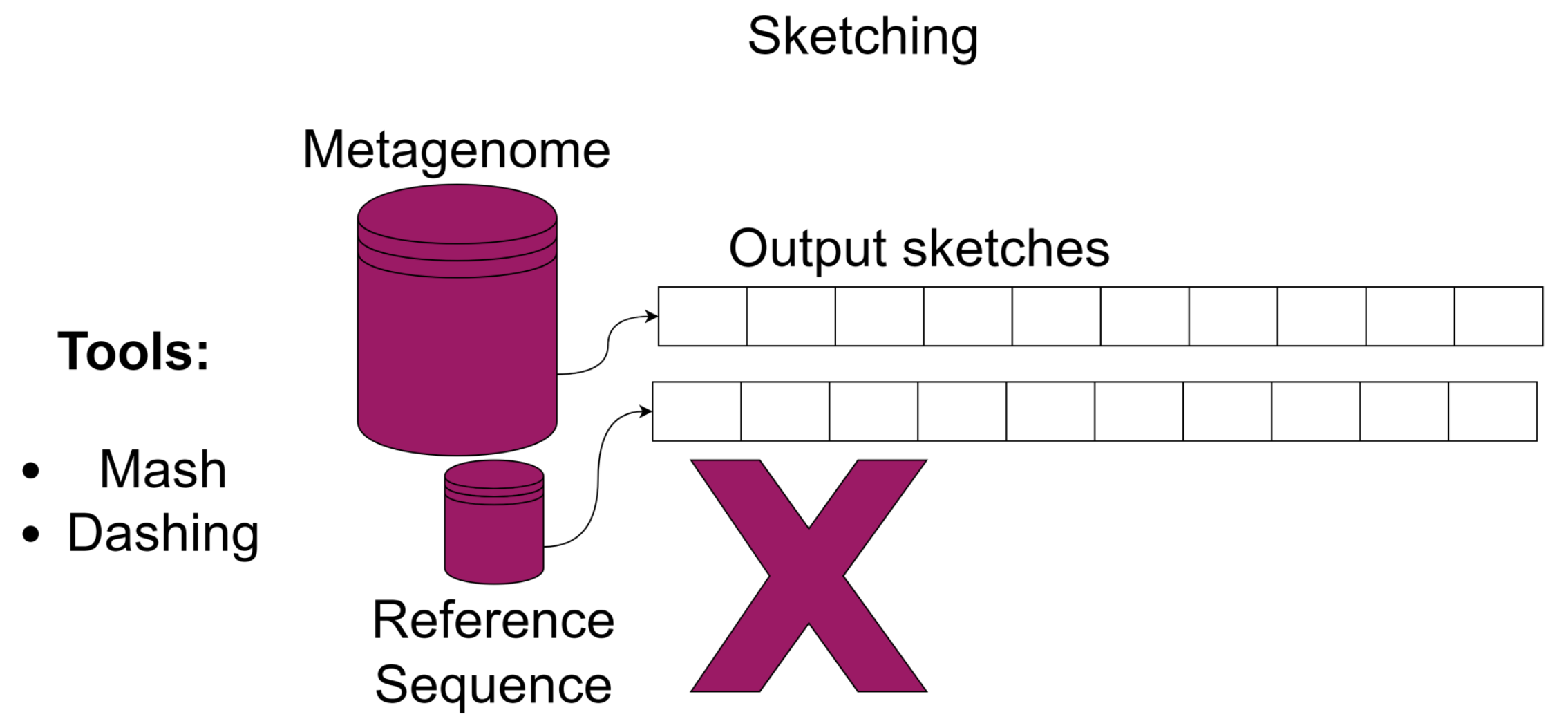
SuperSampler, a memory efficient subsampling strategy for large scale analysis of sequencing data

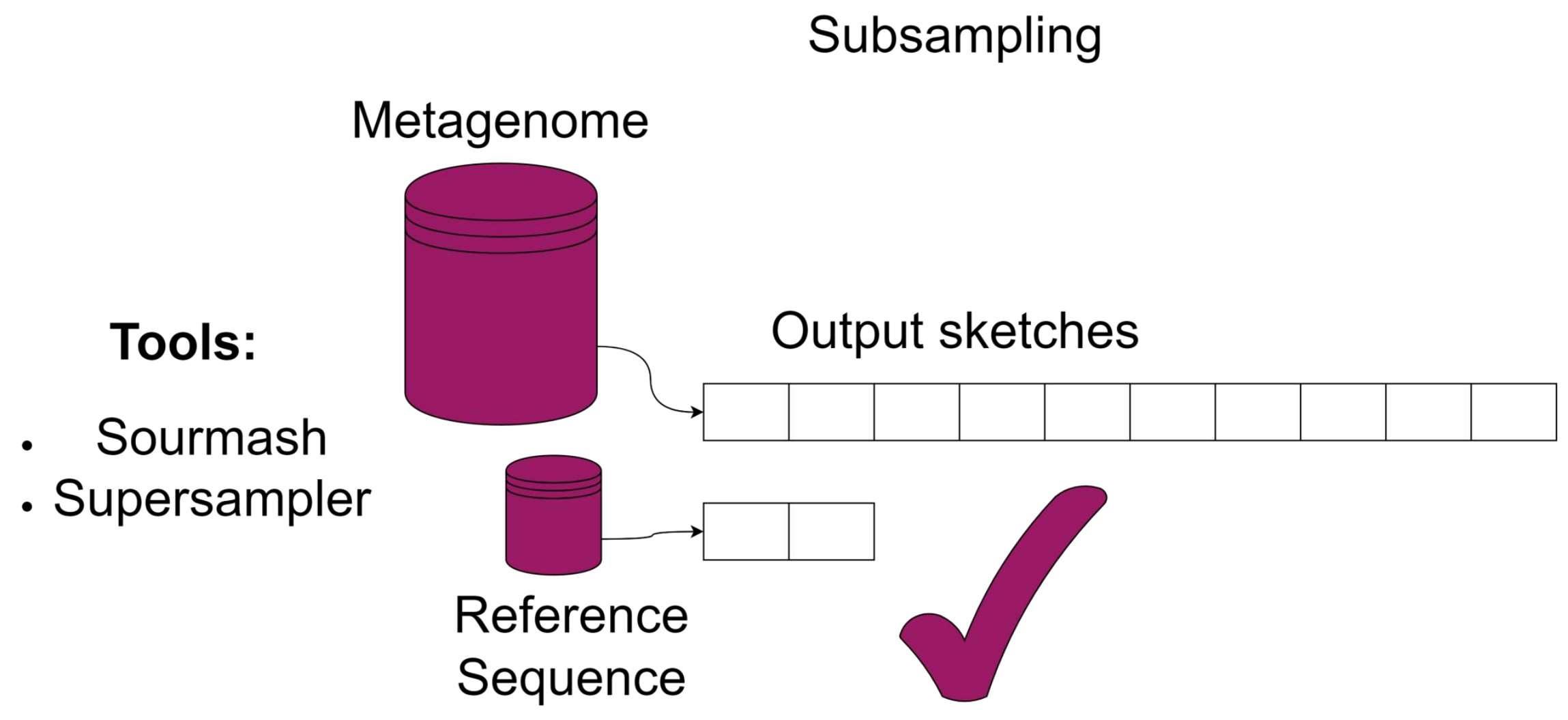
Timothé Rouzé[1], Caleb Smith[1], Antoine Lefevre[1], Antoine Limasset[1]

1. Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

BACKGROUND

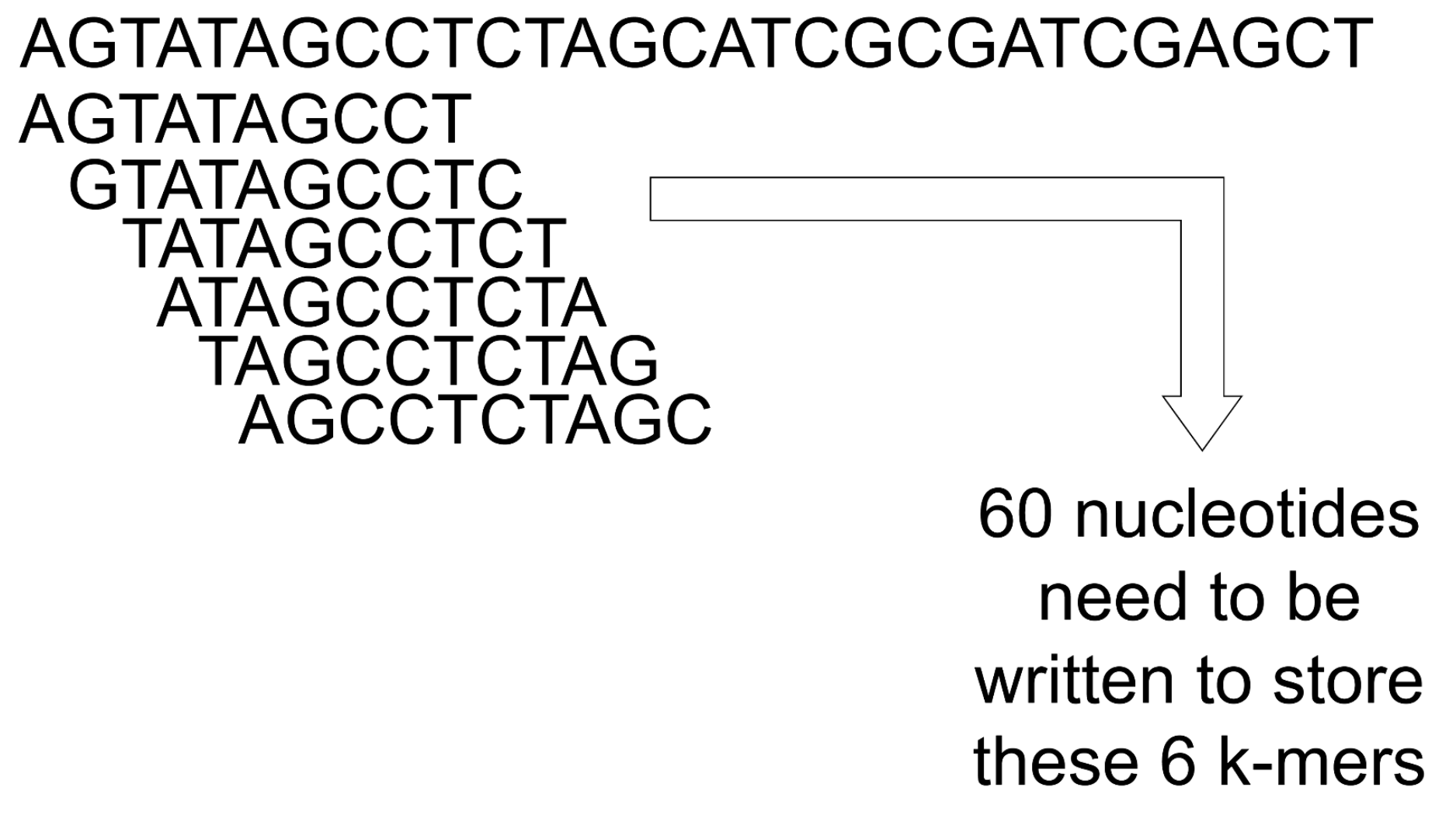
Different scaling strategies exist and among these strategies, two main approaches are prevalent. Sketching produces sketches of the same size whatever the input size, when subsampling produces sketches proportional to the input size. The sketching method causes issues when dealing with datasets that are highly different in sizes. Subsampling addresses this issue while keeping comparability between sketches.



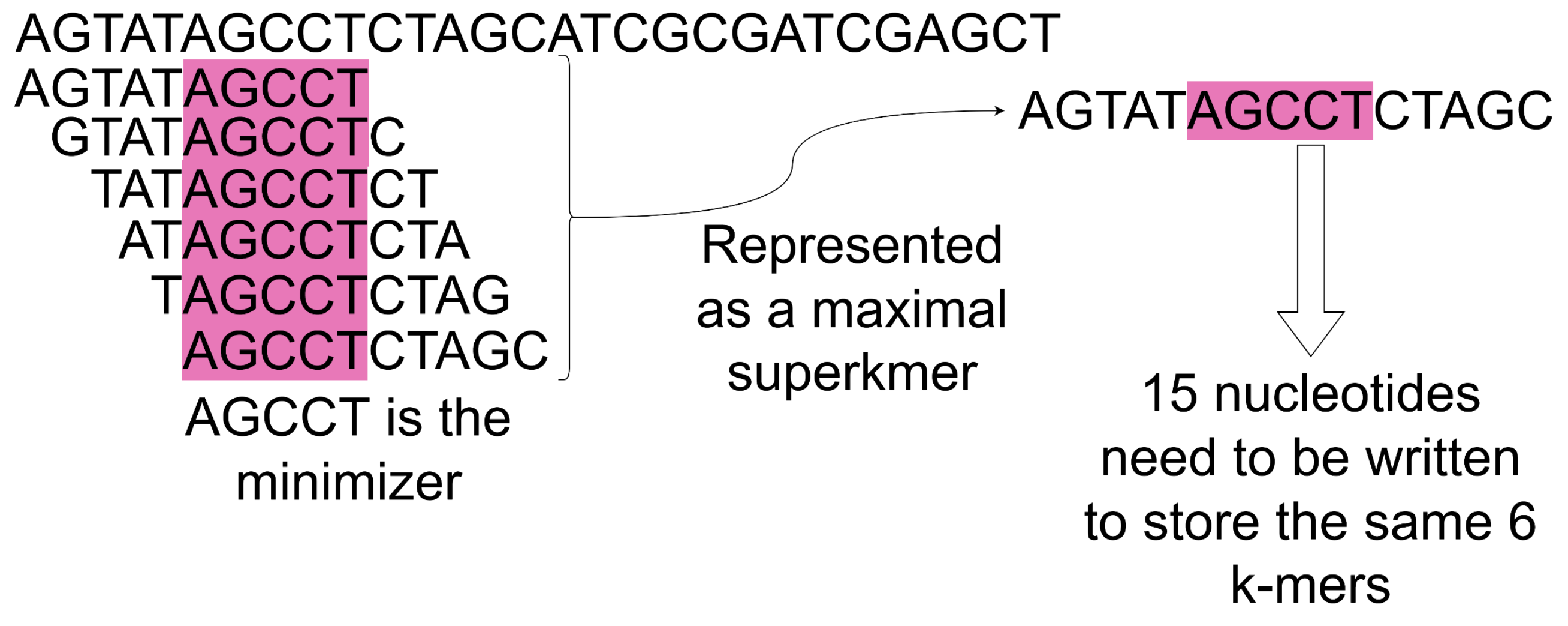


Methods: Supersampler uses superkmers to reduce sketches size

Classic k-mer indexing:


AGTATAGCCTCTAGCATCGCGATCGAGCT
AGTATAGCCT
GTATAGCCTC
TATAGCCTCT
ATAGCCTCTA
TAGCCTCTAG
AGCCTCTAGC
60 nucleotides need to be written to store these 6 k-mers

k-mer indexing with Supersampler:


AGTATAGCCTCTAGCATCGCGATCGAGCT
AGTATAGCCT
GTATAGCCTC
TATAGCCTCT
ATAGCCTCTA
TAGCCTCTAG
AGCCTCTAGC
AGCCT is the minimizer
Represented as a maximal superkmer
AGTATAGCCTCTAGC
15 nucleotides need to be written to store the same 6 k-mers

Estimated memory cost of storing k-mers or superkmers

K-mer indexing memory cost:

$$2 \times K \times \frac{N}{S} \text{ bits}$$

Superkmer indexing memory cost:

$$2 \times \frac{2K-M}{K-M} \times \frac{N}{S} \text{ bits}$$

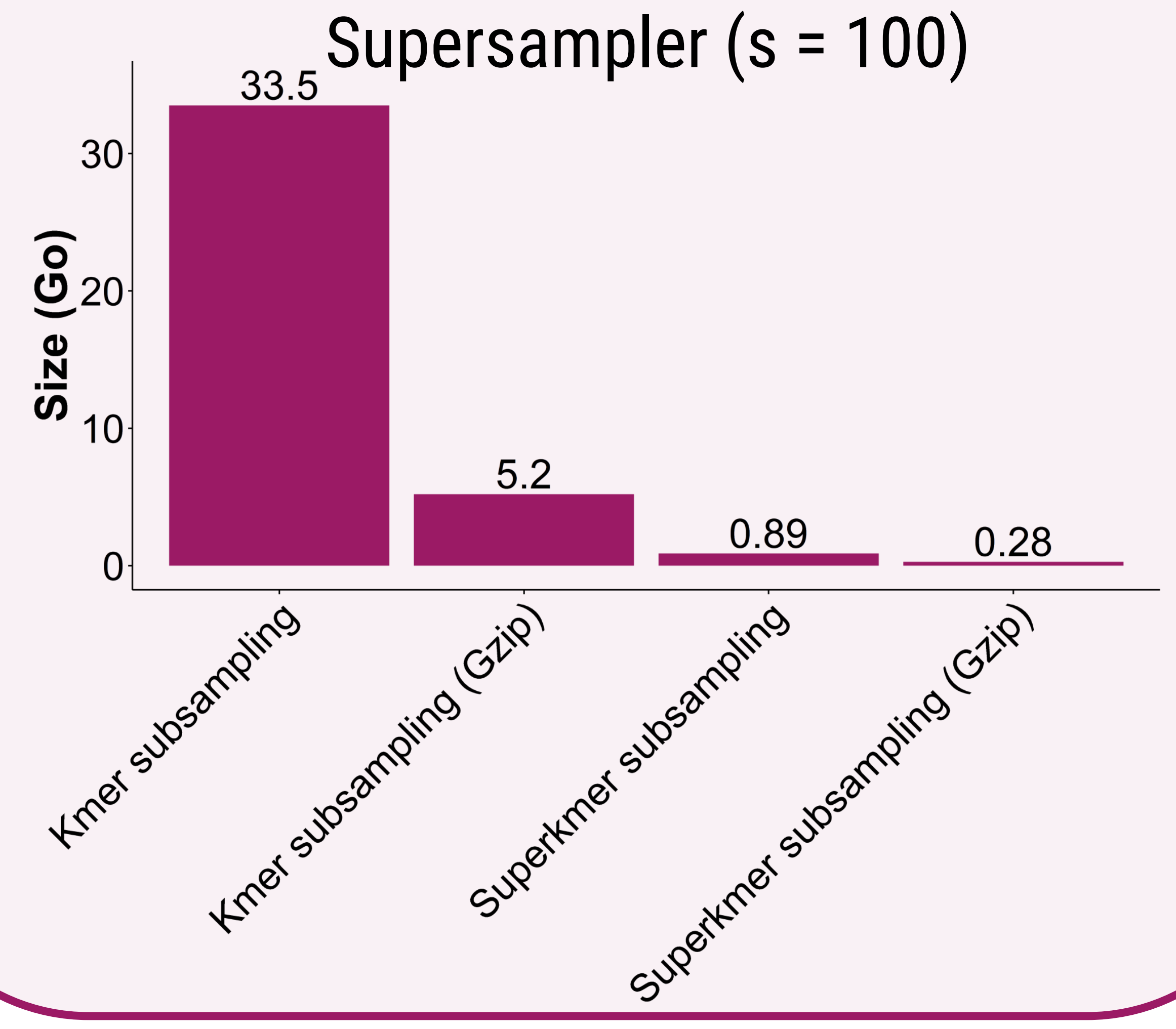
Theoretical size of a Lung fish genome subsampled 1/100

Lungfish genome: 40Gbp

	K = 31	K = 63
k-mers	24,8 Go	50,4 Go
Superkmers	1,88 Go	1,72 Go

Results on lungfish genome

Sketch size comparison between k-mer and superkmer indexing using Supersampler (s = 100)



Sketch sizes and computation time comparison between Sourmash and SuperSampler (s = 100)

	Wall clock (H:m:s)	Output Size	Size Gzipped
Sourmash	1:26:50	4,2 Go	1,6 Go
SuperSampler	1:01:53	0,89 Go	0,28 Go

Conclusion

- Supersampler is more efficient than state of the art subsampling tools both in term of computation time and in term of memory size of the output.

What's next?

- Add multithreading to improve computation speed
- Create a sketch comparison tool
- Statistical comparison between k-mer and superkmer indexing

References:

- Mash: *https://mash.readthedocs.io/en/latest/index.html*
- Dashing: *Baker, D.N., Langmead, B. Dashing: fast and accurate genomic distances with HyperLogLog. Genome Biol* **20,** *265 (2019). https://doi.org/10.1186/s13059-019-1875-0*
- Sourmash: *https://sourmash.readthedocs.io/en/latest/index.html*

Centre de Recherche en Informatique, Signal et Automatique de Lille

Correspondence to:

Timothé Rouzé
Université de Lille,
Lille (France)
timothe.rouze.etu@univ-lille.fr