State-of-the-art long-read overlapping tools comparative analysis

spoiler : they don't find the same overlaps

Pierre MARIJON[1] , Jean-Stéphane VARRÉ[2] and Rayan CHIKHI[2]

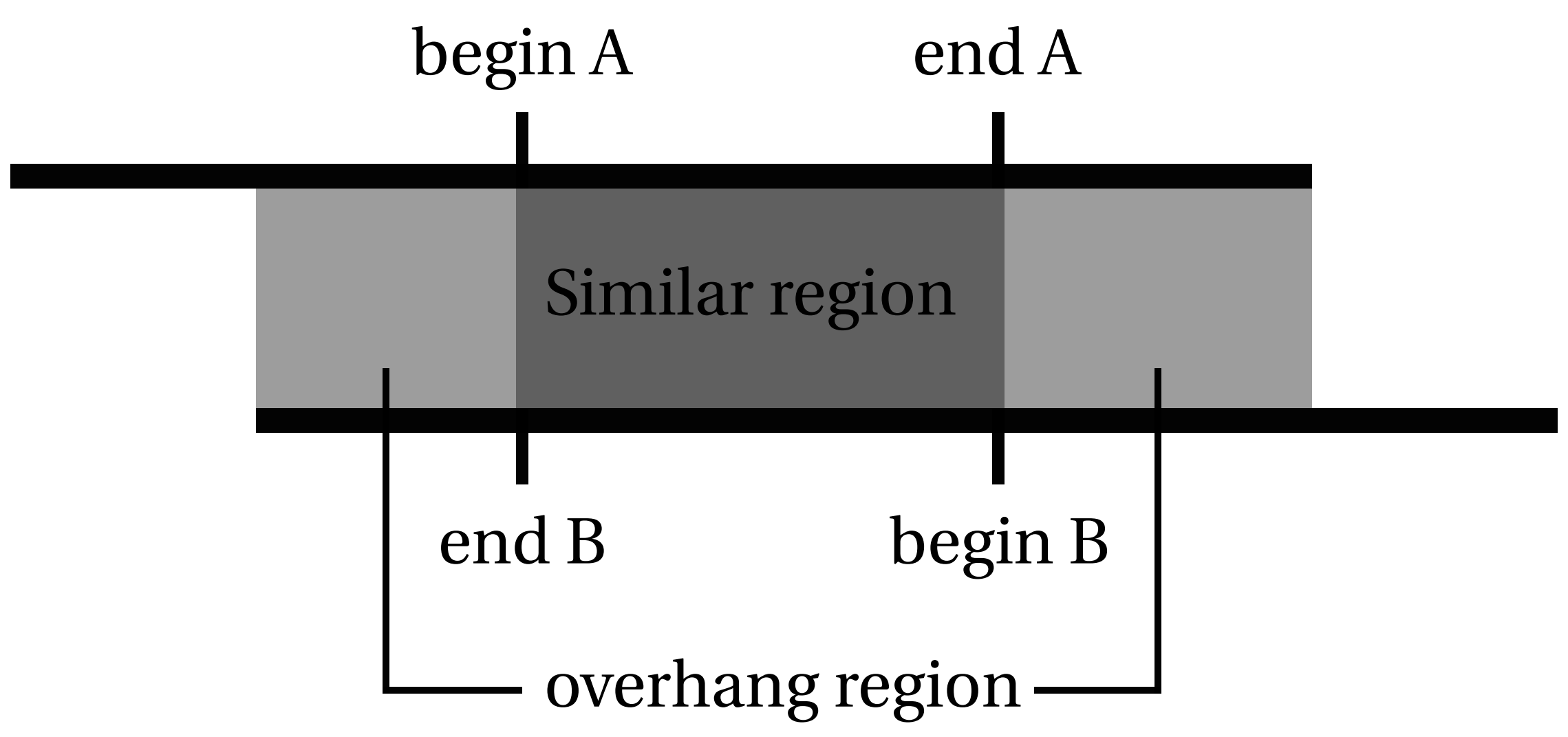
[1] Inria, Université de Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[2] Université de Lille, CNRS, Centrale Lille, Inria, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

Tool	Simulated PB *E. coli*		Simulated ONT *E. coli*		PB P6-C4 *E. coli*		ONT SQK-MAP-006 *E. coli*	
	Sensibility	Precision	Sensibility	Precision	Sensibility	Precision	Sensibility	Precision
BLASR	91.0	**81.9**	**95.2**	**75.1**	66.0	**96.5**	**89.9**	**73.0**
DALIGNER	**92.4**	91.9	94.9	97.6	**83.8**	85.8	**92.9**	91.0
MHAP	91.5	88.0	95.1	86.5	79.8	**79.8**	91.2	82.0
GraphMap	90.1	**96.5**	**90.4**	96.0	71.7	94.0	90.6	93.4
Minimap	**88.9**	94.8	94.6	**99.0**	**59.6**	83.8	91.2	**95.4**

In a previous work, Chu et al.[1] compared 5 long-read overlapping tools on 5 datasets (see left Table).

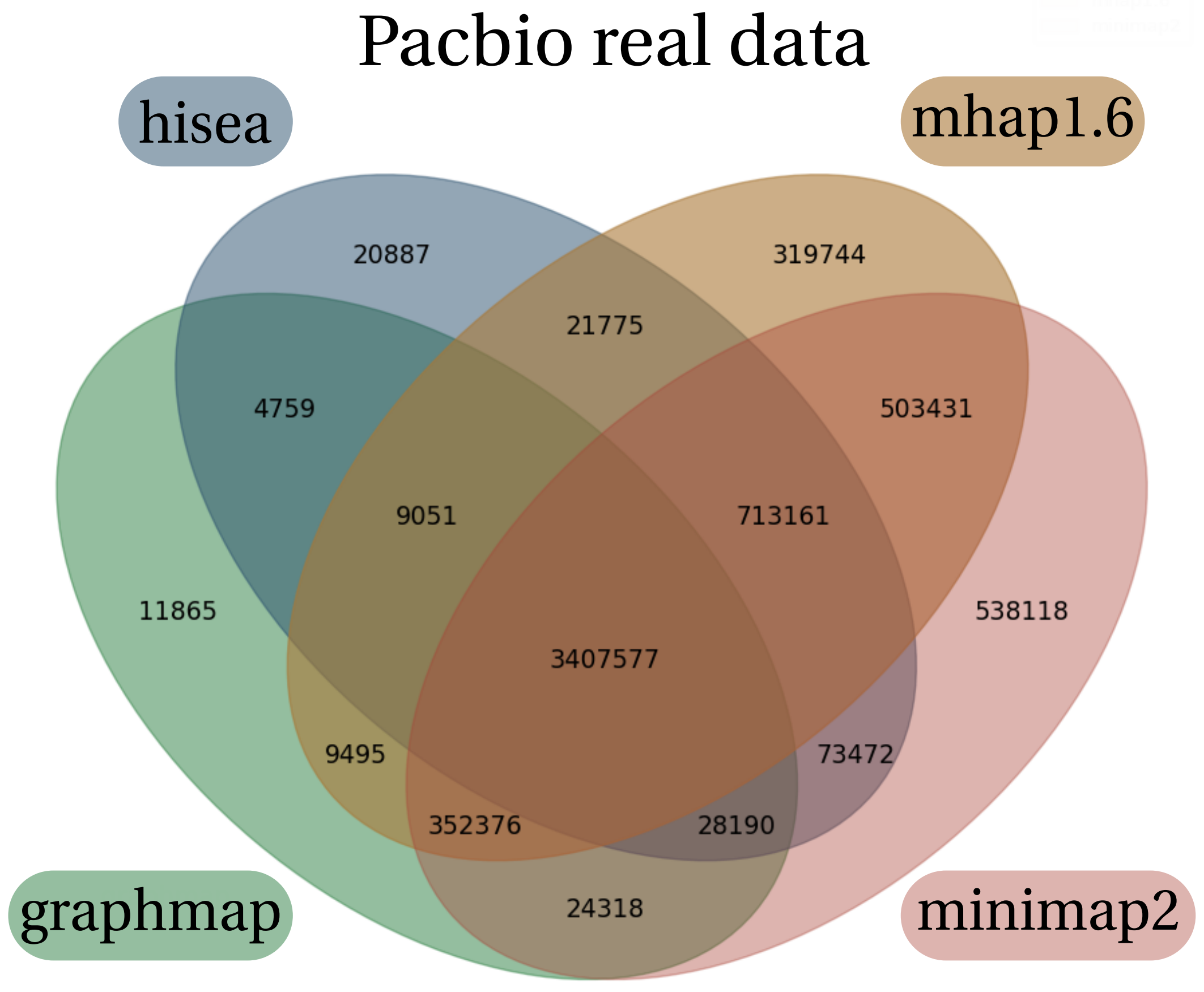
Overlappers showed **better results on synthetic datasets than on real data**.
Loss of sensitivity: 59.6-83.8% on the Pacbio real dataset,
versus 88.9-92.4% on the simulated data.



We will consider 3 types of overlaps, according to the definition found in the **minimap** article[2]:
Internal match: A similar region between two reads $|overhang\ region| > 0.8 \cdot |similar\ region|$
Containment: One read completely contained in another $beginA < beginB < endB < endA$
Classical overlap: A regular suffix-prefix overlap

In this study we will only consider **classical overlaps.** We will store overlaps as **pairs of reads**, without any consideration about overlap length and error rate.

Pacbio real data



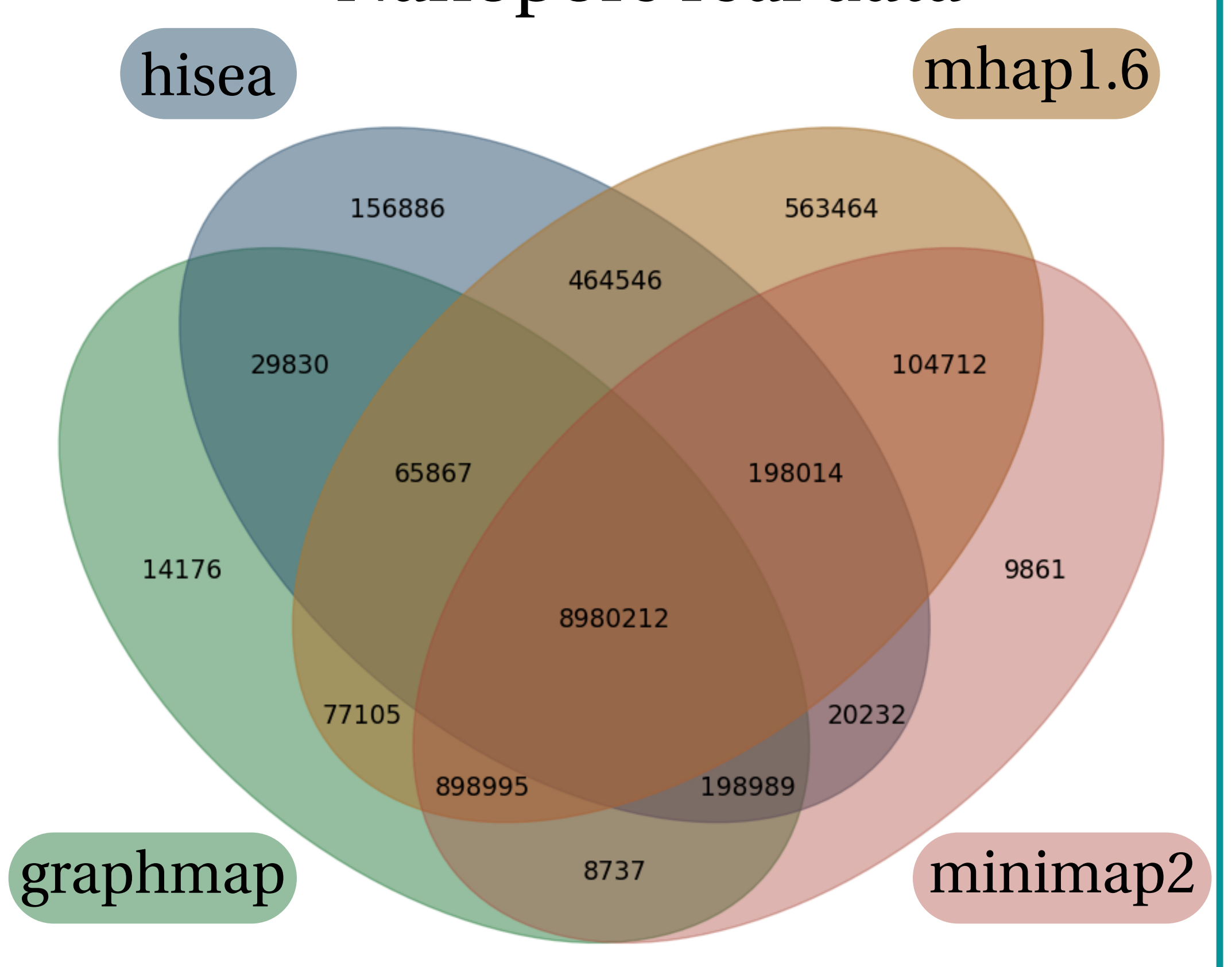
	mhap	minimap2	graphmap	hisea
mhap		0.83	0.70	0.76
minimap2	0.83		0.67	0.74
graphmap	0.70	0.67		0.74
hisea	0.76	0.74	0.74	

In the center of each diagram we can observe the number of overlaps shared by all overlappers. In the Tables below we report the Jaccard similarity coefficient (cardinality of intersection divided by cardinality of union) between two overlappers.

In **Pacbio real data** (left), out of all overlaps found by **minimap2** (5,640,643), **9.54%** of these overlaps are found **only** by this overlapper, for mhap the corresponding value is 5.98% (out of 5,336,610 overlaps).
In **Nanopore real data** (right), out of the 11,352,915 overlaps found by **mhap**, **4.96%** of these are found **only** by this overlapper. For hisea, the corresponding value is 1.55 % (out of 10,114,576 overlaps).

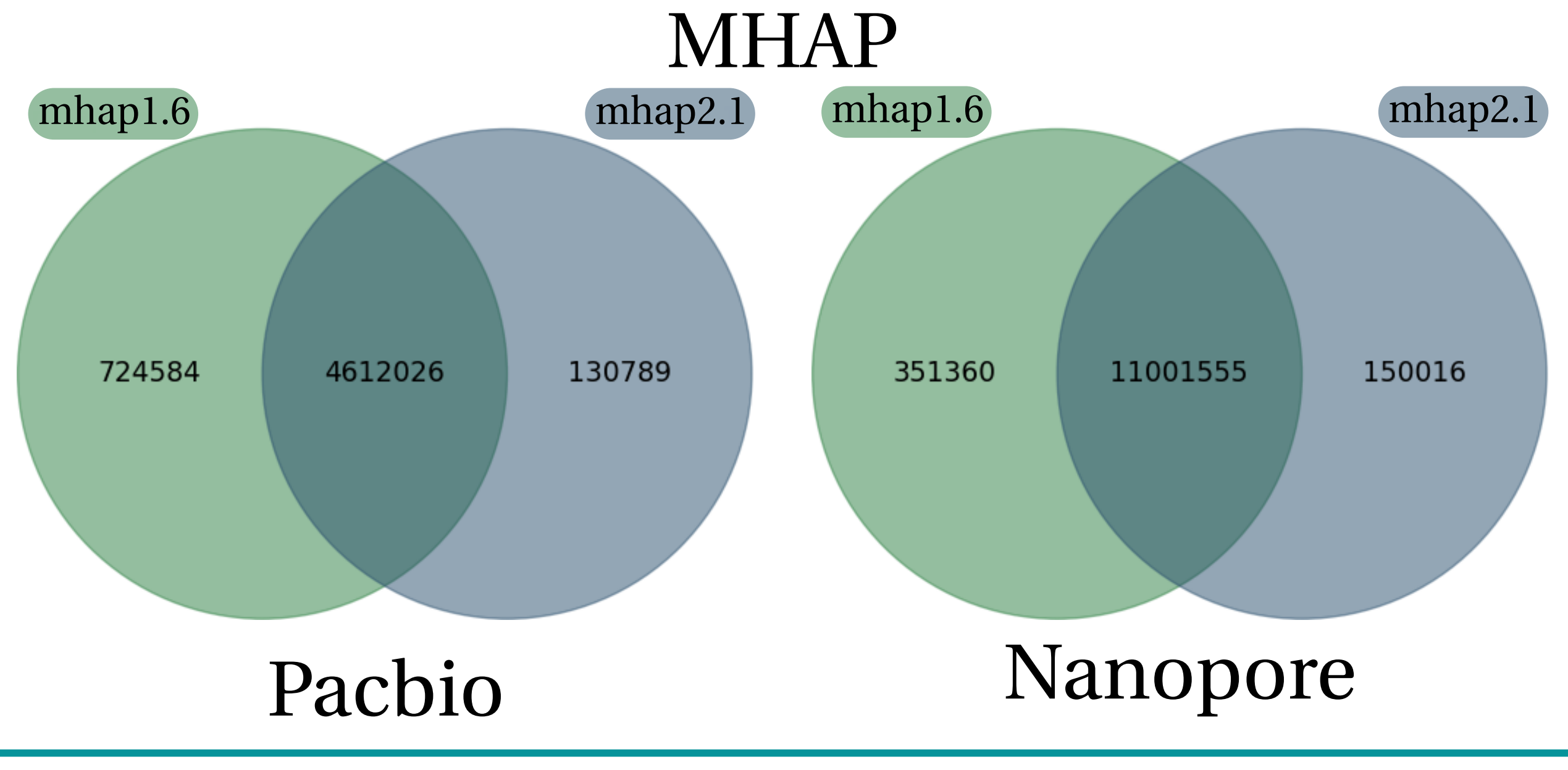
We also compared (below) different versions of MHAP and Minimap. We observe many differences, for example on the Pacbio dataset, the difference between results of **minimap vs. minimap2** is **higher** than **minimap2 vs. hisea** (0.71 and 0.74).

Nanopore real data

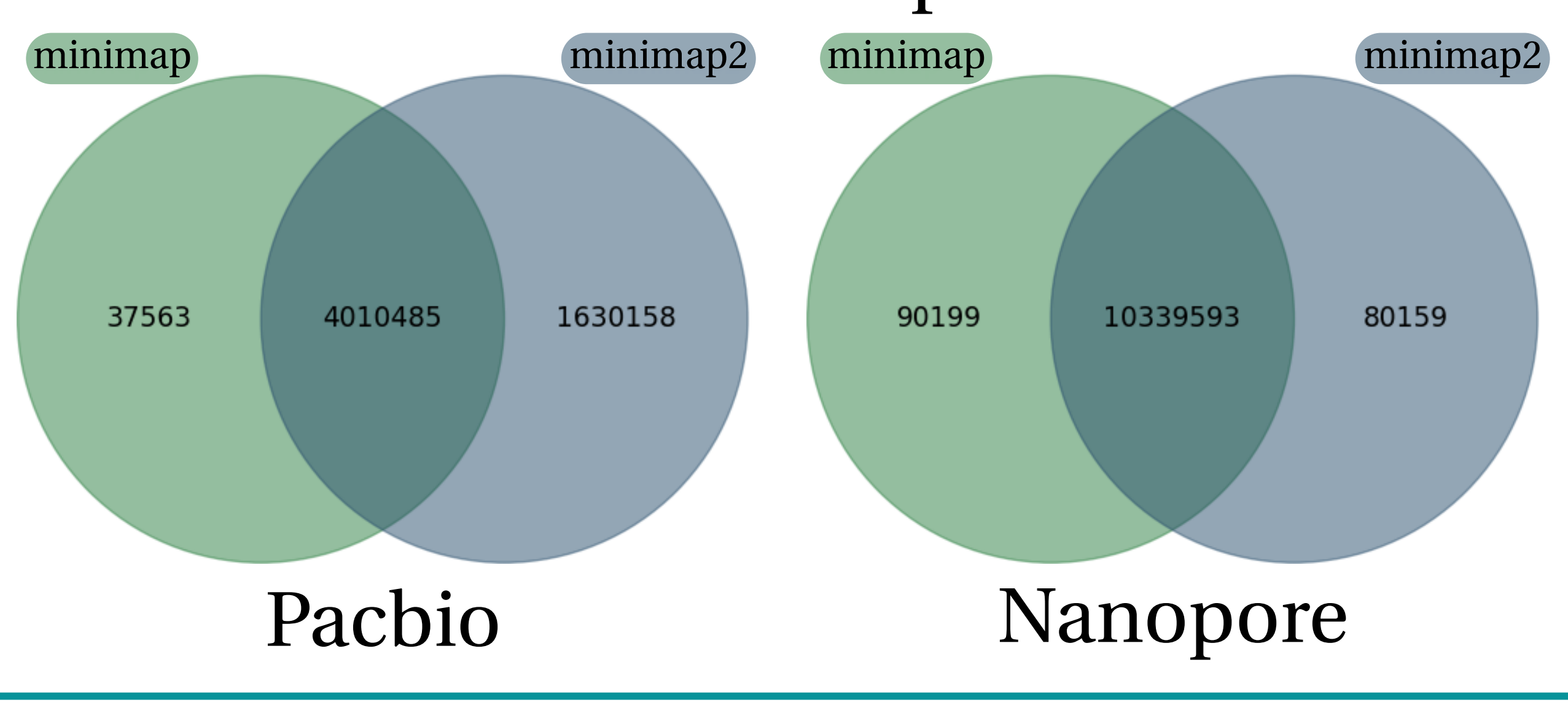


	mhap	minimap2	graphmap	hisea
mhap		0.88	0.85	0.82
minimap2	0.88		0.94	0.84
graphmap	0.85	0.94		0.83
hisea	0.82	0.84	0.83	

MHAP

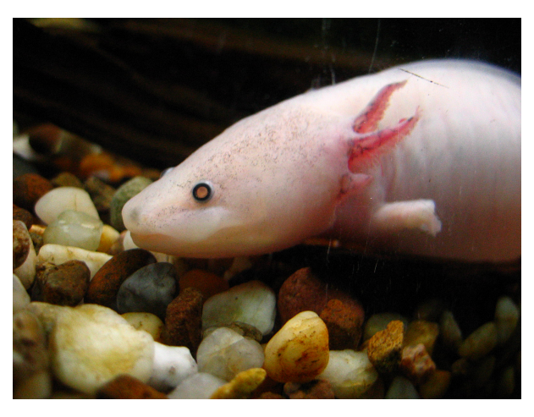


Minimap



Conclusion: Comparison of overlappers based on a quantitative measurement (**sensitivity**, **precision**) is useful but **not perfect.** Two tools with the same sensitivity for a given set could still detect a different set of overlaps, see e.g. mhap and minimap2 for the nanopore set. Moreover, two versions of the same tools can have more different results than two different tools.
Some publications use quality of error-correction, or results of genome assembly, as quality metrics to compare overlappers. It's a good idea but correction and assembly tools make additional choices in the overlaps they keep, and it's not easy to relate assembly or error-correction imperfections and wrong or missed overlaps..

This analysis was made in the context of assembly graphs for third generation assembly. We noticed that different assemblers start by computing different sets of overlaps. Creating a reconciliation tool for overlappers could be a good idea, while keeping in mind that correction and assembly tools seek to reduce the amount of overlaps they consider, through e.g. graph transitivity reduction, Best Overlap Graph, or the MARVEL[3] approach, use to assemble ->



You can find more information on my blog at *https://huit.re/jobim2018_blog*

Source code and instruction to redo analysis are available at *https://huit.re/jobim2018_repetition*

[1] Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art by Chu et al. 2016 doi:10.1093/bioinformatics/btw811

[2] Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences by Heng Li 2016 doi:10.1093/bioinformatics/btw152

[3] The axolotl genome and the evolution of key tissue formation regulators by Nowoshilow et al. 2018 doi:10.1038/nature25458