Comparing and indexing metagenomes at a large scale using random projections

Roland Faure[1], Hasin Abrar[2], Haonan Wu[2], Stephanie Won[2], Adrita Hossain Nakshi[2], Rayan Chikhi[1], David Koslicki[2], Paul Medvedev[2]

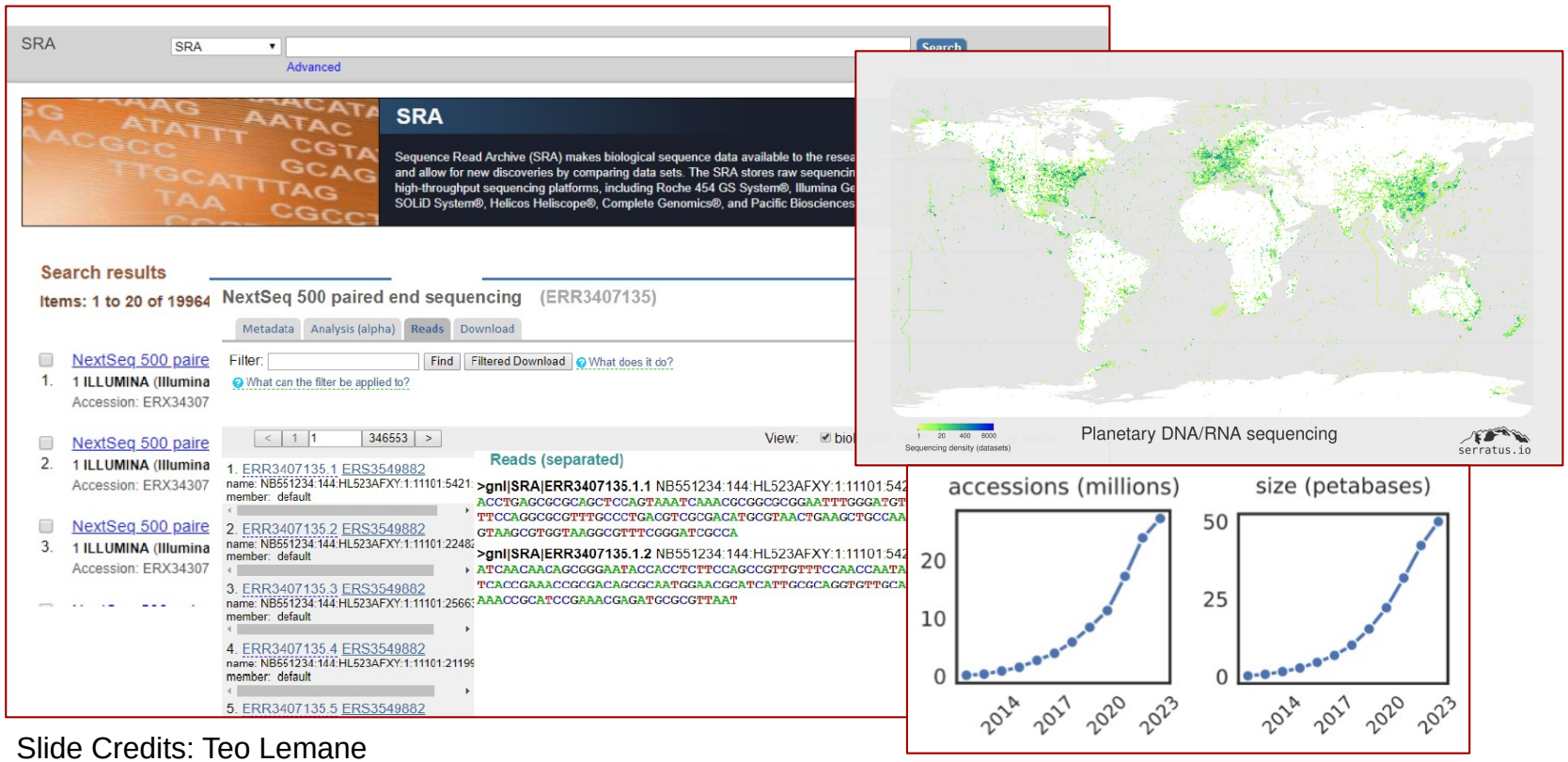
[1]Institut Pasteur

[2]The Pennsylvania State University

SeqBIM 2025

The SRA database

SRA: All public sequencing reads, **50 PBases** (as of Dec 2023)



Slide Credits: Teo Lemane

The Logan project

27 millions accessions, **50PB**

For all accessions

Raw reads → Contigs/Unitigs

```
ACAGCAGT
CAGTCGTA
CGTATAA
```

→

```
ACAGCAGT
    CAGTCGTA
        CGTATAA
```

→ `ACAGCAGTCGTATAA`

Logan

27 millions accessions, **~1PB**

Our problem: comparing all metagenomes

David Koslicki
September 2025

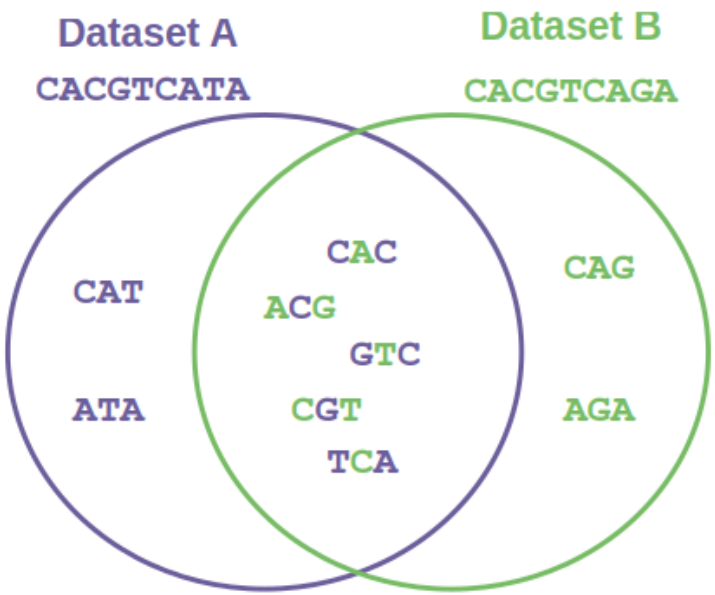
- Phase 1: Compute all-vs-all distance between metagenomes using Logan's unitigs
- Phase 2: Find patterns in metagenomes

It's a big problem!

- 5M metagenomes: 100TB sequences
- All-vs-all means 5M×5M = 25,000 billion comparisons

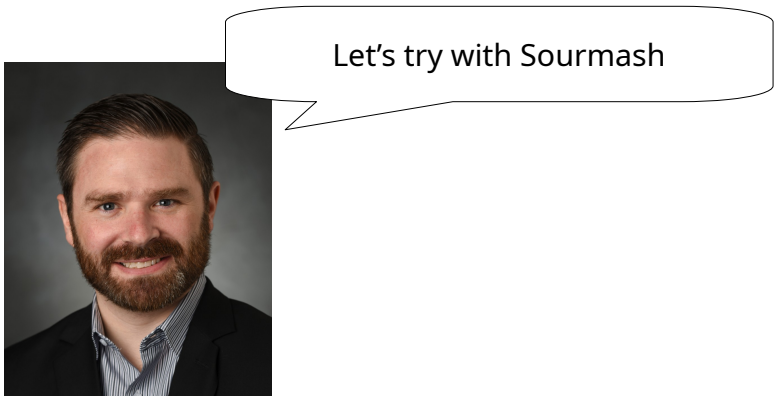
It's a big problem!

- 5M metagenomes: 100TB sequences
- All-vs-all means 5M×5M = 25,000 billion comparisons
- Use Jaccard index as a distance



$J = |A \cap B| / |A \cup B| = 5/9$

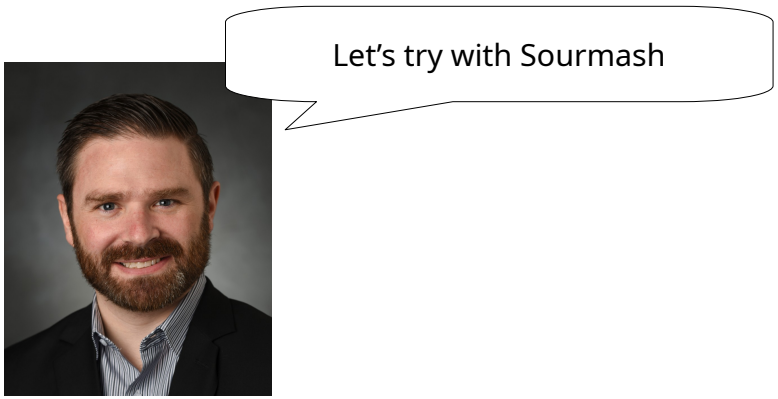
First try: sourmash



David Koslicki
September 2025

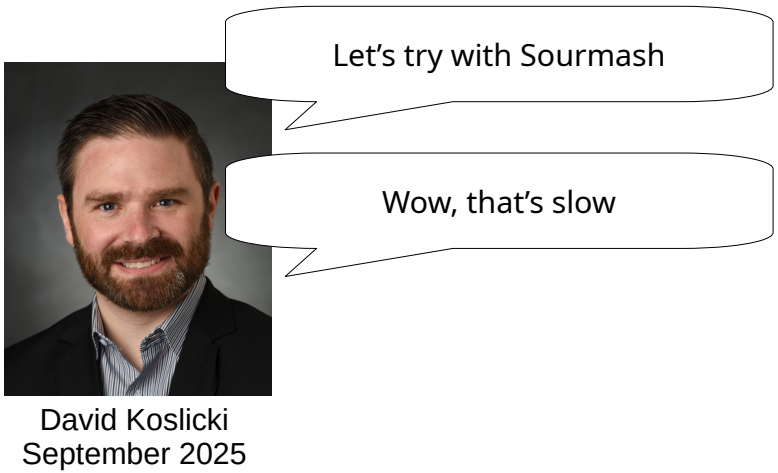
- Sourmash: take 1/1000 k-mer, measure Jaccard on this sketch

First try: sourmash



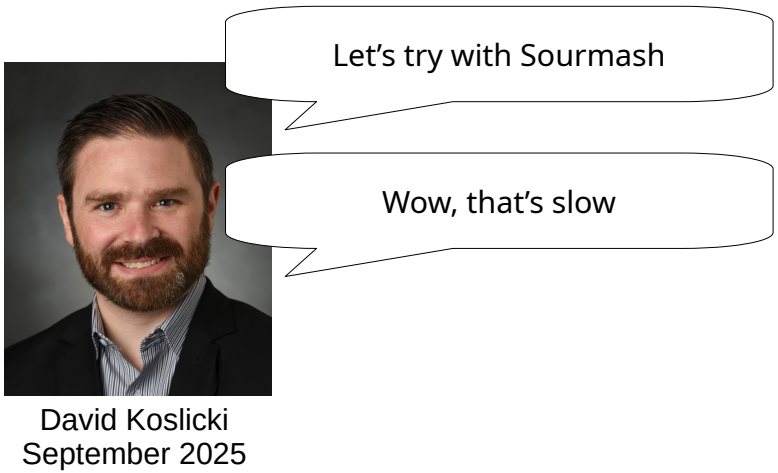
- Sourmash: take 1/1000 k-mer, measure Jaccard on this sketch
- Estimated total time it will take: 15 CPU.years

First try: sourmash



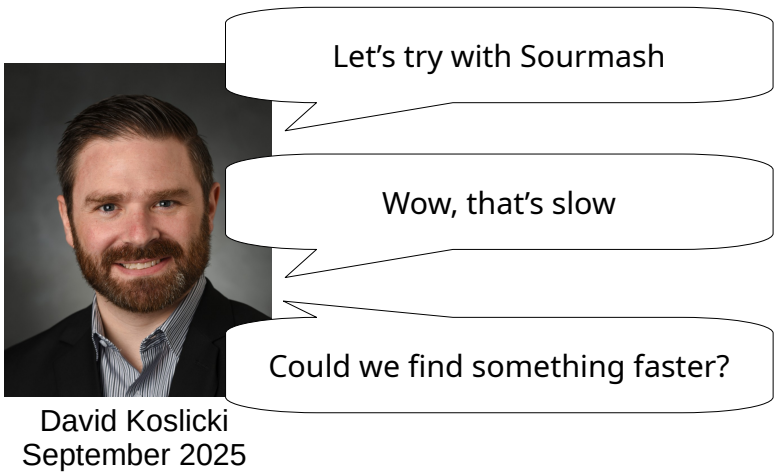
- Sourmash: take 1/1000 k-mer, measure Jaccard on this sketch
- Estimated total time it will take: 15 CPU.years

First try: sourmash



- Sourmash: take 1/1000 k-mer, measure Jaccard on this sketch
- Estimated total time it will take: 15 CPU.years
- The problem: even taking 1/1000 k-mer, sketches get big

First try: sourmash



- Sourmash: take 1/1000 k-mer, measure Jaccard on this sketch
- Estimated total time it will take: 15 CPU.years
- The problem: even taking 1/1000 k-mer, sketches get big

Huge all-vs-all Jaccard computation: existing methods

- Based on **fixed-size** sketches

Huge all-vs-all Jaccard computation: existing methods

- Based on **fixed-size** sketches

- MinHash (e.g. Mash)
- HyperLogLog (e.g. Dashing2)
- DotHash / Random projections (e.g. Hypergen)

Comparing and indexing metagenomes at a large scale using random projections

Roland Faure[1], Hasin Abrar[2], Haonan Wu[2], Stephanie Won[2], Adrita Hossain Nakshi[2], Rayan Chikhi[1], David Koslicki[2], Paul Medvedev[2]

[1]Institut Pasteur

[2]The Pennsylvania State University

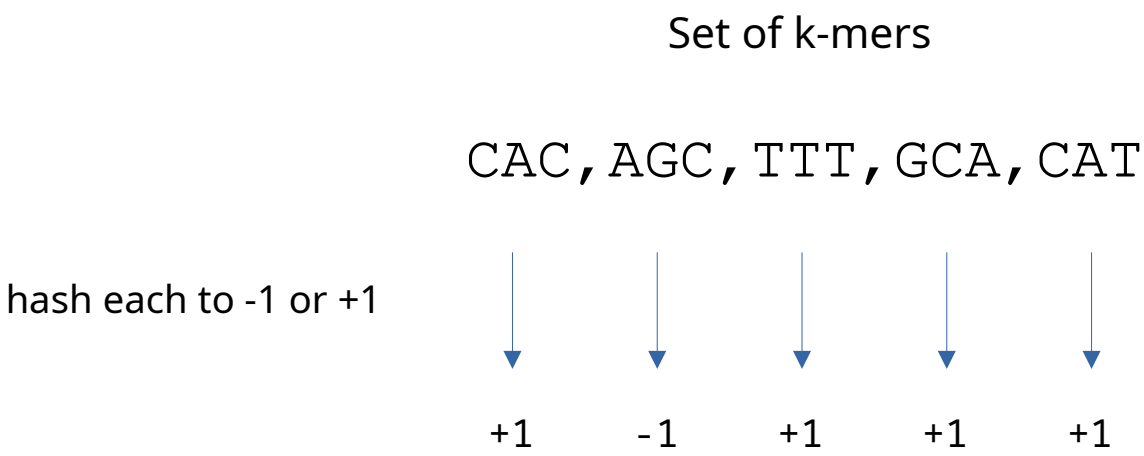
SeqBIM 2025

DotHash/random projections: idea

Set of k-mers

`CAC,AGC,TTT,GCA,CAT`

DotHash/random projections: idea



DotHash/random projections: idea

Set of k-mers

`CAC,AGC,TTT,GCA,CAT`

hash each to -1 or +1

+1 -1 +1 +1 +1

sum

=

3

DotHash/random projections: idea

Set of k-mers

`CAC,AGC,TTT,GCA,CAT`

hash each to -1 or +1

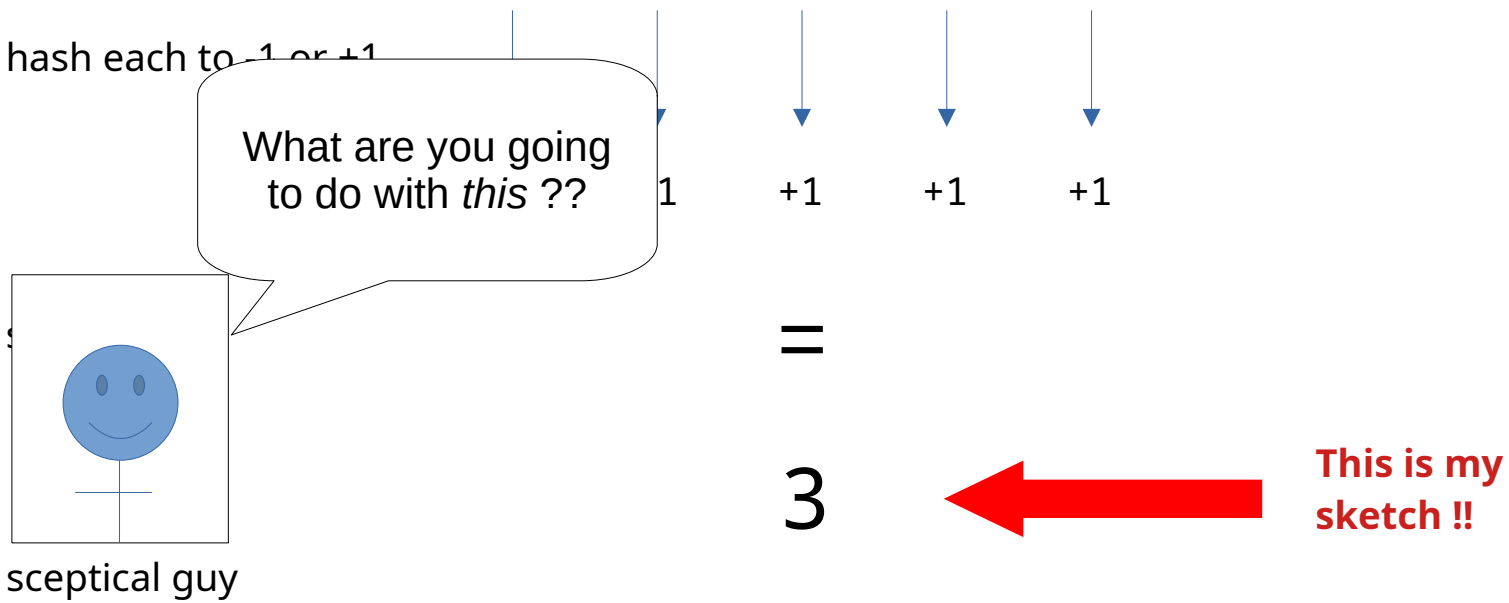
sum

=

3

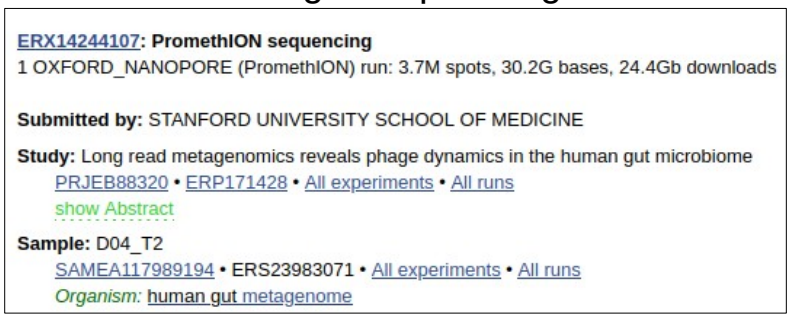
This is my sketch !!

DotHash/random projections: idea



DotHash/random projections: idea

Human gut sequencing

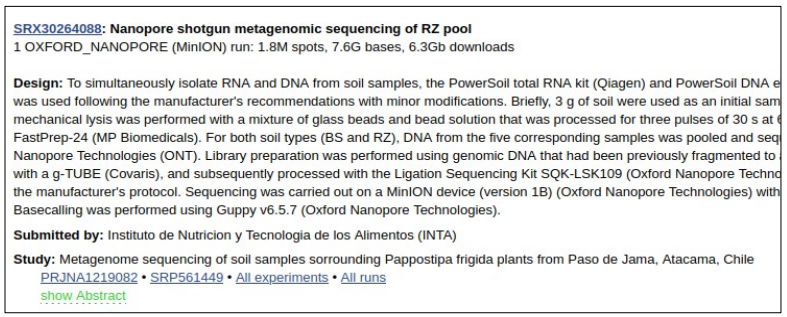

ERX14244107: **PromethION sequencing**
1 OXFORD_NANOPORE (PromethION) run: 3.7M spots, 30.2G bases, 24.4Gb downloads

Submitted by: STANFORD UNIVERSITY SCHOOL OF MEDICINE

Study: Long read metagenomics reveals phage dynamics in the human gut microbiome
PRJEB88320 • ERP171428 • All experiments • All runs
show Abstract

Sample: D04_T2
SAMEA117989194 • ERS23983071 • All experiments • All runs
Organism: human gut metagenome

Soil sequencing


SRX30264088: **Nanopore shotgun metagenomic sequencing of RZ pool**
1 OXFORD_NANOPORE (MinION) run: 1.8M spots, 7.6G bases, 6.3Gb downloads

Design: To simultaneously isolate RNA and DNA from soil samples, the PowerSoil total RNA kit (Qiagen) and PowerSoil DNA e
was used following the manufacturer's recommendations with minor modifications. Briefly, 3 g of soil were used as an initial sam
mechanical lysis was performed with a mixture of glass beads and bead solution that was processed for three pulses of 30 s at
FastPrep-24 (MP Biomedicals). For both soil types (BS and RZ), DNA from the five corresponding samples was pooled and seq
Nanopore Technologies (ONT). Library preparation was performed using genomic DNA that had been previously fragmented to
with a g-TUBE (Covaris), and subsequently processed with the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Techno
the manufacturer's protocol. Sequencing was carried out on a MinION device (version 1B) (Oxford Nanopore Technologies) with
Basecalling was performed using Guppy v6.5.7 (Oxford Nanopore Technologies).

Submitted by: Instituto de Nutricion y Tecnologia de los Alimentos (INTA)

Study: Metagenome sequencing of soil samples sorrounding Pappostipa frigida plants from Paso de Jama, Atacama, Chile
PRJNA1219082 • SRP561449 • All experiments • All runs
show Abstract

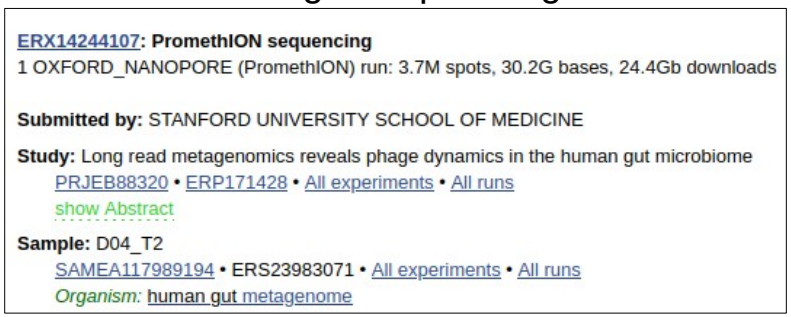
sketch

-276

-98

DotHash/random projections: idea

Human gut sequencing

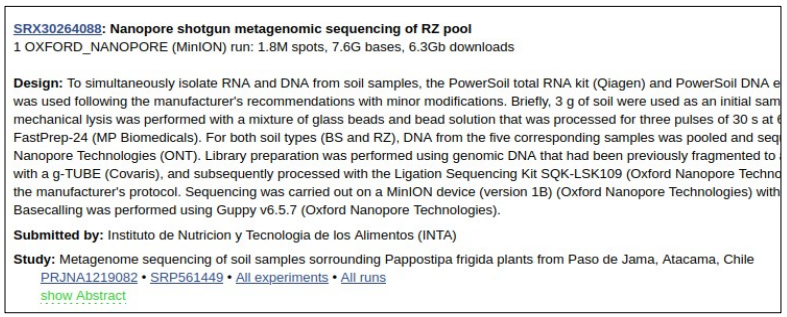

ERX14244107: **PromethION sequencing**
1 OXFORD_NANOPORE (PromethION) run: 3.7M spots, 30.2G bases, 24.4Gb downloads

Submitted by: STANFORD UNIVERSITY SCHOOL OF MEDICINE

Study: Long read metagenomics reveals phage dynamics in the human gut microbiome
PRJEB88320 • ERP171428 • All experiments • All runs
show Abstract

Sample: D04_T2
SAMEA117989194 • ERS23983071 • All experiments • All runs
Organism: human gut metagenome

Soil sequencing


SRX30264088: **Nanopore shotgun metagenomic sequencing of RZ pool**
1 OXFORD_NANOPORE (MinION) run: 1.8M spots, 7.6G bases, 6.3Gb downloads

Design: To simultaneously isolate RNA and DNA from soil samples, the PowerSoil total RNA kit (Qiagen) and PowerSoil DNA e
was used following the manufacturer's recommendations with minor modifications. Briefly, 3 g of soil were used as an initial sam
mechanical lysis was performed with a mixture of glass beads and bead solution that was processed for three pulses of 30 s at
FastPrep-24 (MP Biomedicals). For both soil types (BS and RZ), DNA from the five corresponding samples was pooled and seq
Nanopore Technologies (ONT). Library preparation was performed using genomic DNA that had been previously fragmented to
with a g-TUBE (Covaris), and subsequently processed with the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Techno
the manufacturer's protocol. Sequencing was carried out on a MinION device (version 1B) (Oxford Nanopore Technologies) with
Basecalling was performed using Guppy v6.5.7 (Oxford Nanopore Technologies).

Submitted by: Instituto de Nutricion y Tecnologia de los Alimentos (INTA)

Study: Metagenome sequencing of soil samples sorrounding Pappostipa frigida plants from Paso de Jama, Atacama, Chile
PRJNA1219082 • SRP561449 • All experiments • All runs
show Abstract

sketch

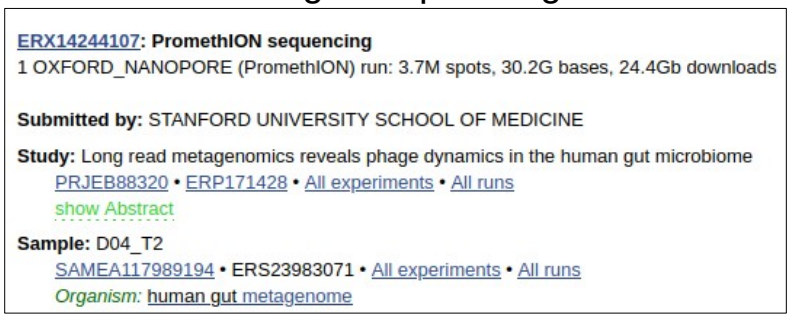
-276

-98

intersection size: -276 x -98 = 27048 k-mers

DotHash/random projections: idea

Human gut sequencing

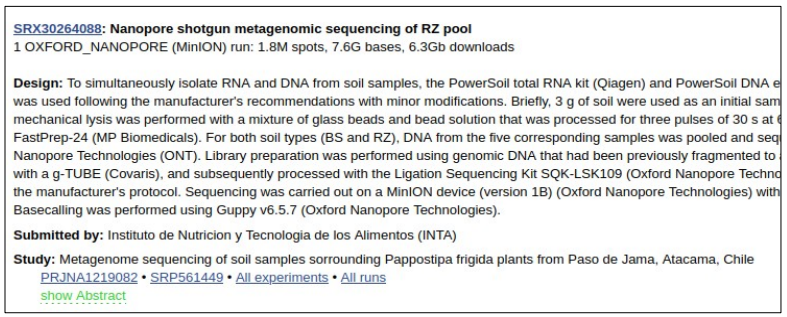

ERX14244107: **PromethION sequencing**
1 OXFORD_NANOPORE (PromethION) run: 3.7M spots, 30.2G bases, 24.4Gb downloads

Submitted by: STANFORD UNIVERSITY SCHOOL OF MEDICINE

Study: Long read metagenomics reveals phage dynamics in the human gut microbiome
PRJEB88320 • ERP171428 • All experiments • All runs
show Abstract

Sample: D04_T2
SAMEA117989194 • ERS23983071 • All experiments • All runs
Organism: human gut metagenome

Soil sequencing


SRX30264088: **Nanopore shotgun metagenomic sequencing of RZ pool**
1 OXFORD_NANOPORE (MinION) run: 1.8M spots, 7.6G bases, 6.3Gb downloads

Design: To simultaneously isolate RNA and DNA from soil samples, the PowerSoil total RNA kit (Qiagen) and PowerSoil DNA e
was used following the manufacturer's recommendations with minor modifications. Briefly, 3 g of soil were used as an initial sam
mechanical lysis was performed with a mixture of glass beads and bead solution that was processed for three pulses of 30 s at
FastPrep-24 (MP Biomedicals). For both soil types (BS and RZ), DNA from the five corresponding samples was pooled and seq
Nanopore Technologies (ONT). Library preparation was performed using genomic DNA that had been previously fragmented to
with a g-TUBE (Covaris), and subsequently processed with the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Techno
the manufacturer's protocol. Sequencing was carried out on a MinION device (version 1B) (Oxford Nanopore Technologies) with
Basecalling was performed using Guppy v6.5.7 (Oxford Nanopore Technologies).

Submitted by: Instituto de Nutricion y Tecnologia de los Alimentos (INTA)

Study: Metagenome sequencing of soil samples sorrounding Pappostipa frigida plants from Paso de Jama, Atacama, Chile
PRJNA1219082 • SRP561449 • All experiments • All runs
show Abstract

sketch

-276

-98

intersection size: -276 x -98 = 27048 k-mers

unbiased estimator

DotHash/random projections: idea

Human gut sequencing

ERX14244107: **PromethION sequencing**
1 OXFORD_NANOPORE (PromethION) run: 3.7M spots, 30.2G bases, 24.4Gb downloads

Submitted by: STANFORD UNIVERSITY SCHOOL OF MEDICINE

Study: Long read metagenomics reveals phage dynamics in the human gut microbiome
PRJEB88320 • ERP171428 • All experiments • All runs
show Abstract

Sample: D04_T2
SAMEA117989194 • ERS23983071 • All experiments • All runs
Organism: human gut metagenome

Soil sequencing

SRX30264088: **Nanopore shotgun metagenomic sequencing of RZ pool**
1 OXFORD_NANOPORE (MinION) run: 1.8M spots, 7.6G bases, 6.3Gb downloads

Design: To simultaneously isolate RNA and DNA from soil samples, the PowerSoil total RNA kit (Qiagen) and PowerSoil DNA e
was used following the manufacturer's recommendations with minor modifications. Briefly, 3 g of soil were used as an initial sam
mechanical lysis was performed with a mixture of glass beads and bead solution that was processed for three pulses of 30 s at
FastPrep-24 (MP Biomedicals). For both soil types (BS and RZ), DNA from the five corresponding samples was pooled and seq
Nanopore Technologies (ONT). Library preparation was performed using genomic DNA that had been previously fragmented to
with a g-TUBE (Covaris), and subsequently processed with the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Techno
the manufacturer's protocol. Sequencing was carried out on a MinION device (version 1B) (Oxford Nanopore Technologies) with
Basecalling was performed using Guppy v6.5.7 (Oxford Nanopore Technologies).

Submitted by: Instituto de Nutricion y Tecnologia de los Alimentos (INTA)

Study: Metagenome sequencing of soil samples sorrounding Pappostipa frigida plants from Paso de Jama, Atacama, Chile
PRJNA1219082 • SRP561449 • All experiments • All runs
show Abstract

↓

-276

sketch

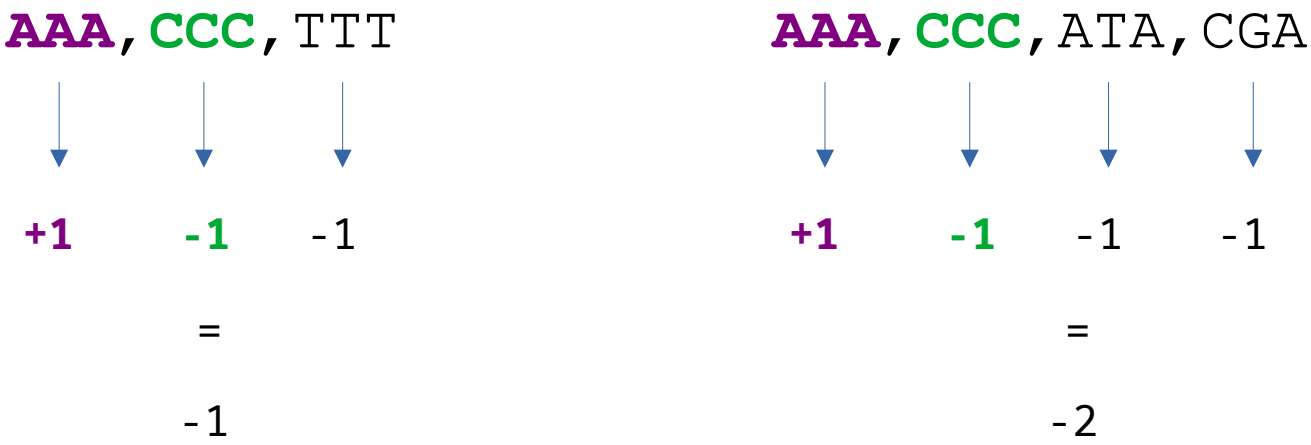
No way!

Rayan Chikhi,
November 2025

intersection size: -276 x -98 = 27048 k-mers

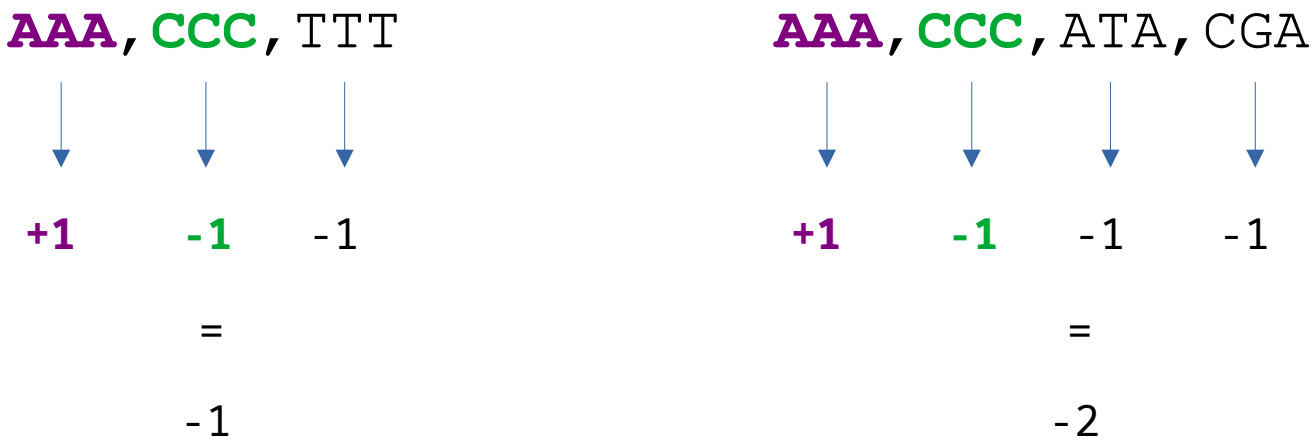
unbiased estimator

DotHash/random projections: idea explanation



-1 x -2 = (+1-1-1)x(+1-1-1-1)

DotHash/random projections: idea explanation



-1 x -2 = (+1-1-1)x(+1-1-1-1)

= 1x1 + -1x-1 + 1x-1 + 1x-1 + 1x-1 + … + -1x-1

DotHash/random projections: idea explanation

-1 x -2 = (+1-1-1)x(+1-1-1-1)

= 1x1 + -1x-1 + 1x-1 + 1x-1 + 1x-1 + … + -1x-1

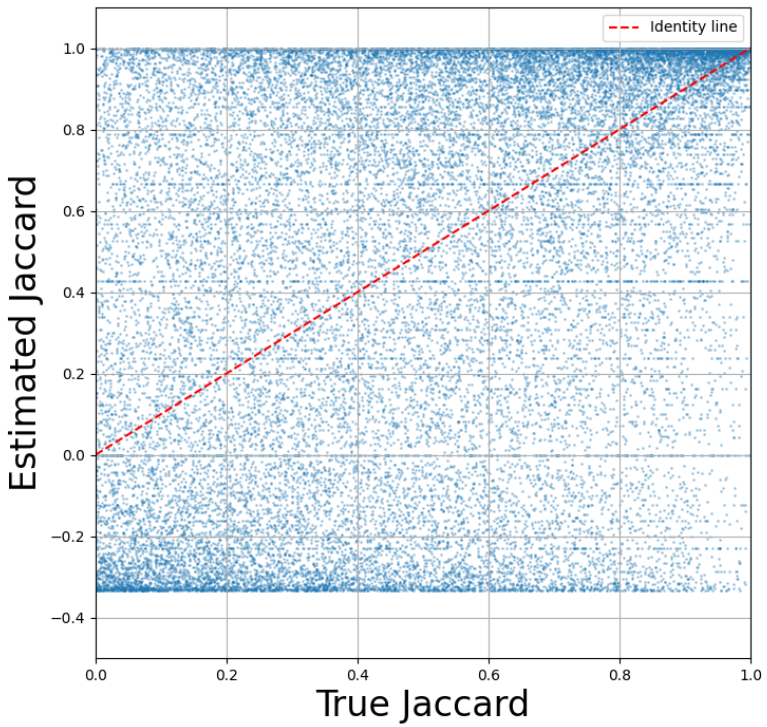
= |A n B| + sum of random -1 and +1, on average 0

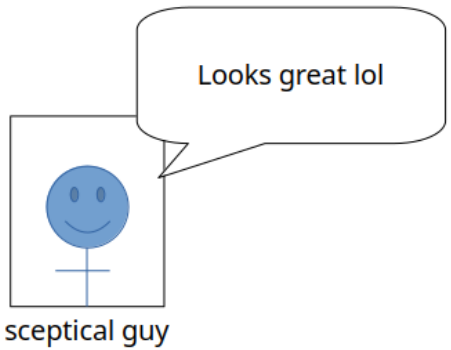
DotHash: in practice

- Let's try to compute the Jaccard of 30k pairs of sets using this method

DotHash: in practice

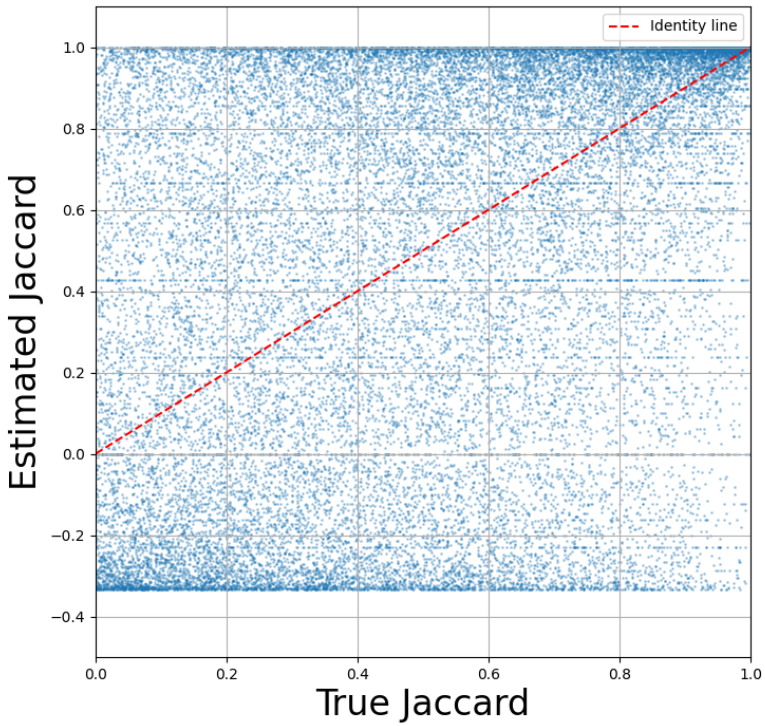
- Let's try to compute the Jaccard of 30k pairs of sets using this method

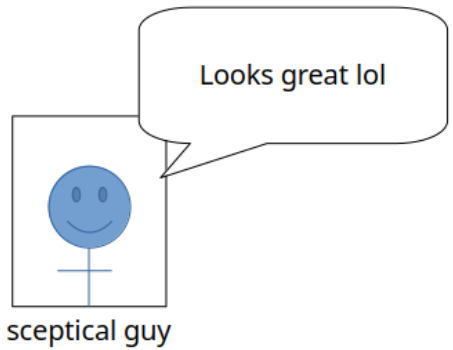




DotHash: in practice

- Let's try to compute the Jaccard of 30k pairs of sets using this method





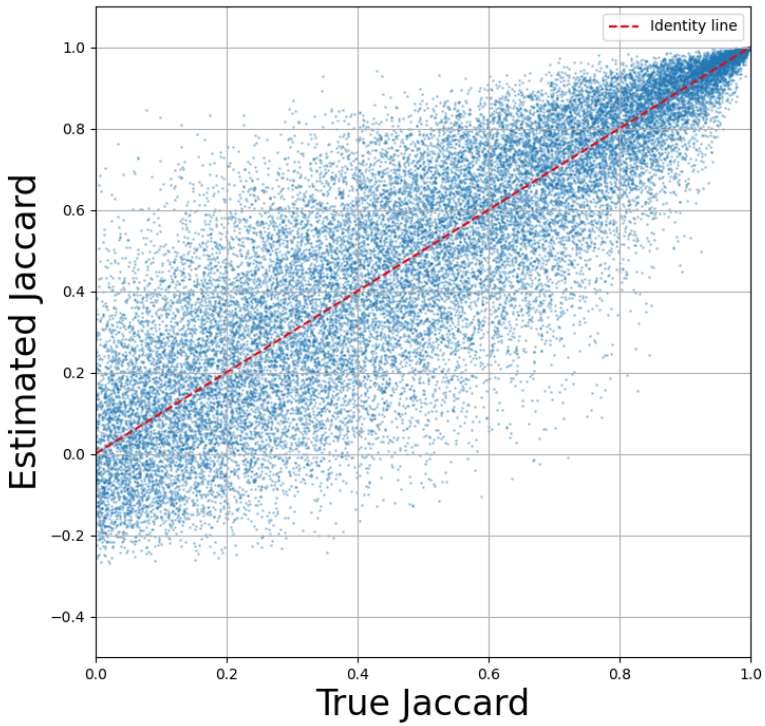
- It works on average, right?

DotHash: in practice

- For each pair of datasets, using 10 different hash functions and taking the mean

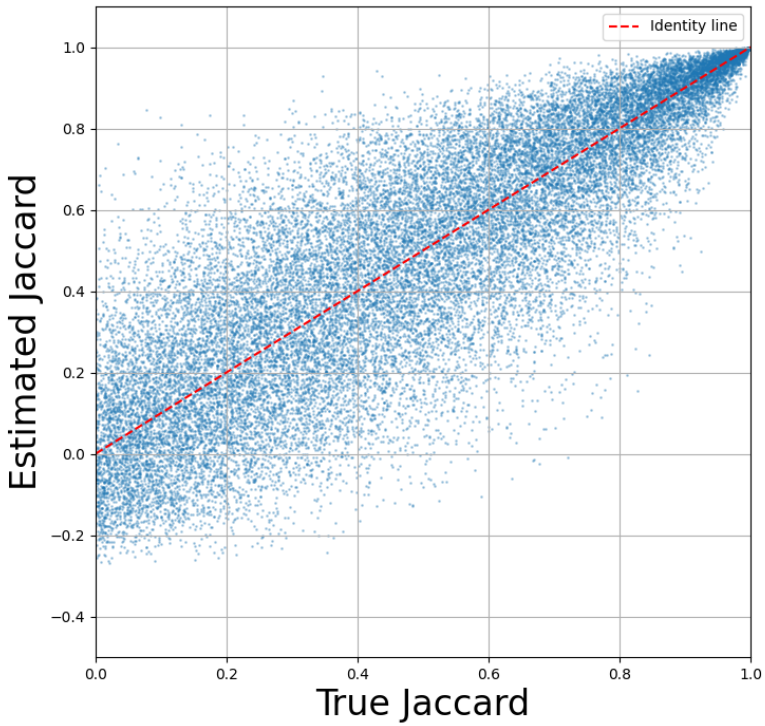
DotHash: in practice

- For each pair of datasets, using 10 different hash functions and taking the mean



DotHash: in practice

- For each pair of datasets, using 10 different hash functions and taking the mean



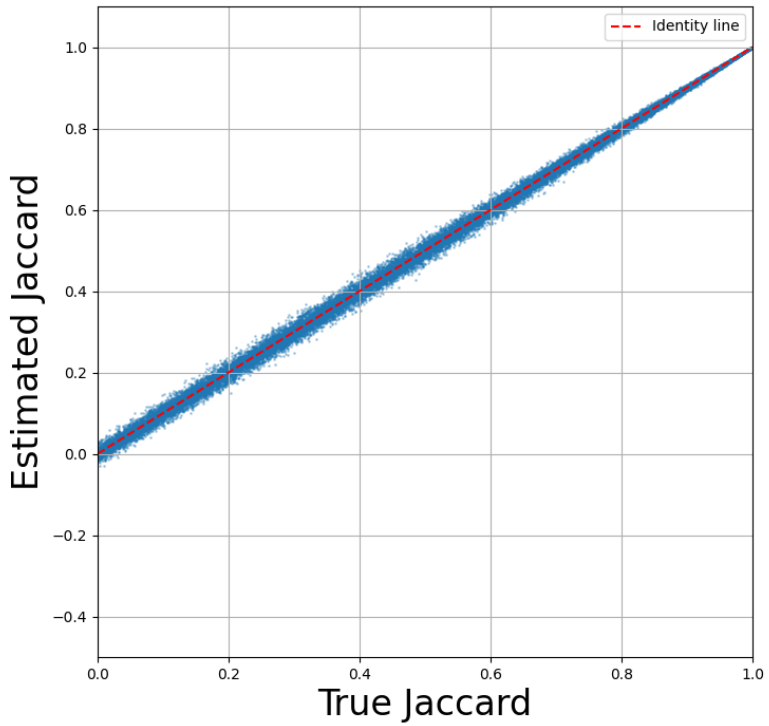
- It works better!

DotHash: in practice

- For each pair of datasets, using 2048 different hash functions and taking the mean

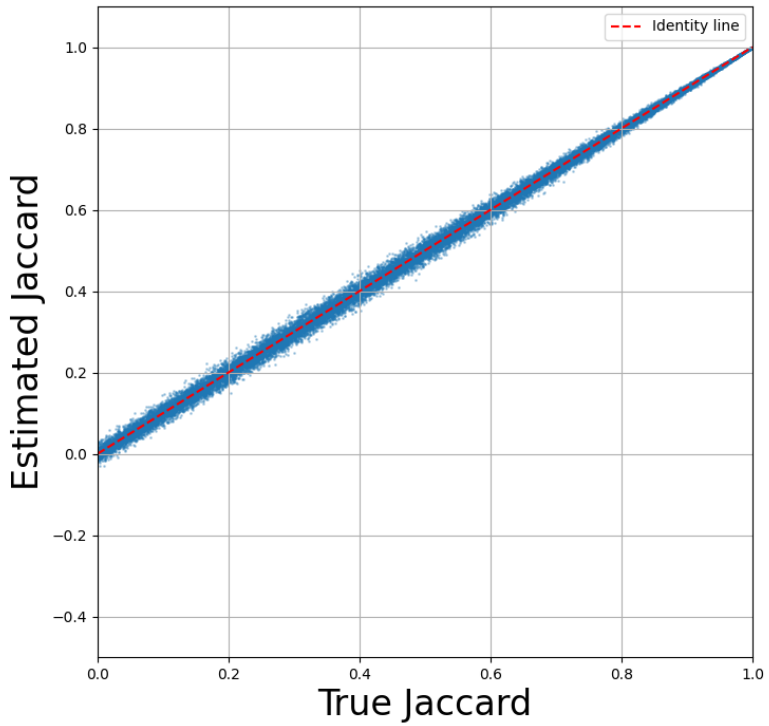
DotHash: in practice

- For each pair of datasets, using 2048 different hash functions and taking the mean



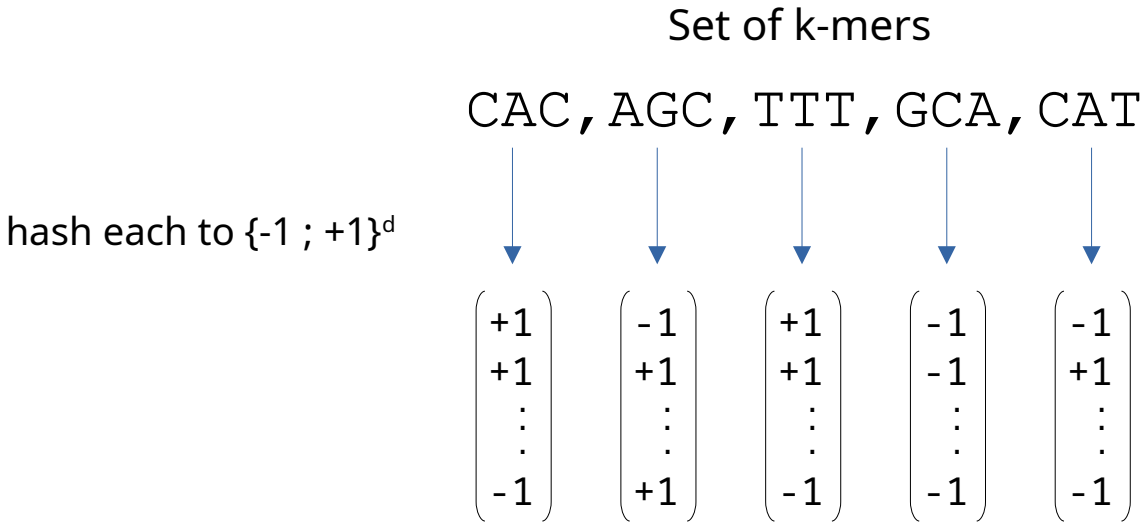
DotHash: in practice

- For each pair of datasets, using 2048 different hash functions and taking the mean

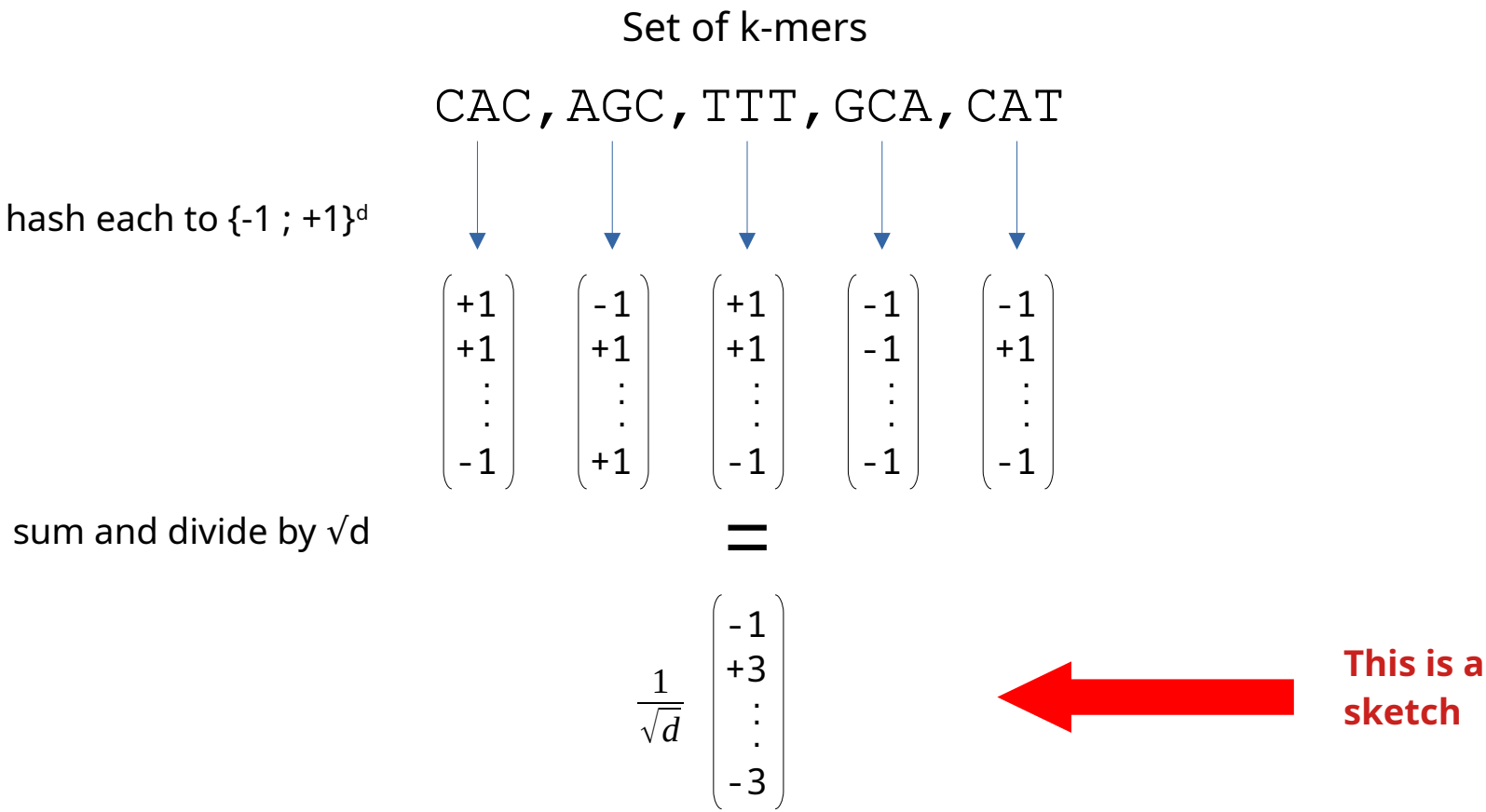


- It works!

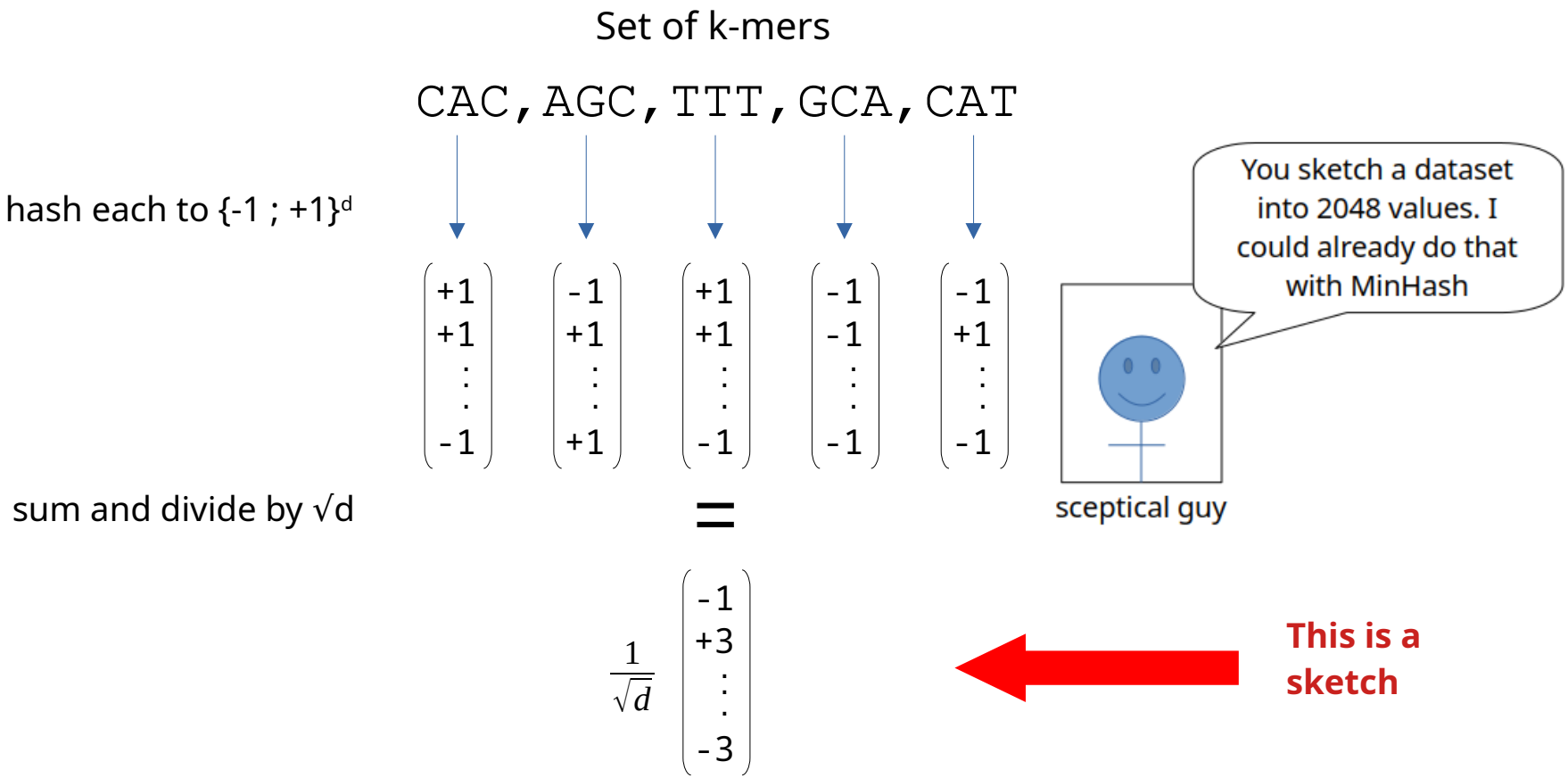
DotHash: the full method



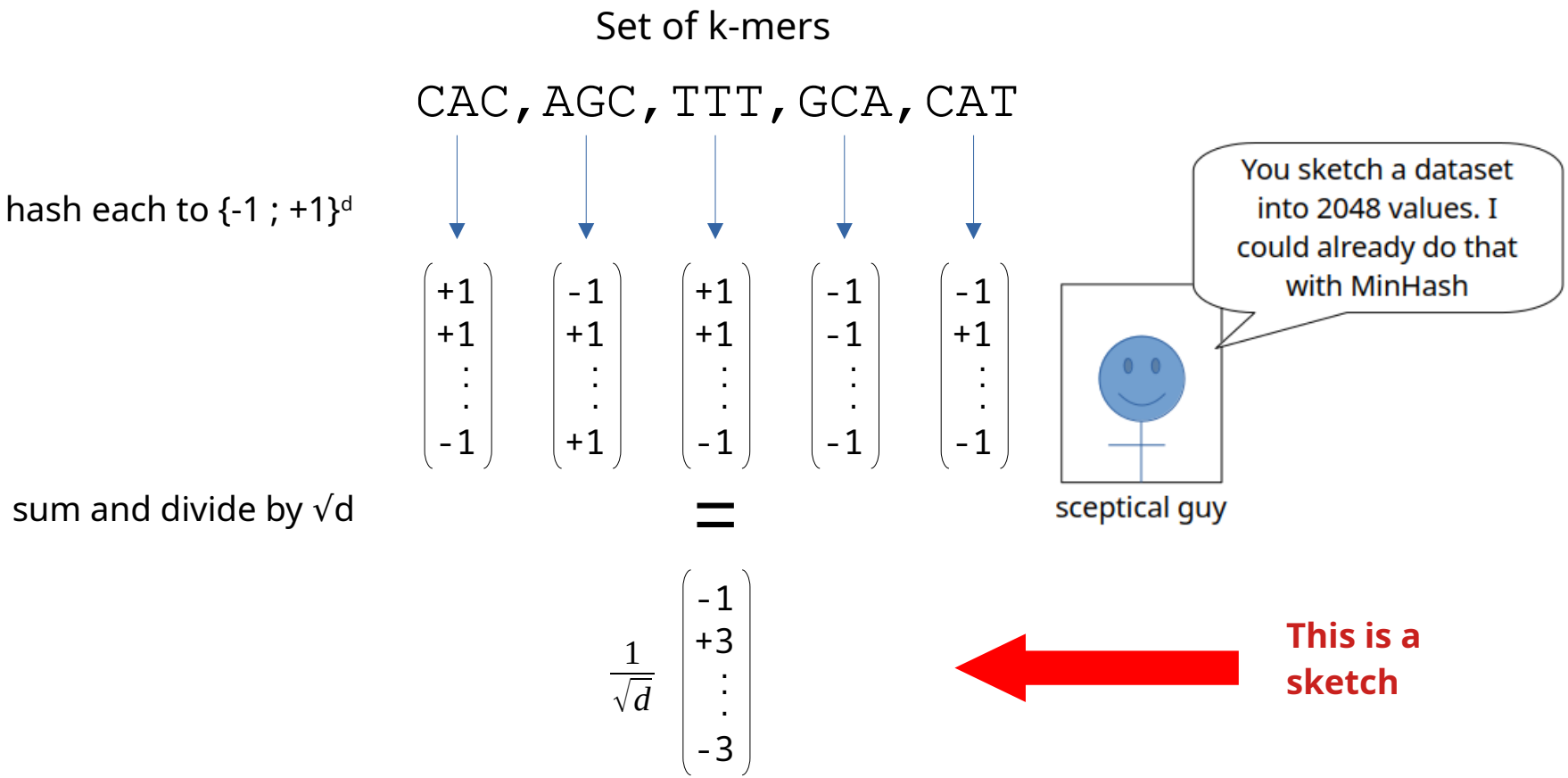
DotHash: the full method



DotHash: the full method



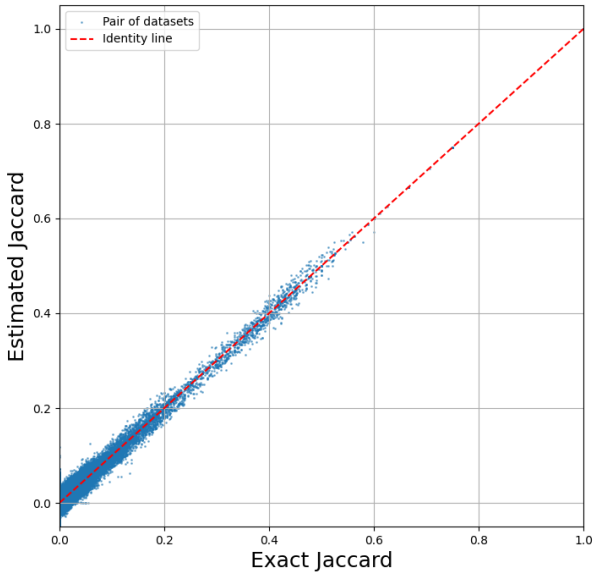
DotHash: the full method



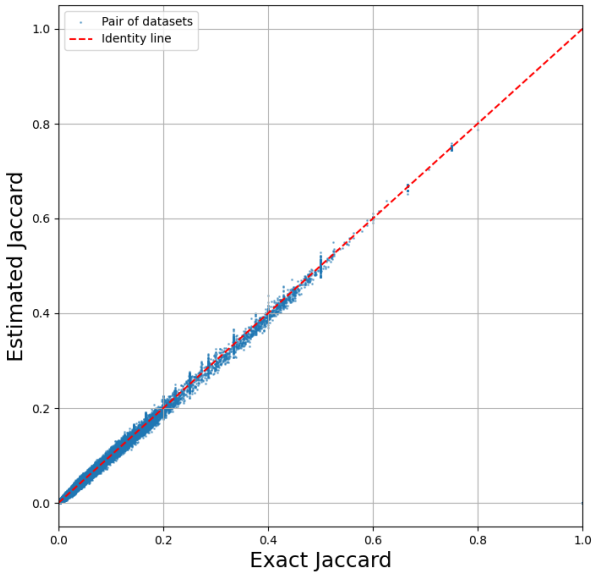
- Let's benchmark the time for all-vs-all jaccard computation

Comparison with other sketching techiques

- True vs estimated Jaccard on 35k real datasets



(a) DotHash (d=1800)

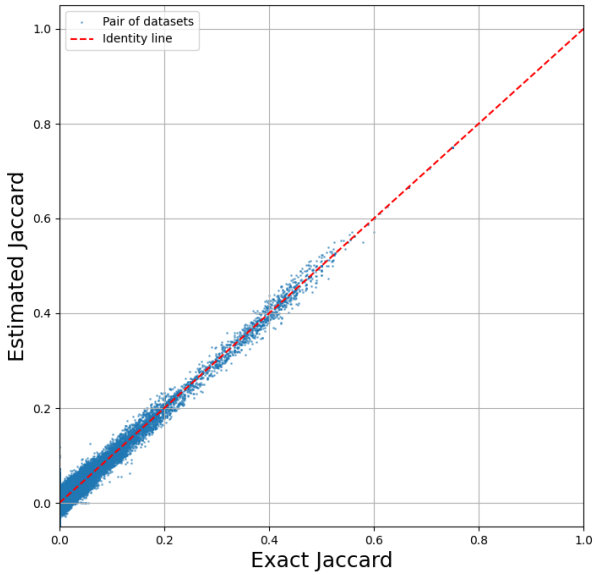


(b) HyperLogLog (S=2048)

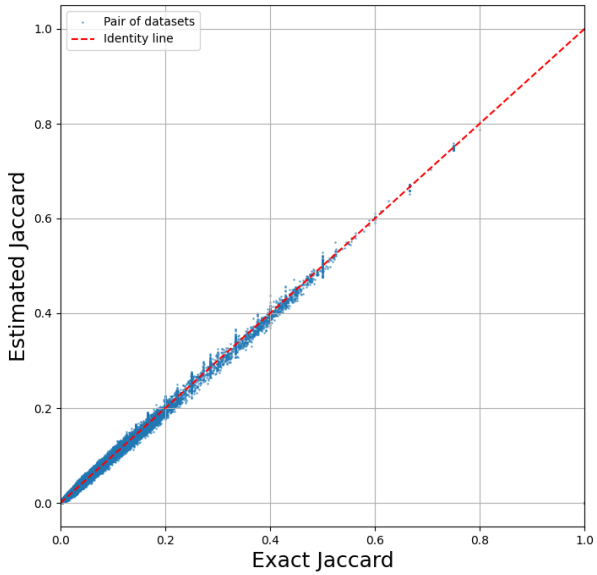
- Different sketching techniques have different error patterns

Comparison with other sketching techiques

- True vs estimated Jaccard on 35k real datasets



(a) DotHash (d=1800)

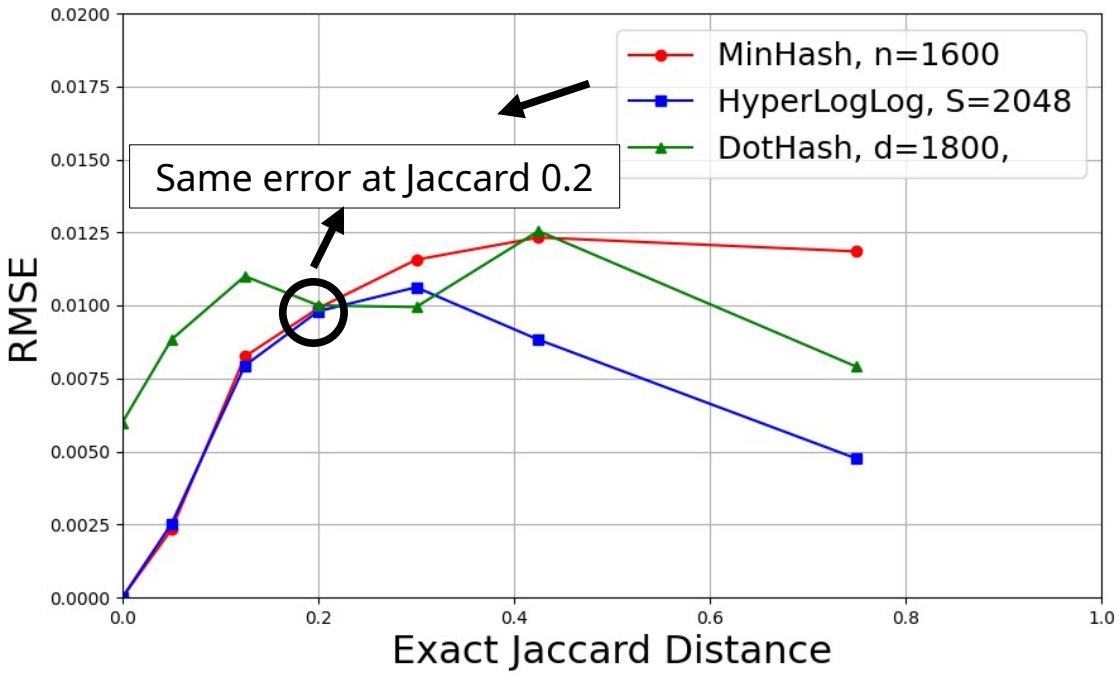


(b) HyperLogLog (S=2048)

- Different sketching techniques have different error patterns
- Let's calibrate the benchmark

Calibrating the benchmark

- We measured the error on 35k real datasets



- Let's benchmark times & memory

Time benchmark of the sketching techiques

- Benchmark on 35k real metagenomes
- Extrapolate the time it would take for the 5M metagenomes and for all Logan

Time benchmark of the sketching techiques

- Benchmark on 35k real metagenomes
- Extrapolate the time it would take for the 5M metagenomes and for all Logan

Method	35k genomes	5M genomes (extrapolated)	Logan (extrapolated)
Sourmash	8 CPU.hours	-	-

Time benchmark of the sketching techiques

- Benchmark on 35k real metagenomes
- Extrapolate the time it would take for the 5M metagenomes and for all Logan

Method	35k genomes	5M genomes (extrapolated)	Logan (extrapolated)
Sourmash	8 CPU.hours	-	-
Mash	4.7 CPU.hours	10 CPU.years	322 CPU.years

Time benchmark of the sketching techiques

- Benchmark on 35k real metagenomes
- Extrapolate the time it would take for the 5M metagenomes and for all Logan

Method	35k genomes	5M genomes (extrapolated)	Logan (extrapolated)
Sourmash	8 CPU.hours	-	-
Mash	4.7 CPU.hours	10 CPU.years	322 CPU.years
Dashing2	12 CPU.minutes	139 CPU.days	13 CPU.years

Time benchmark of the sketching techiques

- Benchmark on 35k real metagenomes
- Extrapolate the time it would take for the 5M metagenomes and for all Logan

Method	35k genomes	5M genomes (extrapolated)	Logan (extrapolated)
Sourmash	8 CPU.hours	-	-
Mash	4.7 CPU.hours	10 CPU.years	322 CPU.years
Dashing2	12 CPU.minutes	139 CPU.days	13 CPU.years
HyperGen	13 CPU.minutes	154 CPU.days	15 CPU.years

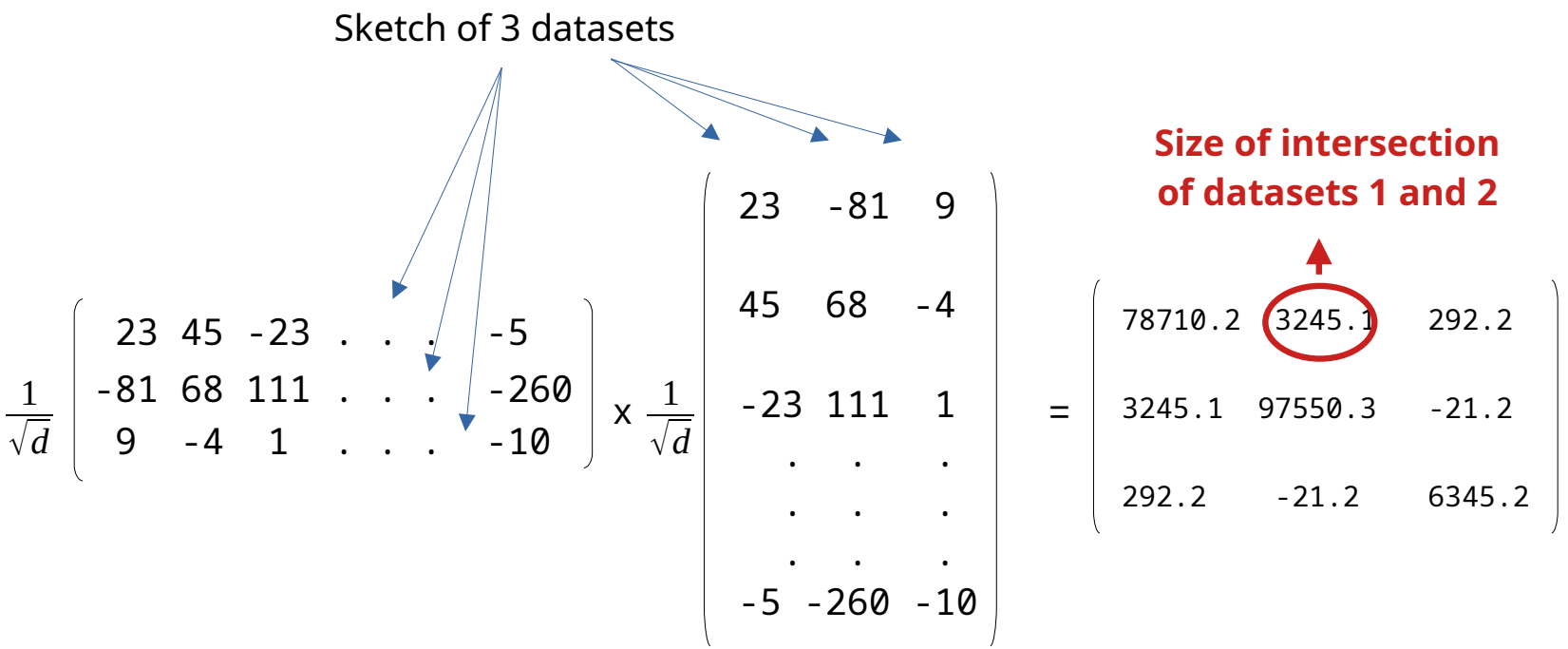
Time benchmark of the sketching techiques

- Benchmark on 35k real metagenomes
- Extrapolate the time it would take for the 5M metagenomes and for all Logan

Method	35k genomes	5M genomes (extrapolated)	Logan (extrapolated)
Sourmash	8 CPU.hours	-	-
Mash	4.7 CPU.hours	10 CPU.years	322 CPU.years
Dashing2	12 CPU.minutes	139 CPU.days	13 CPU.years
HyperGen	13 CPU.minutes	154 CPU.days	15 CPU.years
Custom DotHash	**2 CPU.minutes**	**24 CPU.days**	**2.3 CPU.years**

The strong point of DotHash: practical implementation

- Comparing datasets is a matrix multiplication



- Highly hardware-optimizable, e.g. SIMD, GPU

Back to our original problem

- Comparing all-vs-all jaccard of 5M metagenomes: done in one night!

Back to our original problem

- Comparing all-vs-all jaccard of 5M metagenomes: done in one night!

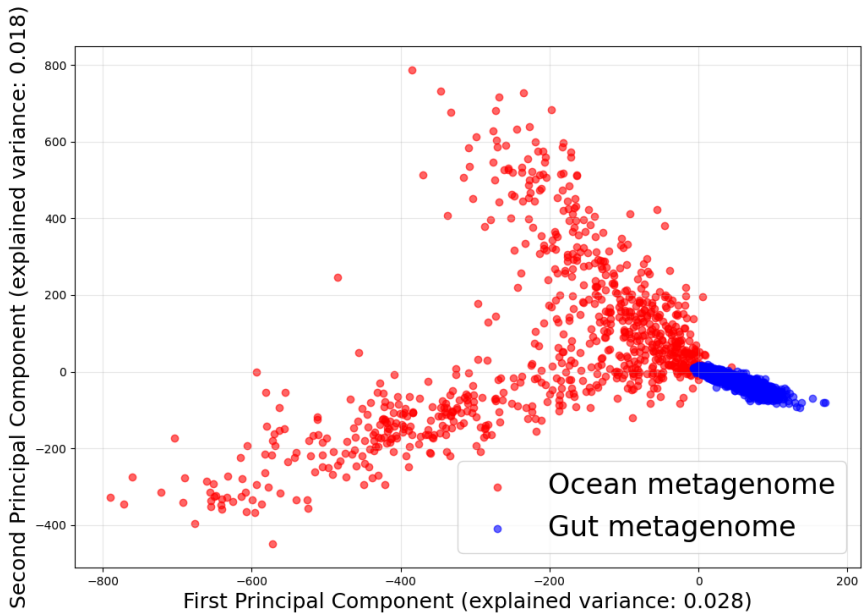
- DotHash vectors can still be interesting!

Manipulating vectors: analysis

- We can run a PCA directly on the sketches!

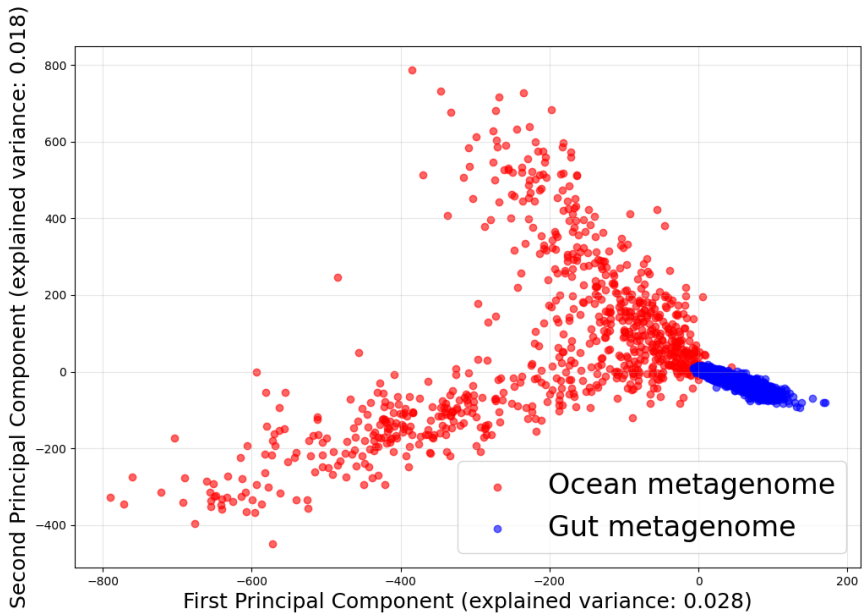
Manipulating vectors: analysis

- We can run a PCA directly on the sketches!
- Example: 7101 metagenomes either from Tara Ocean or American Gut Project



Manipulating vectors: analysis

- We can run a PCA directly on the sketches!
- Example: 7101 metagenomes either from Tara Ocean or American Gut Project



- Also methods for clustering, machine learning, indexing, compressing...

Take-home messages

- DotHash hashes sets into vectors

Take-home messages

- DotHash hashes sets into vectors
- Very computationally efficient

Take-home messages

- DotHash hashes sets into vectors
- Very computationally efficient
- Imprecise for low jaccard distances

Take-home messages

- DotHash hashes sets into vectors
- Very computationally efficient
- Imprecise for low jaccard distances
- Many powerful methods/implementation exist to manipulate vectors

Acknowledgments

Paul Medvedev

David Koslicki

Stephanie Won

Hasin Abrar

Haonan Wu

Rayan Chikhi

Space taken by the sketches

Method	35k genomes	5M genomes (extrapolated)
Mash	408M	52G
Dashing2	548M	245G
Custom DotHash	120M	15G