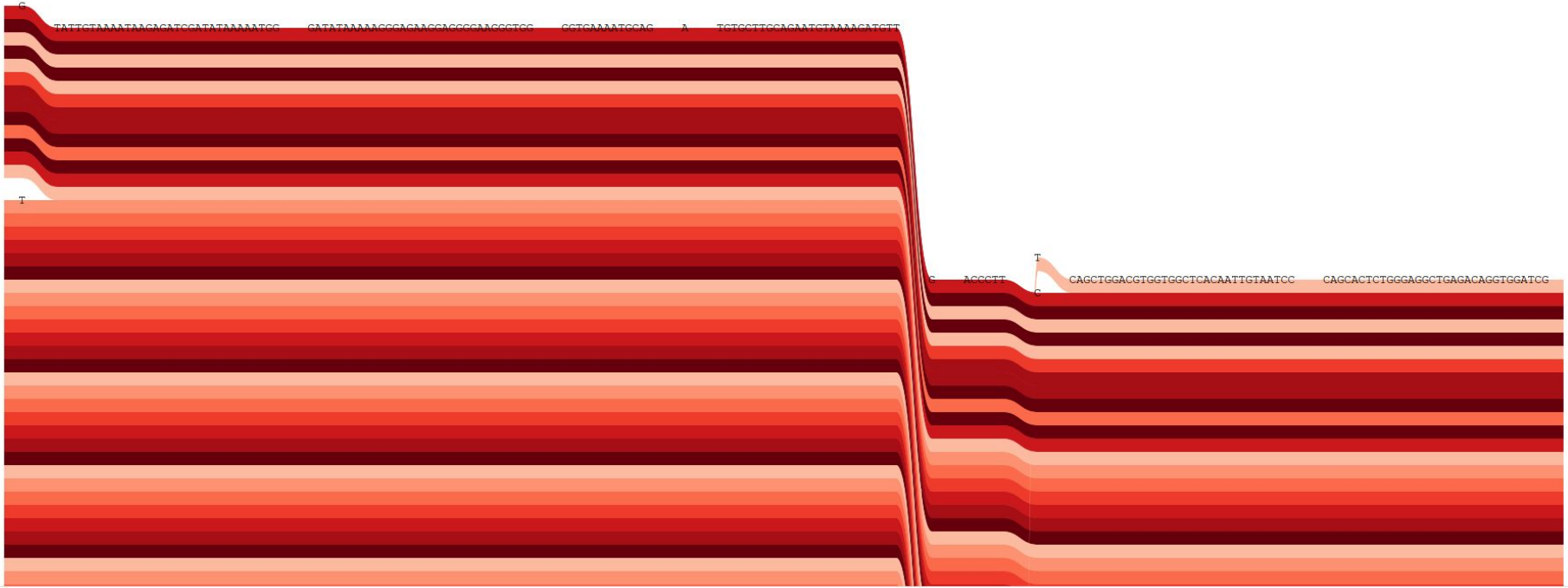
http://biohackathon.org/resource/vg

Variation graph ontology



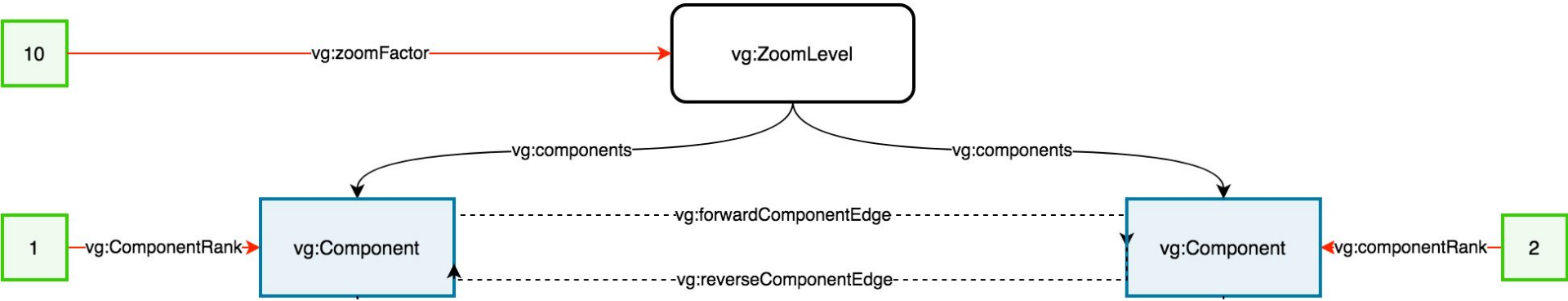


SPARQL Querying a Semantic Genome Graph Japan BH 2019 - Example Visualization[1]

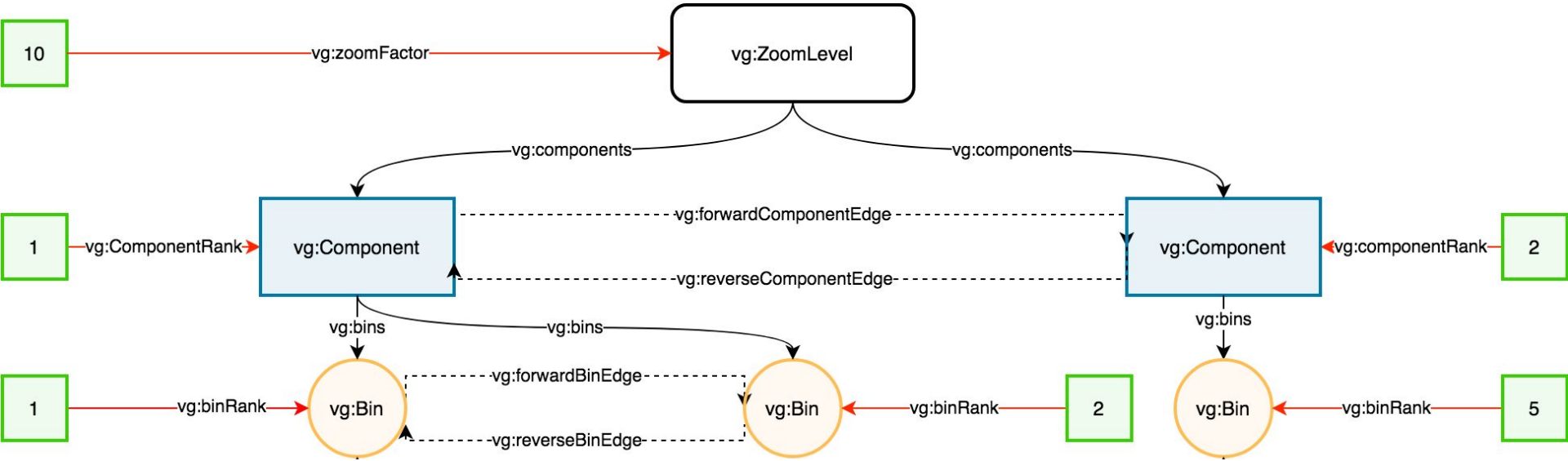


[1]https://twitter.com/simonheumos/status/1169884828860239874

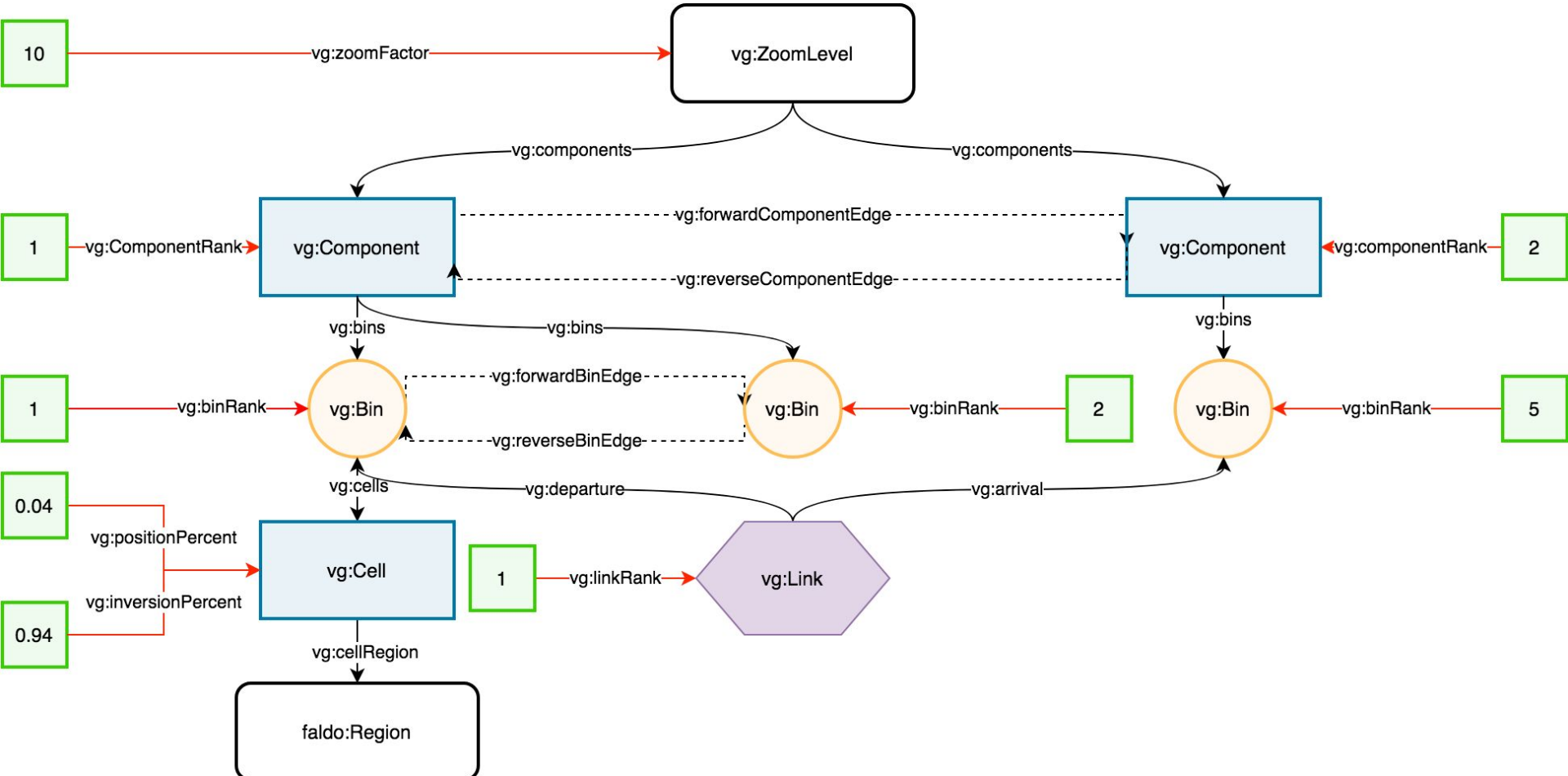
Each ZoomLevel has multiple Components



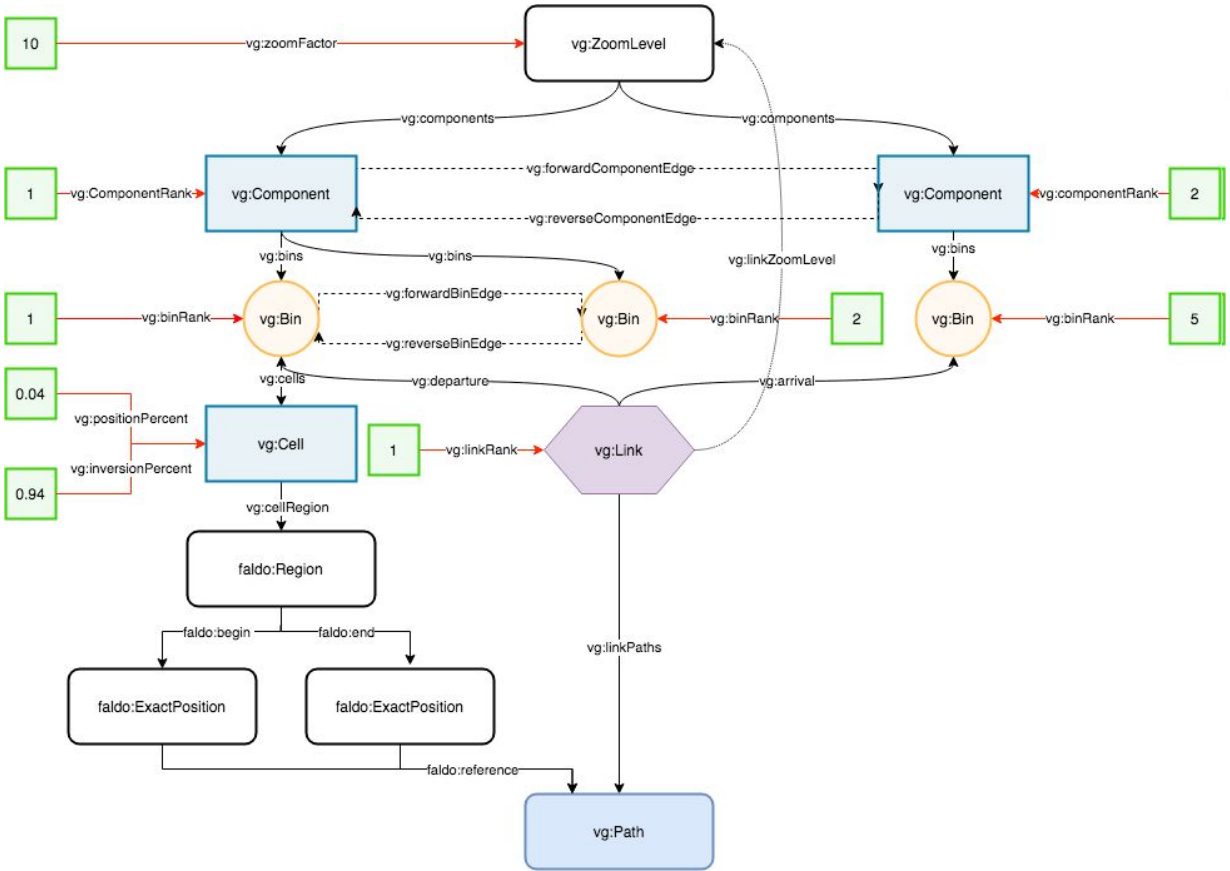
Each Component has multiple Bins

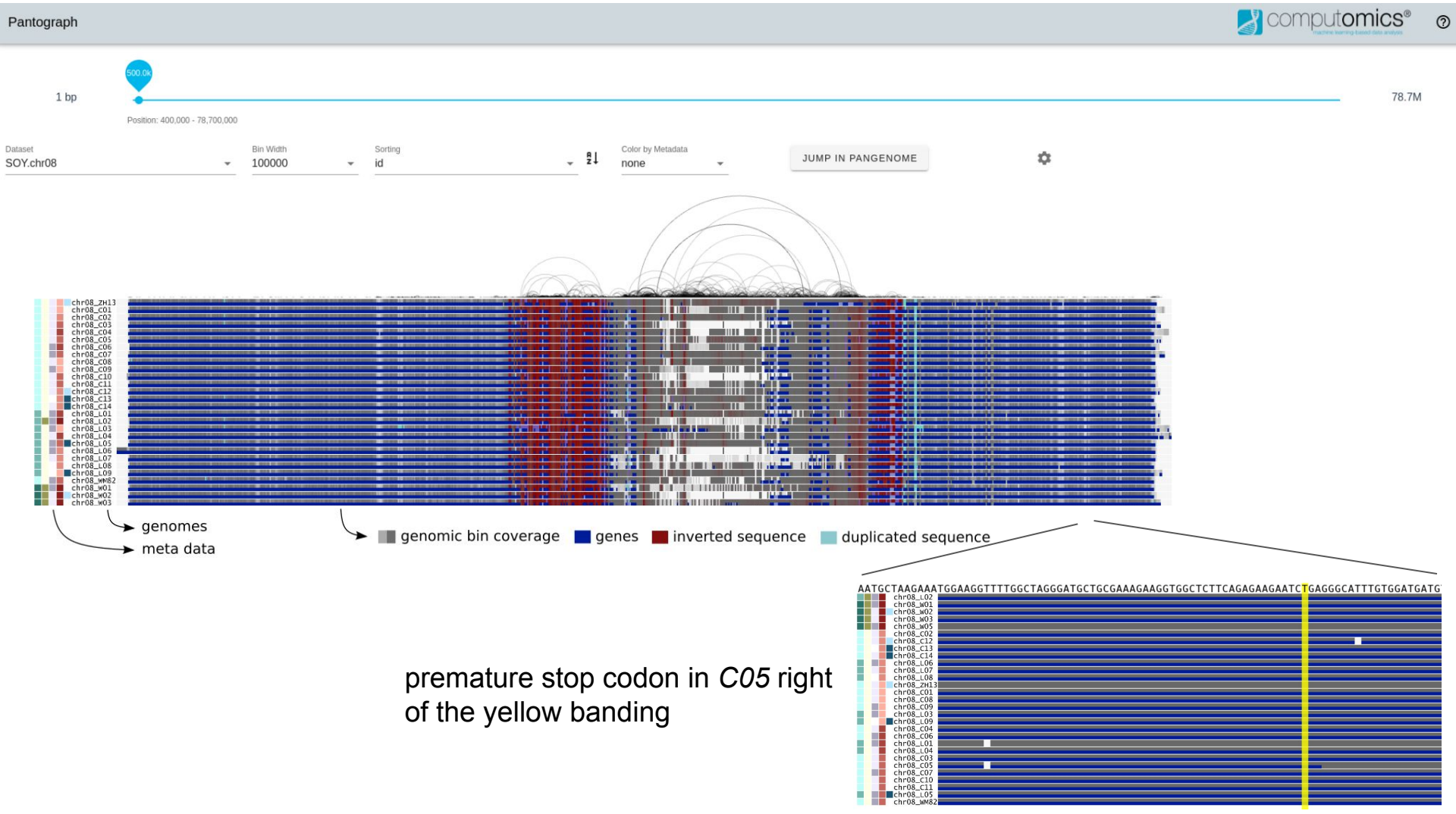


Each Bin is connected by Link, and has Cells



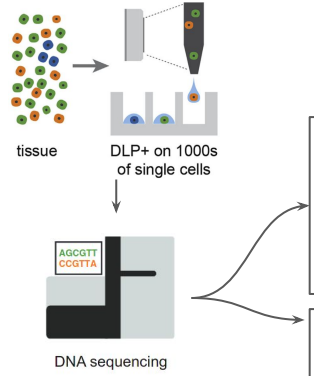
vg ontology is referenced via Link and Cell





premature stop codon in *C05* right of the yellow banding

Brain exercise: Building pangenome graphs from cancer scWGS data



>50k Cells

Figures from Laks et al. 2020.

Reference Guided Assembly!

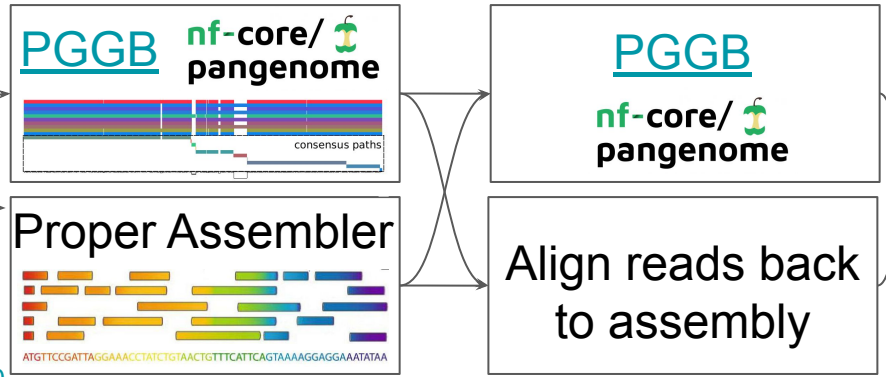


Figure from Seeman et al. 2016.

- PCA
- Variant Calling
- Visualization
- Cancer Genome
- Healthy Genome

Acknowledgments

EU Pangenome Group

Erik Garrison

Andrea Guarracino

Pjotr Prins

Vincenza Colonna

Flavia Villani

David G. Ashbrook

Robert W.Williams

Christian Fischer

HPRC

Sven Nahnsen

Gisela Gabernet

Michael Krone

Oliver Kohlbacher

Antonia Schuster

Friederike Hanssen

HelmholtzZentrum münchen
German Research Center for Environmental Health

Lukas Heumos

Philipp Ehmele

Christian Kubica

Sebastian Vorbrugg

Jörg Hagmann

Jerven Bollemann

Toshiyuki T. Yokoyama

Torsten Pook

BONUS SLIDES

- **Simon:** Detecting complex regions with ODGI depth
- **Christian Kubica:** Evaluation of short read mappers to a pangenome graph with Arabidopsis thaliana
- **Andrea:** Chromosomal communities
- **Flavia Villani:** Chr 19 mouse pangenome graph
- **Jerven:** Zero Extra Costs SPARQLable Pangenome SpOdgi **OR** SapFhir
- **Jörg:** Pantograph Demo
- **Christian Fischer: gfaestus Demo**
- **Simon:** Which contigs travel T2T and are centromere assemblers?

Which contigs travel T2T and are centromere assemblers?

Reminder: odgi tips - Where do our contigs' ends match the reference?

T2T classes:

- **centromere**: One of the contigs' end must match chm13 in the p arm, the other contig's end must match chm13 in the q arm. Both matches are outside of the centromere as given by the chm13 centromere cytobands.
- **centromere_plus_*X***: We artificially extend the size of the centromere by ***X*** nucleotides in both the p and the q arm. The higher ***X***, the more confident we can be that we identified full centromere-assembling contigs!
- **telomere_plus_*X:*** We define a nucleotide length ***X*** which we travel from both chromosome arms. Contigs' ends matching chm13 within these ranges are full T2T contigs!

Detecting complex regions

Human chromosomes have large regions of highly identical repeats:

- Clusters centromeres
- Regions of segmental duplication
- In the acrocentric short arms of chromosomes.

Logsdon et al., Nature 2021: Chr8 carries a modestly sized centromere of approximately 1.5–2.2 Mb, in which AT-rich, 171-base-pair (bp) α-satellite repeats are organized into a well-defined higher-order repeat (HOR) array.

ODGI offers tools to detect and explore such regions.

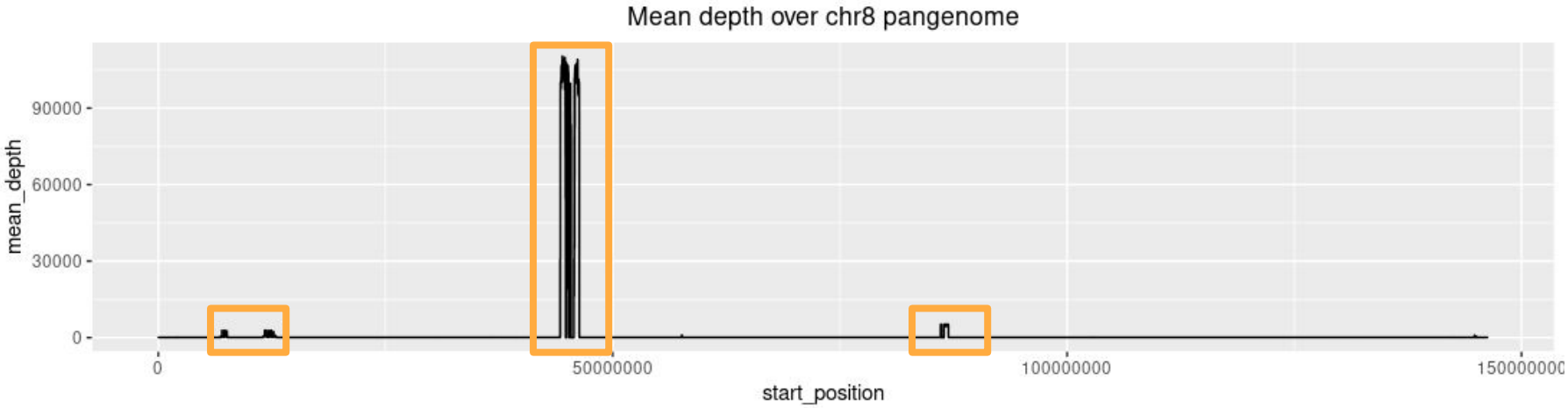
Detecting complex regions - `odgi depth`

Input: Chr8 human pangenome graph made with 44 *de novo* assemblies from the HPRC adding CHM13 and GRCh38 - 90 haplotypes.

Tool: `odgi depth` - calculate node depth: For each node, we record the number of times a node is crossed by all the paths present in the graph.

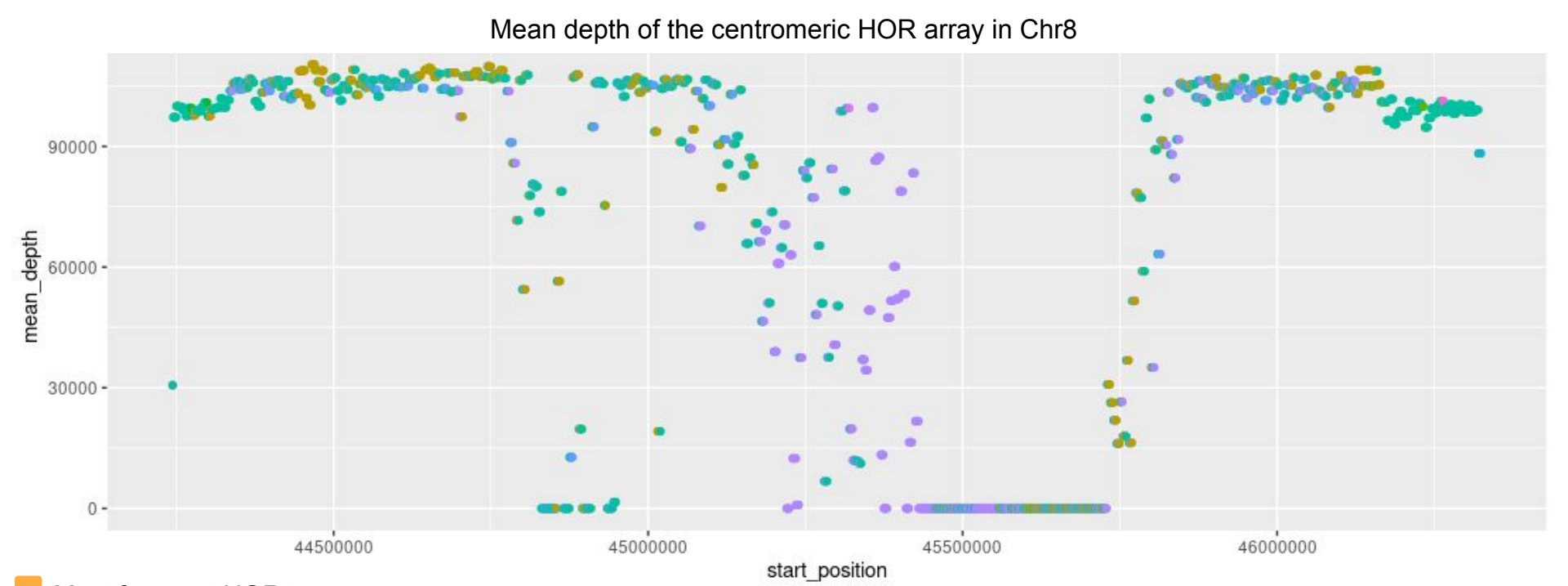
Output: BED file with the mean node depth distribution across windows of the pangenomic CHM13 positions.

Detecting complex regions - Mean depth distribution

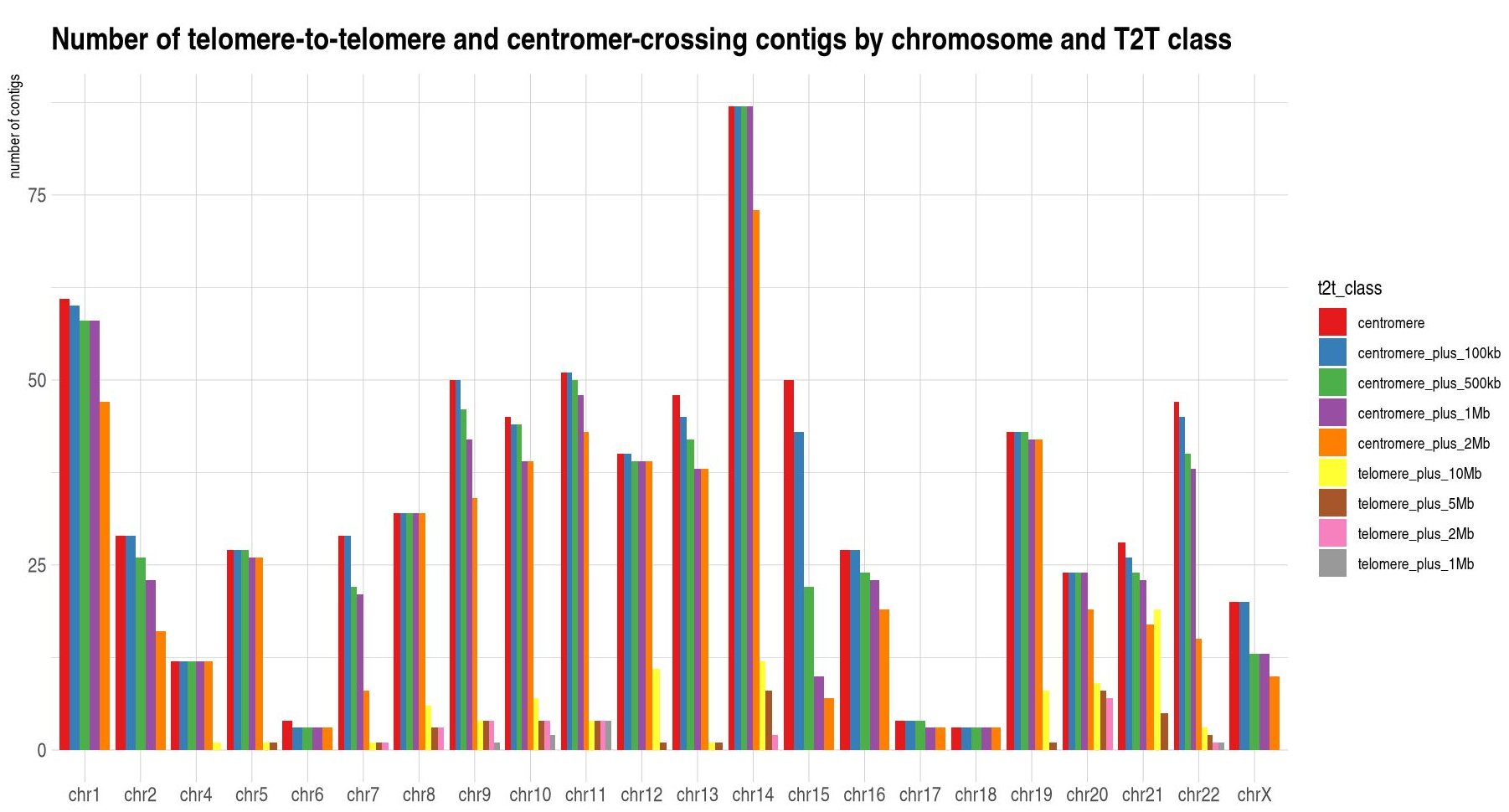


β-defensin locus centromere neocentromere

Chromosome 8 does not only possess a centromeric region, but a complex beta-defensin gene locus and a VNTR that can function as a neocentromere.

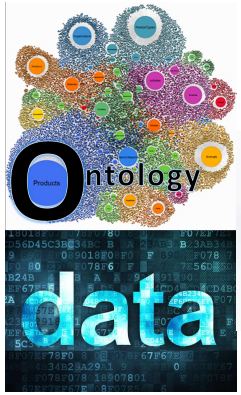


Every letter indicates an alpha-satellite monomer in the HOR: For example A,B,C,D,E,F,G,H,I,J,K would indicate an HOR with 11 alpha-satellite monomers. The mean depth drop falls into the hypomethylated and CENP-A-enriched regions, that have the highest consistent entropy in the entire array. This agrees with the Logsdon et al., Nature 2021 publication.


Number of telomere-to-telomere and centromer-crossing contigs by chromosome and T2T class
number of contigs
0
25
50
75
chr1
chr2
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
t2t_class
centromere
centromere_plus_100kb
centromere_plus_500kb
centromere_plus_1Mb
centromere_plus_2Mb
telomere_plus_10Mb
telomere_plus_5Mb
telomere_plus_2Mb
telomere_plus_1Mb

Zero Extra Costs SPARQLable Pangenome **SpOdgi**

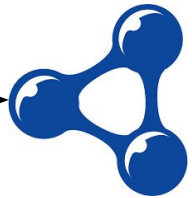
vg Ontology



Variation Graph

SpOdgi

Triples



SPARQL Endpoint

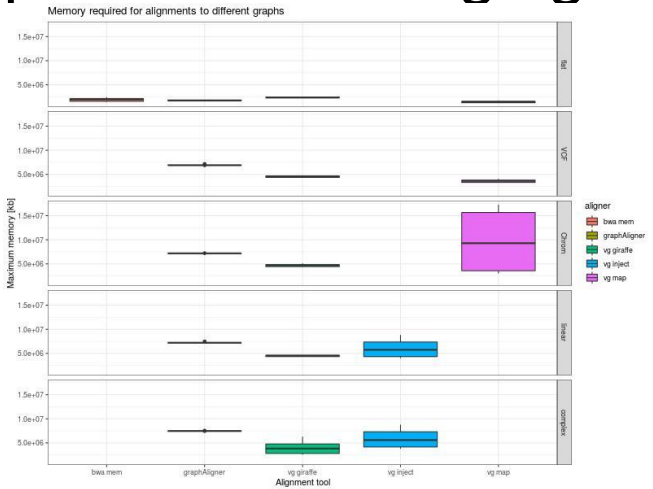
by Jerven Bolleman

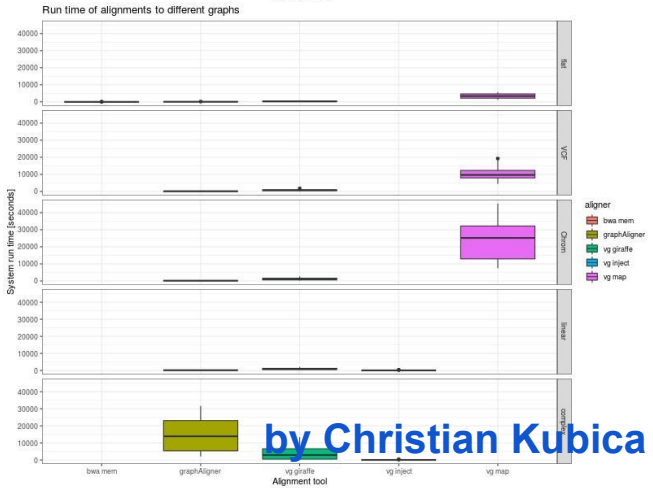
Short read alignments to complex graphs are challenging

Alignments of 12 *A.thaliana* short read sets to graphs of different complexity

- using a set of 12 short read sets (6x N90 100 bp; 6x N90 250 bp)
- Using 5 different alignment tools
  - bwa mem (as control on the flat sequence)
  - graphaligner
  - vg map
  - vg giraffe
  - bwa mem + vg inject (using all genomes concatenated)

Flat graph	VCF graph	Chrom pggb	Linear pggb	Complex pggb
Flat reference genome imported without variation	Variation called from the complex graph used to build a vg graph	Independent chromosome graphs concatenated into one graph.	Graph build from all chromosomes and contigs. Enforcing linearity over compression.	Graph build from all chromosomes and contigs. Trying to compress repeated sequences.
Graph size: 119 mb	Graph size:: 220 mb	Graph size: 197 mb	Graph size:: 195 mb	Graph size: 168 mb
Nodes: 3,829,320 Edges: 3,829,313	Nodes: 5,601,134 Edges: 8,307,979	Nodes: 6,247,306 Edges: 8,491,459	Nodes: 6,282,045 Edges: 8,540,670	Nodes: 6,694,806 Edges: 9,147,110



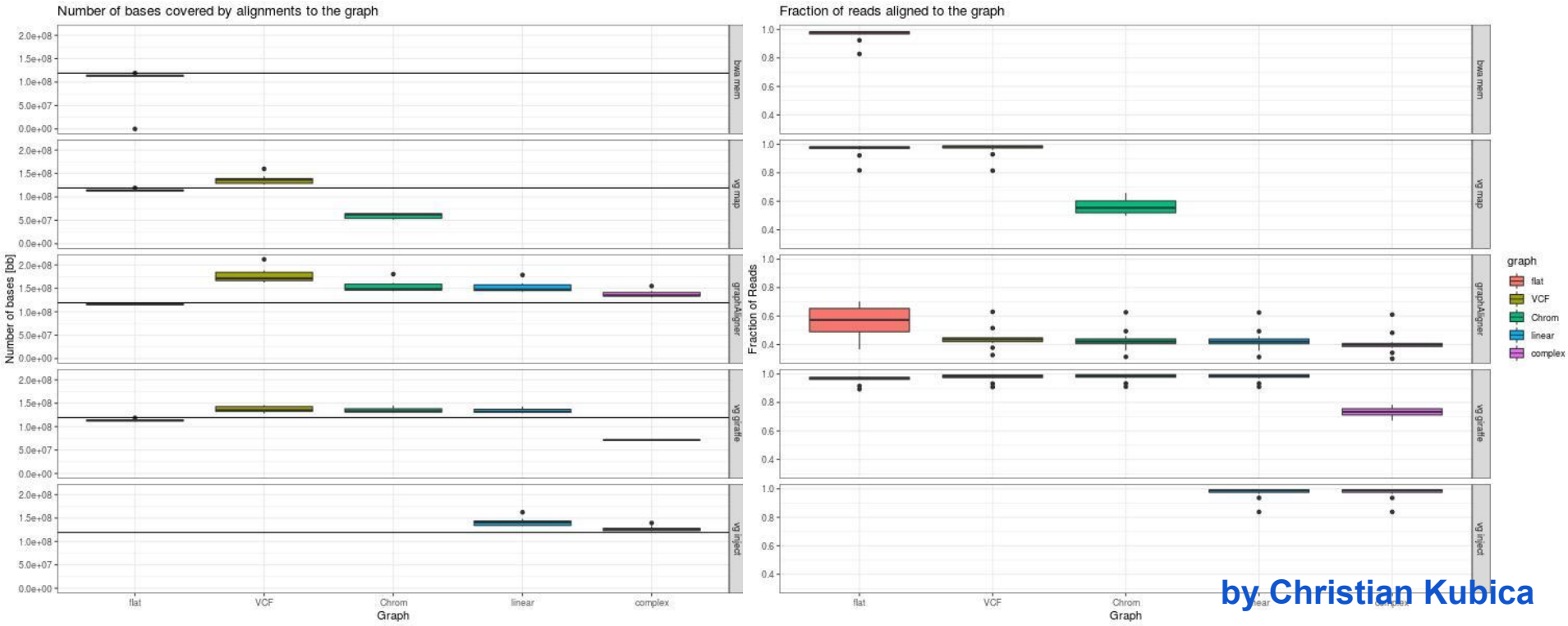


by Christian Kubica

Short read alignments to complex graphs are challenging

vg mappers are not made for WGA derived graphs and lose accuracy. *graphaligner* struggles with short reads.

Current workaround: mapping to flat references and injecting into the graph. Only known variation can be covered



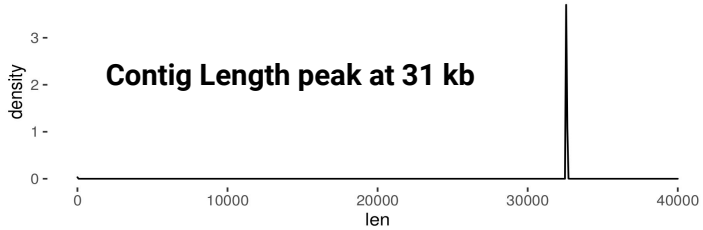
by Christian Kubica

Chromosome 19 mouse pangenome

BXD Family

Ashbrook, D. G., et al. Cell Systems (2021)

148 BXD strains assembled
(10x technology)

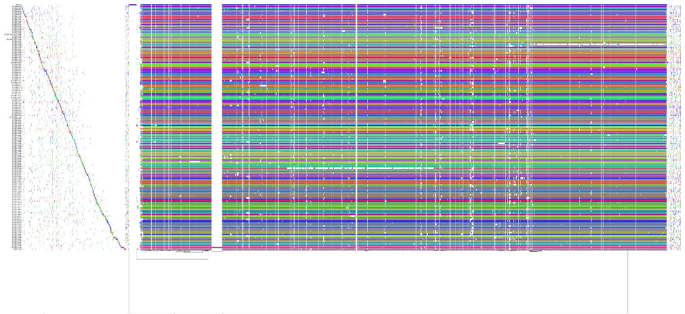


Pangenome of Chr19

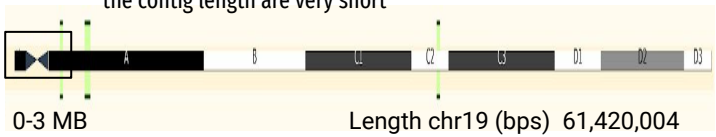
148 BXD strains assembled

Nucleotides from pangenome		
A	87,307,913	14,49 %
C	63,277,613	10,49 %
G	63,151,974	10,47 %
N	**302,016,734**	**50,10 %**
T	87,048,427	14,45 %
tot	602,802,661	

Pangenome made of
5.5M nodes and 8.6M edges
(total length 264,380,676 bp
with 82,695 paths)



Centromeric and telomeric regions are not mapped well because the contig length are very short

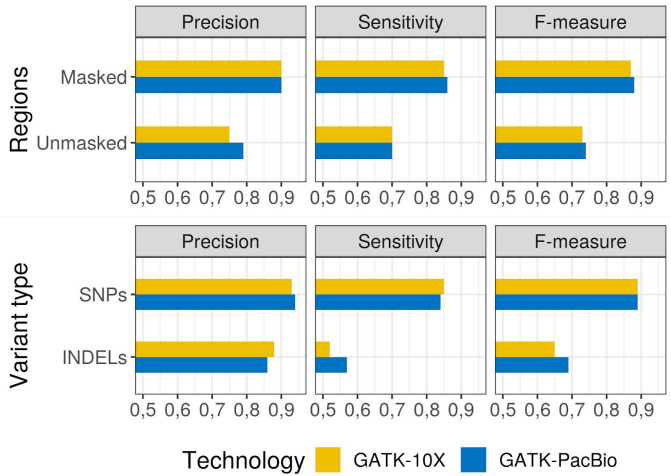


by Flavia Villani

Chromosome 19 mouse pangenome

A Variant Calling of chr19 (DBA/2J sample)

Comparison between the Truth Sets (GATK-10X/ GATK-PacBio) and the Query set (vg-10X)



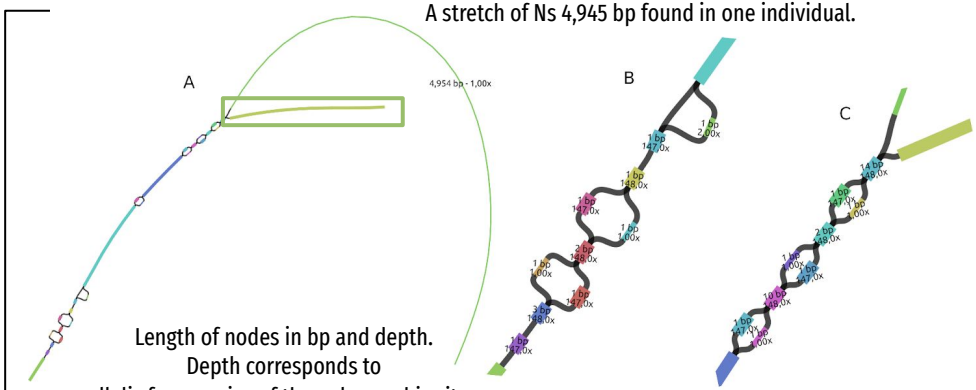
Evaluation

Within masked regions precision and sensitivity of vg calls are 90% and 85-86%, respectively.
For **SNPs** these figures rise to **93-94% and 84-85%**. For **INDELs** we have low values.

B Odgi extract of Gene 2700046G09Rik

Odgi extract

A stretch of Ns 4,945 bp found in one individual.



Length of nodes in bp and depth.
Depth corresponds to allelic frequencies of the polymorphic sites.

https://rrwick.github.io/Bandage/

SINESearch Fri Oct 15 18:26:09 2021

Query sequence : **1** (1455 nt)
Sequence bank : **1** (54395 residues in 217 sequences)
Min overlap : **70 nt**
Min similarity : **65%**

Total hits : **7**

Sequence	identity	overlap	strand	limits	bestfit	
B1	65.4%	130 nt	[R]	850-721	0.911	mouse
MEN	67.8%	115 nt	[R]	851-742	0.921	squirrels
B1-dID	67.3%	101 nt	[R]	842-744	0.931	Sciuridae
pB1	69.0%	87 nt	[R]	838-762	0.940	Rodents
Alu	70.1%	77 nt	[R]	838-768	0.947	Primates
Tu-II	70.1%	77 nt	[R]	833-762	0.947	tree shrews
SP-D-Geo	68.4%	76 nt	[R]	842-768	0.948	Dipodomys ordii

https://sines.eimb.ru/

by Flavia Villani

References

Miga et al. “Telomere-to-telomere assembly of a complete human X chromosome”. *Nature* 585, 79–84 (2020). https://doi.org/10.1038/s41586-020-2547-7

Logsdon et al. “The structure, function and evolution of a complete human chromosome 8”. *Nature* 593, 101–107 (2021). https://doi.org/10.1038/s41586-021-03420-7

Nurk et al. “The complete sequence of a human genome”. *bioRxiv*, 2021.05.26.445798; doi: https://doi.org/10.1101/2021.05.26.445798

Eizenga et al. “Pangenome Graphs”. *Annual Review of Genomics and Human Genetics*, 2020 21:1, 139-162

Garrison et al. “Variation graph toolkit improves read mapping by representing genetic variation in the reference”. *Nature biotechnology*, vol. 36,9 (2018): 875-879. doi:10.1038/nbt.4227

Wick et al. “Bandage: interactive visualisation of *de novo* genome assemblies”. *Bioinformatics*, 31(20) (2015): 3350-3352.

Beyer et al. “Sequence tube maps: making graph genomes intuitive to commuters”. *Bioinformatics*, Volume 35(24) (2019): 5318–5320

Sekar, A., Bialas, A., de Rivera, H. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* 530, 177–183 (2016). https://doi.org/10.1038/nature16549

Laks, Emma et al. “Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing.” Cell vol. 179,5 (2019): 1207-1221.e22. doi:10.1016/j.cell.2019.10.026

References

J. X. Zheng, S. Pawar and D. F. M. Goodman, "Graph Drawing by Stochastic Gradient Descent," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 9, pp. 2738-2748, 1 Sept. 2019, doi: 10.1109/TVCG.2018.2859997.

F. Niu, B. Recht, C. Ré, S. J. Wright. “HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent”, in *NeurIPS Preceedings*, 2011.

T. T. Yokoyama, S. Heumos, J. Seaman, D. Trybushnyi, T. Pook, A. Guarracino, E. Garrison, J. T. Bollemann. “Semantic variation graphs: Ontologies for pangenome graphs”. *International Conference on Intelligent Systems for Molecular Biology* (2020), Online, 13.07. 2020–16.07. 2020. 32.02.12; LK 01.