DSB2022
fimpera: low memory counting Approximate Membership Query
Lucas Robidou

Lucas Robidou, Pierre Peterlongo

Düsseldorf, 2022-06-13

Inria Rennes

context

context

fimpera

context

fimpera

results

Summary

context

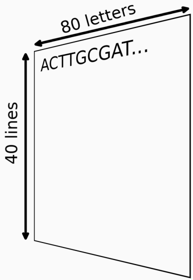
fimpera

results

Some context

My dream:

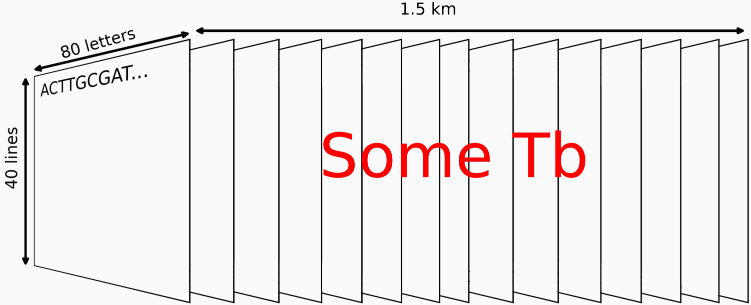
- to index (large) genomic datasets
- to query those indexed datasets



Some context

My dream:

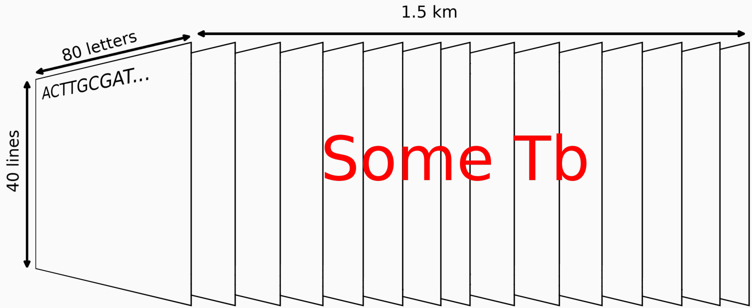
- to index (large) genomic datasets
- to query those indexed datasets



Some context

My dream:

- to index (large) genomic datasets
- to query those indexed datasets



Camille Marchet, Christina Boucher, Simon J Puglisi, Paul Medvedev, Mikaël Salson, and Rayan Chikhi. **Data structures based on k-mers for querying large collections of sequencing data sets.** Genome Research, 31(1):1–12, 2021.

Challenges

- indexation time
- **abundance storage**
- **index size**
- query time
- **false positive rate**

How to compare sequences

- extract every subsequence of size *k* (*k*-mers)
- count *k*-mers
- index them along with their abundance
- query abundance of every *k*-mer from your queried sequence

Summary

context

fimpera

results

Main idea of fimpera

Let’s consider the 13-mer ‘datastructure’. Its 11-mers are:

- ‘datastructu’ (abundance: 5)
- ‘atastructur’ (abundance: 2)
- ‘tastructure’ (abundance: 4)

$\Longrightarrow$ abundance of ‘datastructure‘ can’t be more than 2.

Some notations

Rather than indexing k-mers, **let's index s-mers**, $s < k$.
Let's introduce $z = k - s$, so that a k-mer is made of $z + 1$ smaller s-mers.

A k-mer is said 'found' iif the $z + 1$ s-mers composing it are found in the filter.

Indexation with fimpera

- count *k*-mers

Indexation with fimpera

- count k-mers
- compute s-mers abundance (max of k-mers count)

Indexation with fimpera

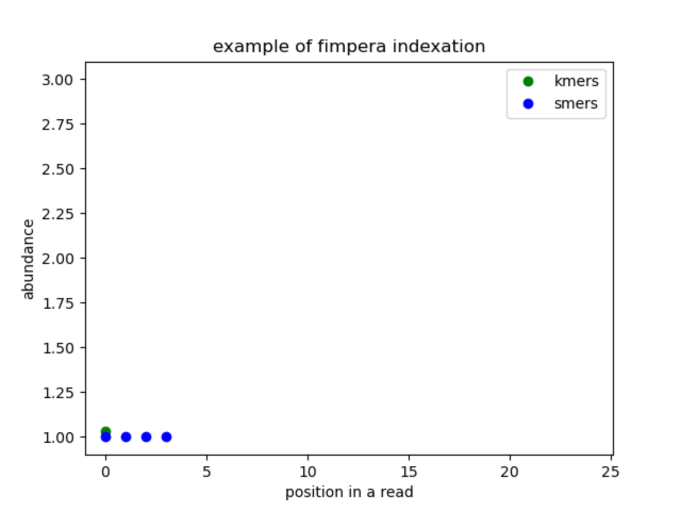
- count *k*-mers
- compute *s*-mers abundance (max of *k*-mers count)
- index *s*-mers along their abundance in a data structure ($*$)

Indexation with fimpera

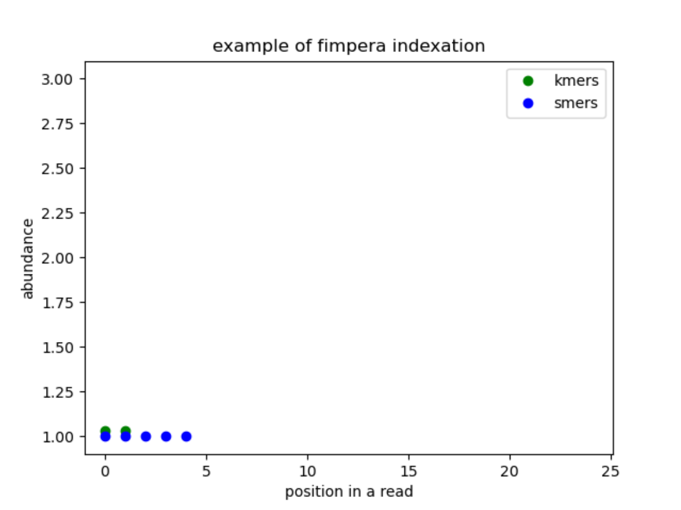
- count *k*-mers
- compute *s*-mers abundance (max of *k*-mers count)
- index *s*-mers along their abundance in a data structure $(*)$

$(*)$ e.g. counting Bloom filter

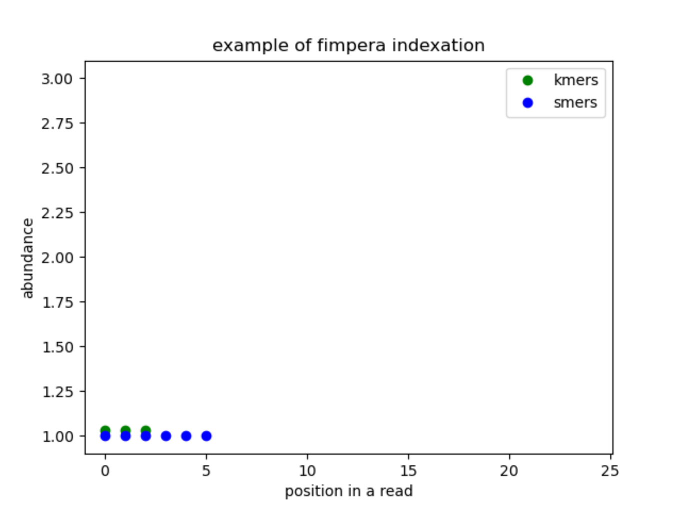
Indexation with fimpera



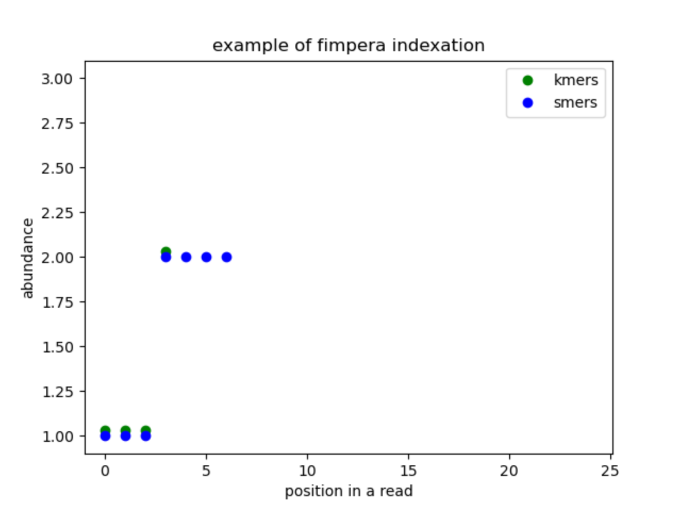
Indexation with fimpera



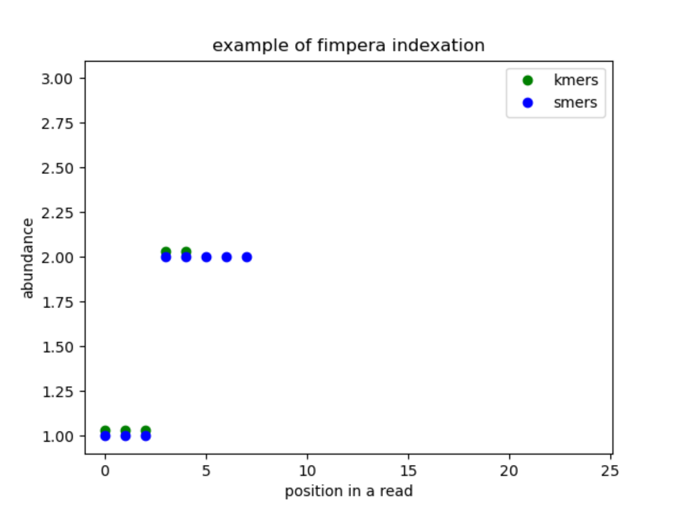
Indexation with fimpera



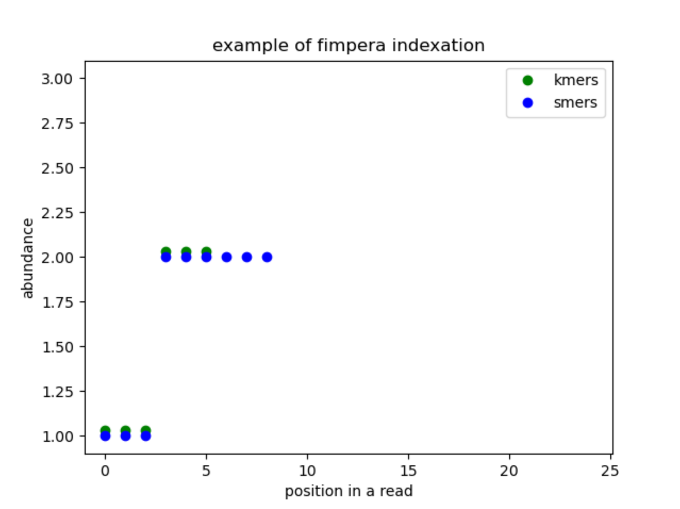
Indexation with fimpera



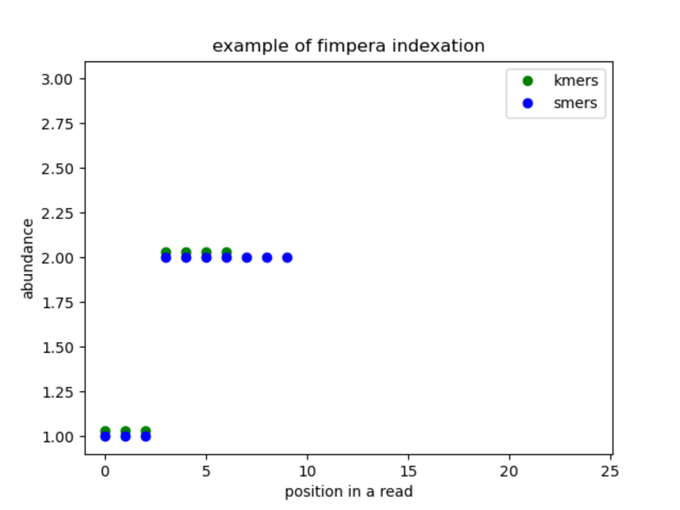
Indexation with fimpera



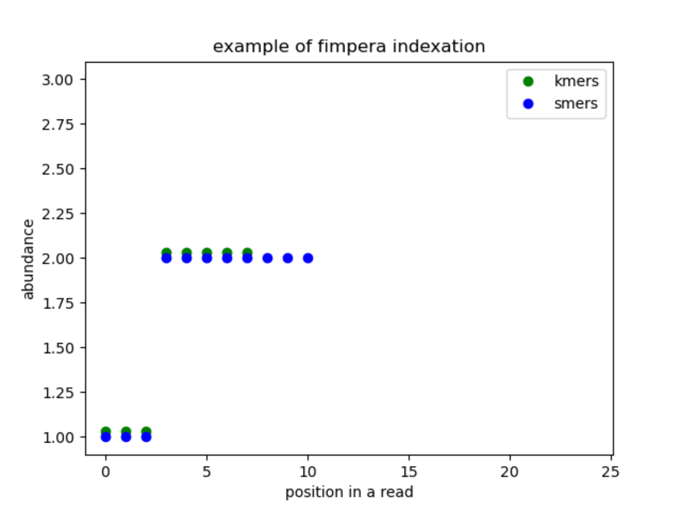
Indexation with fimpera



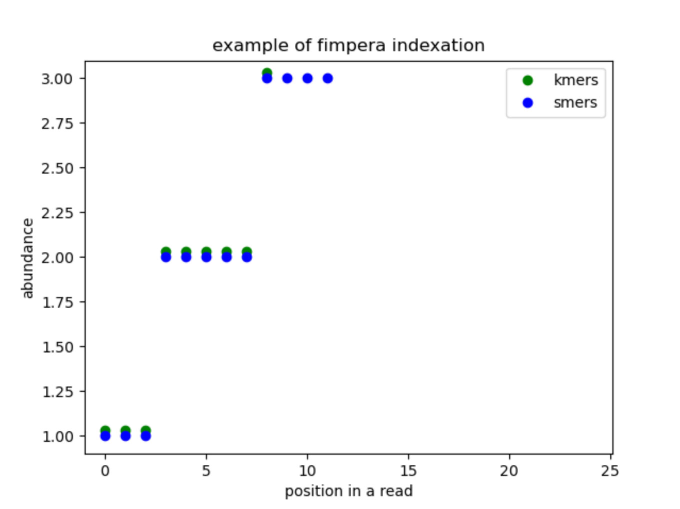
Indexation with fimpera



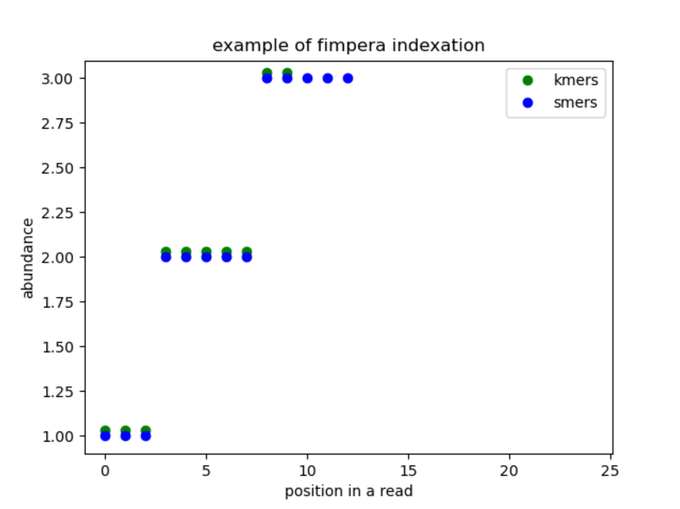
Indexation with fimpera



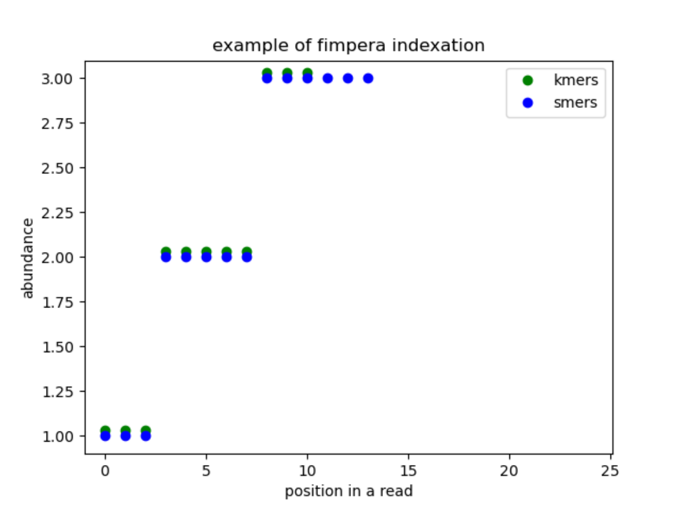
Indexation with fimpera



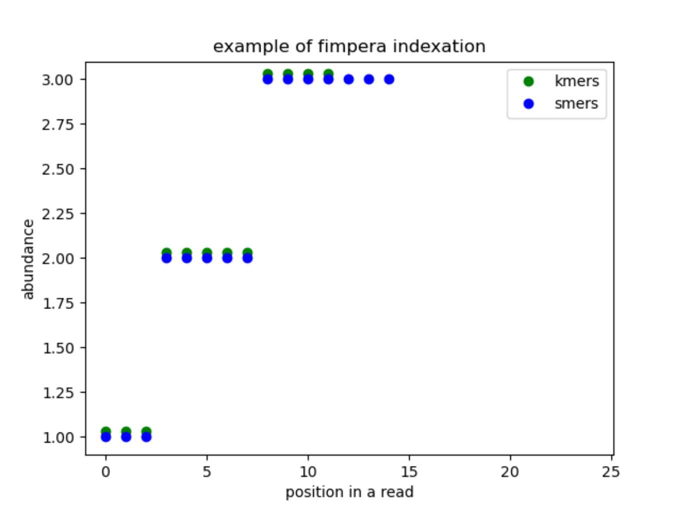
Indexation with fimpera



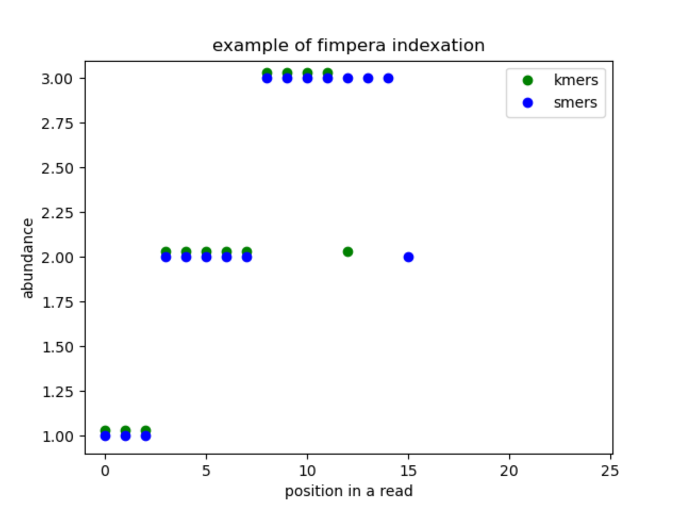
Indexation with fimpera



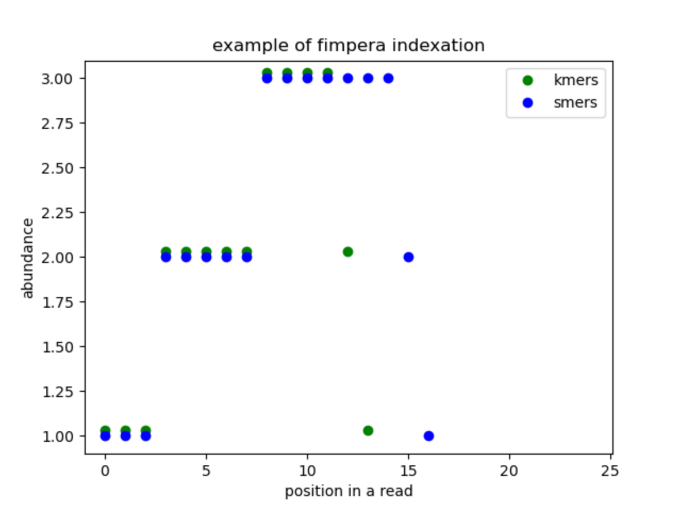
Indexation with fimpera



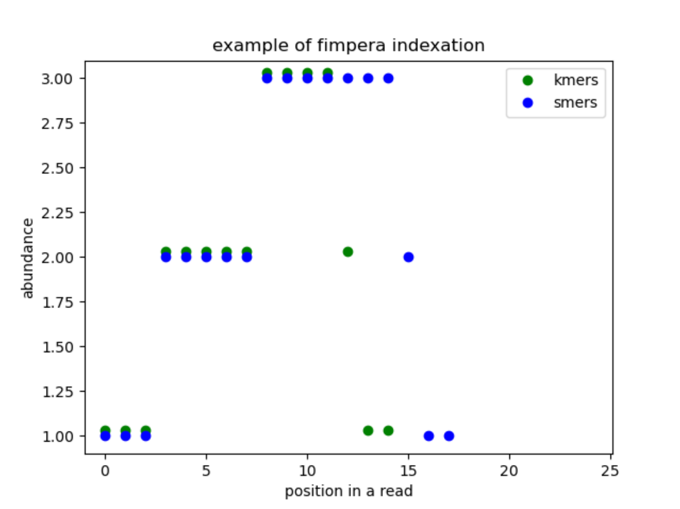
Indexation with fimpera



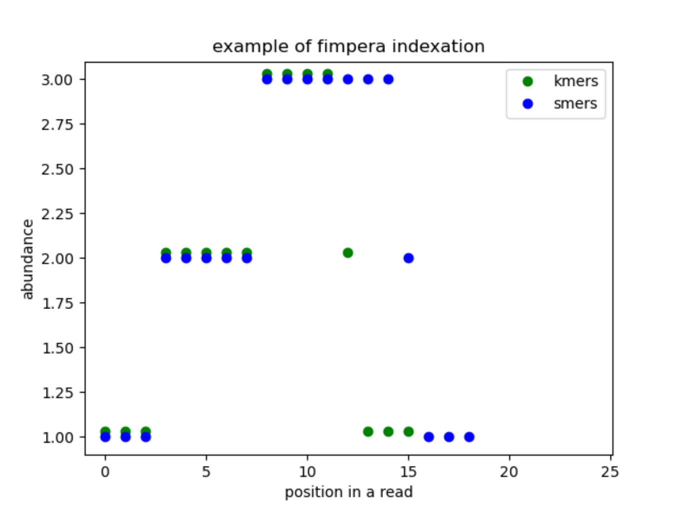
Indexation with fimpera



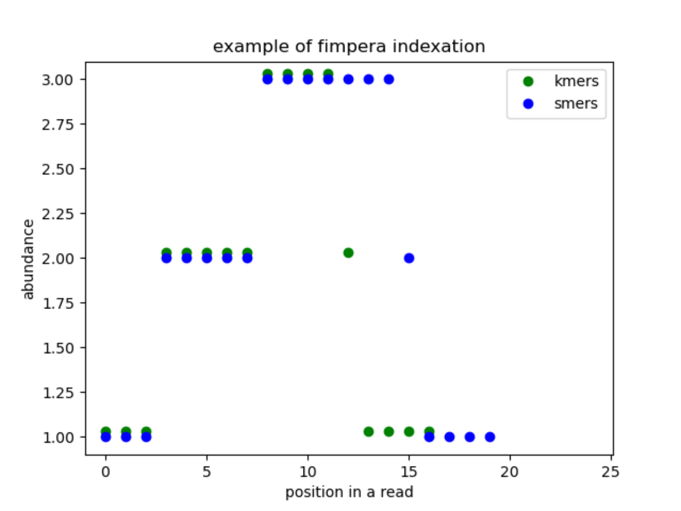
Indexation with fimpera



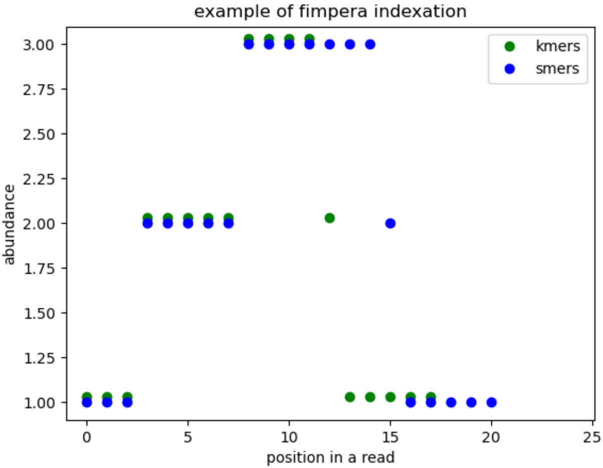
Indexation with fimpera



Indexation with fimpera



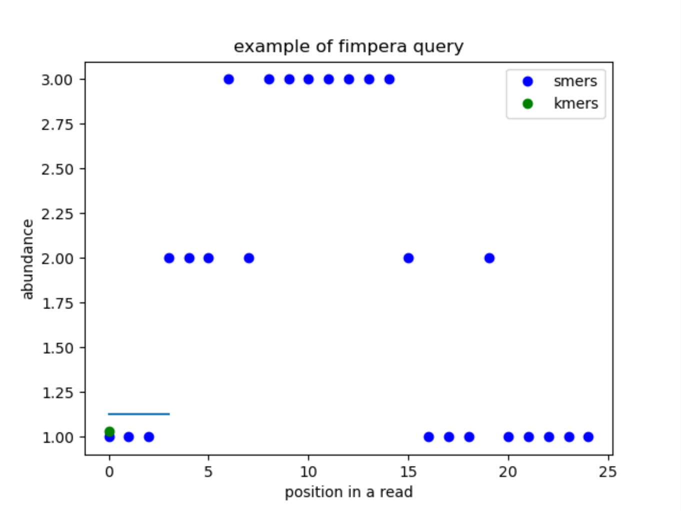
Indexation with fimpera



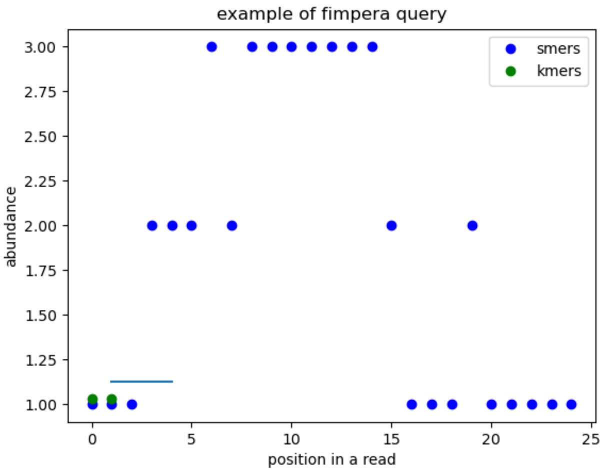
Query on fimpera

- query abundance of every *s*-mers
- compute *k*-mers abundance (min of *s*-mers abundance)

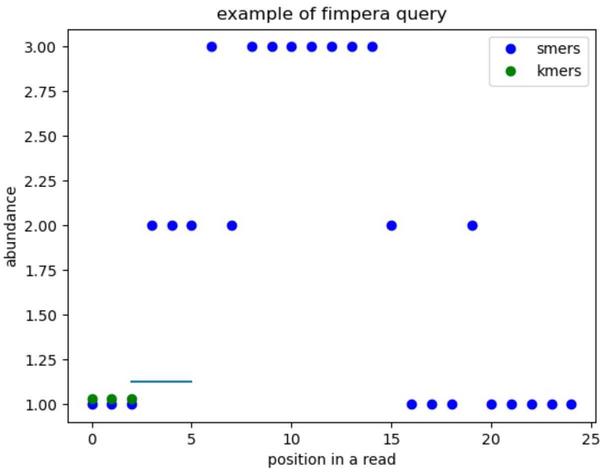
Query on fimpera



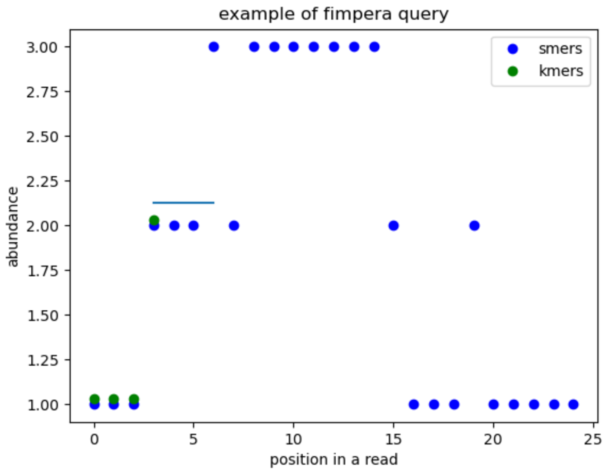
Query on fimpera



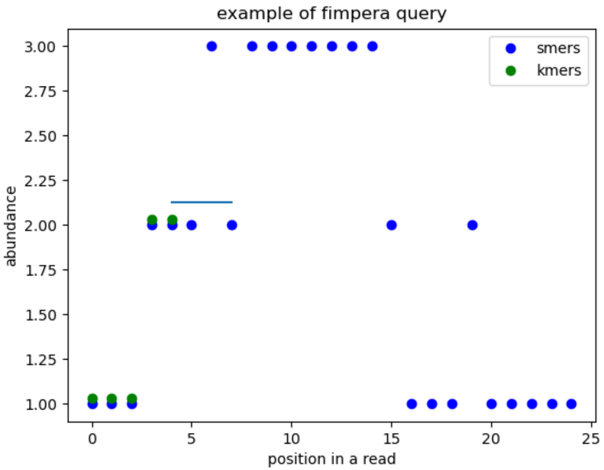
Query on fimpera



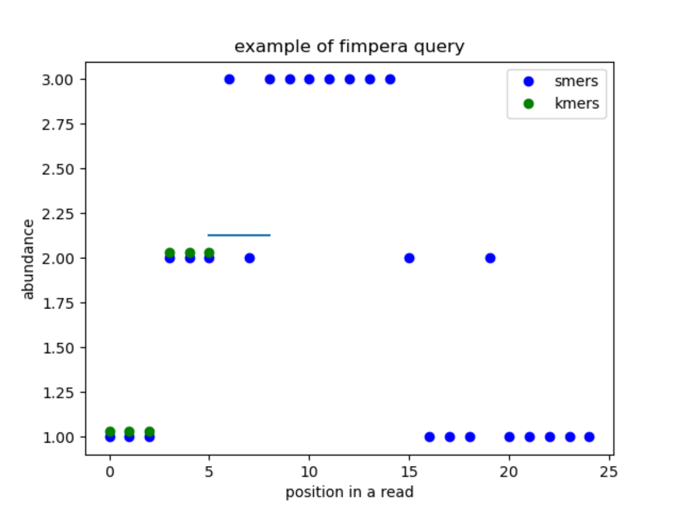
Query on fimpera



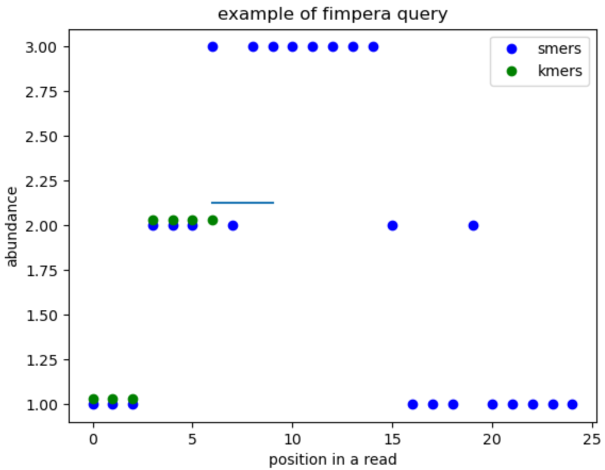
Query on fimpera



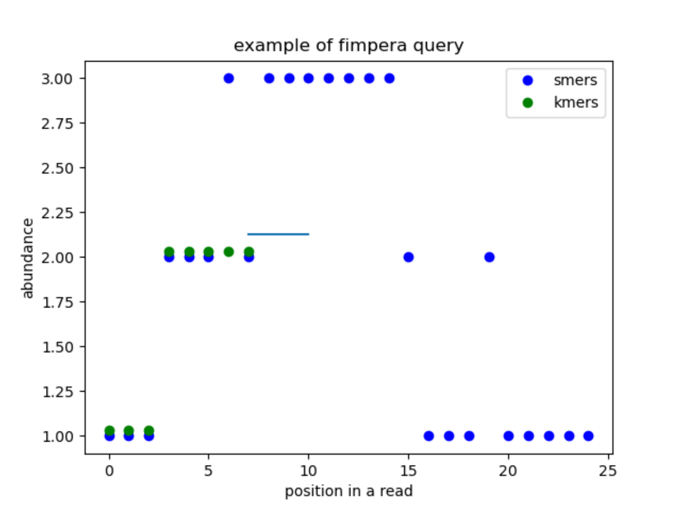
Query on fimpera



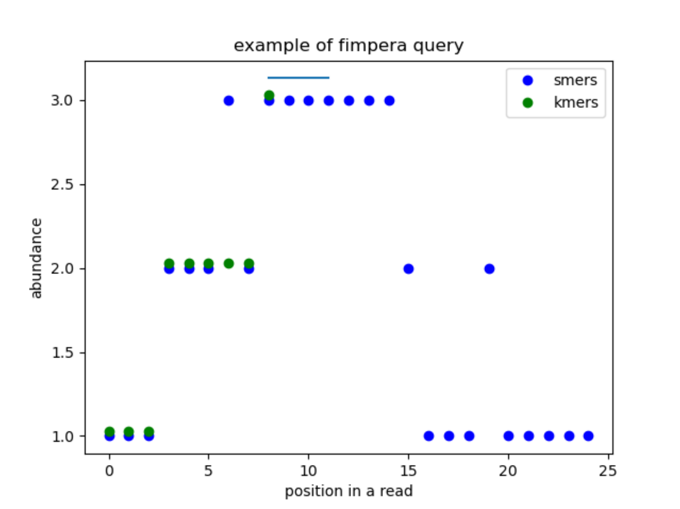
Query on fimpera



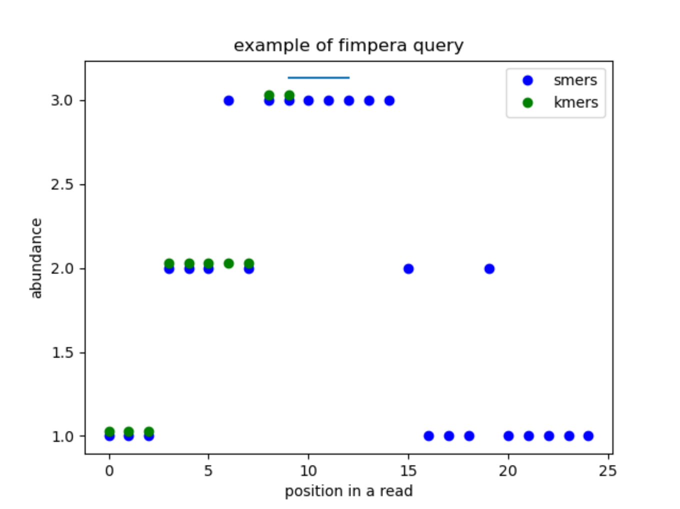
Query on fimpera



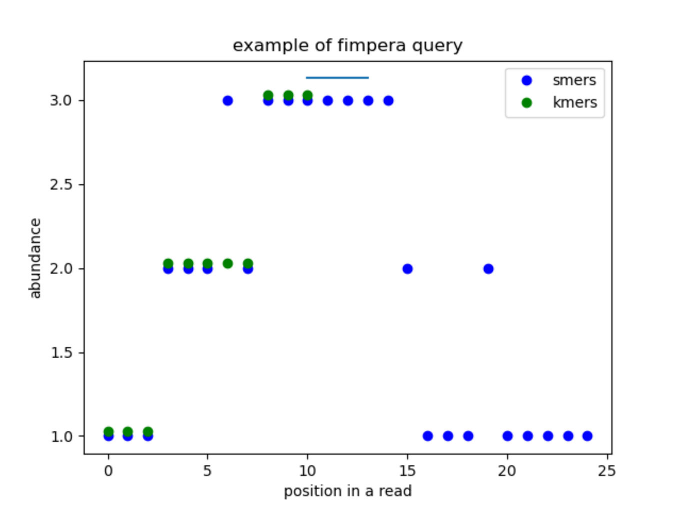
Query on fimpera



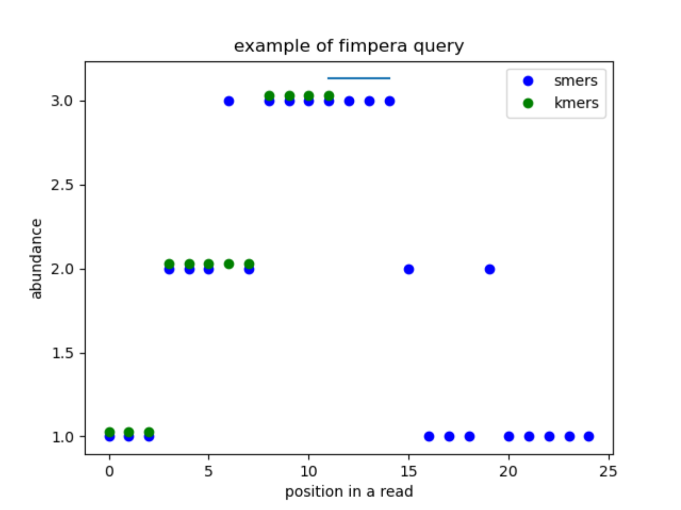
Query on fimpera



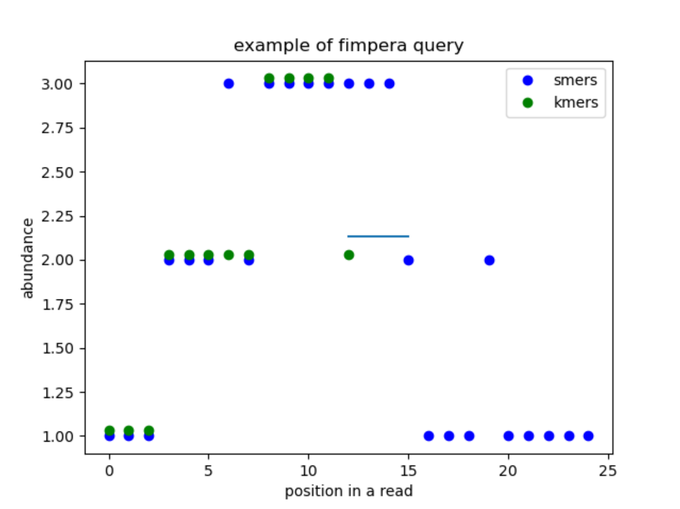
Query on fimpera



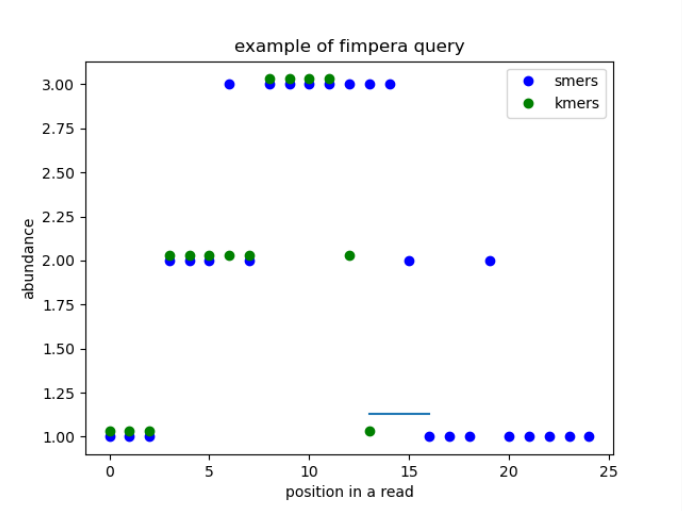
Query on fimpera



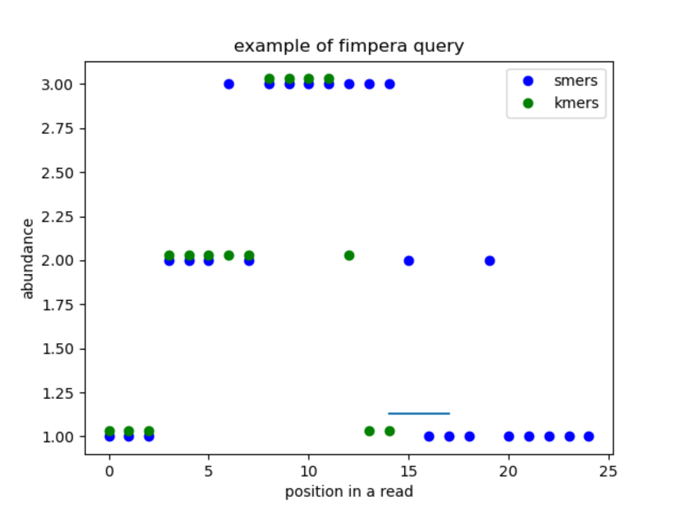
Query on fimpera



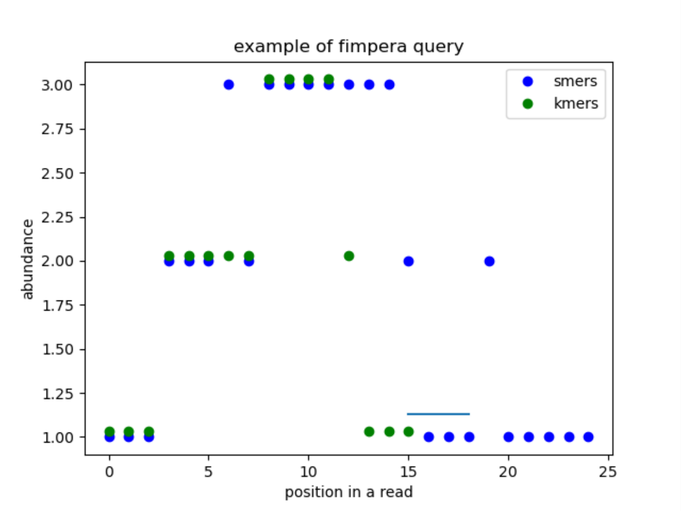
Query on fimpera



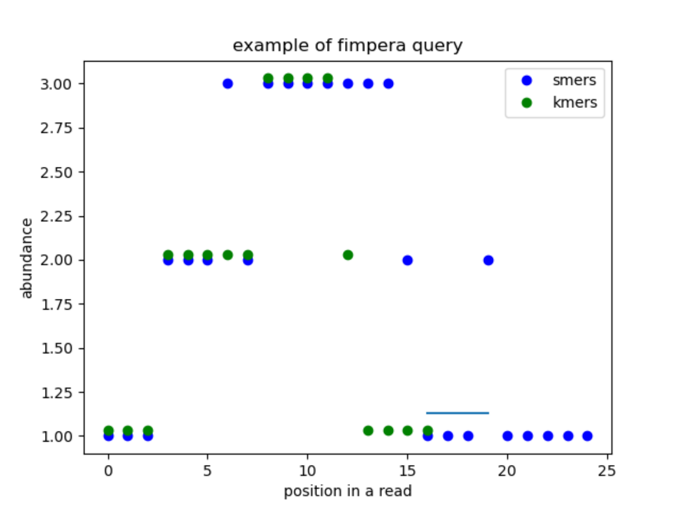
Query on fimpera



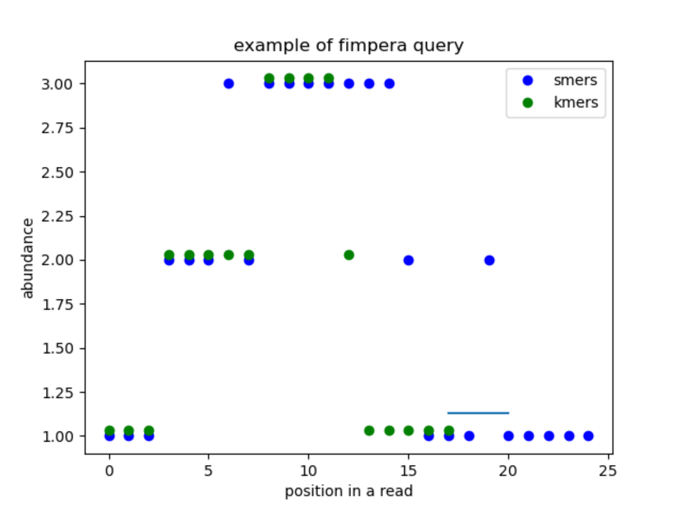
Query on fimpera



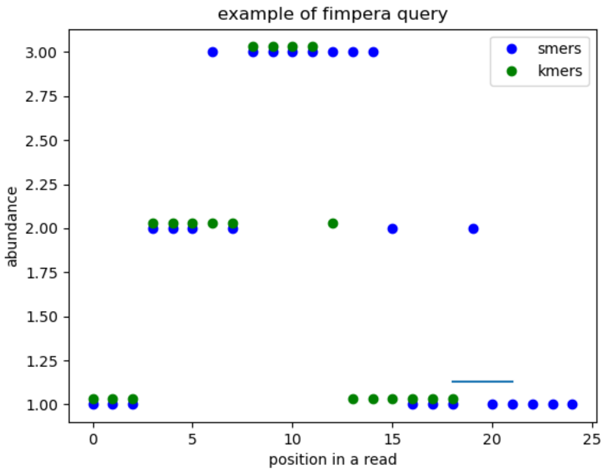
Query on fimpera



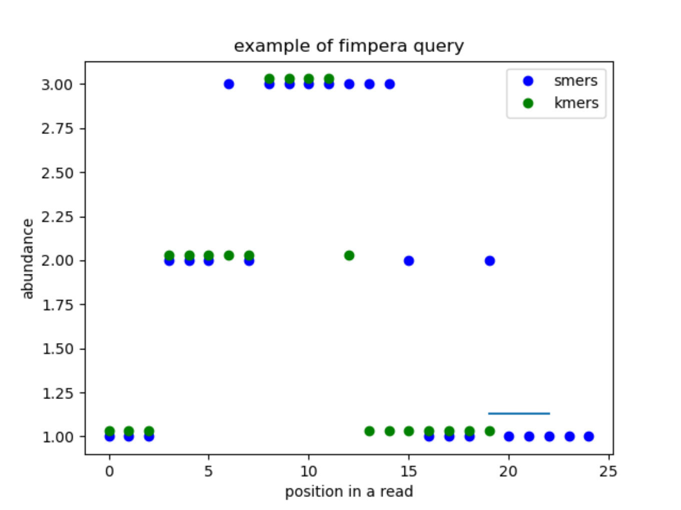
Query on fimpera



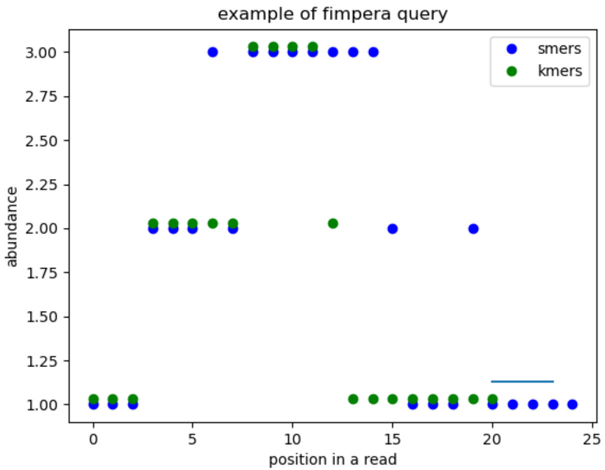
Query on fimpera



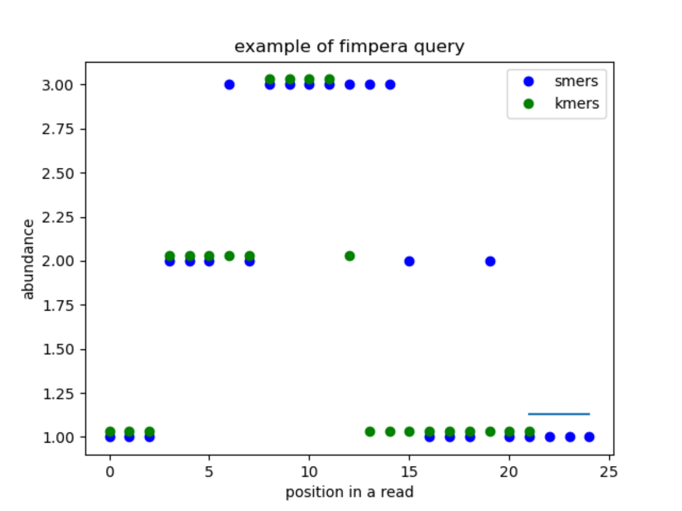
Query on fimpera



Query on fimpera



Query on fimpera



limitation of fimpera

For a chosen k, if z is too high, then fimpera will index and query very small s-mers. In such case, the probability of having indexed all those s-mers is *high*.

Example of construction overestimation

- indexing ‘ACTGAC’ with $s = 3$
- indexed s-mers include ‘GAC’, ‘ACT’ and ‘CTG’
- k-mer ‘GACTG’ would be found with the abundance of ‘ACTGAC’

Summary

context

fimpera

results

Data used

- two fastq files from the TARA ocean dataset (metagenomic)
- one is indexed
- 1,000,000 reads are queried from the second

fimpera’s effect on false positive rate

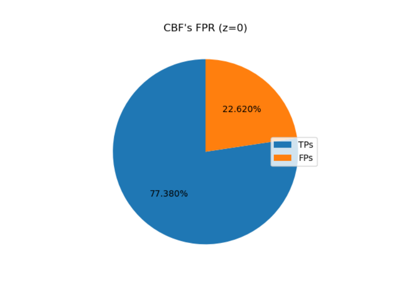


Figure 1: proportion of false positive calls **without** fimpera

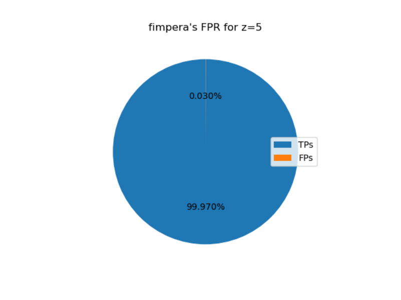


Figure 2: proportion of false positive calls **with** fimpera

fimpera’s effect on abundance correctness

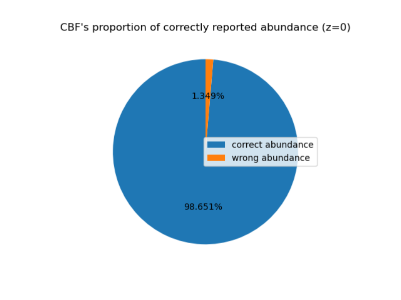


Figure 3: proportion correct abundance calls **without** fimpera

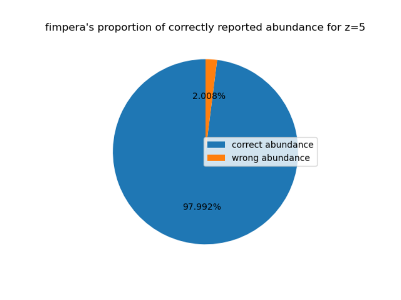


Figure 4: proportion correct abundance calls **with** fimpera

fimpera's effect on abundance error

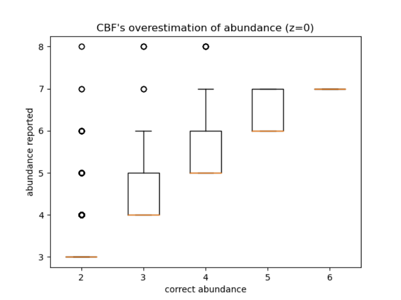


Figure 5: overestimations **without** fimpera

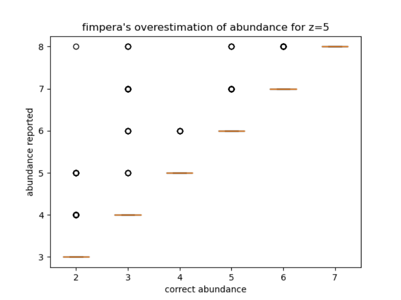


Figure 6: overestimations **with** fimpera