Recent advances in **data structures** for storing **sets of k-mer sets**

Rayan Chikhi
Institut Pasteur

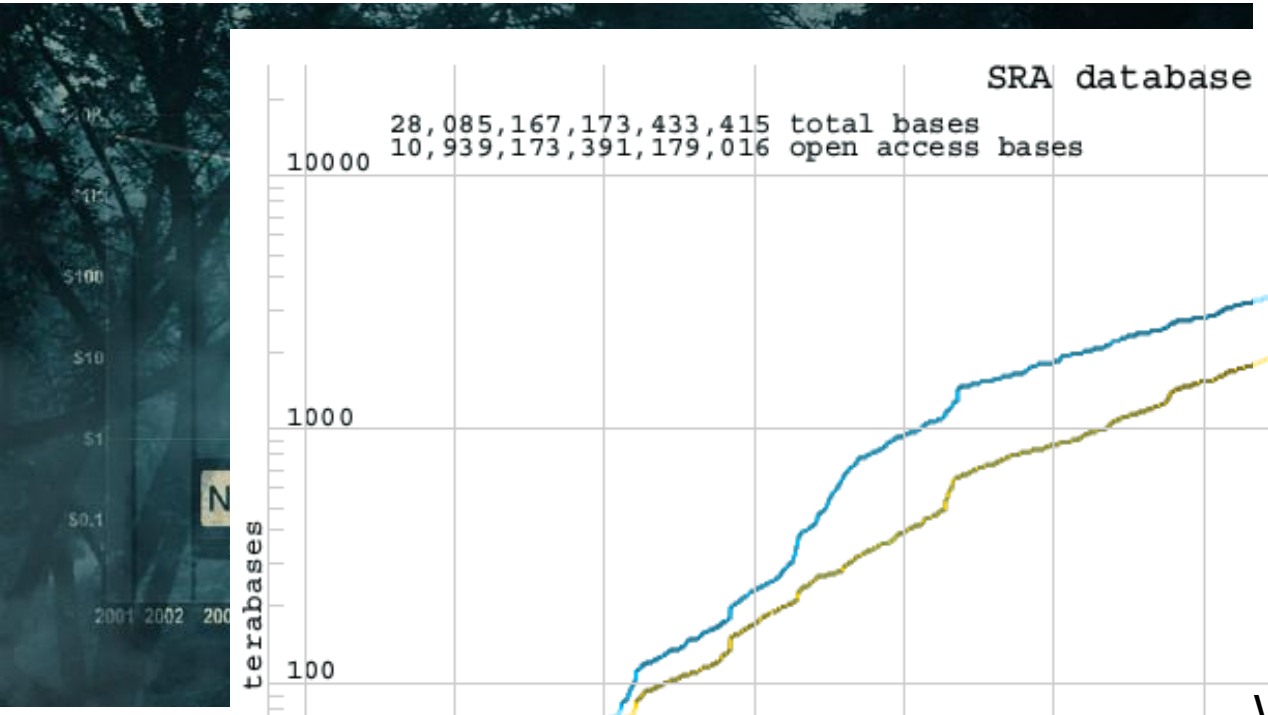
CGSI 2019

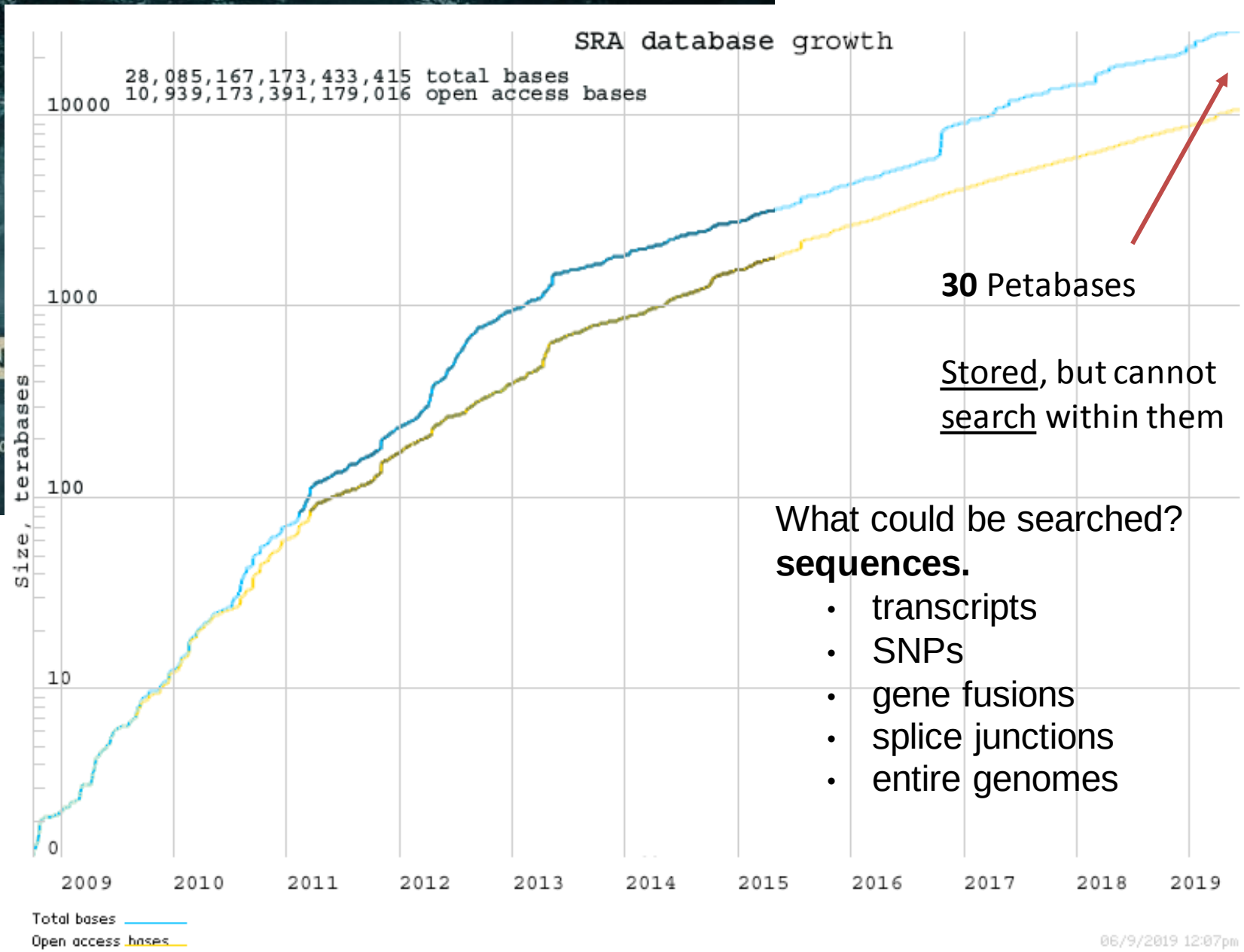
Searching huge databases of sequences

Rayan Chikhi
Institut Pasteur

CGSI 2019

https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/





30 Petabases

<u>Stored</u>, but cannot <u>search</u> within them

What could be searched?

sequences.

- transcripts
- SNPs
- gene fusions
- splice junctions
- entire genomes

YouTube: 100-1000 PB

NCBI SRA database : 30 PB

Institut Pasteur: 8 PB

Your laptop: 0.001 PB

SRA

SRA ▾ Search

Advanced

Help

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Search results

Items: 1 to 20 of 19964

NextSeq 500 paire
1. 1 ILLUMINA (Illumina
Accession: ERX34307

NextSeq 500 paire
2. 1 ILLUMINA (Illumina
Accession: ERX34307

NextSeq 500 paire
3. 1 ILLUMINA (Illumina
Accession: ERX34307

NextSeq 500 paired end sequencing (ERR3407135)

Metadata | Analysis (alpha) | Reads | Download

Filter: Find Filtered Download What does it do?

What can the filter be applied to?

< 1 1 346553 >

View: ☑ biological reads ☐ technical reads

1. ERR3407135.1 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:5421:
member: default

2. ERR3407135.2 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:22482
member: default

3. ERR3407135.3 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:25663
member: default

4. ERR3407135.4 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:21199
member: default

5. ERR3407135.5 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:23504
member: default

Reads (separated)

>gnl|SRA|ERR3407135.1.1 NB551234:144:HL523AFXY:1:11101:5421:1076 *F (Biological)*
ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGGCGCGGAATTTGGGATGTTCCATCAGT
TTCCAGGCGCGTTTGCCCTGACGTCGCGACATGCGTAACTGAAGCTGCCAAATATCACGG
GTAAGCGTGGTAAGGCGTTTCGGGATCGCCA

>gnl|SRA|ERR3407135.1.2 NB551234:144:HL523AFXY:1:11101:5421:1076 *R (Biological)*
ATCAACAACAGCGGGAATACCACCTCTTCCAGCCGTTGTTTCCAACCAATACGCGTTAAT
TCACCGAAACCGCGACAGCGCAATGGAACGCATCATTGCGCAGGTGTTGCAGAATACGGA
AAACCGCATCCGAAACGAGATGCGCGTTAAT

U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

BLAST® » blastn suite

Sequence Read Archive Nucleotide BLAST

blastn

BLASTN programs search SRA databases using a nucleotide query.

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

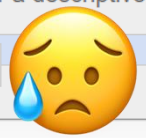
To

Or, upload file Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set



SRA Experiment set (SRX)

Enter an SRA accession (experiment, study, or submission), title, the scientific name or tax id. Only 20 top suggestions will be shown.

What could be gained by large-scale sequence searches?

- **conditions** that express a given **novel isoform**
- **gene fusions** across TCGA
- **metagenomic samples** containing bacterial **strain**
- **Anti-microbial resistance:**
  - detection of drug-resistant genes
  - phylogeny of plasmids carrying AMR genes

https://m.photofunia.com/

Sources: BIGSI article, SBT article, SeqOthello article

Problem definition

- Given many FASTQ files (=*experiments*):
  - Experiment 1
  - …
  - Experiment 100,000
- And a sequence *s*
- **Enumerate all the experiment(s) where *s* appears**

e.g. "ACAGTATGGTTGGGGAAAAG" -> Experiments {23 (human RNAseq), 1523 (human RNAseq), 82499 (human gut metagenome)}

Searching within sequences, techniques

Solution 0: **grep**

Solution 1: build a *huge* dictionary

Solution 2: FM-index

Solution 3 and others: settle for **k-mer searches**

Let’s examine all these solutions..

Solution 0: grep

How fast is grep?

- SRA has .fastq.gz files
- gunzip: 60 MB/sec
- 

pugz: ~400 MB/sec [Kerbiriou, Chikhi 2019]

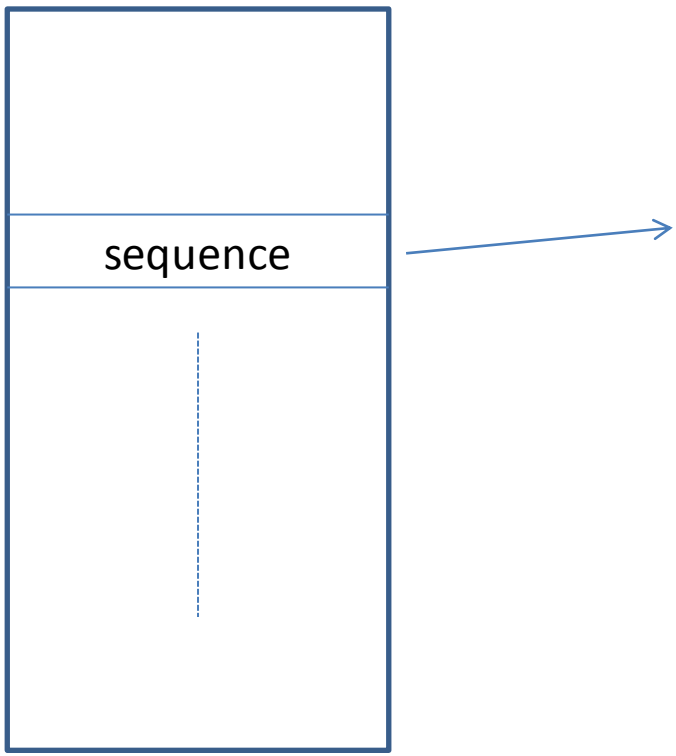
Time:
- 60 GB Illumina human FASTQ: 16 minutes, 1 thread
- 1,000 of them: ~1 day, 10 threads
- SRA (5M files): ~**1 year,** 100 threads

Let's examine all these solutions..

Solution 1: build a *huge* dictionary

But for n experiments of r reads of length l, requires $O(n*r*l^2)$ memory

n=100,000
r=1,000,000,000
l=100



List of experiment(s) that sequence appears in

Let’s examine all these solutions..

Solution 1: build a *huge* dictionary

What if we indexed only small sequences of **fixed length**?

How many in, say, all RefSeq proteobacteria (*E. coli & al*)?

21,883 genomes, 97 GB of FASTAs -> 31 billion sequences [Kerbiriou, personal comm, k=31]

Lower bound of ~ 85 GB to represent [Conway, Bromage 2012; Chikhi et al, 2014]
(as opposed to 31 billion times 31 nucleotides)

That’s only for a small subset of *genomes*, not even FASTQs..

Let’s examine all these solutions..

Solution 2: FM-index

(2017, Genome Research)

Method

Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes

Dirk D. Dolle,[1,6] Zhicheng Liu,[1,2,6] Matthew Cotten,[1] Jared T. Simpson,[3,4] Zamin Iqbal,[5] Richard Durbin,[1] Shane A. McCarthy,[1] and Thomas M. Keane[1,2]

- **Built a BWT** of 2,705 human Illumina WGS error-corrected reads
- Size: ~500 GB + ~5 TB metadata (origin of reads)
- 1 k-mer search ~= 10 ms

k-mer search

- k-mer

Sequence of fixed length k

```
ref:     ACGATGACATGAT
4-mers:  ACGA
          CGAT
           GATG
            ATGA
              …
```

- Membership query

-> *Is x in S?*

- Enables more complex queries

arbitrary sequence *s* <-> decomposition of *s* into k-mers

e.g. SNP <-> set of ~k k-mers

K-mer membership within **1** experiment

solved

An array (Bloom Filter) records the presence of sequences

AAAG
AAGG
AGGC ...
GGCG
GCGT
CGTT
GTTG
TTGA
TGAG
GAGG
AGGT
GGTT

0
0
0
0
0
0
0
0
0
1
0
0
0
0
0
1
0
0
0
0
0

Bloom Filter

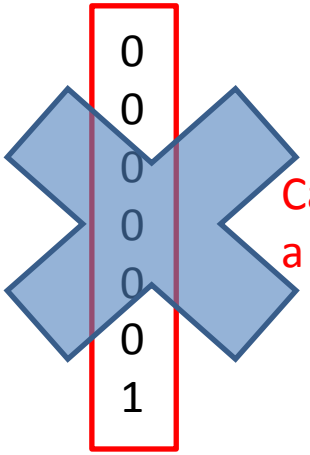
a position ≃ a sequence

« AAAG »?

Recent review on arXiv:
[Chikhi, Medvedev, Holub'19]
Also: CGSI talk 2018

K-mer membership for **1,000,000+** experiments

Open question



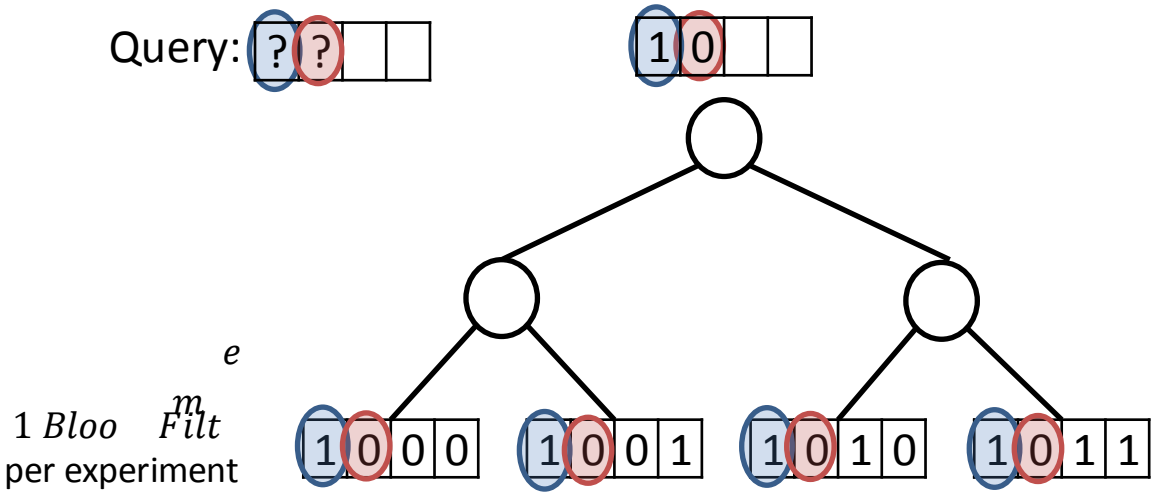
Can't just use
a Bloom Filter

Why:

- *Single* SRA Bloom Filter : ~ petabytes, and pooled
- *1 BF per experiment* : ~ hour-long query (150ms * 5M exps)

Sequence Bloom Trees

[Solomon & Kingsford 2016]

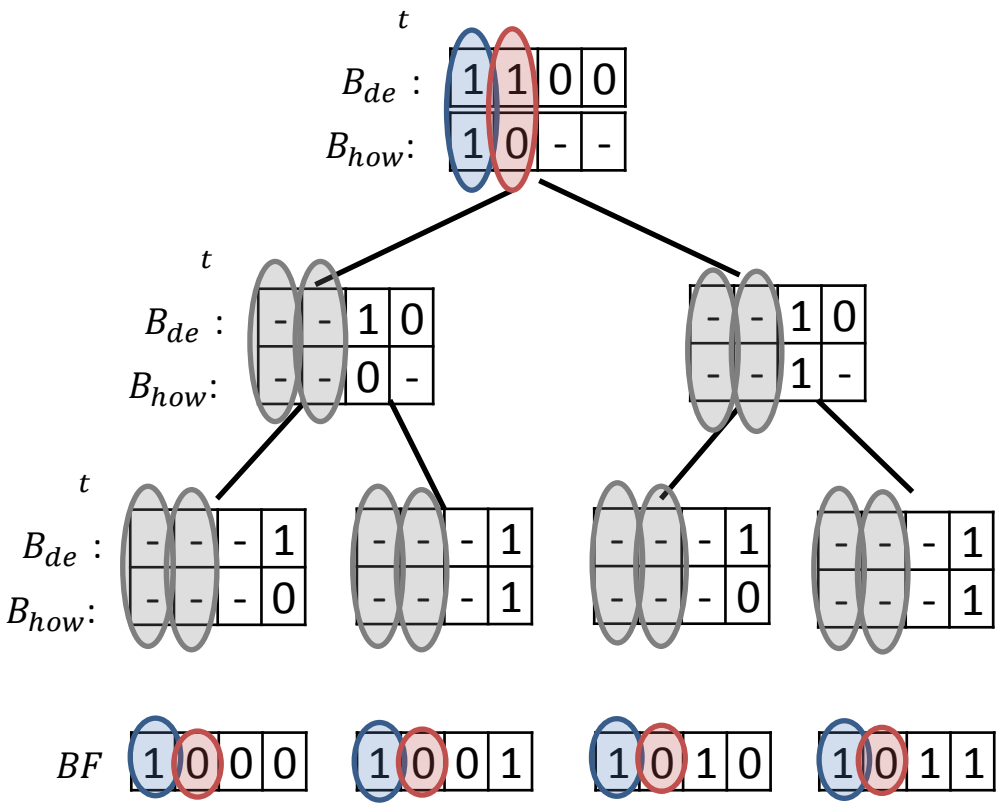


Slide: P. Medvedev

<u>Fast search through pruning</u>

- blue seq matches ALL experiments
- query can stop at root

- AllSome Sequence Bloom Trees
  [Sun, Harris, Chikhi, Medvedev 2017]
- **HowDeSBT**
  [Harris, Medvedev 2019]



Slide: P. Medvedev

SBT performance in a nutshell

Data	Size on disk
2,000 raw experiments	~15,000 GB
AllSome SBT (2017)	142 GB
HowDe SBT (2019)	14 GB

- Smaller space than raw data
- Resides fully on disk
- 1 search = 5 seconds

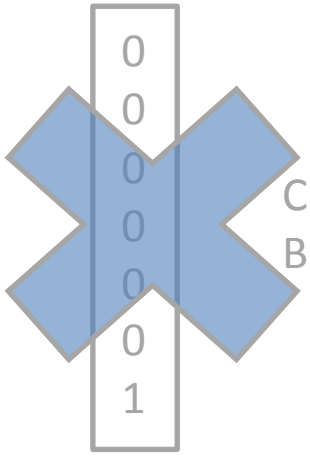
BIGSI

[Bradley, .., Iqbal 2019]

Recall, a few slides ago…

Search within **1,000,000+** experiments

Open question



Can't just use Bloom Filters

- *1 BF per experiment*: very long query time

Actually BIGSI did just this

BIGSI is a **vector** of **Bloom filters**
[Bradley et al, 2019]

(or equivalently, a matrix of bits)

- Rows = k-mers
  Columns = experiments

- All the BFs have the <u>same size</u>

How is this possible? *Microbial data*

- 1 search = 0.3 second

Also the technique behind bing

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	1	1	0	0	1
0	0	0	0	1	0
0	0	0	1	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	1	1	1	1	1
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	1	1	1	1	1
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Many other methods

- SBT: Solomon and Kingsford 2016 *Nature biotechno.*
- AllSomeSBT: Sun et al. 2018 *RECOMB*
- SSBT: Solomon and Kingsford 2018 *RECOMB*
- HowDeSBT: Harris and Medvedev 2019 *RECOMB-Seq*
- BIGSI: Bradley et al. 2019 *Nature biotechno.*
- COBS: Bingman et al. 2019 *arXiv*
- Cortex: Iqbal et al. 2012 *Nature*
- BFT: Holley et al. 2016 *AMB*
- VARI: Muggli et al. 2017 *Bioinformatics*
- VARI-Merge: 2019 *accepted to ISMB*
- Rainbowfish: Almodaresi et al. 2017 *WABI*
- Mantis: Pandey et al. 2018 *Cell*
- Mantis+MST: Almodaresi et al. 2019 *RECOMB*
- SeqOthello: Yu et al. 2018 *Genome Biology*
- Metannot: Mustafa et al. 2018
- Multi-BRWT: Karasikov et al. 2018

Slide: C. Marchet

Review: [Marchet *et al*, in preparation '19]

Short overview in [Chikhi, Medvedev, Holub'19]

Conclusion

- **Sets of k-mer sets**, a powerful representation for sequencing data
- Booming area since 2016
- No method is *really* user-friendly yet

- Many similar experiments: SBT *et al*
- Many experiments of same size: BIGSI *et al*
- *And many others I didn't talk about*
- Versatile method: [???]

A Tera increase in sequencing production in the past 25 years		
Genes & Operons	1990	**Kilo** = 1,000
Bacterial genomes	1995	**Mega** = 1,000,000
Human genome	2000	**Giga** = 1,000,000,000
Human microbiome	2005	**Tera** = 1,000,000,000,000
50K Microbiomes	2015	**Peta** = 1,000,000,000,000,000
what is expected for the next 15 years ? (a Giga?)		
200K Microbiomes	2020	**Exa** = 1,000,000,000,000,000,000
1M Microbiomes	2025	**Zetta** = 1,000,000,000,000,000,000,000
Earth Microbiome	2030	**Yotta** = 1,000,000,000,000,000,000,000,000

Source: @kyrpides

Thank you for your attention!
Any questions?

Acknowledgements: Camille Marchet, Paul Medvedev, Mael Kerbiriou, Mikael Salson, Bob Harris, Chen Sun, Jan Holub, Simon Puglisi, Daniel Gautheret, Antoine Limasset, Pierre Peterlongo, and the wonderful bioinformatics data structure community

@RayanChikhi