UNIVERSITÄT DES SAARLANDES

CLEANIFIER: Removing human DNA contamination with a pangenomic gapped *k*-mer index

Jens Zentgraf, Johanna Elena Schmitz, Sven Rahmann

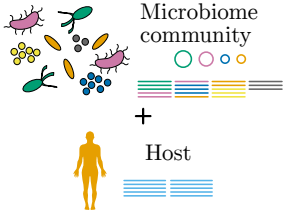
Saarland University

GCB 2025

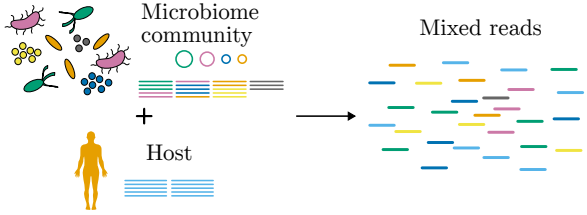
python Package Index

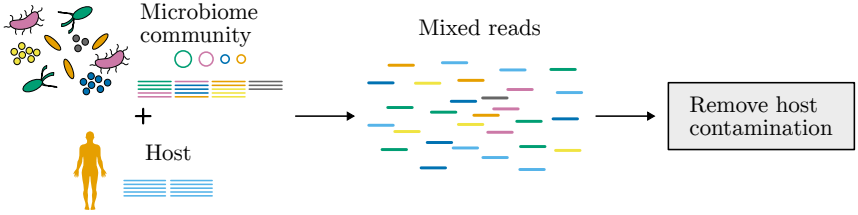
BIOCONDA
Reproducible research.
Install the software from bioconda:
> conda install -c bioconda cleanifier
See bioconda.github.io

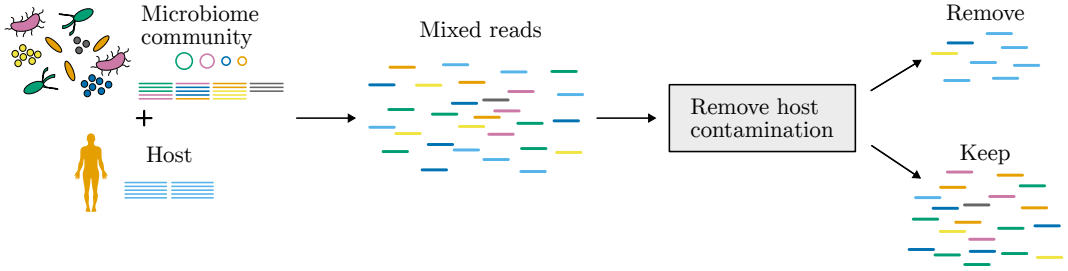
Motivation



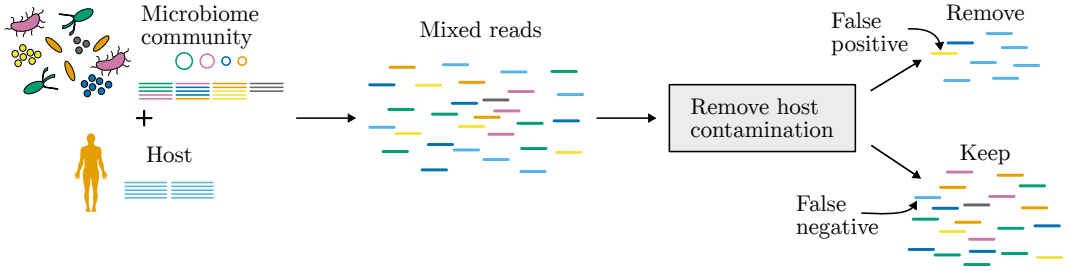
Motivation

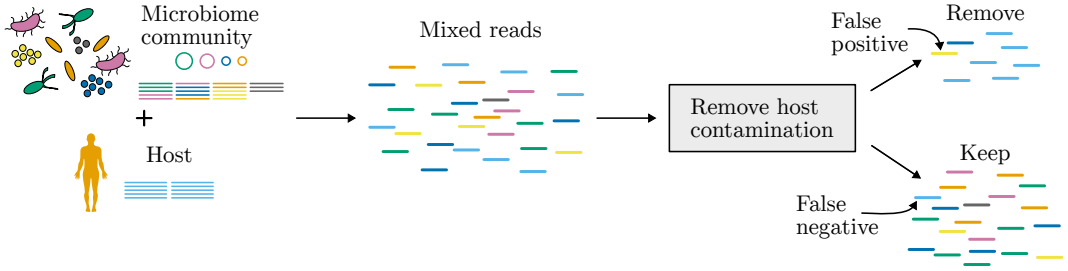



Microbiome community
+
Host
Mixed reads
Remove host contamination


Microbiome community
+
Host
Mixed reads
Remove host contamination
Remove
Keep

Motivation

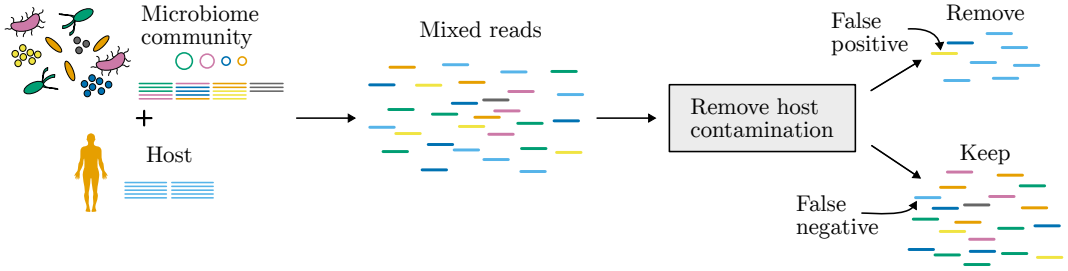

Microbiome community
+
Host
Mixed reads
Remove host contamination
False positive
Remove
Keep
False negative



Privacy

- Human data cannot be made public

Motivation



Privacy

- Human data cannot be made public

Downstream analysis

- Reduce problems in binning and assembly

Definitions

- k-mers
  - Substrings of length k

```
CGATCGACTAGCATCGAACGTACG . . .
k-mer      rc          canonical
```

Definitions

- k-mers
  - Substrings of length k

```
CGATCGACTAGCATCGAACGTACG . . .
k-mer      rc          canonical
CGATC
 GATCG
  ATCGA
   TCGAC
    CGACT
     .....
```

Definitions

- *k*-mers
  - Substrings of length *k*
- Reverse complements (rc)
  - Reverse order
  - `A` $\leftrightarrow$ `T`, `C` $\leftrightarrow$ `G`

```
CGATCGACTAGCATCGAACGTACG . . .
k-mer      rc          canonical
CGATC
 GATCG
  ATCGA
   TCGAC
    CGACT
     .....
```

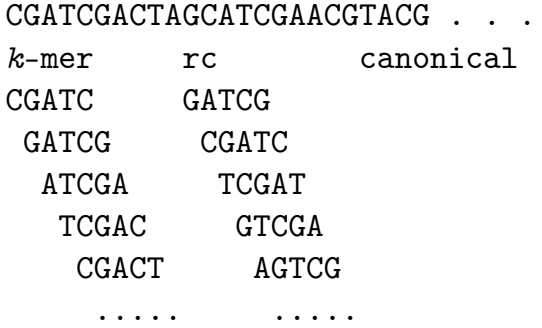
Definitions

- k-mers
  - Substrings of length k
- Reverse complements (rc)
  - Reverse order
  - `A` $\leftrightarrow$ `T`, `C` $\leftrightarrow$ `G`

```
CGATCGACTAGCATCGAACGTACG . . .
k-mer      rc          canonical
CGATC      GATCG
 GATCG      CGATC
  ATCGA      TCGAT
   TCGAC      GTCGA
    CGACT      AGTCG
     .....      .....
```

Definitions

- *k*-mers
  - Substrings of length *k*
- Reverse complements (rc)
  - Reverse order
  - `A` ↔ `T`, `C` ↔ `G`
- Canonical *k*-mer
  - Maximum of *k*-mer and rc(*k*-mer)



Definitions

- k-mers
  - Substrings of length k
- Reverse complements (rc)
  - Reverse order
  - A $\leftrightarrow$ T, C $\leftrightarrow$ G
- Canonical k-mer
  - Maximum of k-mer and rc(k-mer)

```
CGATCGACTAGCATCGAACGTACG . . .
k-mer      rc          canonical
CGATC      GATCG       GATCG
 GATCG      CGATC       GATCG
  ATCGA      TCGAT       TCGAT
   TCGAC      GTCGA       TCGAC
    CGACT      AGTCG       CGACT
     .....      .....       .....
```

Definitions

- k-mers
  - Substrings of length k
- Reverse complements (rc)
  - Reverse order
  - `A` $\leftrightarrow$ `T`, `C` $\leftrightarrow$ `G`
- Canonical k-mer
  - Maximum of k-mer and rc(k-mer)
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - `##_#_##`

```
CGATCGACTAGCATCGAACGTACG . . .
k-mer      rc          canonical
CGATC      GATCG       GATCG
 GATCG      CGATC       GATCG
  ATCGA      TCGAT       TCGAT
   TCGAC      GTCGA       TCGAC
    CGACT      AGTCG       CGACT
     .....      .....       .....
```

Definitions

- k-mers
  - Substrings of length k
- Reverse complements (rc)
  - Reverse order
  - A $\leftrightarrow$ T, C $\leftrightarrow$ G
- Canonical k-mer
  - Maximum of k-mer and rc(k-mer)
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - ##_#_##

```
CGATCGACTAGCATCGAACGTACG . . .
##_#_##
CG T GA
 GA C AC
  AT G CT
   TC A TA
    CG C AG
      . . .
```

Definitions

- k-mers
  - Substrings of length k
- Reverse complements (rc)
  - Reverse order
  - A $\leftrightarrow$ T, C $\leftrightarrow$ G
- Canonical k-mer
  - Maximum of k-mer and rc(k-mer)
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - ##_#_##

```
CGATCGACTAGCATCGAACGTACG . . .
##_#_##
CG T GA
 GA C AC
  AT G CT
   TC A TA
    CG C AG
     . . .
```

- More robust against substitutions
- *Design of Worst-Case-Optimal Spaced Seeds* at WABI 2025

- HOSTILE
  - Alignment based
  - Human reference
  - BOWTIE2 or MINIMAP2

- HOSTILE
  - Alignment based
  - Human reference
  - BOWTIE2 or MINIMAP2
- KRAKEN2
  - *k*-mer based
  - Metagenomic classification
  - Default database (very large)

- HOSTILE
  - Alignment based
  - Human reference
  - BOWTIE2 or MINIMAP2
- KRAKEN2
  - *k*-mer based
  - Metagenomic classification
  - Default database (very large)
- NOHUMAN
  - KRAKEN2 wrapper
  - Custom Human database (smaller)

Comparison

- HOSTILE
  - Alignment based
  - Human reference
  - BOWTIE2 or MINIMAP2
- KRAKEN2
  - *k*-mer based
  - Metagenomic classification
  - Default database (very large)
- NOHUMAN
  - KRAKEN2 wrapper
  - Custom Human database (smaller)

- HRRT
  - *k*-mer based
  - Min-hash based (Jaccard Similarity)
  - Include all *k*-mers from human derived eukaryotic species
  - Exclude all *k*-mers in non-eukaryotic species

Comparison

- HOSTILE
  - Alignment based
  - Human reference
  - BOWTIE2 or MINIMAP2
- KRAKEN2
  - *k*-mer based
  - Metagenomic classification
  - Default database (very large)
- NOHUMAN
  - KRAKEN2 wrapper
  - Custom Human database (smaller)

- HRRT
  - *k*-mer based
  - Min-hash based (Jaccard Similarity)
  - Include all *k*-mers from human derived eukaryotic species
  - Exclude all *k*-mers in non-eukaryotic species
- DEACON
  - Minimizer based
  - Pangenome approach

- HOSTILE
  - Alignment based
  - Human reference
  - BOWTIE2 or MINIMAP2
- KRAKEN2
  - *k*-mer based
  - Metagenomic classification
  - Default database (very large)
- NOHUMAN
  - KRAKEN2 wrapper
  - Custom Human database (smaller)

- HRRT
  - *k*-mer based
  - Min-hash based (Jaccard Similarity)
  - Include all *k*-mers from human derived eukaryotic species
  - Exclude all *k*-mers in non-eukaryotic species
- DEACON
  - Minimizer based
  - Pangenome approach
- CLEANIFIER
  - Gapped *k*-mer based
  - Pangenome approach

_####### _### _####### _###### $k = 29, w = 33$

Human references

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference
- 1000 Genome Project
  - Extract all variants
    (Substitution, Insertion and Deletion)
  - Allele frequency of at least 1%

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference
- 1000 Genome Project
  - Extract all variants (Substitution, Insertion and Deletion)
  - Allele frequency of at least 1%
- Human Pangenome Refence Consotrium
  - 47 assemblies

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference
- 1000 Genome Project
  - Extract all variants (Substitution, Insertion and Deletion)
  - Allele frequency of at least 1%
- Human Pangenome Refence Consotrium
  - 47 assemblies
- IPD-IMGT/HLA database
  - HLA is highly variable

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference
- 1000 Genome Project
  - Extract all variants (Substitution, Insertion and Deletion)
  - Allele frequency of at least 1%
- Human Pangenome Refence Consotrium
  - 47 assemblies
- IPD-IMGT/HLA database
  - HLA is highly variable
- cDNA transcripts
  - Deal with RNA

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference
- 1000 Genome Project
  - Extract all variants (Substitution, Insertion and Deletion)
  - Allele frequency of at least 1%
- Human Pangenome Refence Consotrium
  - 47 assemblies
- IPD-IMGT/HLA database
  - HLA is highly variable
- cDNA transcripts
  - Deal with RNA

Data structures

1. Bucketed Cuckoo hash table
   - Exact data structure
   - Stores the k-mers
   - Size 13.85 GB

_####### _### _####### _###### $k = 29, w = 33$

Human references

- T2T reference
- 1000 Genome Project
  - Extract all variants (Substitution, Insertion and Deletion)
  - Allele frequency of at least 1%
- Human Pangenome Refence Consotrium
  - 47 assemblies
- IPD-IMGT/HLA database
  - HLA is highly variable
- cDNA transcripts
  - Deal with RNA

Data structures

1. Bucketed Cuckoo hash table
   - Exact data structure
   - Stores the k-mers
   - Size 13.85 GB
2. Windowed Cuckoo filter
   - Probabilistic set membership data structure
   - Store a fingerprint (of p bits) instead of the k-mer
   - False positive rate of $2^{-p} = 2^{-14}$
   - Size 6.9 GB

Data structure

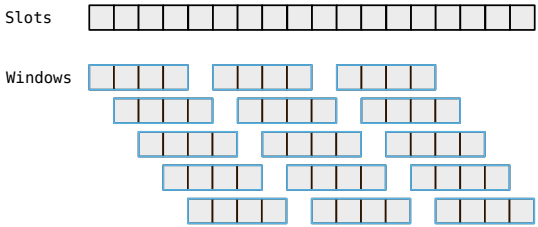
- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989

Data structure

- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989



Data structure

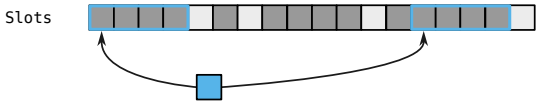
- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989

Data structure

- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

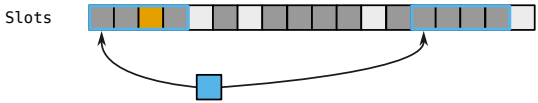
- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989



Data structure

- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

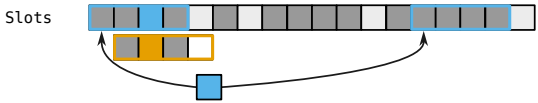
- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989



Data structure

- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

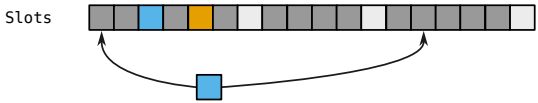
- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989



Data structure

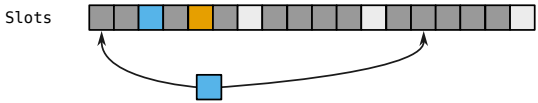
- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}

- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989

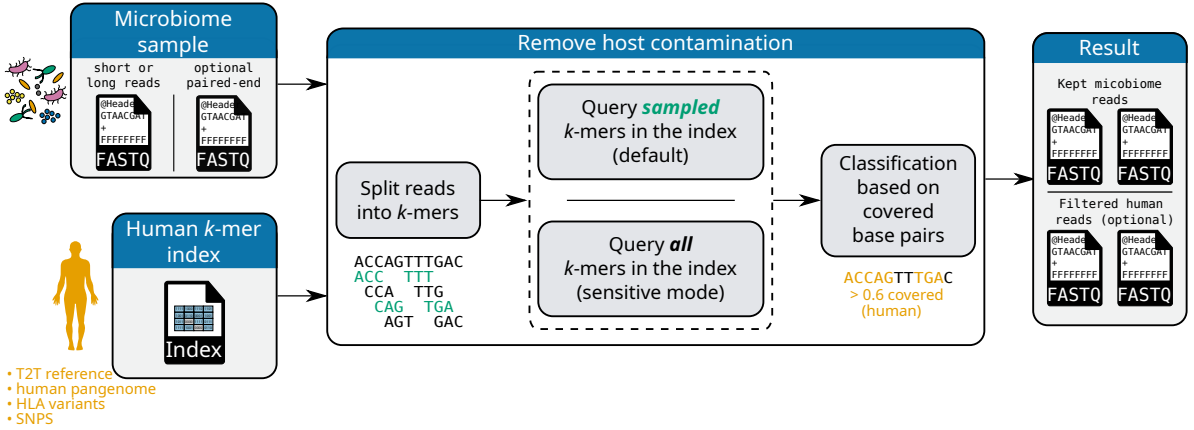


Data structure

- Probabilistic set membership data structure
- Store a fingerprint (of p bits) instead of the k-mer
- False positive rate of 2^{-p}
- d hash functions
- Window size of ℓ
- High fill rates possible
  - $d = 2$, $\ell = 4$ fill rate of ≈ 0.9989



Smaller and More Flexible Cuckoo Filters
at ALENEX 2026


Microbiome sample
short or long reads
optional paired-end
@Heade
GTAACGAT
+
FFFFFFFF
FASTQ
Human k-mer index
Index
• T2T reference
• human pangenome
• HLA variants
• SNPS
Remove host contamination
Split reads into k-mers
ACCAGTTTGAC
ACC TTT
CCA TTG
CAG TGA
AGT GAC
Query sampled k-mers in the index (default)
Query all k-mers in the index (sensitive mode)
Classification based on covered base pairs
ACCAGTTTGAC
> 0.6 covered (human)
Result
Kept micobiome reads
Filtered human reads (optional)

Sensitive mode

- Query all gapped k-mers
- All k bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

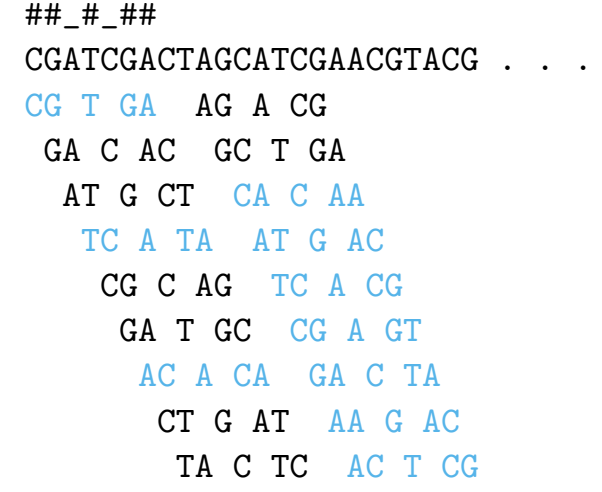
- Query all gapped k-mers
- All k bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

```
##_#_##
CGATCGACTAGCATCGAACGTACG . . .
```

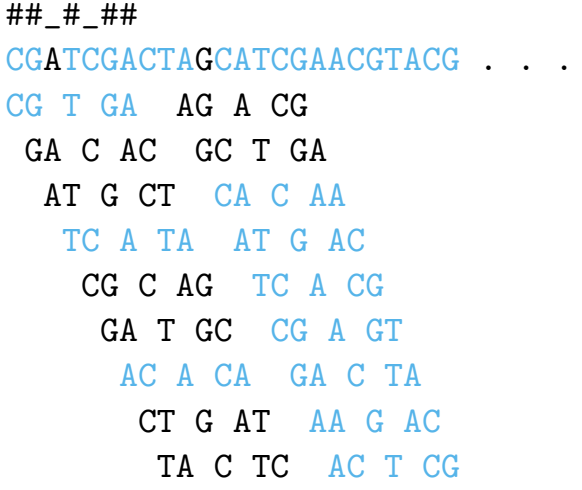
- Query all gapped k-mers
- All k bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

```
##_#_##
CGATCGACTAGCATCGAACGTACG . . .
CG T GA  AG A CG
 GA C AC  GC T GA
  AT G CT  CA C AA
   TC A TA  AT G AC
    CG C AG  TC A CG
     GA T GC  CG A GT
      AC A CA  GA C TA
       CT G AT  AA G AC
        TA C TC  AC T CG
```

- Query all gapped k-mers
- All k bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not



- Query all gapped k-mers
- All k bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not



Sampling mode

- Query every $\lfloor w/2 \rfloor$ gapped k-mer
- All w bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

Sampling mode

- Query every $\lfloor w/2 \rfloor$ gapped k-mer
- All w bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

```
##_#_##
CGATCGACTAGCATCGAACGTACG . . .
```

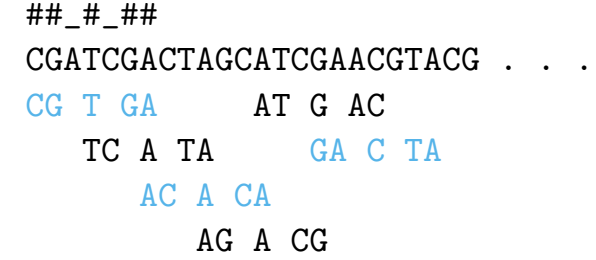
Sampling mode

- Query every $\lfloor w/2 \rfloor$ gapped k-mer
- All w bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

```
##_#_##
CGATCGACTAGCATCGAACGTACG . . .
CG T GA     AT G AC
   TC A TA     GA C TA
      AC A CA
         AG A CG
```

Sampling mode

- Query every $\lfloor w/2 \rfloor$ gapped k-mer
- All w bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

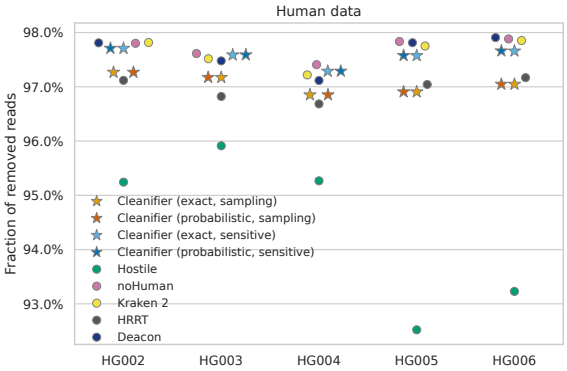


Sampling mode

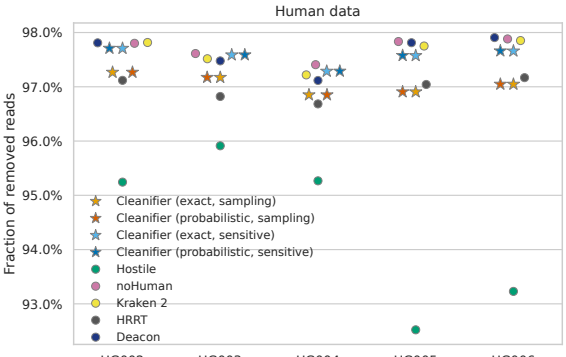
- Query every $\lfloor w/2 \rfloor$ gapped k-mer
- All w bases count as covered
- Check how many bases are covered by a human gapped k-mer
- Threshold to decide if human or not

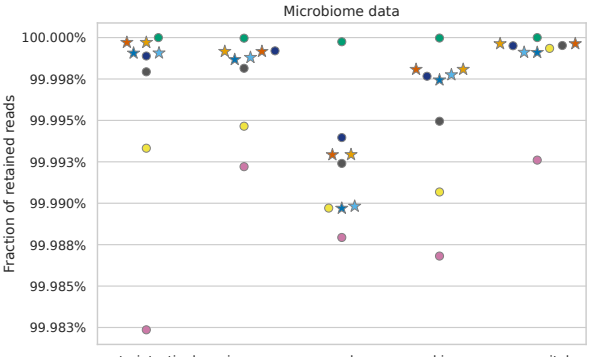
```
##_#_##
CGATCGACTAGCATCGAACGTACG . . .
CG T GA      AT G AC
   TC A TA      GA C TA
      AC A CA
         AG A CG
```

Evaluation

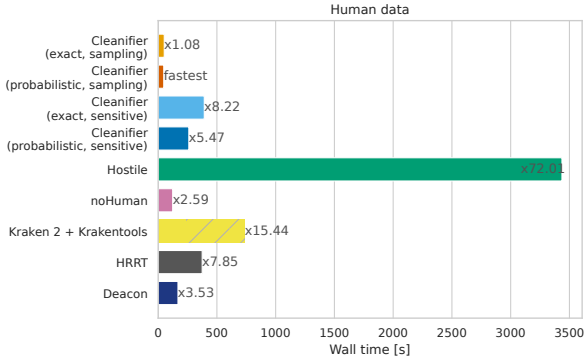

Human data
Fraction of removed reads
98.0%
97.0%
96.0%
95.0%
94.0%
93.0%
Cleanifier (exact, sampling)
Cleanifier (probabilistic, sampling)
Cleanifier (exact, sensitive)
Cleanifier (probabilistic, sensitive)
Hostile
noHuman
Kraken 2
HRRT
Deacon
HG002
HG003
HG004
HG005
HG006

Evaluation

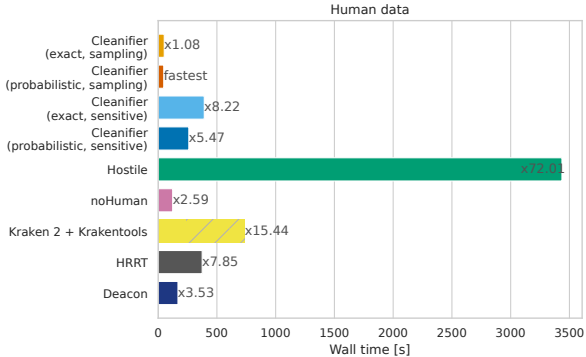

Human data
Fraction of removed reads
98.0%
97.0%
96.0%
95.0%
94.0%
93.0%
Cleanifier (exact, sampling)
Cleanifier (probabilistic, sampling)
Cleanifier (exact, sensitive)
Cleanifier (probabilistic, sensitive)
Hostile
noHuman
Kraken 2
HRRT
Deacon
HG002
HG003
HG004
HG005
HG006

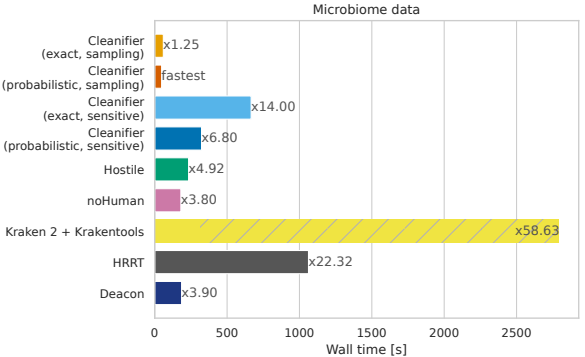

Microbiome data
Fraction of retained reads
100.000%
99.998%
99.995%
99.993%
99.990%
99.988%
99.985%
99.983%
gastrointestinal
airways
oral
skin
urogenital

Evaluation


Human data
Cleanifier (exact, sampling)
Cleanifier (probabilistic, sampling)
Cleanifier (exact, sensitive)
Cleanifier (probabilistic, sensitive)
Hostile
noHuman
Kraken 2 + Krakentools
HRRT
Deacon
x1.08
fastest
x8.22
x5.47
x72.01
x2.59
x15.44
x7.85
x3.53
0
500
1000
1500
2000
2500
3000
3500
Wall time [s]

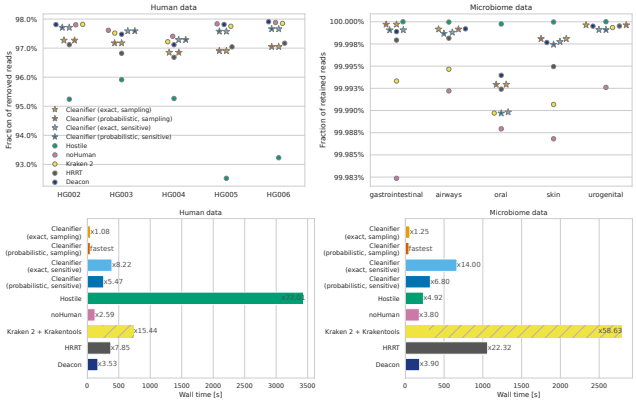
Evaluation


Human data
Cleanifier (exact, sampling)
Cleanifier (probabilistic, sampling)
Cleanifier (exact, sensitive)
Cleanifier (probabilistic, sensitive)
Hostile
noHuman
Kraken 2 + Krakentools
HRRT
Deacon
x1.08
fastest
x8.22
x5.47
x72.01
x2.59
x15.44
x7.85
x3.53
0
500
1000
1500
2000
2500
3000
3500
Wall time [s]


Microbiome data
Cleanifier (exact, sampling)
Cleanifier (probabilistic, sampling)
Cleanifier (exact, sensitive)
Cleanifier (probabilistic, sensitive)
Hostile
noHuman
Kraken 2 + Krakentools
HRRT
Deacon
x1.25
fastest
x14.00
x6.80
x4.92
x3.80
x58.63
x22.32
x3.90
0
500
1000
1500
2000
Wall time [s]

Cleanifier

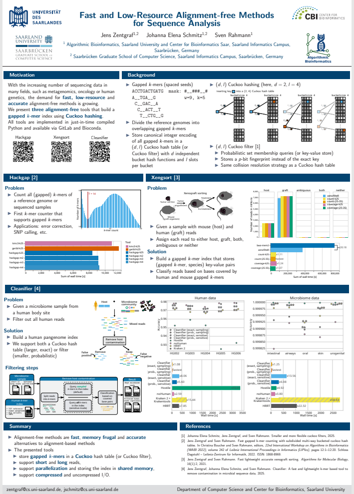
- High accuracy
- Low memory footprint (supports shared memory)
- Fast filtering
- Supports short and long reads



python Package Index

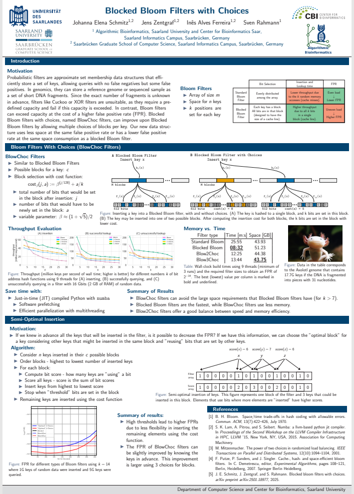
BIOCONDA
Reproducible research.
Install the software from bioconda:
> conda install -c bioconda cleanifier
See bioconda.github.io

P037



Johanna Elena Schmitz

P075



Inês Alves Ferreira

P079

Moein Karami