ODGI: scalable tools for pangenome graphs

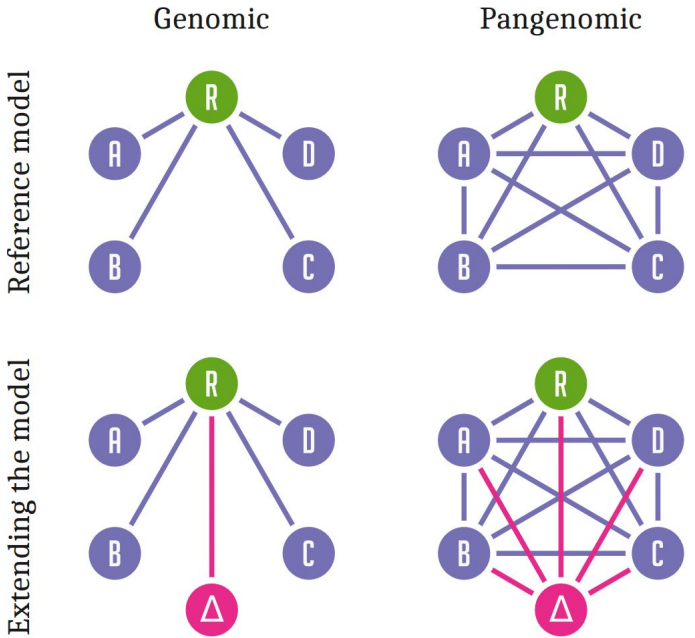
German Conference on Bioinformatics (GCB) 2021
7 September 2021

Andrea Guarracino, Simon Heumos,
Pjotr Prins, Erik Garrison

De novo assembly and pangenomes

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**.



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

Pangenome graphs

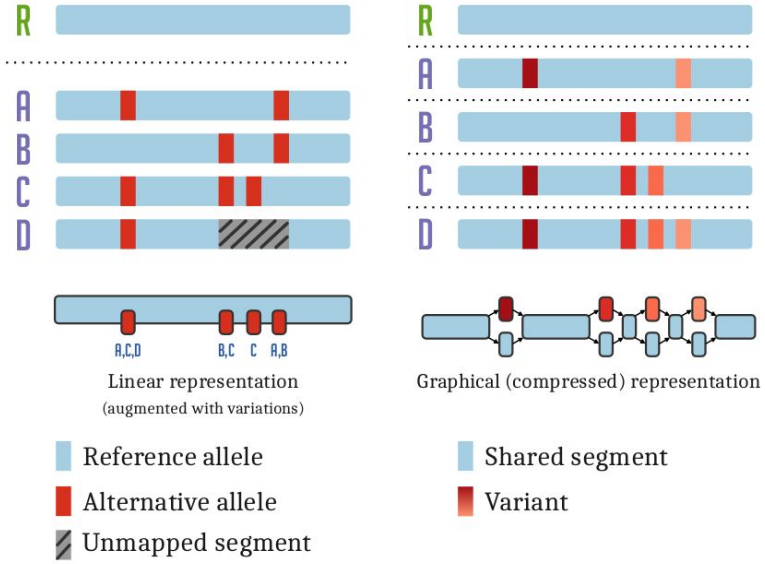


Figure from Eizenga et al., 2020.

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

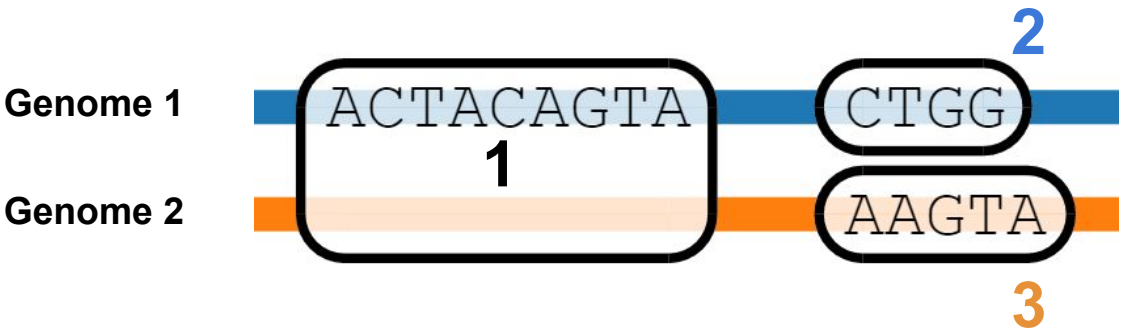
Pangenome graphs compress redundant sequences into a smaller data structure that is still representative of the full set.

Variation graphs

Genome 1: **ACTACAGTA**CTGG Path: **1** 2

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

Linear sequences are **paths** through nodes.



Graph topology is not directly shown.

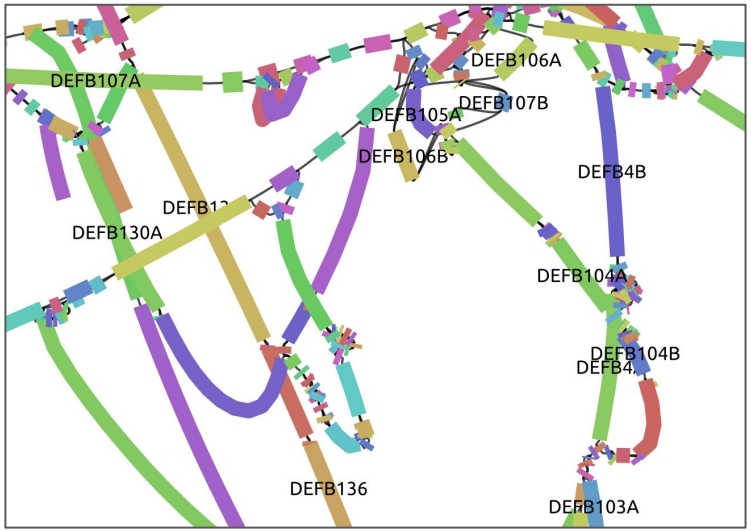
The nodes represent DNA sequences.

Sketch made using SequenceTubeMap.

Vertebrate pangenome graphs are complex

A major challenge is writing software that can deal with graphs representing **hundreds of eukaryotic genomes**.

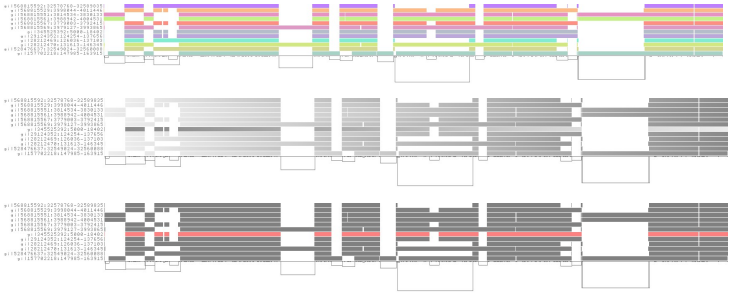
Highly repetitive regions (centromeres, segmental duplications, and acrocentric chromosomes) increases the complexity of the operations performed on graphs.



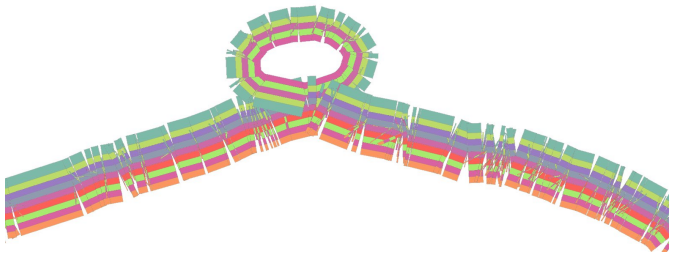
Beta-defensin *locus* of a pangenome graph made with 44 human *de novo* assemblies from the Human Pangenome Reference Consortium (HPRC) dataset. Figure made with Bandage.

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

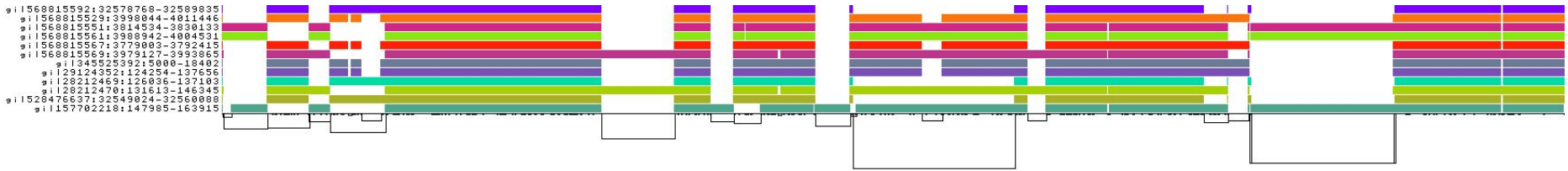


ODGI offers more than 30 tools for graph interrogation, manipulation, and visualization.

Visualizing pangenome graphs in 1D - `odgi viz`

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

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



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

Visualizing pangenome graphs in 1D - `odgi viz`

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

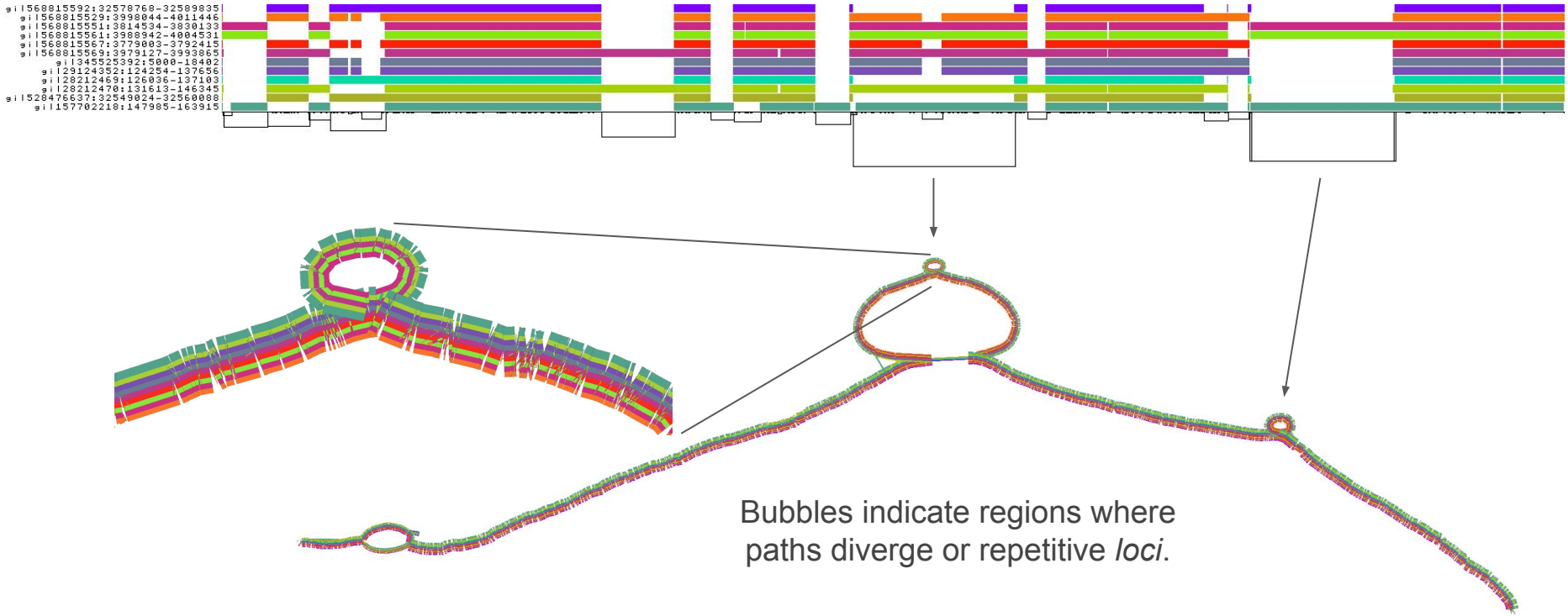
Colored by path position (light = start, dark = end)

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

Colored by orientation (black = forward, red =reverse)

Visualizing pangenome graphs in 2D - `odgi draw`

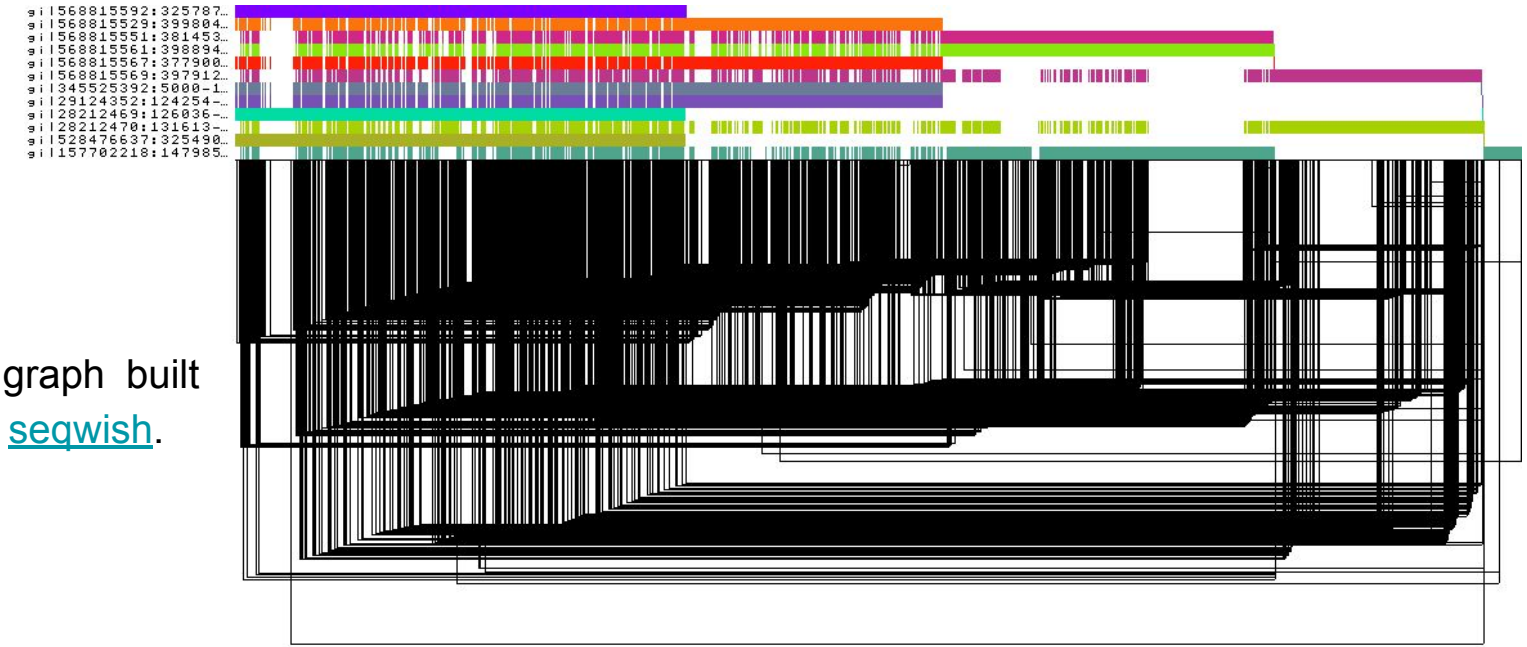
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Finding latent structures in pangenome graphs

Pangenome graphs built from raw sets of alignments may have complex structures which can introduce difficulty in downstream analyses.

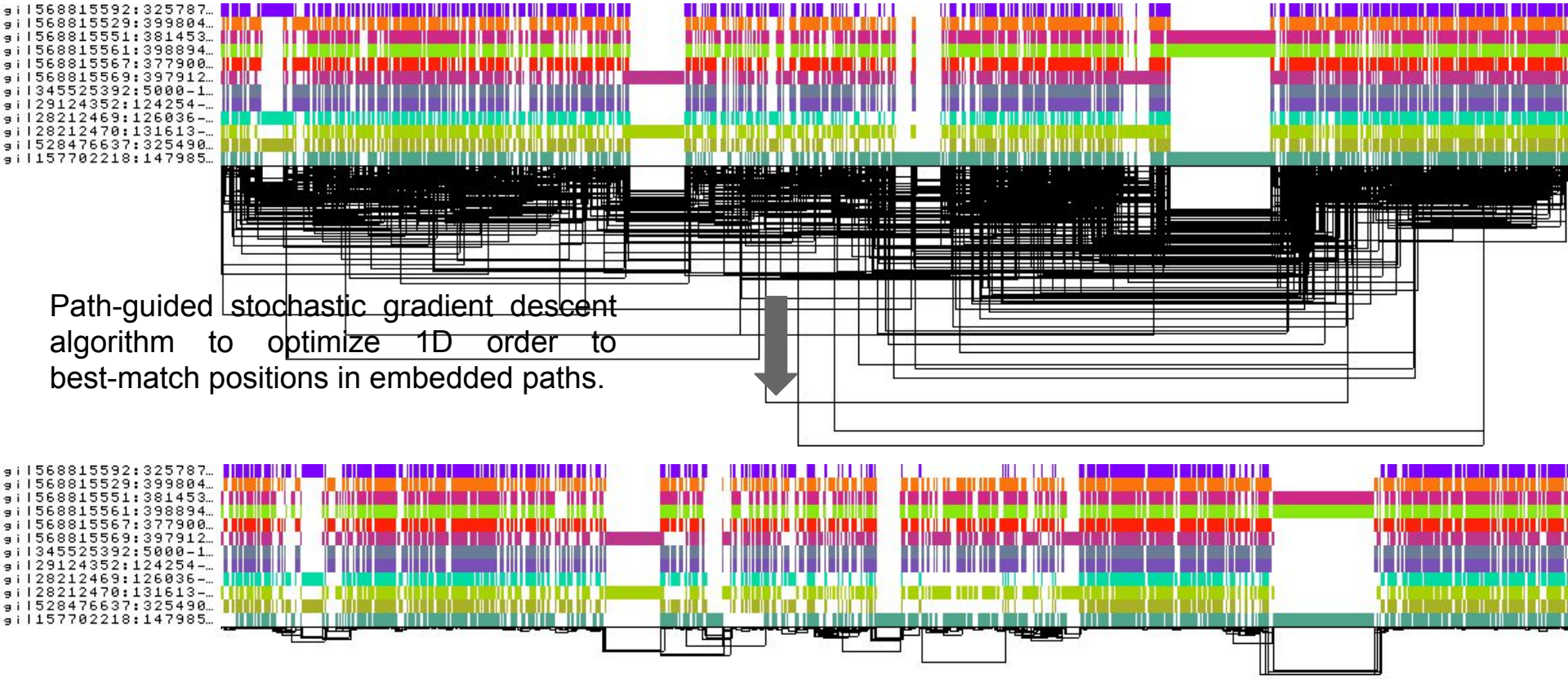
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Raw graph built using seqwish.

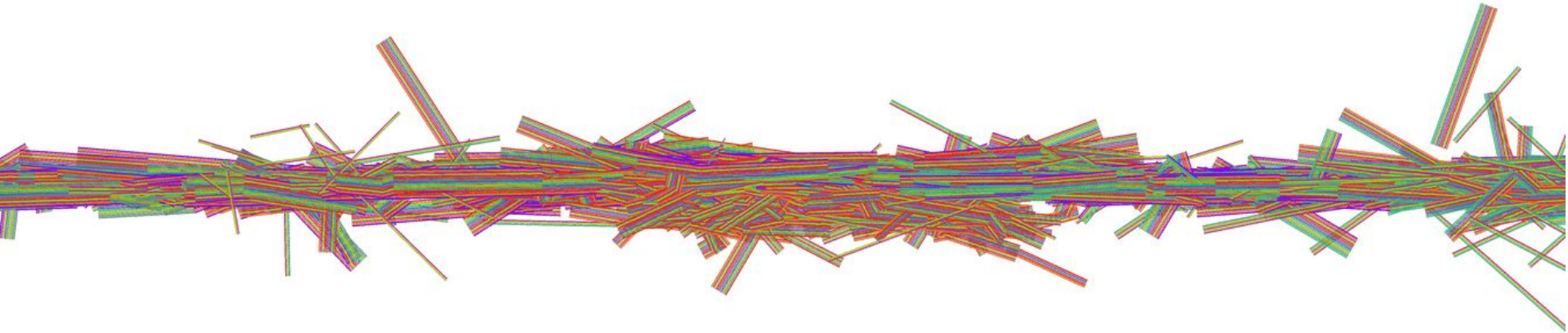
Finding latent structures in pangenome graphs - `odgi sort`

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



Finding latent structures in pangenome graphs - `odgi layout`

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

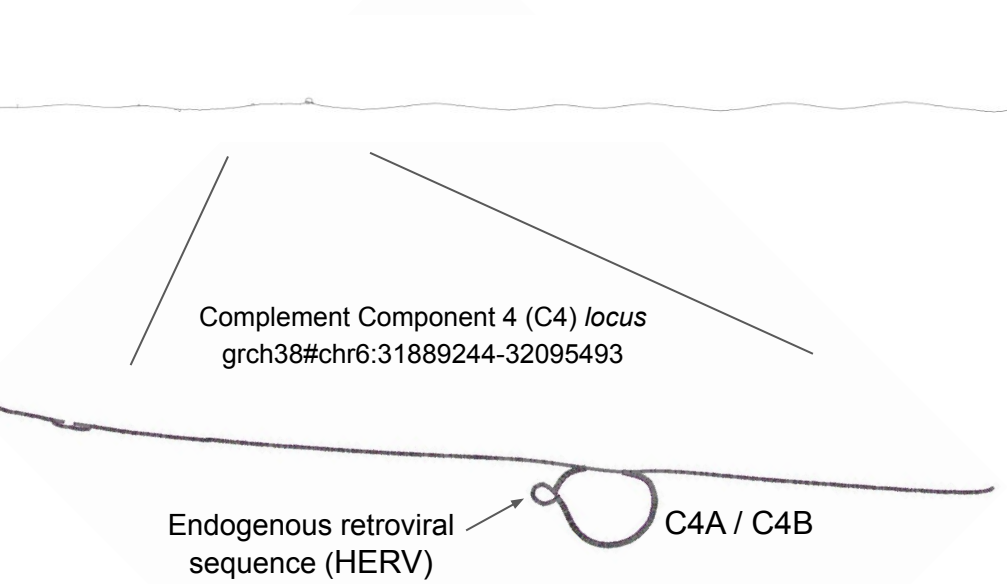


Path-guided stochastic gradient descent algorithm to optimize 2D layout. Path-labeled rendering with `odgi draw`.

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.



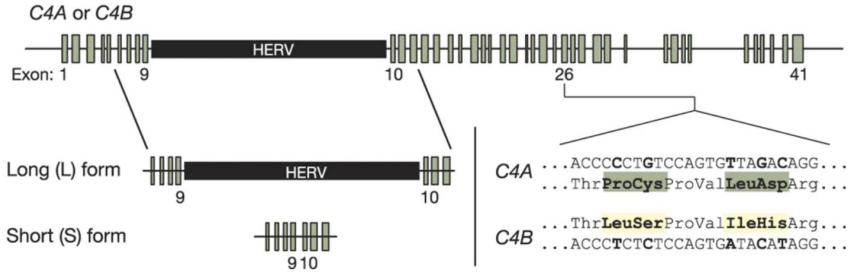
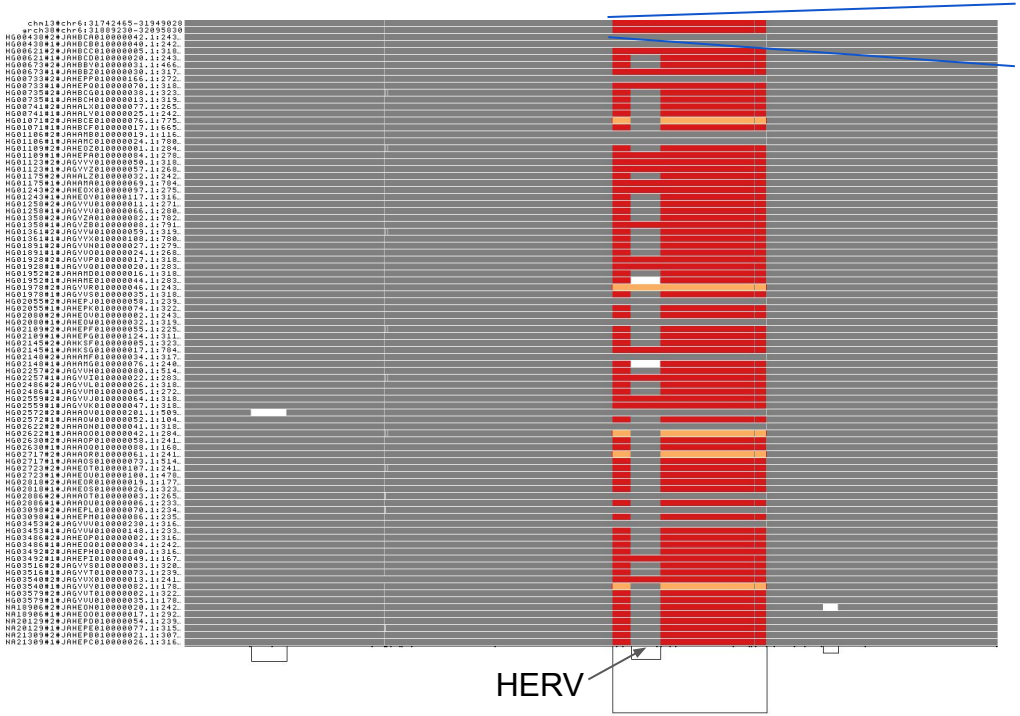


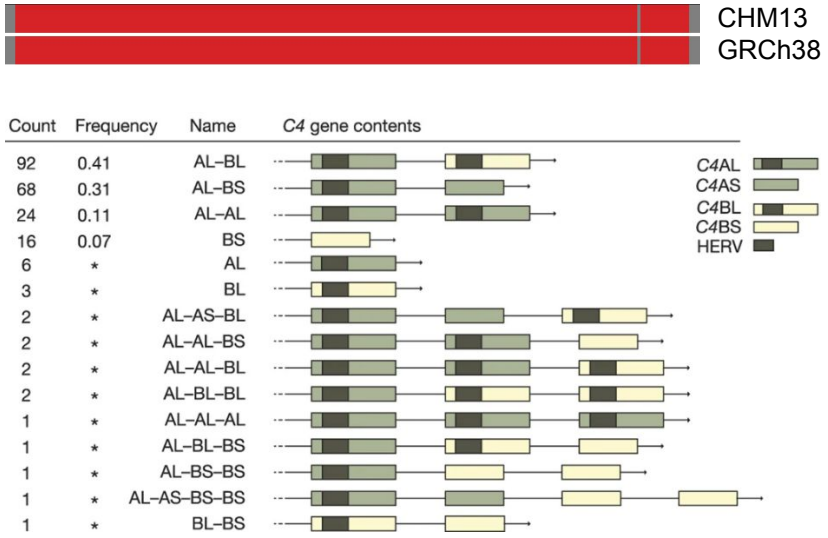
Figure from Sekar et al., 2016.

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)

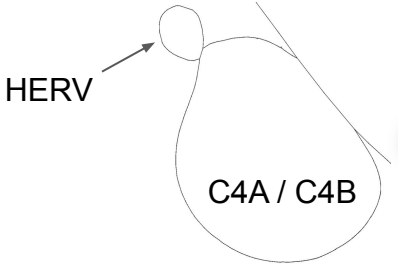




222 chromosomes from 111 genetically unrelated individuals. Figure from Sekar et al., 2016.

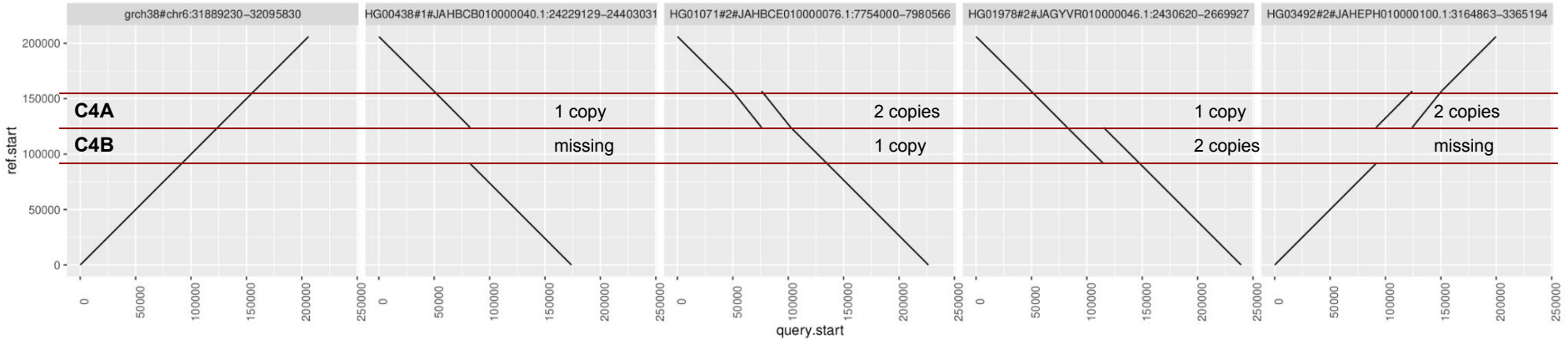
Untangling pangenome graphs - `odgi untangle`

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.

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.

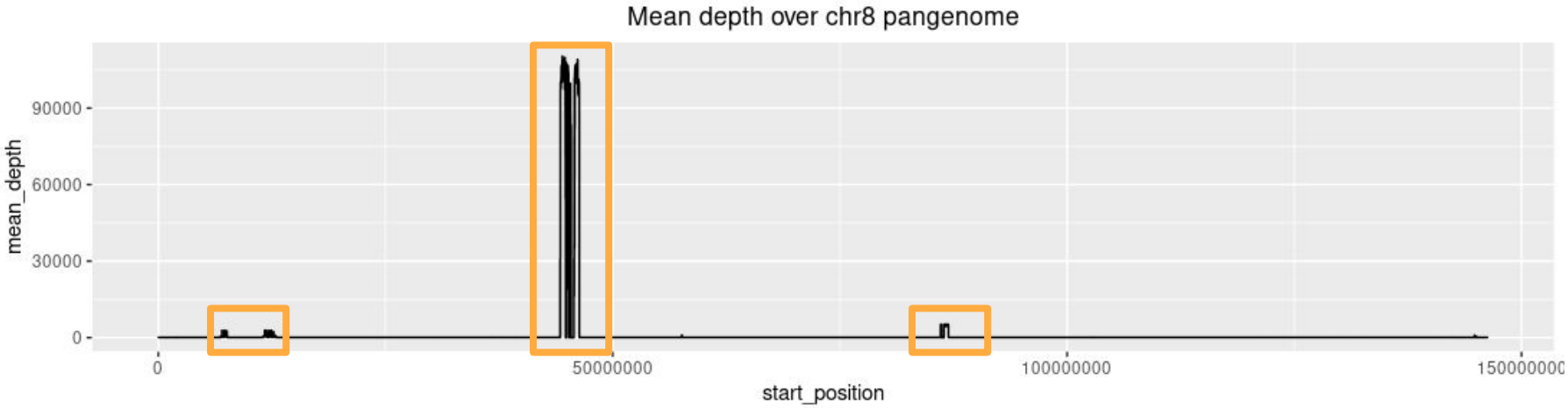
Detecting complex regions - `odgi depth`

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.

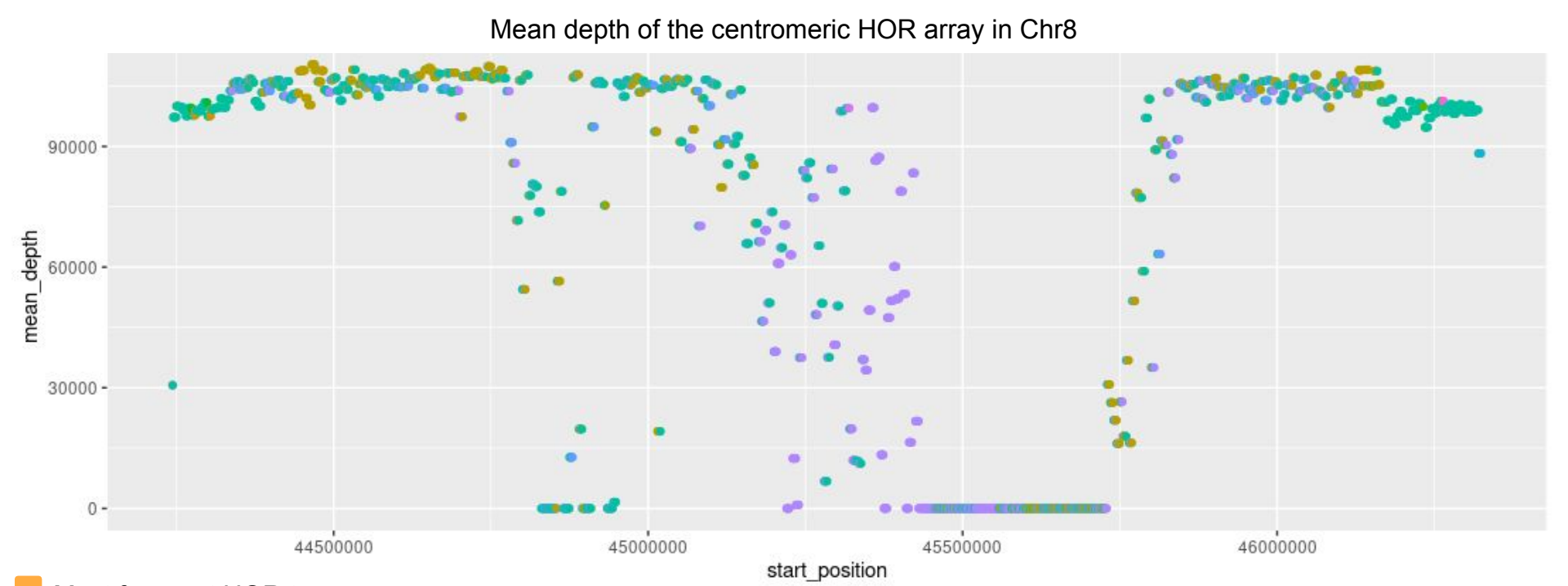
Output: BED file with the mean node depth distribution across windows of the pangenomic CHM13 positions.

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.

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

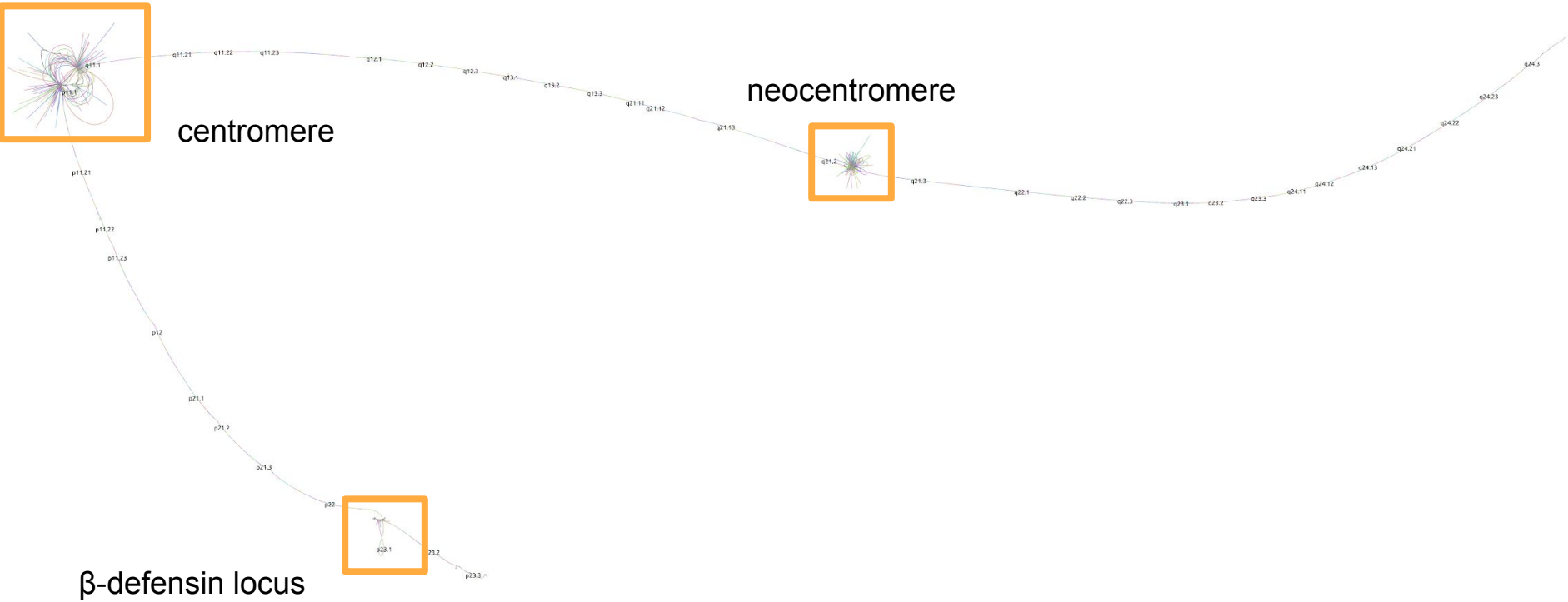
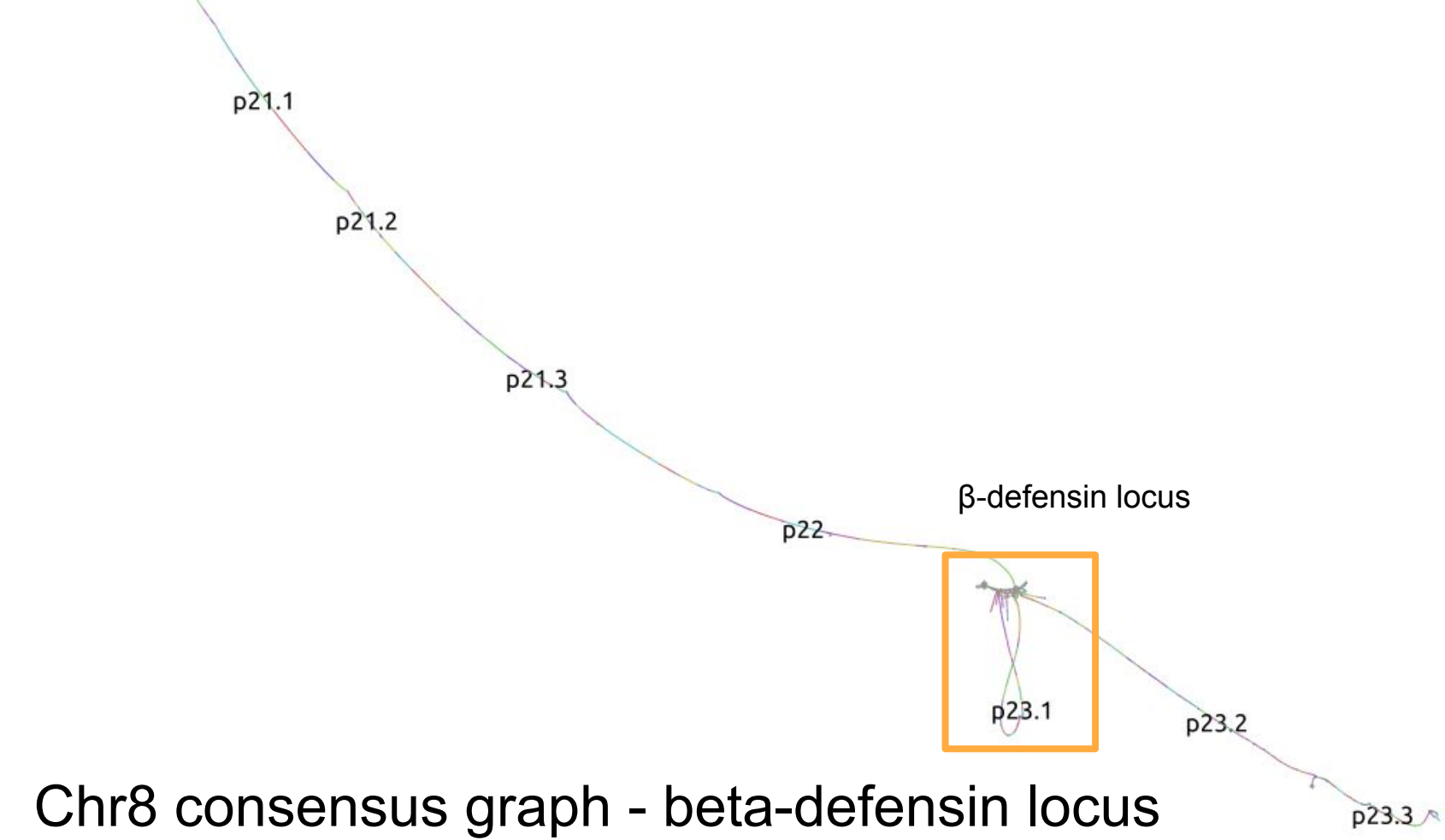


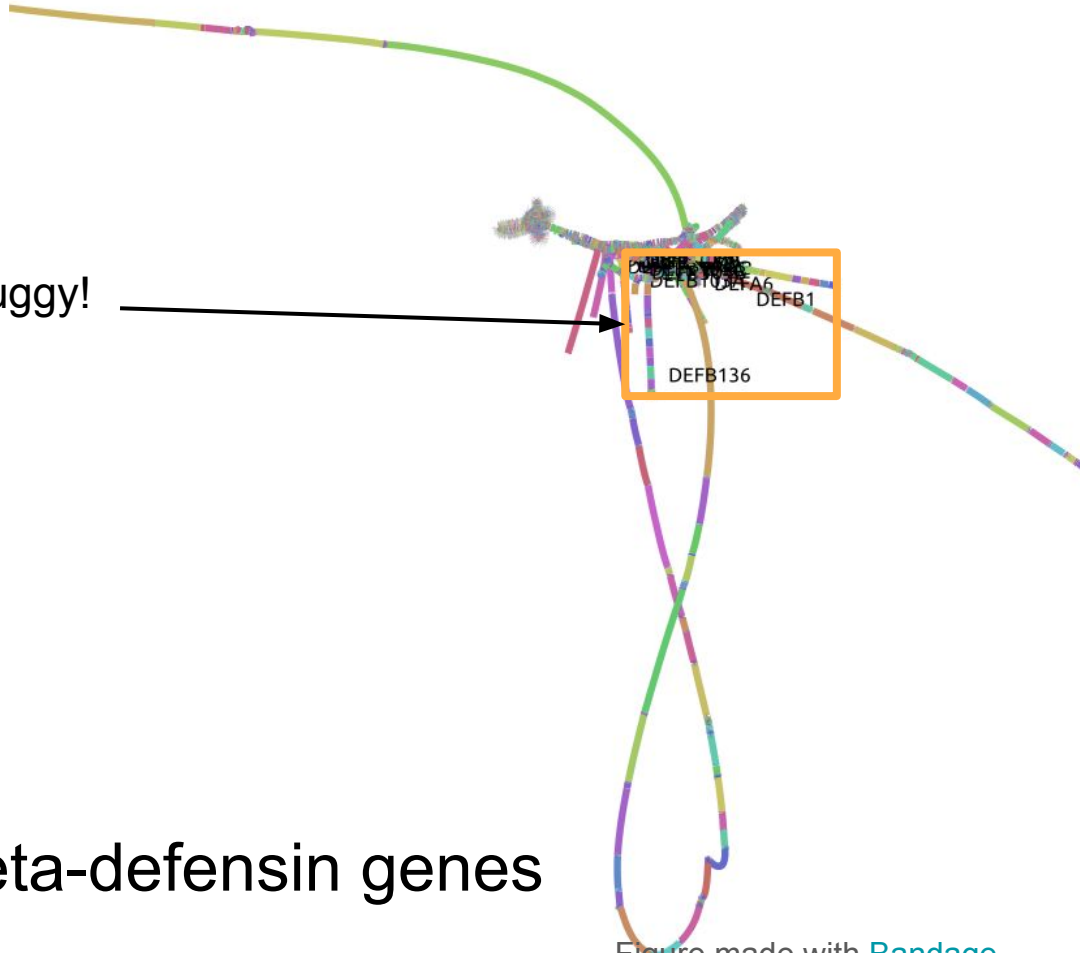
Figure made with Bandage.



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

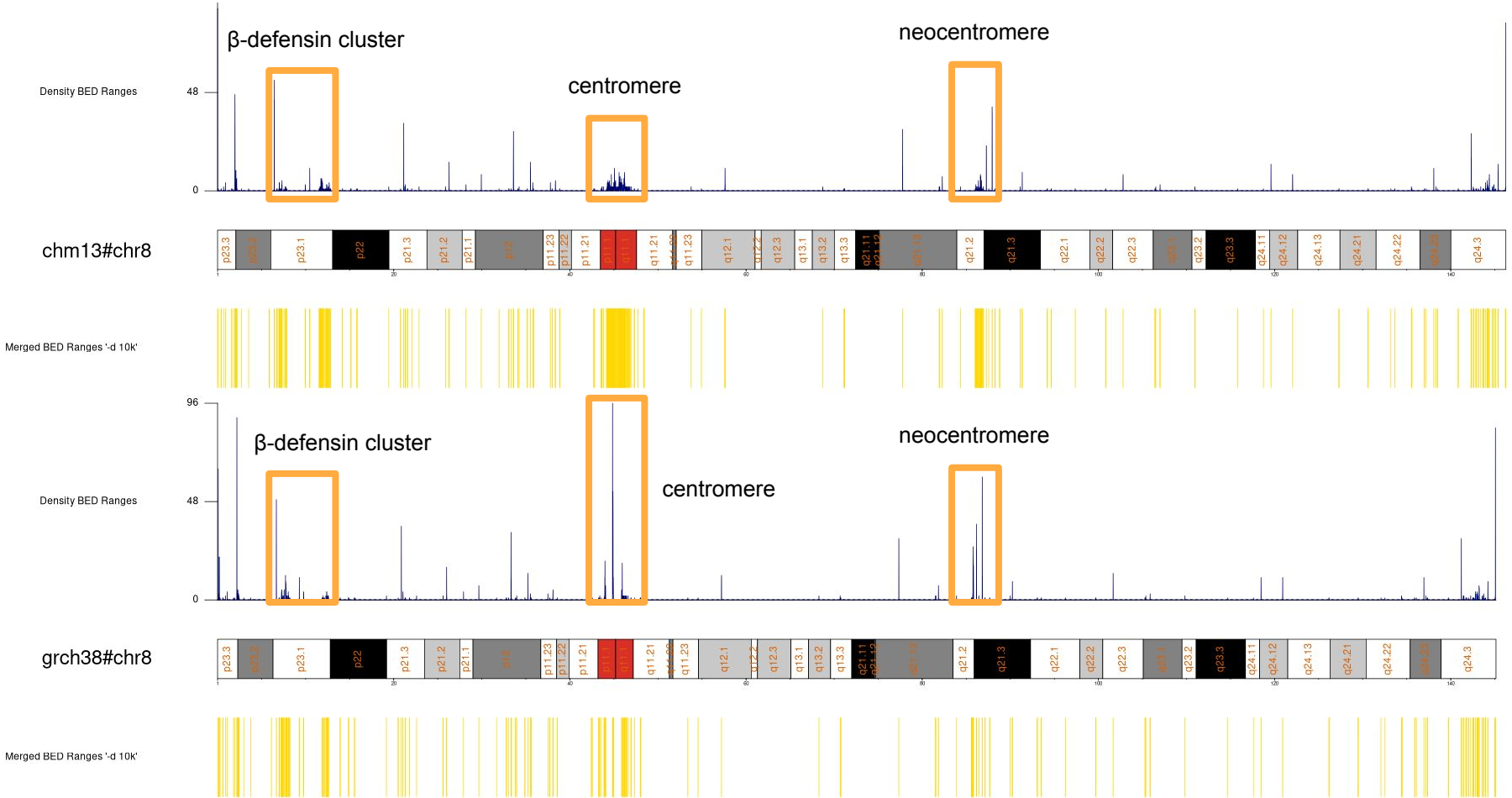
Where do our contigs' ends match the references? - Detecting regions that are difficult to assemble.

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: 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

Acknowledgments

EU Pangenome Group

Vincenza Colonna

Flavia Villani

David G. Ashbrook

Robert W.Williams

Christian Fischer

Sven Nahnsen

Gisela Gabernet

Michael Krone

Oliver Kohlbacher

HPRC

Christian Kubica

Sebastian Vorbrugg

Jörg Hagmann

ODGI recipe

ODGI repository

ODGI documentation

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