RECOMB 2021
PADOVA (ITALY) AUG 29 - SEP 1

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# SPRISS: Approximating Frequent k-mers by Sampling Reads, and Applications


***D. Santoro***, L. Pellegrina, and F. Vandin
University of Padova

**diego.santoro@phd.unipd.it** , {leonardo.pellegrina,fabio.vandin}@unipd.it

# Background

**Dataset $D$ of reads**

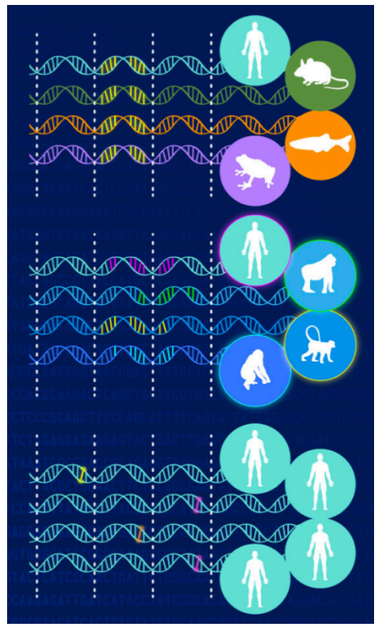
```
>seq1_RTW_read1
ATCGACGTACGATGCACGCATGACG
>seq1_RTW_read2
ACGATGCATTGCATCGACTCGAATG
>seq1_RTW_read3
TTGCTAGTGTACCTGATGCATTGCA
>seq1_RTW_read4
CGACTCGAATACGATGCATTGCATG
>seq1_RTW_read5
GTACCTGATTGCTAGTTGCATTGCA
…
```

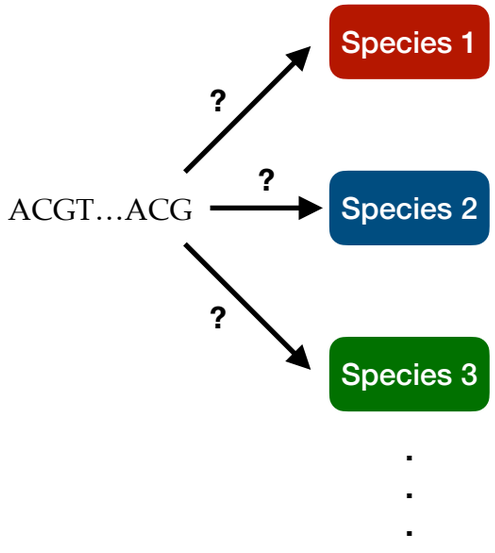
***k*-mer $P$**: substring of length $k$ of a read

**frequency $f_D(P)$ of $P$ on $D$** = fraction of $k$-mers of $D$ equal to $P$

The **study of k-mers and their frequencies** from datasets of reads is a **crucial step** for:

- Comparison of metagenomic datasets;
- Read classification in metagenomics;
- Genome comparison;
- Error correction for genome assembly;
- …





# Motivation

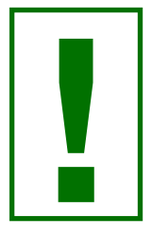
## Exact *k*-mer counters exist 🙂

**Jellyfish** (Marçais et al., 2011)
**BFCounter** (Melsted and Pritchard, 2011)
**DSK** (Rizk et al., 2013)
**KAnalyze** (Audano and Vannberg, 2014)
**Turtle** (Roy et al., 2014)
**KMC** (Kokot et al., 2017)
**Squeakr** (Pandey et al., 2017)
…

⟶ **counting all *k*-mers is computationally expensive on massive modern datasets** 😕

For some applications:

- Comparisons of metagenomic datasets
- Discovery of discriminative k-mers



just **frequent *k*-mers are of interest**

# Motivation

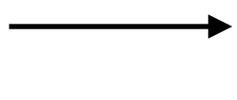
**Frequent $k$-mer $P$**: given $\theta \in (0,1]$, $P$ appears in $D$ with frequency $f_D(P) \geq \theta$

**Def.**

**Frequent k-mer counting problem**

Given $\theta$, extract frequent k-mers (and their frequencies) $FK(D, k, \theta)$ from $D$

**computationally expensive**
**on massive modern datasets** 😕



**approximation methods**

# State of the art

**Frequent k-mers approximations with theoretical guarantees: relatively unexplored**

***SAKEIMA*** [Pellegrina, Pizzi, Vandin. RECOMB 2019 - JCB 2020]

- Approximations of frequent $k$-mers with guarantees by ***sampling k-mers***
- Require to scan the entire dataset $D$
- Needs to be reimplemented for more efficient exact $k$-mer counters
  (It is built on *Jellyfish*)


**This work, *SPRISS***

- Approximations of frequent $k$-mers with guarantees by ***sampling reads***
- Require to scan just a sample of reads of $D$
- No need to reimplement it for more efficient exact $k$-mer counters

# Computational problem

**Computational problem**

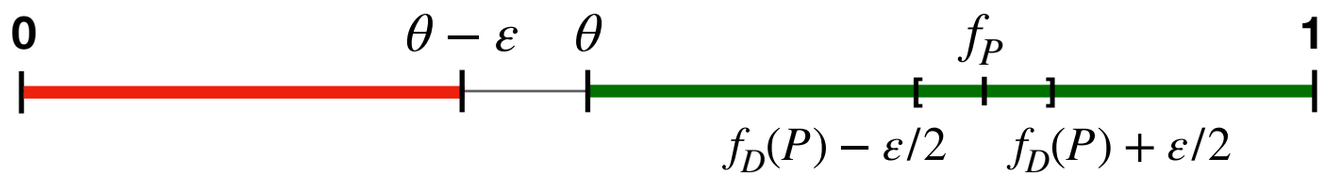
*Input:* $D, k, \theta, \varepsilon, \delta$

*Output:* **$\varepsilon$-approximation** of frequent $k$-mers $FK(D, k, \theta)$ with probability $\geq 1 - \delta$

**Def.**

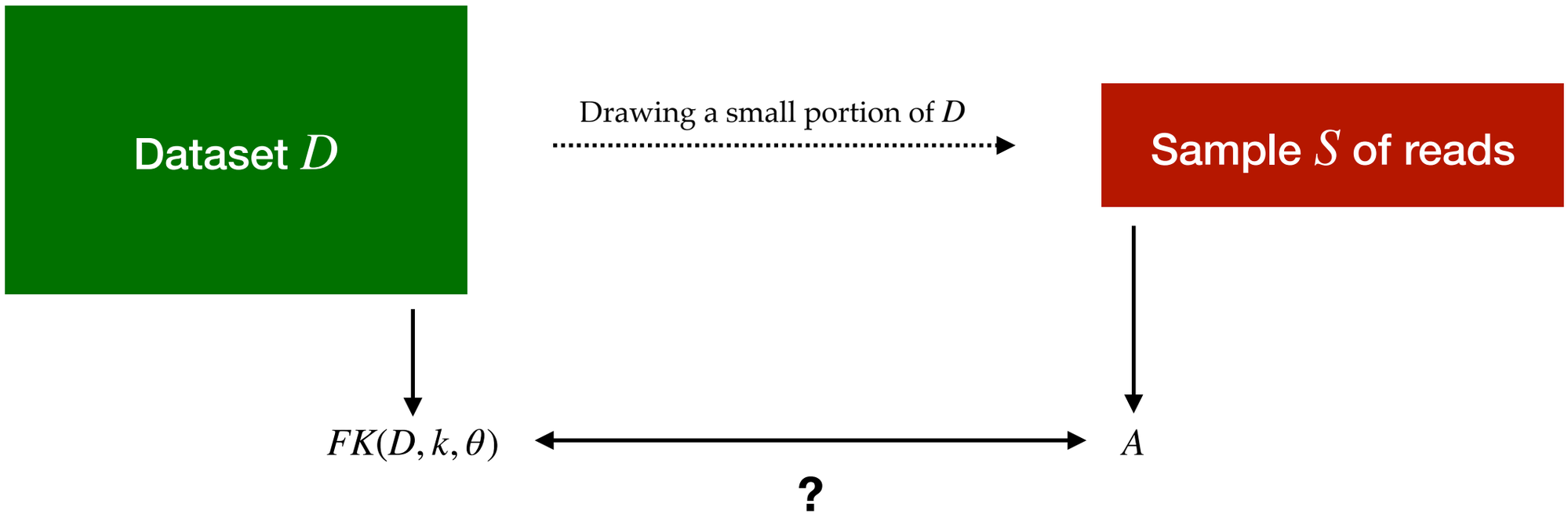
Given $\varepsilon \in (0,\theta)$, a set $A = \{(P, f_P)\}$ is an **$\varepsilon$-approximation** of $FK(D, k, \theta)$ if:

1. $A$ contains **no false negatives**
2. $A$ does not contain $k$-mers s.t. $f_D(P) < \theta - \varepsilon$
3. All $k$-mers in $A$ are s.t. $|f_D(P) - f_P| \leq \varepsilon/2$



# SPRISS: main idea

**Approximation** $A$ of frequent $k$-mers $FK(D, k, \theta)$ by analyzing a **sample** $S$ of $D$




**Challenges:**

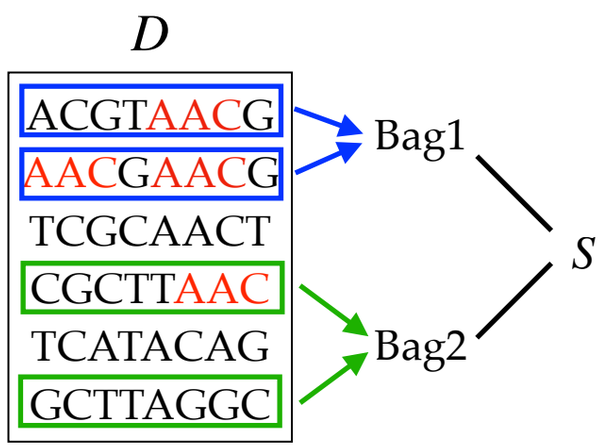
**1.** Find a rigorous relation between $A$ and $FK(D, k, \theta)$.

**2**. Identify a sample size which is sufficient to guarantee good estimates from $S$.

**3**. Reads introduce dependencies among $k$-mers

# SPRISS

**Sampling strategy**: sample $S$ is a collection of $m$ ***bags of*** $\ell$ ***reads*** sampled independently and uniformly at random, with replacement, from $D$




$P = \text{AAC}$

$f_S(P) \propto$ #occurrences of $P$ in $S$ (i.e., 4)

$\hat{f}_S(P) \propto$ #bags where $P$ appears (i.e., 2)

**Main steps:**

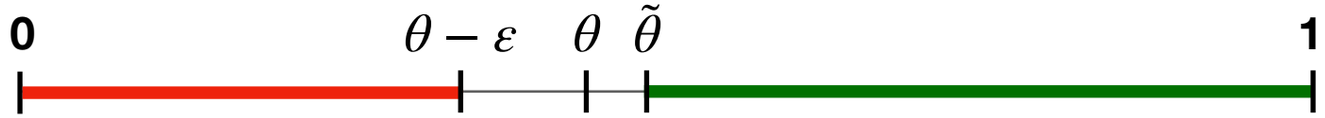
1. **Compute sample** $S$
2. **Compute** $f_S(P)$ **and** $\hat{f}_S(P)$ **using any exact k-mer counter**
3. **Output** $A = \{(P, f_S(P)) : \hat{f}_S(P)) \geq \theta - \varepsilon/2\}$

# Main contribution

**Thm**

**If** $m \geq \frac{2}{\varepsilon^2}\left(\frac{1}{\ell\ell_{\mathcal{D},k}}\right)^2\left(\lfloor \log_2 \min(2\ell\ell_{\max,\mathcal{D},k}, \sigma^k)\rfloor + \ln\left(\frac{1}{\delta}\right)\right)$ **then**

$A = \{(P, f_S(P)) : \hat{f}_S(P)) \geq \theta - \varepsilon/2\}$ **is *almost* an $\varepsilon$-approximation of $FK(D, k, \theta)$, with probability $\geq 1 - \delta$**



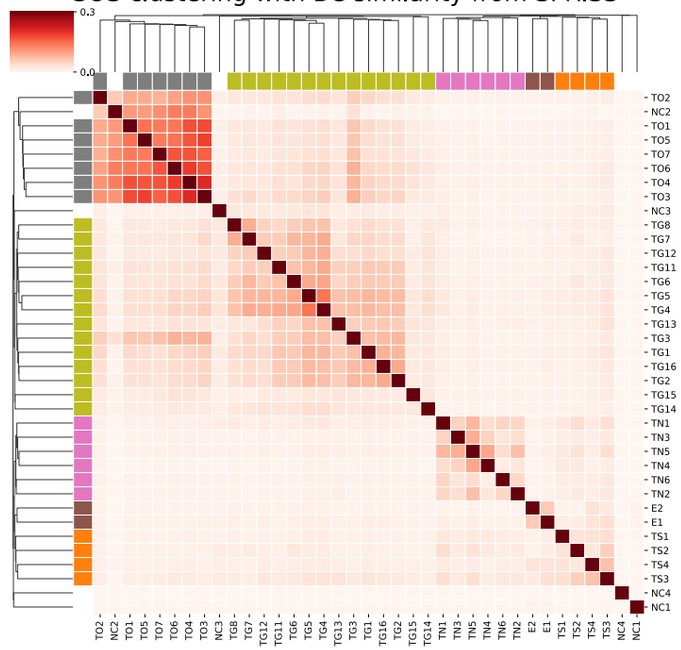

**Proof:** based on the ***pseudodimension,*** a key tool from statistical learning theory, of k-mers

# Experimental results

**Implementation**: C++ (based on *KMC* exact counter)
**Machine**: 512 GB of RAM and 2 Intel(R) Xeon(R) CPU E5-2698 v3 @2.3GHz

**Experimental results:**

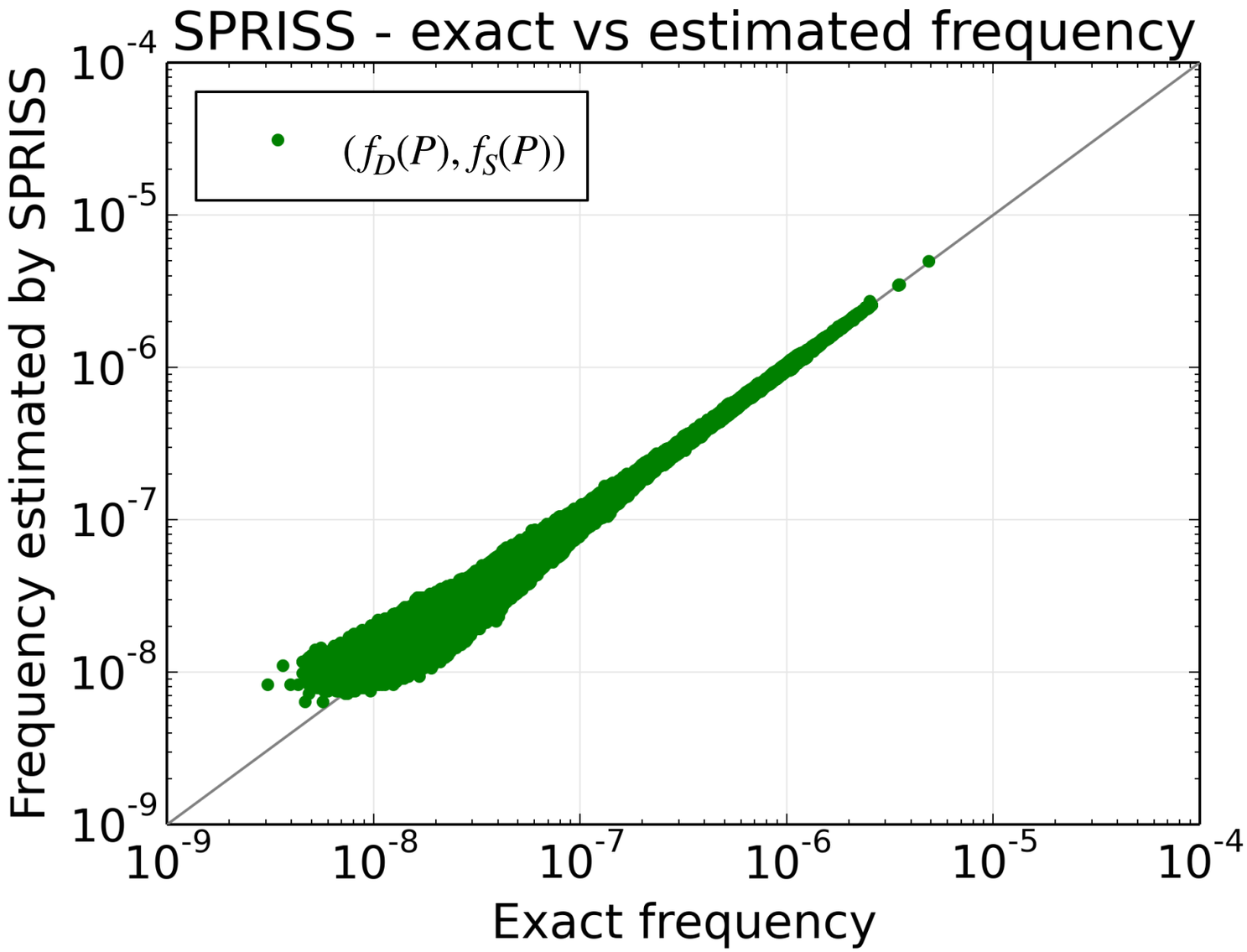
1. **Accuracy** of the estimates
2. **Resources**
3. **Comparing metagenomic datasets**
4. **Discriminative k-mers approximations**



# Accuracy

6 large datasets from Human Microbiome Project (HMP) - $\approx 10^8$ reads

$k = 31$

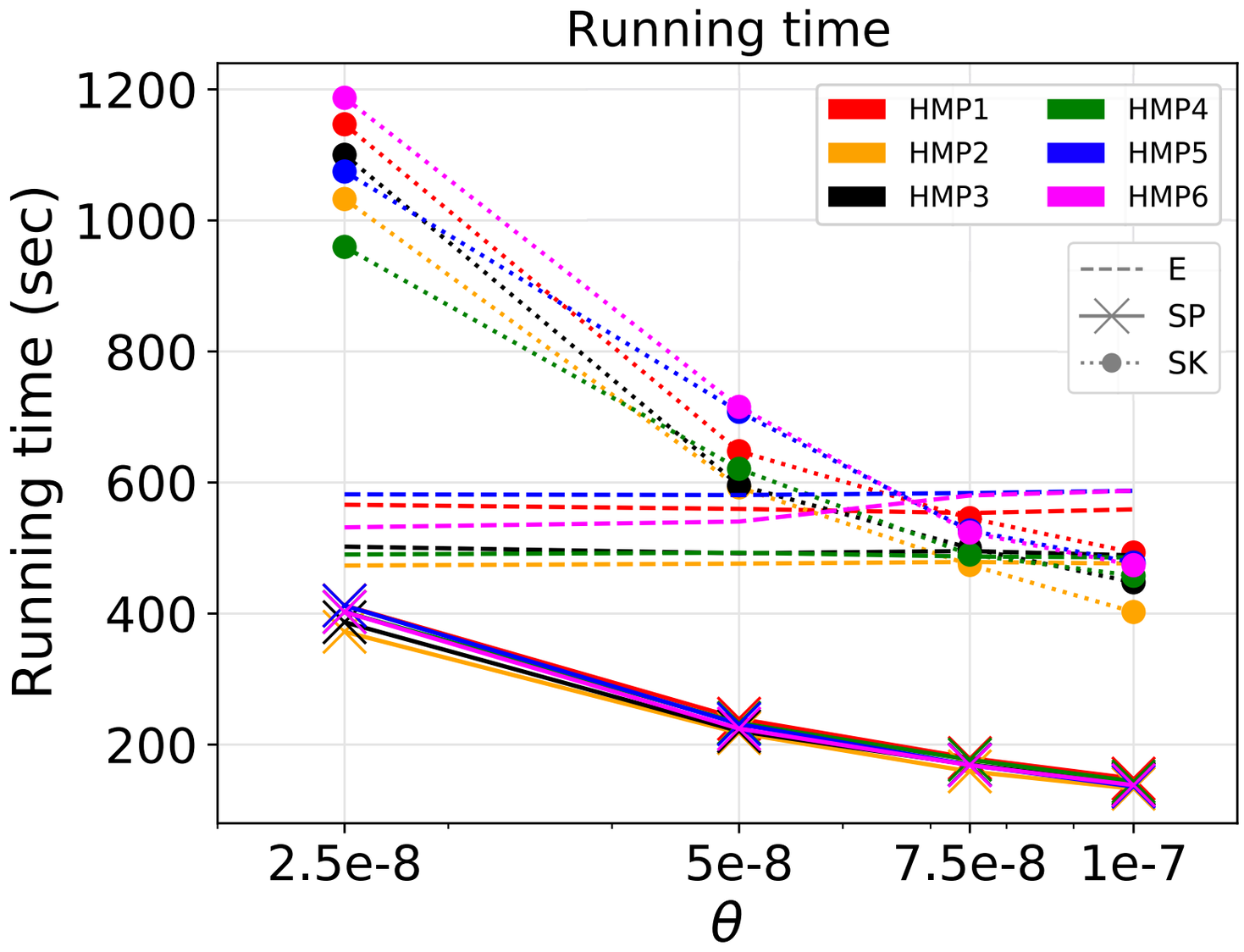



$\theta = 2.5 \cdot 10^{-8}$

**30% of $D$ analyzed**

# Resources

**SP** = *SPRISS* **SK** = *SAKEIMA* (built on *Jellyfish*) **E** = exact approach (*KMC*)

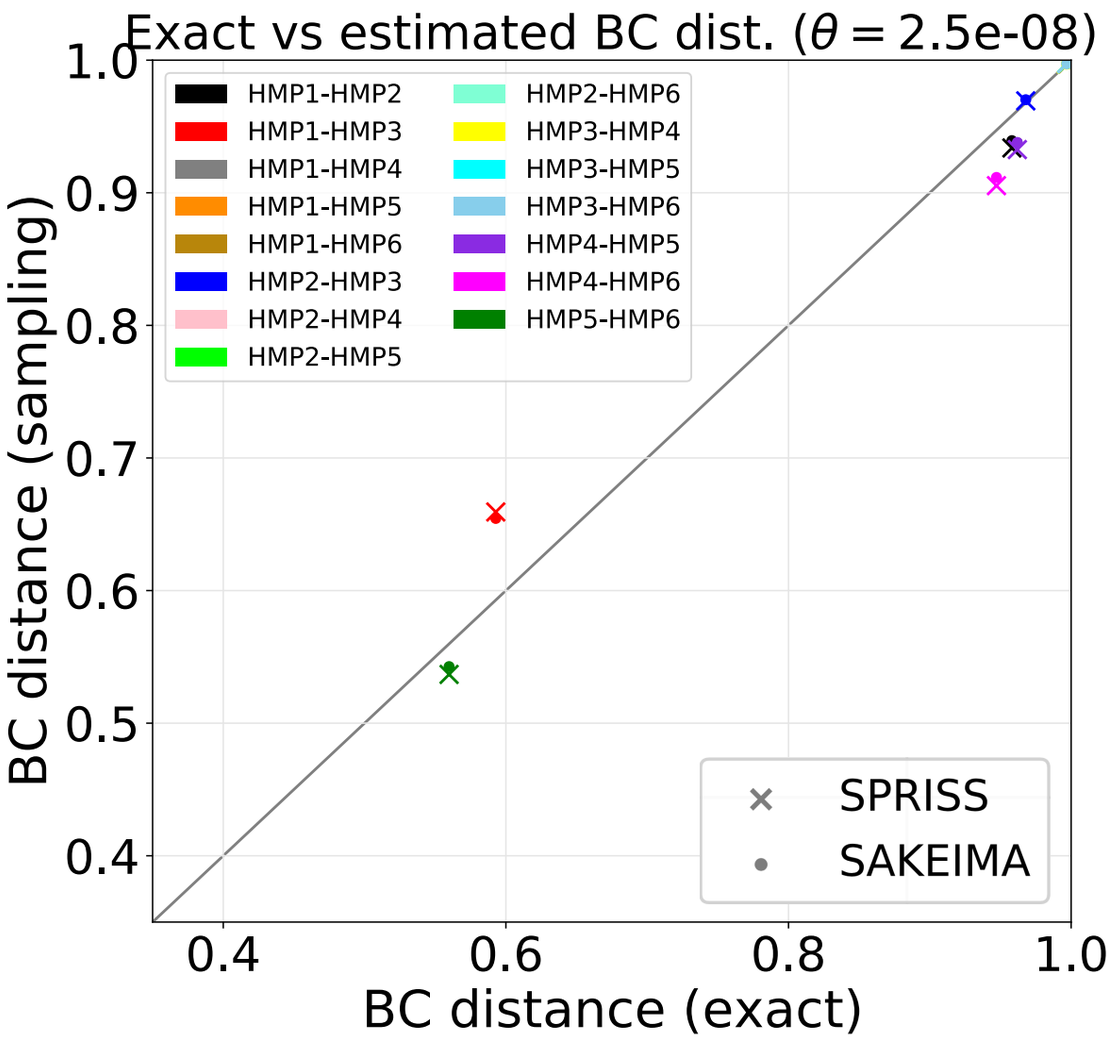


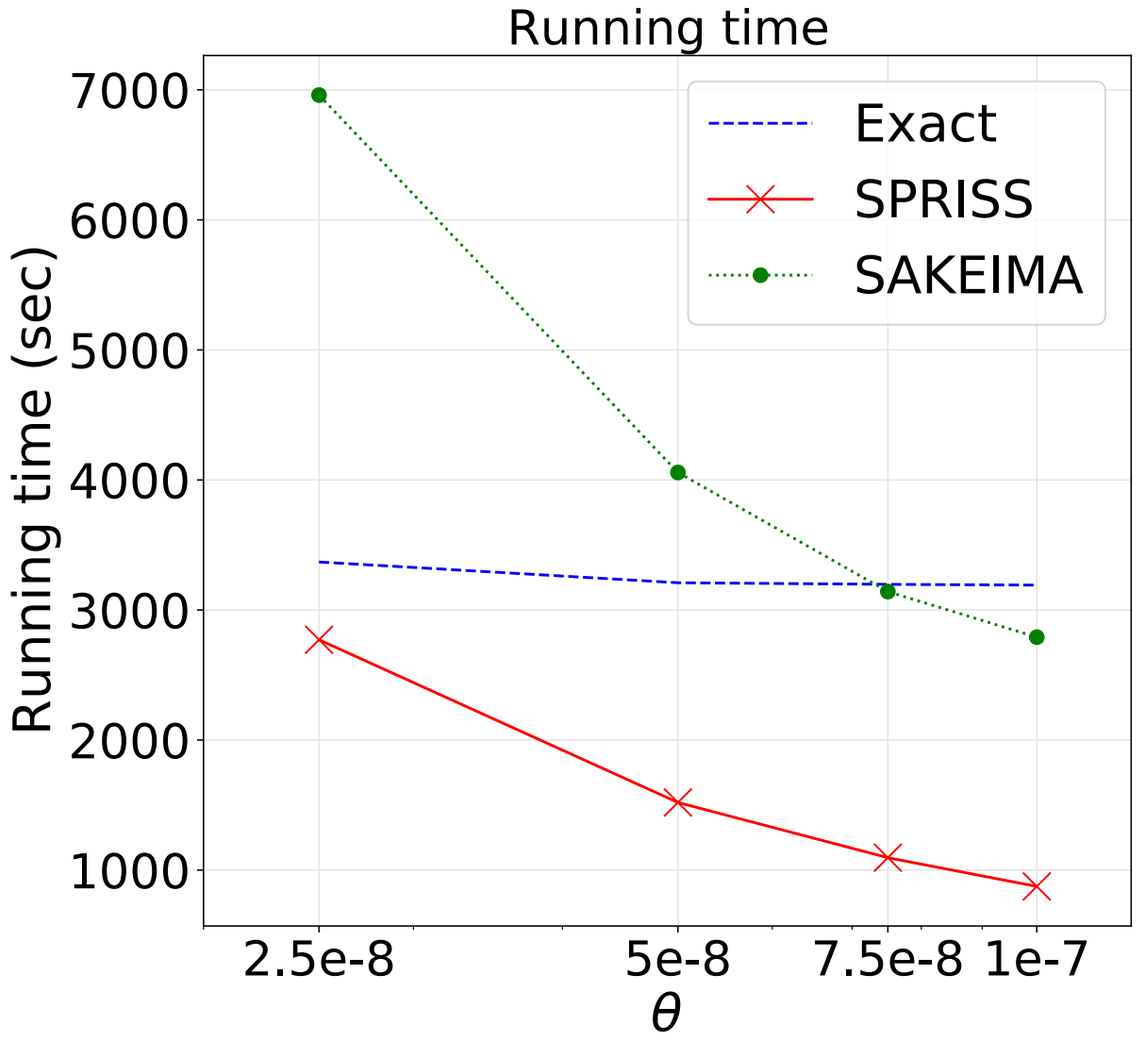

**SPRISS analyzes at most 34% of each dataset $D$**

# Comparing metagenomic datasets

**Bray-Curtis (BC) distance** of $D_1$ and $D_2$ as a function of frequent $k$-mers

**Estimation of BC by using approximations of** $FK(D_1, k, \theta)$ **and** $FK(D_2, k, \theta)$

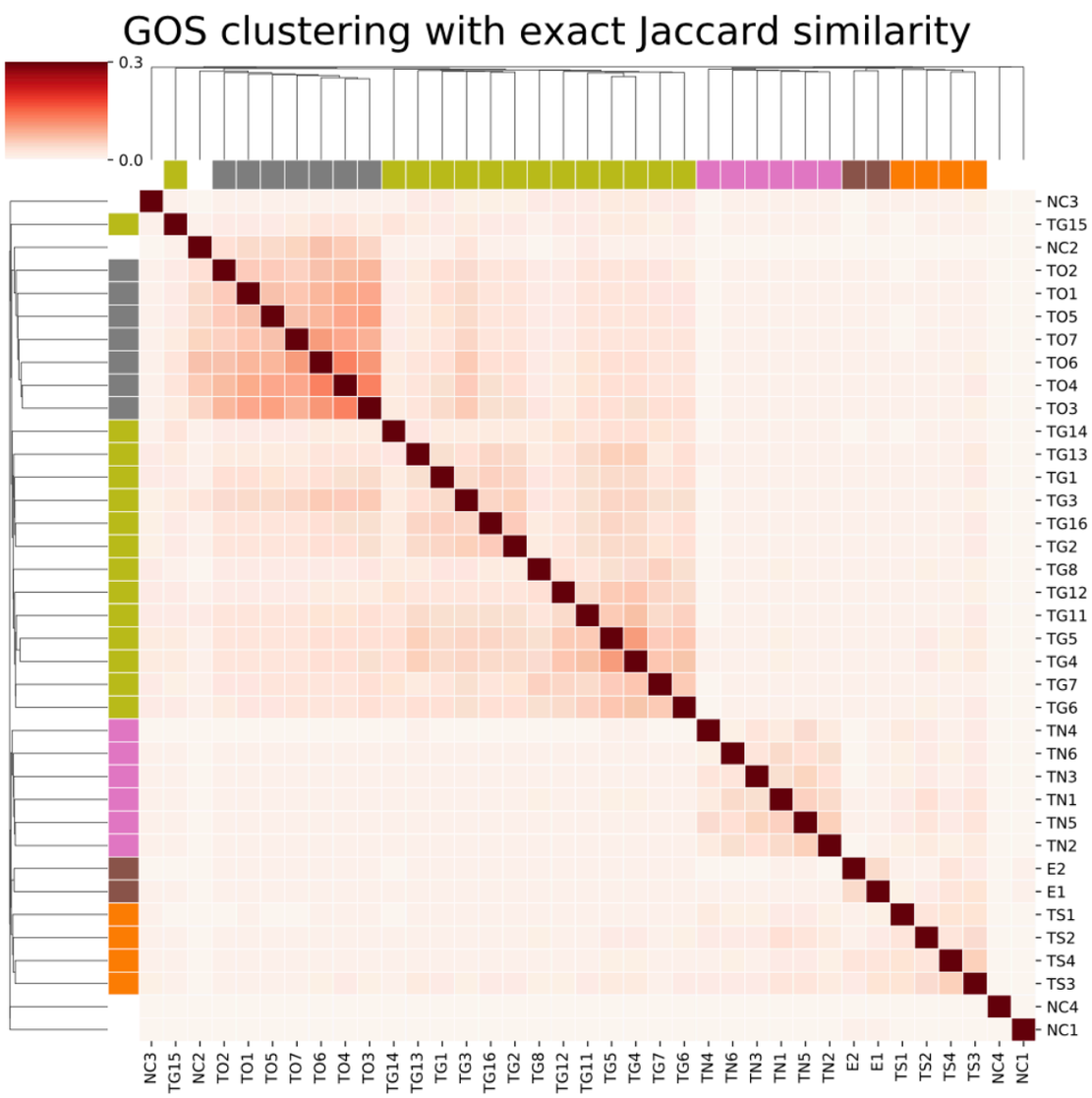


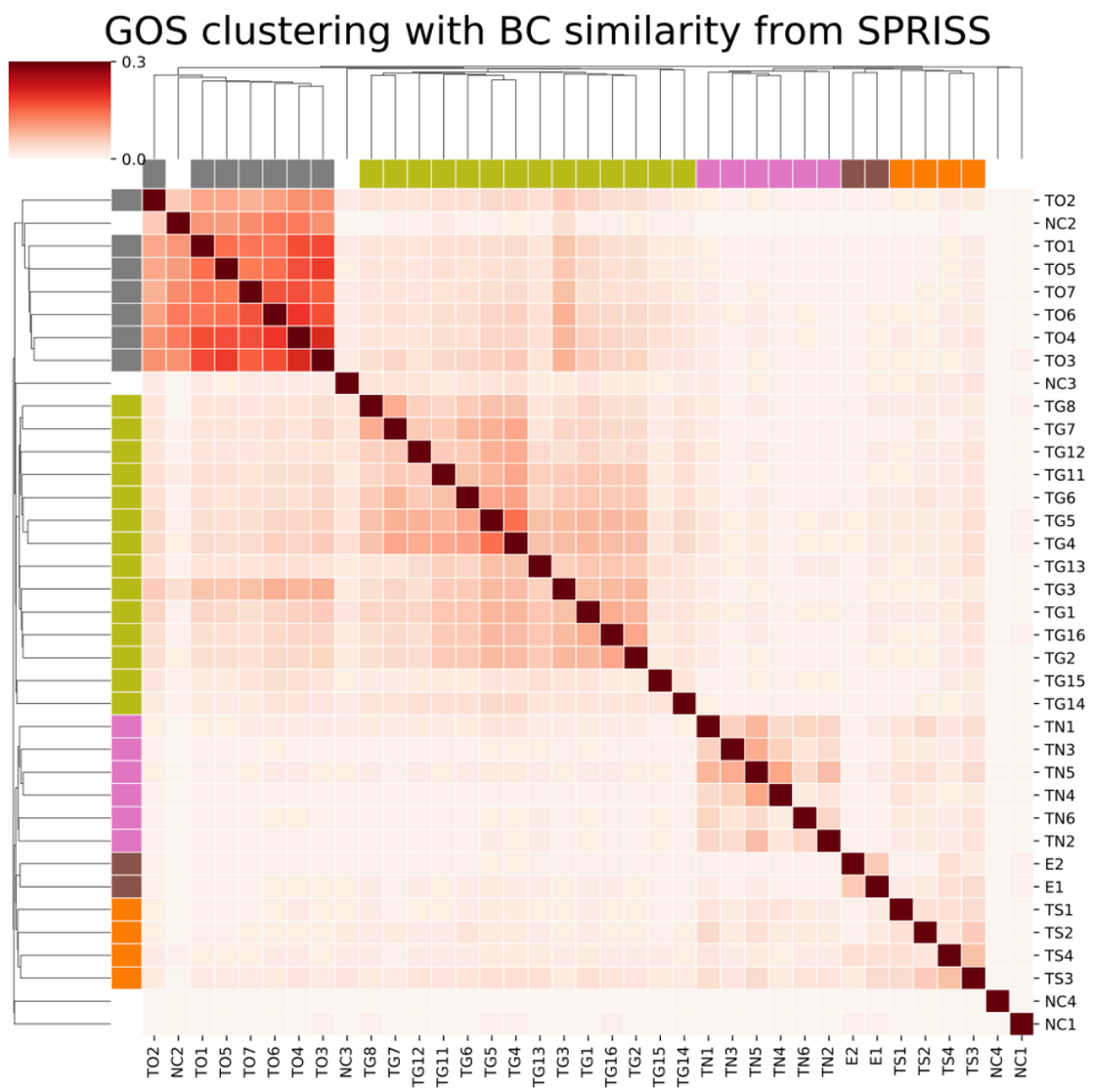



# Comparing metagenomic datasets

37 datasets from Global Ocean Sampling (GOS) Expedition - $\approx 10^5$ reads

$k = 21$




**Inside-vs-outside cluster signal increases of 50% using BC estimates**

**SPRISS requires 40% of the time of exact BC approach**

# Discriminative $k$-mers

**Def.**

Given two datasets $D_1$ and $D_2$, the $D_1$-**discriminative** $k$-**mers** are s.t.:

1. $P \in FK(D_1, k, \theta)$, and
2. $f_{D_1}(P) \geq \rho f_{D_2}(P), \rho > 1$

$D_1$

ACGTAACG
AACGAACG
CGCTTAAC
TCATACAG
GCTTAGGC
TCGCAACT

$P = \text{AAC}$

$D_2$

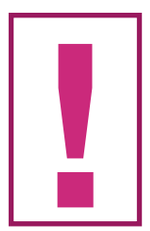
TTGACATG
GTCGTACG
ATCCCGAG
CTGATCAA
CCATACTC
ACTCATCG

**Estimation of discriminative $k$-mers by using SPRISS's approximations**

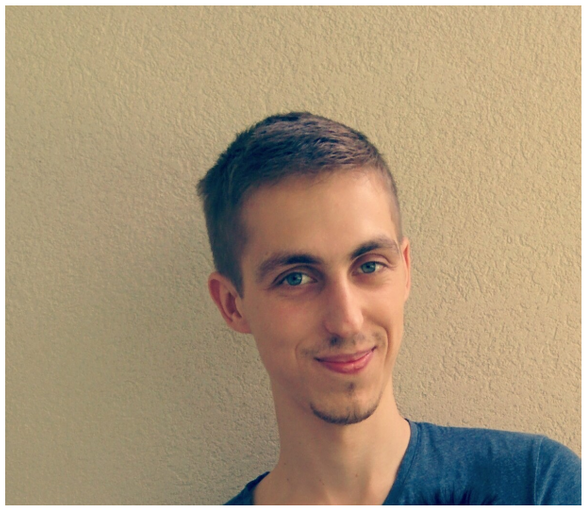
2 large datasets from (Liu et al., 2017) - $\approx 4 \times 10^8$ reads, $k = 31, \rho = 2$

Using just **5% of reads** of $D_1$ and $D_2$

- false negative rate is **< 0.03**
- Running time gain: **90%**



## Acknowledgements:



*Leonardo Pellegrina*


*Fabio Vandin*

## Fundings:

# Thanks for your attention! 🙂

**SPRISS available at:**

https://github.com/VandinLab/SPRISS

https://arxiv.org/abs/2101.07117

**Recipe of SPRISS:**

3 ounces of Prosecco
2 ounces of Aperol
1 ounce of Club Soda
Garnish: orange slice

https://www.liquor.com/recipes/aperol-spritz/


**Diego Santoro**

diego.santoro@phd.unipd.it

University of Padova