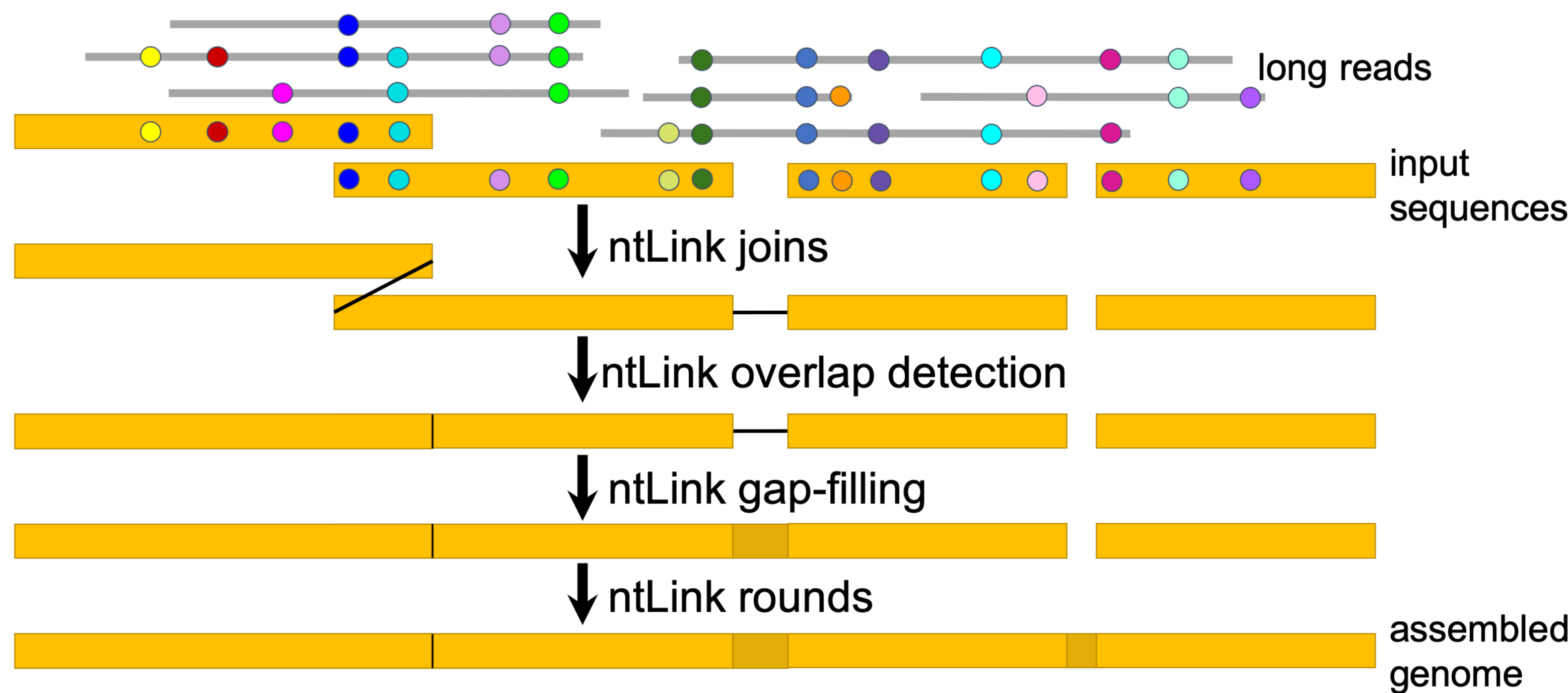


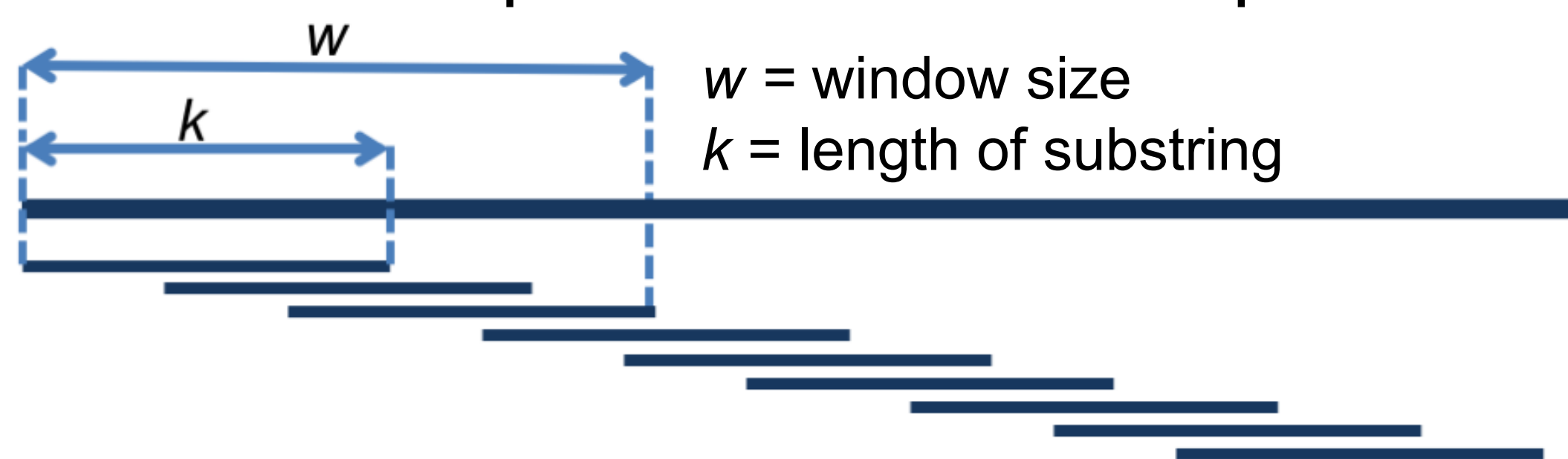
GoldRush-Link

- Essential step in new *de novo* assembler GoldRush
 - Powered by **ntLink**¹ long read scaffolder
- ntLink** improvements:
 - Overlap detection
 - Gap-filling
 - Liftover-based rounds



Minimizer sketches

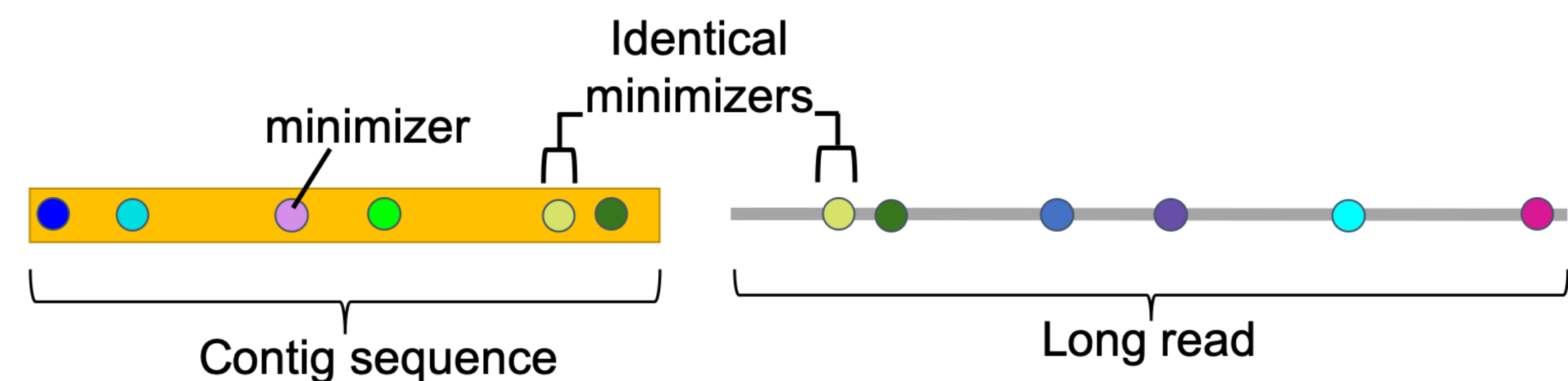
Reduce computational cost of sequence data storage and manipulation²



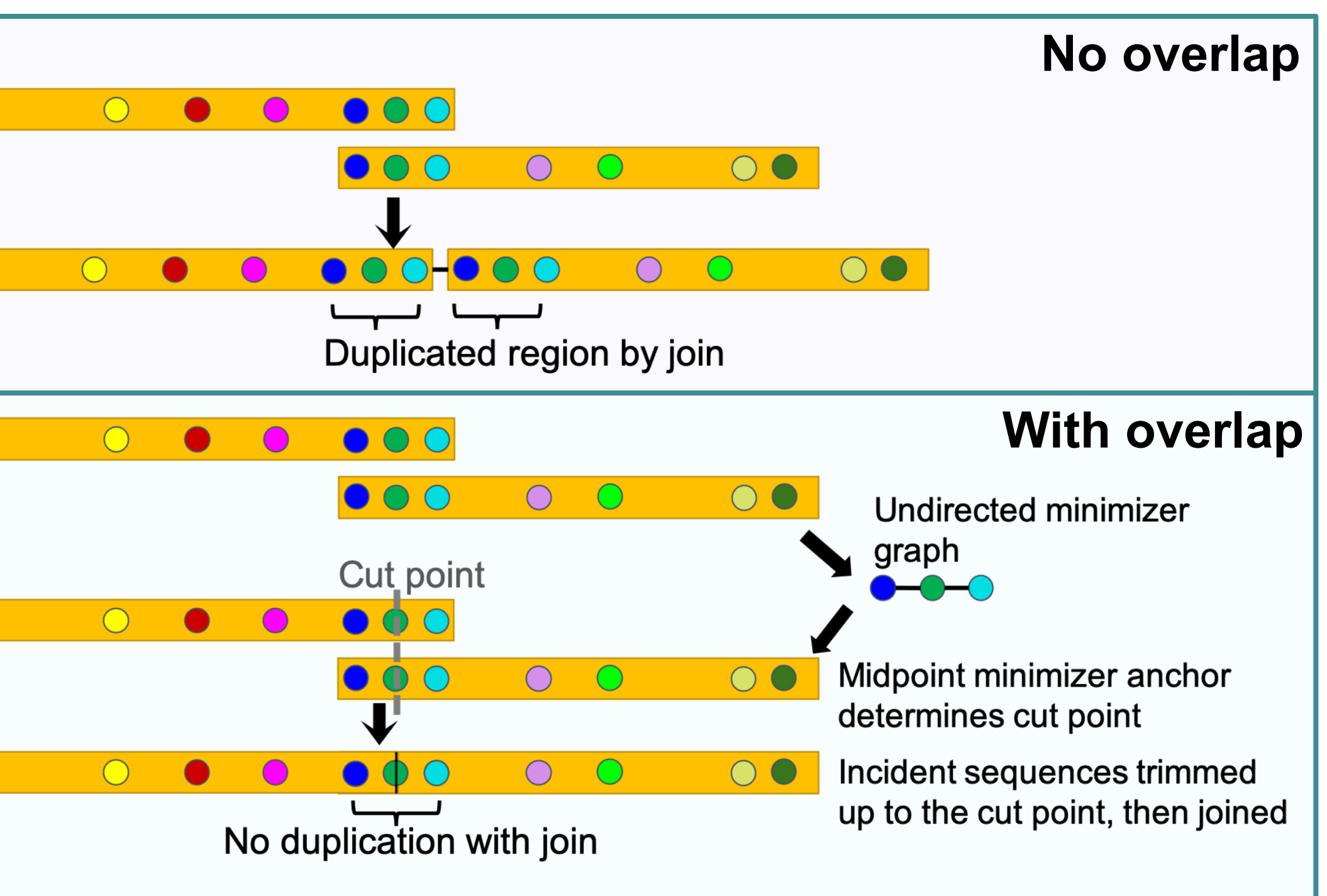
For each window of w adjacent k -mers:
 • Compute hash values of each k -mer
 • Window's minimizer = smallest hash value
 Generates ordered list of minimizers per sequence

Methods

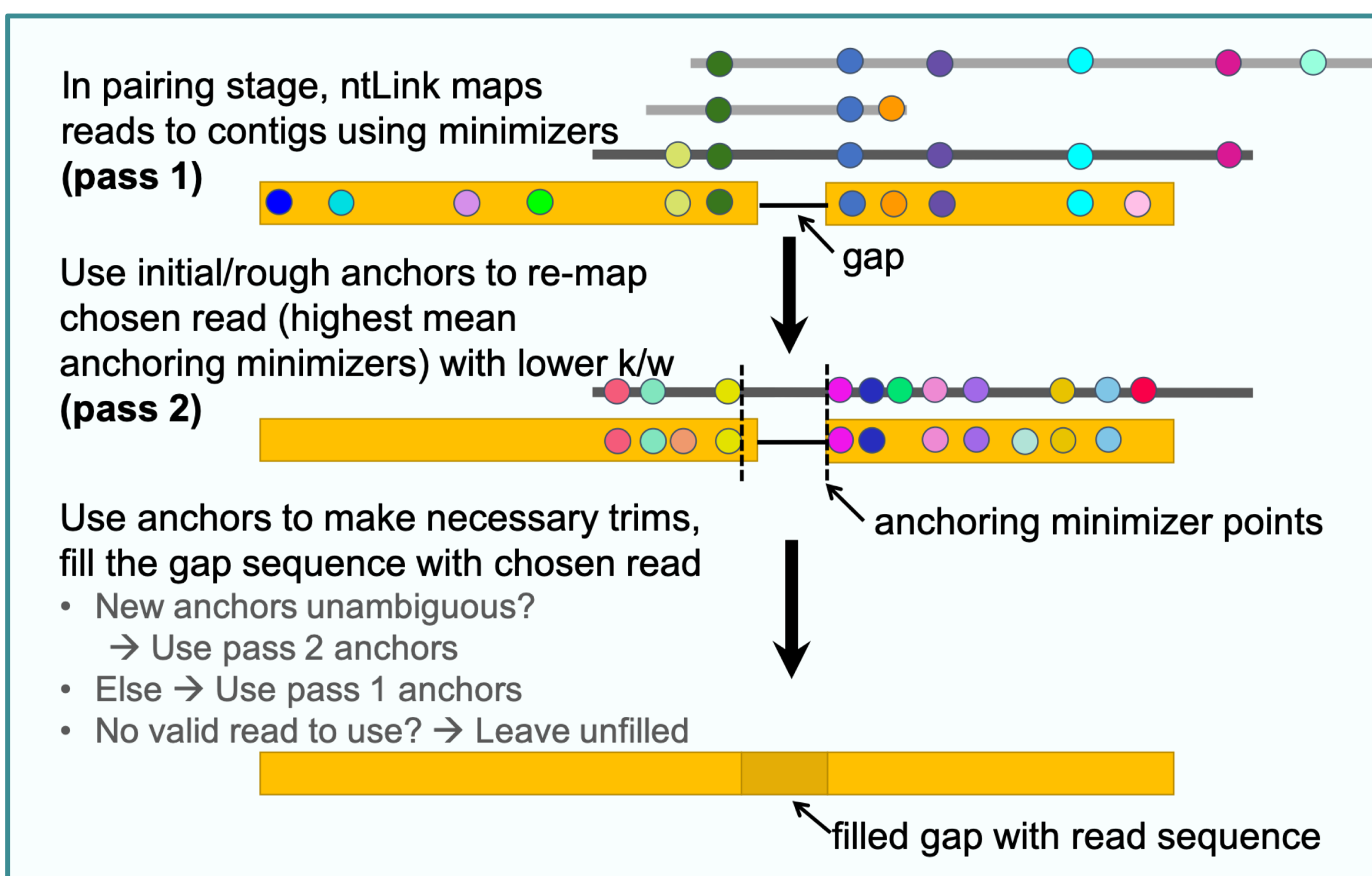
Schematic key:



Overlap detection

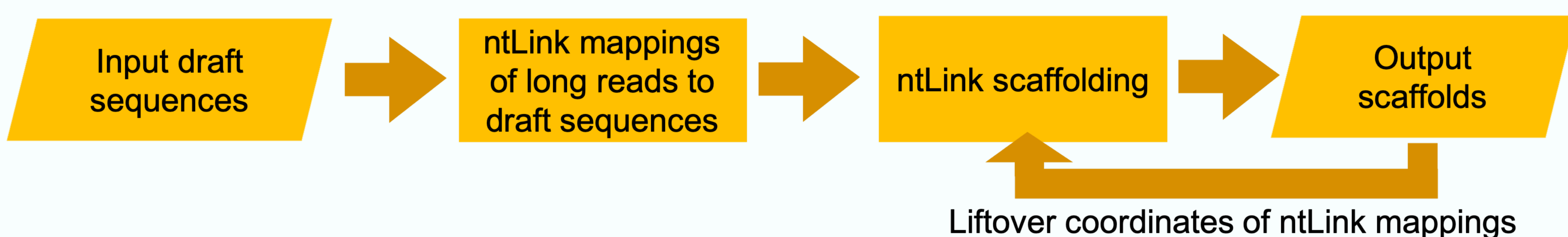


Gap-filling



Liftover-based rounds

- Running additional rounds of ntLink can produce more contiguous final assemblies
- Re-mapping the reads at each round is costly
 → Liftover mapping coordinates after each ntLink round based on scaffold composition



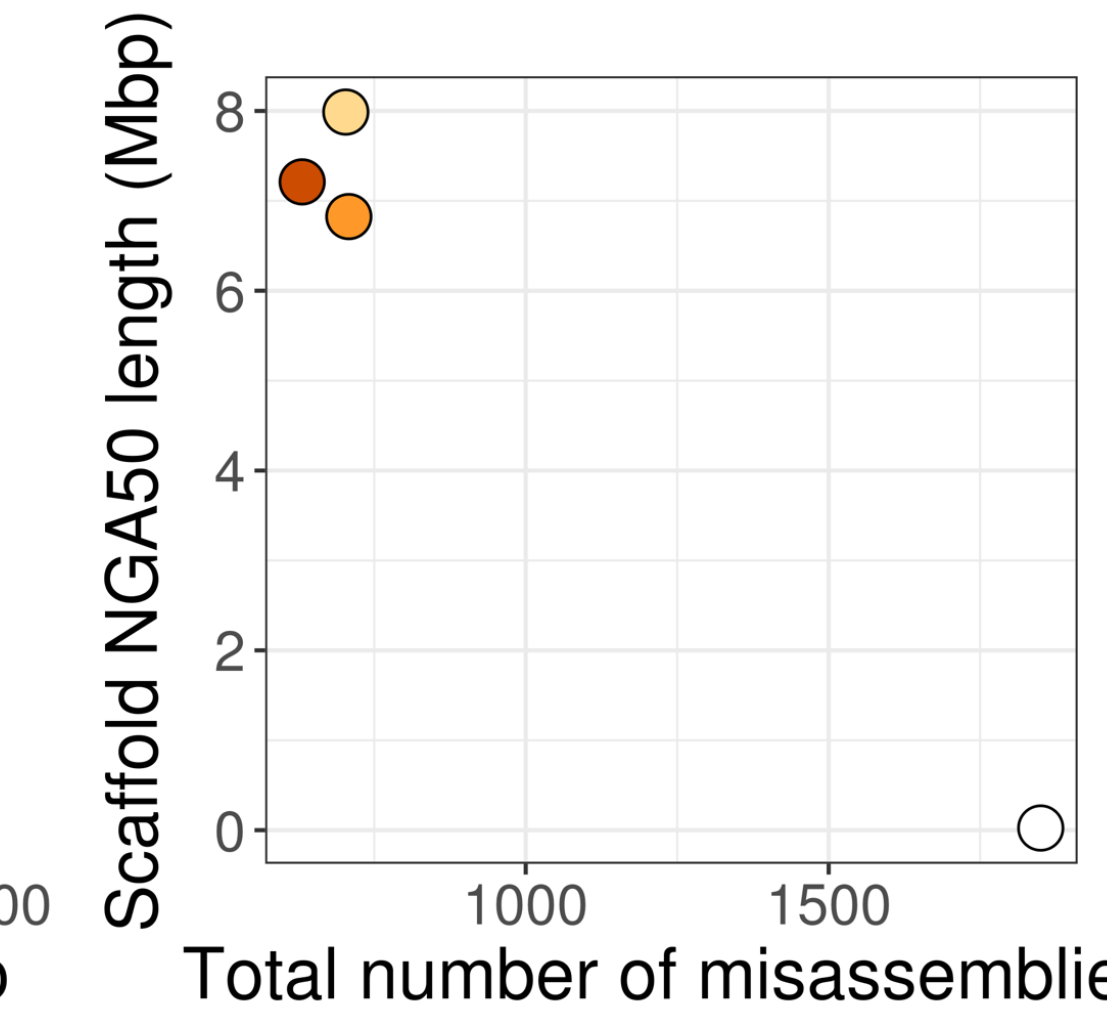
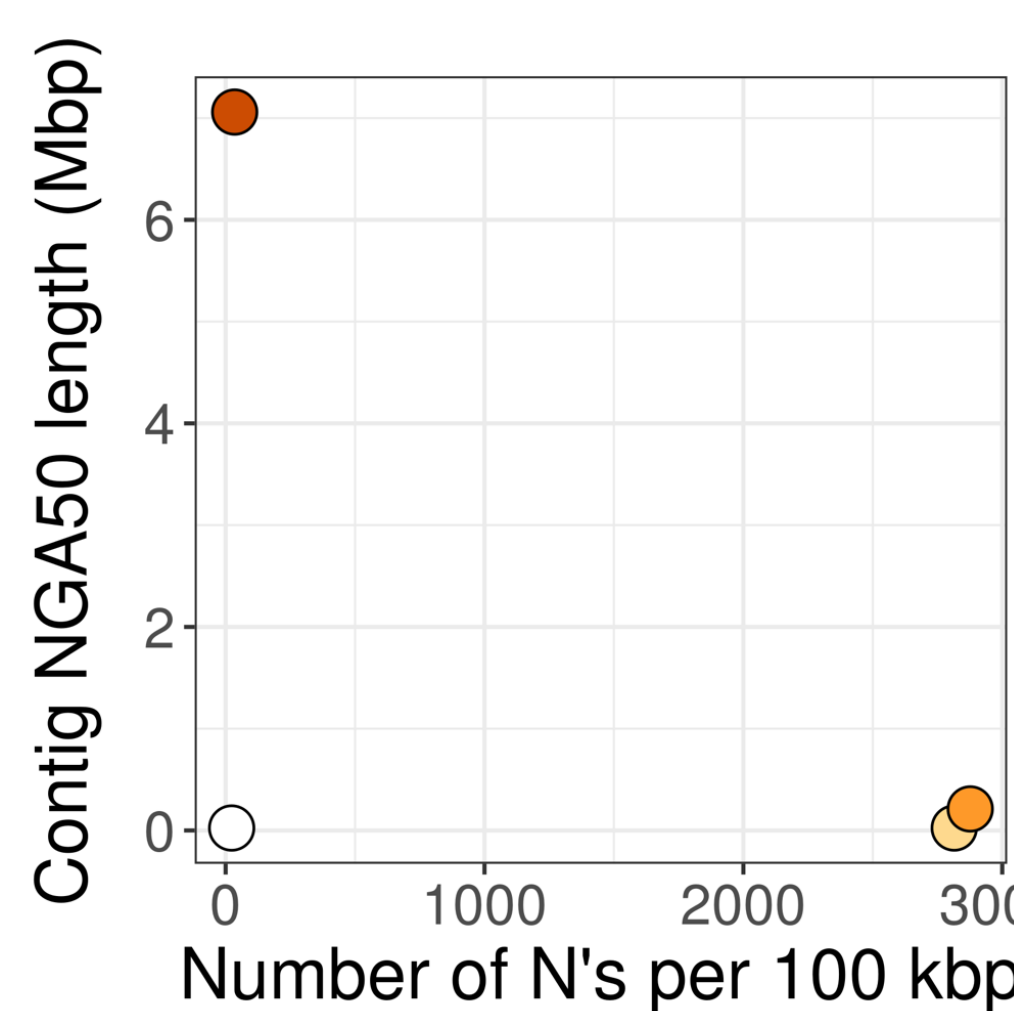
Results

ntLink assemblies using Oxford Nanopore long reads for two human individuals:

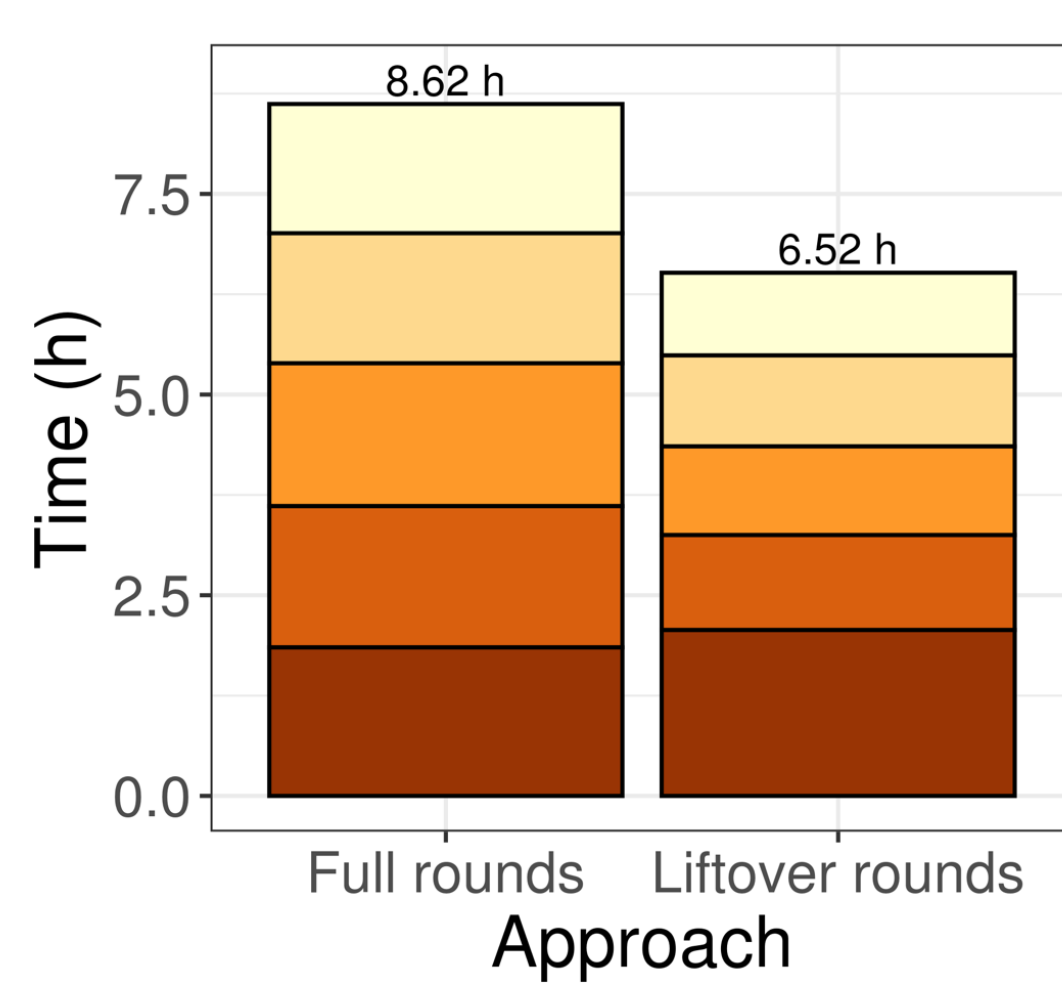
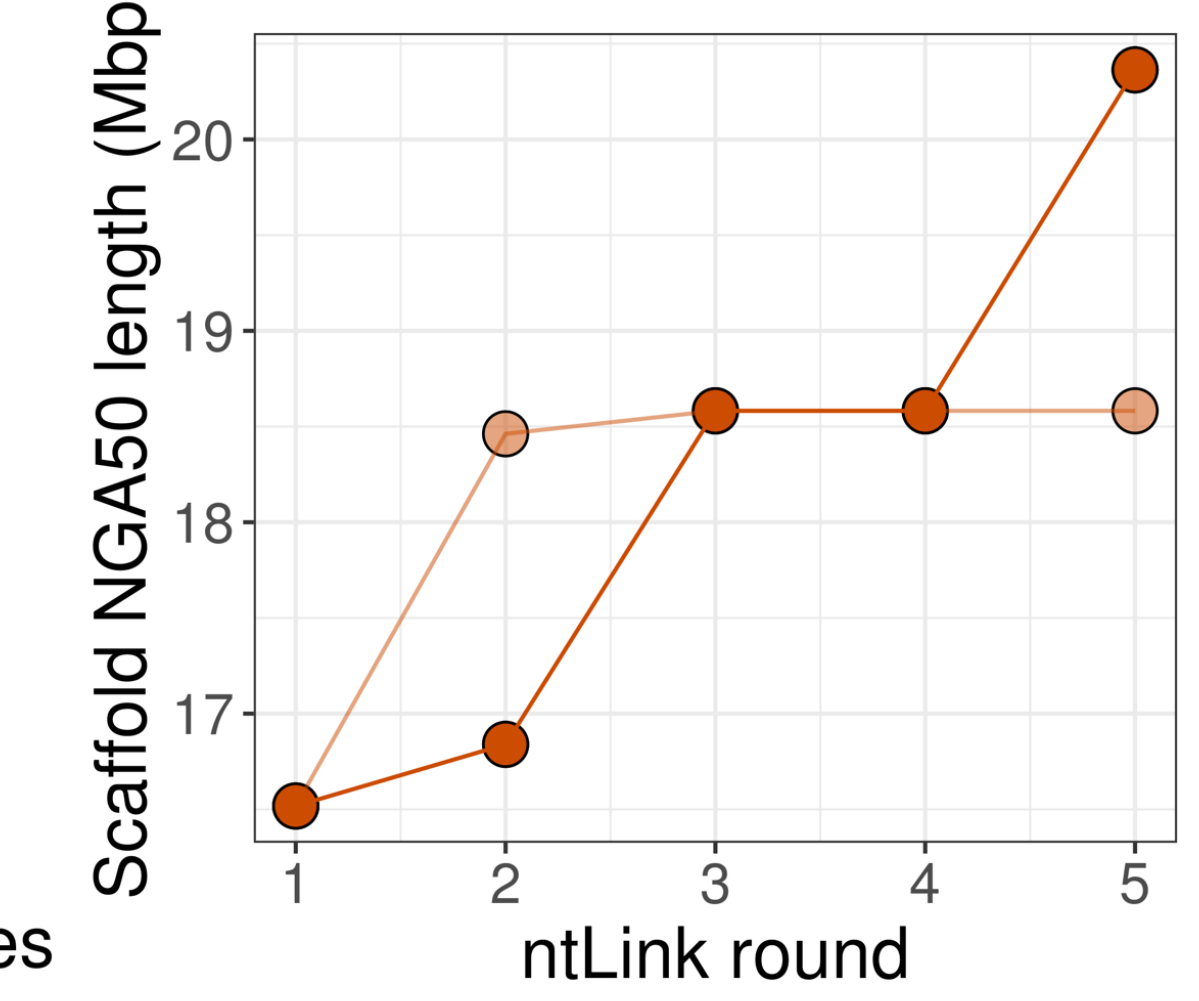
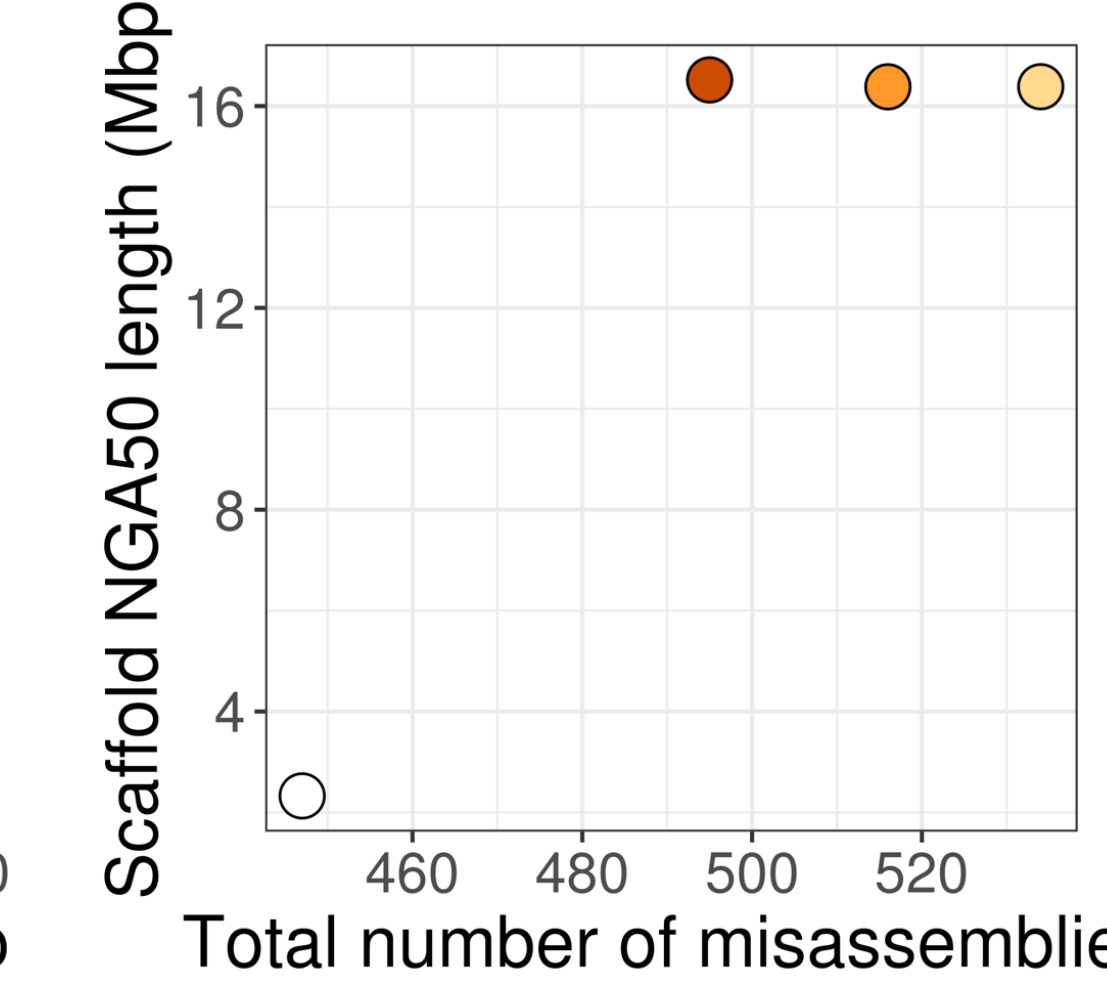
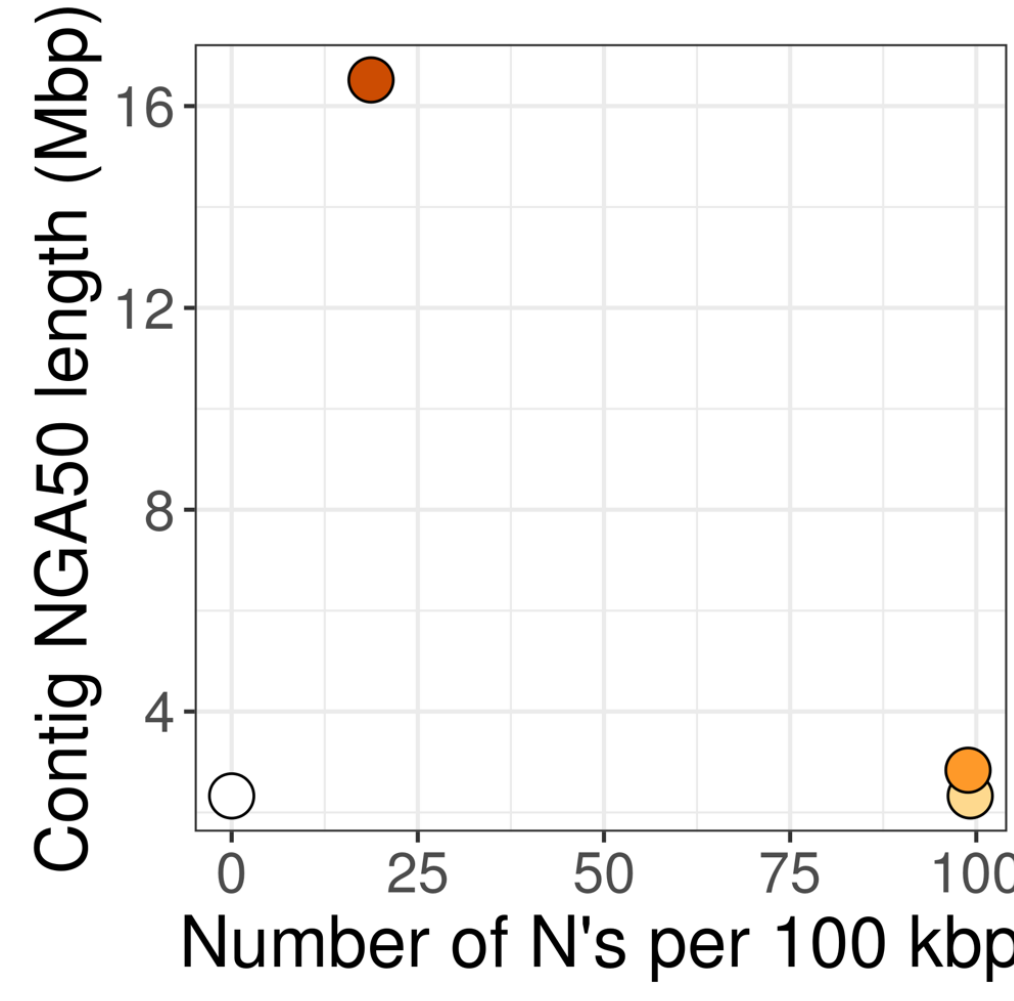
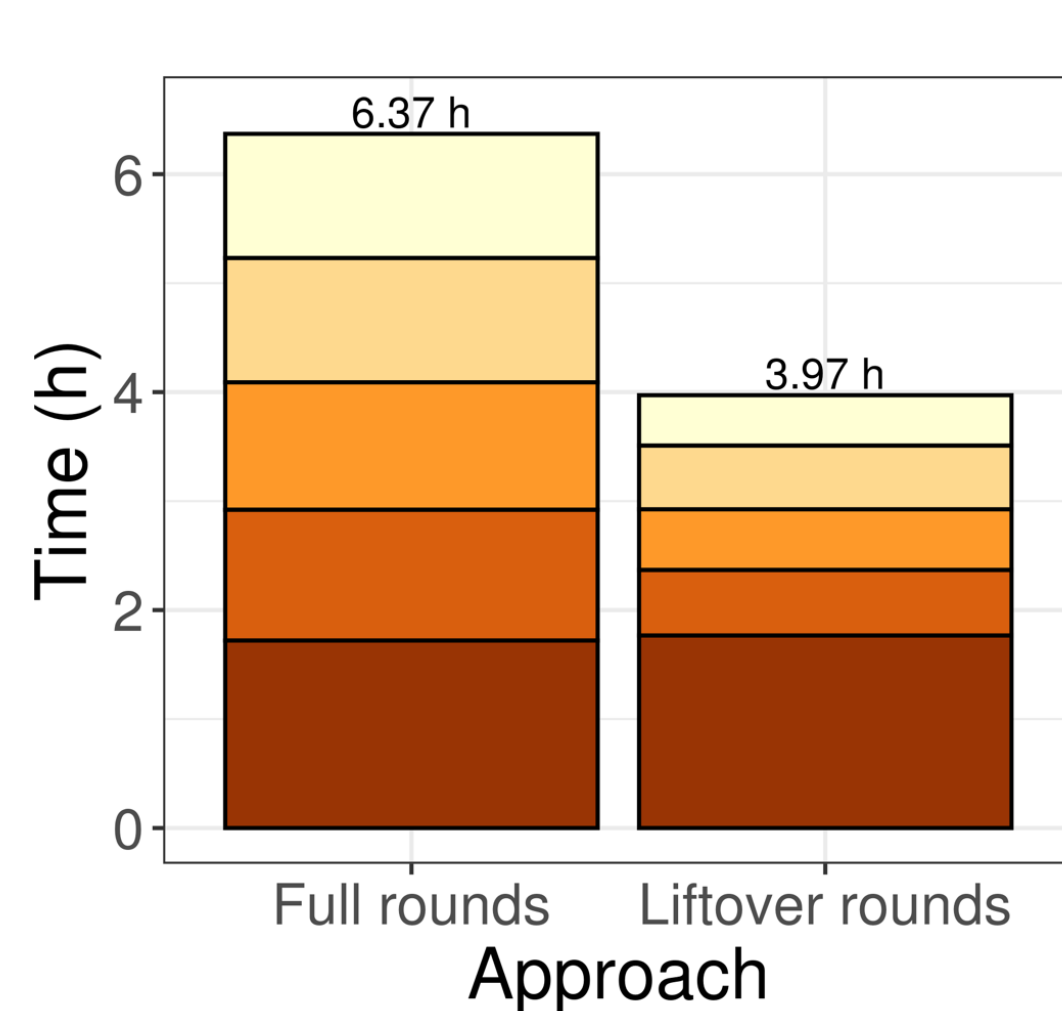
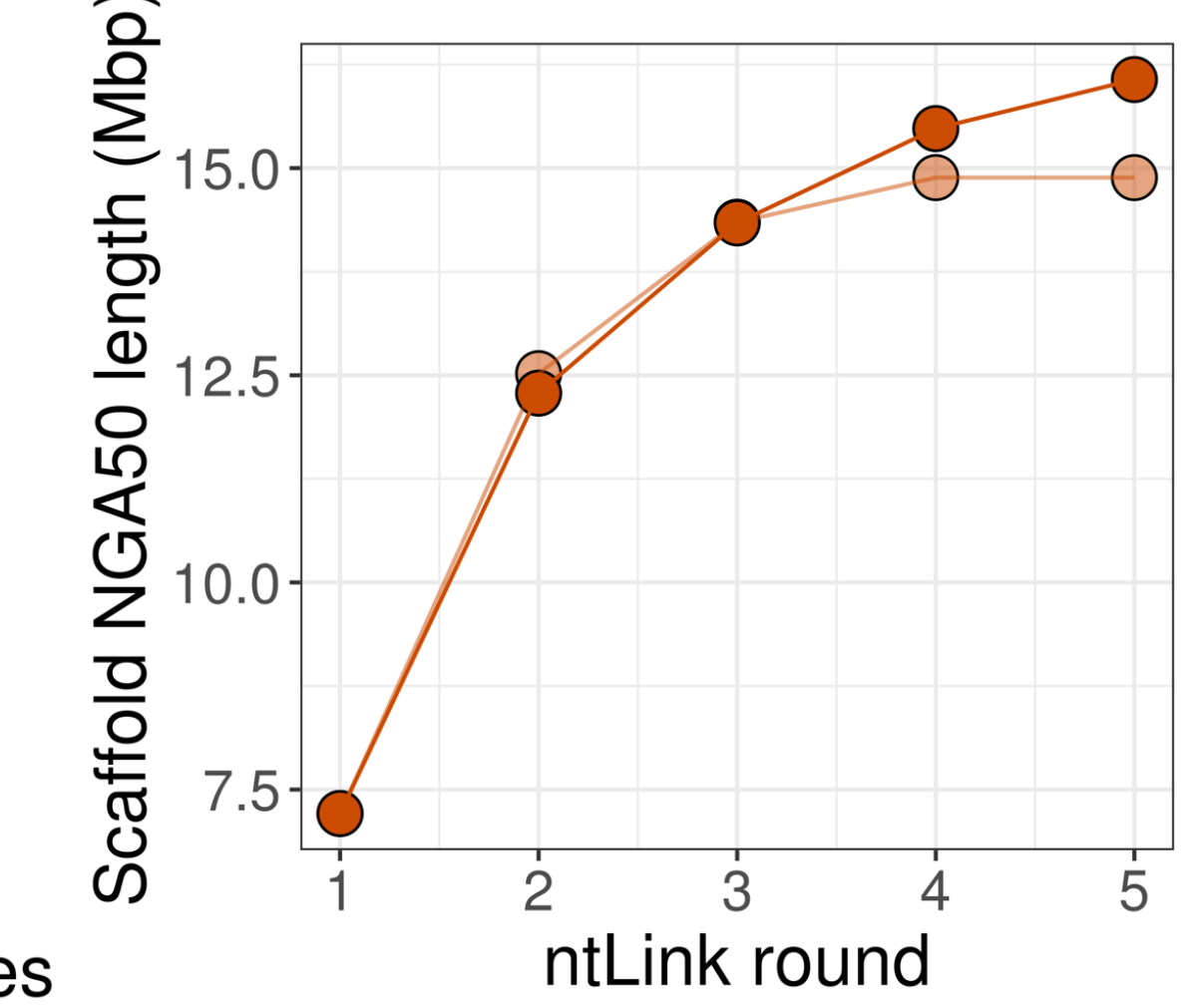
| Individual | Fold coverage | Baseline assembly |
|------------|---------------|--------------------------------|
| NA24385 | 67 | GoldRush ⁴ goldtigs |
| NA19240 | 49 | Shasta ³ |

GoldRush goldtigs: Polished, corrected golden path reads
Golden path reads: ~1x read representation of the genome

Overlap detection and gap-filling features



Liftover-based rounds



○ Baseline ○ no overlap, no gap-fill ● overlap, no gap-fill ● overlap, gap-fill

- Using the overlap and gap-filling features substantially increases the "contig" contiguity
- Both features also reduce the overall number of misassemblies

- Running additional ntLink rounds yields further contiguity gains
- Using the liftover functionality vs. naïve rounds results in higher contiguity and faster runtimes

Conclusions

- Multiple new features were added to the **ntLink** long read scaffolder
 - Overlap detection, gap-filling and liftover-based rounds
- These improvements were made with our *de novo* long read assembler, GoldRush, in mind, but are also applicable to the general usage of **ntLink**

References

- Coombe L, et al. 2021. LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* **22**: 2021.06.17.448848.
- Roberts M, et al. 2004. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**: 3363–3369.
- Shafin K, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. *Nat Biotechnol* **38**: 1044–1053
- Wong J, et al. 2022. GoldRush-Path: a *de novo* assembler for long reads with linear time complexity. *ISMB HITSeq talk*.

Software Availability

<https://github.com/bcgsc/ntlink>
<https://github.com/bcgsc/goldrush>
 conda install -c bioconda ntlink

Genome British Columbia
 GenomeCanada
 National Institutes of Health