Supersampler

Efficient subsampling strategy for metagenomic data analysis

Timothé Rouzé, Camille Marchet, Antoine Limasset

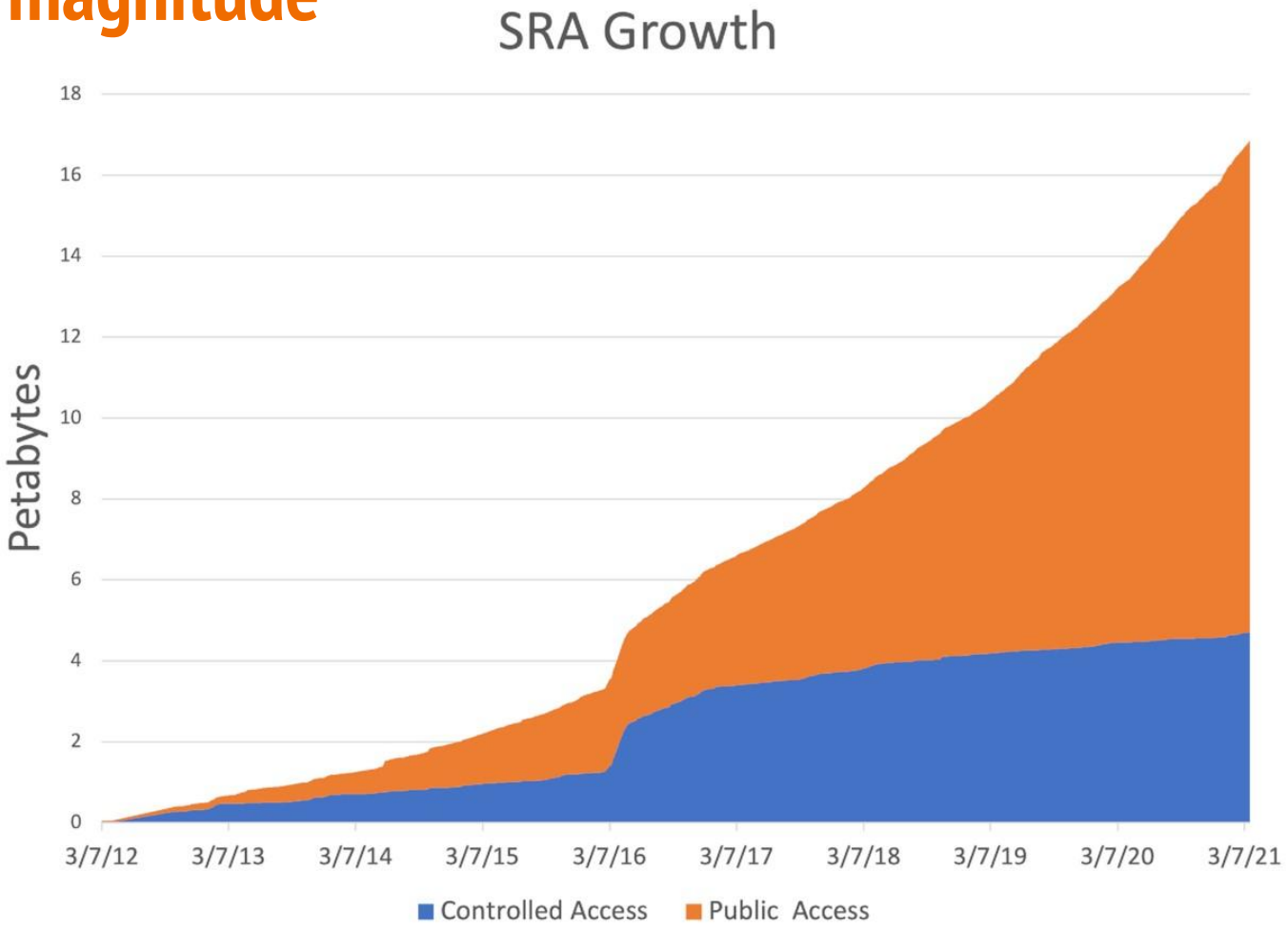
17/11/2022

SeqBim 2022

Univ. Lille, CNRS, CRIStAL

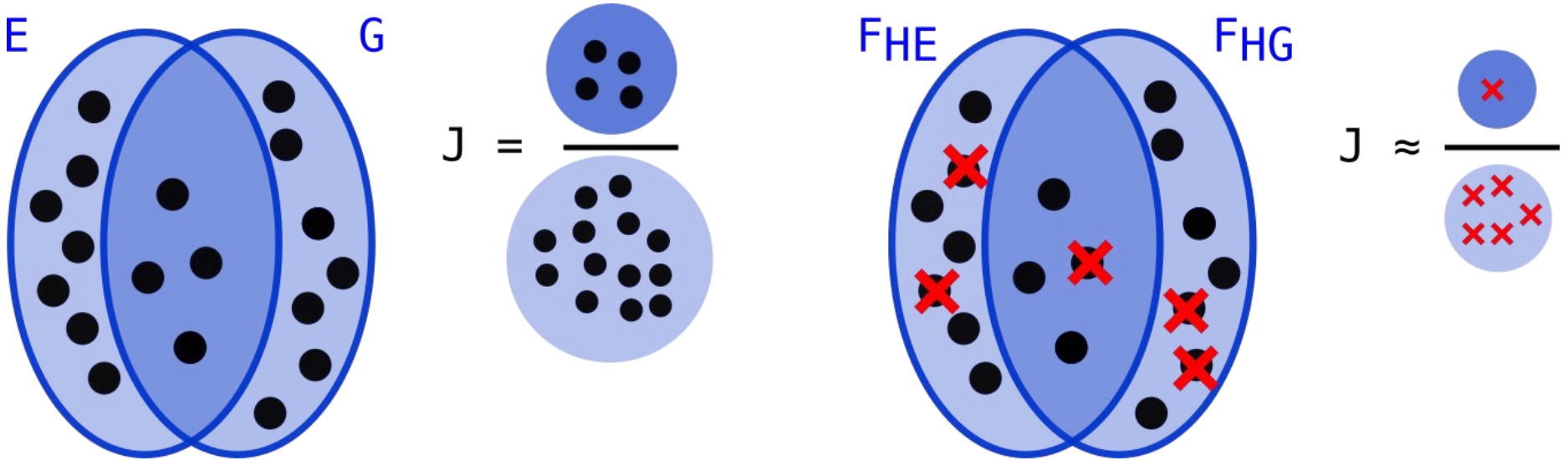
Context

Orders of magnitude

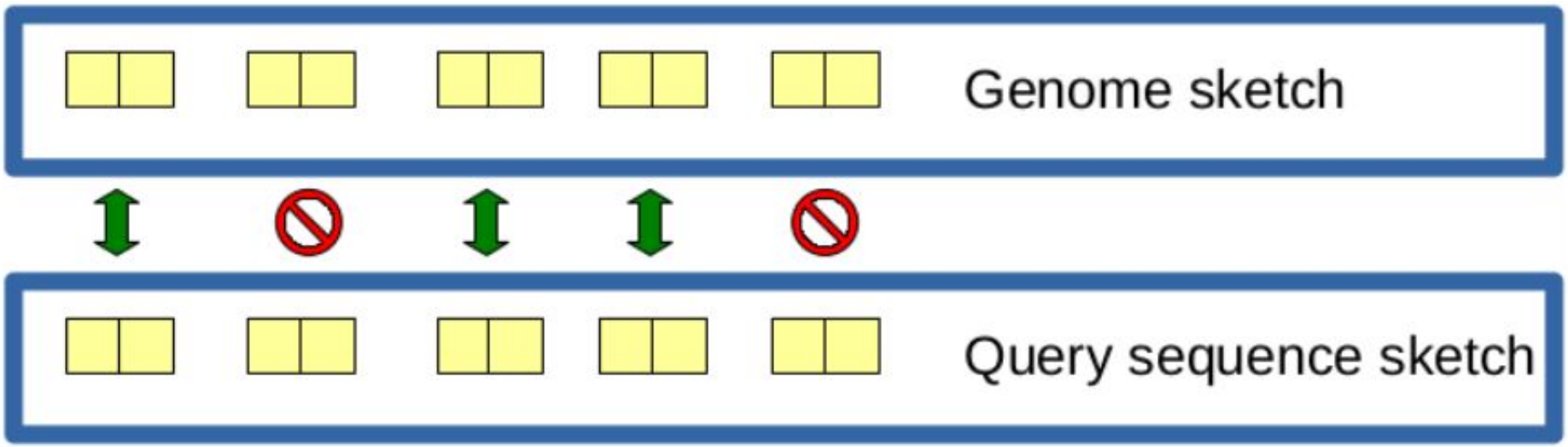


Sketching

Sketching



Sketching



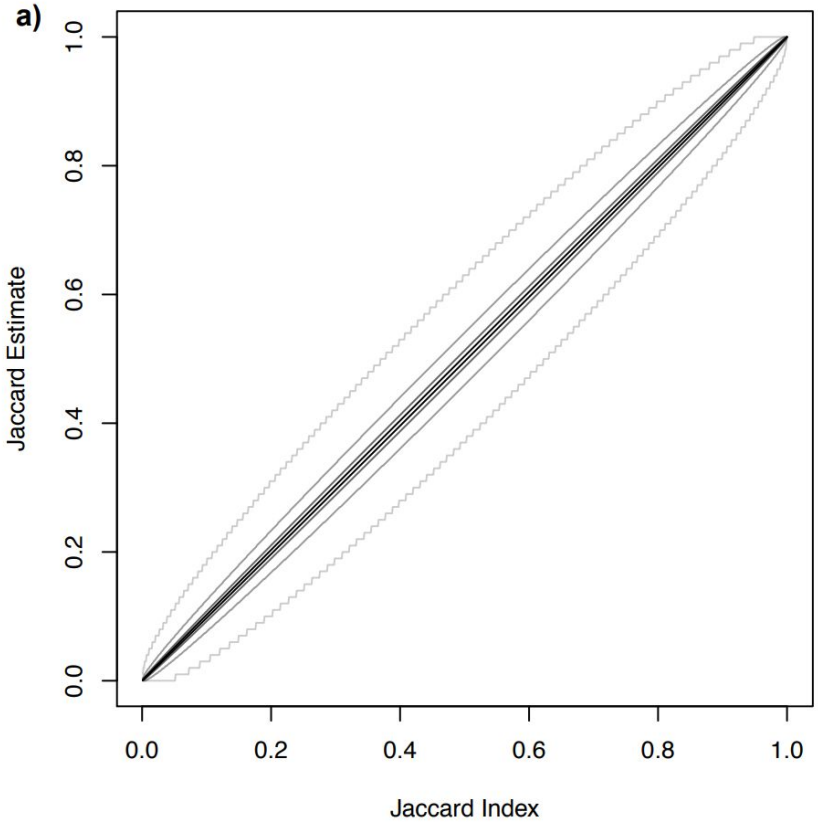
Mash

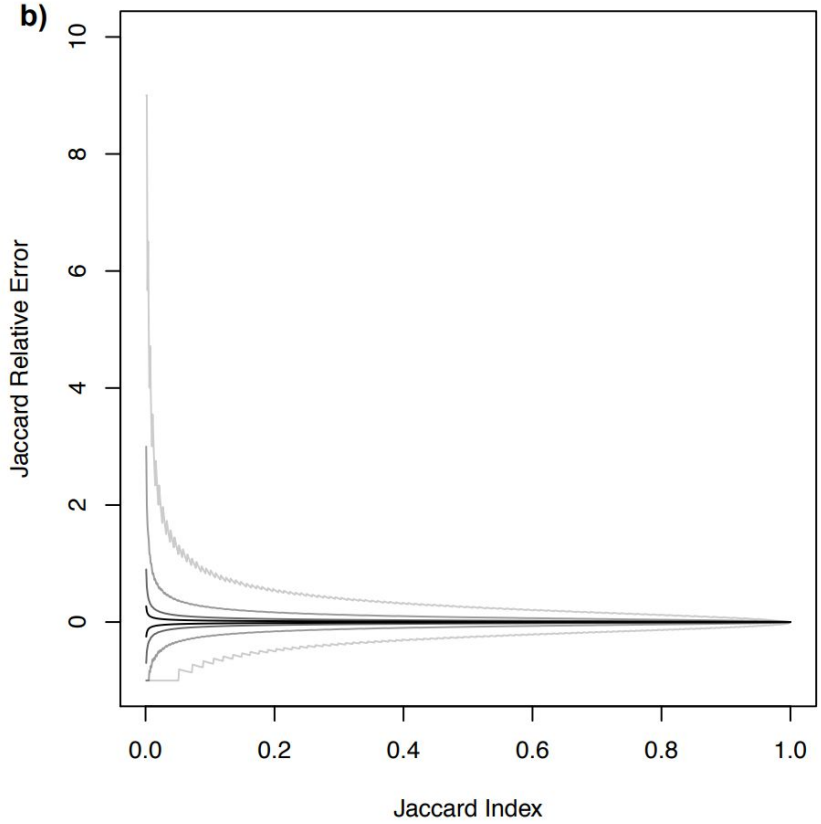
Table 1 Example Mash error bounds for a k-mer size of 21 and increasing sketch sizes

	Mash distance							
Sketch size	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
100	0.0271	0.0868	–	–	–	–	–	–
500	0.0098	0.0245	0.0473	–	–	–	–	–
1000	0.0068	0.0158	0.0323	0.0630	–	–	–	–
5000	0.0029	0.0065	0.0124	0.0235	0.0460	–	–	–
10,000	0.0020	0.0046	0.0086	0.0159	0.0300	0.0726	–	–
50,000	0.0009	0.0020	0.0037	0.0065	0.0116	0.0219	0.0396	0.0822
100,000	0.0006	0.0014	0.0026	0.0046	0.0081	0.0143	0.0250	0.0492
500,000	0.0003	0.0006	0.0011	0.0020	0.0035	0.0060	0.0105	0.0187
1,000,000	0.0002	0.0004	0.0008	0.0014	0.0024	0.0042	0.0072	0.0128

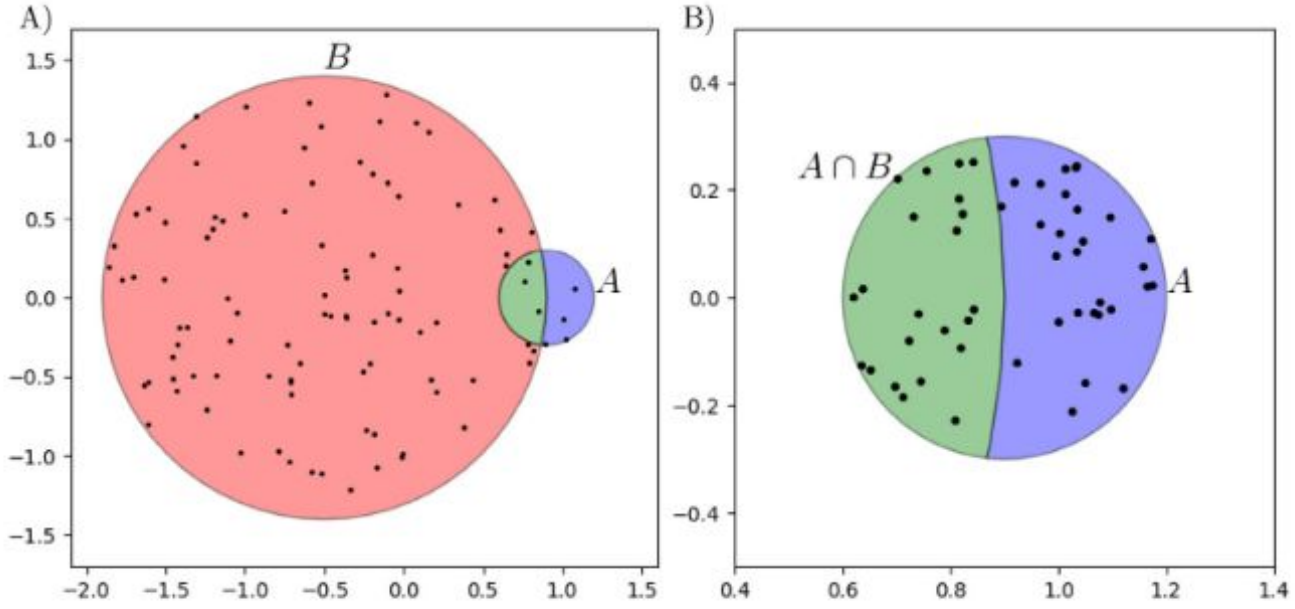
Ondov et al. (2016)

Drawback


a)
Jaccard Estimate
0.0
0.2
0.4
0.6
0.8
1.0
Jaccard Index

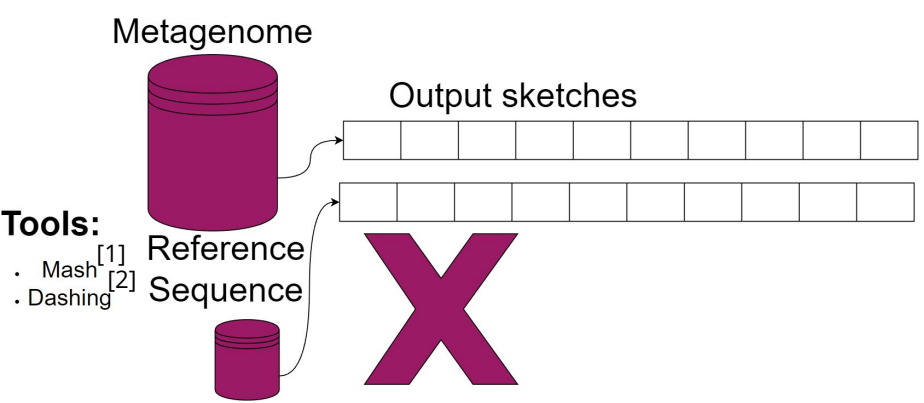

b)
Jaccard Relative Error
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
Jaccard Index

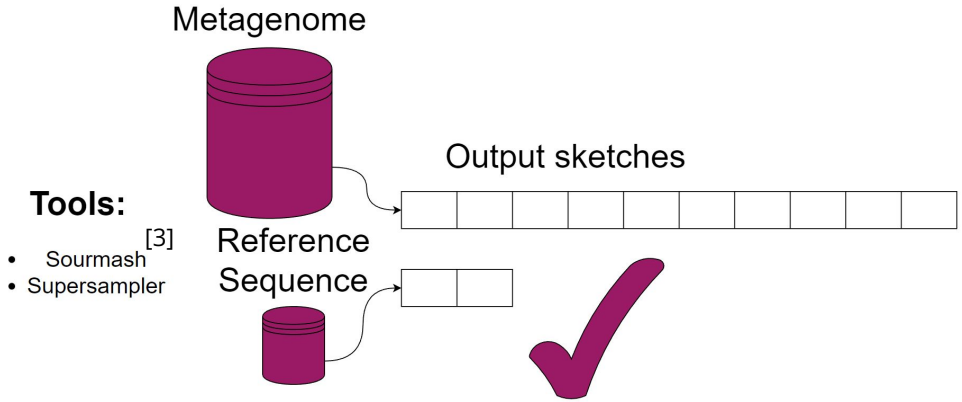
Jaccard index vs Containment index



Koslicki & Zabeti (2019)

Subsampling

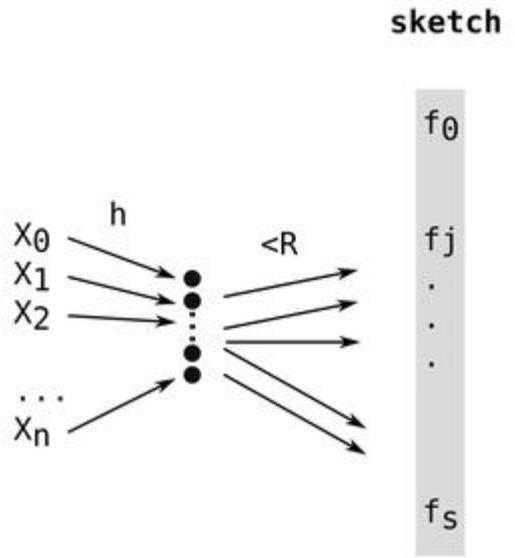




- ➢ Sketch size adaptable to dataset size
- ➢ Containment index

[1] Ondov et al. (2016)
[2] Baker et al. (2019)
[3] Irber et al. (2022)

Sourmash [1]



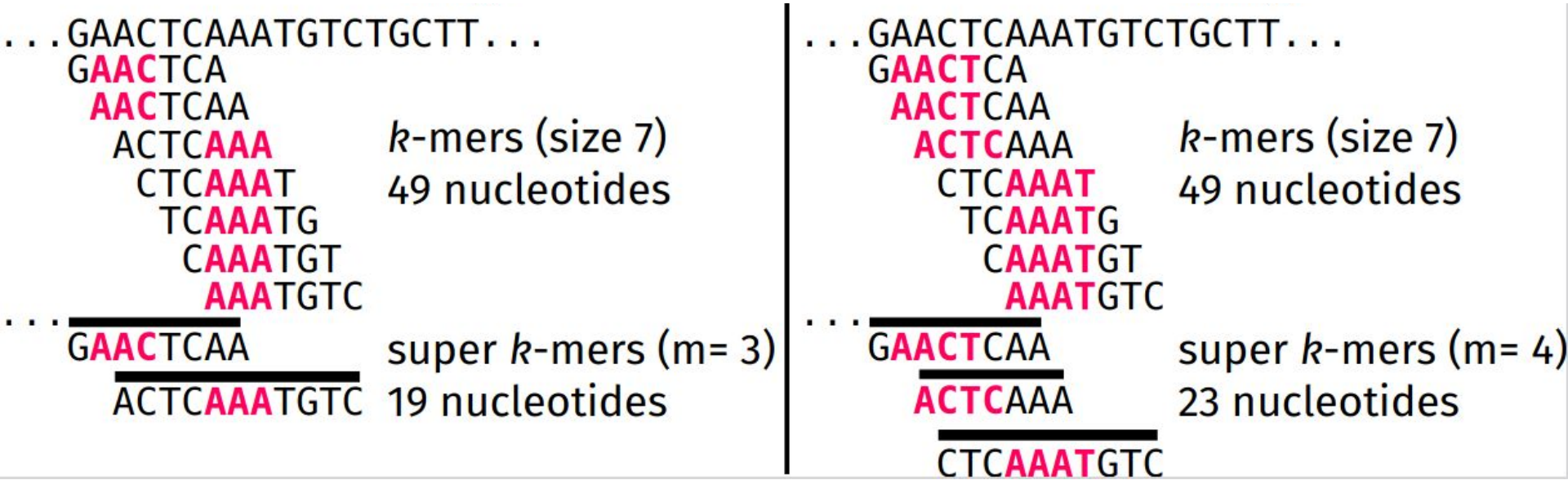
FracMinHash

- K-mer selected with fixed probability P
- Selected k-mers are stored in a 32 bits hash

[1] Irber et al. (2022)

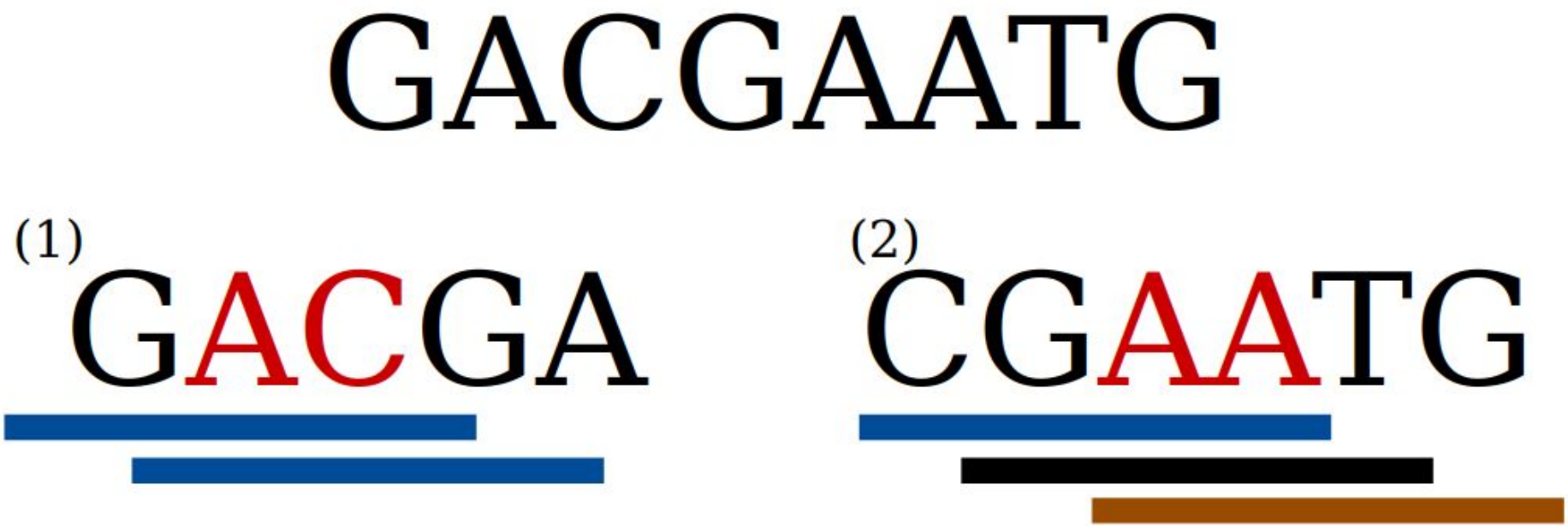
Minimizers and Super-k-mers

➢ Minimizer: smallest sequence of size m of a k-mer



Maximal super-k-mers

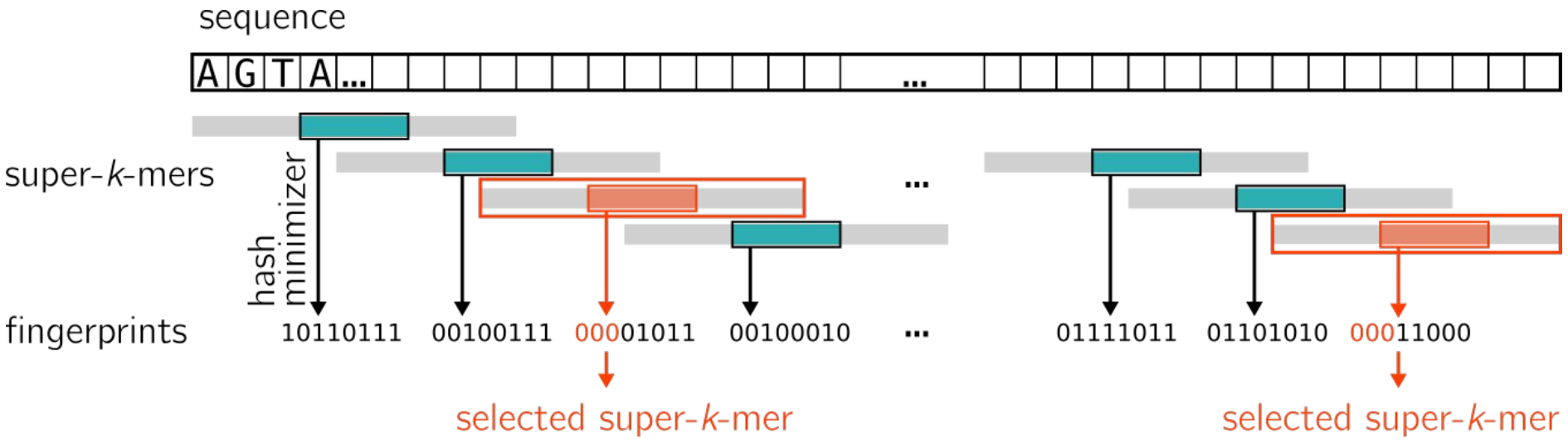
- Maximal super-k-mer size = 2k-m



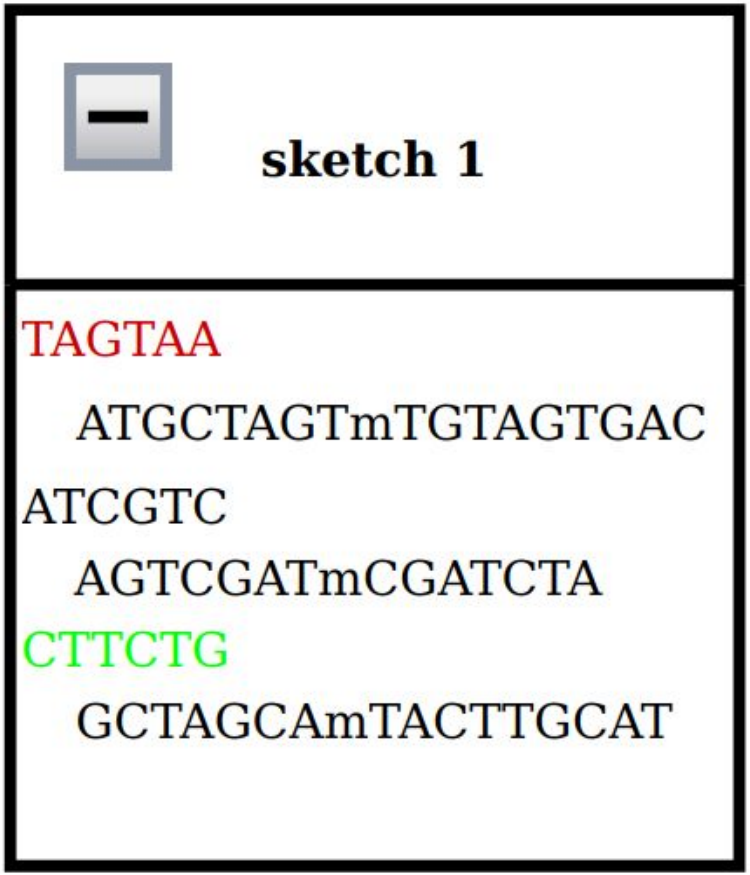
K = 4, m = 2

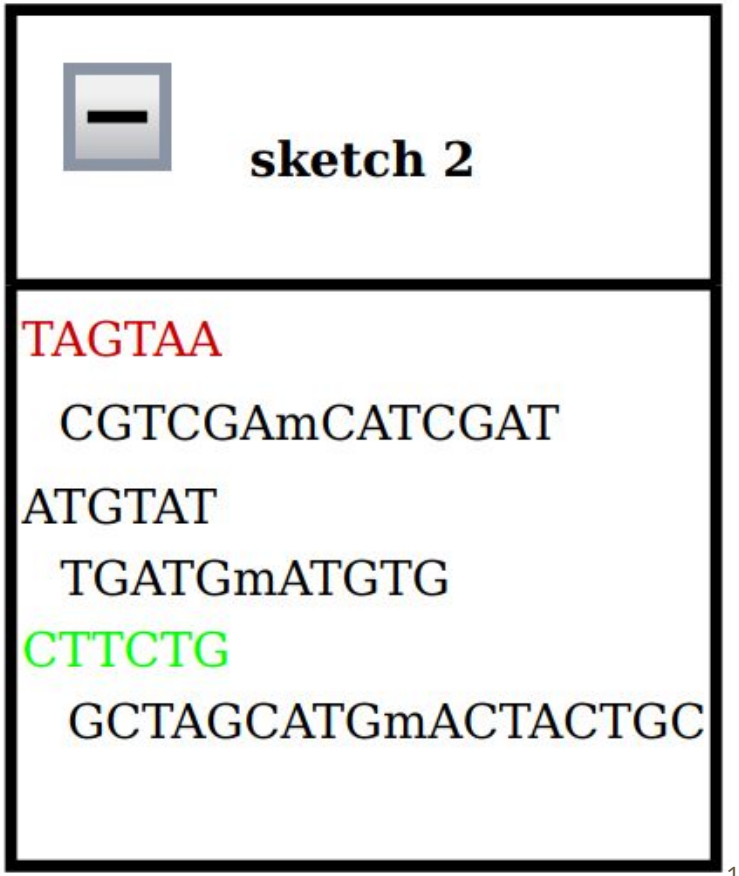
Supersampler

Scaled superkmers

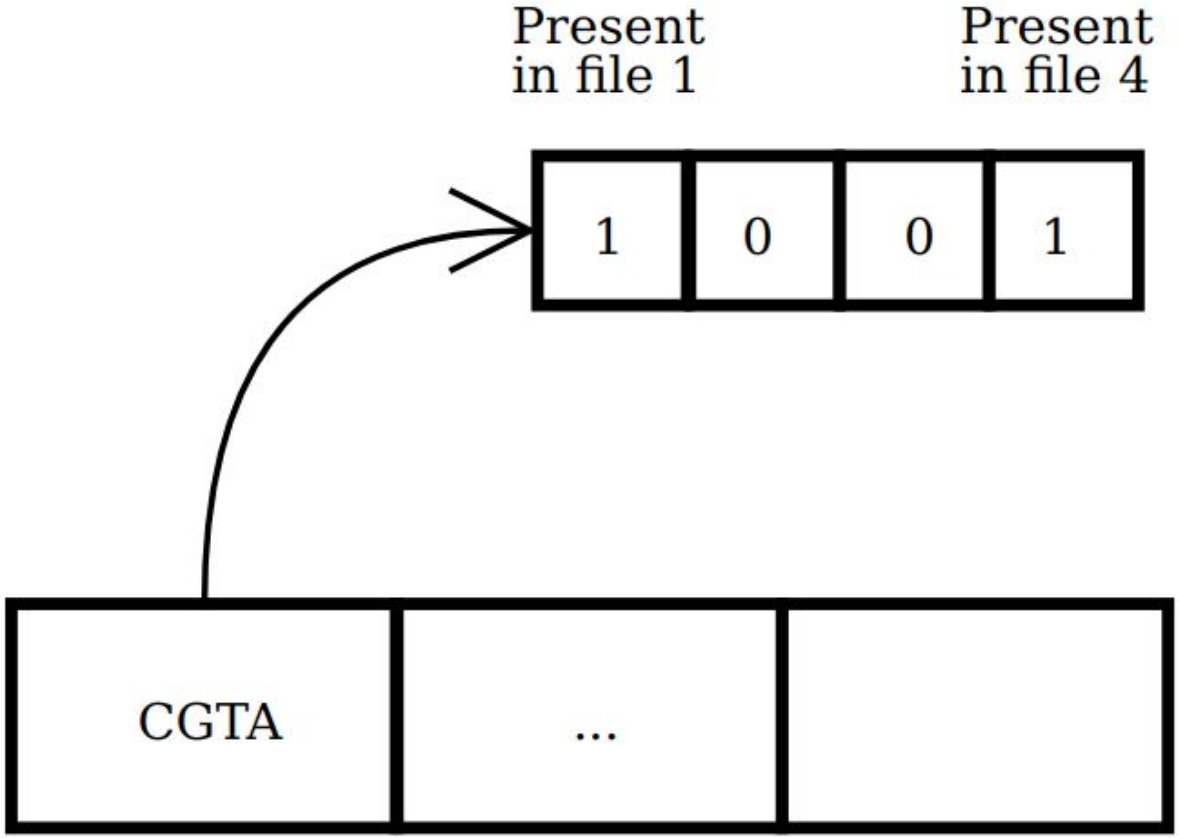


Sketch comparison





Color map



Comparator

- Outputs containment index and Jaccard index
- Outputs shared kmers
- O(#hits) complexity.
- Light RAM usage (one bucket at a time)

Results

Results: lowering memory footprint

Subsampling 1/1000

	Genome size (.gz)	**Supersampler (.gz)**	Sourmash (.gz)
Axolotl	8.6 GBytes	**12.7 MBytes**	126 MBytes
Human	928 MBytes	**1.5 MBytes**	19 MBytes

Subsampling 1/10000

	Genome size (.gz)	**Supersampler (.gz)**	Sourmash (.gz)
Axolotl	8.6 GBytes	**1.58 MBytes**	13 MBytes
Human	928 MBytes	**159 kBytes**	1.8 MBytes

Results: time and memory footprints

Human
3 billion k-mers

	Supersampler K=31, m=11, s=1000	**Supersampler K=31, m=11, s=10000**	Sourmash (k = 31, s=1000)	Sourmash (k=31, s=10000)
Maximal super-k-mer rate	0.943	0.999	NA	NA
bits/ k-mer	**3.45**	**2.43**	32	32
Time (min:sec)	**0:41**	**0:36**	1:39	1:42

Results: time and memory footprints

Axolotl assembled genome
32 billion k-mers

	Supersampler K=31, m=11, s=1000	Supersampler K=31, m=11, s=10000	Sourmash k=31, s=1000	Sourmash K=31, s=10000
Maximal super-k-mer rate	0.968	0.999	NA	NA
bits/ k-mer	**3.56**	**3.53**	32	32
Time (min:sec)	**17:01**	**16:32**	18:54	18:24

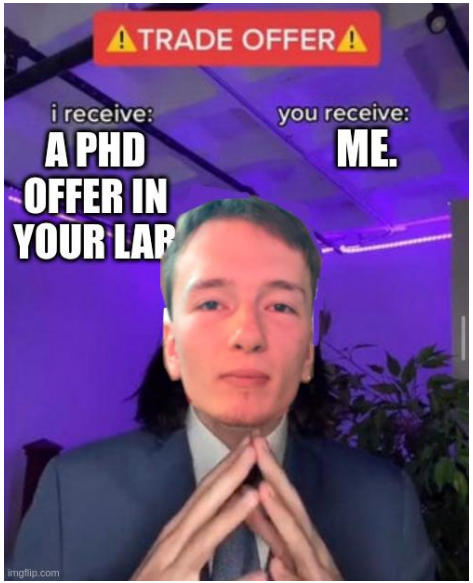
Take home messages

Supersampler

- As fast as sourmash + up to 10X less memory usage
- Accurate sketch comparison, results consistent with current knowledge [1]

Future

- Seed indexing strategy
- Faster sketch comparison
- Add less accurate (FP) but lighter version



[1] Hera et al. (2022)

References

- Ondov, B.D., Treangen, T.J., Melsted, P. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x
- 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
- David Koslicki, Hooman Zabeti, Improving MinHash via the containment index with applications to metagenomic analysis, Applied Mathematics and Computation, Volume 354, 2019, Pages 206-215, ISSN 0096-3003, https://doi.org/10.1016/j.amc.2019.02.018.
- Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers, Luiz Irber, Phillip T. Brooks, Taylor Reiter, N. Tessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, C. Titus Brown, bioRxiv 2022.01.11.475838; doi: https://doi.org/10.1101/2022.01.11.475838
- Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances Mahmudur Rahman Hera, N. Tessa Pierce-Ward, David Koslicki bioRxiv 2022.01.11.475870; doi: https://doi.org/10.1101/2022.01.11.475870