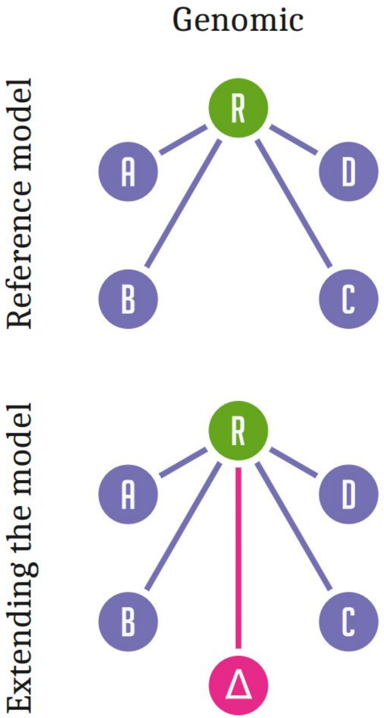
Pangenome Graphs

IBMI PhD Talks 2023
22 February 2023

Simon Heumos

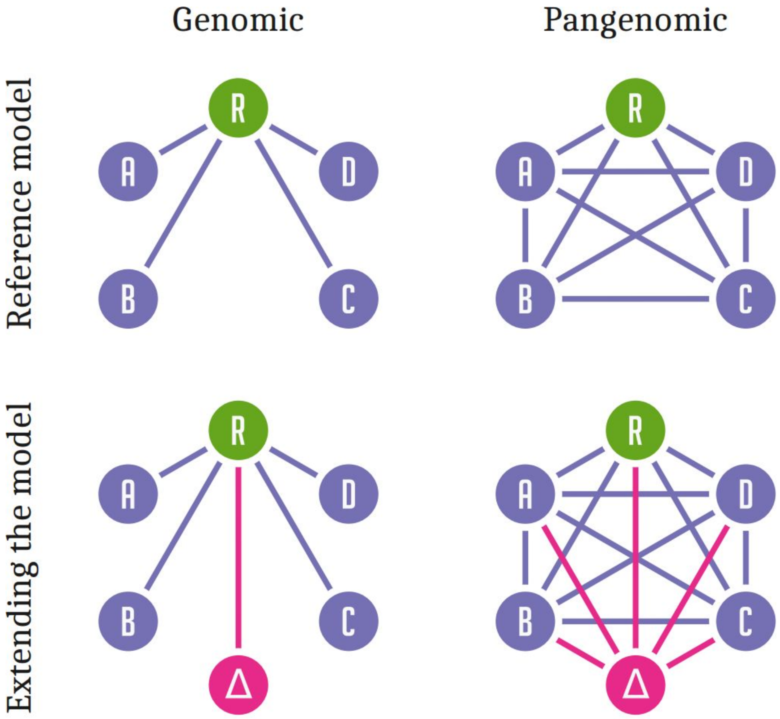
De novo assembly and a pangenomic model



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.

De novo assembly and a pangenomic model

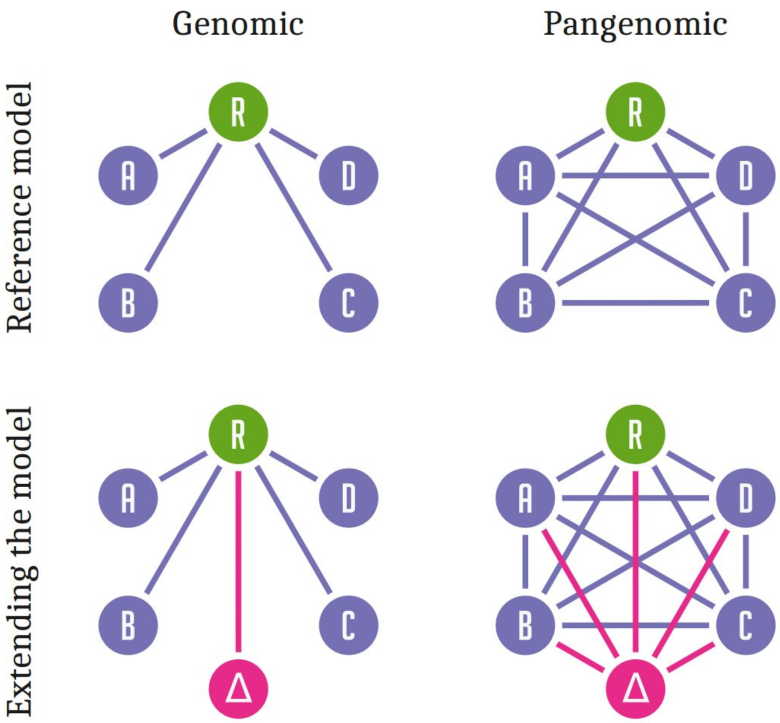


Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

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.

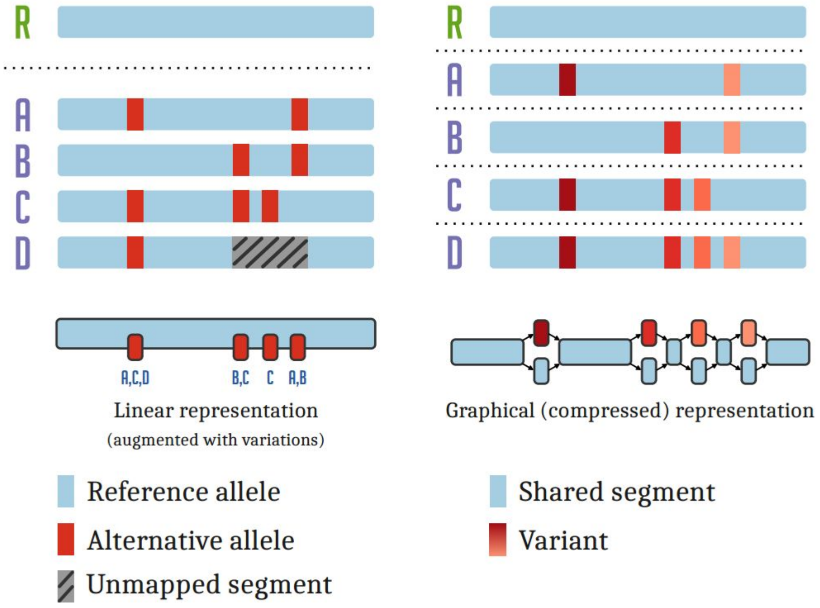


Figure from Eizenga et al., 2020.

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.

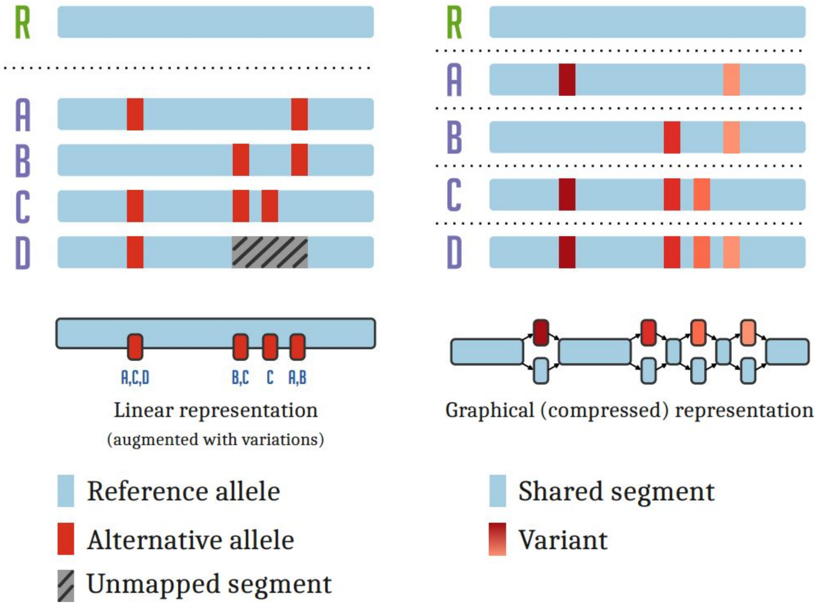


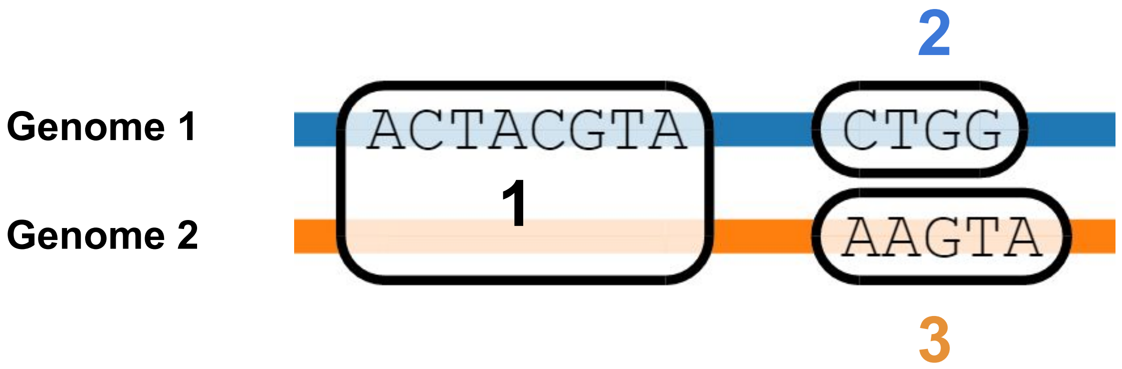
Figure from Eizenga et al., 2020.

Variation graphs

🟦 Genome 1: **ACTACGTACTGG** Path: **1 2**

🟧 Genome 2: **ACTACGTAAAGTA** Path: **1 3**

Linear sequences are **paths** through nodes.



Graph topology is not directly shown.

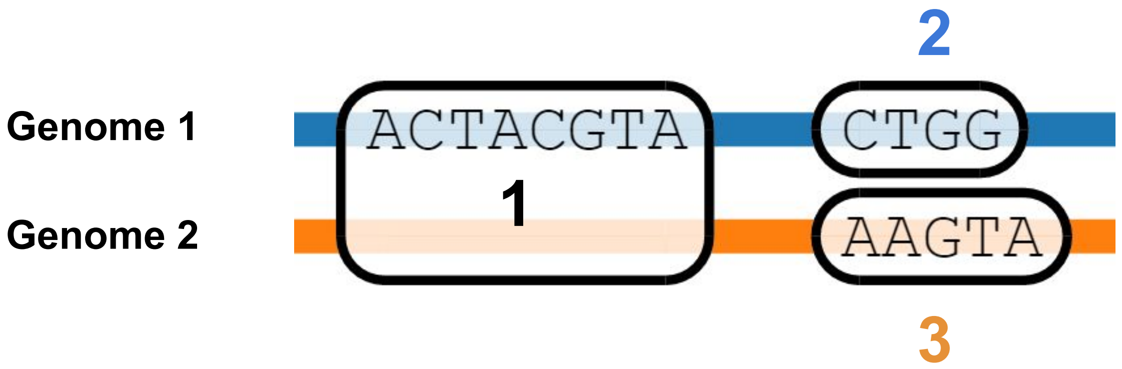
The nodes represent DNA sequences.

Sketch made using SequenceTubeMap.

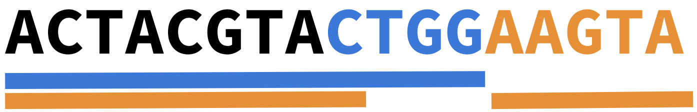
Paths can be contigs, haplotypes, reads, or whole chromosomes.

Towards a 1D visualization

- Genome 1: **ACTACGTACTGG** Path: **1 2**
- Genome 2: **ACTACGTAAAGTA** Path: **1 3**



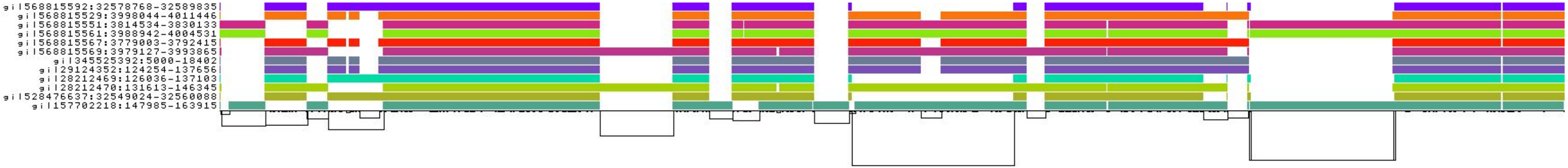
Concatenate nucleotides to a pangenome sequence.



Presence - absence matrix encodes actual genomic sequence.

1D Graph visualization explained

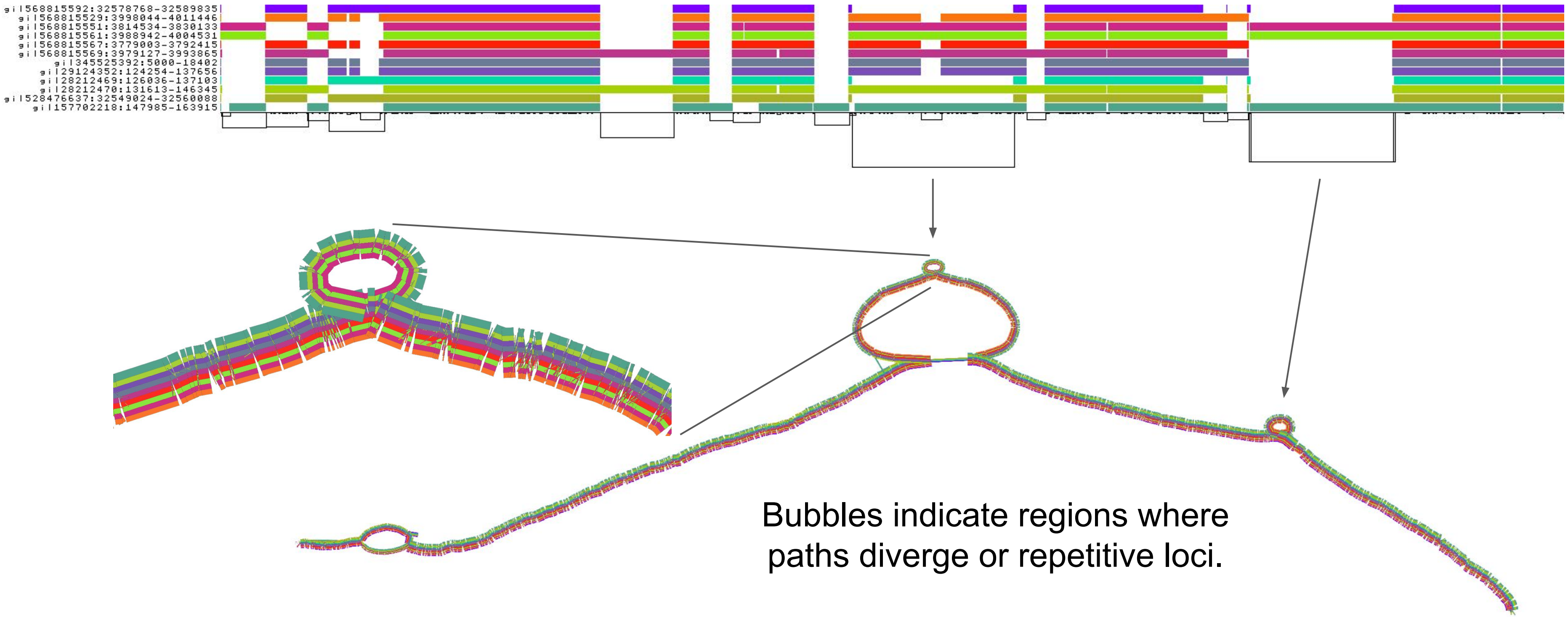
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.


gi|568815592:32578768-32589835
gi|568815529:3998044-4011446
gi|568815551:3814534-3830133
gi|568815561:3988942-4004531
gi|568815567:3779003-3792415
gi|568815569:3979127-3993865
gi|345525392:5000-18402
gi|29124352:124254-137656
gi|28212469:126036-137103
gi|28212470:131613-146345
gi|528476637:32549024-32560088
gi|157702218:147985-163915

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

2D Graph visualization explained

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


gi|568815592:32578768-32589835
gi|568815529:3998044-4011446
gi|568815551:3814534-3830133
gi|568815561:3988942-4004531
gi|568815567:3779003-3792415
gi|568815569:3979127-3993865
gi|345525392:5000-18402
gi|29124352:124254-137656
gi|28212469:126036-137103
gi|28212470:131613-146345
gi|528476637:32549024-32560088
gi|157702218:147985-163915

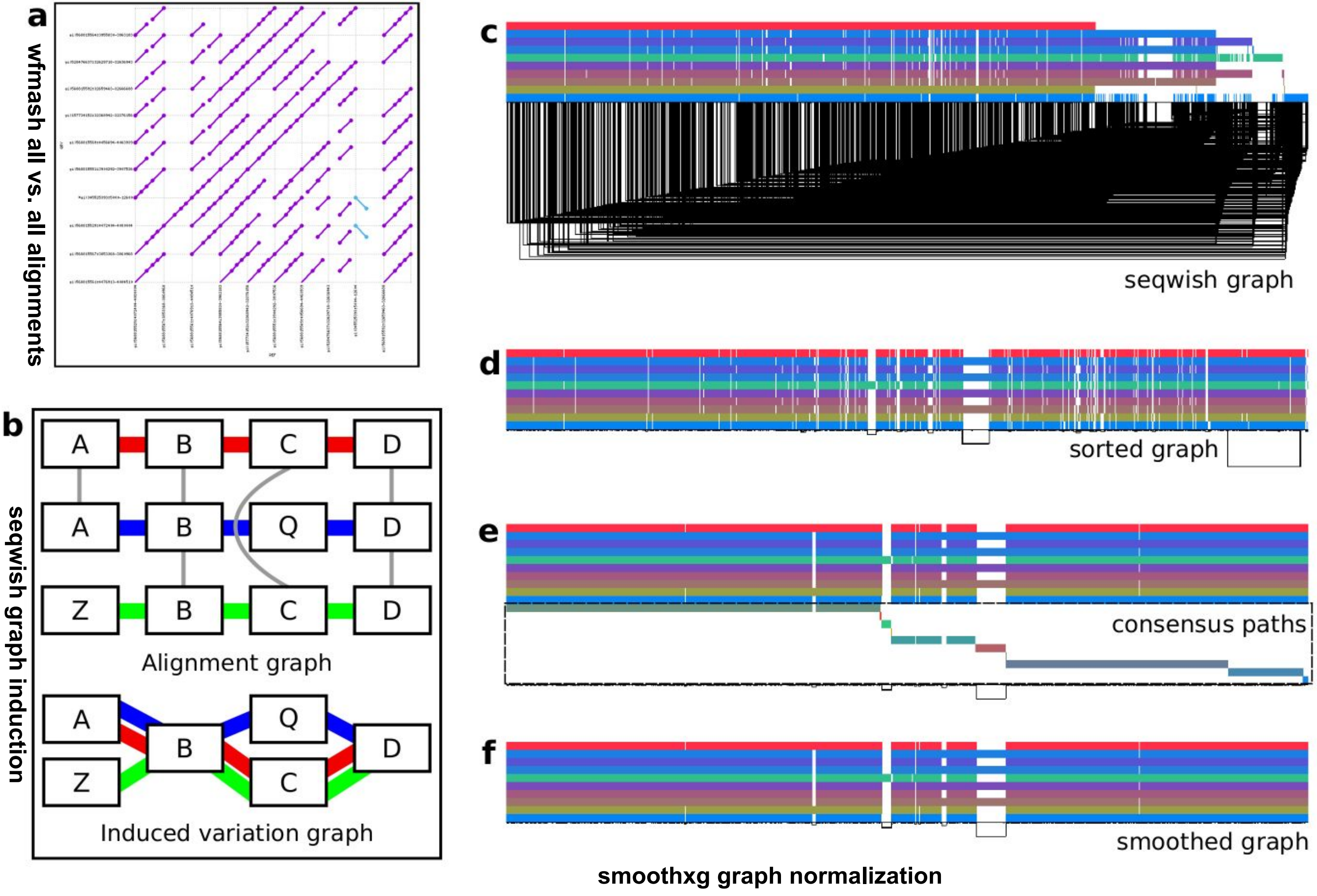
Bubbles indicate regions where paths diverge or repetitive loci.

Building Pangenome Graphs

Solving the whole genome alignment problem in 3 steps.

a) all-to-all alignment
b) graph induction
c-f) normalization

implemented in the PanGenome Graph Builder (PGGB)



Erik Garrison

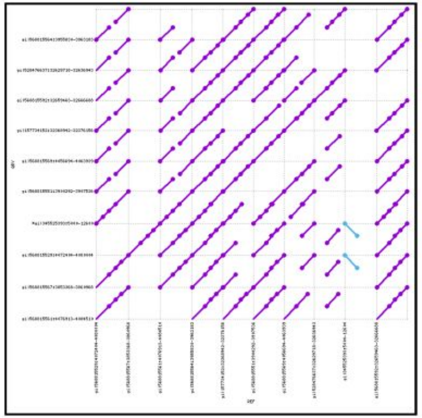
Andrea Guarracino

Building Pangenome Graphs

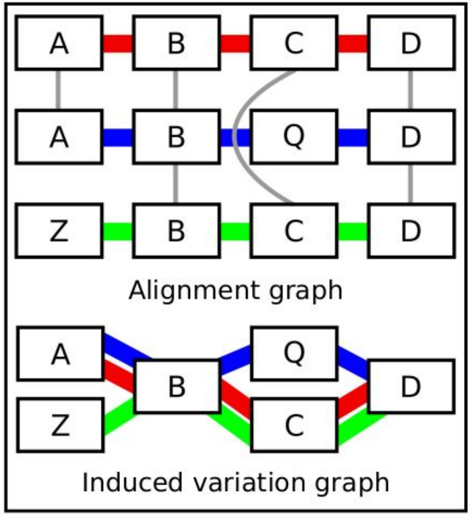
Solving the whole genome alignment problem in 3 steps.

a) all-to-all alignment
b) graph induction
c-f) normalization

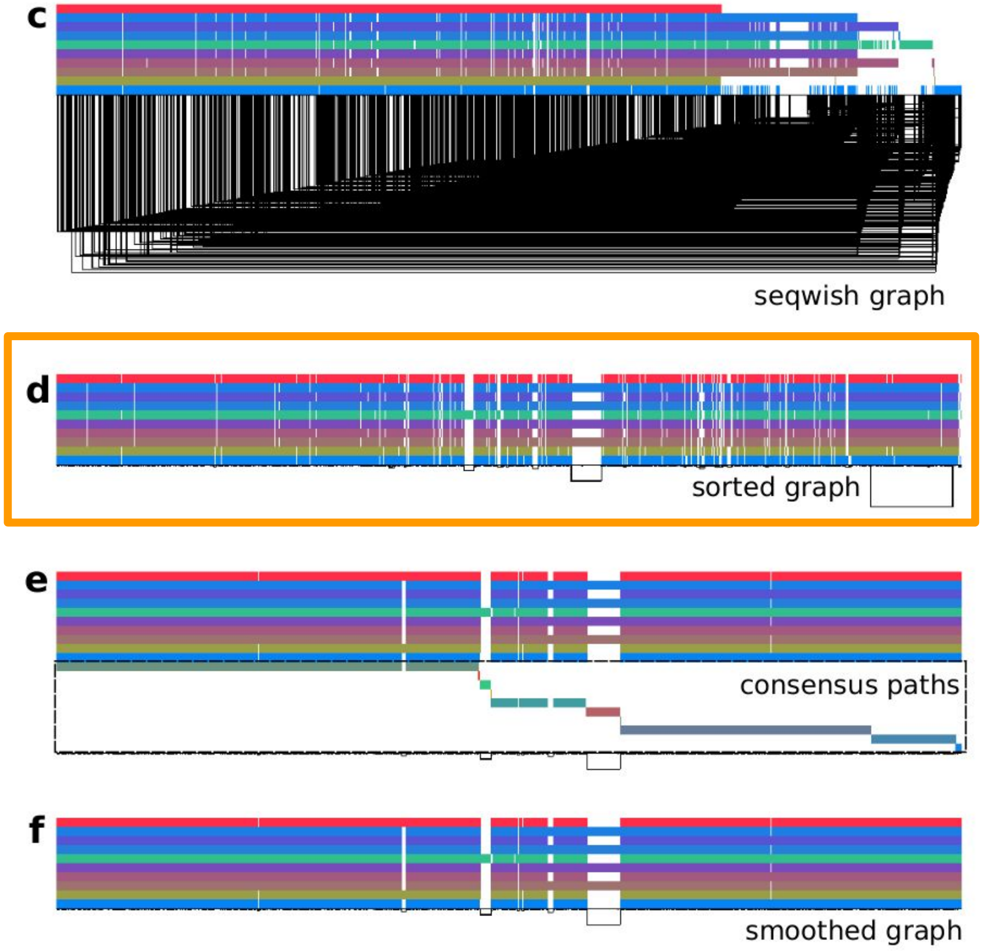
implemented in the PanGenome Graph Builder (PGGB)



wfmash all vs. all alignments



seqwish graph induction



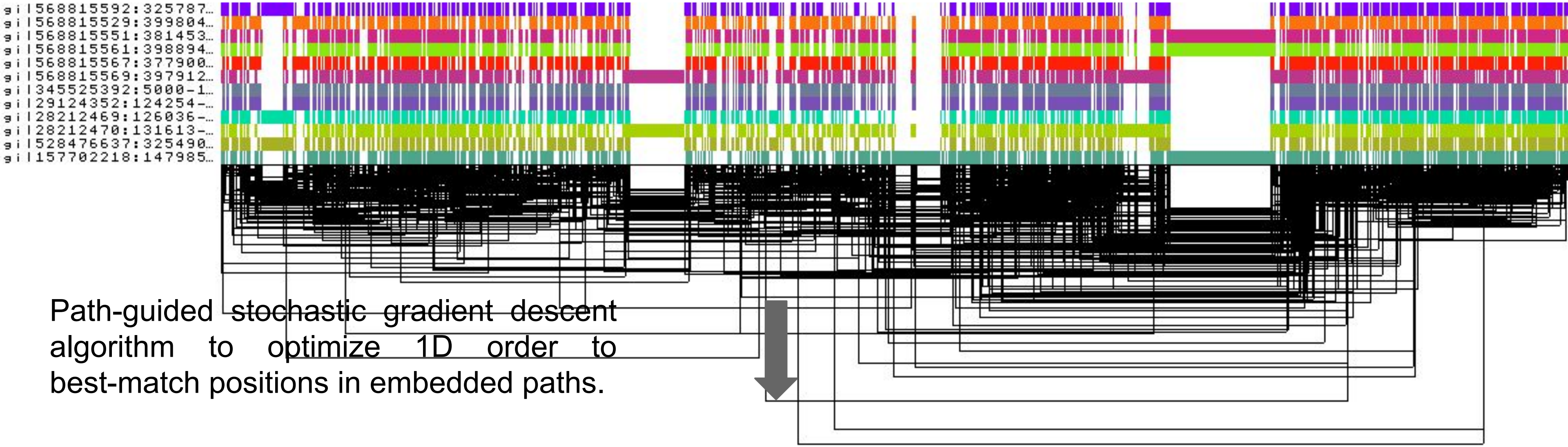
smoothxg graph normalization

Erik Garrison

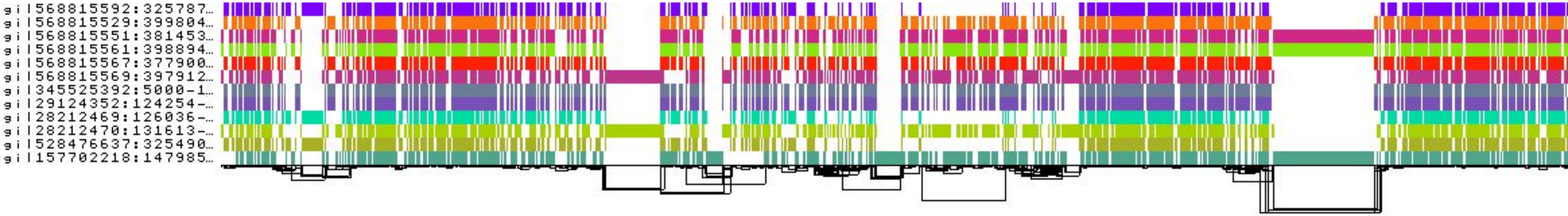
Andrea Guarracino

Path-Guided Stochastic Gradient Descent (PG-SGD) in 1D

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

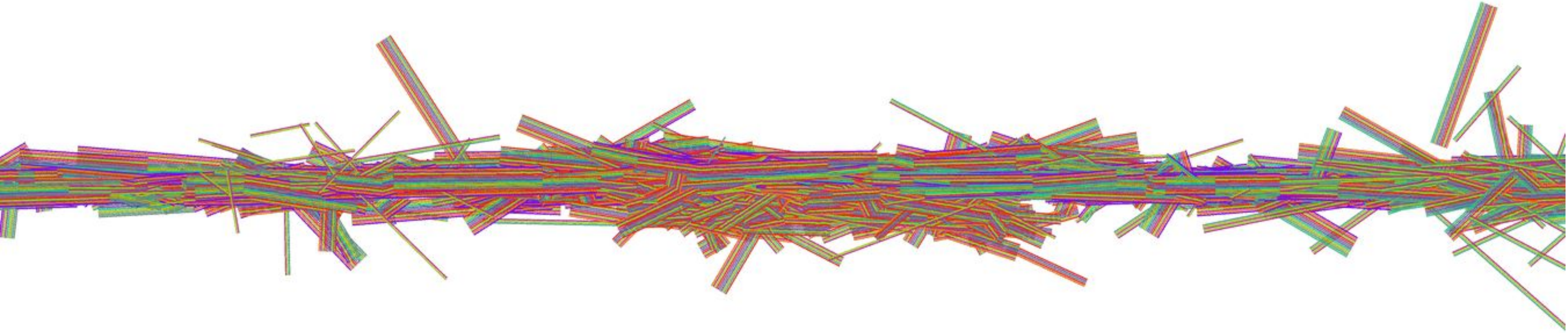


Path-guided stochastic gradient descent algorithm to optimize 1D order to best-match positions in embedded paths.



Bonus: 2D Graph layout by PG-SGD

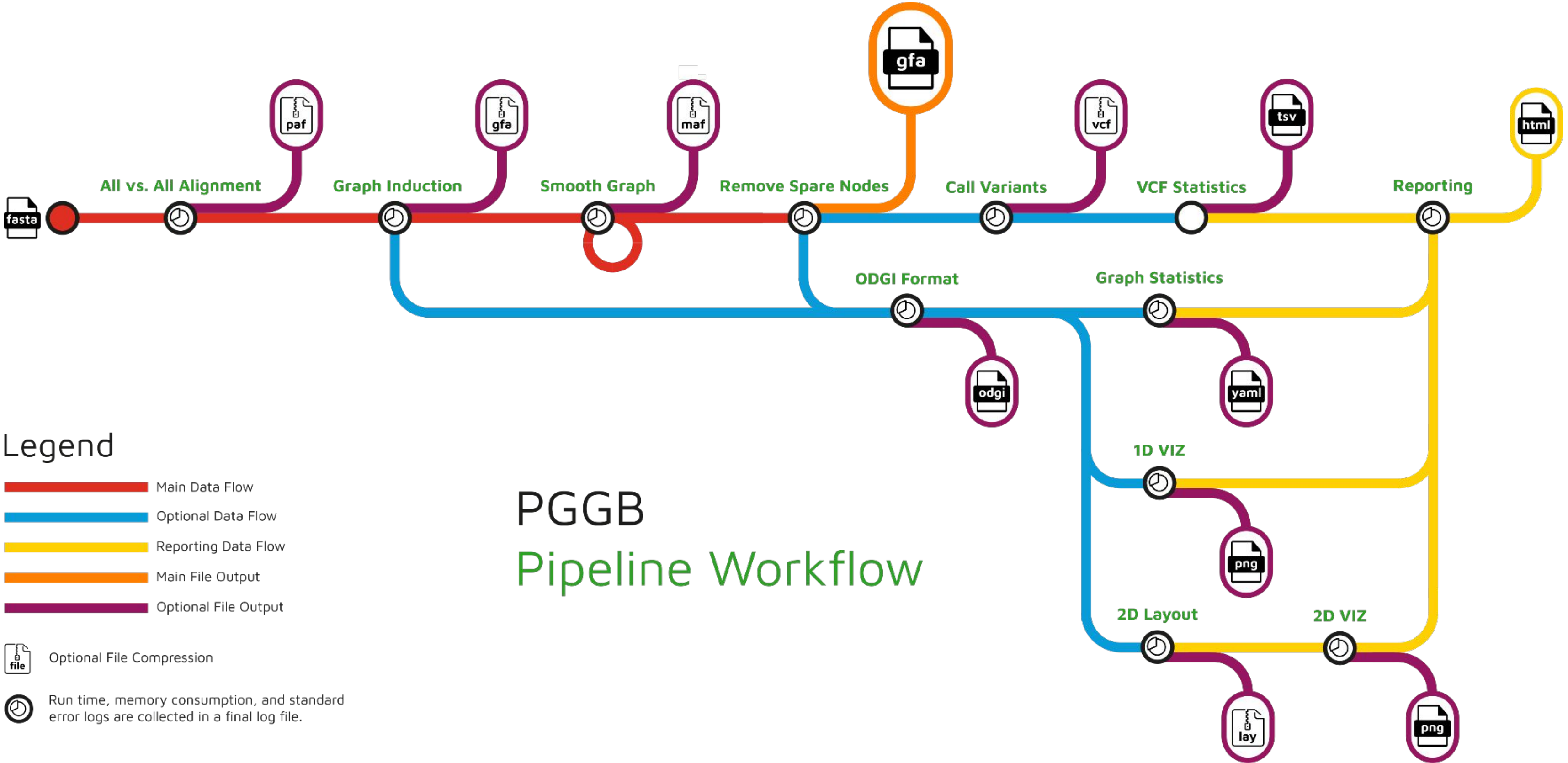
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



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.

The PanGenome Graph Builder (PGGB) - overview



fasta

All vs. All Alignment

paf

Graph Induction

gfa

Smooth Graph

maf

Remove Spare Nodes

gfa

Call Variants

vcf

VCF Statistics

tsv

Reporting

html

ODGI Format

odgi

Graph Statistics

yaml

1D VIZ

png

2D Layout

lay

2D VIZ

png

PGGB

Pipeline Workflow

Legend

🟥 Main Data Flow

🟦 Optional Data Flow

🟨 Reporting Data Flow

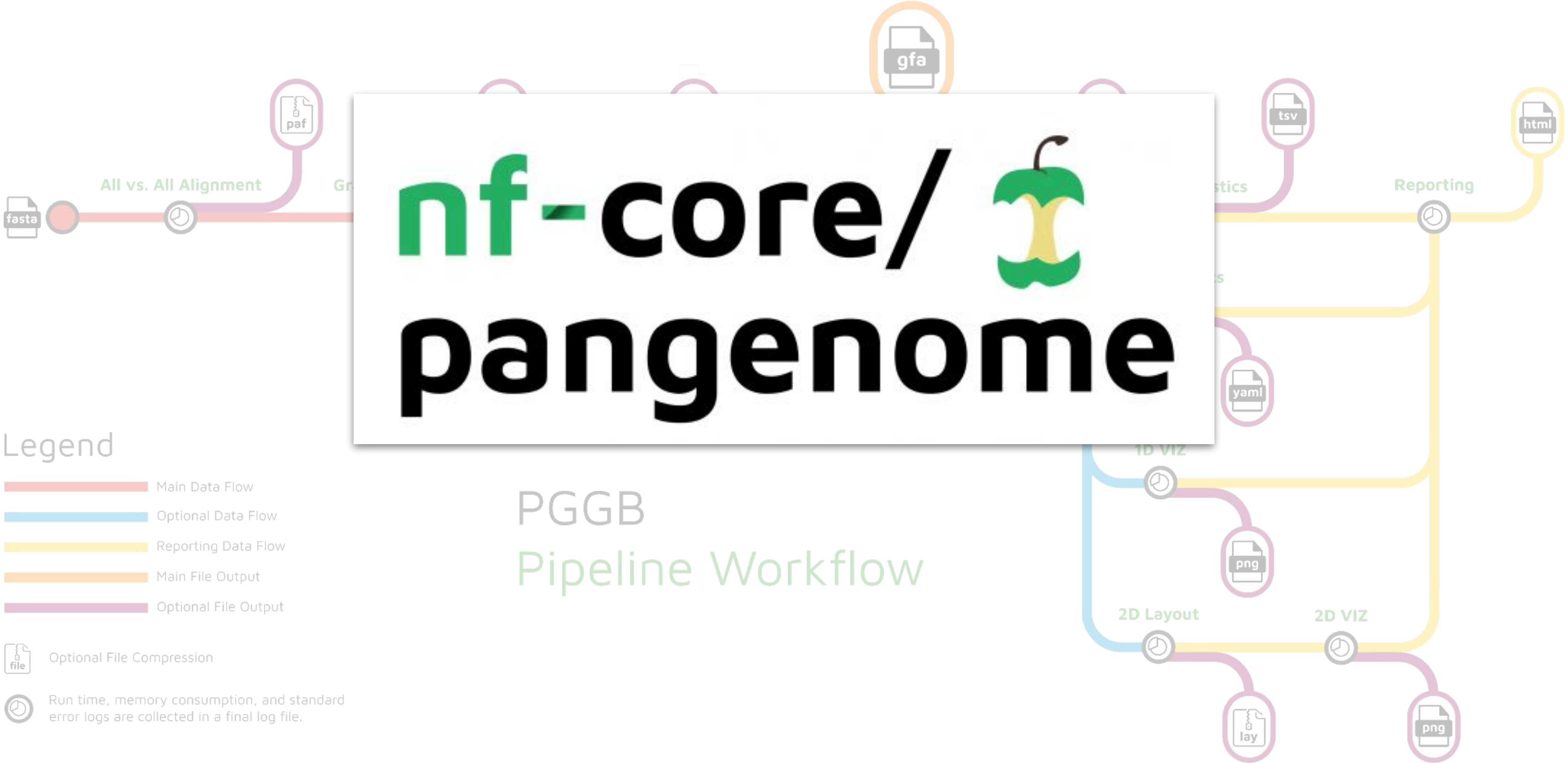
🟧 Main File Output

🟪 Optional File Output

file Optional File Compression

◷ Run time, memory consumption, and standard error logs are collected in a final log file.

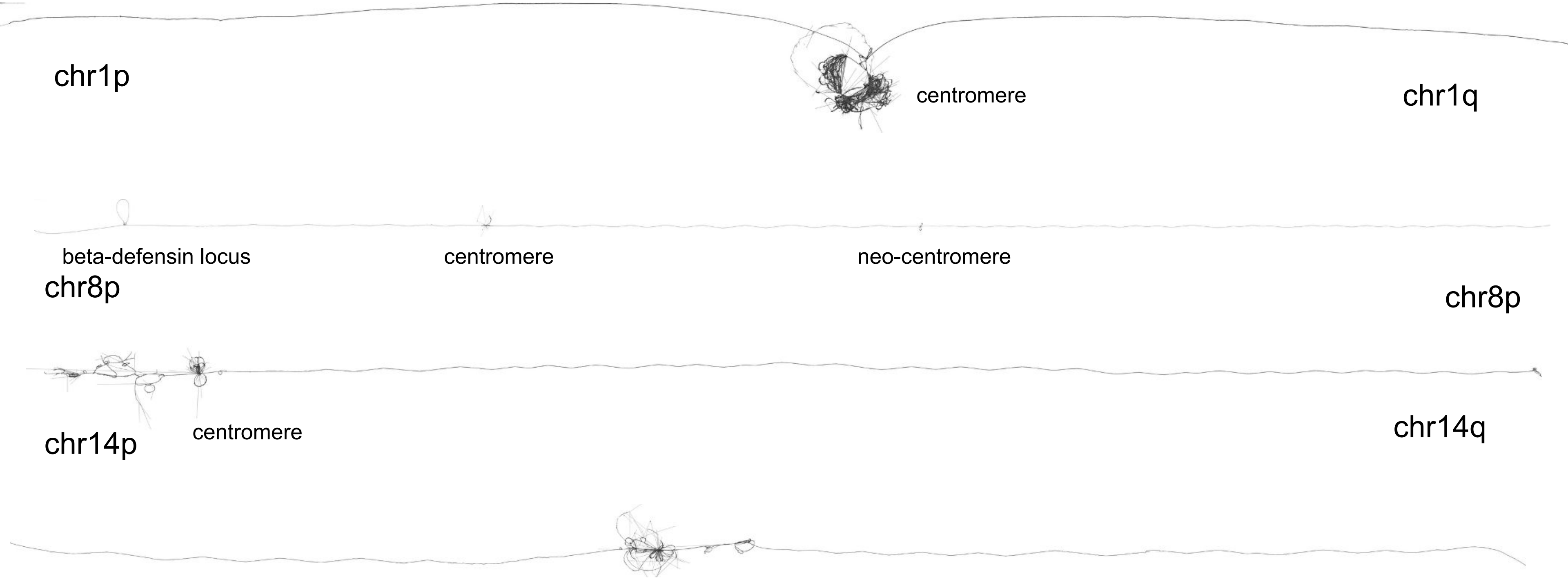
The PanGenome Graph Builder (PGGB) - overview



A 90 Haplotypes Human Pangenome Graph

Erik Garrison

(learned 2D visualization of PGGB HPRC chromosomes)



chr1p

centromere

chr1q

beta-defensin locus

centromere

neo-centromere

chr8p

chr8p

chr14p

centromere

chr14q

chr20p

centromere

chr20q

ZIM
Zentrales
Innovationsprogramm
Mittelstand

Acknowledgements

THE UNIVERSITY OF TENNESSEE
HEALTH SCIENCE CENTER.

Erik Garrison

Andrea Guarracino

Pjotr Prins

Vincenza Colonna

Flavia Villani

David G. Ashbrook

Robert W.Williams

Christian Fischer

HUMAN PANGENOME

HPRC

QBiC

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sven Nahnsen

Oliver Kohlbacher

Michael Krone

Gisela Gabernet

Friederike Hanssen

Antonia Schuster

de.NBI

HelmholtzZentrum münchen
German Research Center for Environmental Health

Lukas Heumos

Philipp Ehmele

MAX PLANCK
CAMPUS TÜBINGEN

Christian Kubica

Sebastian Vorbrugg

computomics®
molecular data analysis

Jörg Hagmann

SIB
Swiss Institute of
Bioinformatics

Jerven Bollemann

CBMS
Computational Biology and Medical Sciences

Toshiyuki T. Yokoyama

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Torsten Pook

Universität Stuttgart

Franziska Huth

Visualizing MHC and C4 pangenome graphs.

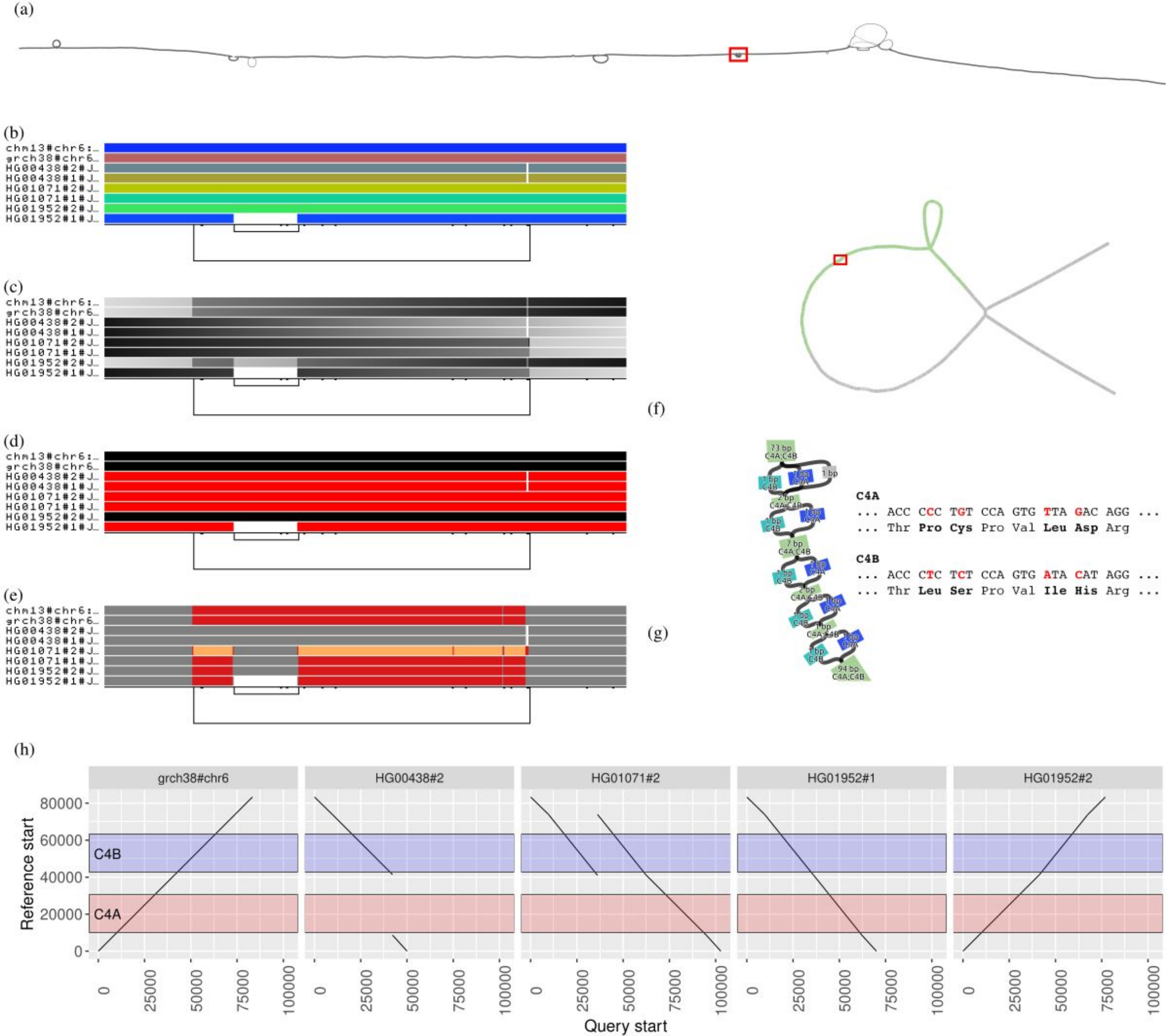
(a) `odgi extract`, `odgi layout`, and `odgi draw`

(b)-(e) `odgi viz` of 8 paths

(f) `Bandage` layout annotated with `odgi position` in green including the HERV sequence

(g) Annotated `Bandage` layout indicating single nucleotide differences in C4A and C4B

(h) `odgi untangle` output showing copy number state with respect to CHM13



Features of a 90-haplotype human pangenome graph of exon 1 of the huntingtin gene (*HTTexon1).*

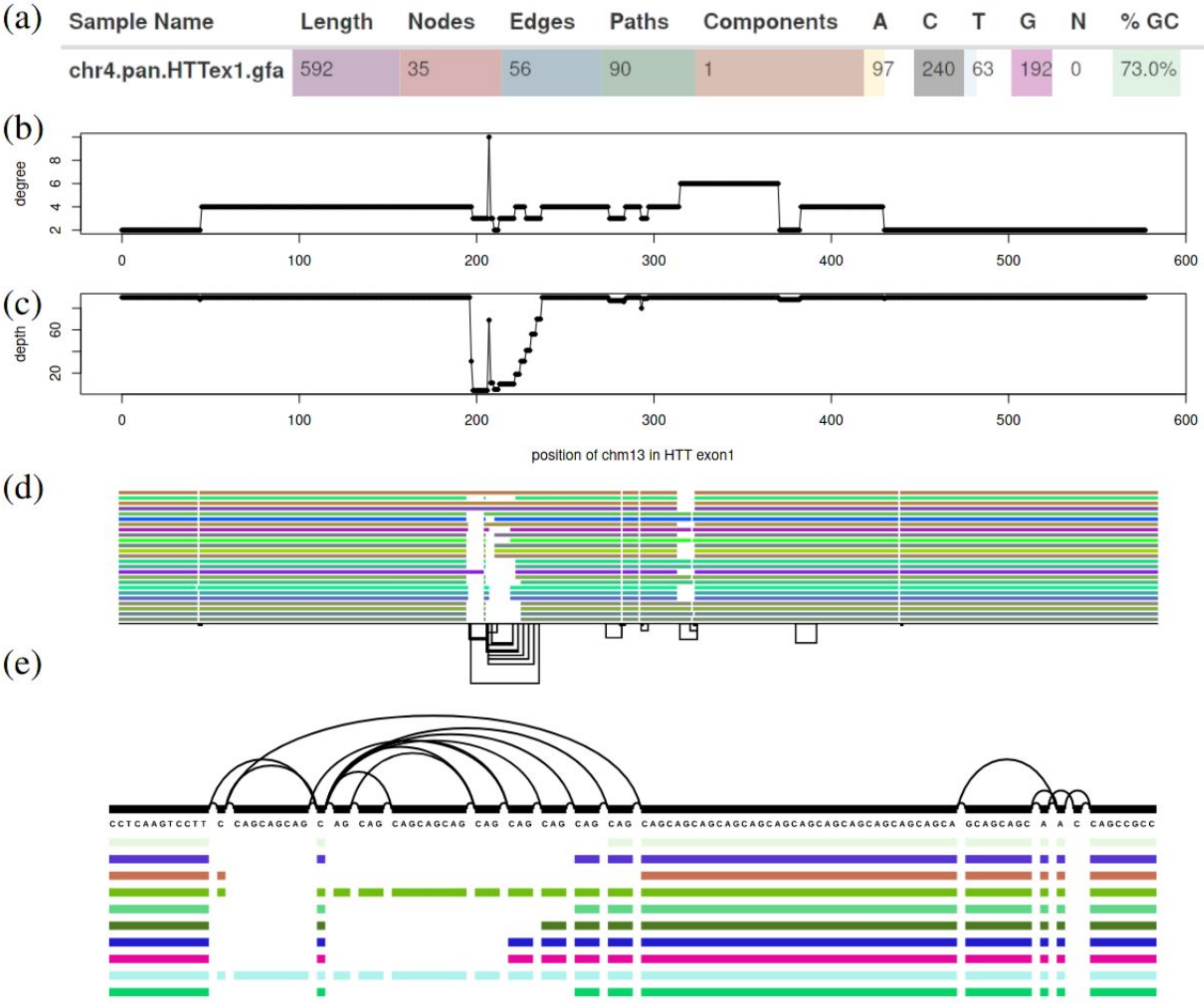
(a) ODGI `MultiQC` report excerpt

(b) Node degree distribution of CHM13 in the graph

(c) Node depth distribution of CHM13 in the graph

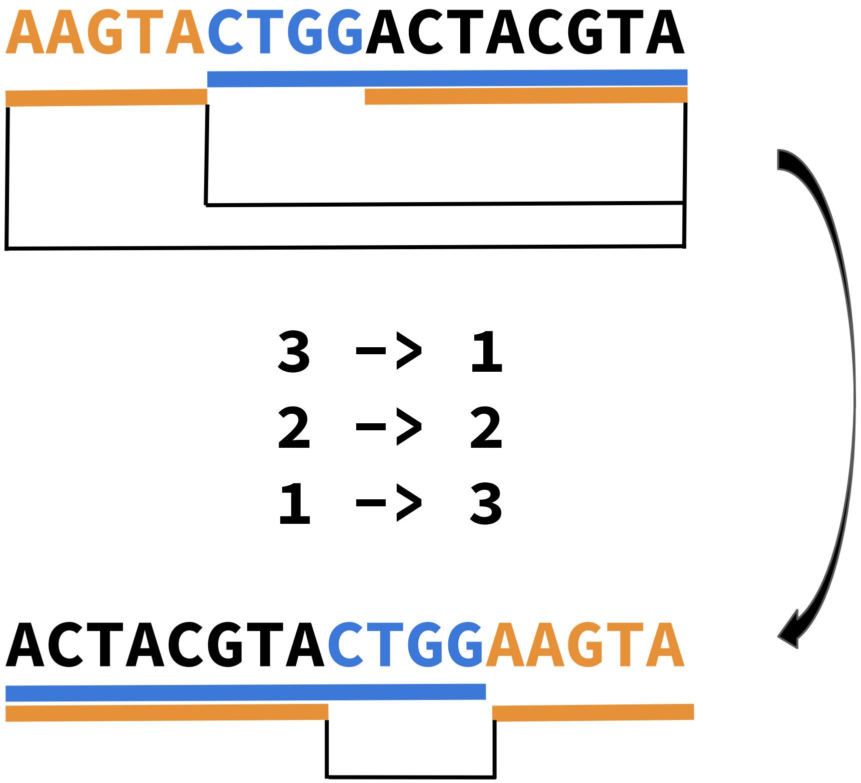
(d) `odgi viz` – 23 largest gene alleles

(e) `vg viz` – nucleotide level zoomed in on 10 largest gene alleles



Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N	% GC
chr4.pan.HTTex1.gfa	592	35	56	90	1	97	240	63	192	0	73.0%

1D Sorting process explained



AAGTACTGGACTACGTA

3 -> 1
2 -> 2
1 -> 3

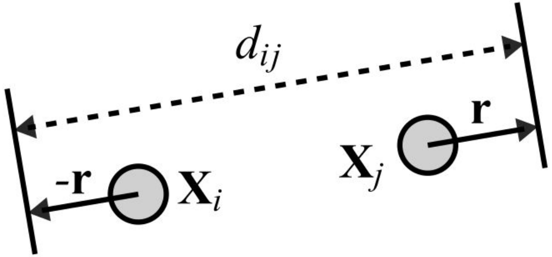
ACTACGTACTGGAAGTA

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

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

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.

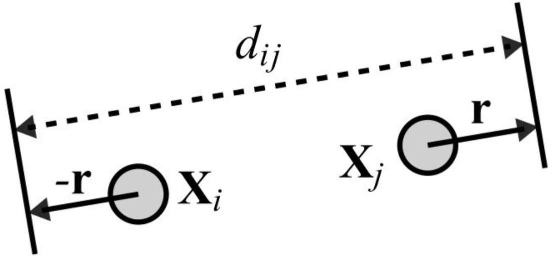
- The first node X_i of a pair is a uniform path step pick from all nodes.

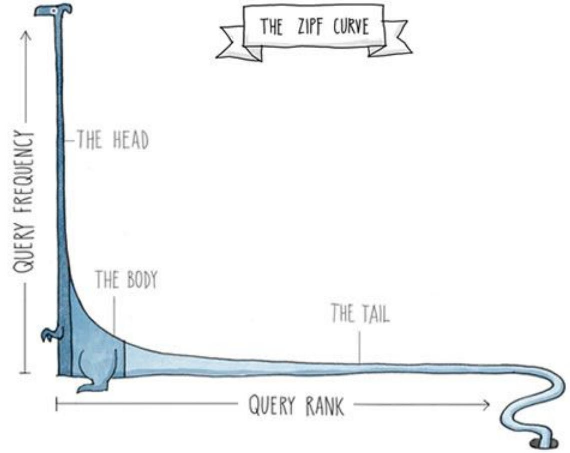
1D Graph Sorting by P-SGD - The Algorithm Explained

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.

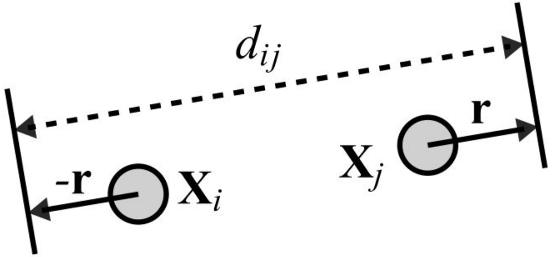




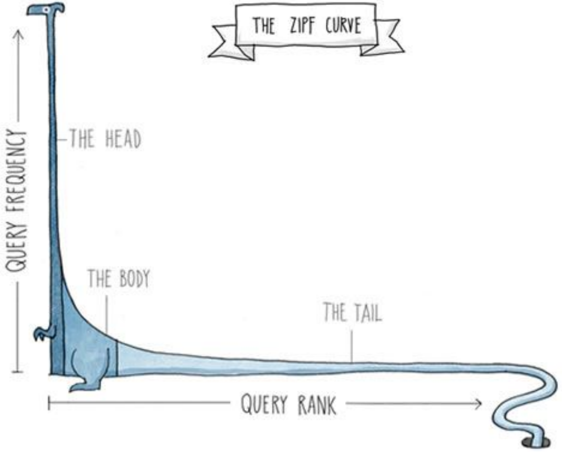
1D Graph Sorting by P-SGD - The Algorithm Explained

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.



THE ZIPF CURVE
QUERY FREQUENCY
THE HEAD
THE BODY
THE TAIL
QUERY RANK