Future impact of ALPACA

Research prospects by
Luca Denti

Open problems on pangenomes

(currently under investigation)

Research prospects by
Luca Denti

1) Pangenomic analysis in transcriptomics *(aka pantranscriptomics)*

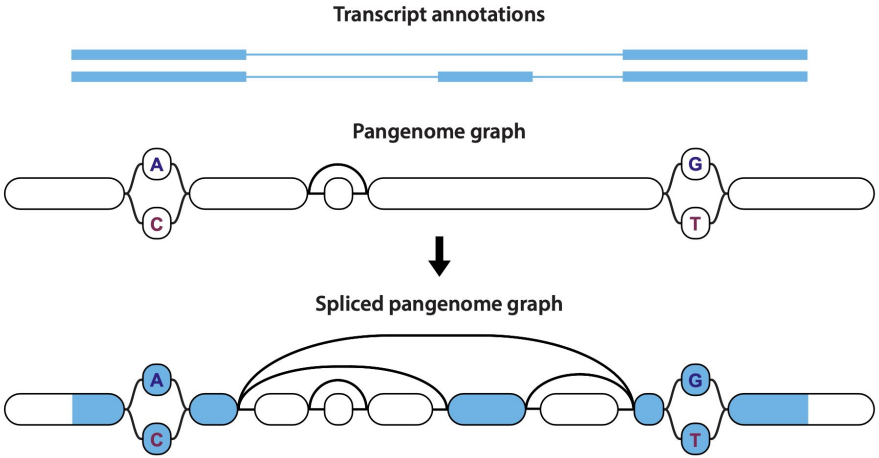
RSEM is to RPVG* as rMATS is to pantas

PLOS COMPUTATIONAL BIOLOGY

Differential quantification of alternative splicing events on spliced pangenome graphs

Simone Ciccolella, Davide Cozzi, Gianluca Della Vedova, Stephen Njuguna Kuria, Paola Bonizzoni, Luca Denti

doi: https://doi.org/10.1101/2023.11.06.565751



*Sibbesen et al. ***Haplotype-aware pantranscriptome analyses using spliced pangenome graphs.*** Nature Methods (2023)

1) Pangenomic analysis in transcriptomics *(aka pantranscriptomics)*

`RSEM` is to `RPVG`* as `rMATS` is to `pantas`

PLOS COMPUTATIONAL BIOLOGY

Differential quantification of alternative splicing events on spliced pangenome graphs

Simone Ciccolella, Davide Cozzi, Gianluca Della Vedova, Stephen Njuguna Kuria, Paola Bonizzoni, Luca Denti

doi: https://doi.org/10.1101/2023.11.06.565751

Based on **spliced alignment** to spliced pangenomes with `vg mpmap` (`GCSA2`-based)

- efficiency
  - *graph needs to be **heavily pruned** (haplotype-aware transcripts from dros. n.o. genes, >256GB RAM)*
  - *one replicate required more than **3 days** to align*
- backward extensions limited to 256bp (short reads) - *longer patterns can return **false positives***
- inaccuracies when aligning over **novel splicing sites**

*Sibbesen et al. ***Haplotype-aware pantranscriptome analyses using spliced pangenome graphs.*** Nature Methods (2023)

1) Pangenomic analysis in transcriptomics *(aka pantranscriptomics)*

`RSEM` is to `RPVG`* as `rMATS` is to `pantas`

PLOS COMPUTATIONAL BIOLOGY

Differential quantification of alternative splicing events on spliced pangenome graphs

Simone Ciccolella, Davide Cozzi, Gianluca Della Vedova, Stephen Njuguna Kuria, Paola Bonizzoni, Luca Denti

doi: https://doi.org/10.1101/2023.11.06.565751

Based on **spliced alignment** to spliced pangenomes with `vg mpmap` (`GCSA2`-based)

↳ *`ropebwt3`?*

- ➢ efficiency
  - *graph needs to be **heavily pruned** (haplotype-aware transcripts from dros. n.o. genes, >256GB RAM)*
  - *one replicate required more than **3 days** to align*
- ➢ backward extensions limited to 256bp (short reads) - *longer patterns can return **false positives***
- ➢ inaccuracies when aligning over **novel splicing sites**

*Sibbesen et al. ***Haplotype-aware pantranscriptome analyses using spliced pangenome graphs.*** Nature Methods (2023)

2) Long reads and pangenomes

Common short reads practice (e.g., `Giraffe-DV`):

1. Align to graph → 2. Surject to reference → 3. Reference-based approach

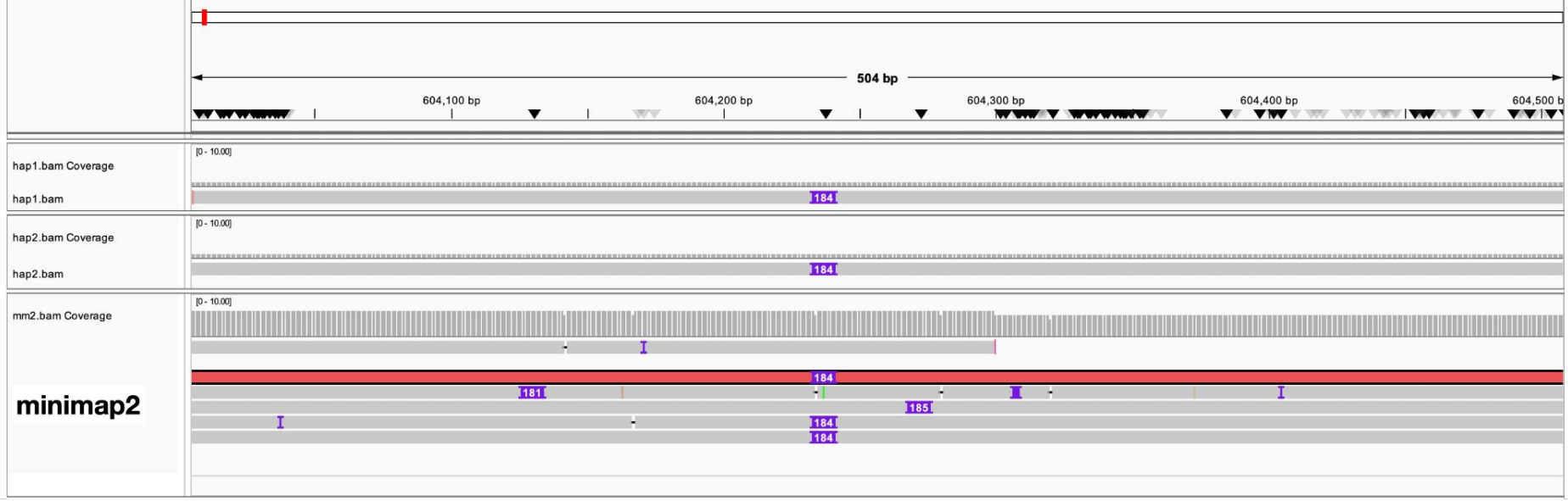
2) Long reads and pangenomes

Common short reads practice (e.g., `Giraffe-DV`):

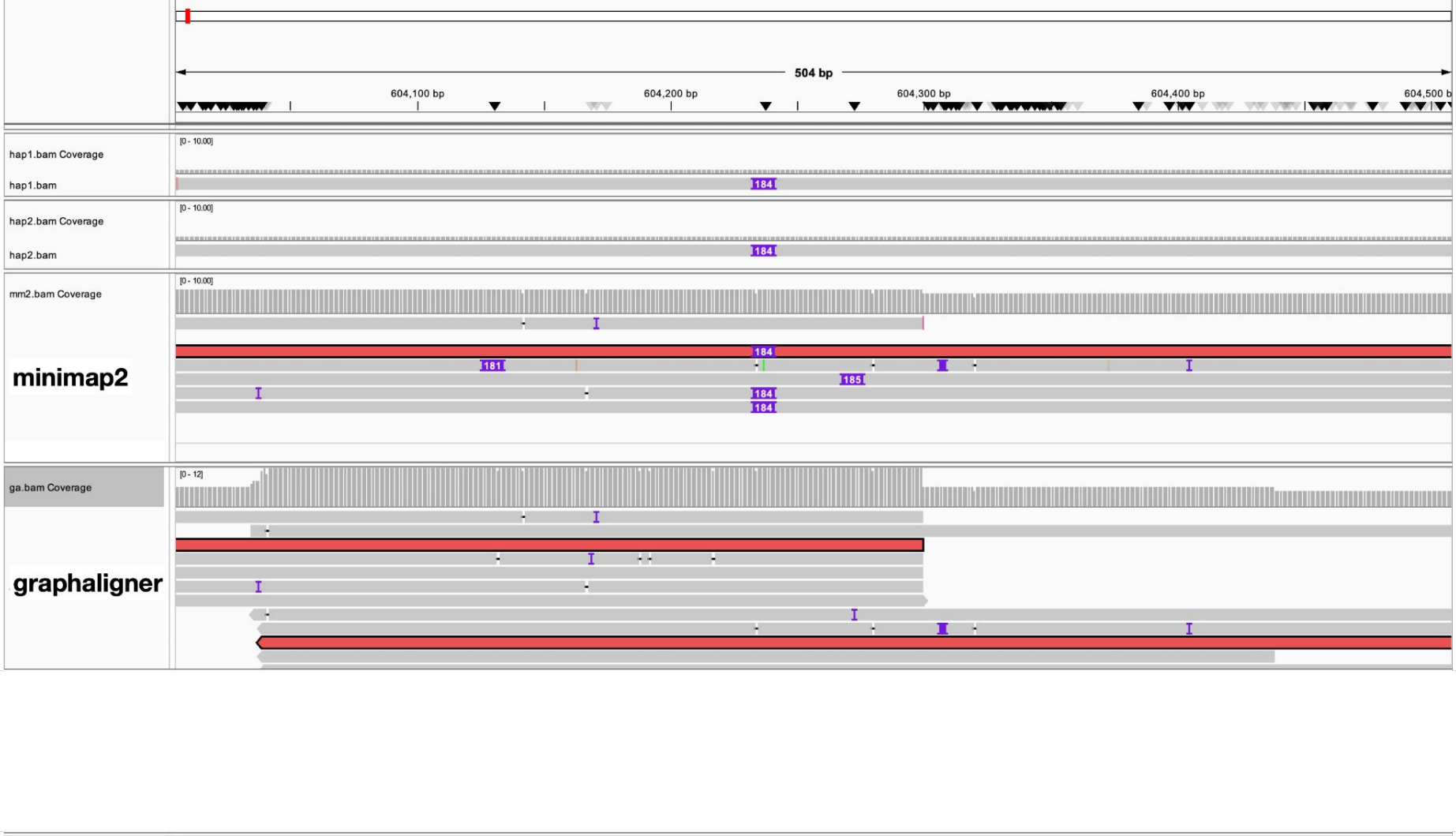
> *1. Align to graph → 2. Surject to reference → 3. Reference-based approach*

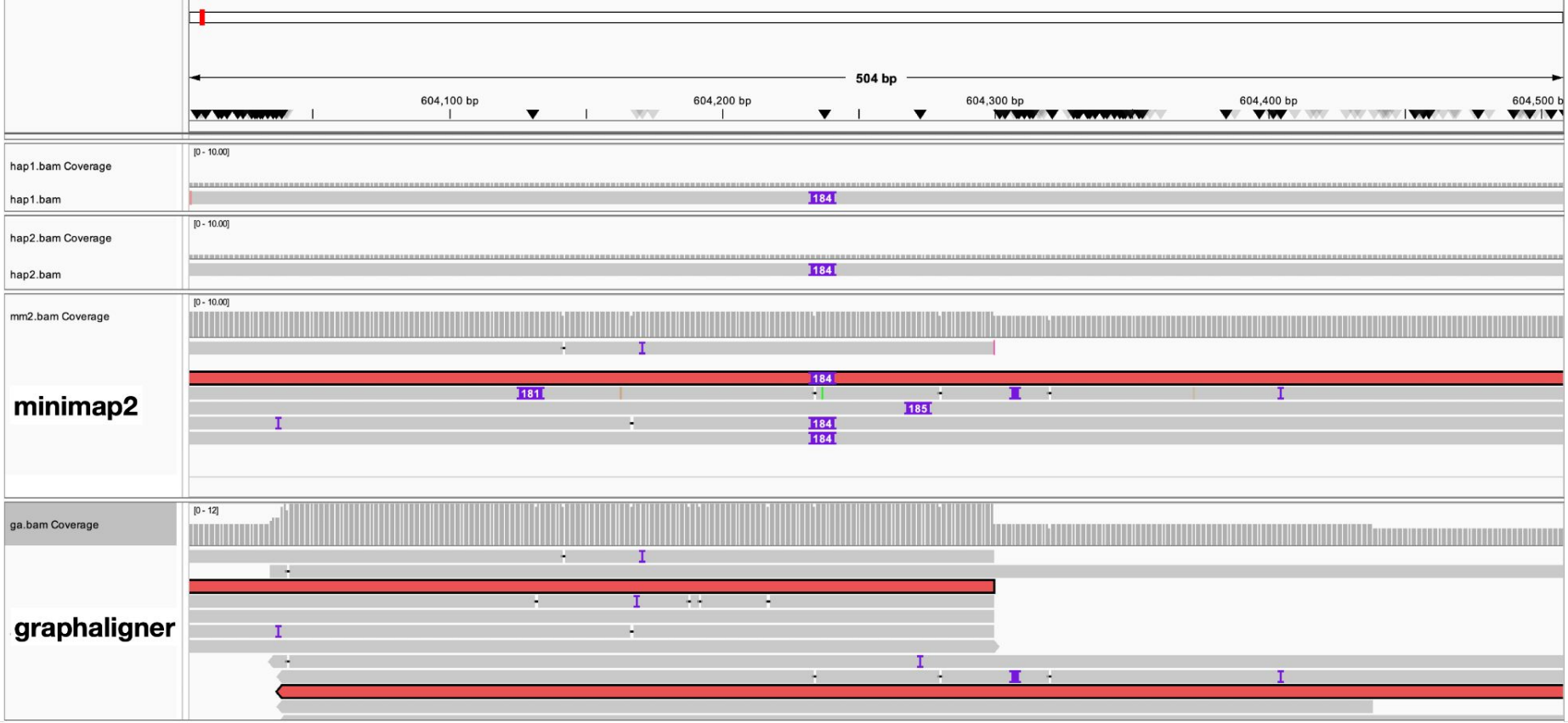
This is **not applicable to long reads** *(for now and at least for SV calling)*:

- ➢ improve long read alignments (1)
- ➢ design new pangenomic approaches for downstream analyses (2,3 / 3)
- ➢ design alignment-free methodologies (1,2,3)


504 bp
604,100 bp
604,200 bp
604,300 bp
604,400 bp
604,500 b
hap1.bam Coverage
[0 - 10.00]
hap1.bam
184
hap2.bam Coverage
[0 - 10.00]
hap2.bam
184
mm2.bam Coverage
[0 - 10.00]
minimap2
184
181
185
184
184

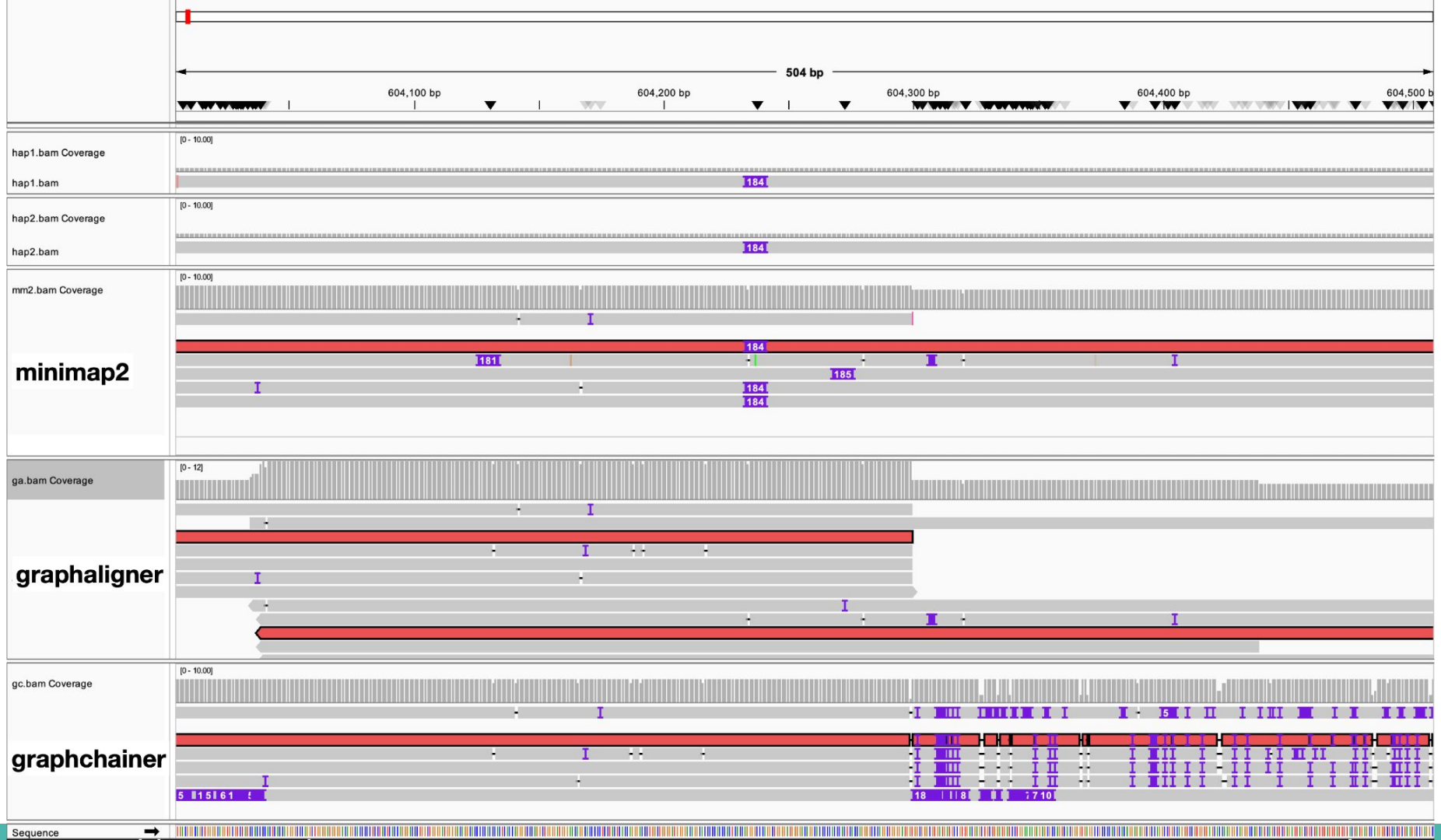
Sequence


504 bp
604,100 bp
604,200 bp
604,300 bp
604,400 bp
604,500 b
hap1.bam Coverage
[0 - 10.00]
hap1.bam
184
hap2.bam Coverage
[0 - 10.00]
hap2.bam
184
mm2.bam Coverage
[0 - 10.00]
minimap2
184
181
185
184
184
ga.bam Coverage
[0 - 12]
graphaligner
Sequence



3304M 1I 4602M 43I 912M 1I 2660M 1I248M1I864M1I **790M184I391M** 1I90M

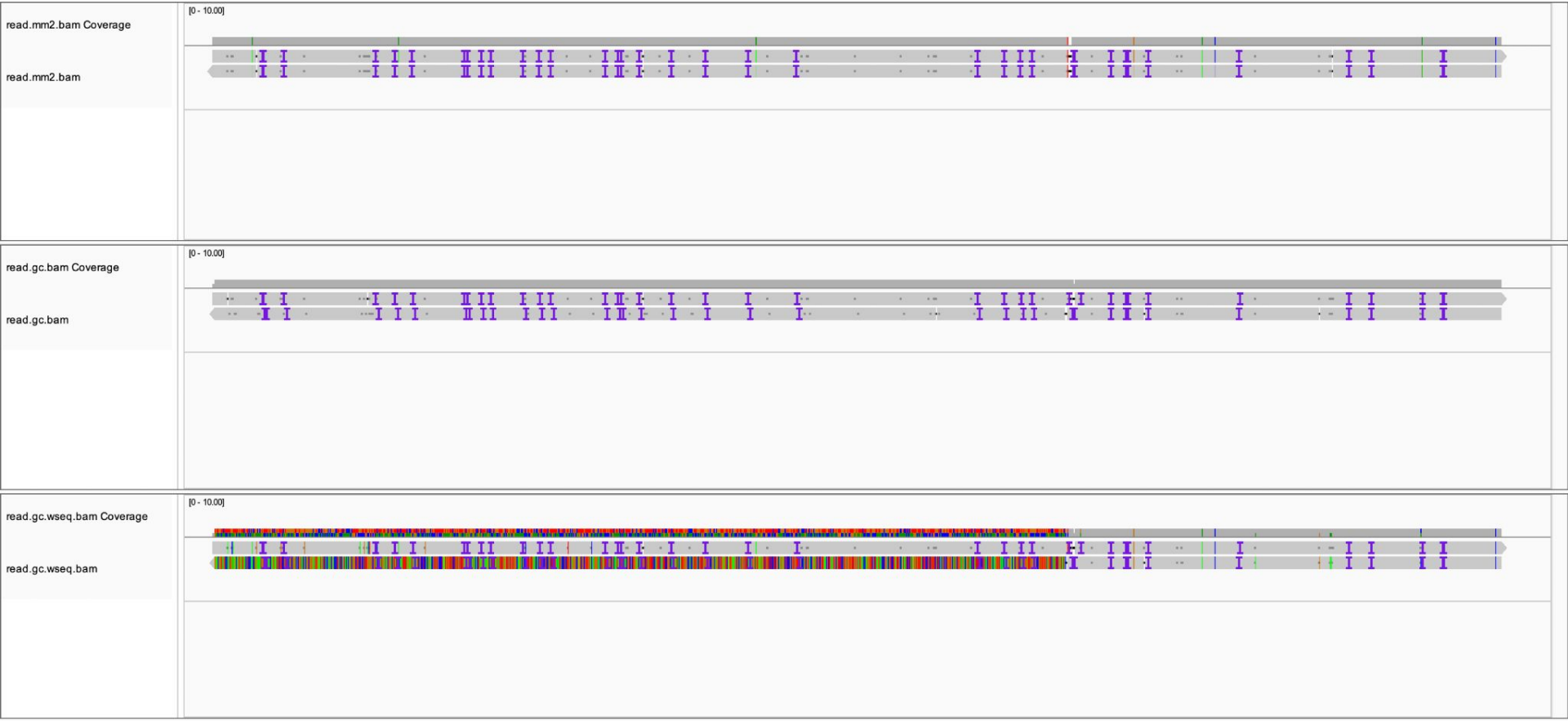
3304M 1I 4624M 34I 1M7I1M2I894M1I 2654M 1I248M1I874M1I **842M : 590M**1I90M


504 bp
604,100 bp
604,200 bp
604,300 bp
604,400 bp
604,500 b
hap1.bam Coverage
hap1.bam
hap2.bam Coverage
hap2.bam
mm2.bam Coverage
minimap2
ga.bam Coverage
graphaligner
gc.bam Coverage
graphchainer
Sequence

3304M 1I 4602M 43I 912M 1I 2660M 1I248M1I864M1I 790M184I391M 1I90M

3304M 1I 4624M 34I 1M7I1M2I894M1I 2654M 1I248M1I874M1I 842M : 590M1I90M

3304M1I4599M11I1M24I1M3I1M5I912M1I2660M1I248M1I864M1I851M1D2M1I9M2I2M2I1M1I2M1I2M1I9M2D5M1D4M1D9M1I6M1I2M1I10M1D2M1D17M1I9M3I4M1I3M1I6M1I6M1I6M2D5M1I5M1I13M1I12M1I9M1I8M1I1M1I4M1I3M2D6M1I1M1I2M1I3M1I4M1I5M1D7M1D3M1D5M1D7M1I3M1D9M1D1M1D1M1D2M3D2M2D14M1I3M1I1M1I2M1D2M1D2M2D2M1D6M1I4M1D2M1D2M1I10M1I2M1I3M2D27M2I1M1I20M1D17M1D2M2D3M1I1M2I80M2I4M1I10M1I1M1I1M2I25M2I14M1I21M1I17M1I3M1D4M2I2M


read.mm2.bam Coverage
[0 - 10.00]
read.mm2.bam
read.gc.bam Coverage
[0 - 10.00]
read.gc.bam
read.gc.wseq.bam Coverage
[0 - 10.00]
read.gc.wseq.bam

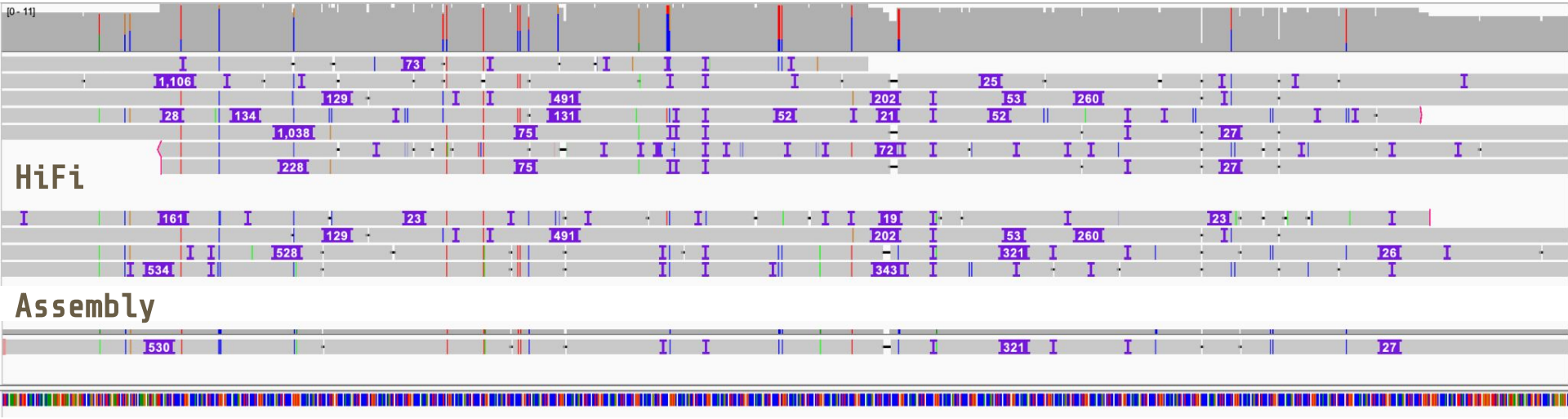
3) Reference-bias vs Alignment-bias

*"No single genome can represent the **diversity** in the human population. Using a single reference genome creates **reference biases**, adversely affecting variant discovery, gene–disease association studies and the accuracy of genetic analyses."*

https://humanpangenome.org/

*"Any downstream analysis is affected by the **choices** made by the aligner"*

3) Reference-bias vs Alignment-bias (CGR)



(mainly due to scores/penalties and query lengths)

- What should we “fix”?
- Is alignment to pangenomes going to improve/facilitate this?
- Is this just a problem of “representation”?

4) How to represent SVs *(or any variant)* wrt a pangenome?

- One VCF against each path *(genome)* of the pangenome?
- FASTA with *svtigs** ?
- Can we find a more compact way?
- What about a GAF-like file format for variations?

*Soylev, ..., Marschall. ***SVarp: pangenome-based structural variant discovery.*** *bioRxiv (2024)*