Multiple Comparative Metagenomics using Multiset k-mer Counting

"Simka"

Pierre Peterlongo

SUMMER SCHOOL 2016 IN METAGENOMICS

WARNING – This summer school

"Its content will focus on the taxonomic assignment and the functional analysis of metatranscriptomic and metagenomic data.

Summer School 2016 in Metagenomics

FRANCE GÉNOMIQUE

elixir

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Institut Pasteur

Metagenomics, the sequencing of DNA directly from a sample without first culturing and isolating the organisms, has become the principal tool of "meta-omic" analysis. It can be used to explore the diversity, function, and ecology of microbial communities.

The aim of these 4 days workshop will be to give researchers and students an overview of the tools and bioinformatics techniques available for the analysis of next generation sequence data from microbial communities. Its content will focus on the taxonomic assignment and the functional analysis of metatranscriptomic and metagenomic data. The format will comprise a mixture of lectures and hands-on practical tutorials where students will process example data sets in real-time.

WARNING – This summer school

"Its content will focus on the ~~taxonomic assignment~~ and the ~~functional analysis~~ of metatranscriptomic and metagenomic data.

Summer School 2016 in Metagenomics

FRANCE GÉNOMIQUE

elixir

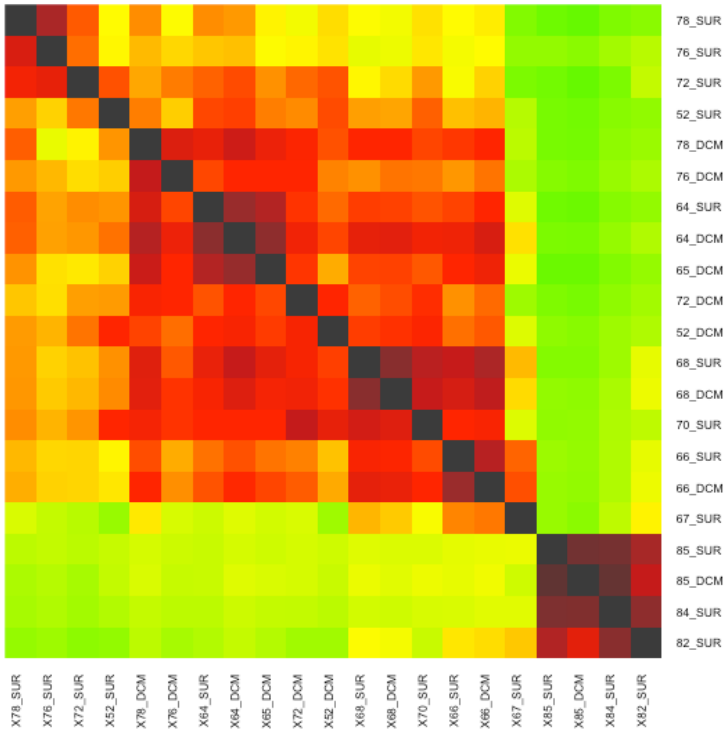
ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Institut Pasteur

Metagenomics, the sequencing of DNA directly from a sample without first culturing and isolating the organisms, has become the principal tool of "meta-omic" analysis. It can be used to explore the diversity, function, and ecology of microbial communities.

The aim of these 4 days workshop will be to give researchers and students an overview of the tools and bioinformatics techniques available for the analysis of next generation sequence data from microbial communities. Its content will focus on the taxonomic assignment and the functional analysis of metatranscriptomic and metagenomic data. The format will comprise a mixture of lectures and hands-on practical tutorials where students will process example data sets in real-time.

“comparing metagenomic samples”



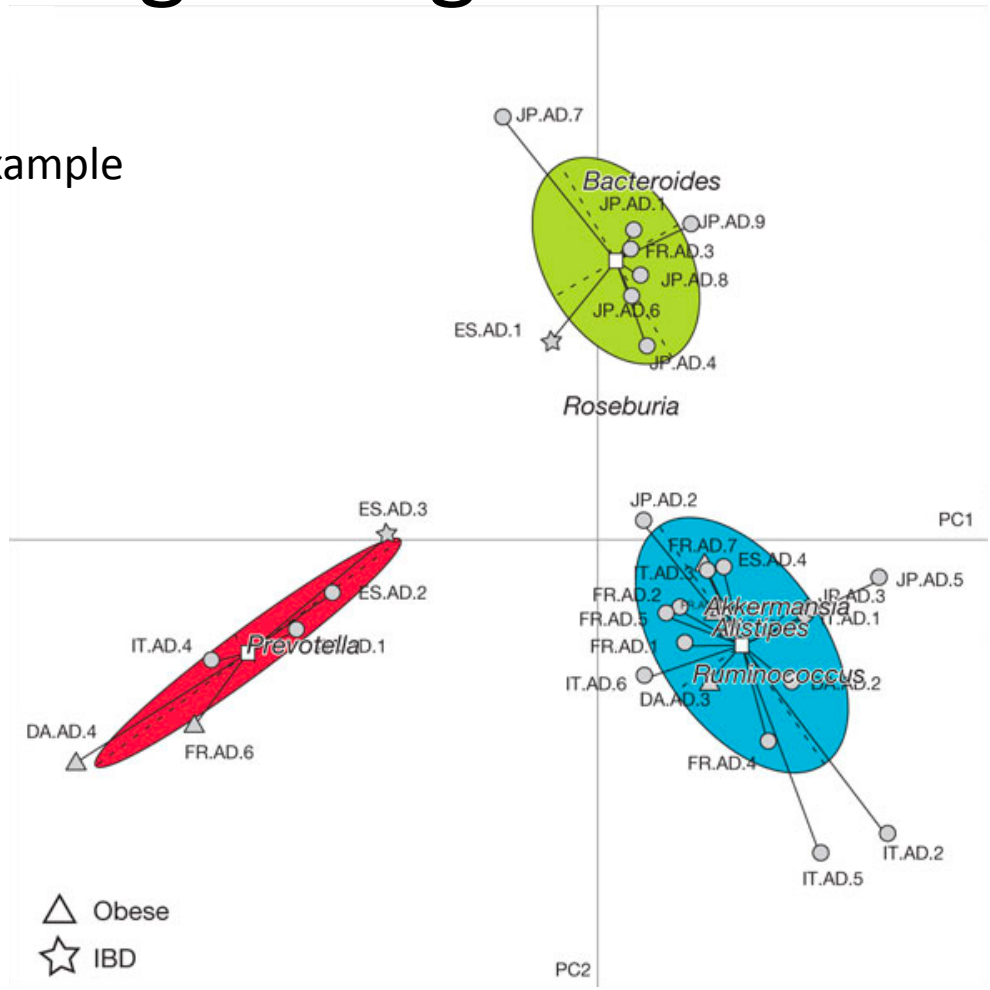
Similarity measure for each couple of datasets

“comparing metagenomic samples”

Mettre ici des publis: Tara, HMP, New york metro, …

“comparing metagenomic samples”

Human enterotypes example

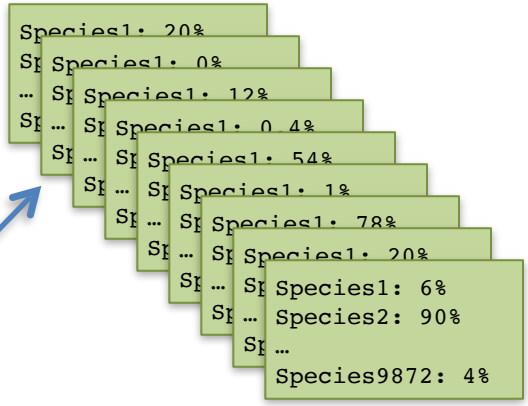


Arumugam *et al.* Enterotypes of the human gut microbiome. Nature, 2011

“reference-based comparison of metagenomic samples”

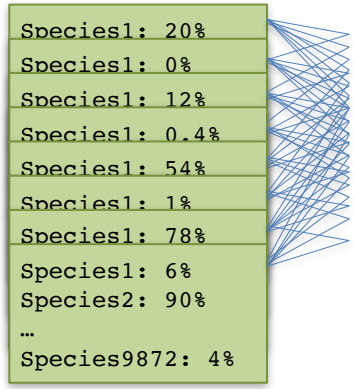
- From reads to taxonomic composition

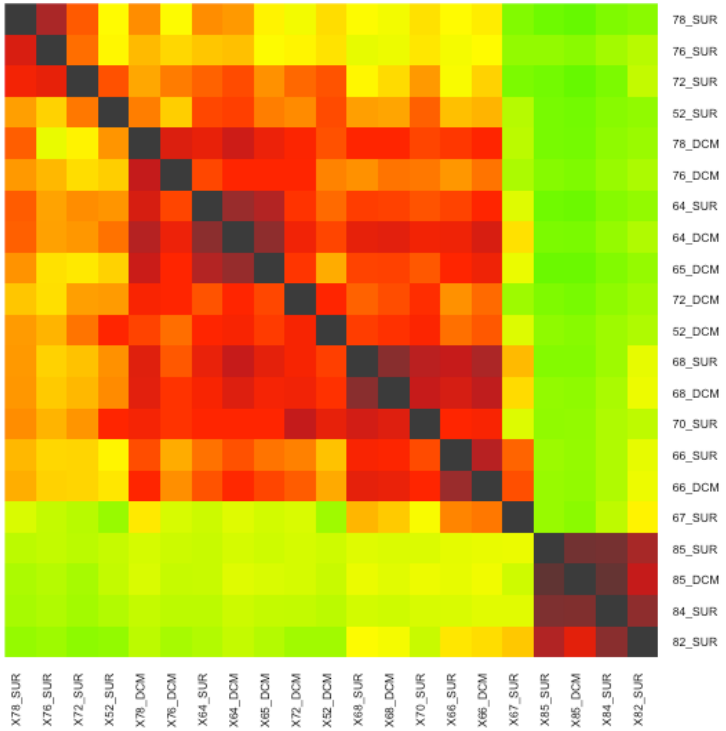
Credit: http://niniejolie.eklablog.com/



“reference-based comparison of metagenomic samples”

- Compare taxonomy composition





"reference-based comparison of metagenomic samples"

- Reference based limitations
  - Databases not representative of diversity

Credit: http://niniejolie.eklablog.com/

"we have only sequenced 10^{-22}% of the total DNA on Earth

(Nature Review Microbiol. editorial, 2011)

“Tara ocean example”

Tara

…

- Hundreds of samples
- Billion of reads

Tara

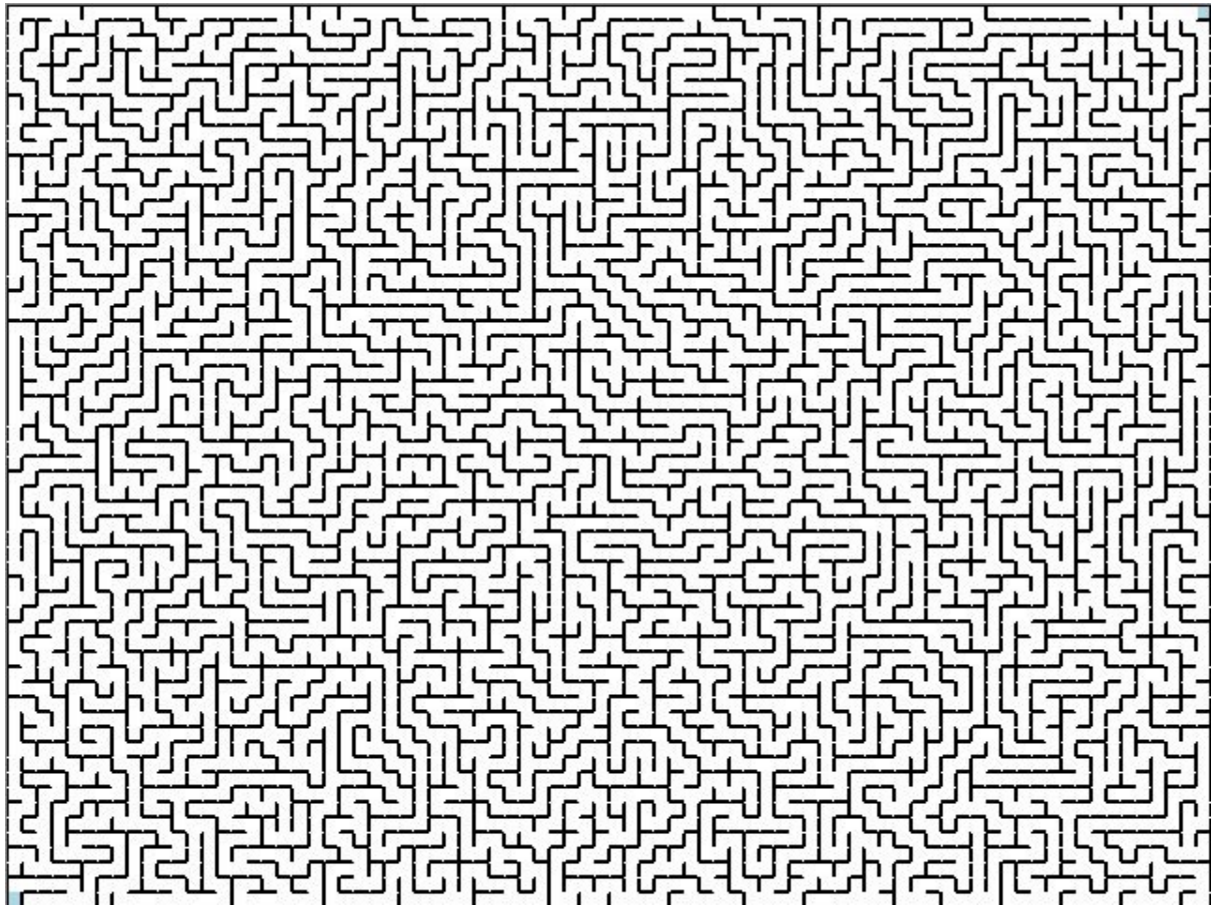
```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

Assembly

Mapping

(**Sea water:**
< 5% assembled reads
< 10% mapped reads)

“de-novo comparison: a complex problem”



Credit: http://junesblog.org/

“de-novo comparison: a complex problem”

- Comparing two reads is simple
  - Classical edit distance problem:

```
TACGGGACTGAT-CAGACGTCAA
 ||||   ||||| || || ||||
 ACGG--CTGATTCATACTTCAAGG
```

Computation ≈1 µsec

Derives similarity measure

“de-novo comparison: a complex problem”

- Comparing two reads is simple
- Comparing
  - 100 read sets
  - each composed of 100 millions reads is (terribly) complex

“de-novo comparison: a complex problem”

- Comparing reads is simple
- Comparing
  - 100 read sets
  - each composed of 100 millions reads is (terribly) complex

$(100.10^6)^2 \times 100^2$ Comparisons

```
TACGGGACTGAT-CAGACGTCAA
 ||||  |||||  ||  ||  ||||
 ACGG--CTGATTCATACTTCAAGG
```

“de-novo comparison: a complex problem”

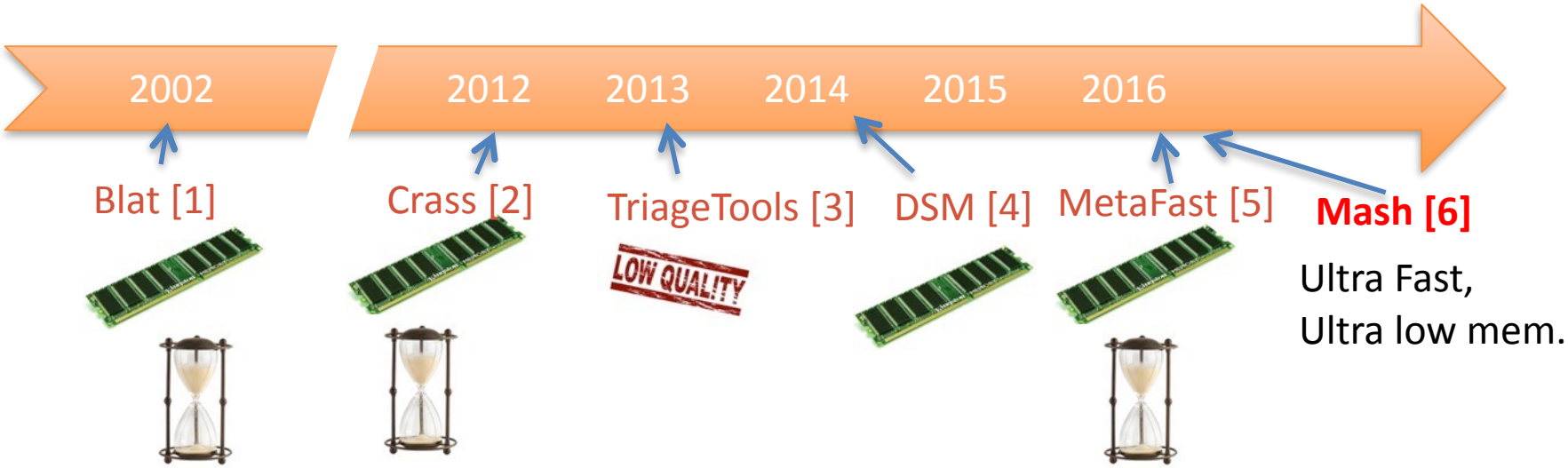
- Comparing reads is simple
- Comparing
  - 100 read sets
  - each composed of 100 millions reads is (terribly) complex

$(100.10^6)^2 \times 100^2 \times 1\ \mu sec$

$= 10^{20}\ \mu sec$ = 3 billions centuries

What tool for *de novo* comparative metagenomics

de novo comparative metagenomics- *State of the art*



[1] Kent 2002
Computation of pairwise distances
[2] Dutilh *et. al.* 2012
Assembly based
[3] Seth *et. al.* 2014
Small kmer based

[4] Fimereli *et al.* 2013
Small k-mer based
[5] Ulyantsev *et. al.* 2016
Unitig & clustering based
[6] Ondov *et. al 2016*
Subsampling – 0/1
Jaccard

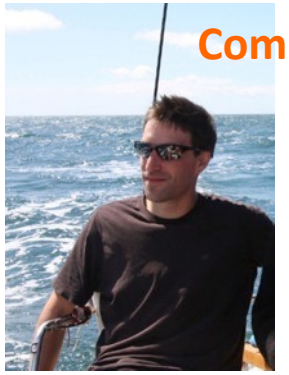
de novo comparative metagenomics- our proposals

Compareads [5]

Commet [6]

Simka [7]

SRC_linker [8]



Nicolas Maillet

Gaetan Benoit

Camille Marchet

Antoine Limasset

[5] Compareads [Maillet *et. al.* 2012]
[6] Commet [Maillet *et. al.* 2014]

[7] Simka [Benoit *et al.* 2016]
[8] SRC_linker [Marchet *et. al.* 2016]

Main **algorithmic idea** of our tools

Main idea

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

Read to read comparisons = way too long

```
TACGGGACTGAT-CAGACGTCAA
 ||||  |||||  ||  ||  ||||
 ACGG--CTGATTCATACTTCAAGG
```

Similarity = 74%

Main idea

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

All ideas based on ***alignment-free*** methods

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

***k-mer* based:**

```
TACGGGACTGAT-CAGACGTCAA
 ||||  |||||  ||  ||  ||||
ACGG--CTGATTCATACTTCAAGG
```

Similarity = 4 shared *k*-mers:

ACGG CTGA TGAT TCAA

56% of positions covered by a shared kmer

Main idea

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

All ideas based on ***alignment-free*** methods

***k-mer* based:**

TACGGACTGATCAGACGTCAA
ACGGCTGATTCATACTTCAAGG

Similarity = 4 shared *k*-mers:
ACGG CTGA TGAT TCAA

56% of positions covered by a shared kmer

Main idea

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

```
>read0
ACACGTCAACGGCTACGGCAGACTAACGACTCAGTC
>read1
ACGAGCATCAGGGGACGTATGTTATGCAGTTCCAGG
>read2
GTCAGGCGATCAGGGCATCTACGGGCATCTACAGTT
>read3
ACGTCATGGCATTCAGCAGCTCATGCGGGCGATATT
…
>read1,000,000,000
TTGCATCGCAGGCCAGGCATCATGGCGGCATTTTTT
```

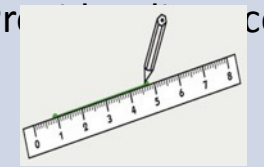
All ideas based on ***alignment free*** methods

```
TACGGGACTGAT-CAGACGTCAA
 ||||  |||||  ||  ||  ||||
 ACGG--CTGATTCATACTTCAAGG
```

alignment → Similarity = 74%

k-mer based → 56% of positions covered by a shared kmer

Our (alignment free) proposals

		Commet	SRC_linker	Simka
Pr[illegible]ce 		✓	✗	✓
Scales up 	Mem	✓	✗	✓
	Time	✗	✗	✓
		— vs 	— vs —	 vs 

Simka focus

Gaetan Benoit

Sophie Schbath

Dominique Lavenier

Claire Lemaitre

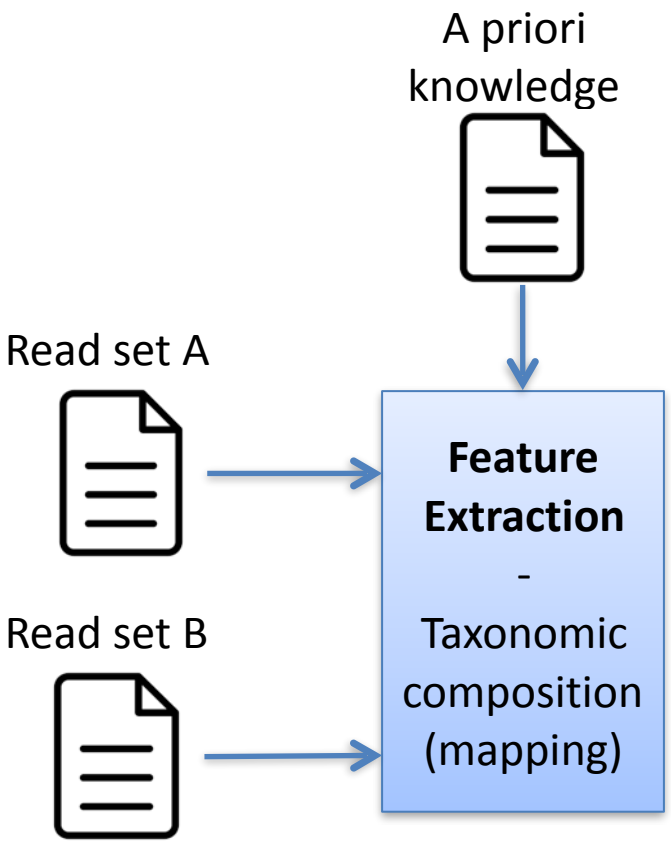
Mahendra Mariadassou

Erwan Drezen

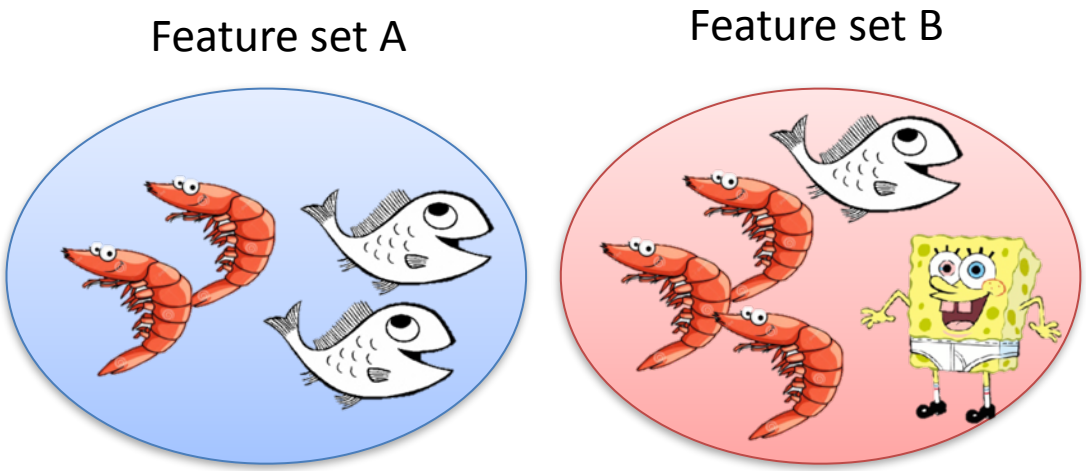
Multiple Comparative Metagenomics using Multiset k-mer Counting

Gaëtan Benoit[1,*], Pierre Peterlongo[1], Mahendra Mariadassou[3], Erwan Drezen[1,4], Sophie Schbath[3], Dominique Lavenier[1], Claire Lemaitre[1]

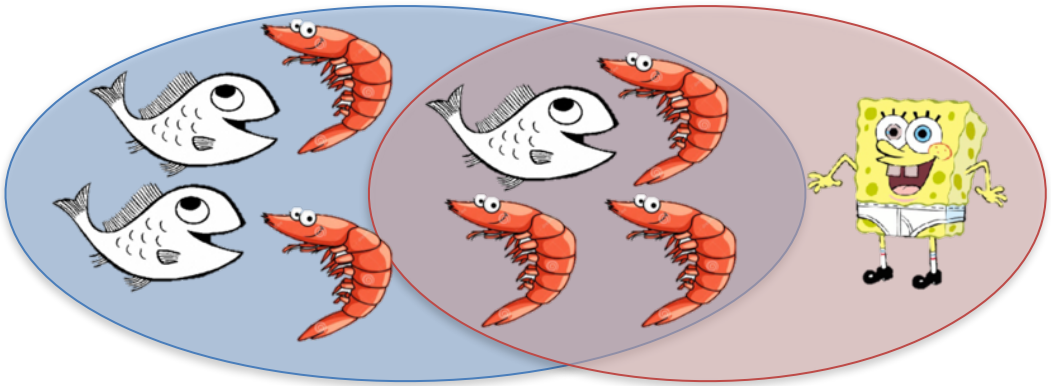
Comparing 2 read sets – using a priori knowledge



Taxonomic composition

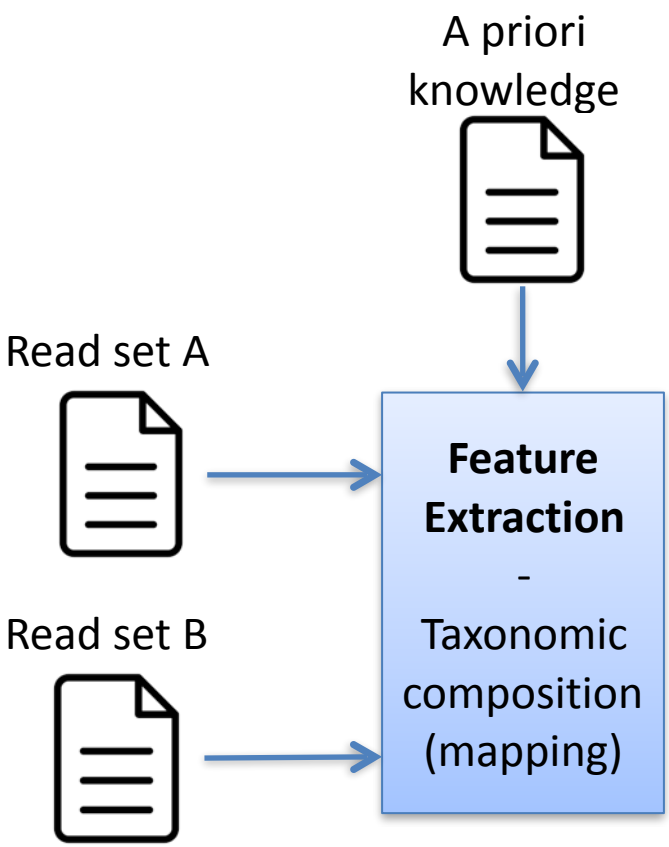


Distance computation



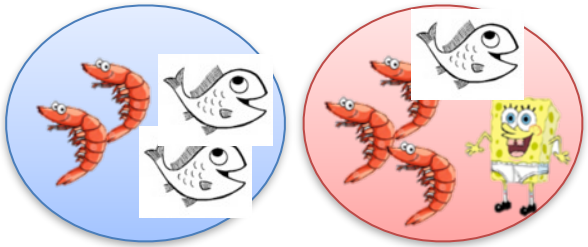
Comparing 2 read sets – using a priori knowledge

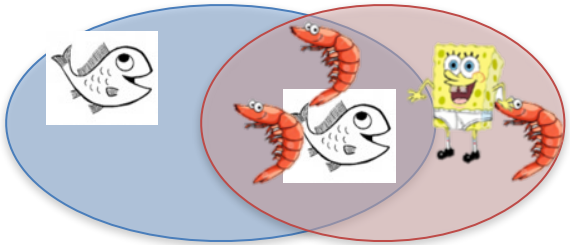
$$Jaccard(A,B) = \underbrace{\sum_{s \in A \cap B} S_A + S_B}_{\text{n (intersection)}} / \underbrace{\sum_{s \in A \cup B} S_A + S_B}_{\text{U (union)}}$$



Taxonomic composition
Abundance table

	A	B
	2	1
	2	3
	0	1





Jaccard(A,B) = 4 / 6

Comparing N read sets – using a priori knowledge

A priori knowledge

Read set A

Read set B

Read set N

Feature Extraction
-
Taxonomic composition (mapping)

Taxonomic composition
Abundance table

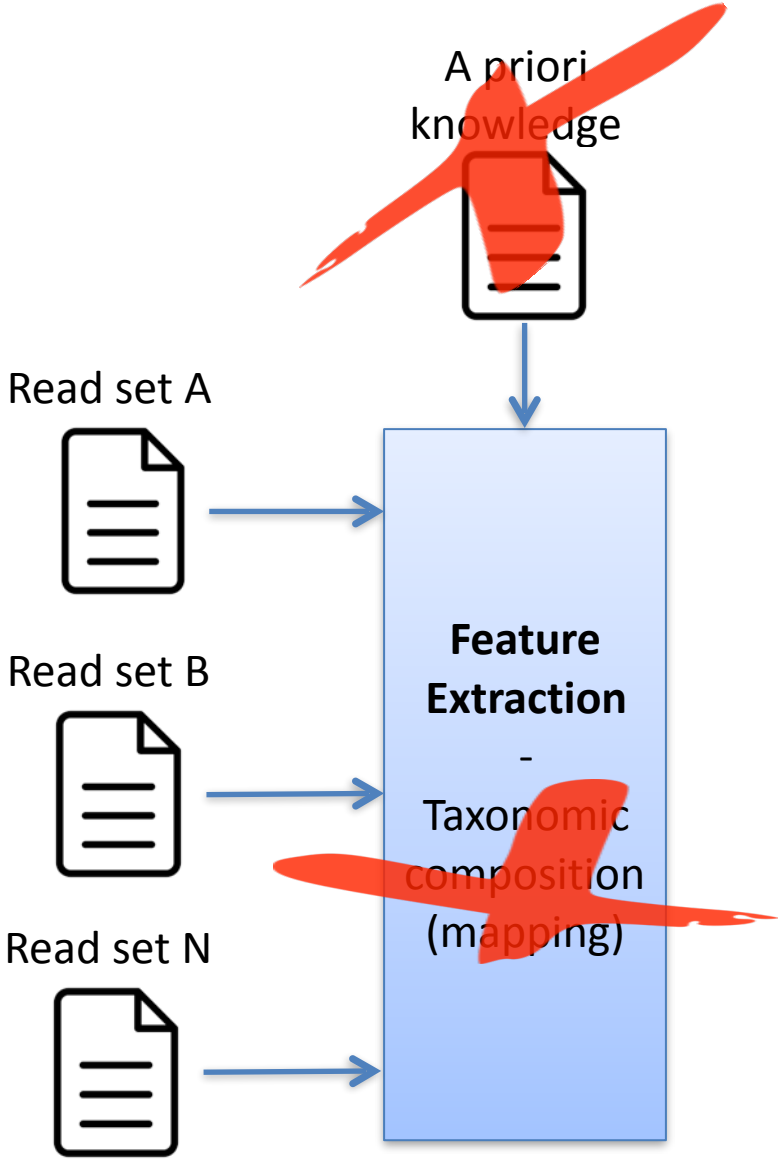
	A	B	N
	2	0	0
	2	3	3
	0	1	1

N read sets:
N x nb features

	A	B	N
A		0.3	0.1
B			0.5
N			

Symmetrical square distance matrix of size N

Comparing N read sets – ***de novo***

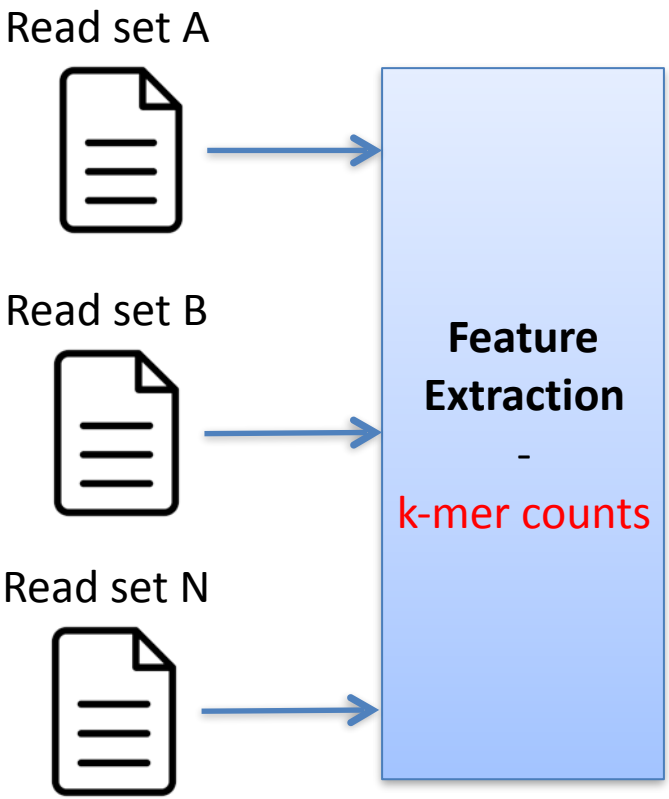


	A	B	N
?	?	?	?
?	?	?	?
?	?	?	?

	A	B	N
A		?	?
B			?
N			

Symmetrical square distance matrix of size N

Comparing N read sets – ***de novo***



	A	B	N
?	?	?	?
?	?	?	?
?	?	?	?

Comparing N read sets – ***de novo***

Kmer composition
Abundance table

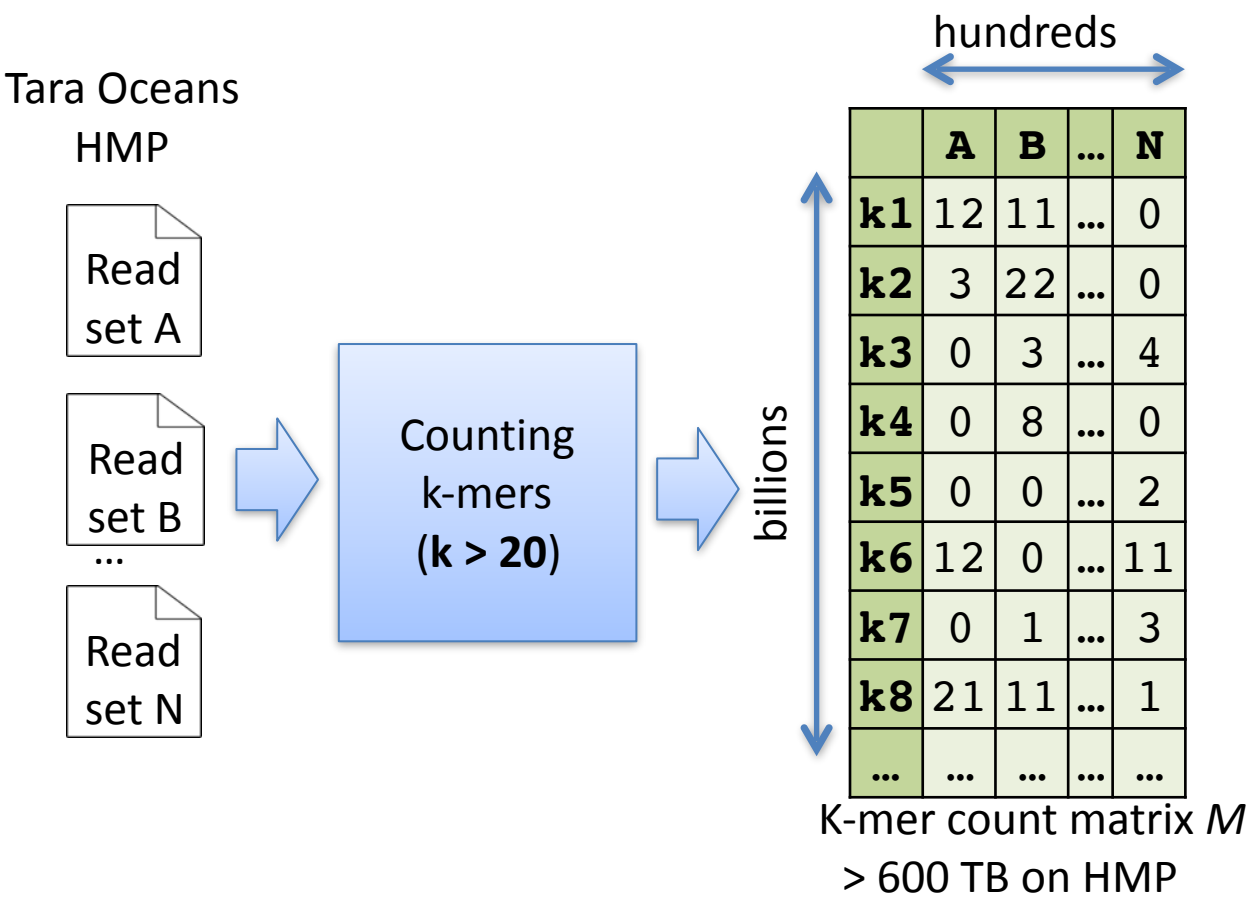
	A	B	N
ACGAG	2	4	0
CGAGC	2	1	9
GAGCT	0	0	5

Read set A

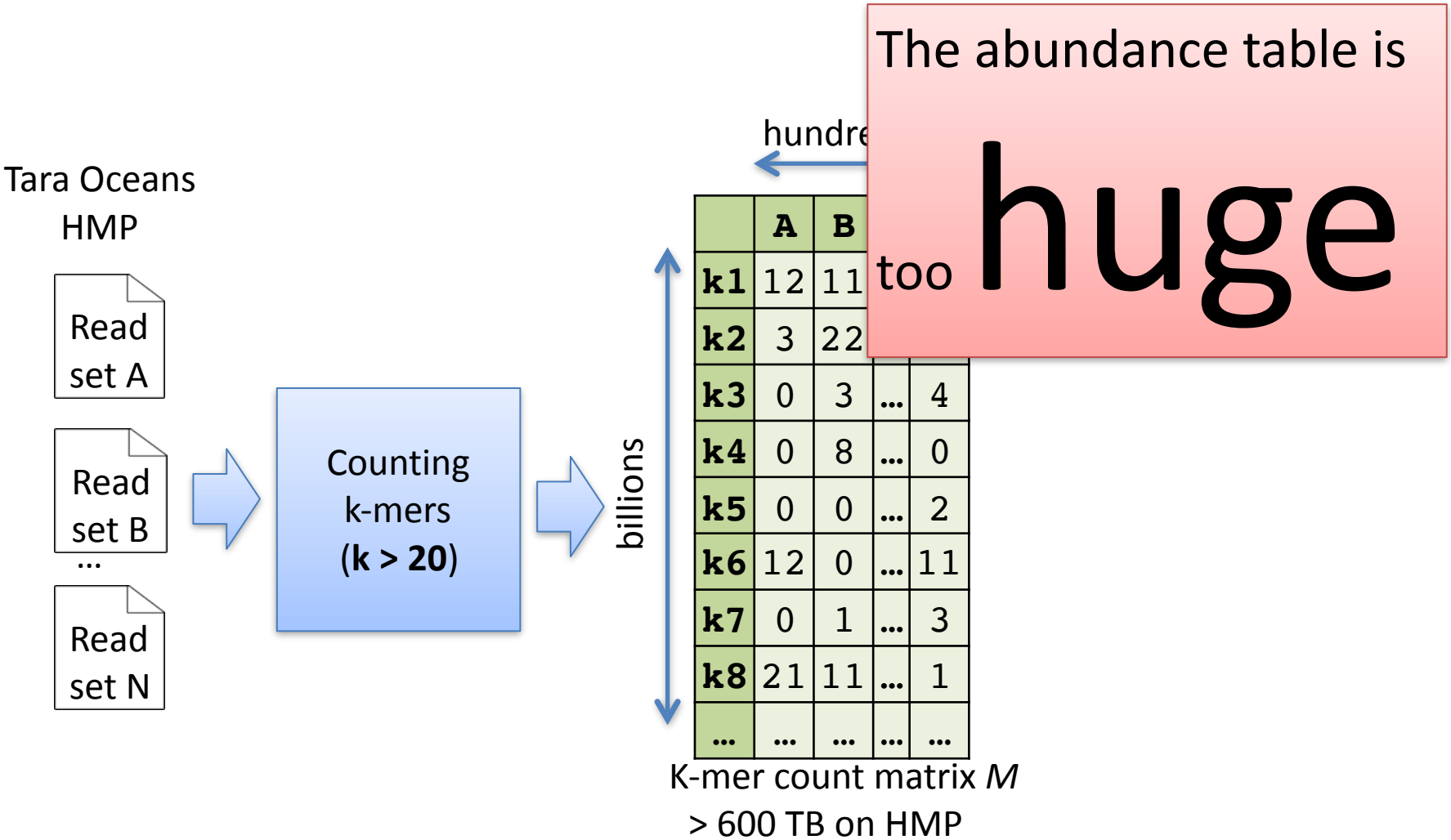
Read set B

Read set N

Issues


Tara Oceans
HMP
Read set A
Read set B
...
Read set N
Counting k-mers (k > 20)
hundreds
billions
A B ... N
k1 12 11 ... 0
k2 3 22 ... 0
k3 0 3 ... 4
k4 0 8 ... 0
k5 0 0 ... 2
k6 12 0 ... 11
k7 0 1 ... 3
k8 21 11 ... 1
... ...
K-mer count matrix *M*
> 600 TB on HMP

Issues


The abundance table is
too huge
hundre
Tara Oceans
HMP
Read set A
Read set B
...
Read set N
Counting k-mers (k > 20)
billions
A
B
k1 12 11
k2 3 22
k3 0 3 ... 4
k4 0 8 ... 0
k5 0 0 ... 2
k6 12 0 ... 11
k7 0 1 ... 3
k8 21 11 ... 1
... ...
K-mer count matrix M
> 600 TB on HMP

Ecology distance computation

Most of the ecology distances are **additive over the lines** of the abundance table

$$Jaccard(A,B) = \frac{\sum_{s \in A \cap B} S_A + S_B}{\sum_{s \in A \cup B} S_A + S_B}$$

	A	B
k1	2	2
k2	1	0
k3	0	1
k4	1	1

Distance computation
(n=intersection u=union)
n=2 u=2

Ecology distance computation

Most of the ecology distances are **additive over the lines** of the abundance table

$$Jaccard(A,B) = \frac{\sum_{s \in A \cap B} S_A + S_B}{\sum_{s \in A \cup B} S_A + S_B}$$

	A	B
k1	2	2
k2	1	0
k3	0	1
k4	1	1

Distance computation
(n=intersection u=union)

N=2 u=2

N=2 u=3

Ecology distance computation

Most of the ecology distances are **additive over the lines** of the abundance table

$$Jaccard(A,B) = \frac{\sum_{s \in A \cap B} S_A + S_B}{\sum_{s \in A \cup B} S_A + S_B}$$

	A	B
k1	2	2
k2	1	0
k3	0	1
k4	1	1

Distance computation
(n=intersection u=union)

n=2 u=2
n=2 u=3
n=2 u=4

Ecology distance computation

Most of the ecology distances are **additive over the lines** of the abundance table

$$Jaccard(A, B) = \frac{\sum_{s \in A \cap B} S_A + S_B}{\sum_{s \in A \cup B} S_A + S_B}$$

Distances can be computed **one line at a time**

	A	B
k1	2	2
k2	1	0
k3	0	1
k4	1	1

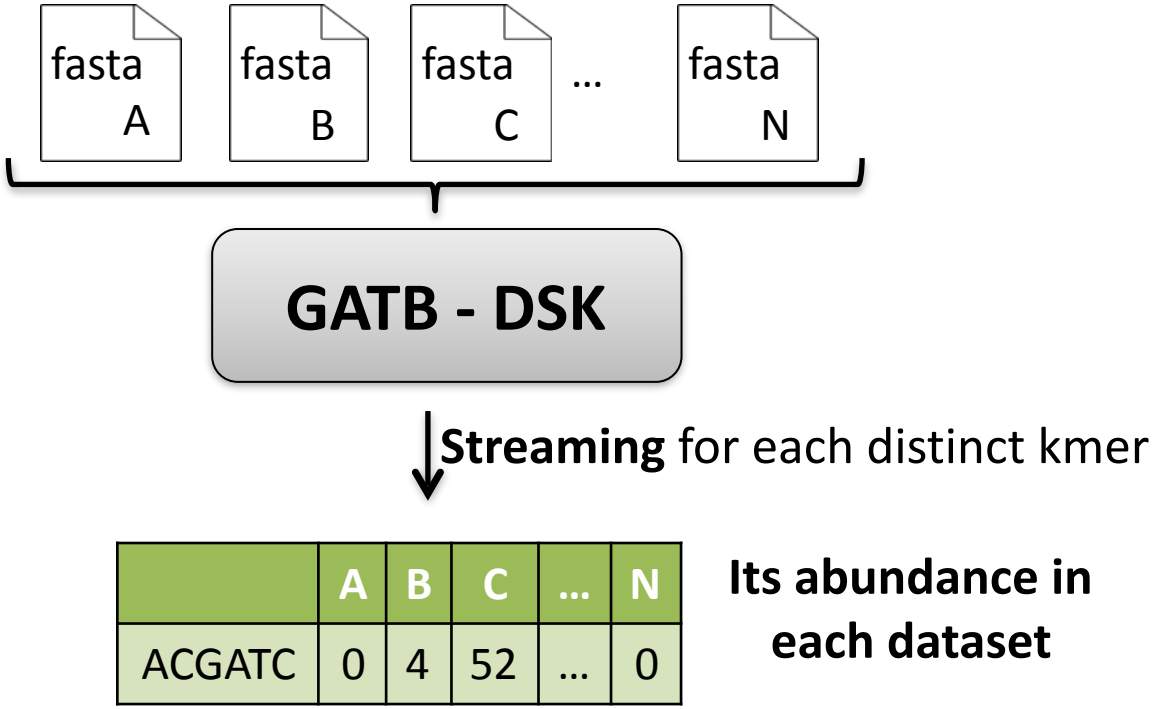
Distance computation
(n=intersection u=union)

n=2	u=2
n=2	u=3
n=2	u=4
n=3	u=5

Jaccard(A, B) = 3 / 5

Multiset kmer counting

- Count the kmers of N datasets **simultaneously**
  - Based on KMC2 algorithm (Deorowicz *et al.* Bioinformatics 2015)
  - Available in GATB library (Drezen et al. Bioinformatics 2014)



From reads to counted kmers

GATB - DSK

basic idea

Can't be more simple:

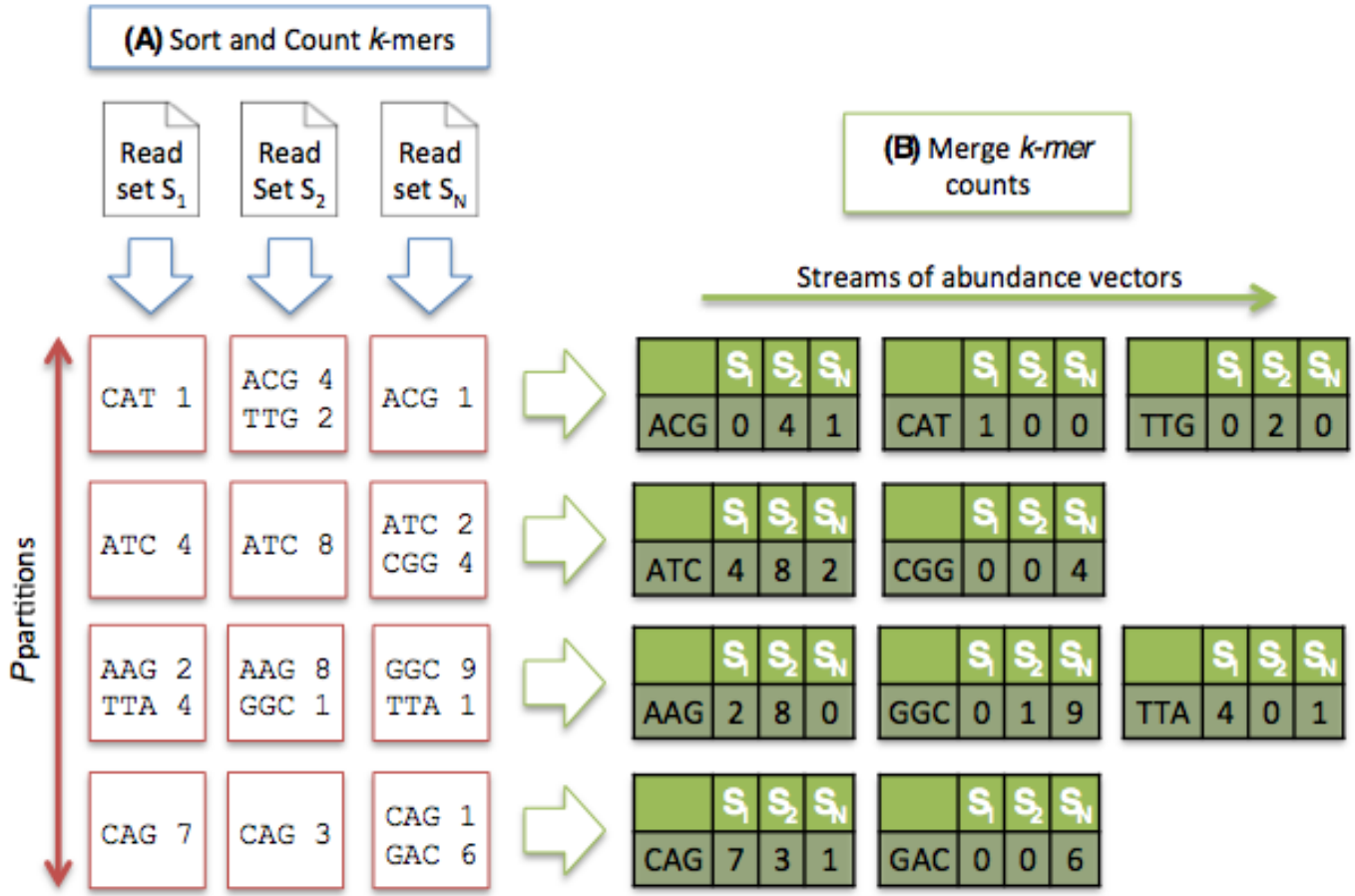
1. Read *k*mers and write them in a file
2. Sort file
3. Identical kmers occur consecutively, count them

1.
```
CAGG
ACGG
CAGG
TTAC
CAGG
ACGG
```

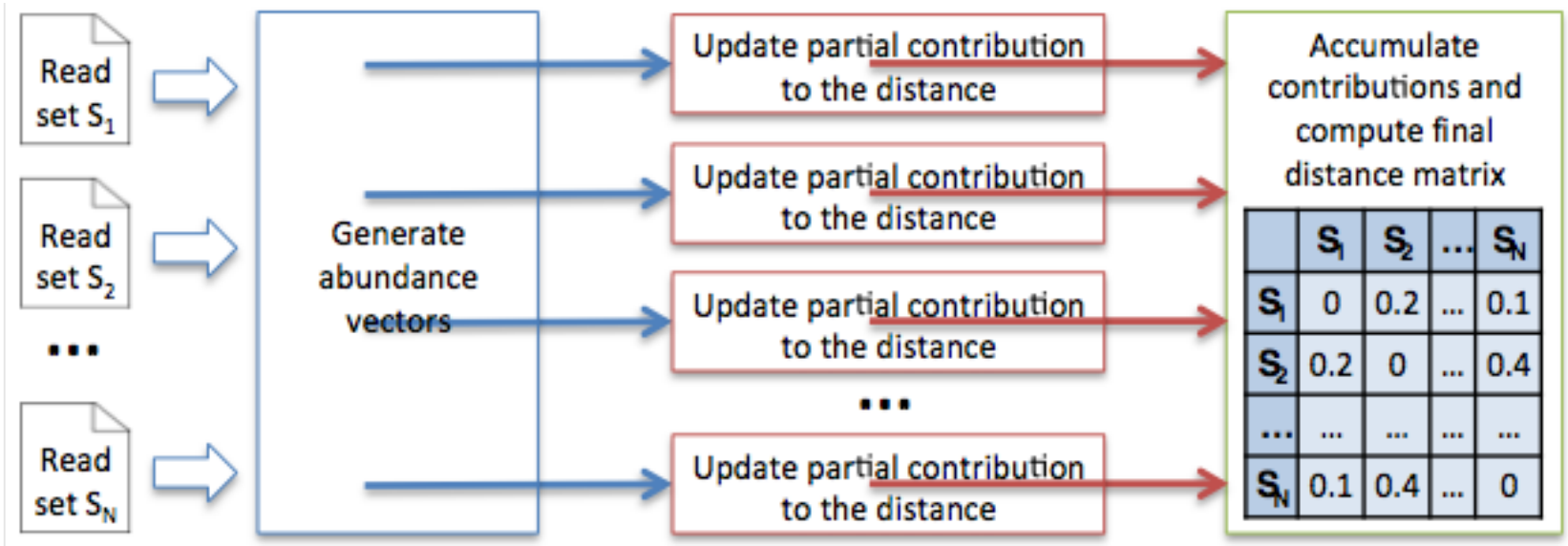
2.
```
ACGG
ACGG
CAGG
CAGG
CAGG
TTAC
```

3.
```
ACGG 2
CAGG 3
TTAC 1
```

GATB - DSK


(A) Sort and Count k-mers
Read set S1
Read Set S2
Read set SN
(B) Merge k-mer counts
Streams of abundance vectors
P partitions
CAT 1
ACG 4
TTG 2
ACG 1
ATC 4
ATC 8
ATC 2
CGG 4
AAG 2
TTA 4
AAG 8
GGC 1
GGC 9
TTA 1
CAG 7
CAG 3
CAG 1
GAC 6
S1 S2 SN
ACG 0 4 1
CAT 1 0 0
TTG 0 2 0
ATC 4 8 2
CGG 0 0 4
AAG 2 8 0
GGC 0 1 9
TTA 4 0 1
CAG 7 3 1
GAC 0 0 6

Dealing with a **streaming** of abundance vectors



Simka performances

Full HMP project (690 samples, 32 billions reads)

- Computation time: < 14h
- 64 GB memory
- ~1T disk (< half of input size)

Simka computes several distances

Name	Definition	C_{S_i}	$f(x,y,X,Y)$	$g(x)$
Quantitative distances				
Chord	$\sqrt{2-2\sum_w \frac{N_{S_i}(w)N_{S_j}(w)}{C_{S_i}C_{S_j}}}$	$\sqrt{\sum_w N_{S_i}(w)^2}$	$\frac{xy}{XY}$	$\sqrt{2-2x}$
Hellinger	$\sqrt{2-2\sum_w \frac{\sqrt{N_{S_i}(w)N_{S_j}(w)}}{\sqrt{C_{S_i}C_{S_j}}}}$	$\sum_w N_{S_i}(w)$	$\frac{\sqrt{xy}}{\sqrt{XY}}$	$\sqrt{2-2x}$
Whittaker	$\frac{1}{2}\sum_w \frac{\lvert N_{S_i}(w)C_{S_j} - N_{S_j}(w)C_{S_i}\rvert}{C_{S_i}C_{S_j}}$	$\sum_w N_{S_i}(w)$	$\frac{\lvert xY - yX\rvert}{XY}$	$\frac{x}{2}$
Bray-Curtis	$\sum_w \frac{\lvert N_{S_i}(w) - N_{S_j}(w)\rvert}{C_{S_i}+C_{S_j}}$	$\sum_w N_{S_i}(w)$	$\frac{\lvert x-y\rvert}{X+Y}$	x
Kulczynski	$1-\frac{1}{2}\sum_w \frac{(C_{S_i}+C_{S_j})\min(N_{S_i}(w),N_{S_j}(w))}{C_{S_i}C_{S_j}}$	$\sum_w N_{S_i}(w)$	$\frac{(X+Y)\min(x,y)}{XY}$	$1-\frac{x}{2}$
Jensen-Shannon	$\sqrt{\frac{1}{2}\sum_w \left[\frac{N_{S_i}(w)}{C_{S_i}}\log\frac{2C_{S_j}N_{S_i}(w)}{C_{S_j}N_{S_i}(w)+C_{S_i}N_{S_j}(w)} + \frac{N_{S_j}(w)}{C_{S_j}}\log\frac{2C_{S_i}N_{S_j}(w)}{C_{S_j}N_{S_i}(w)+C_{S_i}N_{S_j}(w)}\right]}$	$\sum_w N_{S_i}(w)$	$\frac{x}{X}\log\frac{2xY}{xY+yX}+\frac{y}{Y}\log\frac{2yX}{xY+yX}$	$\sqrt{\frac{x}{2}}$
Canberra	$\frac{1}{a+b+c}\sum_w \left\lvert\frac{N_{S_i}(w)-N_{S_j}(w)}{N_{S_i}(w)+N_{S_j}(w)}\right\rvert$	–	$\left\lvert\frac{x-y}{x+y}\right\rvert$	$\frac{1}{a+b+c}x$
Jaccard	$1-\sum_w \frac{(N_{S_i}(w)+N_{S_j}(w))1_{\{N_{S_i}(w)N_{S_j}(w)>0\}}}{C_{S_i}+C_{S_j}}$	$\sum_w N_{S_i}(w)$	$\frac{(x+y)1_{\{xy>0\}}}{X+Y}$	x
Qualitative distances				
Chord/Hellinger	$\sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)}$	–	–	–
Whittaker	$\frac{1}{2}\left(\frac{b}{a+b}+\frac{c}{a+c}+\left\lvert\frac{a}{a+b}-\frac{a}{a+c}\right\rvert\right)$	–	–	–
Bray-Curtis/Sorensen	$\frac{b+c}{2a+b+c}$	–	–	–
Kulczynski	$1-\frac{1}{2}\left(\frac{a}{a+b}-\frac{a}{a+c}\right)$	–	–	–
Ochiai	$1-\frac{a}{\sqrt{(a+b)(a+c)}}$	–	–	–
Jaccard	$\frac{b+c}{a+b+c}$	–	–	–
Abundance-based (AB) variants of qualitative distances				
AB-Jaccard	$1-\frac{UV}{U+V-UV}$	–	–	–

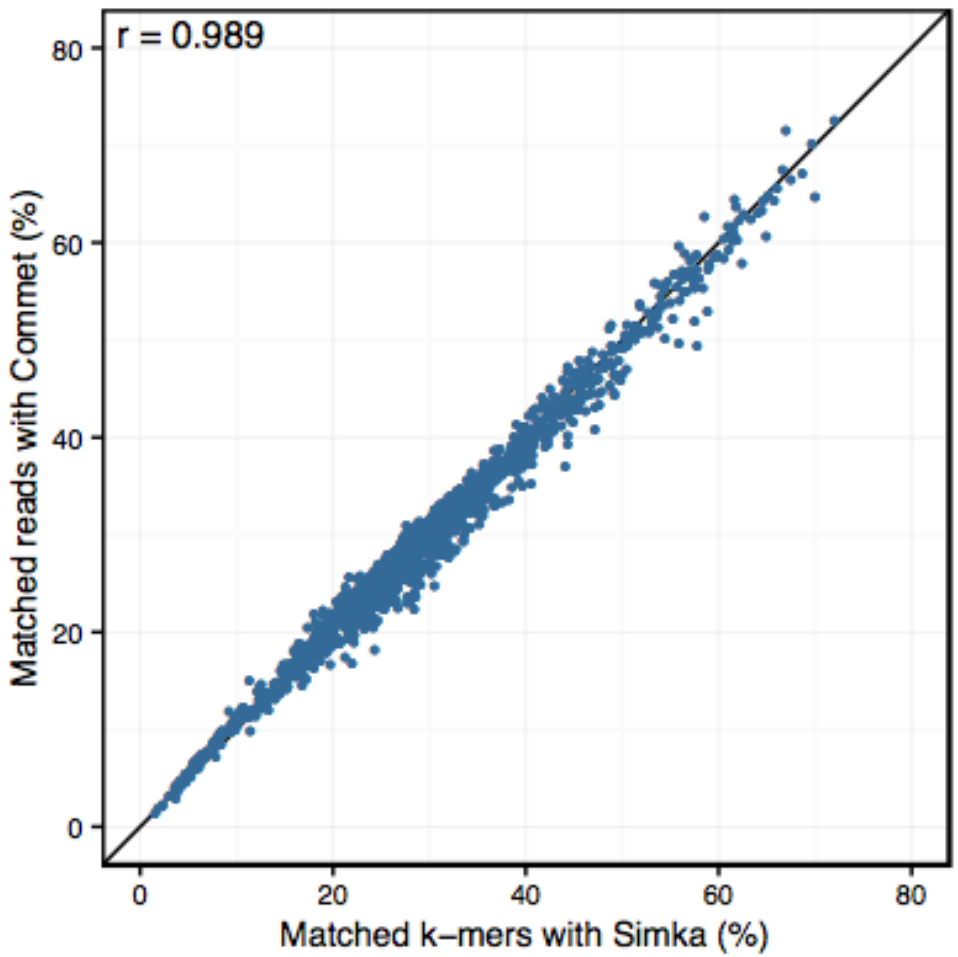
Are kmer distances good enough?

```
TACGGGACTGAT-CAGACGTCAA
 ||||   ||||| || || ||||
 ACGG--CTGATTCATACTTCAAGG
```

alignment → Similarity = **74%**

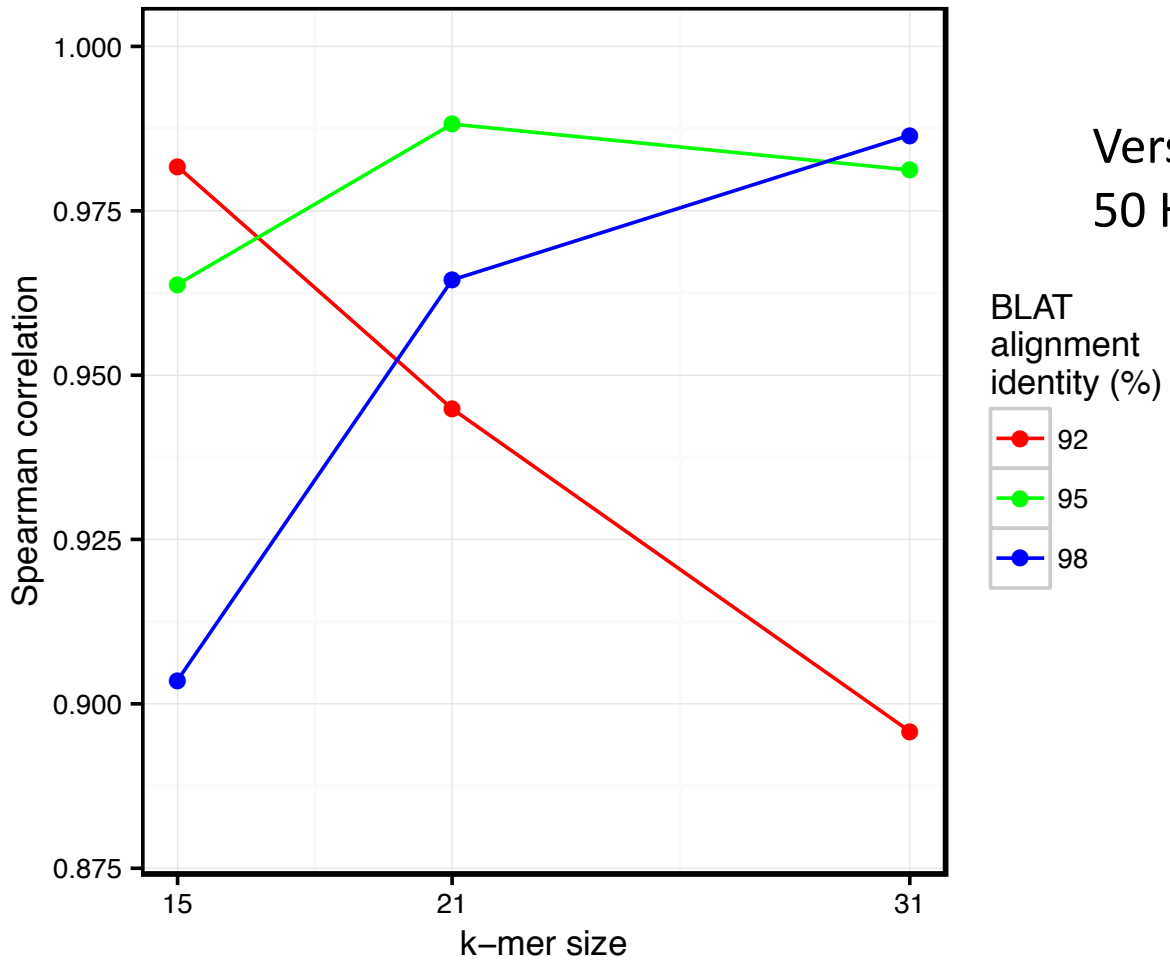
k-mer based → **56%** of positions covered by a shared kmer

Are kmer distances good enough?



Versus **Commet**
50 HMP samples (among 690)

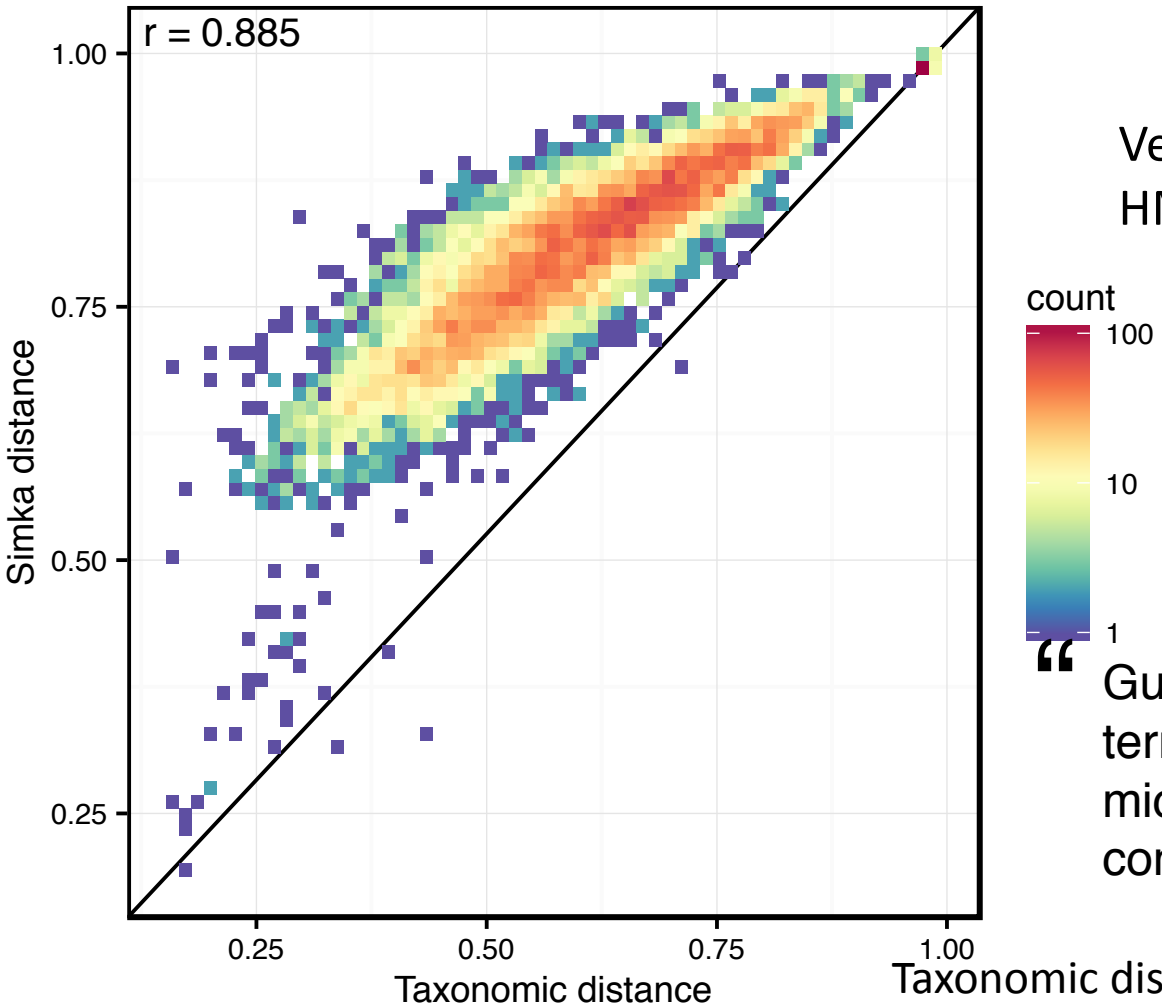
Are kmer distances good enough?



Versus **Blat**

50 HMP samples (among 690)

Are kmer distances good enough?



Versus **Taxonomical dist.**
HMP **GUT** samples

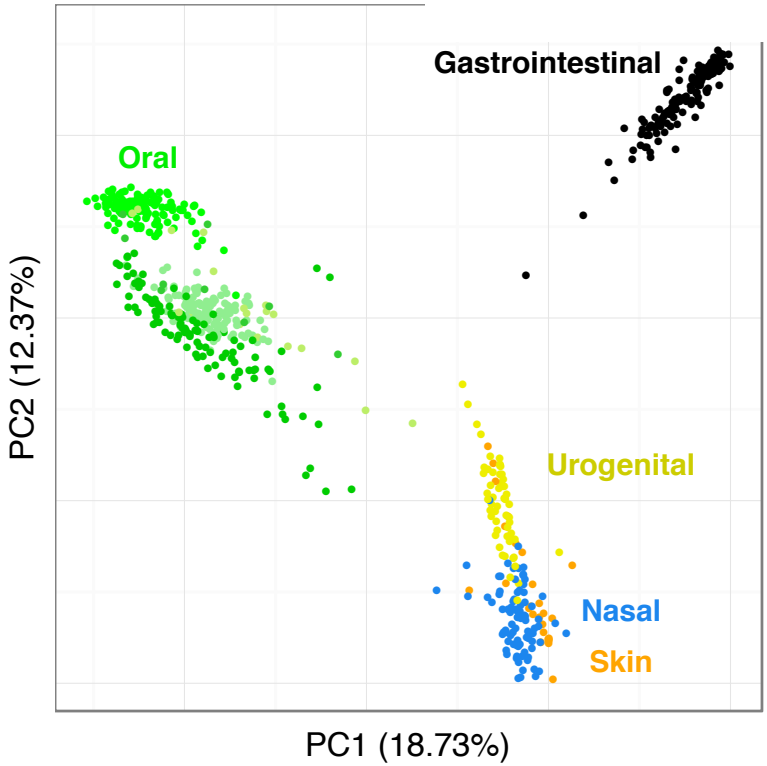
Mash Spearman = 0.51:

“Gut samples differ more in terms of relative abundances of microbes than in terms of composition

Taxonomic distances are obtained from http://www.hmpdacc.org/HMSCP/

Tests on full HMP project

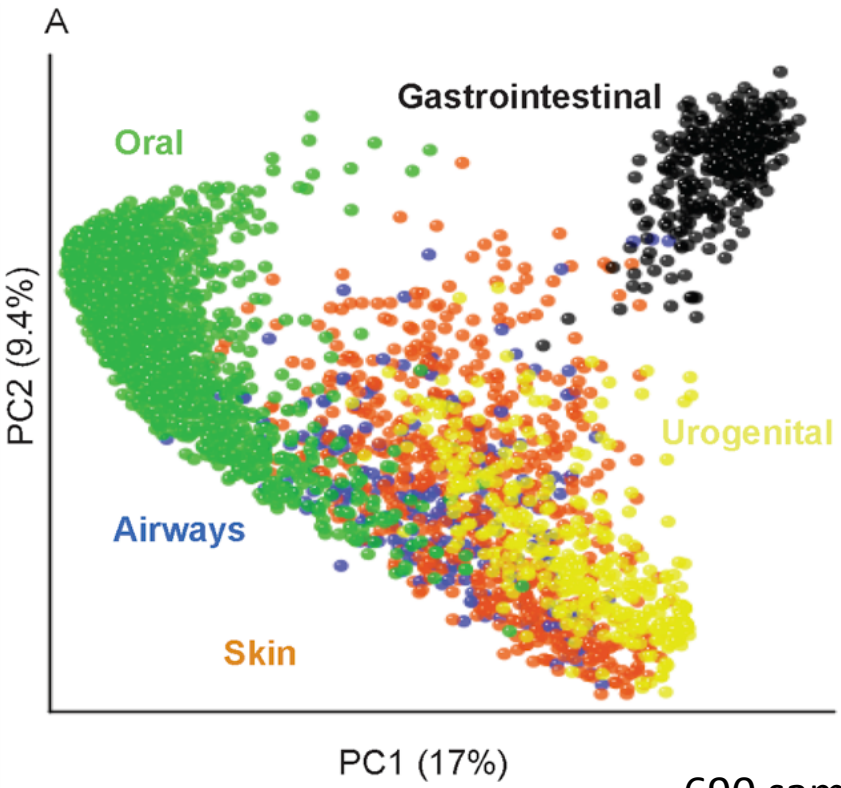
Simka (14h, 62GB)



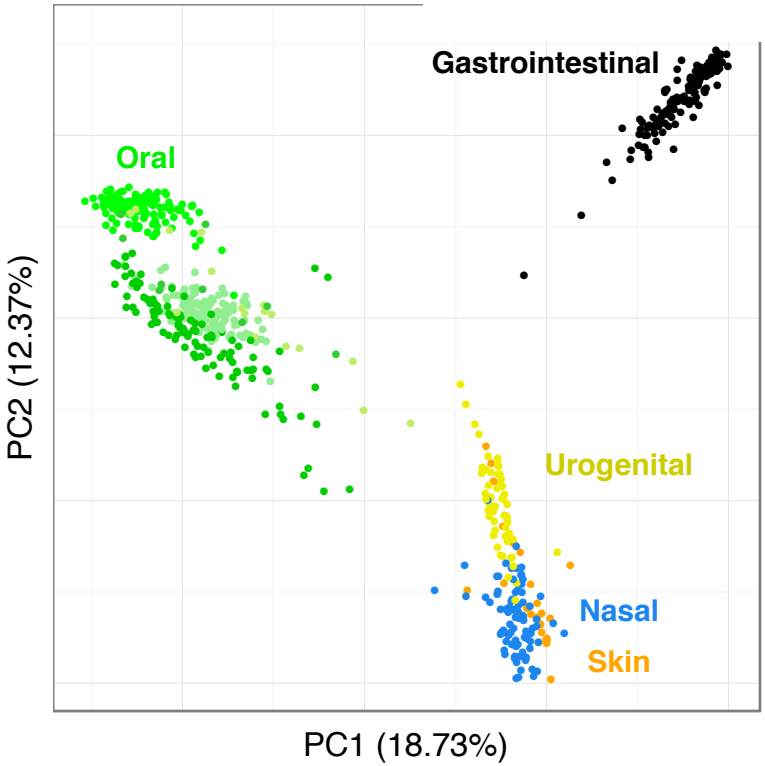
- 690 samples - 3727 GB
- 2x16 billions paired reads

Tests on full HMP project

HMP – (OTU) [1]



Simka (14h, 62GB)

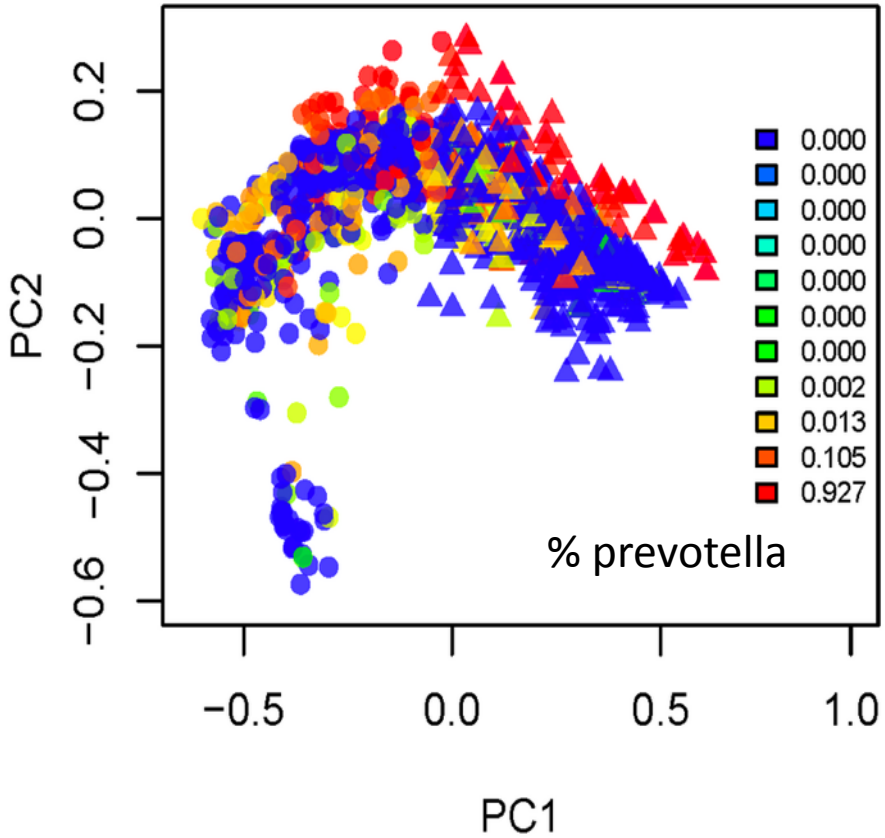


- 690 samples - 3727 GB
- 2x16 billions paired reads

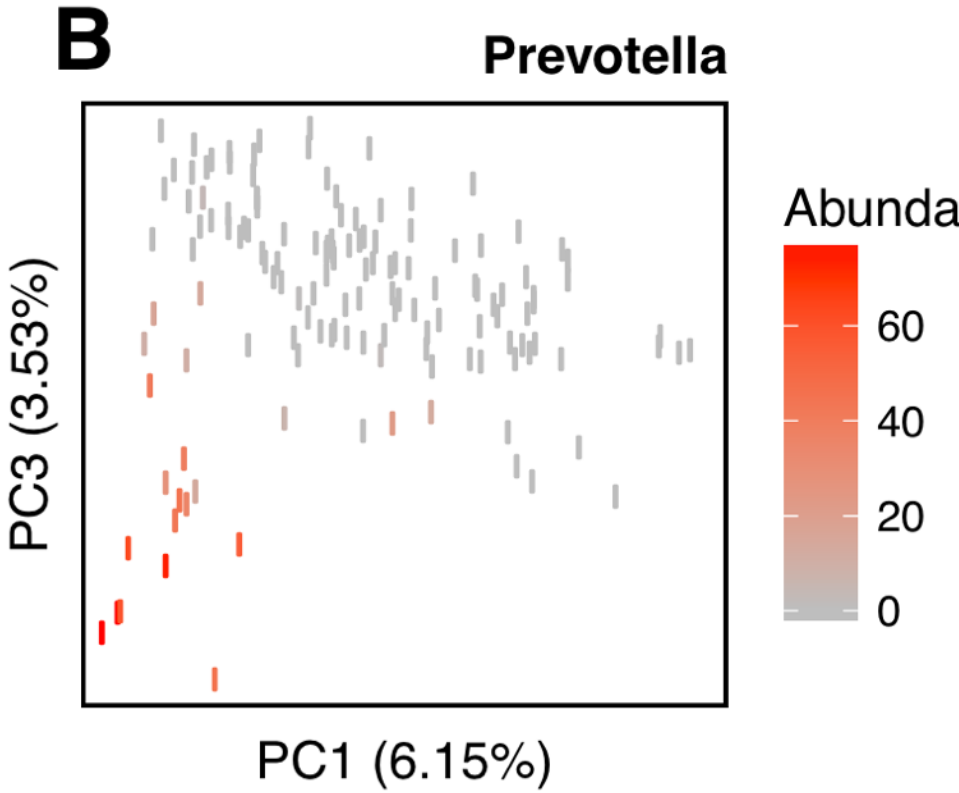
[1] Koren et al 2013
A Guide to Enterotypes across the Human Body […]

Tests on HMP project – GUT enterotypes

HMP (OTU) [1]

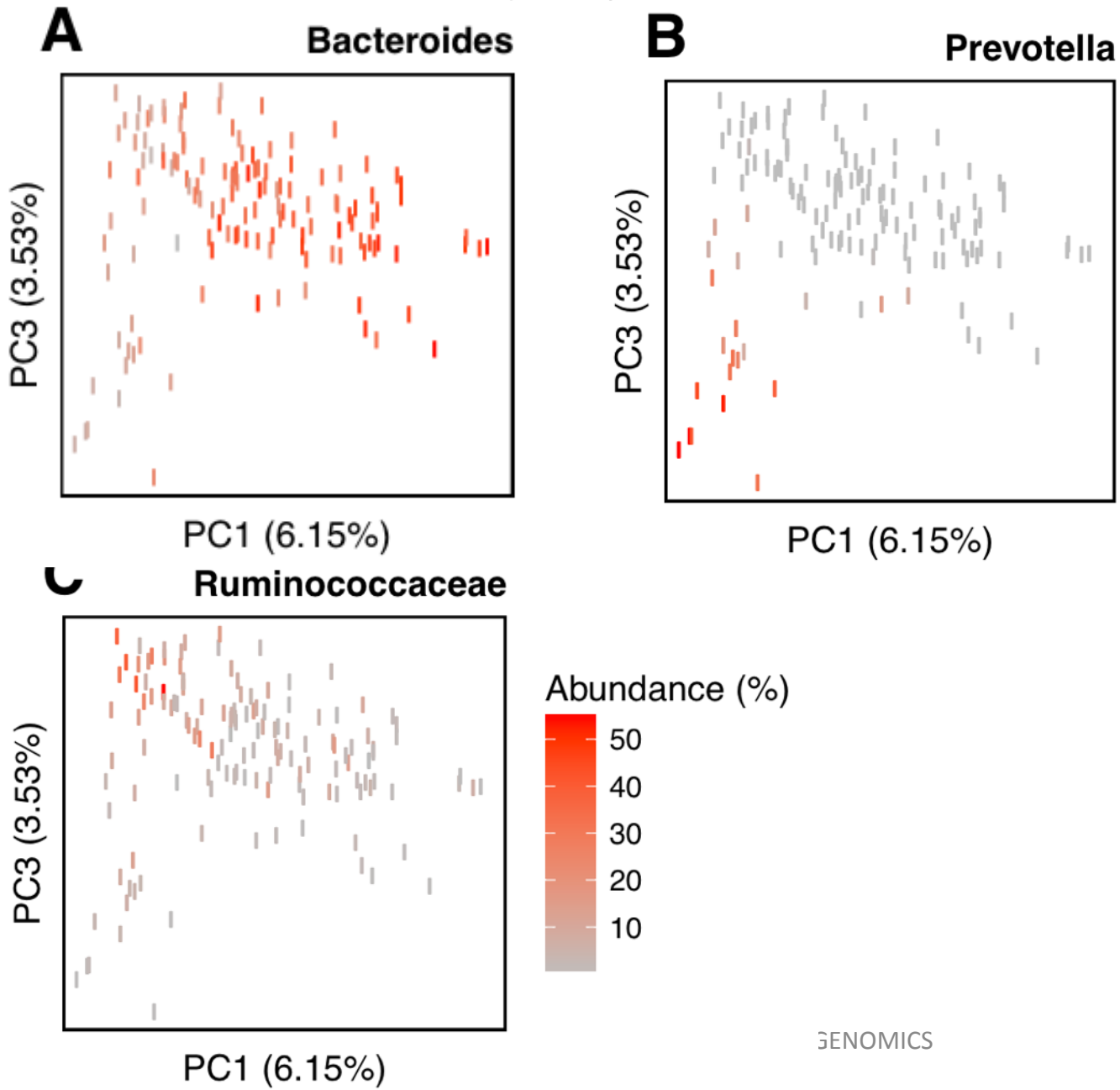


Simka



[1] Koren et al 2013
A Guide to Enterotypes across the Human Body […]

Tests on HMP project – GUT enterotypes



Simka: Take Home Message

- Push button
- Many ecological distances
- Ultra Fast computation
- Acceptable memory footprint

- Does not provide links
- *k*-mers instead of species
  - *under.: horizontal gene*
  - *over.: ≠ genome sizes*

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., & Lemaitre, C.
Multiple Comparative Metagenomics using Multiset k-mer Counting.
arXiv id: 1604.02412 – Peer J. review process –

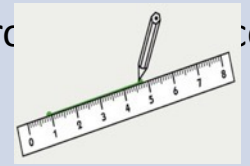
Simka: Take Home Message

- Push button
- Many ecological distances
- Ultra Fast computation
- Acceptable memory footprint
- **Validates *k-mer* based distances**

- Does not provide links
- *k*-mers instead of species
  - *under.: horizontal gene*
  - *over.: ≠ genome sizes*

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., & Lemaitre, C.
Multiple Comparative Metagenomics using Multiset k-mer Counting.
arXiv id: 1604.02412 – Peer J. review process –

Future

	Commet	SRC_linker	Simka	?
Pr  ce	✓	✗	✓	✓
Scale  Mem	✓	✗	✓	✓
Time	✗	✗	✓	✓
	— vs 	— vs —	 vs 	— vs —

Future

Algorithmic

- *k*-mer subsampling
- Dynamic addition of new sets

Applications

- Tara analyses
- …

Thanks!

Tools

- **Commet**

  github.com/pierrepeterlongo/commet

- **Simka**

  github.com/GATB/simka

- **SRC_linker**

  github.com/GATB/rconnector

pierre.peterlongo@inria.fr


Olivier Jaillon


Antoine Limasset


Lucie Bittner


Thomas Vanier


Claire Lemaitre


Mahendra Mariadassou


Nicolas Maillet


Gaetan Benoit


Sophie Schbath


Dominique Lavenier


Guillaume Collet


Erwan Drezen


Camille Marchet

FP & FN pregnancy test example

- False Negative:

  "You are not pregnant"

- False Positive:

  "You are pregnant"