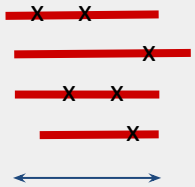
Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms

Nadège Guiglielmoni, Antoine Houtain, Alessandro Derzelle, Karine Van Doninck, Jean-François Flot

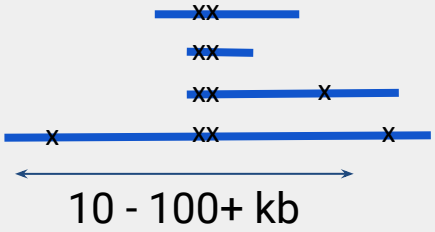
JOBIM 2020

Long-reads and genome assembly

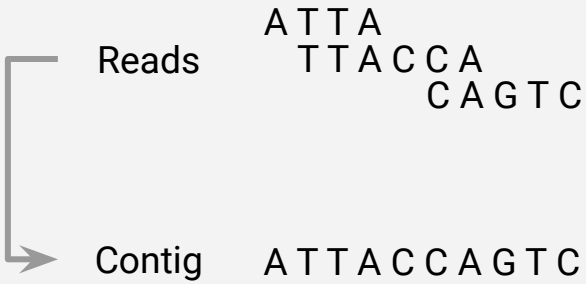
PacBio reads



Nanopore reads



Assembly process



Long-read assemblers

Canu	Ra	Shasta
Flye	Raven	wtdbg2
NextDeNovo		

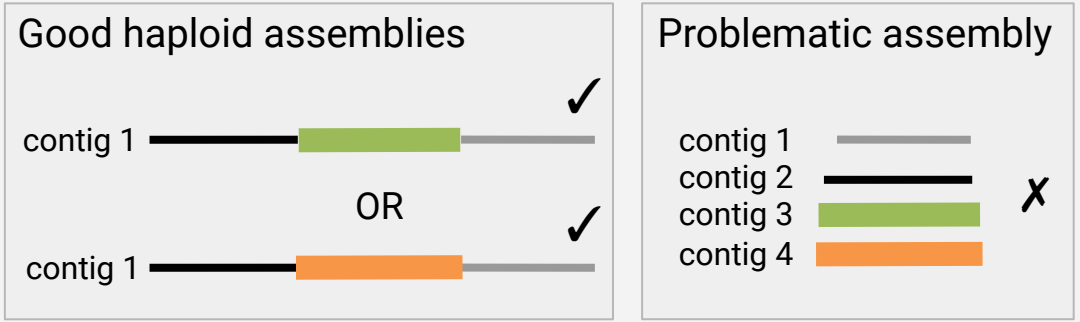
The problem of highly heterozygous regions

Haplotype 1 ATTACCAGTCTCAA**TGGATGGCTACTC**TTTGACGATAGCT

Haplotype 2 ATTACCAGTCTCAA**AGGCTGCTAGTG**TTTGACGATAGCT

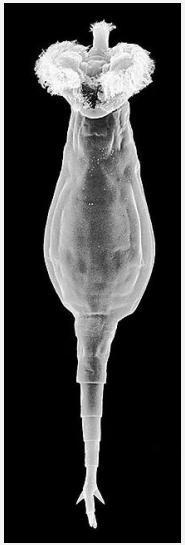
Assembly process

Assembly output

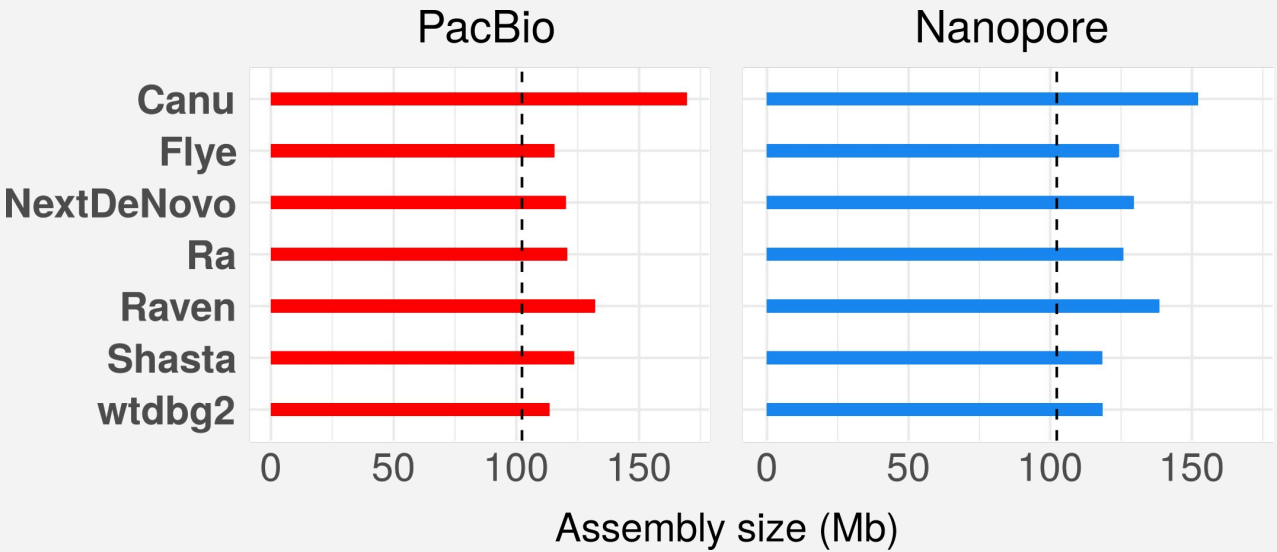


Symptoms of uncollapsed haplotypes

Assemblies of bdelloid rotifer *Adineta vaga* → expected haploid size 102.3 Mb



Who Needs Sex (or Males) Anyway?
Liza Gross, PloS Biology, 2007



Strategies to reduce uncollapsed haplotypes

- ➔ Strategy 1: choose a better assembler
- ➔ Strategy 2: removing uncollapsed haplotypes

  Tool: Purge Haplotigs

  Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies, Roach et *al.*, BMC Bioinformatics, 2018
- ➔ Strategy 3: select longest reads for assembly

2 long-read datasets for *Adineta vaga*:	**PacBio**	23.5 Gb, 230X
	Nanopore	17.5 Gb, 171X

Evaluation criteria

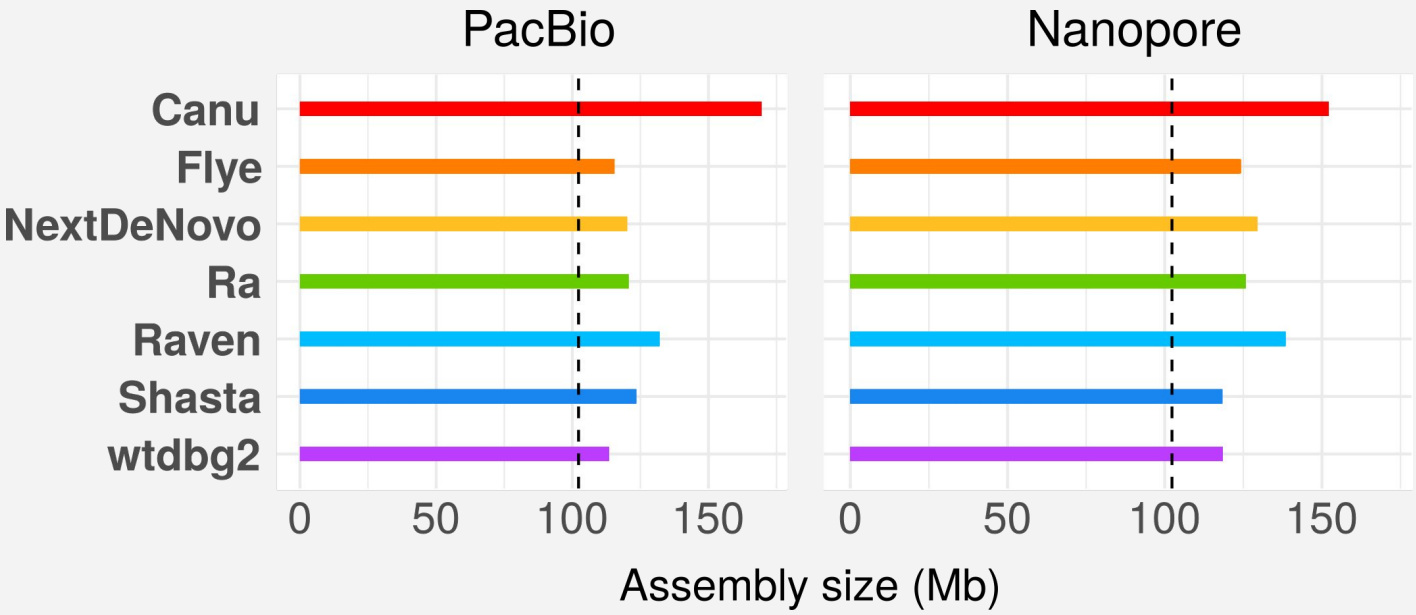
- **Assembly size**: sum of the lengths of all contigs, compared to the **estimated size of 102.3 Mb**

- **BUSCO score**: **number of orthologs** from a specific lineage (Metazoa) retrieved in the assembly, **either in a single-copy or duplicated**

- **Coverage**: number of reads covering a given position in a contig

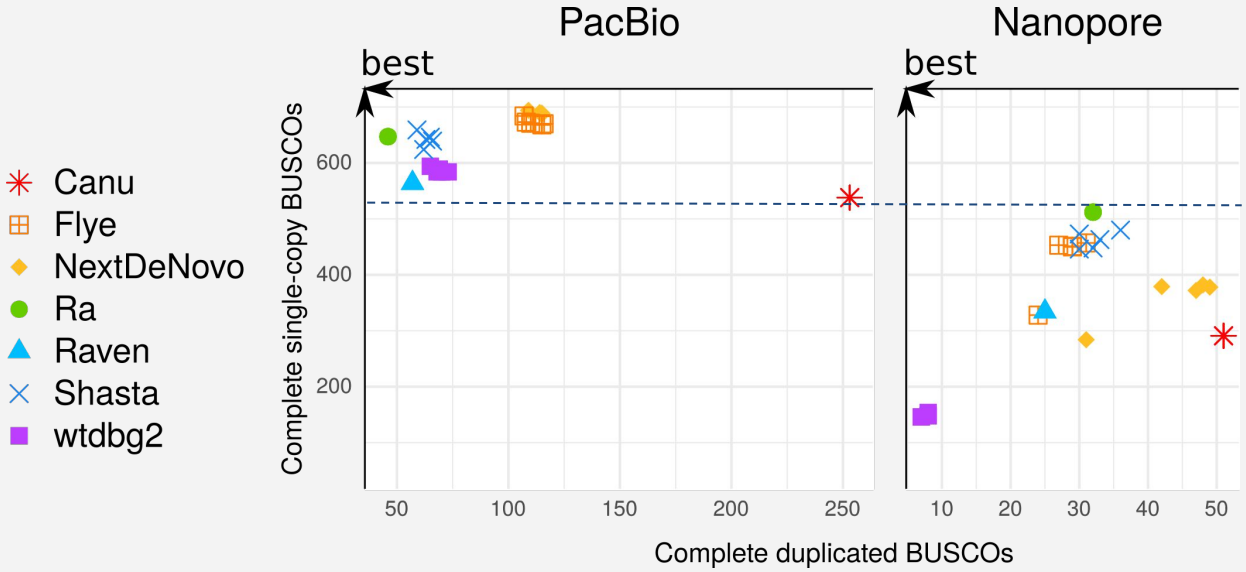
Evaluation criteria: assemblies of all reads

Assembly size: sum of the lengths of all contigs, compared to the estimated size of 102.3 Mb



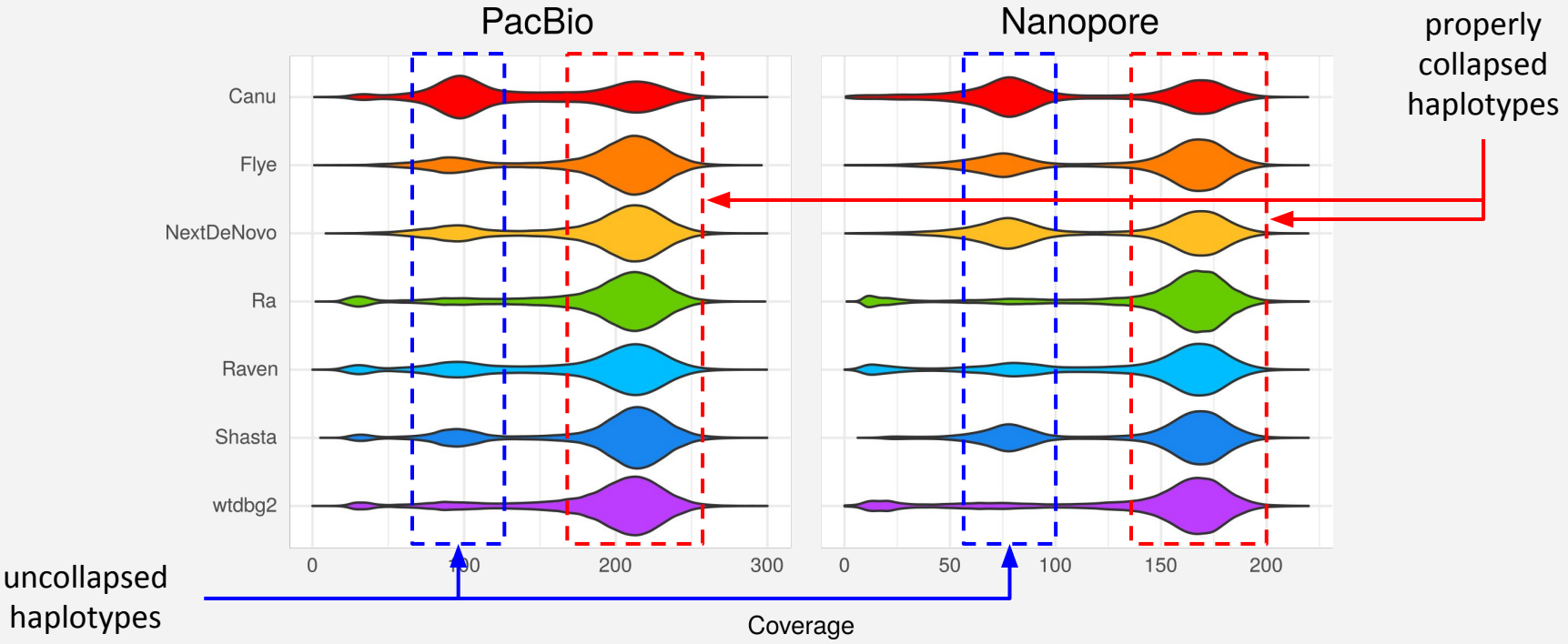
Evaluation criteria: assemblies of all reads

BUSCO score: number of orthologs from a specific lineage (Metazoa, 954 features) retrieved in the assembly, either in a single-copy or duplicated



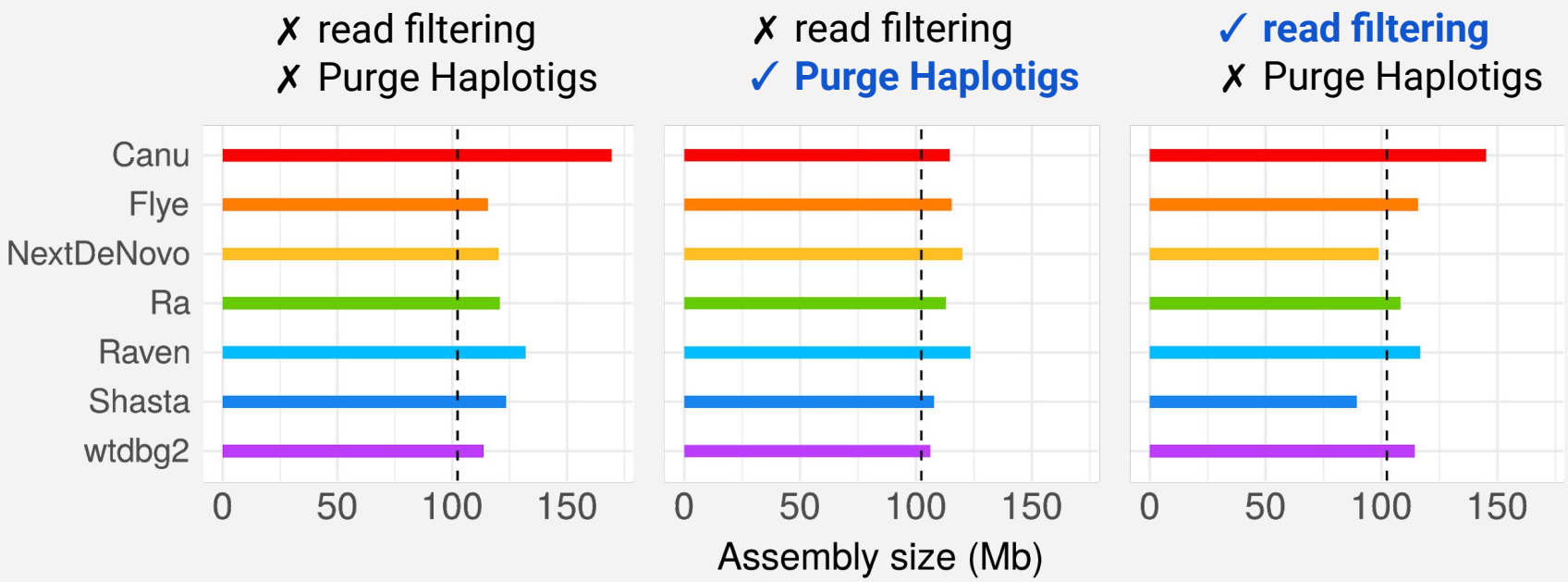
Evaluation criteria: assemblies of all reads

Coverage: number of reads covering a given position in a contig, based on long reads mapping



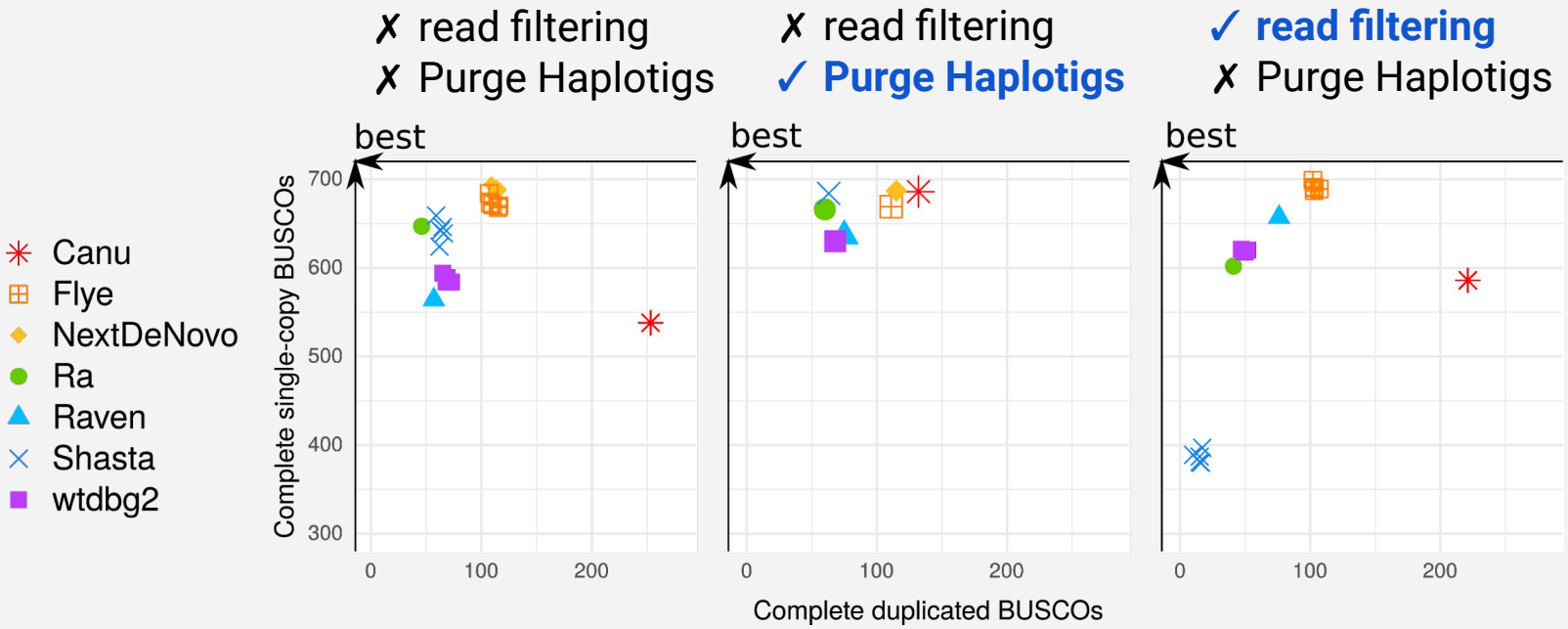
Collapsing haplotypes: PacBio assemblies

Assembly size: sum of the lengths of all contigs, compared to the estimated size of 102.3 Mb



Collapsing haplotypes: PacBio assemblies

BUSCO score: number of orthologs from a specific lineage (Metazoa, 954 features) retrieved in the assembly, either in a single-copy or duplicated



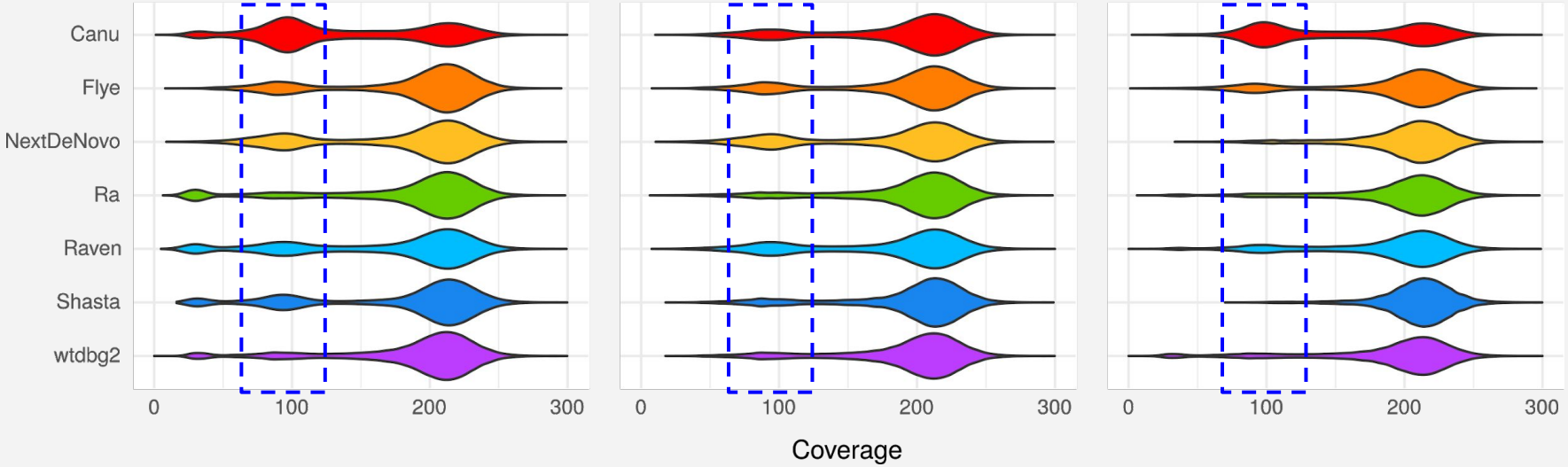
Collapsing haplotypes: PacBio assemblies

Coverage: number of reads covering a given position in a contig, based on PacBio reads mapping

✗ read filtering
✗ Purge Haplotigs

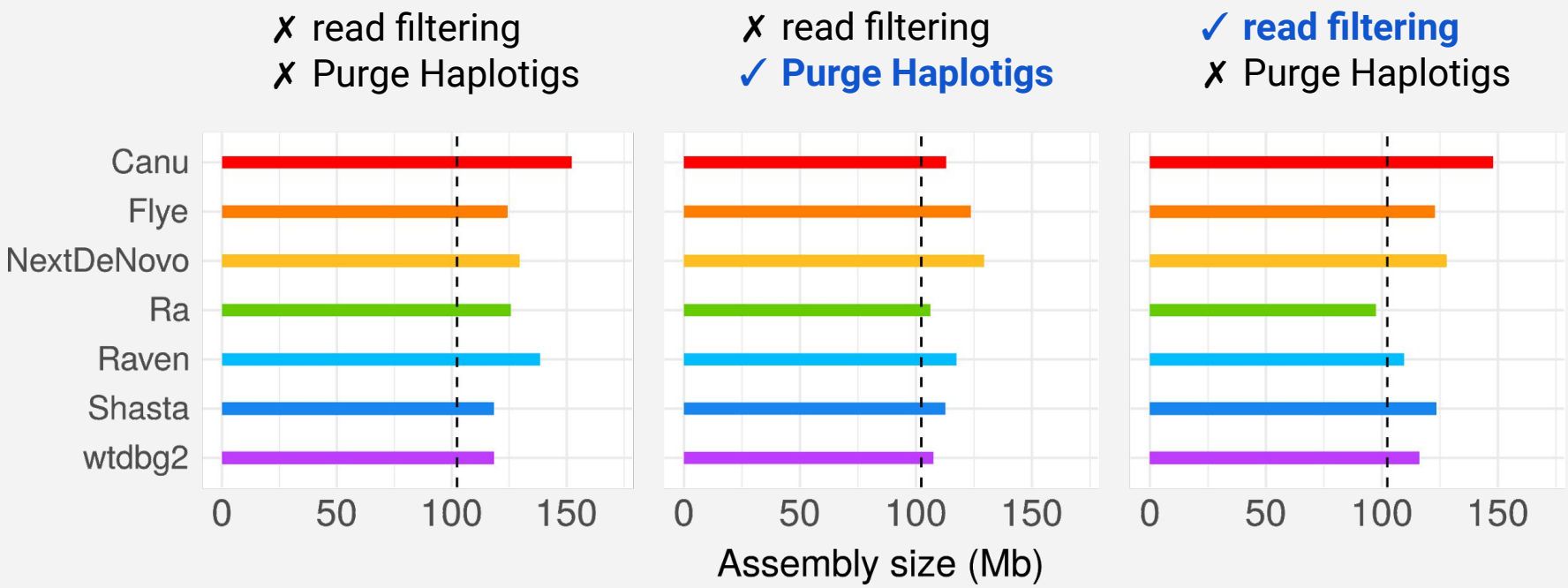
✗ read filtering
✓ **Purge Haplotigs**

✓ **read filtering**
✗ Purge Haplotigs



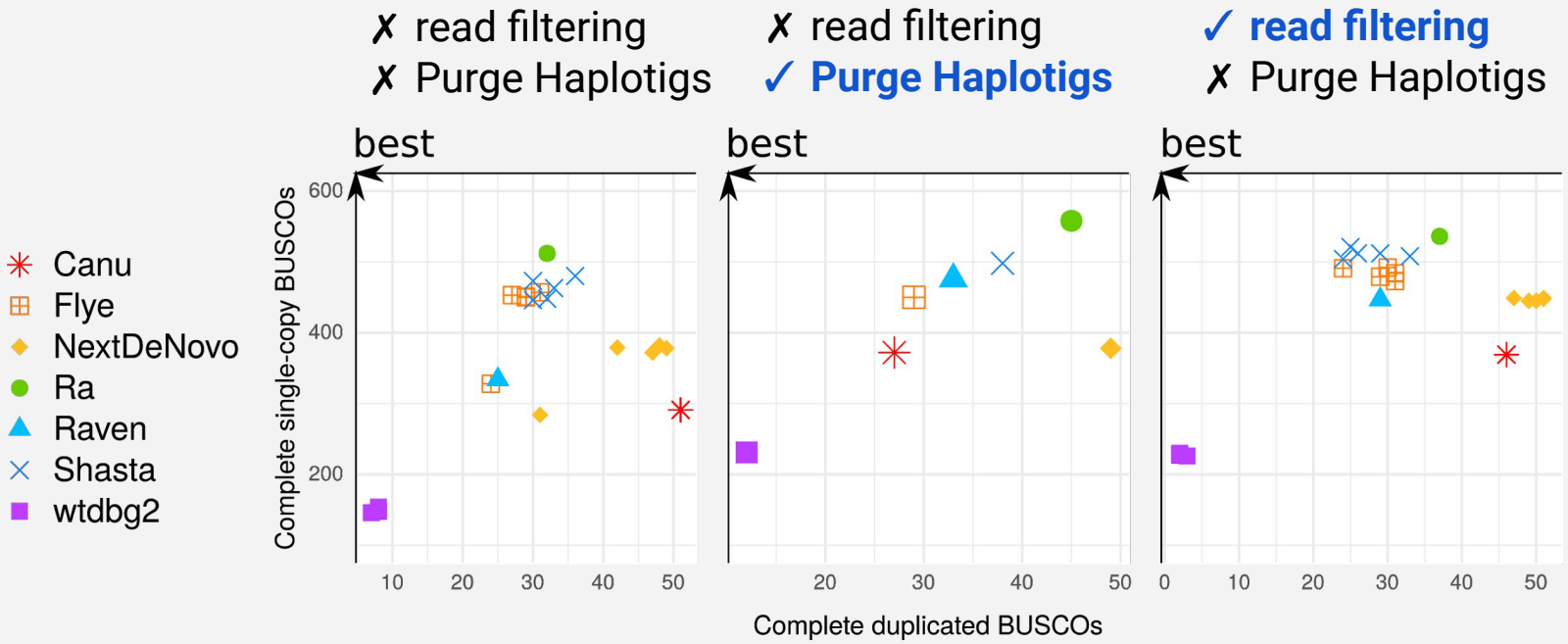
Collapsing haplotypes: Nanopore assemblies

Assembly size: sum of the lengths of all contigs, compared to the estimated size of 102.3 Mb



Collapsing haplotypes: Nanopore assemblies

BUSCO score: number of orthologs from a specific lineage (metazoa, 954 features) retrieved in the assembly, either in single-copy or duplicated



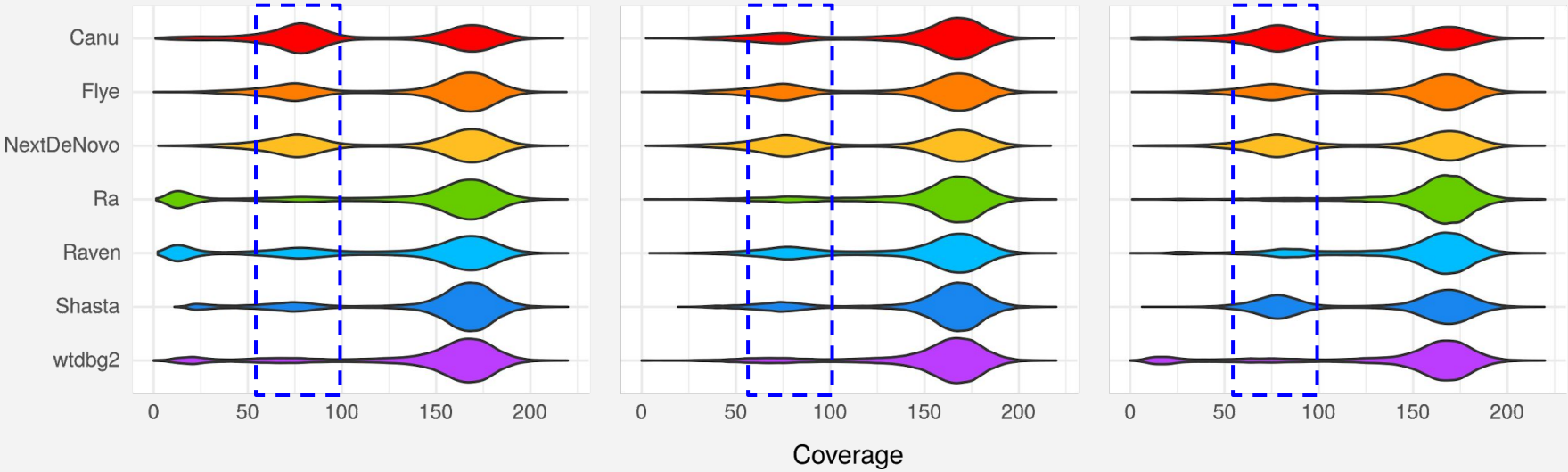
Collapsing haplotypes: Nanopore assemblies

Coverage: number of reads covering a given position in a contig, based on Nanopore reads mapping

✗ read filtering
✗ Purge Haplotigs

✗ read filtering
✓ **Purge Haplotigs**

✓ **read filtering**
✗ Purge Haplotigs



Take-home message

➔ Strategy 1: some assemblers are better at collapsing haplotypes (Ra, wtdbg2)

Take-home message

- ➔ Strategy 1: some assemblers are better at collapsing haplotypes (Ra, wtdbg2)
- ➔ Strategy 2: removing uncollapsed haplotypes works better on some assemblies than others

Take-home message

- ➔ Strategy 1: some assemblers are better at collapsing haplotypes (Ra, wtdbg2)
- ➔ Strategy 2: removing uncollapsed haplotypes works better on some assemblies than others
- ➔ Strategy 3: Read filtering improves structure and does not decrease quality

Take-home message

- ➔ Strategy 1: some assemblers are better at collapsing haplotypes (Ra, wtdbg2)
- ➔ Strategy 2: removing uncollapsed haplotypes works better on some assemblies than others
- ➔ Strategy 3: Read filtering improves structure and does not decrease quality
- ➔ Need for better assessment of assemblies ≠ contiguity

Take-home message

- ➔ Strategy 1: some assemblers are better at collapsing haplotypes (Ra, wtdbg2)
- ➔ Strategy 2: removing uncollapsed haplotypes works better on some assemblies than others
- ➔ Strategy 3: Read filtering improves structure and does not decrease quality
- ➔ Need for better assessment of assemblies ≠ contiguity
- ➔ There is not one measure to pick the best assembly

Acknowledgements

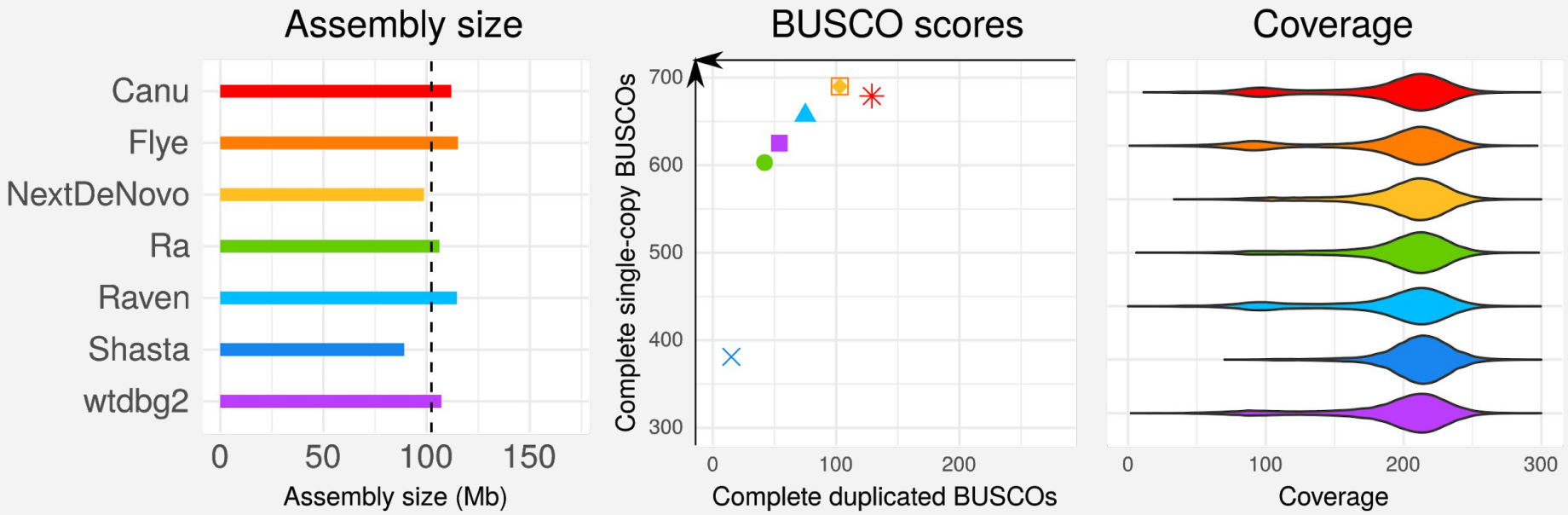
EBE, Université libre de Bruxelles
Jean-François Flot

Université de Namur
Karine Van Doninck
Antoine Houtain
Alessandro Derzelle
Paul Simion

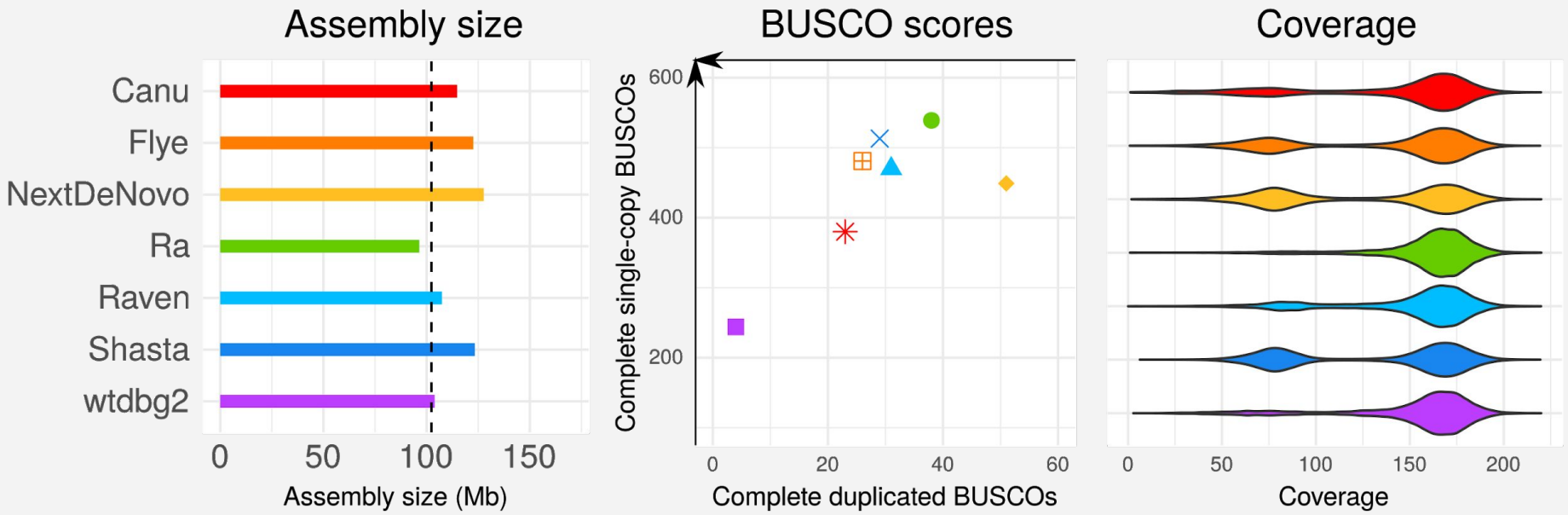
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840

Thank you for your attention!
Questions?

PacBio: Purge Haplotigs + read filtering



Nanopore: Purge Haplotigs + read filtering



Performance