

Advances in k-mer matrix construction for analysis of large sequencing collections

Téo Lemane¹ Rayan Chikhi² Pierre Peterlongo¹

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

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

SeqBIM

26/11/2021



SeqBIM



- An holistic representation of sequence content across sequencing samples.



| | 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 |

- Sequence similarity between metagenomic sequencing samples¹
- RNA-Seq analyses²
- Bacterial GWAS³
- Read samples indexing⁴
- k-mer-based variants detection⁵

¹Benoit et al., "Multiple comparative metagenomics using multiset k-mer counting"

²Audoux et al., "DE-kupl: Exhaustive capture of biological variation in RNA-seq data through k-mer decomposition"

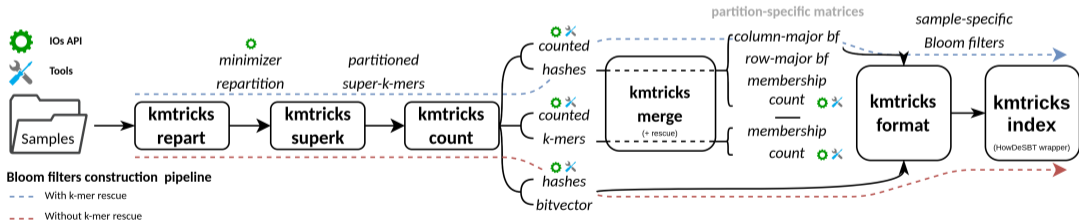
³Jaillard et al., "A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events"

⁴Lemane et al., "kmtricks: Efficient construction of Bloom filters for large sequencing data collections"

⁵Rahman et al., "Association mapping from sequencing reads using k-mers"

Main features

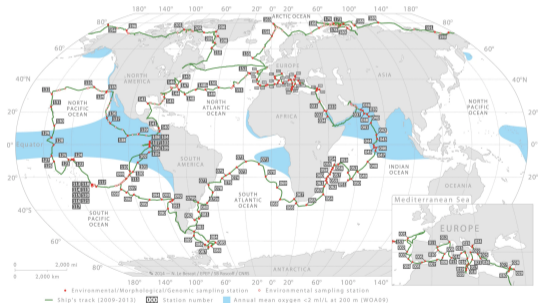
- k-mer matrix construction
- Bloom matrix construction
- k-mer filtering



Application on Tara Ocean bacterial metagenome

241 sampling stations

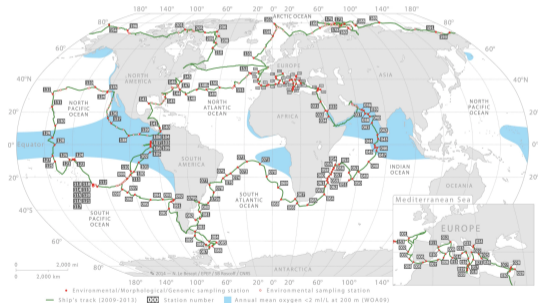
- 712 short read samples, +6TB of compressed data
- 266 billions of distinct k-mers



Application on Tara Ocean bacterial metagenome

241 sampling stations

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



| | Time (min) | Mem (GB) | Disk (TB) |
|-----------------|------------|----------|-----------|
| kmtricks | 2248 | 43.4 | 2.2 |
| Jellyfish-HowDe | >10000 | 80.6 | ≈ 1.1 |
| KMC3-HowDe | >8500 | 213 | ≈ 1.1 |

k-mer rescue:


- Save rare but shared k-mers


k-mer rescue:

- Save rare but shared k-mers

Filtered cells


| | |
|------------------------------|-------------|
| Expected errors | 98 billion |
| Using the hard ab. threshold | 756 billion |
| Using the rescue strategy | 86 billion |

Availability:  <https://github.com/tlemane/kmtricks>

Availability:  <https://github.com/tlemane/kmtricks>

CLI

- **pipeline:** `kmtricks pipeline --file in.fof --run-dir kdir`
- **modules:** `kmtricks count --id D1 --run-dir kdir --mode kmer`
- **tools:** `kmtricks aggregate --run-dir kdir --matrix kmer --format text > matrix.txt`

Availability:  <https://github.com/tlemane/kmtricks>


CLI

- **pipeline:** `kmtricks pipeline --file in.fof --run-dir kdir`
- **modules:** `kmtricks count --id D1 --run-dir kdir --mode kmer`
- **tools:** `kmtricks aggregate --run-dir kdir --matrix kmer --format text > matrix.txt`

API

```
KmerMerger merger(...);  
while (merger.next()) {  
    // matrix streaming  
}
```

```
Repartition repart(...);  
repart.get_partition(kmer.minimizer());
```

Availability:  <https://github.com/tlemane/kmtricks>

CLI

- **pipeline:** `kmtricks pipeline --file in.fof --run-dir kdir`
- **modules:** `kmtricks count --id D1 --run-dir kdir --mode kmer`
- **tools:** `kmtricks aggregate --run-dir kdir --matrix kmer --format text > matrix.txt`

API

```
KmerMerger merger(...);  
while (merger.next()) {  
    // matrix streaming  
}
```

```
Repartition repart(...);  
repart.get_partition(kmer.minimizer());
```

PLUGIN

- Easily extend kmtricks features

kmtricks plugin: a stupid example

Implementation: explugin.cpp

```
#include <kmtricks/plugin.hpp>
class ExPlugin : public km::IMergePlugin
{
    public:
        bool process_kmer(...) override {
            if (counts[0] > 42)
                return true; // keep row
            return false;    // discard row
        }
};
```

Usage

```
kmtricks --plugin libexplugin.so [kmtricks args...]
```

Complete examples are available on the github wiki.

- Efficient and flexible k-mer matrix toolbox
 - Tara Ocean: 36h instead of week
- Supports large datasets
 - Tara Ocean
 - Applied on large human cohorts at Institut Pasteur
- Comes with a set of utilities/API for downstream analysis

- Efficient and flexible k-mer matrix toolbox
 - Tara Ocean: 36h instead of week
- Supports large datasets
 - Tara Ocean
 - Applied on large human cohorts at Institut Pasteur
- Comes with a set of utilities/API for downstream analysis

Future work:

- Support findere¹ algorithm
 - "Free" multi-hash Bloom filter
 - Reduce index size and query time

¹Robidou and Peterlongo, "findere : Fast and Precise Approximate Membership Query"

A reversed approach

- Identification of all sequences (k-mers) associated with the phenotype
- Characterization of SVs in these sequences

A reversed approach

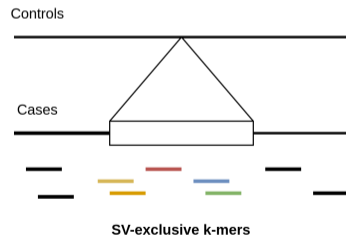
- Identification of all sequences (k-mers) associated with the phenotype
- Characterization of SVs in these sequences

| | 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 |

A reversed approach

- Identification of all sequences (k-mers) associated with the phenotype
- Characterization of SVs in these sequences

| | 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 |



Statistical test:

- Likelihood ratio assuming Poisson distribution ¹

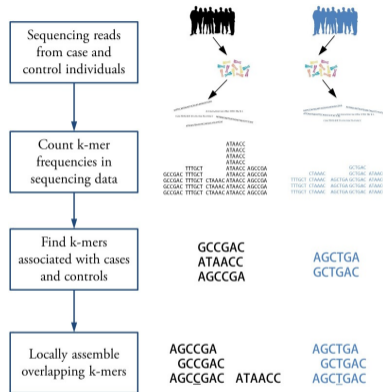
FWER/FDR control:

- Bonferroni
- Benjamini-Hochberg

Correction of population stratification:

- PCA on a random subset of counted k-mers ^{2 3}

Implemented in **HAWK** (Hitting Association With K-mers)



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

²Price et al., "Principal components analysis corrects for stratification in genome-wide association studies"

³Patterson, Price, and Reich, "Population Structure and Eigenanalysis"

Pros:

- Good recall
- Considers population stratification

Pros:

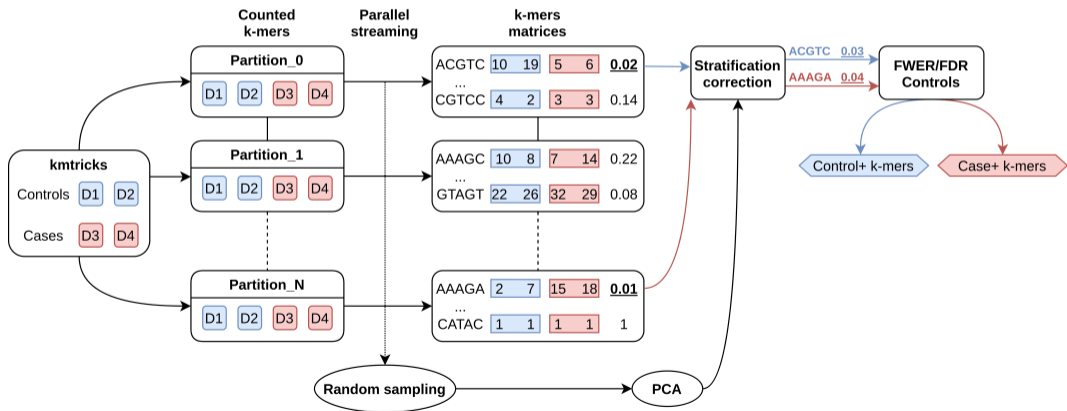
- Good recall
- Considers population stratification

Cons (kmdiff motivations):

- Doesn't scale up
- Limited to 31-mers
- Outputs are limited to significant k-mers
- Not very user-friendly

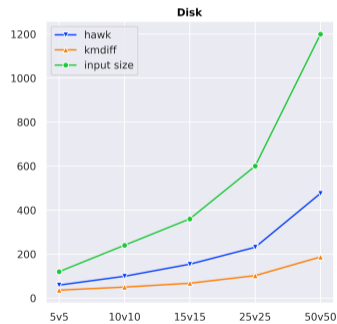
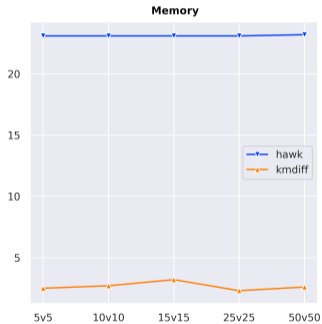
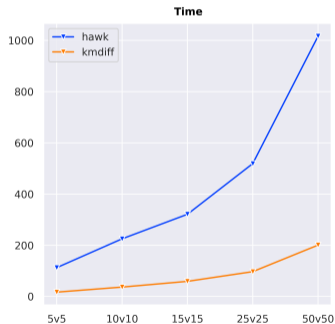
kmdiff overview (WIP)

- Basically: kmtricks + HAWK statistical methods
 - Provides same functionalities but more efficiently



Simulated data:

- Human chromosome 1, 20X, 1% errors, 100bp
- Insertion, deletion, inversion, $|SV| = 500 \pm 200$



*Use an older (slower) version of kmtricks

- Scalable
- More flexible
 - Unlimited k-mer size (for recent sequencing data types)
 - Designed to add new models, like kmtricks it will probably support plugin later

- Scalable
- More flexible
 - Unlimited k-mer size (for recent sequencing data types)
 - Designed to add new models, like kmtricks it will probably support plugin later

Future work:

- Application on real data (Alzheimer, Parkinson)
- **SVs characterization** (i.e significant k-mer set to VCF)

Thank you!