

# Mapping long reads to segmental duplications

Timofey Prodanov

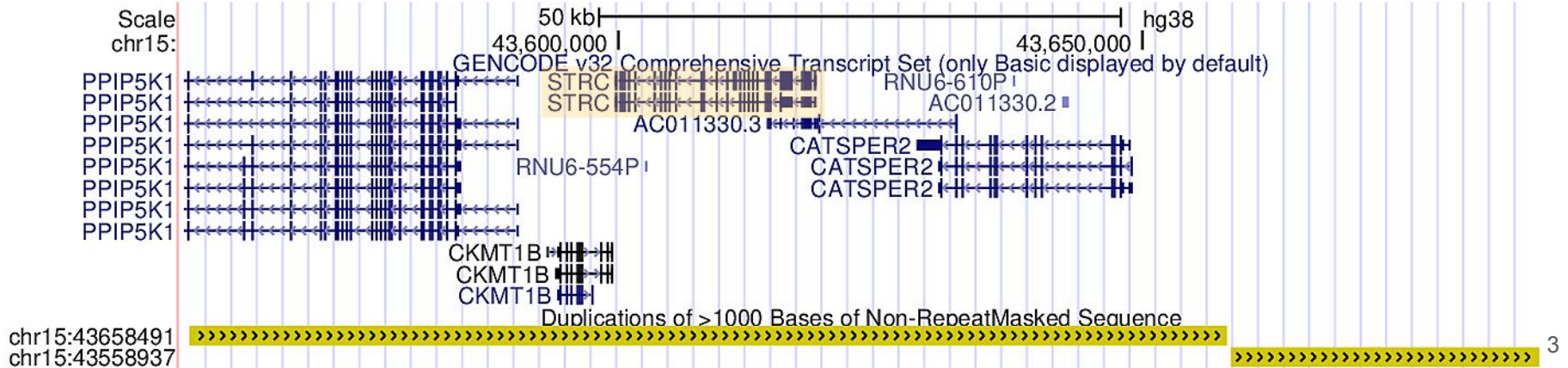
Vikas Bansal

# Segmental duplications

- The human genome is highly repetitive and contains long segmental duplications
- Many of them are longer than 10kb and greater than 98% sequence similarity to their other copies
- These duplications cover almost 4% of the human genome
- Overlap more than 600 protein-coding genes
- Some of the duplicated genes are implicated in Mendelian or complex diseases

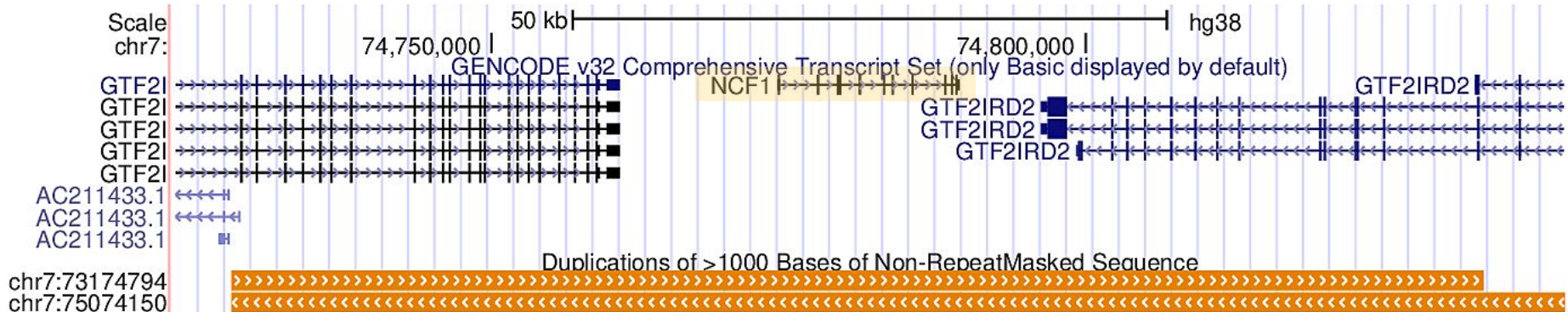
# STRC gene

- Encodes *stereocilin* protein, involved in hearing
- Mutations may lead to hearing problems, incl. hearing loss
- Completely lies within 101 kb duplication, 98% seq. similarity
- Duplication is tandem (one copy follows another)



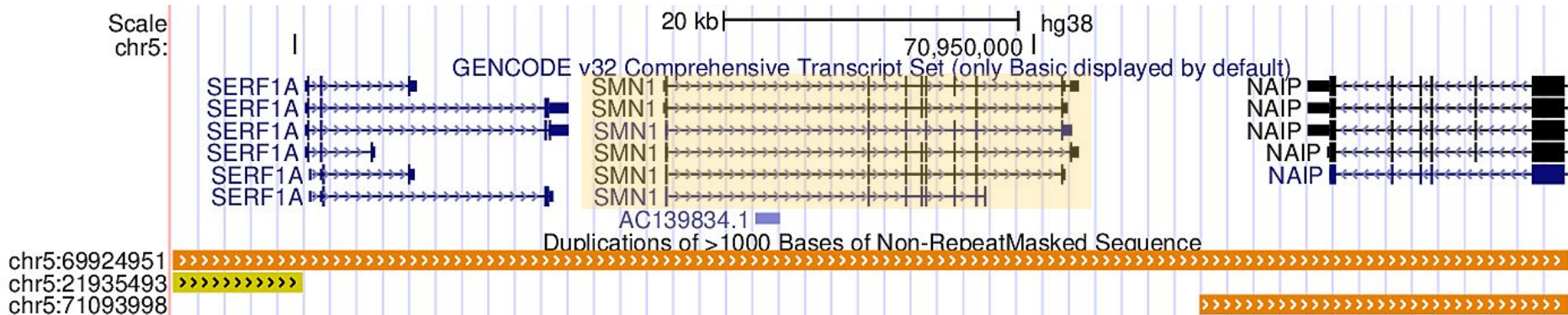
# NCF1 gene

- Encodes *neutrophil cytosolic factor 1* protein
- Plays a role in the immune system
- Mutations are associated with the *Chronic granulomatous disease*, and overall weaken immune system



# SMN1/2 genes


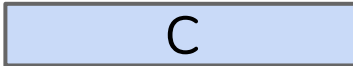
- Encodes *survival motor neuron* protein
- Mutations and copy number variations can lead to *spinal muscular atrophy*
- Lies within 205 kb duplication with 99.8% seq. similarity







# Paralogous Sequence Variants (PSVs)

PSV – small sequence difference between repeat copies





Often coincide with a polymorphism. For example:

Reference: Copy 1   
Copy 2 

Reliable PSV (genotype consistent with reference):

Copy 1   
  
Copy 2   


Unreliable PSV:

Copy 1   
  
Copy 2   


# Sequencing technologies

## Short-read sequencing (Illumina)

- Length 100 bp - 300 bp, usually paired reads
- Very low error rate: < 1%

## Long-read sequencing

PacBio CLR, PacBio HiFi, Oxford Nanopore (ONT), Ultralong ONT.

- Length ~ 10-15 kb
- Error rate ~ 12-15%.

Exceptions:

- PacBio HiFi: error rate < 1%
- Ultralong ONT: mean length ~ 50kb

# Segmental duplications features

- Many regions are unmappable using short reads
- > 50% of duplicated genes show extensive copy number variation
- Many PSVs are unreliable



# Existing long-read aligners

- Minimap2
- BLASR

Common algorithm:

- Find exact matches with the reference (minimizers)
- Complete the alignment between the minimizers

Other modifications:

- NGMLR: accounts for small structural variations
- Winnowmap: uses frequent minimizers to map to extra-long tandem repeats

# Mapping to segmental duplications

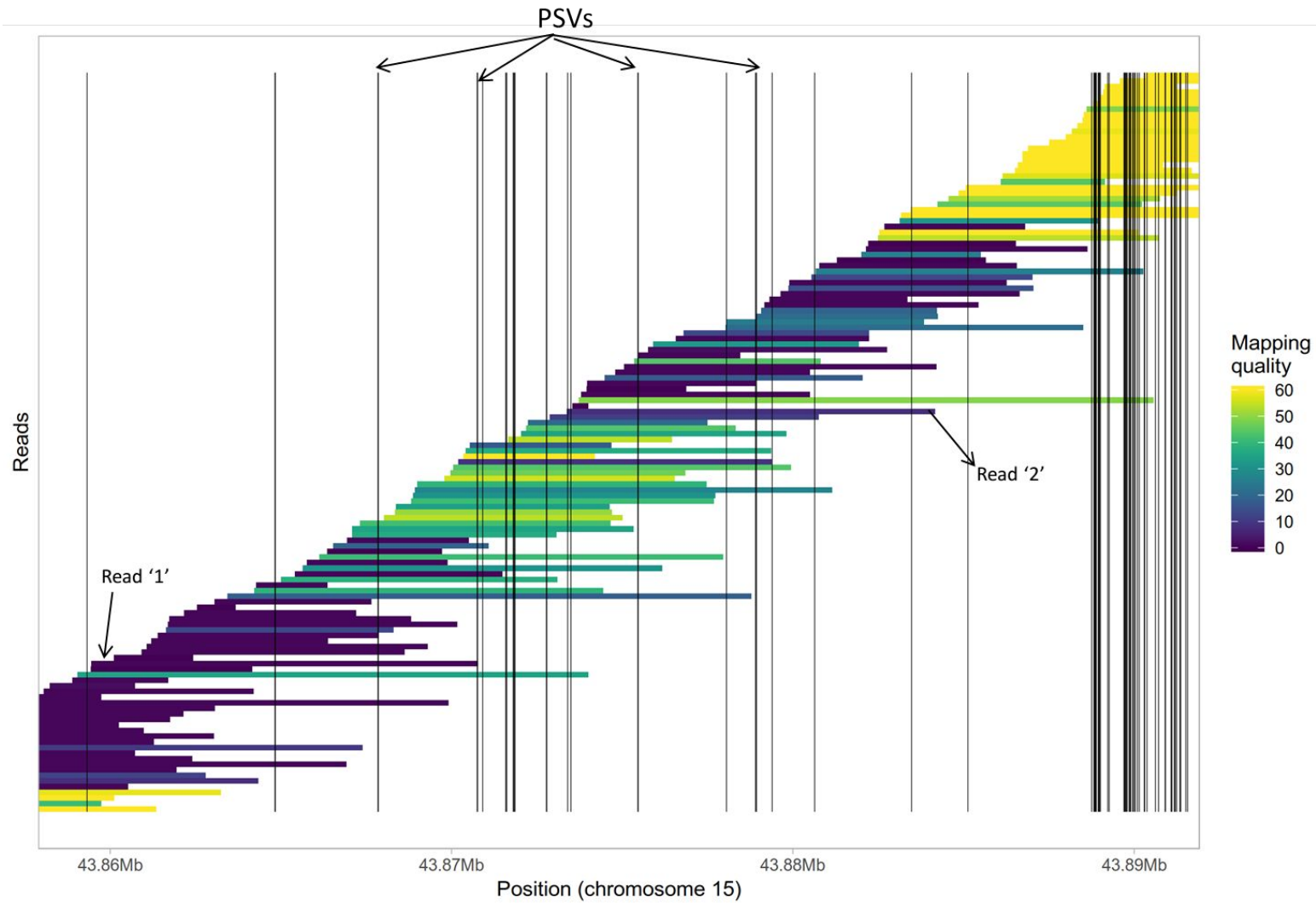
Each alignment location get a **score**,

Location with the best score – **primary** alignment,

**Mapping quality** is based on difference between location scores.

Standard aligners use PSVs implicitly, therefore

- All PSVs are assumed reliable,
- Hard to say if alignment score difference is random.



# Problem statement

Given a set of reads aligned to segmental duplications, we need to find

- New read alignments,
- PSV genotypes

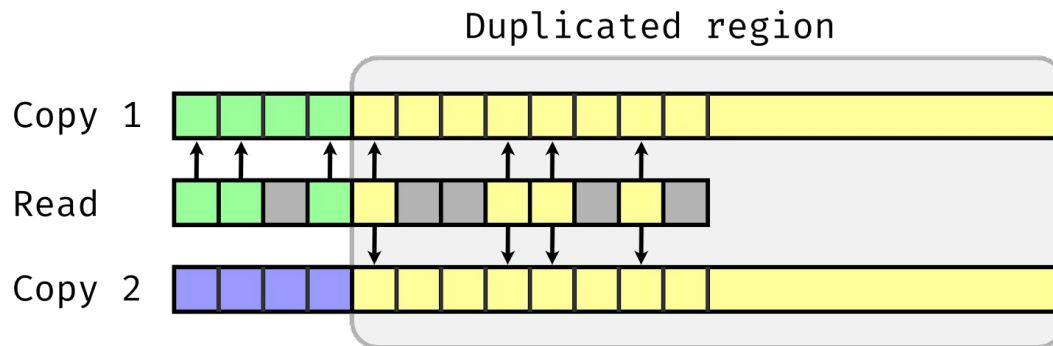
that maximize the agreement between PSVs and read alignments.

# Method overview

1. Construct PSV database.  
Only keep long duplications with high sequence similarity (101 Mb for hg38)
2. For each read overlapping segmental duplications:
  - a. Find and filter candidate alignment locations
  - b. Construct local read-PSV alignments
3. Iteratively:
  - a. Find best read locations and their probabilities
  - b. Genotype PSVs.

## Find and filter candidate locations (2.a)

- Find candidate locations using initial alignment and alignments between copies.
- Use LCS to discard some locations  
(read overlaps unique region or diverse region of the duplication).



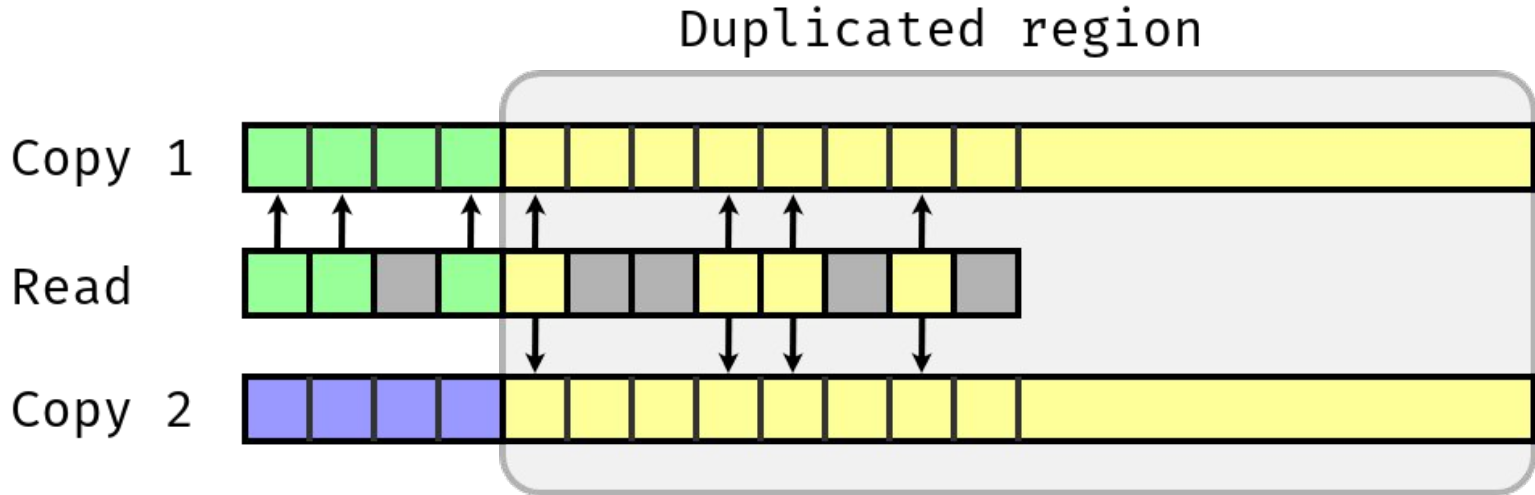
$$\text{LCS}(\text{read}, \text{copy 1}) \setminus \text{LCS}(\text{read}, \text{copy 2}) = 3$$

$$\text{LCS}(\text{read}, \text{copy 2}) \setminus \text{LCS}(\text{read}, \text{copy 1}) = 0$$

# Filtering locations

Calculate LCS using fast algorithm LCSk++.

Use LCS because less problems with low complexity regions, and we can use smaller k.

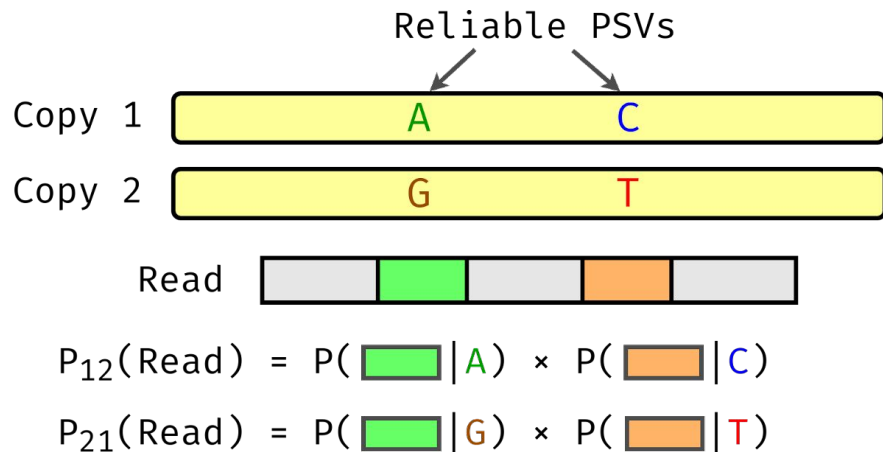


## Read-location probabilities (3.a)

For each read calculate alignment probability around each PSV

Read-location probability = product of all read-PSV probabilities

Scale PSVs by their reliability



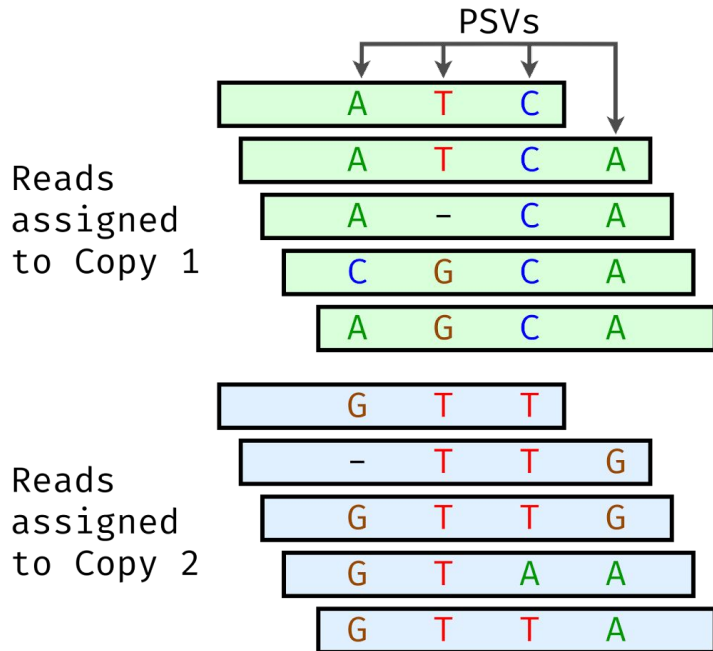


# Estimating PSV genotypes (3.b)

Use reads with high probability for one of the copies

Estimate most likely genotypes

Reliable PSVs have reference allele on all copies



	Reference:			
Copy 1	A	G	C	A
Copy 2	G	T	T	G
	Most likely genotypes:			
Copy 1	A/A	T/G	C/C	A/A
Copy 2	G/G	T/T	T/T	A/G
Reliable:	yes	no	yes	no

# Accuracy on simulated data

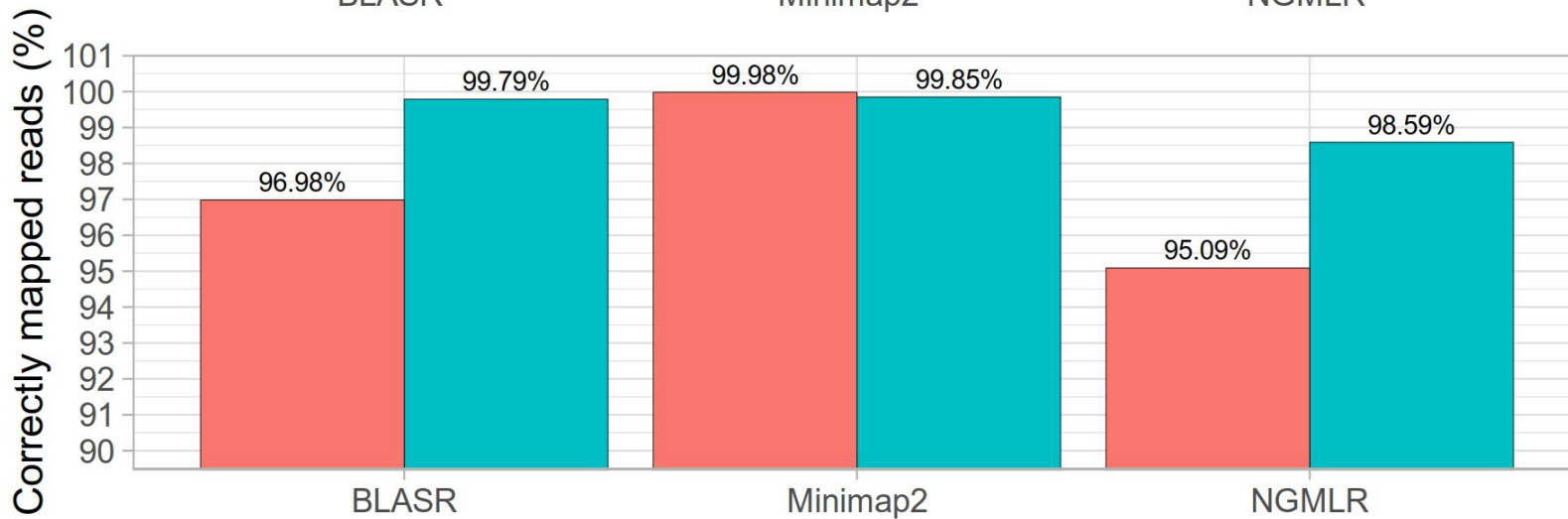
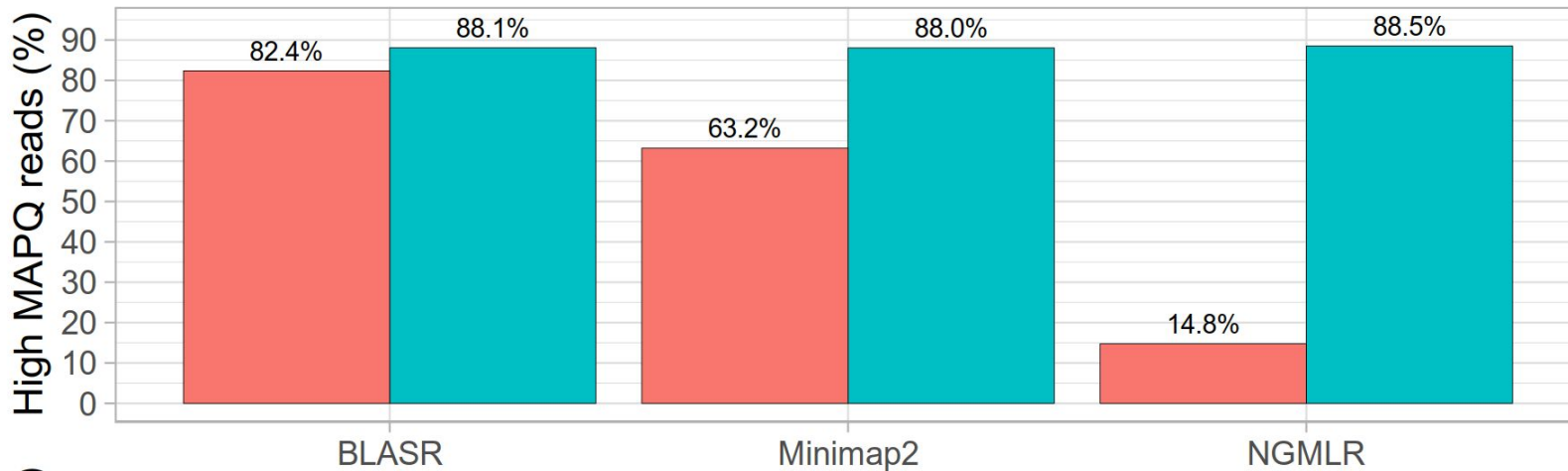
We used SimLoRD to generate reads.

Mapped with Minimap2 and BLASR,  
then remap with DuploMap.

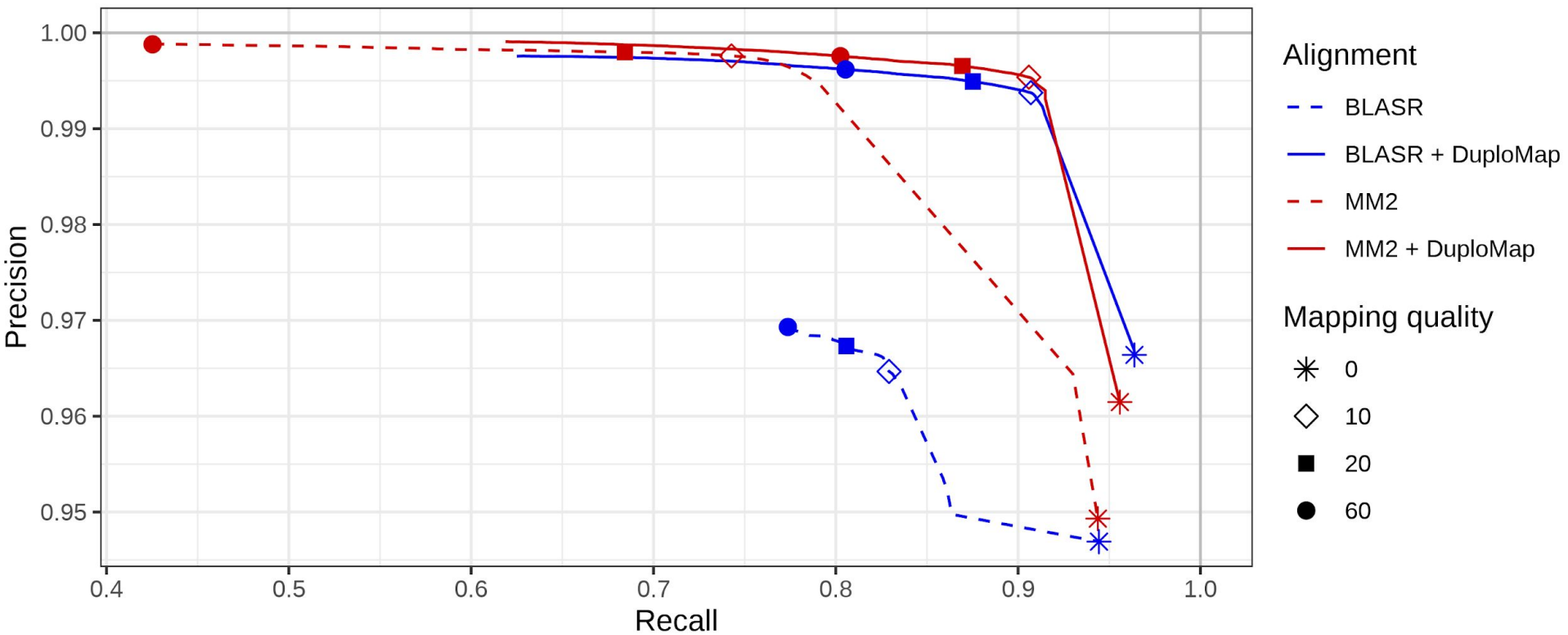
Two metrics:

- Precision =  $\frac{\text{reads correctly mapped to segm.dupl. with high MQ}}{\text{reads mapped to segm.dupl. with high MQ}}$
- Recall =  $\frac{\text{reads correctly mapped to segm.dupl. with high MQ}}{\text{all reads generated within segm.dupl.}}$

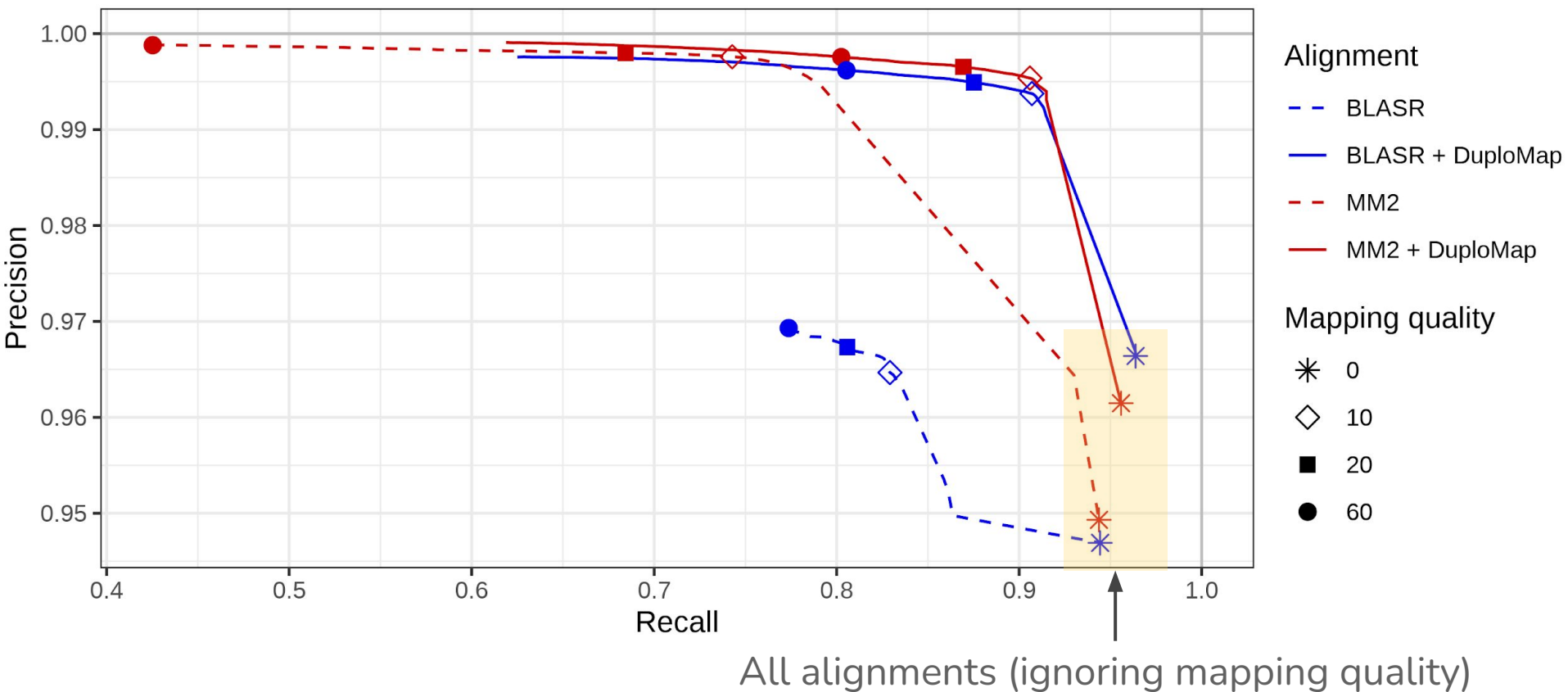
# Simulated reads (mean length ~ 8.5kb). Mapping evaluation.



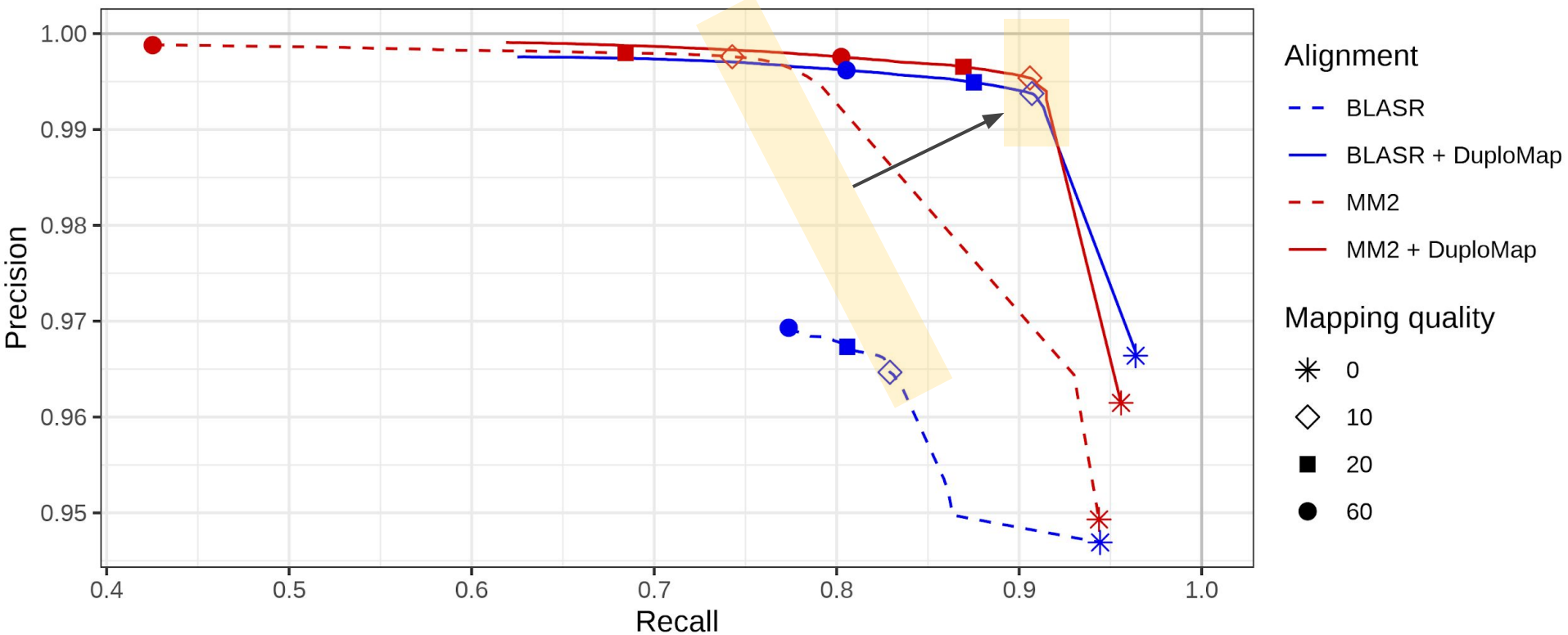
# Precision-recall curves



# Precision-recall curves



# Precision-recall curves

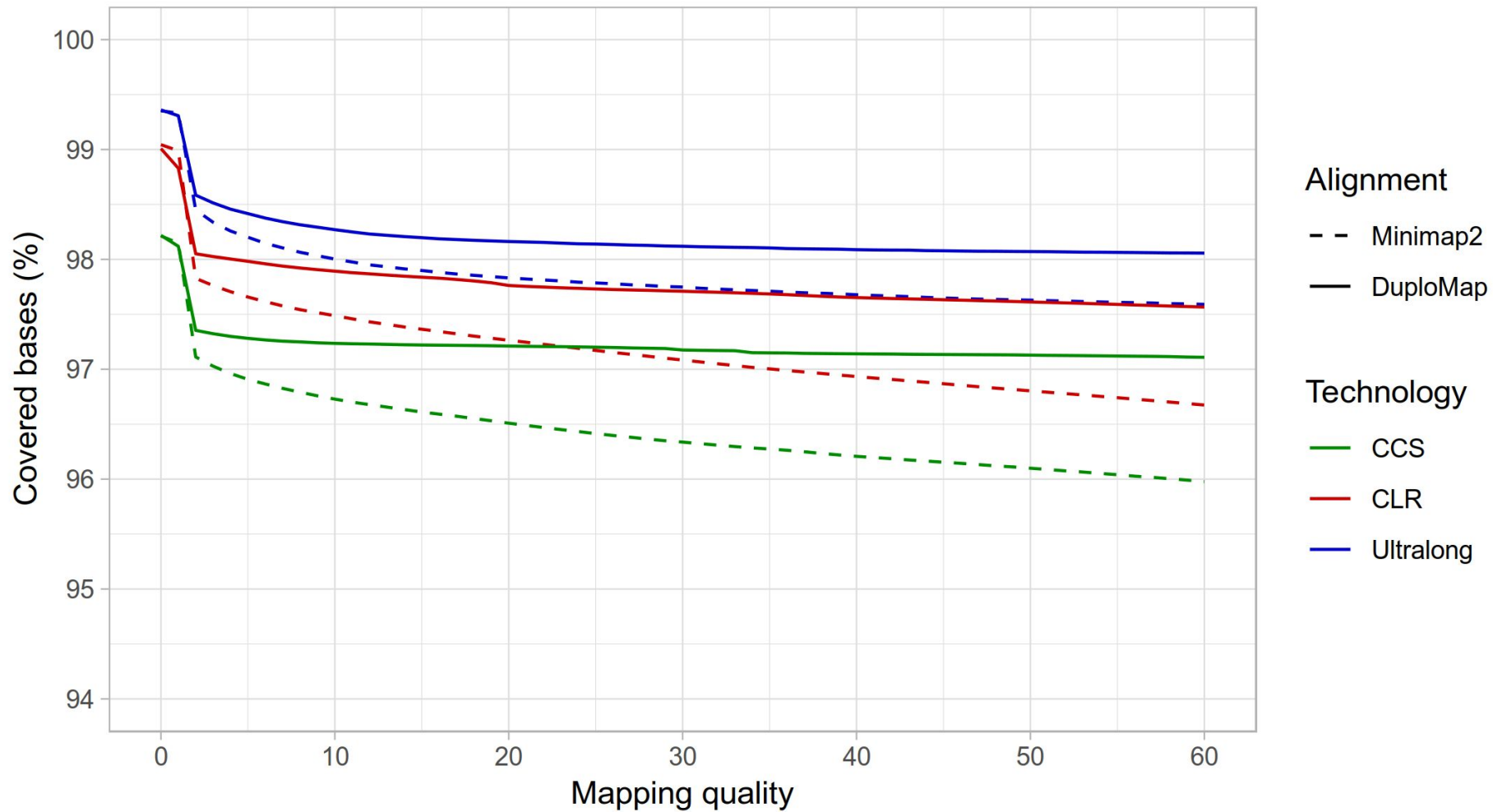


# Real data

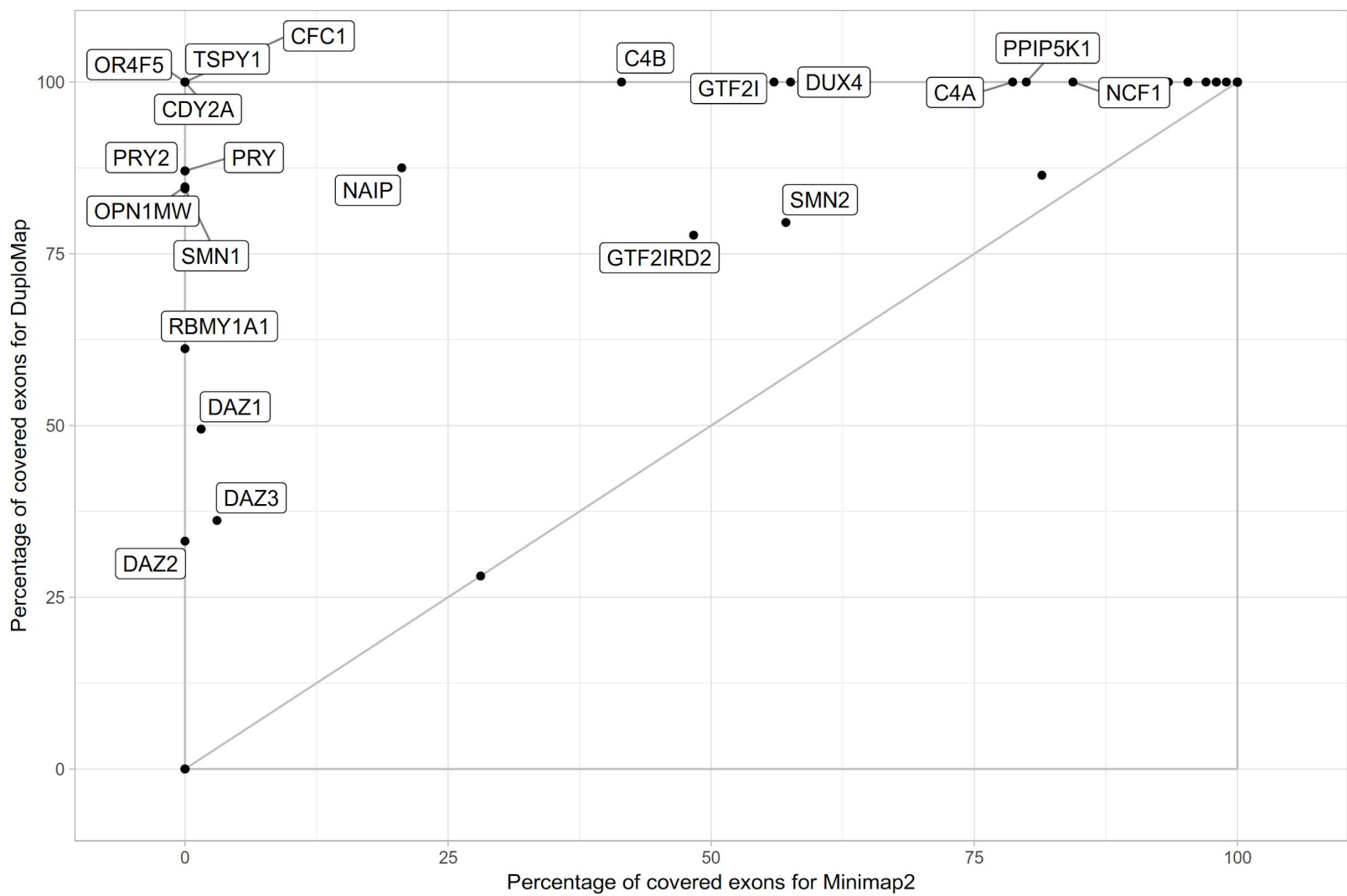
Looked at 8 datasets with 4 sequencing technologies  
(PacBio CLR, PacBio HiFi, ONT, Ultralong ONT)

Genome	Sequencing technology	Read length (N50)	MM2 (%)		$\Delta$ MM2+Duplomap (%)	
			MQ $\geq$ 10	MQ $\geq$ 20	MQ $\geq$ 10	MQ $\geq$ 20
HG002	PacBio CLR	11.3k	59.4	52.9	+8.4	+10.7
HG003	PacBio CLR	11.0k	59.9	53.5	+9.8	+11.3
HG004	PacBio CLR	10.9k	65.1	58.3	+8.7	+10.5
HG002	PacBio HiFi	13.5k	65.7	58.9	+14.9	+19.5
HG005	PacBio HiFi	10.4k	64.2	56.6	+15.8	+20.7
HG001	PacBio HiFi	10.0k	71.6	63.7	+15.0	+21.2
HG001	ONT	13.8k	63.5	55.7	+3.9	+7.8
HG002	ONT	54.3k	64.5	58.0	-1.5	+1.7

# HG002. Whole genome. hg38

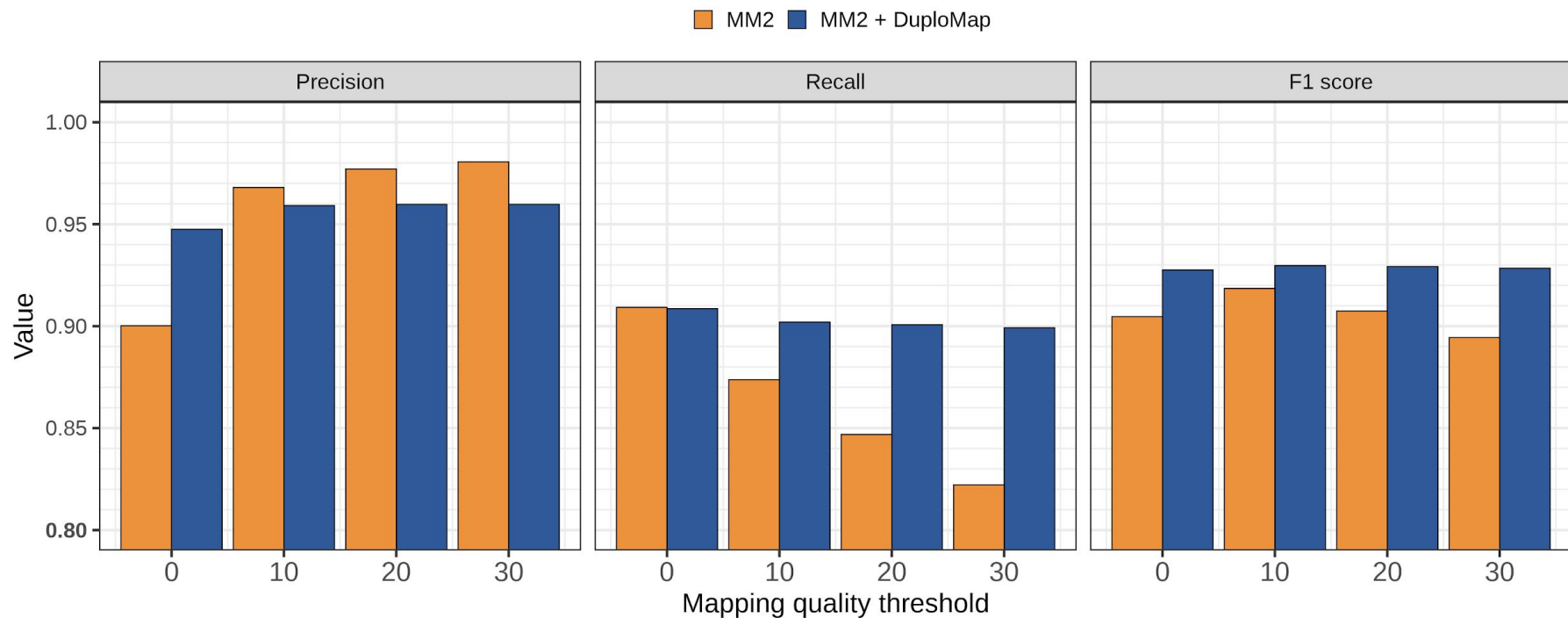






# Variant calling after realignment (PacBio HiFi)

Compare with a high-confidence benchmark variant calls for HG002

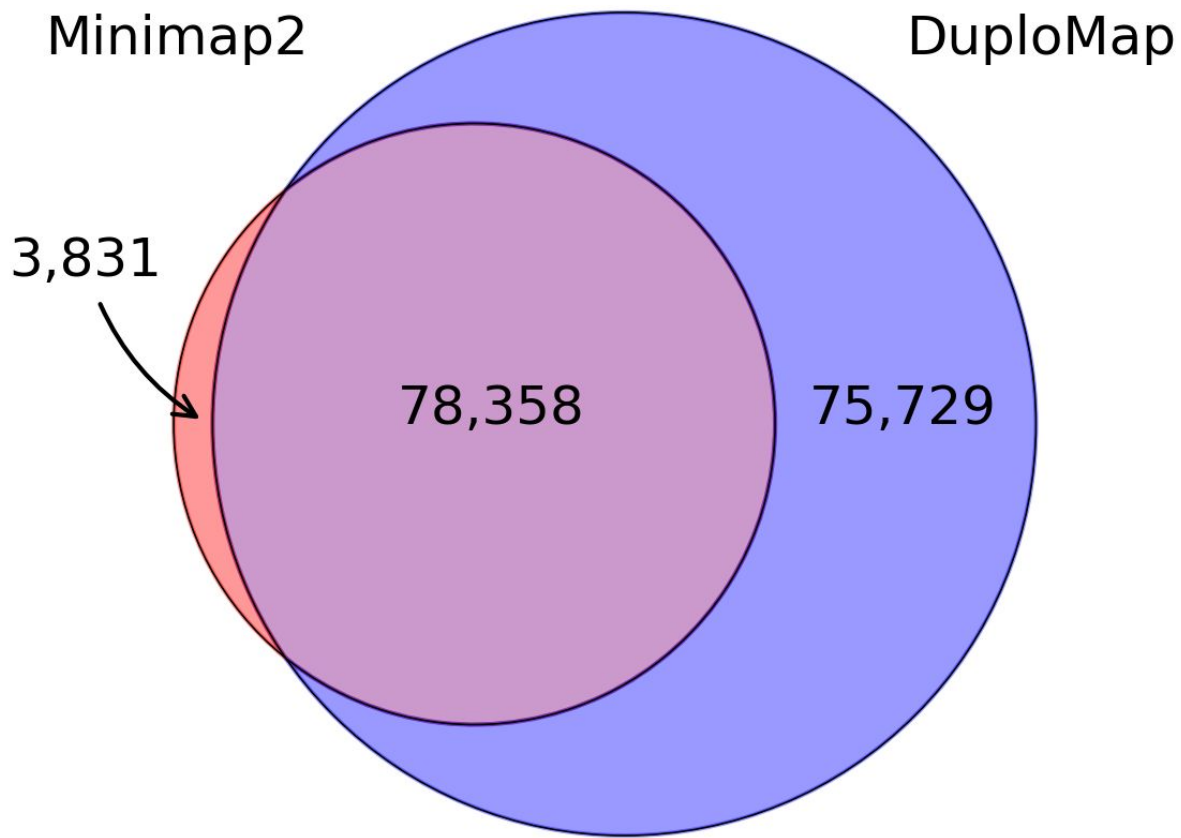




# Unreliable PSVs

PacBio HiFi data:

- ~ 43% SNPs in segmental duplications overlap PSVs,
- ~ 23% PSVs with high quality genotypes were unreliable.



More than half of the called variants intersect PSVs

# Conclusions

Possible to optimize long-read alignments by identifying reliable PSVs.

## Limitations:

- Does not account for copy number variations,
- Ultralong ONT does not show big improvements,
- Assembly-dependent: new telomere-to-telomere assembly should improve results.