Graph layout by path-guided stochastic gradient descent

International Genome Graph Symposium 2022
6 July 2022

Simon Heumos*, Andrea Guarracino*,
Sven Nahnsen, Erik Garrison

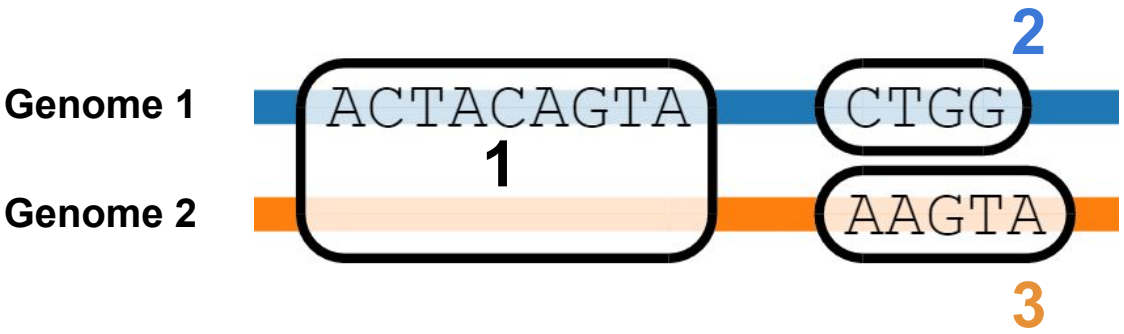
*contributed equally

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.

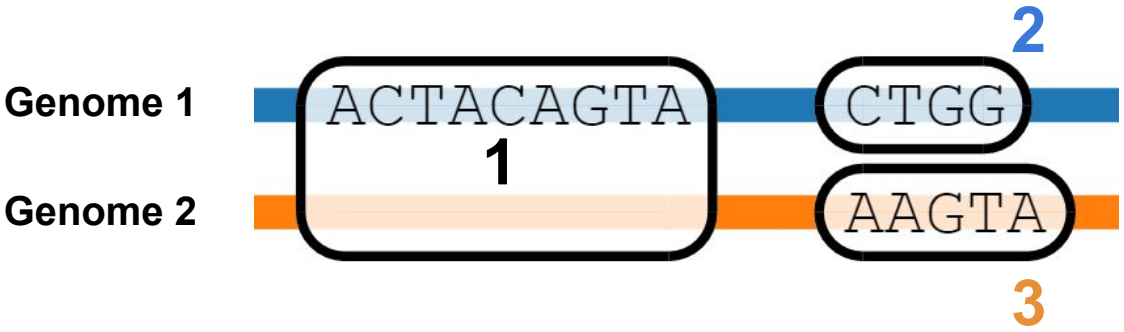
Sketch made using SequenceTubeMap.

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

Towards a 1D visualization

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

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



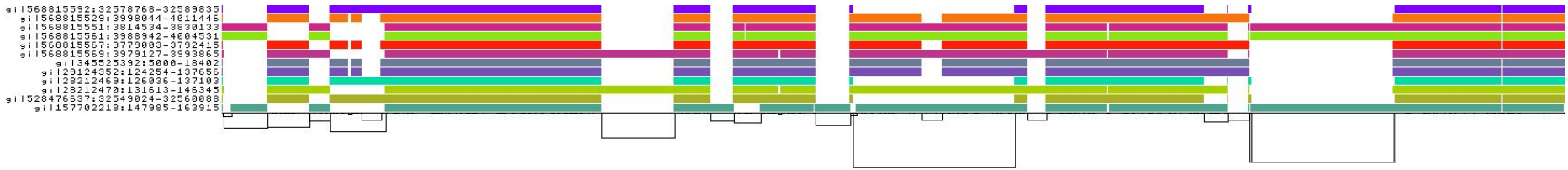
Concatenate nucleotides to a pangenome sequence.

ACTACAGTACTGGAAGTA

Presence - absence matrix encodes actual genomic sequence.

1D Graph visualization explained

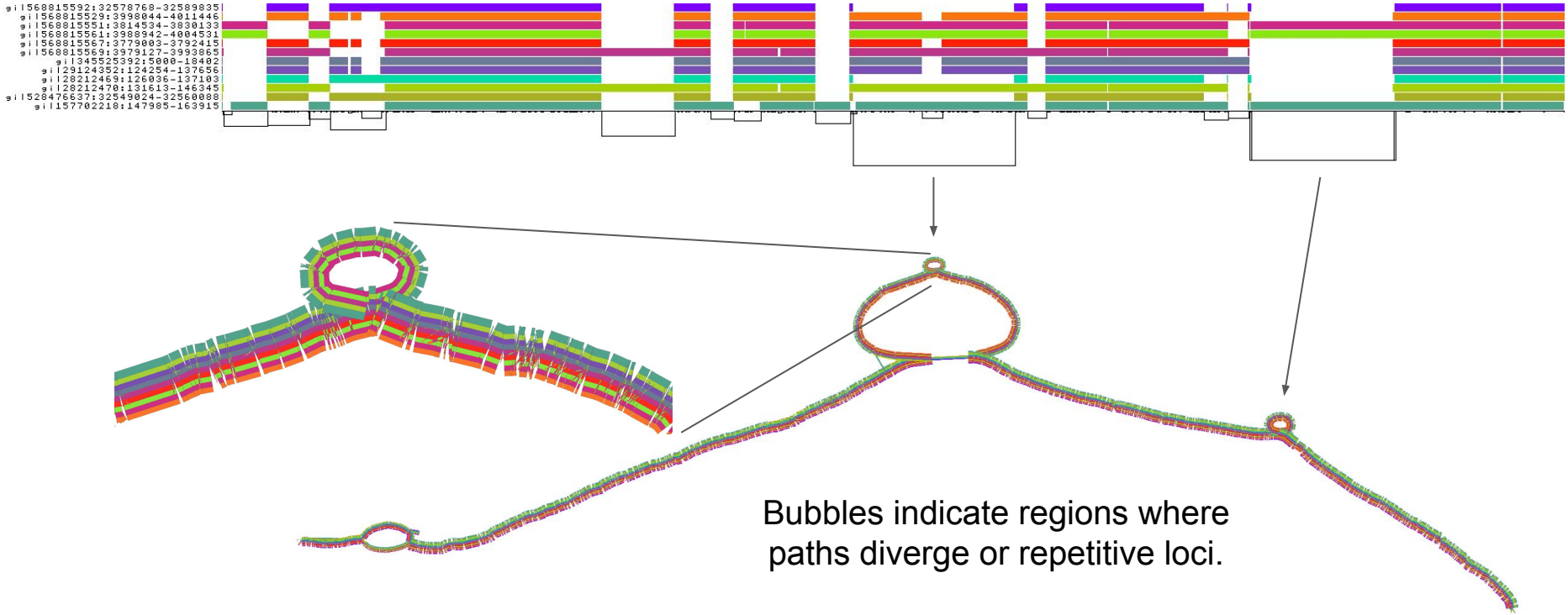
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



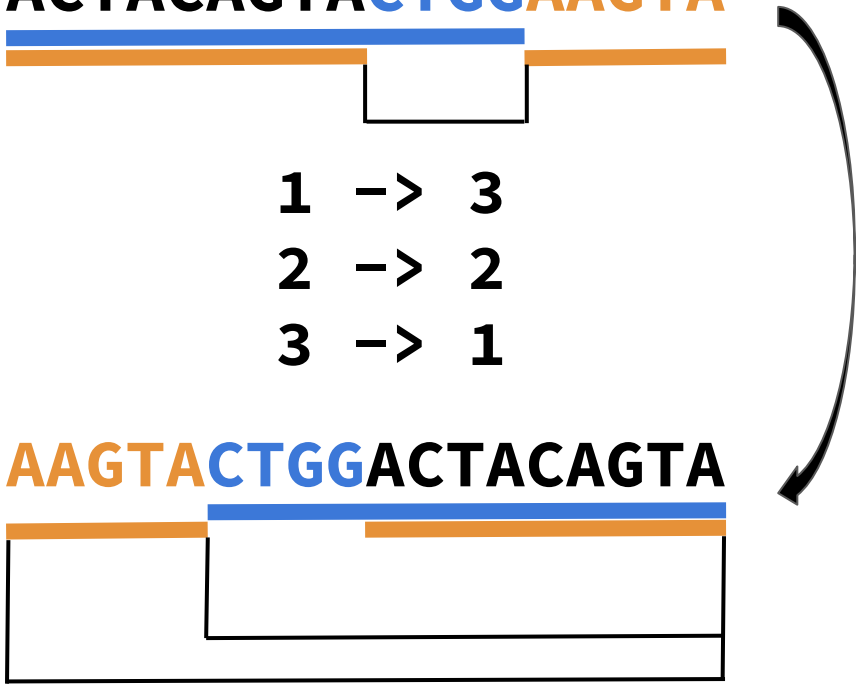
- Graph nodes are arranged from left to right forming the pangenome sequence
- Colored bars are the paths vs. the pangenome sequences in a binary matrix
- Path names are left
- The black lines under the paths are the links representing the graph topology
  If sequences traverse nodes from left to right, we don’t draw a link

2D Graph visualization explained

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



1D Sorting process explained

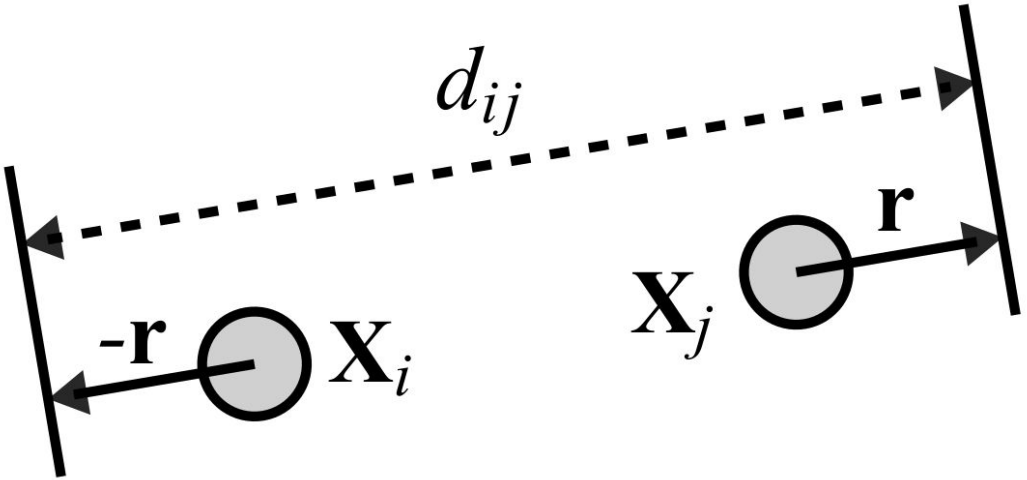



```
S    1    ACTACAGTA
S    2    CTGG
S    3    AAGTA
P    Genome1  1+,2+
P    Genome2  1+,3+
L    1    +    2    +
L    1    +    3    +
```

```
S    1    AAGTA
S    2    CTGG
S    3    ACTACAGTA
P    Genome1  3+,2+
P    Genome2  3+,1+
L    3    +    2    +
L    3    +    1    +
```

Graph drawing by stochastic gradient descent

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.

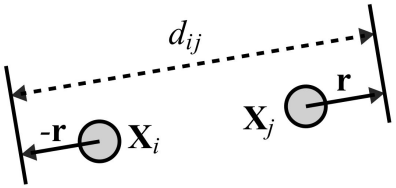


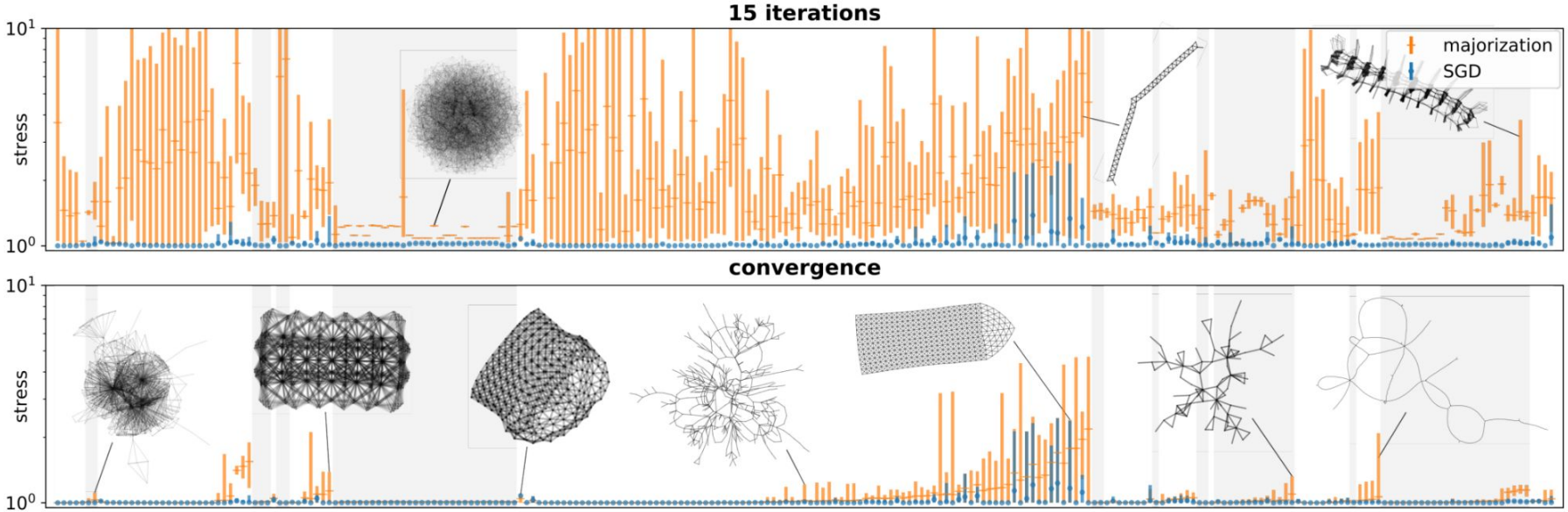
We organize the graph by moving nodes so their position in the layout better-matches their distance through paths in the graph.

We learn the graph layout.

Graph drawing by stochastic gradient descent

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.





SGD Layout of very large (sparse) variation graphs

SGD graph drawing is fast, but obtaining the distances in the graph is extremely slow.

We might need the distance between all pairs of nodes!

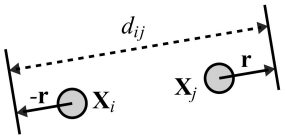
So we use a trick that works on our sparse (and large 1M to 1G node) graphs: index the genome paths through them and use that to quickly obtain distances.

→*Path-guided SGD*

To go really fast we apply HOGWILD!-SGD:

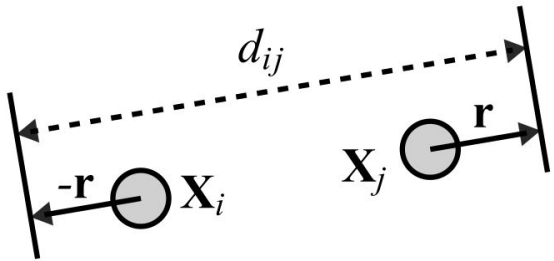
Algorithm 1: Stochastic Gradient Descent

SGD (G)**:**
  inputs: graph $G = (V, E)$
  output: k-dimensional layout $\mathbf{X}$ with n vertices
  $d_{\{i,j\}} \leftarrow ShortestPaths(G)$
  $\mathbf{X} \leftarrow RandomMatrix(n, k)$
  for η *in annealing schedule* **:**
    foreach $\{i, j : i < j\}$ *in random order* **:**
      $\mu \leftarrow w_{ij}\eta$
      if $\mu > 1$ **:**
        $\mu \leftarrow 1$
      $\mathbf{r} \leftarrow \frac{||\mathbf{X}_i - \mathbf{X}_j|| - d_{ij}}{2} \frac{\mathbf{X}_i - \mathbf{X}_j}{||\mathbf{X}_i - \mathbf{X}_j||}$
      $\mathbf{X}_i \leftarrow \mathbf{X}_i - \mu\,\mathbf{r}$
      $\mathbf{X}_j \leftarrow \mathbf{X}_j + \mu\,\mathbf{r}$



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.

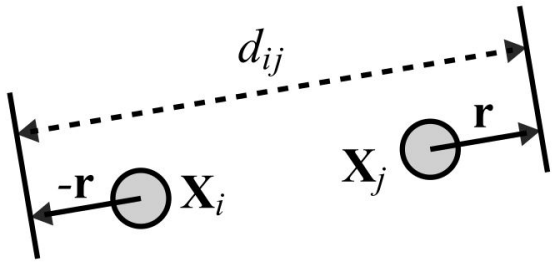
- The first node X_i of a pair is a uniform path step pick from all nodes.

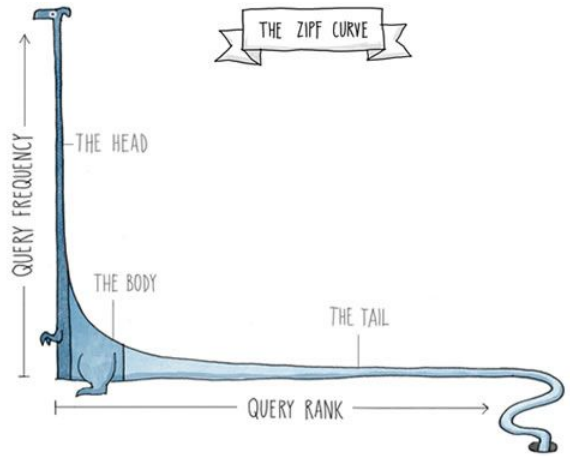
1D Graph Sorting by P-SGD - The Algorithm Explained

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.

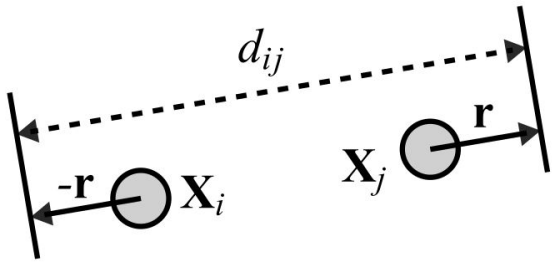




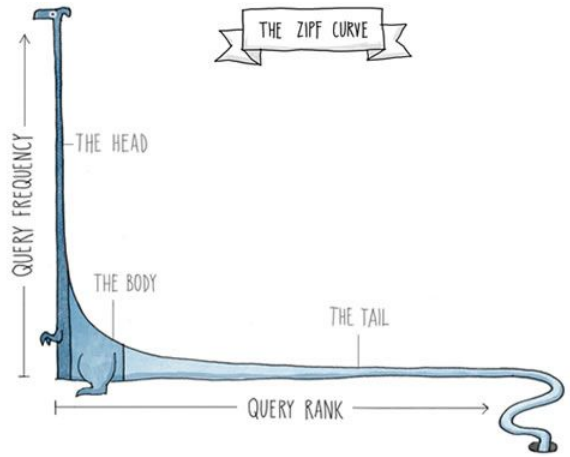
1D Graph Sorting by P-SGD - The Algorithm Explained

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.



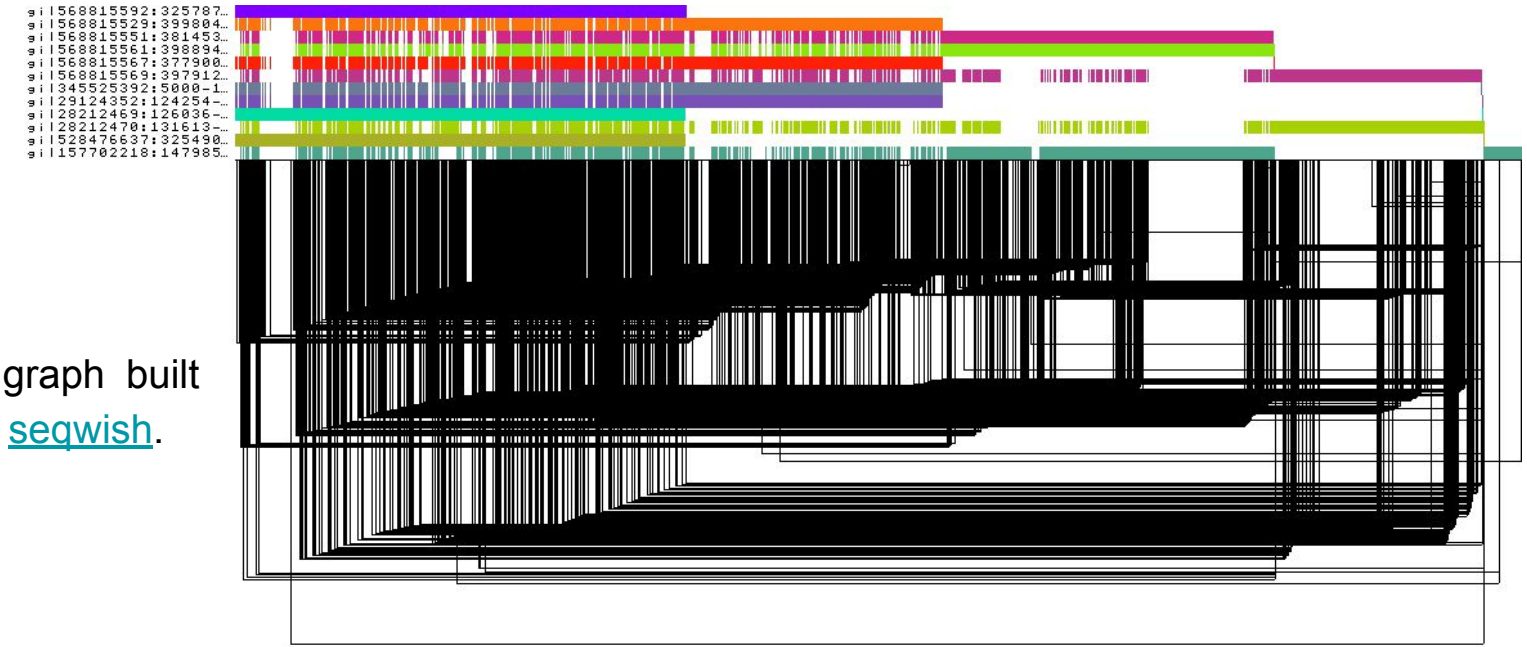
- The first node X_i of a pair is a uniform path step pick from all nodes.
- The second node X_j of a pair is sampled from the same path following a Zipfian distribution.
- The path nucleotide distance of the nodes in the pair guides the actual layout distance d_{ij} update of these nodes. The magnitude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



Finding latent structures in pangenome graphs

A “raw” graph produced by seqwish is sorted by the order in which we see sequences in the input genomes. This is basically disordered.

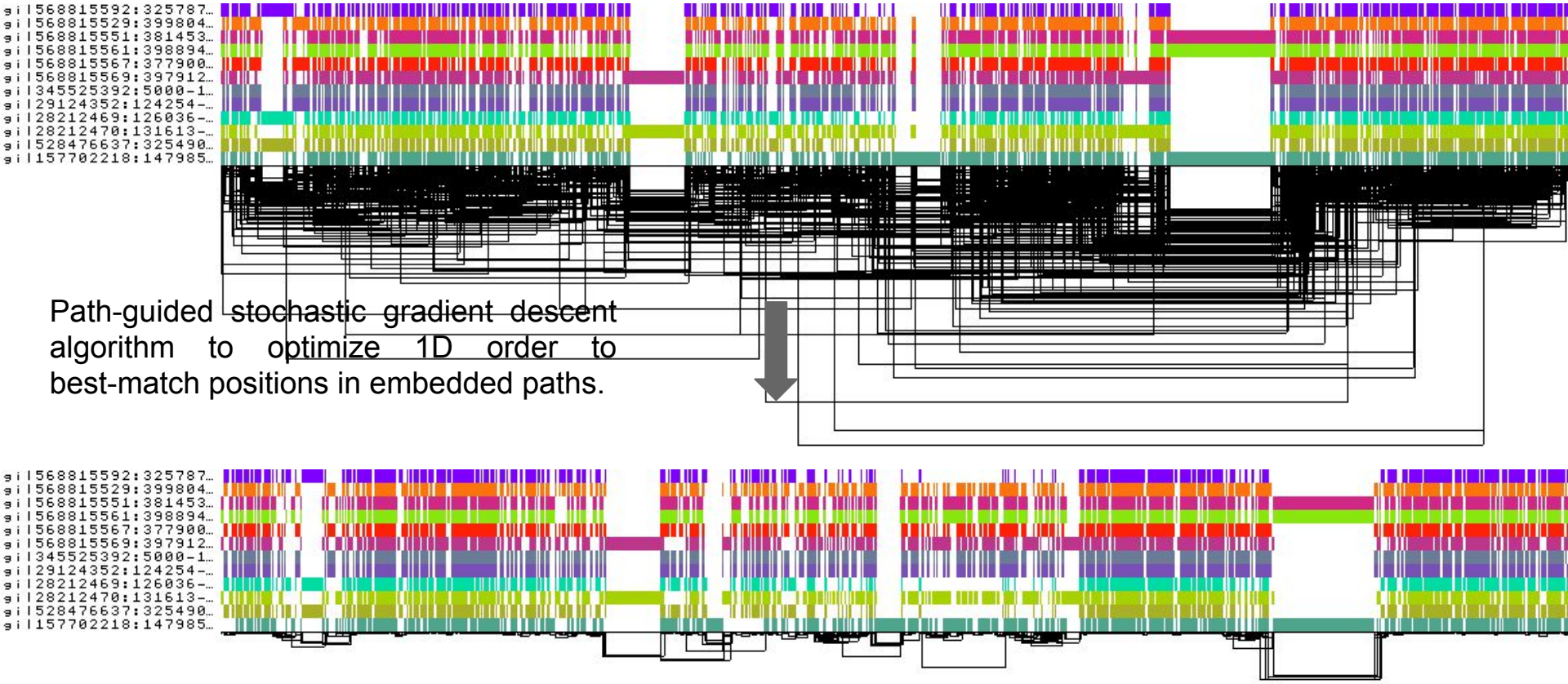
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Raw graph built using seqwish.

Learned *1D* models of pangenome graphs

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



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.

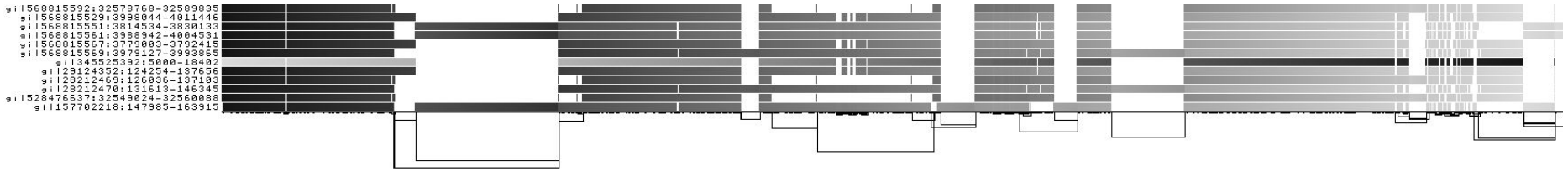
Reference-based PG-SGD

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

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

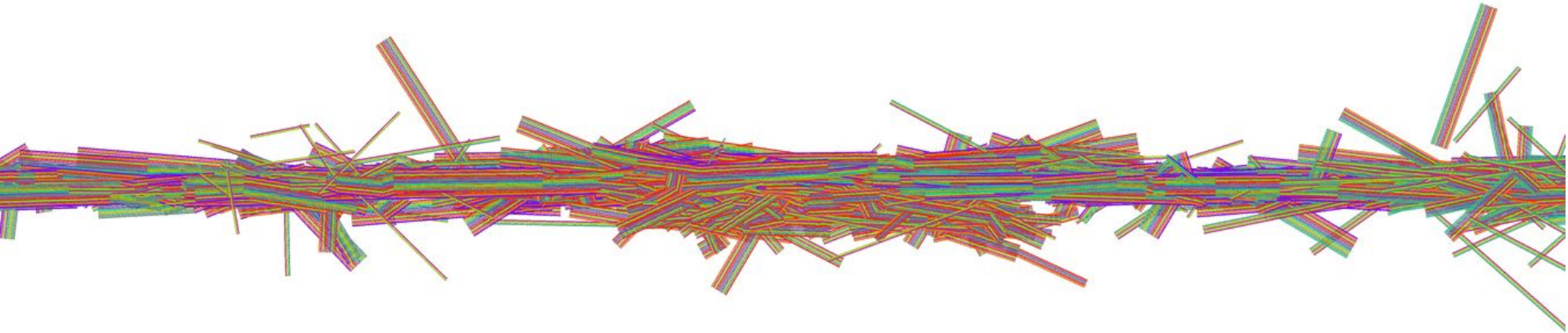
Default PG-SGD

Reference-based PG-SGD with `gi|345525392:5000-18402` as the reference



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.


Oggi relax
Buona domenica
a tutti
Fun
Family
Fizziana Mosso

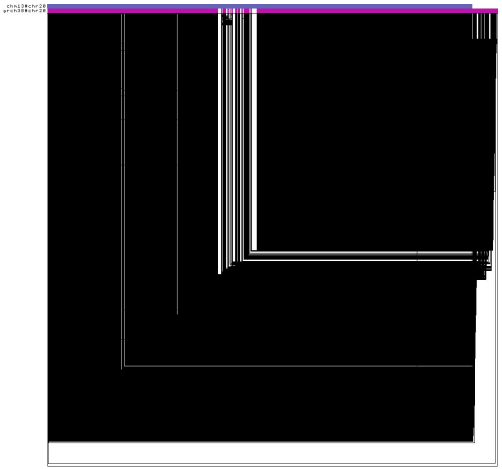

oggi relax
Buongiorno
Il ranocchio

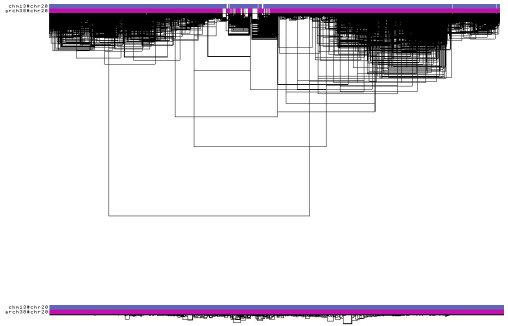

OGGI
RELAX
ilBuongiorno.it
SUNDAY

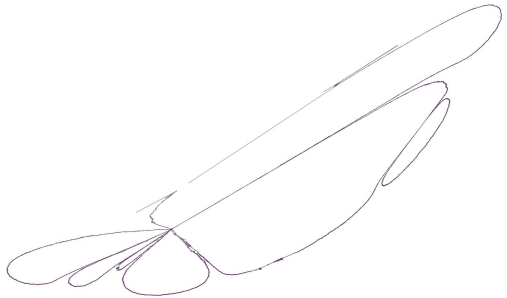
odgi relax

Problem

- Pangenome graphs from all vs. all alignments might contain spurious or abnormal alignments
- Such alignments introduce very large SVs
- Example: Chr20 pangenome graph of grch38 and chm13:



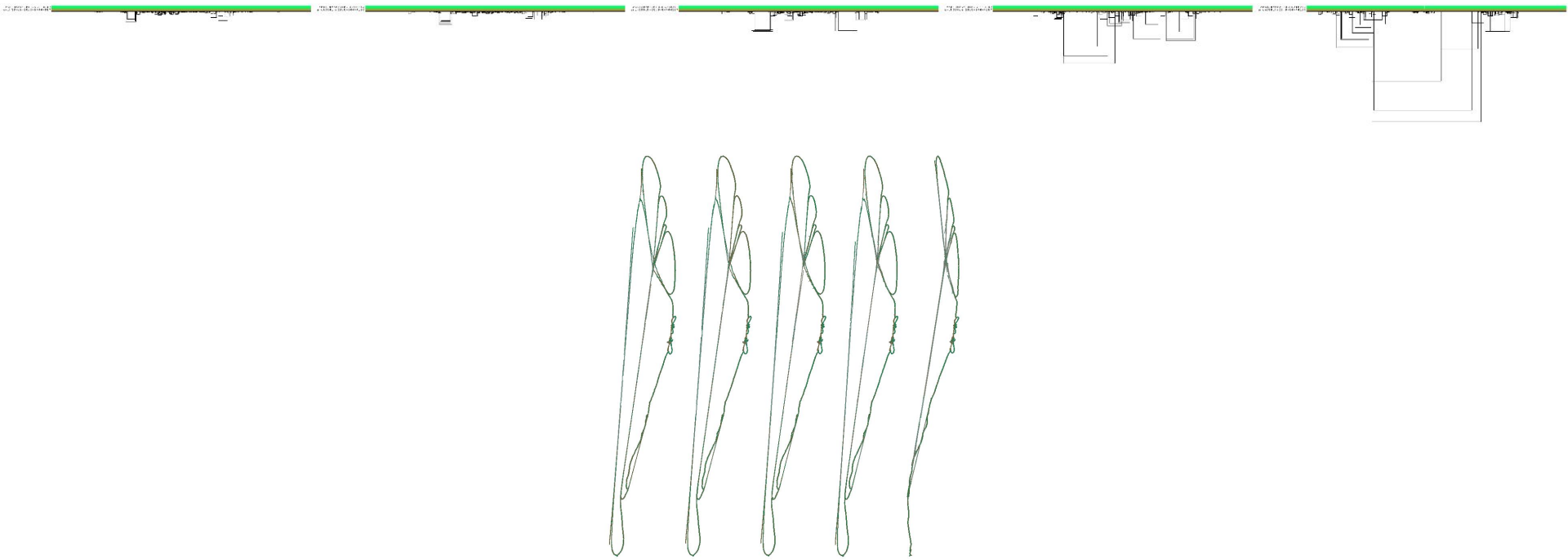




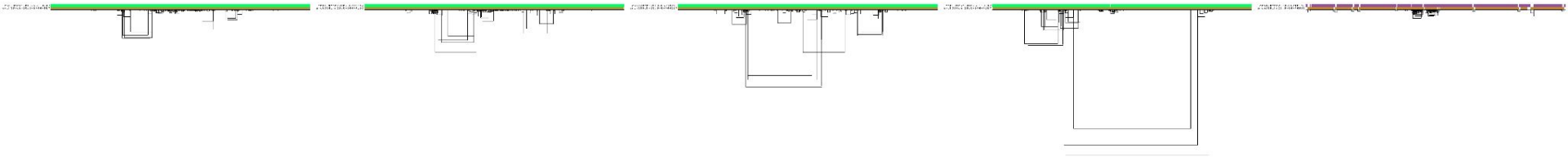
Possible Approach - Detect Tension and Relax

- Use our **1D** layout to detect *tension* in windows across the pangenome
- Measure discrepancy of:
  path layout position *versus* expected path nucleotide position
- The higher the *tension ratio*, the greater the possibility of a biologically infeasible alignment
- Take the ranges with the ***n*** highest *tension ratio* of each path and *relax* them with odgi extract:
  - Remove paths steps in given BED ranges
  - Lace paths back together, adding the sequences of the ranges as private insertions

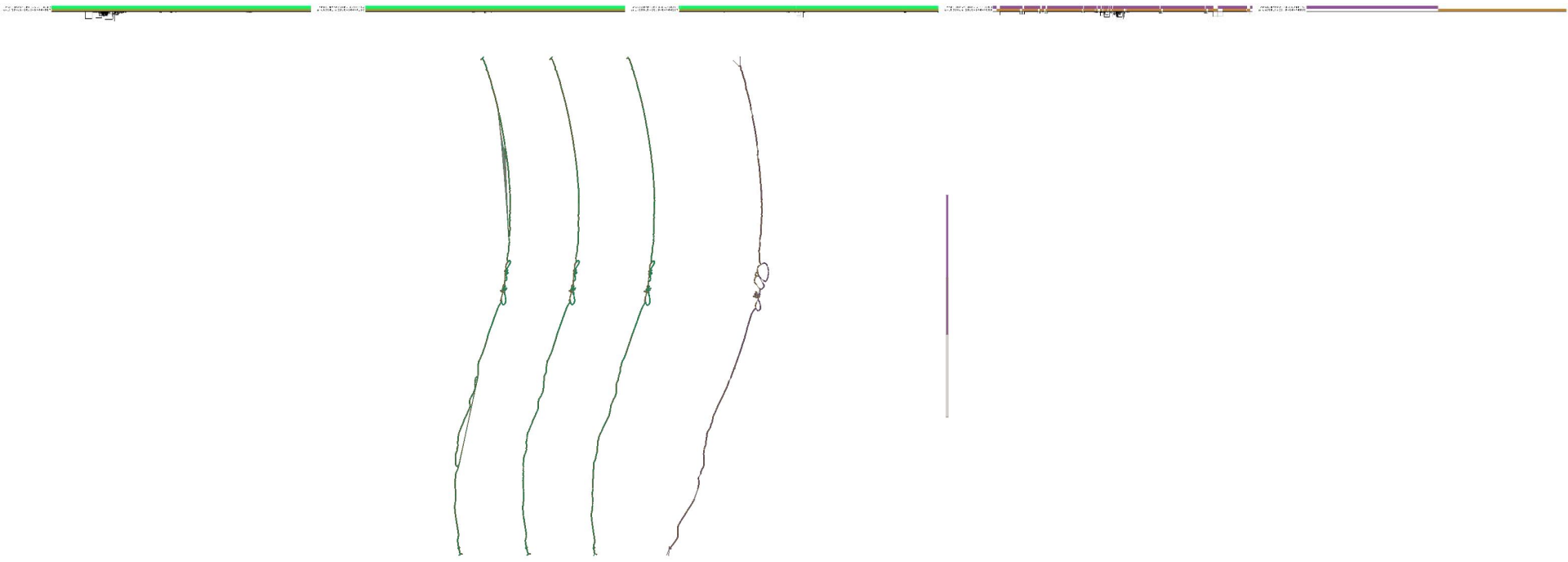
Relaxing - w100 - n10,20,100,1000,10000



Relaxing - w1000 - n10,20,100,1000,10000



Relaxing - w10000 - n10,20,100,1000,10000



Relaxing - w100000 - n10,20,100,1000,10000

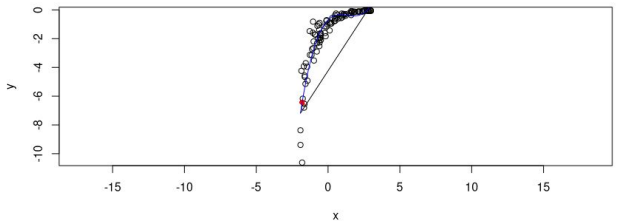
Relaxing - w-node - n10,20,100,1k,10k,50k,100k

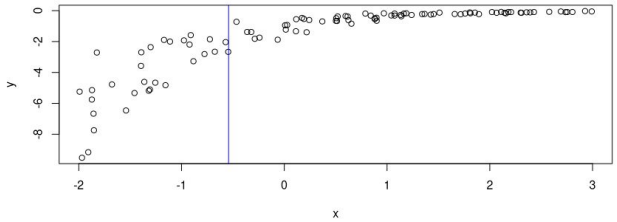
Relaxing ,100,1k,10k,50k,100k

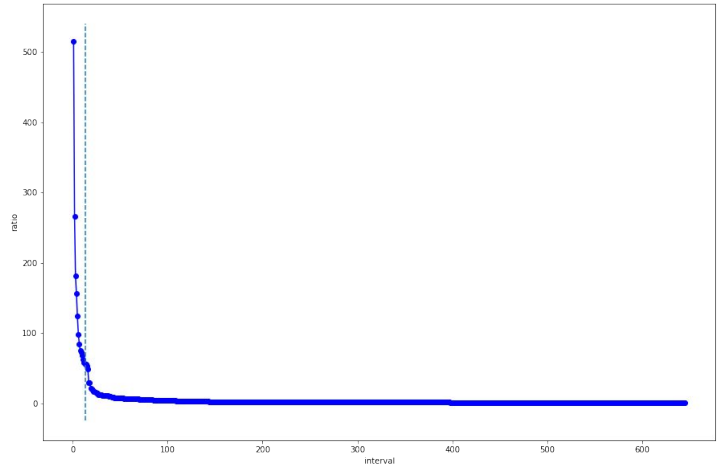
Finding a reasonable ***n***

Finding the knee

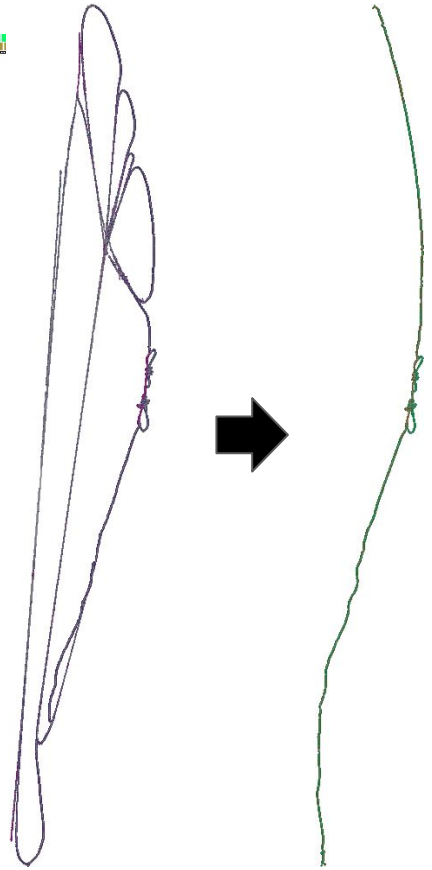
- R:
  - 😞 MASS
  - 😞 akmedoids
  - 😞 inflection
- Python:
  - 😍 kneed: https://ieeexplore.ieee.org/document/5961514





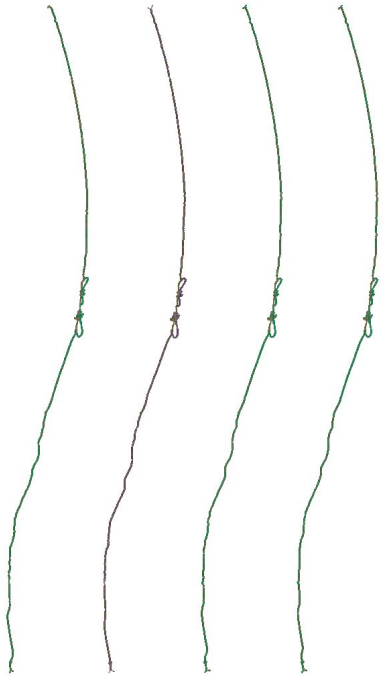


Relaxing kneed:1, w10000, GRCh38:92, CHM13:94

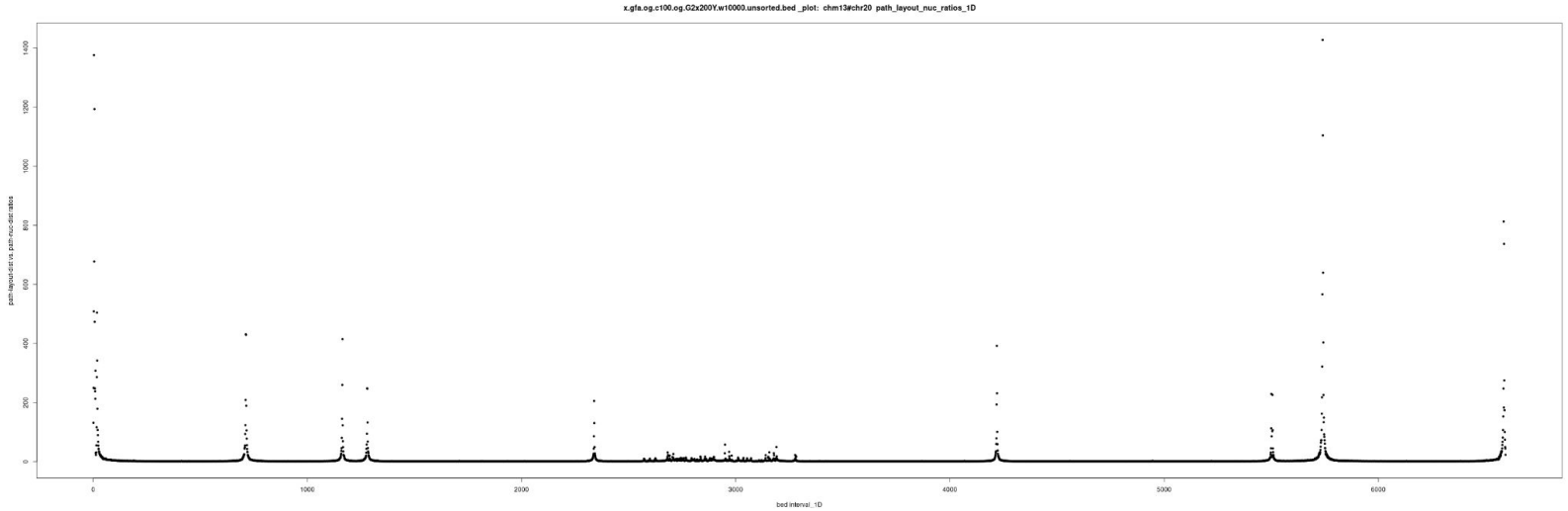


Finding a reasonable ***S*** or ratio cutoff

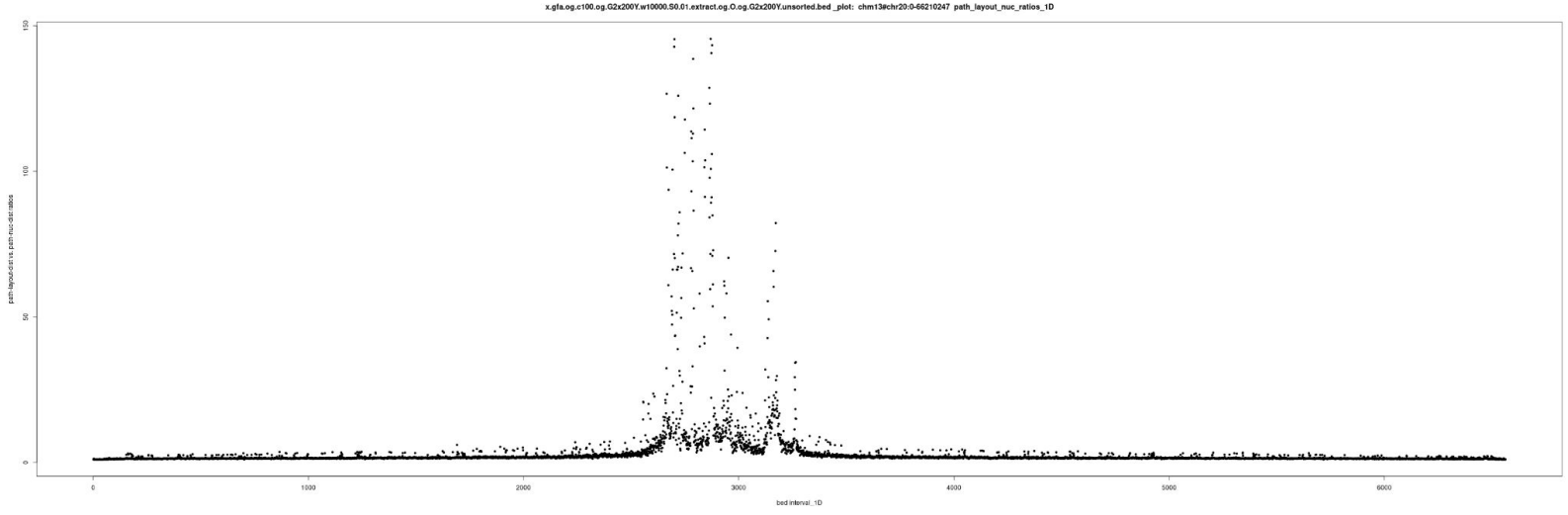
Relaxing - w10000 - S0.01,0.1,1,10



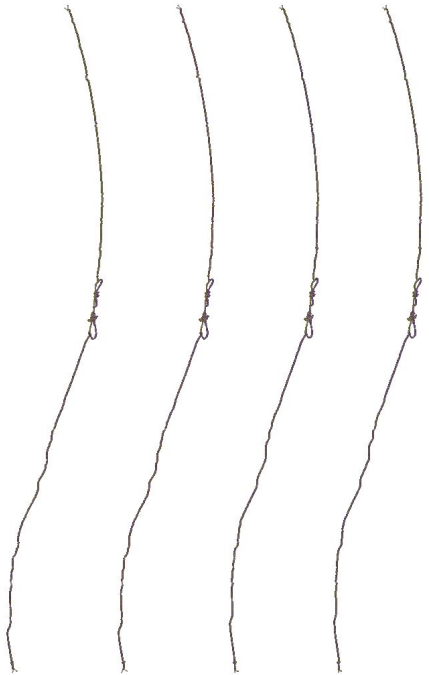
before relax



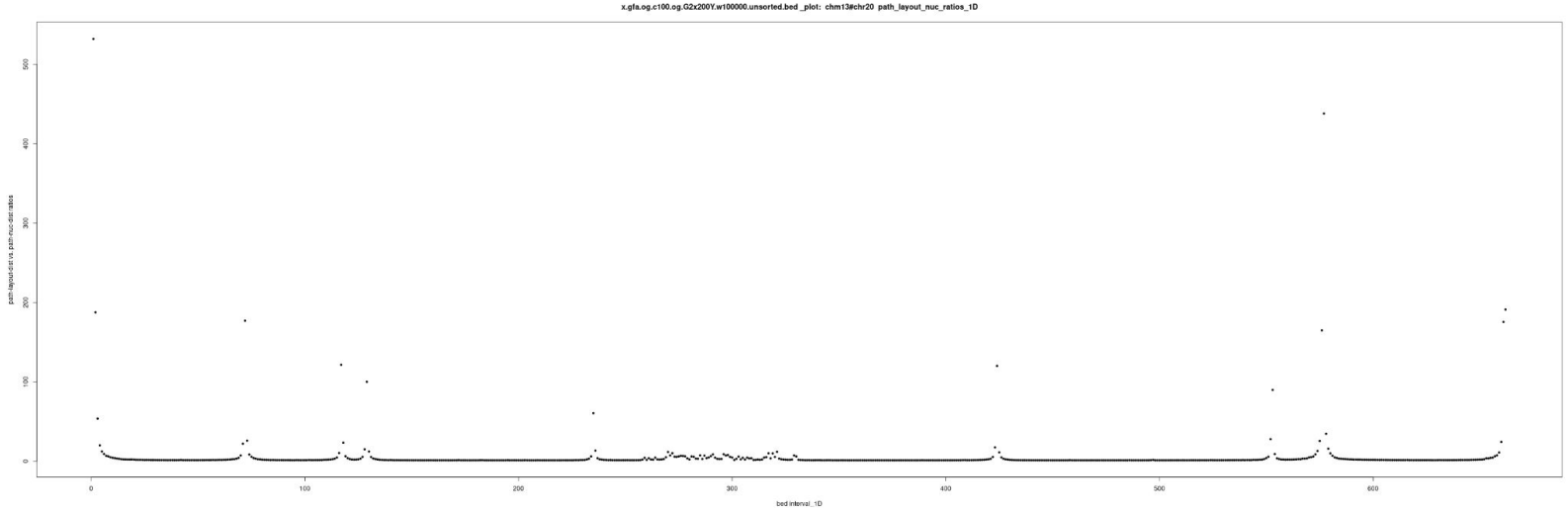
after relax



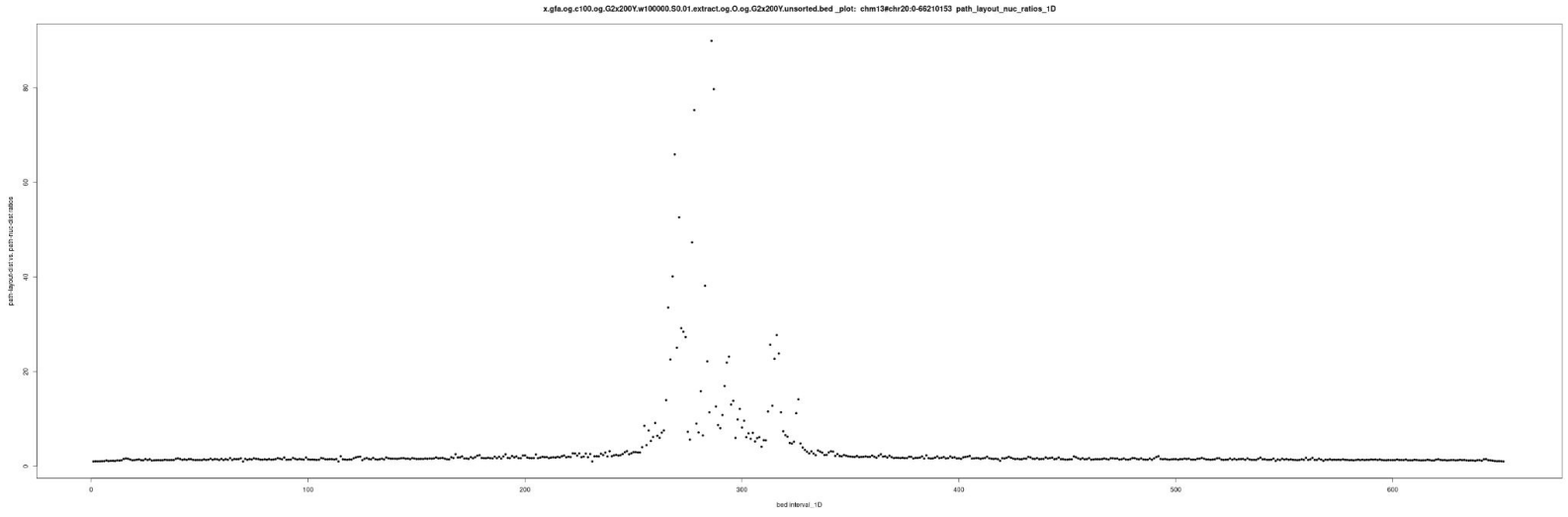
Relaxing - w100000 - S0.01,0.1,1,10



before relax



after relax



Acknowledgements

EU Pangenome Group

Erik Garrison

Andrea Guarracino

Pjotr Prins

Vincenza Colonna

Flavia Villani

David G. Ashbrook

Robert W.Williams

Christian Fischer

HPRC

Sven Nahnsen

Oliver Kohlbacher

Michael Krone

Gisela Gabernet

Friederike Hanssen

Antonia Schuster

HelmholtzZentrum münchen
German Research Center for Environmental Health

Lukas Heumos

Philipp Ehmele

Christian Kubica

Sebastian Vorbrugg

Jörg Hagmann

Jerven Bollemann

Toshiyuki T. Yokoyama

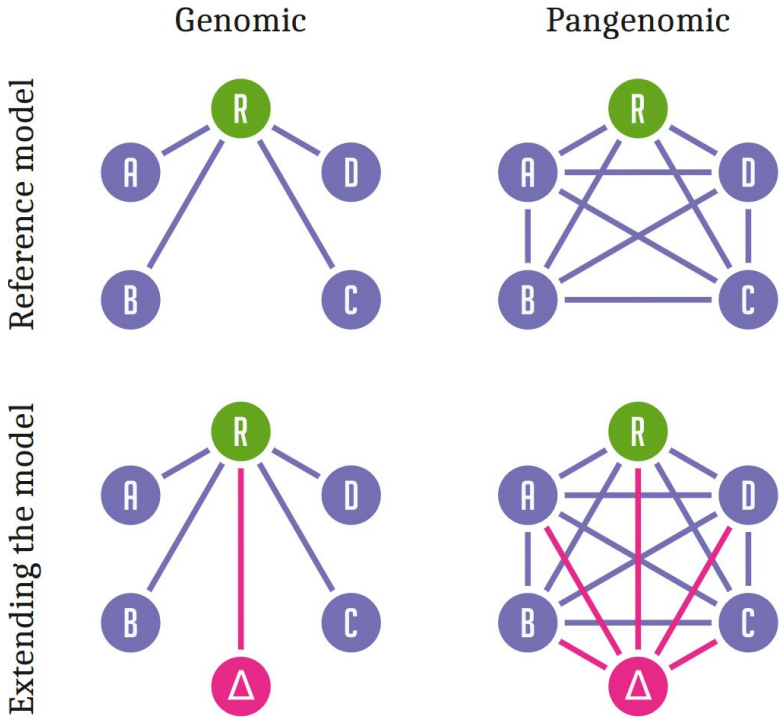
Torsten Pook

Franziska Huth

More slides with results could be

- Metrics to evaluate the 1D sort:
  - Explain with figures how our metrics work
  - Explain how the ALIBI metric works
- Performance comparison of 1D sorting algorithms:
  - The Flow Procedure
  - ALIBI
  - random
  - toposort
  - Pipeline of sorts
  - Reference-based sorting
- Performance of 2D layout: ~10 hours for a layout of a 90 haplotype chr8 human pangenome graph
  - Compare odgi layout + draw vs. Bandage
  - 2D ref based sorting
- Applications:
  - Graph linearization -> 1D viz, smoothxg, pggb
  - gfaestus
  - Pantograph, waragraph
  - Current research status with ODGI tension

De novo assembly and a pangenomic model

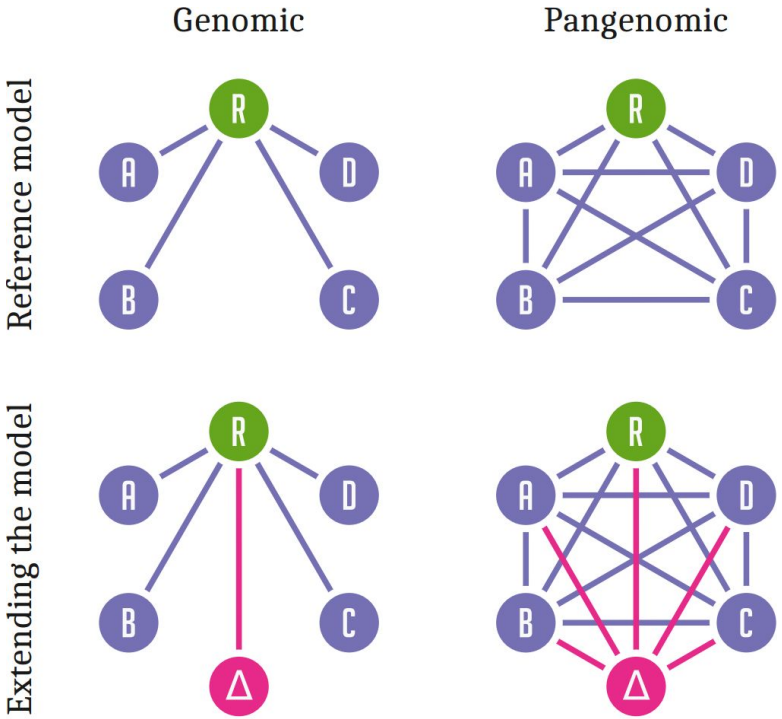


Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

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.

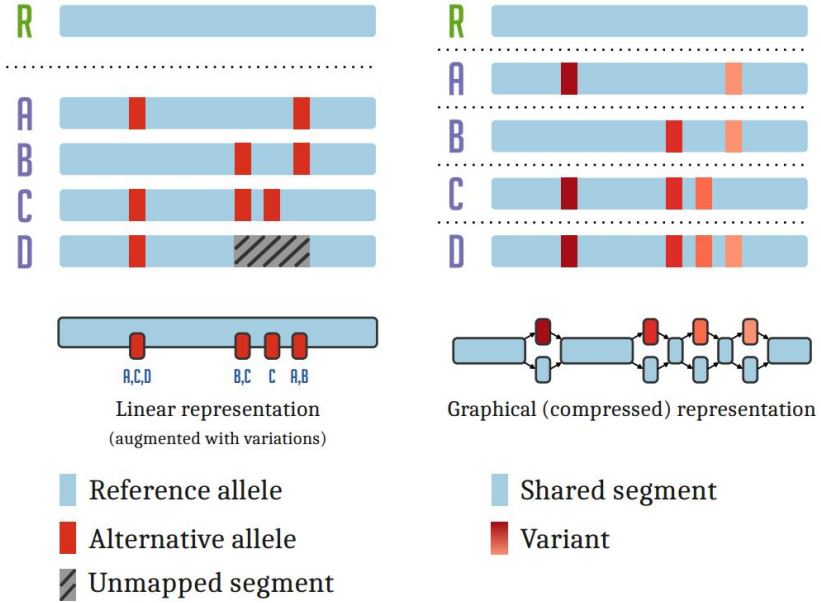


Figure from Eizenga et al., 2020.

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.

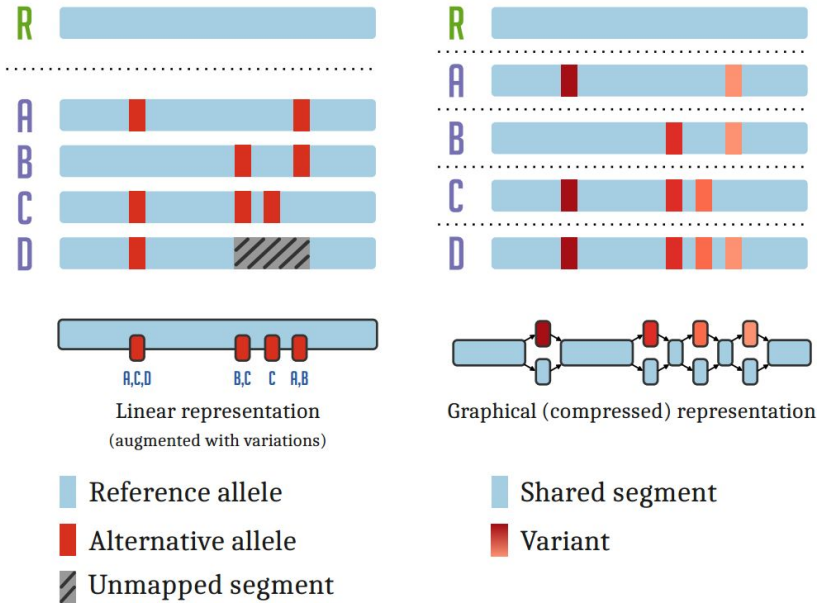
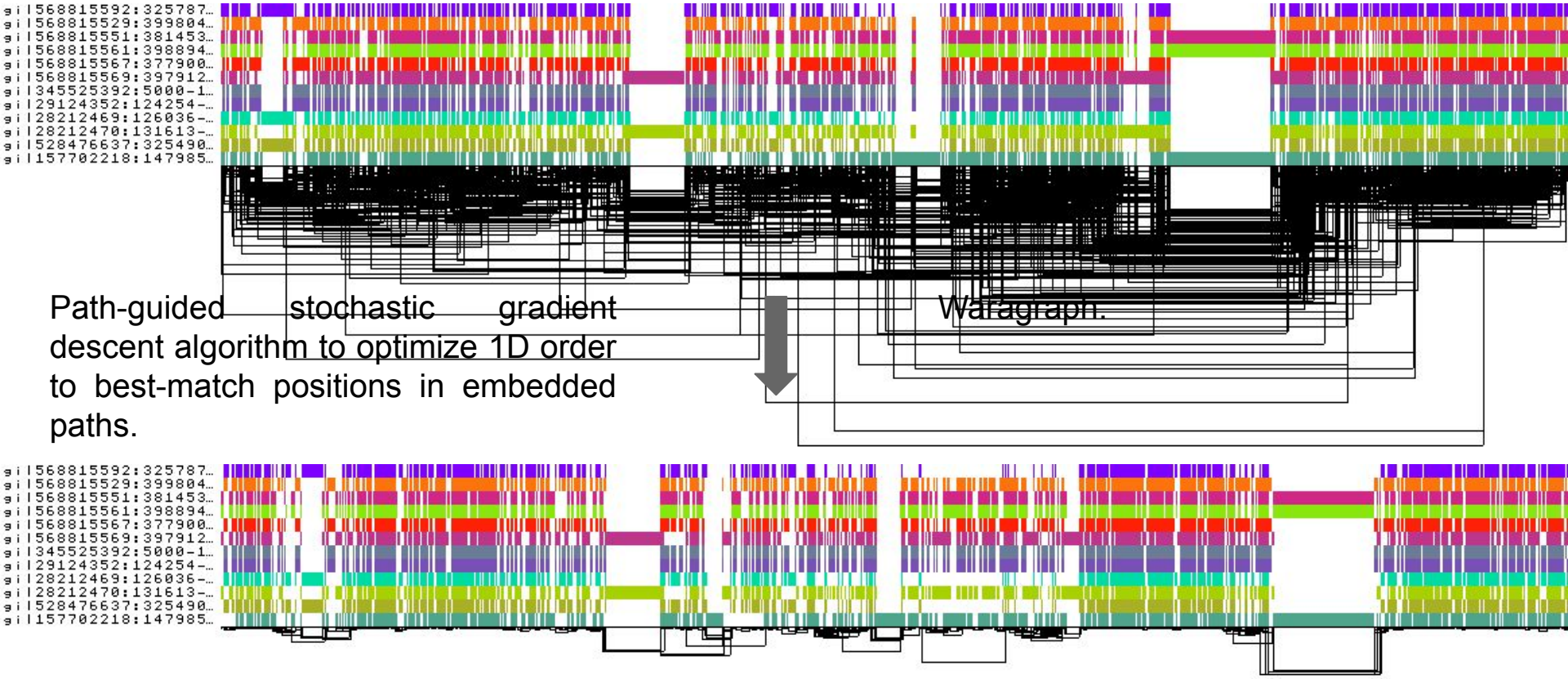


Figure from Eizenga et al., 2020.

1D Graph sorting by PG-SGD - Hogwild!

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

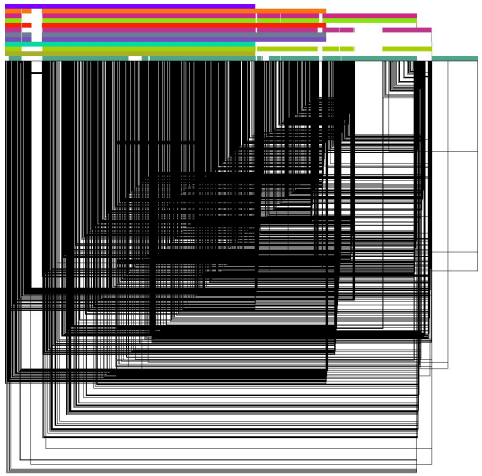


1D Graph Sorting and 2D Layouting by Path-Guided Stochastic Gradient Descent (PG-SGD)

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

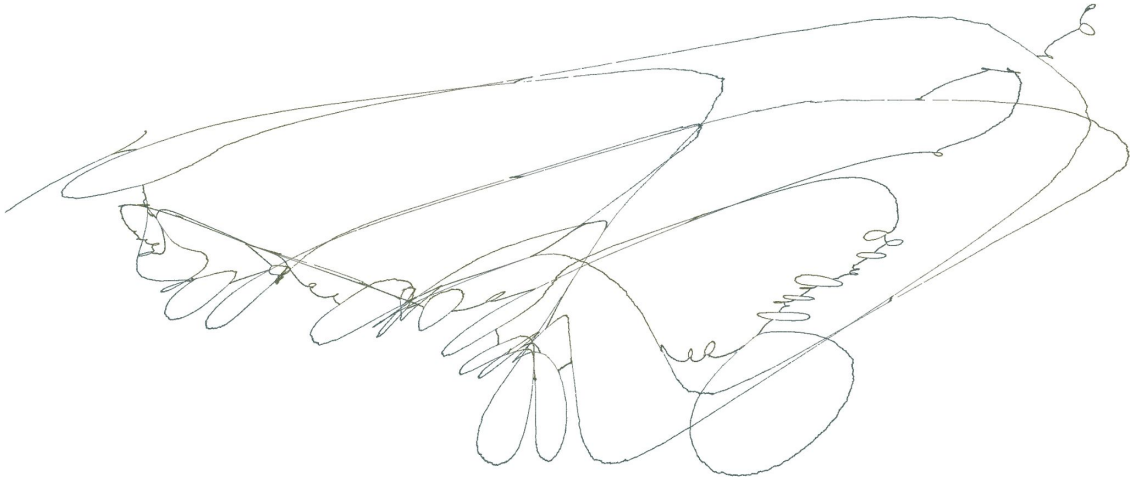


Raw graph built using seqwish.

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

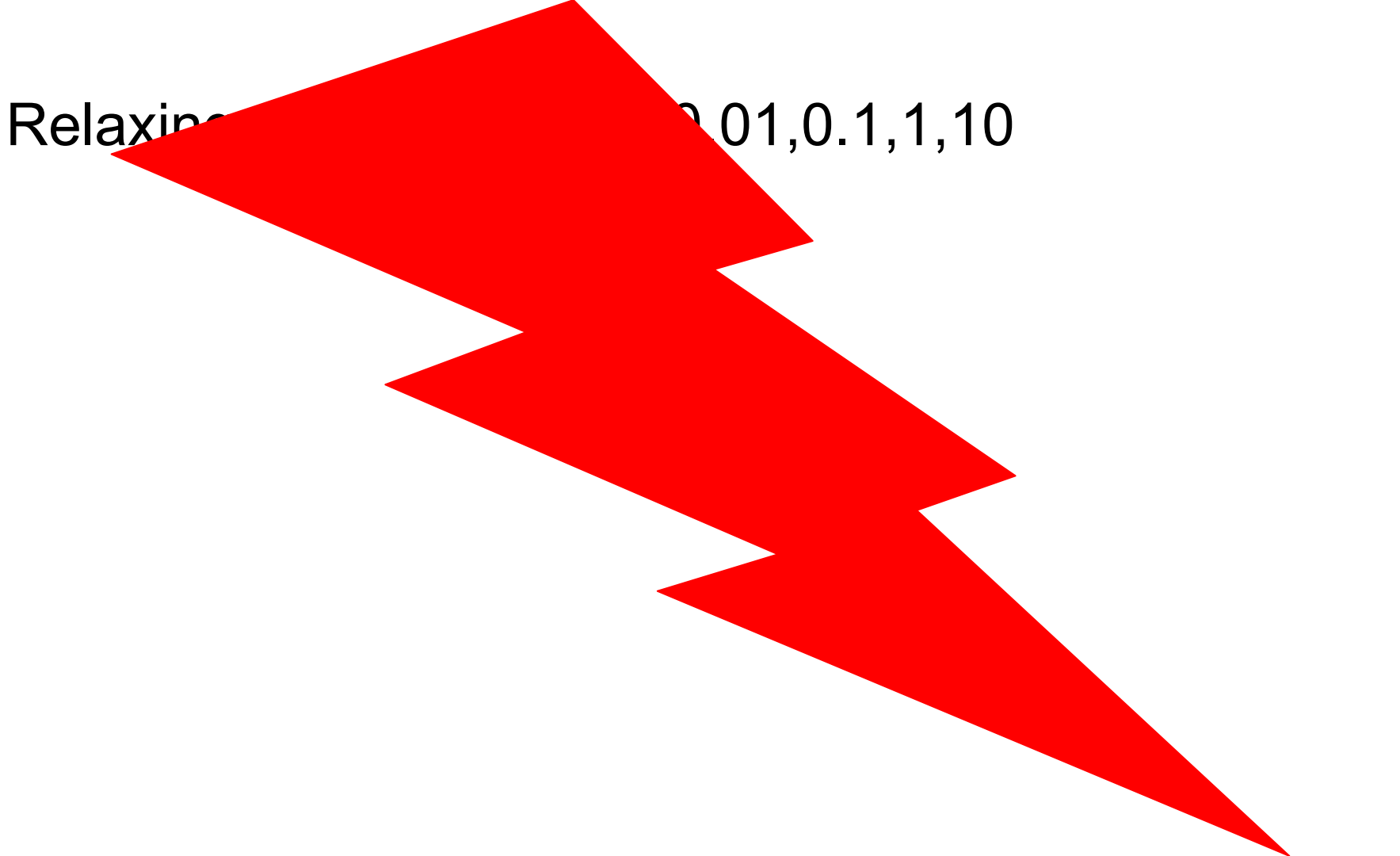
w	S	path	n	min_ratio
10000	0.01	chm13#chr20	25	247.89
10000	0.01	grch38#chr20	69	67.28
10000	0.1	chm13#chr20	25	247.89
10000	0.1	grch38#chr20	69	67.28
10000	1	chm13#chr20	94	49.25
10000	1	grch38#chr20	92	47.01
10000	10	chm13#chr20	223	16.40
10000	10	grch38#chr20	219	16.72
100000	0.01	chm13#chr20	16	25.81
100000	0.01	grch38#chr20	13	55.74
100000	0.1	chm13#chr20	16	25.81
100000	0.1	grch38#chr20	13	55.74
100000	1	chm13#chr20	25	13.27
100000	1	grch38#chr20	22	16.33
100000	10	chm13#chr20	25	13.27
100000	10	grch38#chr20	27	12.30

Vicia faba locus

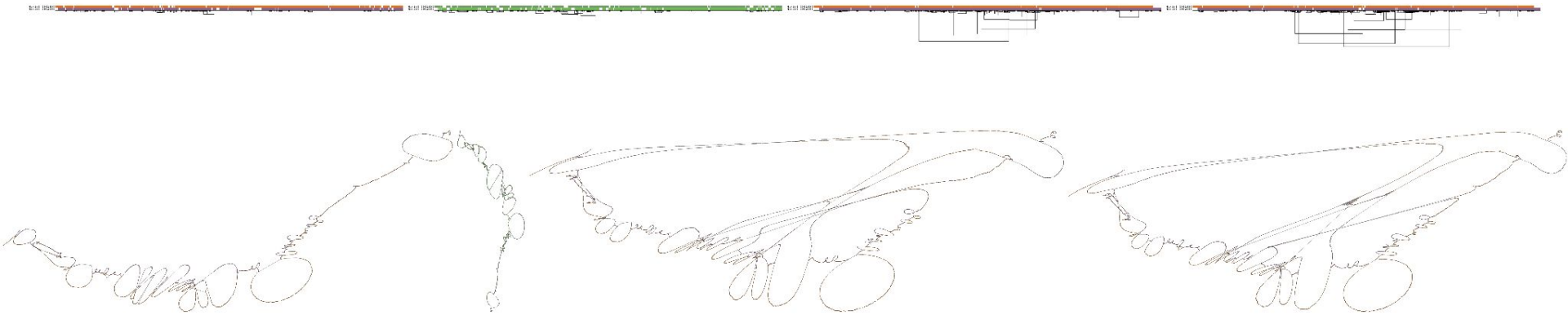


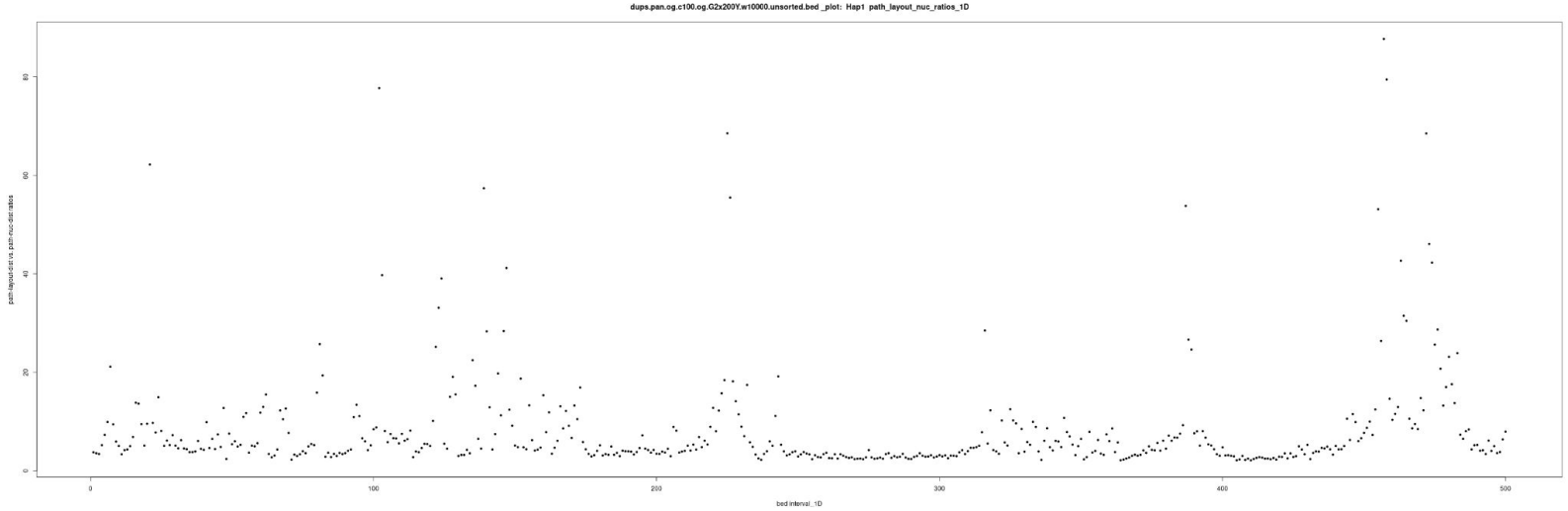
Relaxing - w100,1000 - S0.01,0.1,1,10

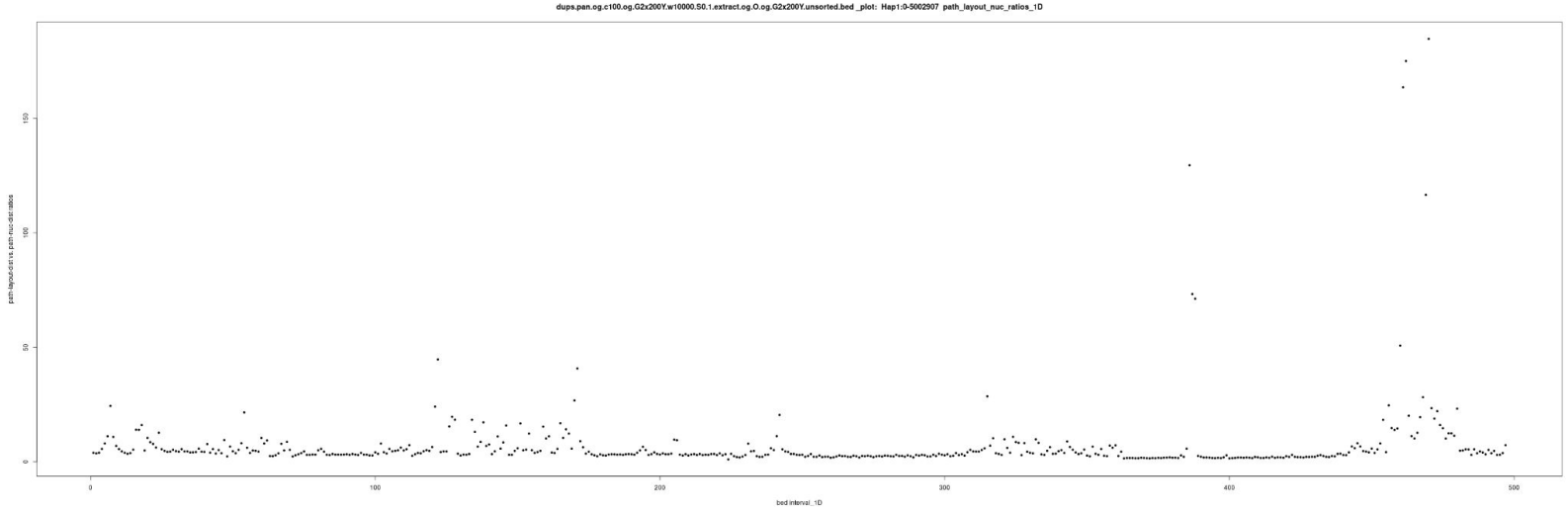
Relaxing 01,0.1,1,10



Relaxing - w10000 - S0.01,0.1,1,10

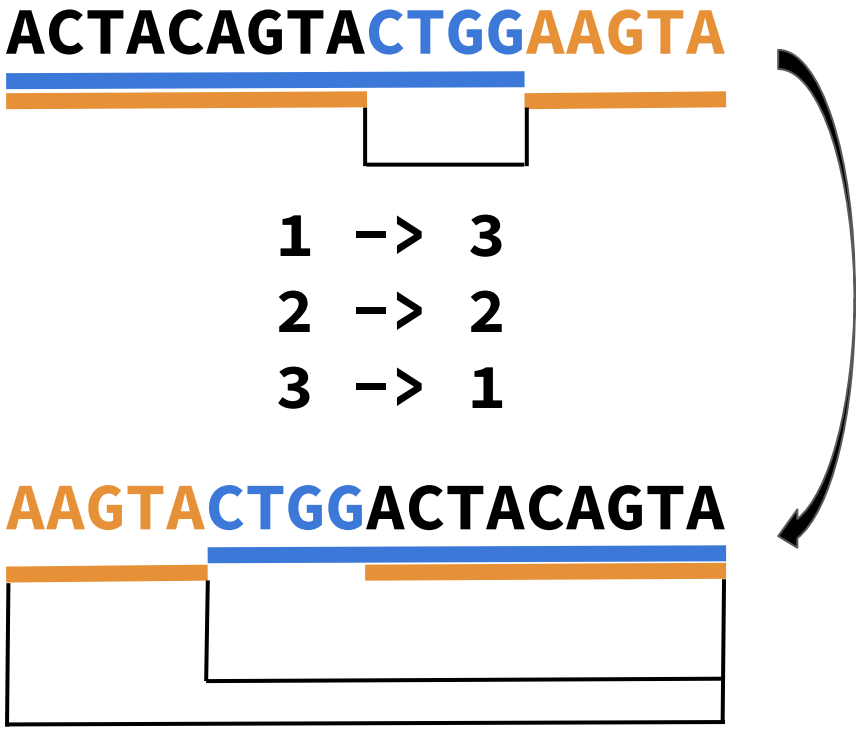



dups.pan.og.c100.og.G2x200Y.w10000.unsorted.bed _plot: Hap1 path_layout_nuc_ratios_1D
path-layout-dist vs. path-nuc-dist-ratios
bed interval_1D
0
20
40
60
80
0
100
200
300
400
500


dups.pan.og.c100.og.G2x200Y.w10000.S0.1.extract.og.O.og.G2x200Y.unsorted.bed _plot: Hap1:0-5002907 path_layout_nuc_ratios_1D
path-layout-dist vs. path-nuc-dist-ratios
bed interval_1D
0
50
100
150
0
100
200
300
400
500

Relaxing - w100000 - S0.01,0.1,1,10

Measure 1D sorting goodness




```
S    1    ACTACAGTA
S    2    CTGG
S    3    AAGTA
P    Genome1  1+,2+
P    Genome2  1+,3+
L    1    +    2    +
L    1    +    3    +
```

```
S    1    AAGTA
S    2    CTGG
S    3    ACTACAGTA
P    Genome1  3+,2+
P    Genome2  3+,1+
L    3    +    2    +
L    3    +    1    +
```


```
S    1    ACTACAGTA
S    2    CTGG
S    3    AAGTA
P    Genome1  1+,2+
P    Genome2  1+,3+
L    1    +    2    +
L    1    +    3    +
```

```
S    1    AAGTA
S    2    CTGG
S    3    ACTACAGTA
P    Genome1  3+,2+
P    Genome2  3+,1+
L    3    +    2    +
L    3    +    1    +
```