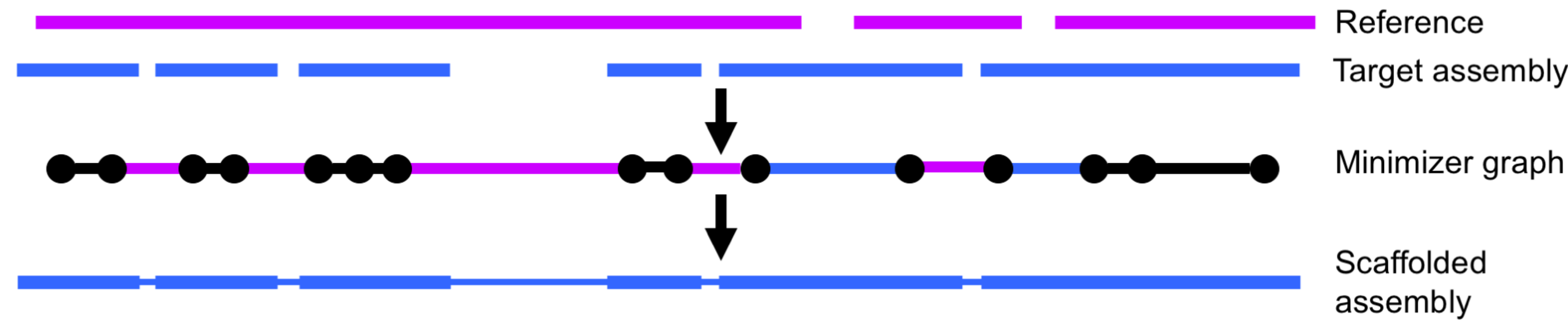# ntJoin: Fast and lightweight assembly-guided scaffolding using minimizer graphs

Lauren Coombe, Vladimir Nikolić, Justin Chu, Inanc Birol and René L. Warren

lcoombe@bcgsc.ca
www.birollab.ca

BC CANCER RESEARCH — Provincial Health Services Authority

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

## Introduction

ntJoin

- Scaffolding using one (or more) references
- Uses minimizer graph approach for fast, lightweight, and scalable scaffolding

Flexible usage

Target	Reference
Draft assembly	Reference-grade genome
Short read assembly	Long read assembly
Draft assembly	Assembly from related species

Minimizer sketches

- Reduce computational cost of sequence data storage and manipulation[1]

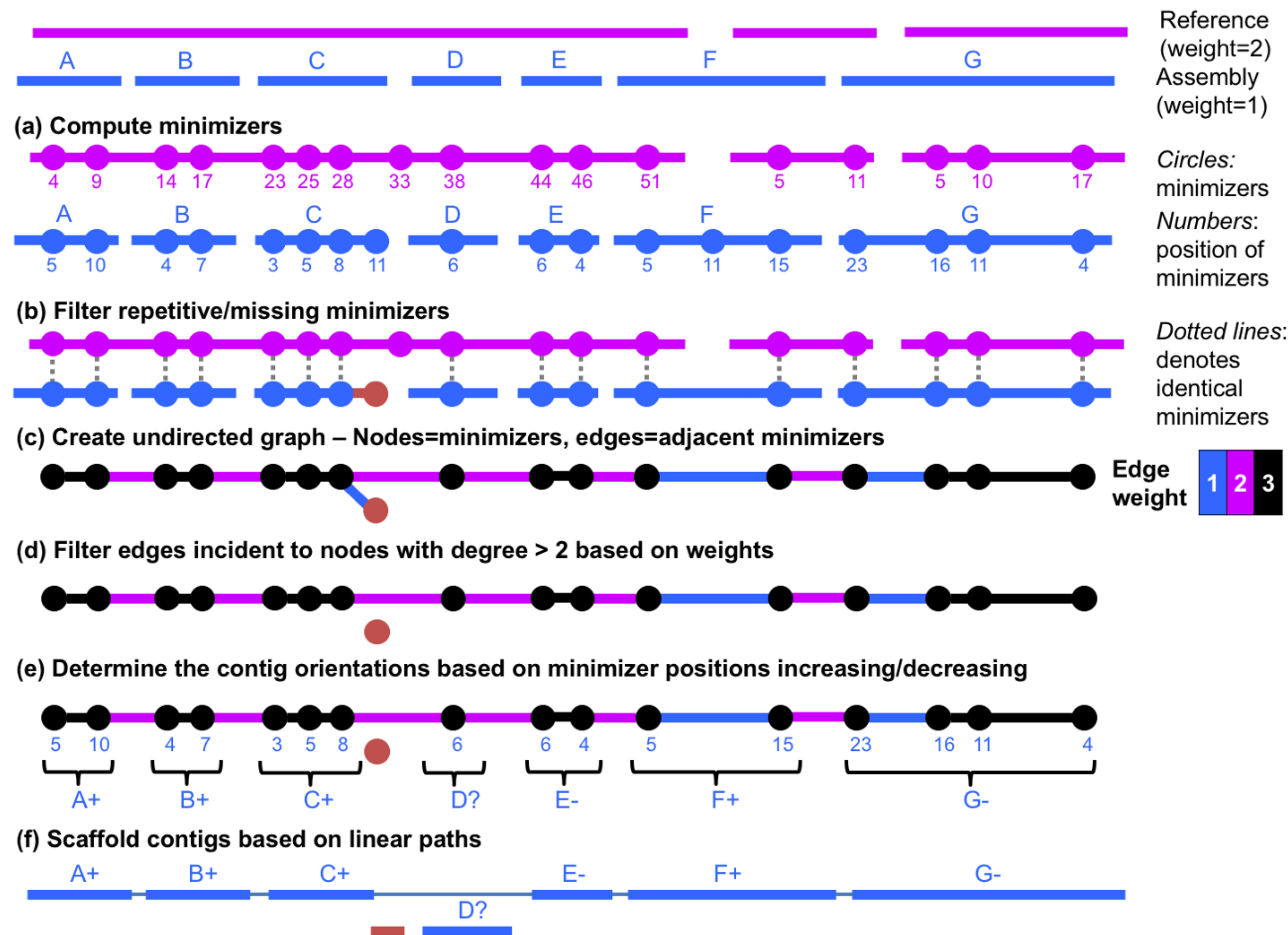w = window size
k = length of substring
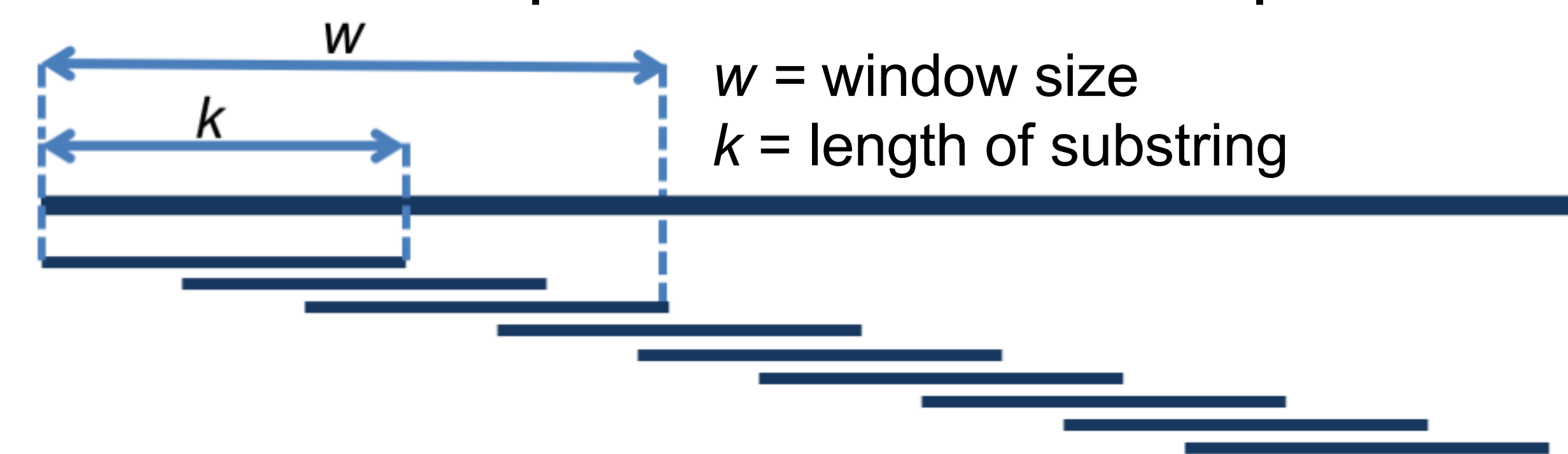
For each window of w adjacent k-mers:
- Compute hash value of each k-mer
- Window's minimizer = smallest hash value
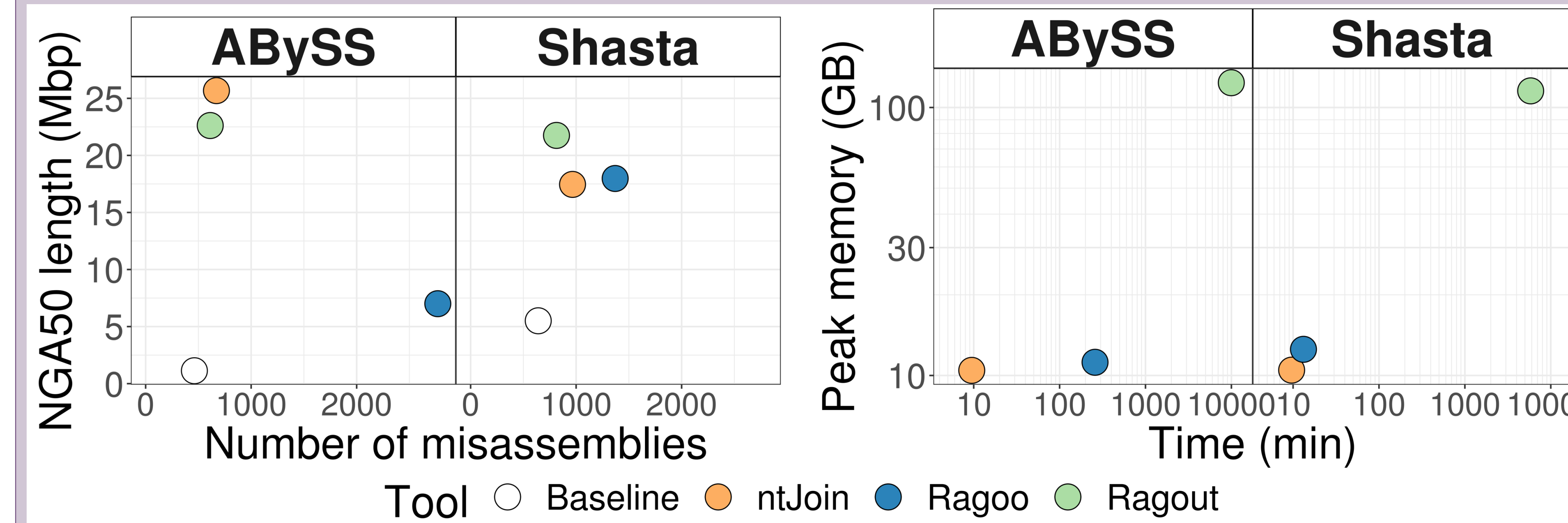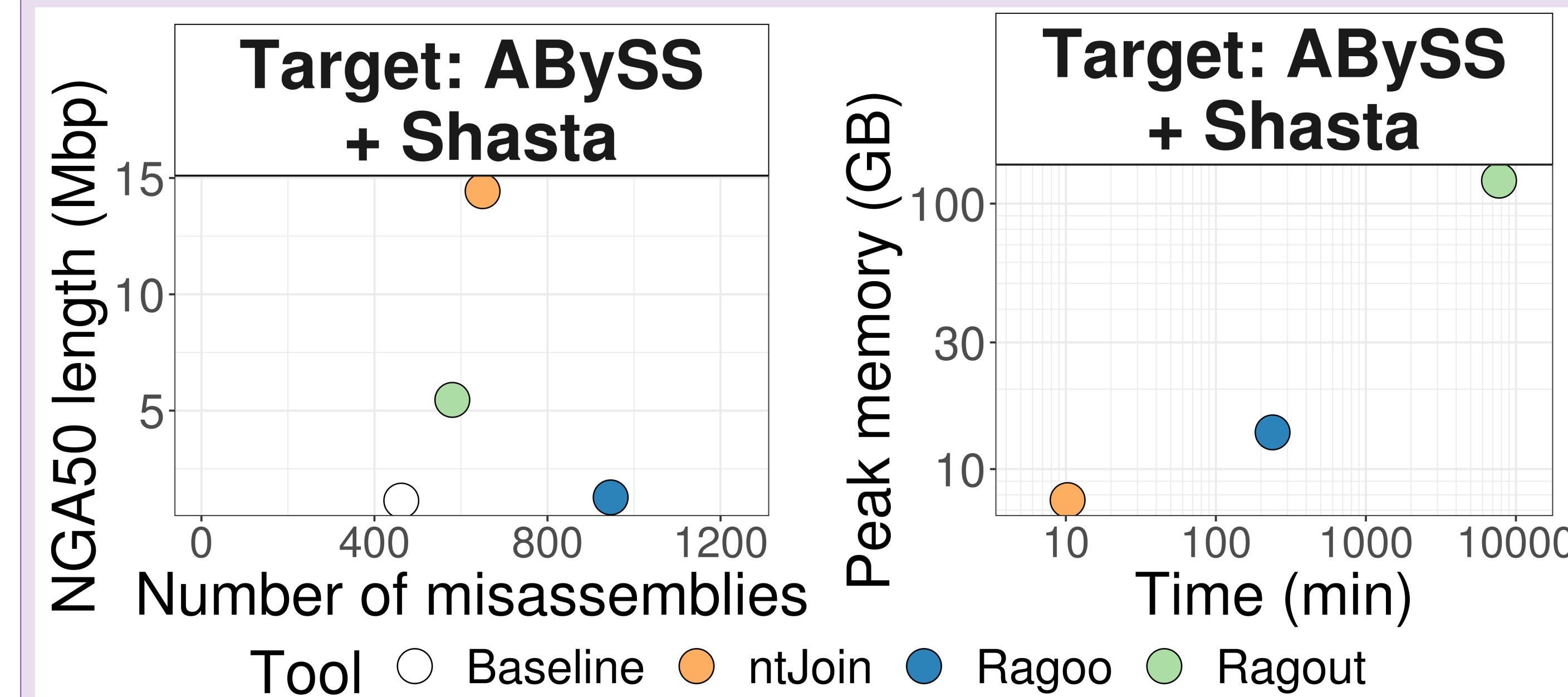Generates ordered list of minimizers per sequence

Method

Reference (weight=2)
Assembly (weight=1)

(a) Compute minimizers

Circles: minimizers
Numbers: position of minimizers

(b) Filter repetitive/missing minimizers

Dotted lines: denotes identical minimizers

(c) Create undirected graph – Nodes=minimizers, edges=adjacent minimizers

Edge weight: 1 2 3

(d) Filter edges incident to nodes with degree > 2 based on weights

(e) Determine the contig orientations based on minimizer positions increasing/decreasing

(f) Scaffold contigs based on linear paths

Results

Scaffolding draft assemblies with a reference-grade assembly

| Target assemblies	ABySS[2] (NA12878 short reads)
	Shasta[3] (NA12878 nanopore reads)
Reference	Human reference GRCh38

Scaffolding a short read assembly with a draft long read assembly

Target assembly	ABySS (NA12878 short reads)
Reference	Shasta (NA12878 nanopore reads)

- Achieves contiguity of a long read assembly and the base quality of the short read assembly

Scaffolding a draft assembly with an assembly from a related species
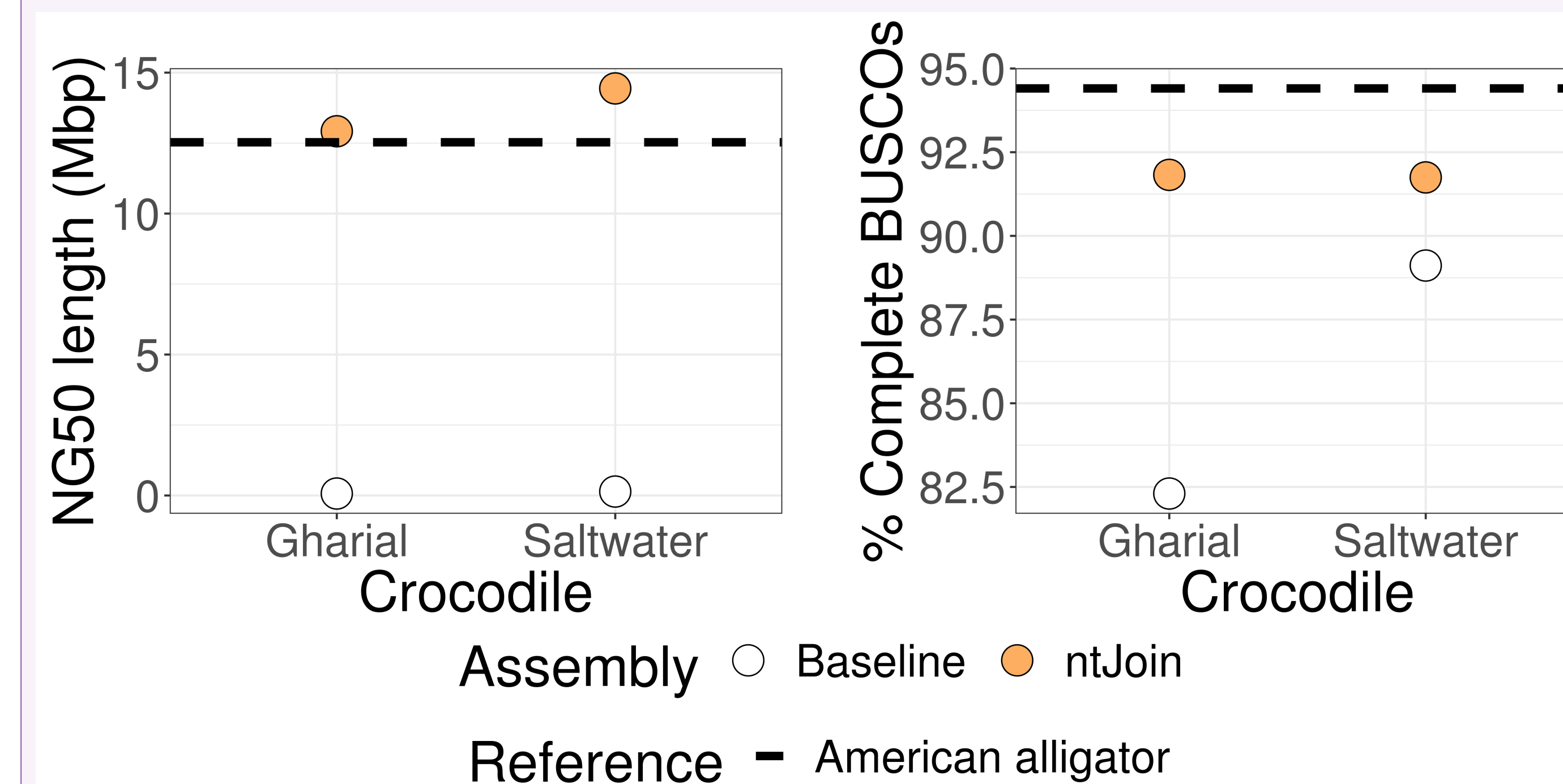
Approximate divergence between crocodile and alligator species: 80 million years

Target assemblies	Saltwater crocodile (*Crocodylus porosus*)
	Gharial crocodile (*Gavialis gangeticus*)
Reference	American alligator (*Alligator mississippiensis*)

Conclusions

ntJoin:
- Produces high-quality assemblies with minimizer graph-based scaffolding quickly and efficiently
- Numerous different applications, including hybrid assembly and population genomics studies
- *Recently published*: Coombe, L., et al. (2020) Bioinformatics: doi:10.1093/bioinformatics/btaa253

References

1. Roberts M, et al. (2004) Bioinformatics; 20:3363-3369.
2. Jackman SD, et al. (2017) Genome research; 27:768–77.
3. Shafin K, et al. (2020) Nature Biotechnology; doi:10.1038/s41587-020-0503-6.
4. Alonge M, et al. (2019) Genome Biology: 20: 224.
5. Kolmogorov M, et al. (2018) Genome research: 28:1720–32.
6. Simão FA, et al. (2015) Bioinformatics;31:3210–2.

Software Availability

https://github.com/bcgsc/ntjoin
brew install brewsci/bio/ntjoin
conda install -c bioconda ntjoin