

kmtricks: creating bloom filters for indexing large sequencing data collections

Téo Lemane¹ Paul Medvedev^{2, 3, 4} Rayan Chikhi⁵ Pierre Peterlongo¹

¹Univ. Rennes, Inria, CNRS, Rennes, France

²Department of Computational Biology, Institut Pasteur, Paris, France

³Department of Computer Science and Engineering, The Pennsylvania State University, USA

⁴Department of Biology, The Pennsylvania State University, USA

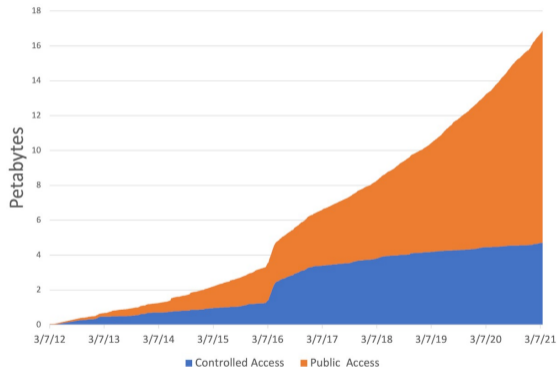
⁵Huck Institutes of the Life Sciences, The Pennsylvania State University, USA

6th july 2022



A big and well-known problem

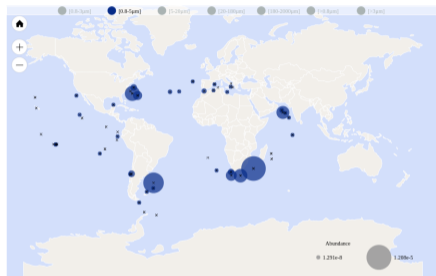
SRA Growth



K. Katz *et al.*, 2022

- Tara Ocean Project: 250 billions metaG reads
- 100,000 Genome Project: 20 PB
- SRA: 🚀

Ocean Gene Atlas



<https://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/>

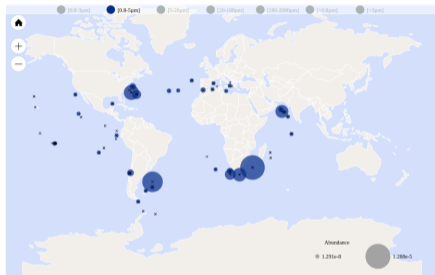
¹B. Solomon and C. Kingsford. Fast search of thousands of short-read sequencing experiments. *Nature Biotechnology*, 2016.

²Y. Yu, et al. Seqothello: querying rna-seq experiments at scale. *Genome Biology*, 2018.

³N. Luhmann, et al. Blastfrost: Fast querying of 100,000 s of bacterial genomes in bifrost graphs. *Genome Biology*, 2021.

⁴R. Wittler. Alignment-and reference-free phylogenomics with colored de Bruijn graphs. *Algorithms for Molecular Biology*, 2020.

Ocean Gene Atlas



<https://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/>

And others:

- RNA-Seq
 - Expressed isoform according to tissues¹
 - Gene fusion²
- Microbial genomics
 - Antimicrobial resistance³
- Genome dynamics
 - Phylogeny⁴
- ...

¹B. Solomon and C. Kingsford. Fast search of thousands of short-read sequencing experiments. Nature Biotechnology, 2016.

²Y. Yu, et al. Seqothello: querying rna-seq experiments at scale. Genome Biology, 2018.

³N. Luhmann, et al. Blastfrost: Fast querying of 100,000 s of bacterial genomes in bifrost graphs. Genome Biology, 2021.

⁴R. Wittler. Alignment-and reference-free phylogenomics with colored de Bruijn graphs. Algorithms for Molecular Biology, 2020.

Given an experiments set, and a sequence of interest, which sample contains this sequence ?

Given an experiments set, and a sequence of interest, which sample contains this sequence ?

In terms of k -mers:

- A query Q matches a sample S if at least a fraction θ of Q 's k -mers are present in S .

Given an experiments set, and a sequence of interest, which sample contains this sequence ?

In terms of k -mers:

- A query Q matches a sample S if at least a fraction θ of Q 's k -mers are present in S .

Q = ACGTAGCT

K = 5



ACGT

CGTA

GTAG

TAGC

AGCT



ACGT

CGTA

GTAG

TAGC

AGCT

● ●

● ● ●

● ●

●

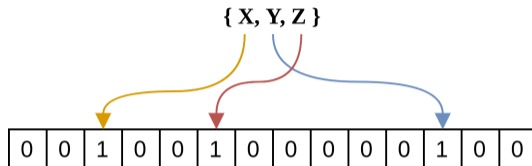
● ● ● ●

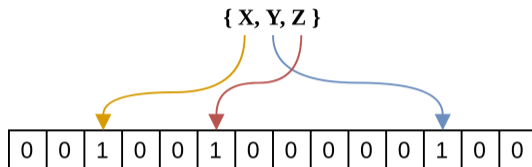
- **BFT** (Holley *et al.*, 2016)
- **Sequence Bloom Tree**
 - **SBT** (Solomon & Kingsford, 2016)
 - **AllSomeSBT** (Sun *et al.*, 2017)
 - **SSBT** (Solomon & Kingsford, 2018)
 - **HowDeSBT** (Harris & Medvedev, 2019)
- Mantis (Pandey *et al.*, 2018)
- SeqOthello (Yu *et al.*, 2018)
- **BIGSI** (Bradley *et al.*, 2019)
- **COBS** (Bingmann *et al.*, 2019)
- REINDEER (Marchet *et al.*, 2020)
- Metagraph (Karasikov *et al.*, 2021)

Review of k-mer indexing:

Data structure based on k-mers for querying large collections of sequencing datasets (Marchet *et al.*, 2019)

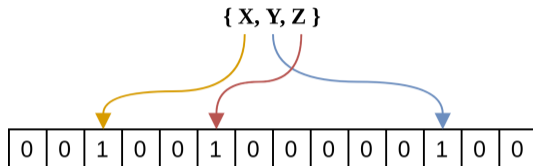
Bloom filters





Construction

- Count k -mers
- For each k -mer: compute hashes and set corresponding bits



Construction

- Count k -mers
- For each k -mer: compute hashes and set corresponding bits

Issues

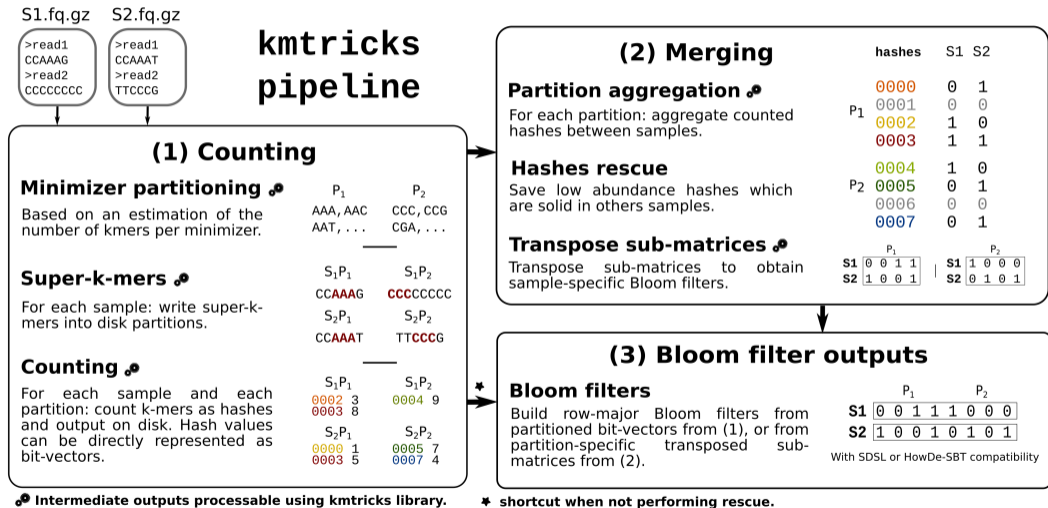
- k -mer counting is a huge bottleneck
- Bad data locality

kmtricks: Bloom filters matrix construction

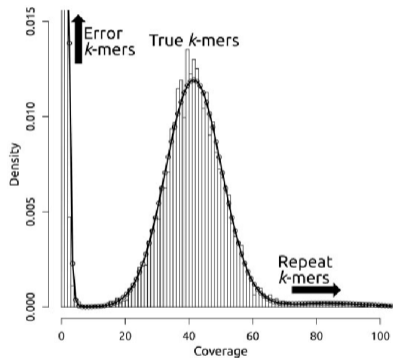
S1	S2	S3	S4	S5	S6	S7	S8	S9	...	Sn
0	1	1	0	0	1	1	1	1	...	1
0	1	1	1	1	1	0	0	1	...	0
0	1	1	1	1	1	0	0	0	...	0
1	0	0	1	1	0	1	1	1	...	1
1	0	1	0	0	1	0	0	1	...	0
0	0	1	0	0	1	1	0	0	...	0
0	1	0	0	1	1	1	1	1	...	1
0	0	0	0	1	0	1	1	1	...	0
1	1	1	0	1	1	0	0	1	...	0
1	1	0	0	1	0	0	1	0	...	0
1	1	0	1	1	1	1	1	0	...	0
0	0	0	1	1	0	1	0	0	...	0

kmtricks: Bloom filters matrix construction

	S1	S2	S3	S4	S5	S6	S7	S8	S9	...	Sn
P1	0	1	1	0	0	1	1	1	1	...	1
	0	1	1	1	1	1	0	0	1	...	0
	0	1	1	1	1	1	0	0	0	...	0
	0
P2	1	0	0	1	1	0	1	1	1	...	1
	1	0	1	0	0	1	0	0	1	...	0
	0	0	1	0	0	1	1	0	0	...	0
	1
...											
Pn	0	1	0	0	1	1	1	1	1	...	1
	0	0	0	0	1	0	1	1	1	...	0
	1	1	1	0	1	1	0	0	1	...	0
	0

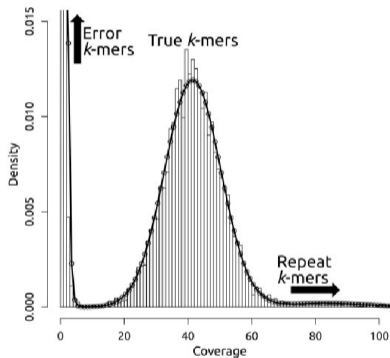


Hard abundance threshold vs kmtricks rescue strategy



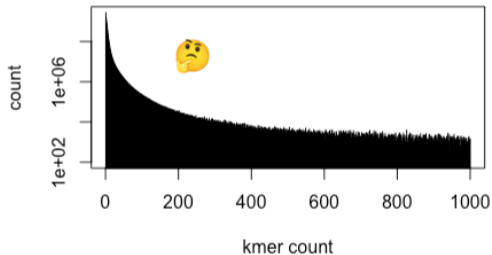
D. Laehnemann *et al.*, 2015

Hard abundance threshold vs kmtricks rescue strategy



D. Laehnemann *et al.*, 2015

Kmer histogram of Tara sample 76-SUR-CCKK



k-mer filtering

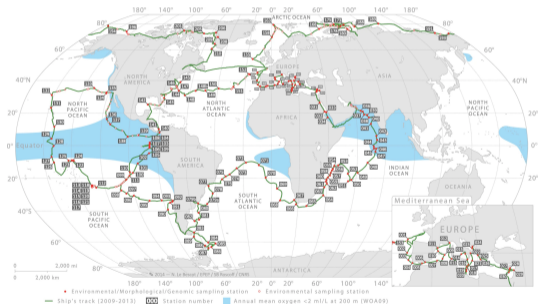
The holistic view of *k*-mers abundances across samples allows custom errors screening

ab. threshold	Counted <i>k</i> -mers					Post filtration result				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
k1	<u>2</u>	0	<u>2</u>	<u>5</u>	<u>2</u>	<u>1</u>	0	1	1	1
k2	<u>4</u>	1	<u>6</u>	2	0	1	0	1	0	0

hard-min=1, share-min=3

241 sampling stations

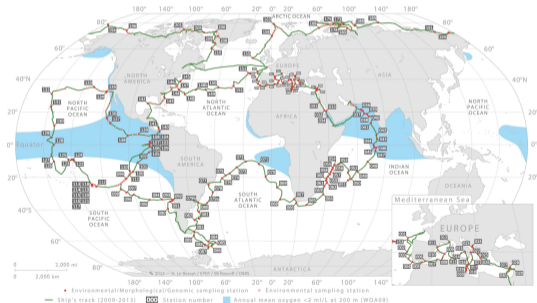
- 712 short read samples, +6TB of compressed data
- 266 billions of distinct k-mers



Pesant *et al.*, 2015

241 sampling stations

- 712 short read samples, +6TB of compressed data
- 266 billions of distinct k-mers



Pesant *et al.*, 2015

Benchmark environment

- 128 threads
- 970 MB/s and 216 MB/s sequential read/write

	Time (min)	Memory (GB)	Disk (TB)
kmtricks	1433	83.4	1.5
Jellyfish ^a + makebf	$\approx 8071^b$	80.6 ^b	$\approx 0.8^b$
KMC3 ^a + makebf	$\approx 5310^b$	100 ^b	$\approx 0.8^b$

^aStopped after 72h computation. ^bExtrapolated estimation.

Hard abundance threshold vs kmtricks rescue strategy

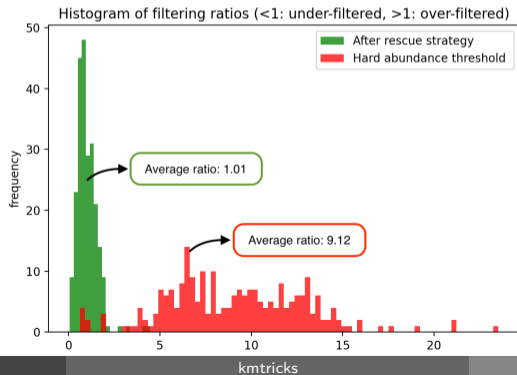
- For each sample, we know the error rate of the sequencer used thanks to the Genoscope benchmarks.

$$\frac{\textit{k-mers seen only once}}{\textit{expected number of wrong k-mers}} \quad \frac{\textit{k-mers discarded by kmtricks}}{\textit{expected number of wrong k-mers}}$$

Hard abundance threshold vs kmtricks rescue strategy

- For each sample, we know the error rate of the sequencer used thanks to the Genoscope benchmarks.

$$\frac{\textit{k-mers seen only once}}{\textit{expected number of wrong k-mers}} \quad \frac{\textit{k-mers discarded by kmtricks}}{\textit{expected number of wrong k-mers}}$$



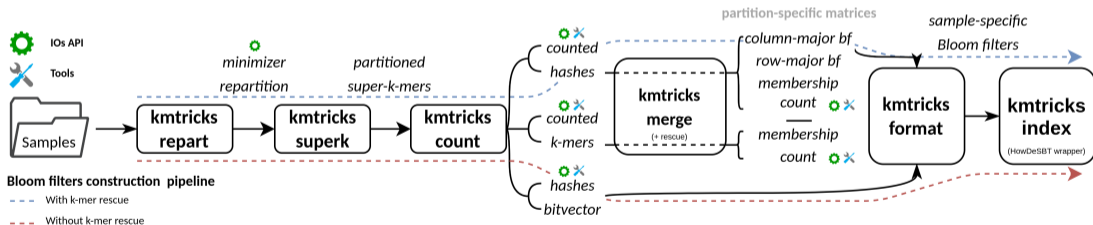
kmtricks: a k -mer matrix toolbox

CLI

- **pipeline:** end-to-end matrix construction
- **modules:** step-by-step matrix construction
- **tools:** work with kmtricks outputs

C++ API and plugin support

- I/O
- Matrix streaming
- Extend features, e.g. matrix filtering



<https://github.com/tleman/kmtricks>

Example of application: Differential k -mer analysis

kmdiff: large-scale and user-friendly differential kmer analyses

	Control			Case		
	1	2	3	4	5	6
k0	2	1	3	2	3	1
k1	3	8	9	1	0	2
k2	7	5	8	0	0	0
k3	4	4	6	3	7	5
k4	2	0	6	8	9	9
..						
kn	6	4	8	2	2	3

- Uses kmtricks streaming features along with a state-of-the-art statistical model¹ to find differentially represented k -mers between two cohorts


¹A. Rahman *et al.*, "Association mapping from sequencing reads using k -mers", 2018

Example of application: Differential k -mer analysis

kmdiff: large-scale and user-friendly differential kmer analyses

	Control			Case		
	1	2	3	4	5	6
k0	2	1	3	2	3	1
k1	3	8	9	1	0	2
k2	7	5	8	0	0	0
k3	4	4	6	3	7	5
k4	2	0	6	8	9	9
..						
kn	6	4	8	2	2	3

- Uses kmtricks streaming features along with a state-of-the-art statistical model¹ to find differentially represented k -mers between two cohorts
- 40vs40 human Illumina WGS (+3TB gz)
 - 9h, 11GB ram (vs 138h, 85GB ram)¹
- Applications
 - GWAS on non-model species

 <https://github.com/tlemane/kmdiff>

¹A. Rahman *et al.*, "Association mapping from sequencing reads using k -mers", 2018

- Efficient and flexible Bloom/ k -mer matrix toolbox
- Supports medium/large datasets like Tara Ocean
- Comes with a set of utilities/API/plugins for downstream analysis
- Obviously, still very insufficient to hope to scale up on very large databases like SRA

Future work:

- Characterization of rare rescued k -mers
- Take advantage of the partitioned structure of Bloom filters for a more efficient construction/query of the HowDeSBT tree

Now available in Bioinformatics Advances: T. Lemane, P. Medvedev, R. Chikhi, P. Peterlongo, “kmtricks: Efficient and flexible construction of Bloom filters for large sequencing data collections”, Bioinformatics Advances, 2022