



HAL
open science

Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets

Gaëtan Benoit

► **To cite this version:**

Gaëtan Benoit. Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets. RCAM, Oct 2015, Paris, France. hal-01231795

HAL Id: hal-01231795

<https://inria.hal.science/hal-01231795>

Submitted on 18 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Simka

Fast kmer-based method for estimating the similarity between numerous metagenomic datasets

Gaëtan BENOIT

PHD student - ANR Hydrogen

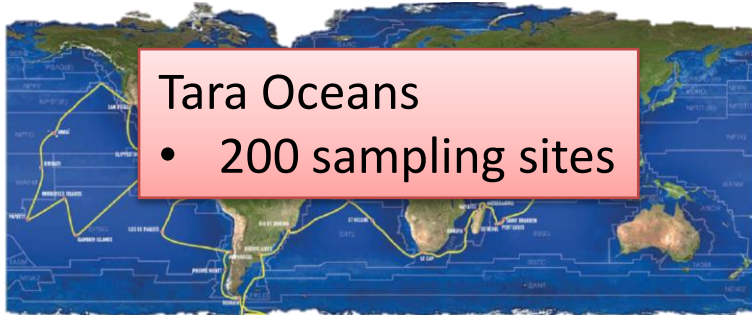
GenScale bioinformatics research group

IRISA/INRIA – Rennes - FRANCE

24/09/2015



Context



Tara Oceans

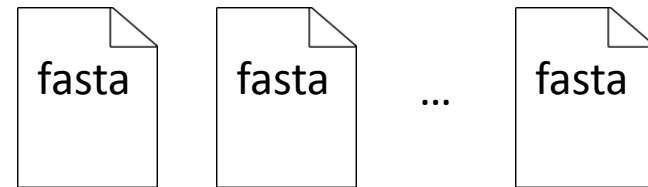
- 200 sampling sites

N metagenomic samples

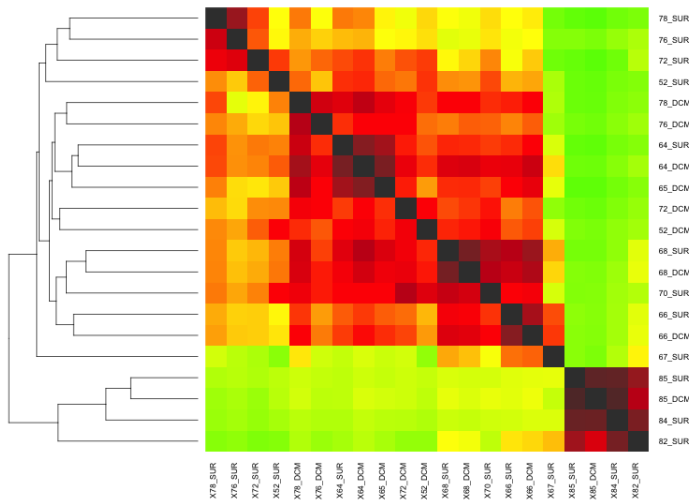


Whole genome sequencing

De novo comparative metagenomics



N MetaG datasets



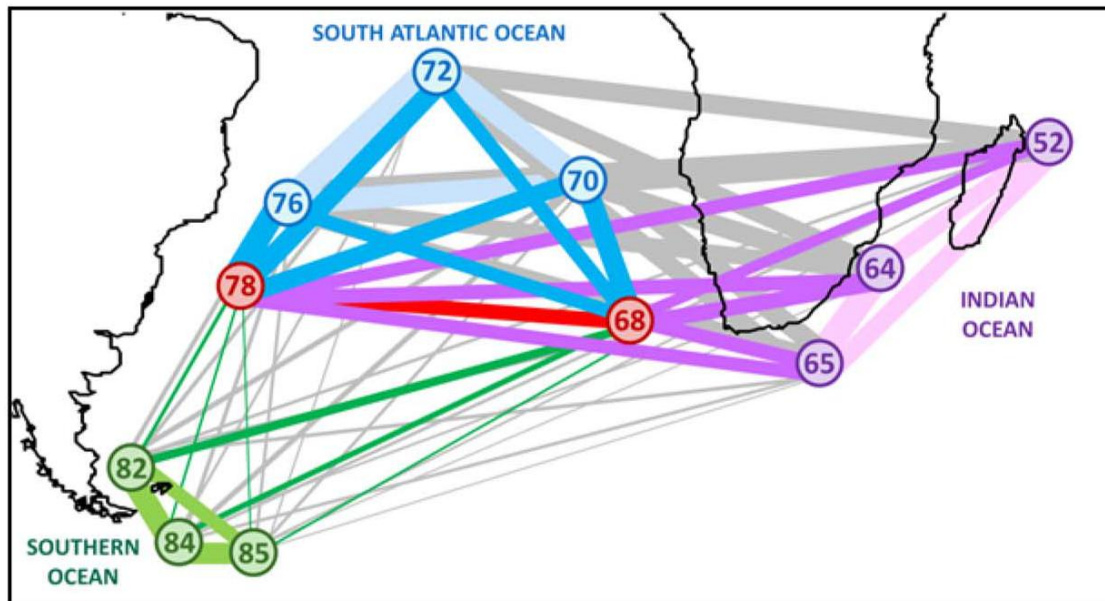
Similarity heatmap N^2

Tara Oceans

- > 1000 datasets
- > 100 millions reads each

De novo comparative metagenomics

- Sea water
 - < 5% assembled reads
 - < 10% known species
- Compare **environmental data** and **genomic contents**
 - Environmental characteristics of Agulhas rings affect interocean plankton transport. Villar *et al.* Science 2015

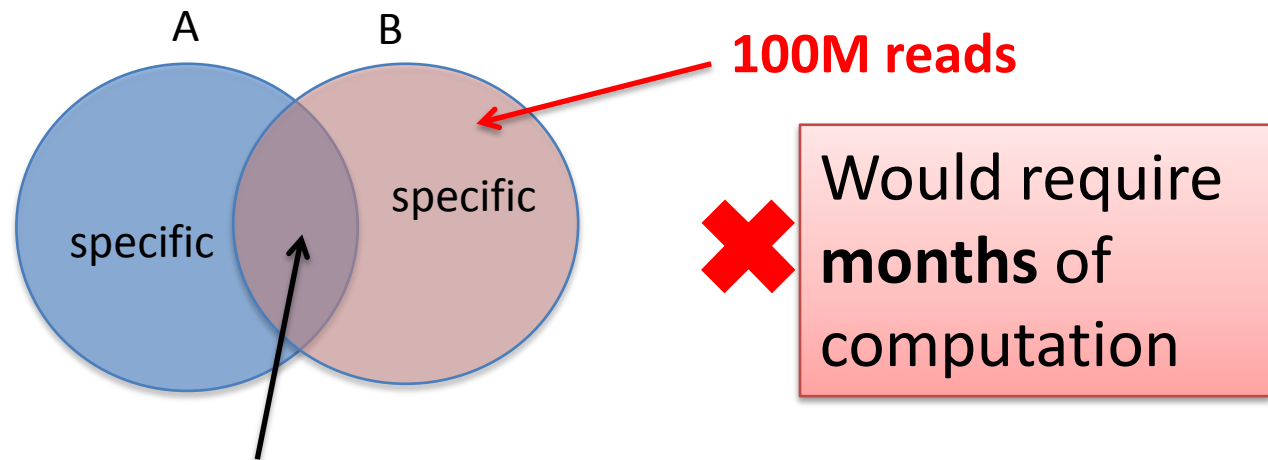


Similarities on

- Barcodes
 - WGS
- same conclusions

Similarity between 2 samples

Idea: Similarity is given by the **size of the intersection**



Intersection = number of **similar reads**

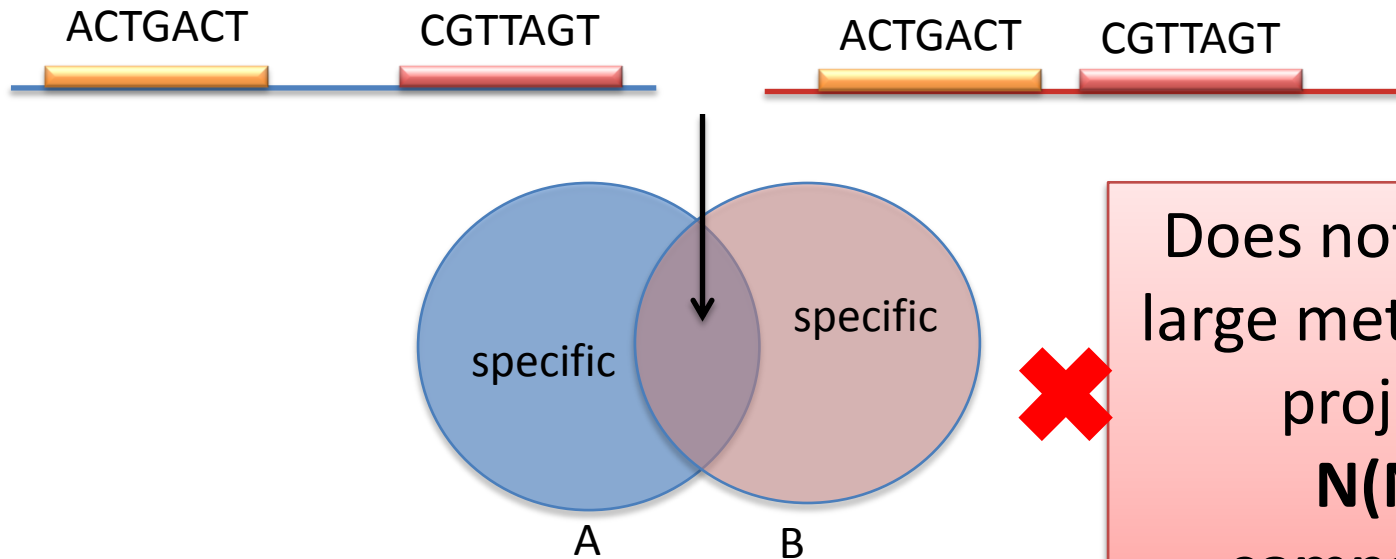
Similar: alignment score > 90%

(Blast-like approaches, all-vs-all sequence alignment)

State of the art

Commet (Maillet *et al.* IEEE BIBM 2014)

- “**similar**”: share at least t non-overlapping ***kmers*** (words of size k)
- Example: $t=2$:



Does not scale on
large metagenomic
projects :
 $N(N-1)$
comparisons

- Computes one intersection in **few hours**
 - Alignment free method
 - Only **indexing** and **querying kmers**

Simka: comparing numerous datasets

New representation of datasets

- **Not** viewed as a **read-set** anymore
- But as a **bag** of its **overlapping kmers**

read-set



read A A A A

A A A A

A A A

Kmers
(k=6)

A A C G A A

In practise k > 30

Output of kmer
counting
algorithm

kmer-bag

all
occurrences
of kmers

A A A A C G

A A A C G A

A A C G A A

...

A A A A C G



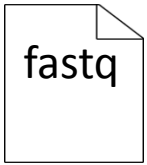
A A A A C G 2

A A A C G A 1

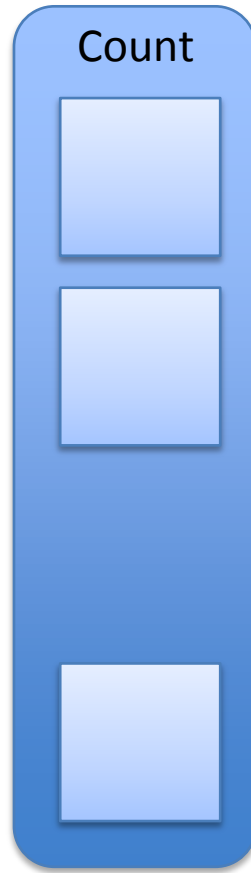
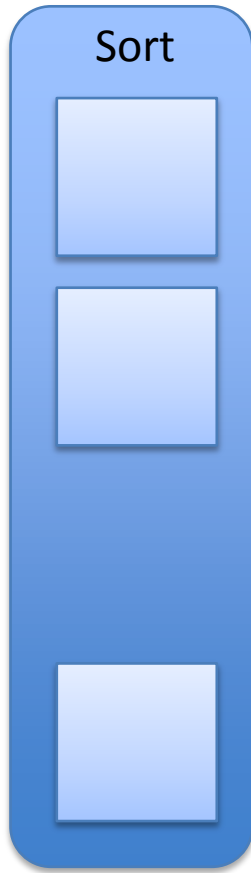
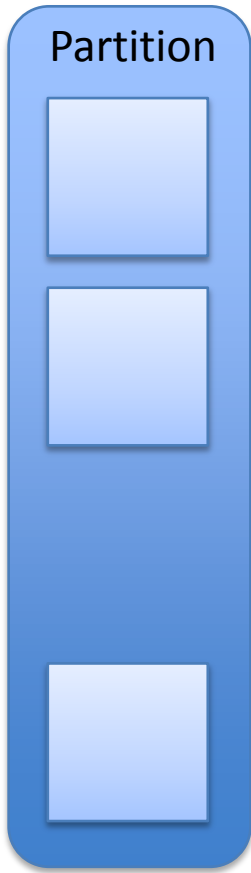
A A C G A A 1

...

Distinct kmers
with their abundance



A



B

Simka: comparing numerous datasets

New similarity function

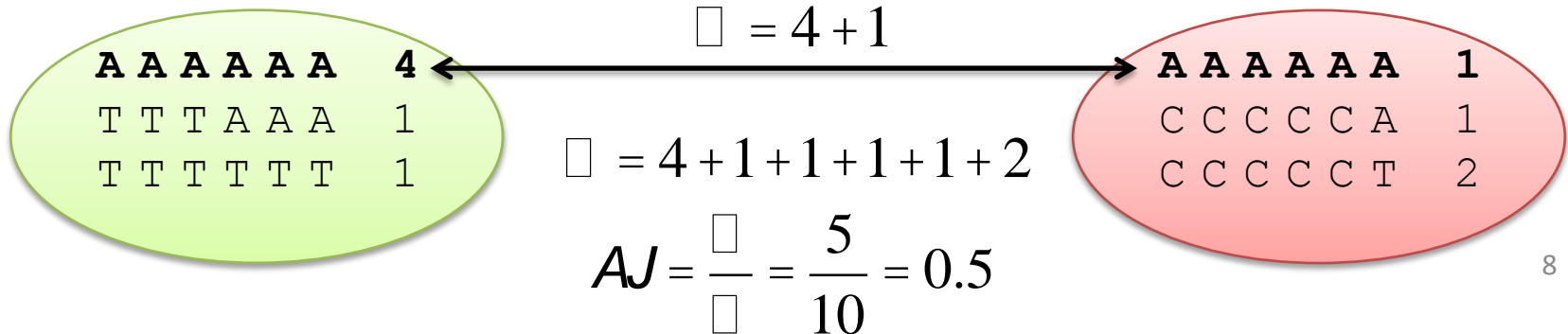
- Pairwise similarity measure based on **shared kmers**

Abundance-based Jaccard similarity

$$AJ(A, B) = \frac{\sum_{w \in \hat{A} \cap \hat{B}} N_A(w) + N_B(w)}{\sum_{w \in \hat{A} \cup \hat{B}} N_A(w) + N_B(w)}$$

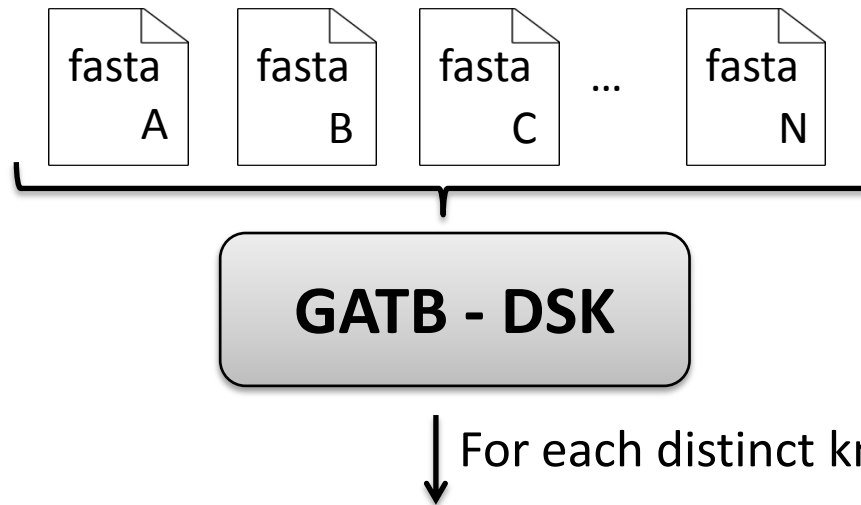
We need

- Fast kmer counting algorithm
- Compute faster the $N(N-1)$ comparisons



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)



	A	B	C	...	N
ACGATC	0	4	5	...	0

Its abundance in each dataset

- We can now count the kmers of large datasets quickly
- Compute all the intersections simultaneously

Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

Simka algorithm

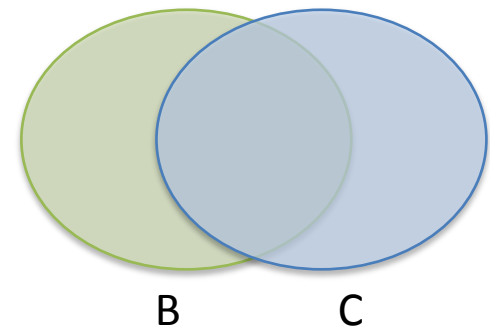
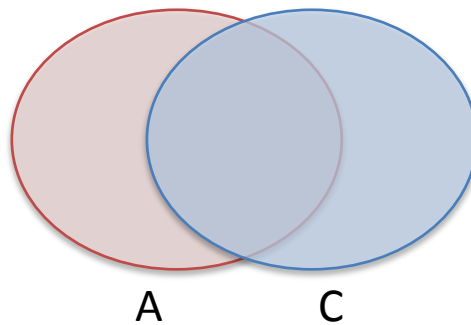
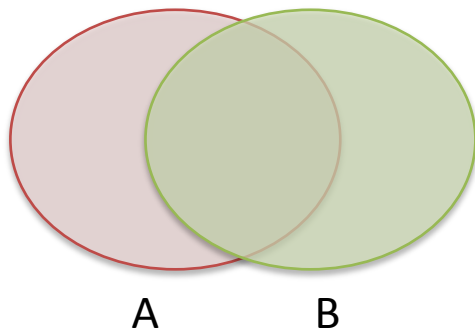
- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8

For each pair(i, j)



Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

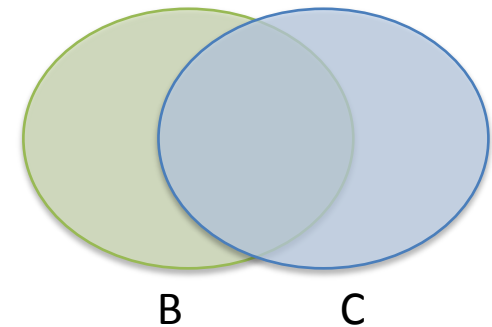
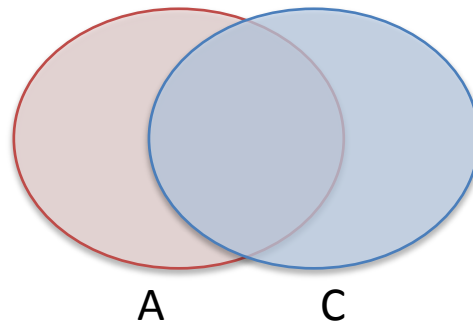
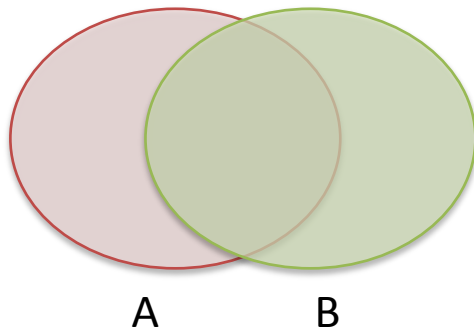
Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8



For each pair (i, j)

If kmer is shared $N_i > 0$ and $N_j > 0$



Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

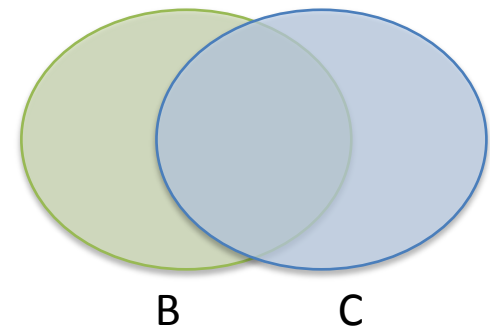
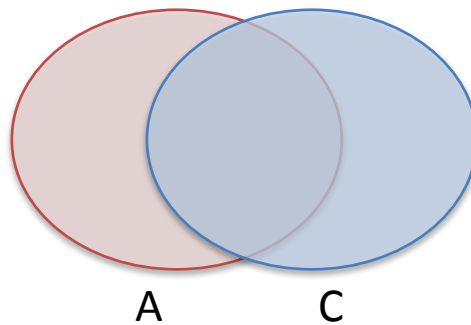
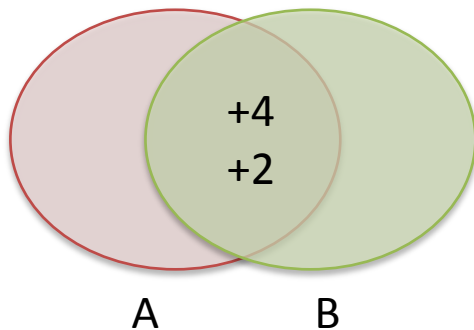
Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8

For each pair (i, j)

If kmer is shared $N_i > 0$ and $N_j > 0$

Update intersection $n[i, j] += N_i + N_j$



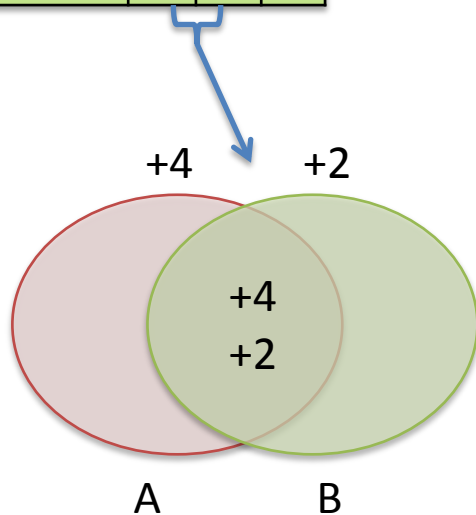
Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8



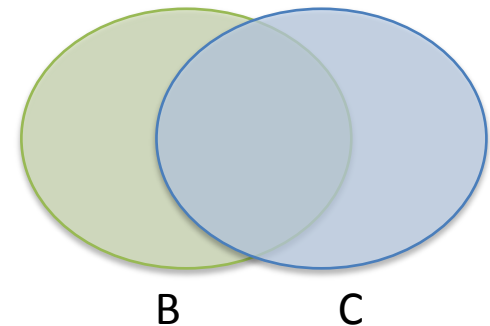
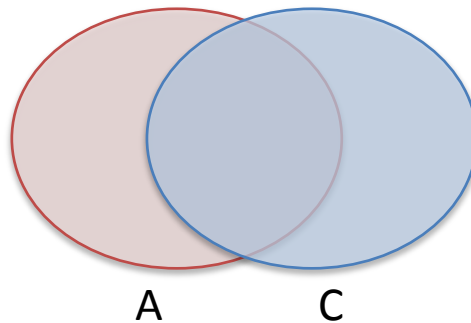
$$AJ(A, B) = 6/6 = 1$$

For each pair (i, j)

If kmer is shared $N_i > 0$ and $N_j > 0$

Update intersection $n[i, j] += N_i + N_j$

Update union $u[i, j] += N_i + N_j$



Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

Kmer shared by A, B and C

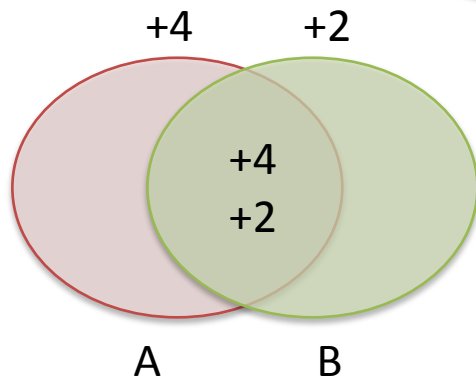
	A	B	C
ACGATC	4	2	8

For each pair(i, j)

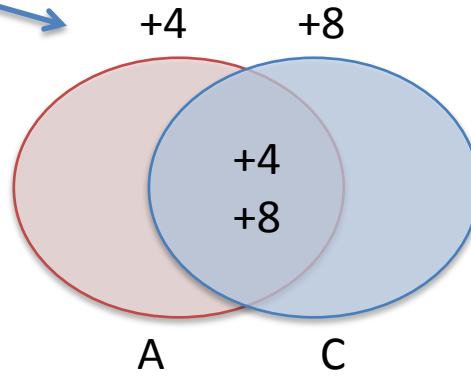
If kmer is shared $N_i > 0$ and $N_j > 0$

Update intersection $n[i,j] += N_i + N_j$

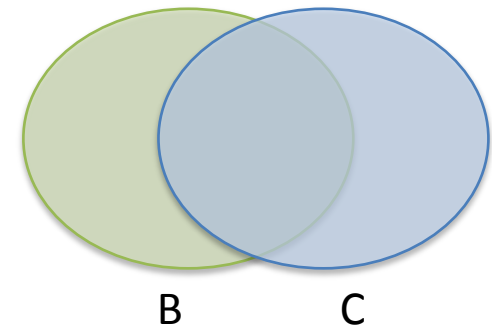
Update union $u[i,j] += N_i + N_j$



$$AJ(A,B) = 6/6 = 1$$



$$AJ(A,C) = 12/12 = 1$$



Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

Kmer shared by A, B and C

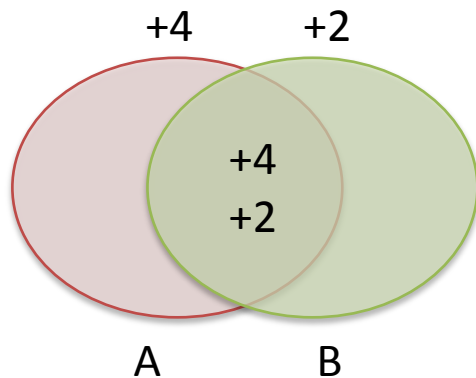
	A	B	C
ACGATC	4	2	8

For each pair(i, j)

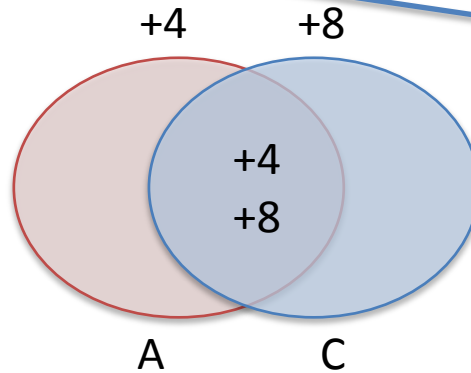
If kmer is shared $N_i > 0$ and $N_j > 0$

Update intersection $n[i,j] += N_i + N_j$

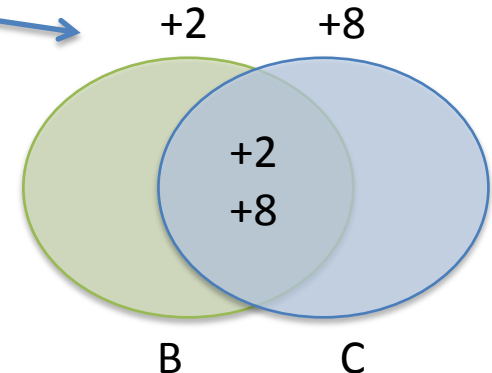
Update union $u[i,j] += N_i + N_j$



$$AJ(A,B) = 6/6 = 1$$



$$AJ(A,C) = 12/12 = 1$$



$$AJ(B,C) = 10/10 = 1$$

Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

Kmer specific to A

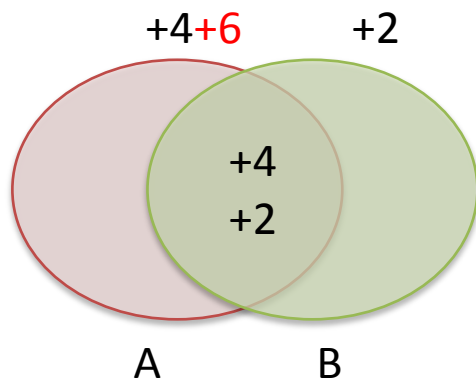
	A	B	C
ACGATC	6	0	0

For each pair(i, j)

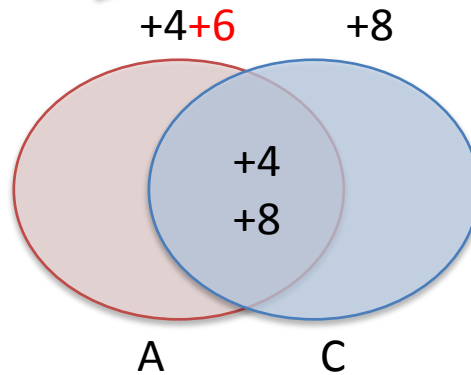
If kmer is shared $N_i > 0$ and $N_j > 0$

Update intersection $n[i,j] += N_i + N_j$

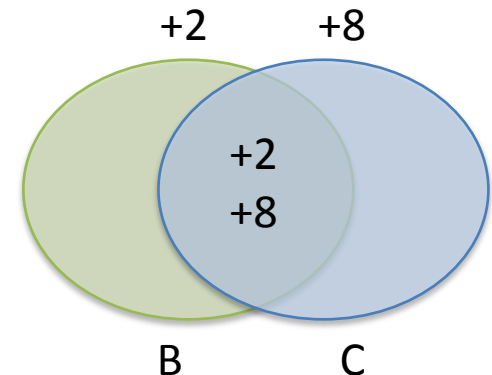
Update union $u[i,j] += N_i + N_j$



$$AJ(A,B) = 6/12 = 0.5$$



$$AJ(A,C) = 12/18 = 0.66$$



$$AJ(B,C) = 10/10 = 1$$

Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

$$AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}$$

$N(N-1)$

For each pair(i, j)

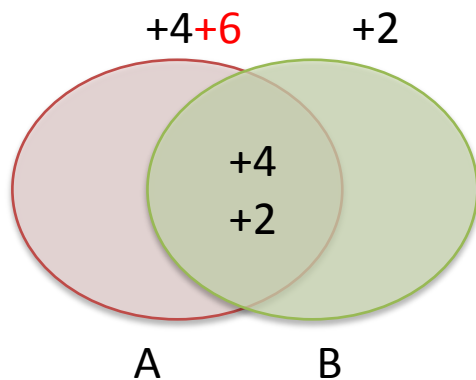
If kmer is shared $N_i > 0$ and $N_j > 0$

Update intersection $n[i,j] += N_i + N_j$

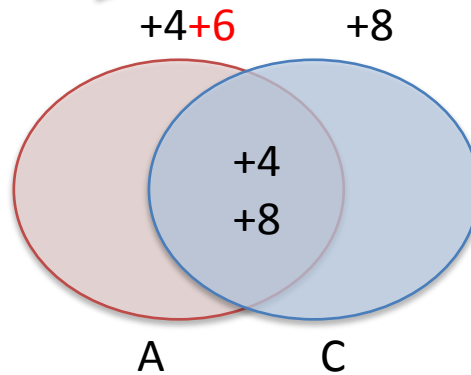
Update union $u[i,j] += N_i + N_j$

Kmer specific to A

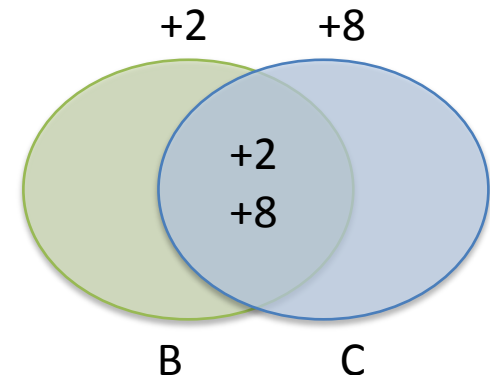
	A	B	C
ACGATC	6	0	0



$$AJ(A,B) = 6/12 = 0.5$$



$$AJ(A,C) = 12/18 = 0.66$$



$$AJ(B,C) = 10/10 = 1$$

Execution time

On 21 Tara samples, 2.1G reads, 210GB data

- Commet (state of the art)
 - On cluster (one node)
 - **A few weeks**
- Simka
 - On cluster (one node): **3h**
 - Counting kmers: 75%
 - Update unions/intersections **$N(N-1)$** : 25%
 - On standard computer: 10h
 - 4 GB memory, 4 cores

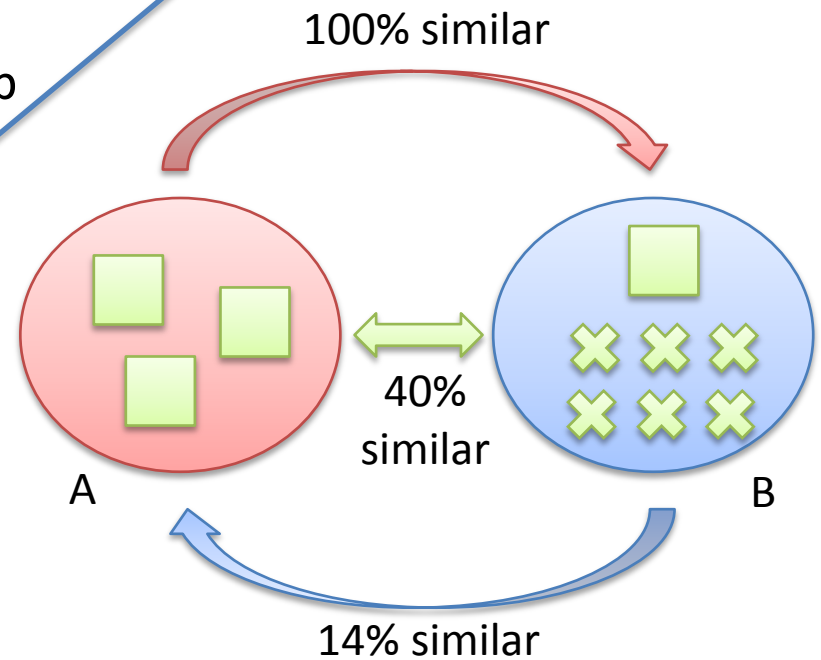
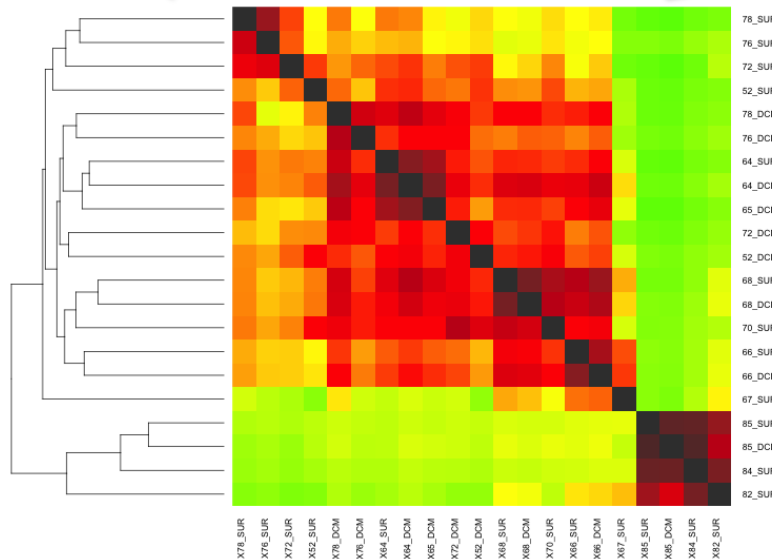
Representation of similarity

$$AJ_{sym}(i, j) = \frac{\hat{a}_{\hat{w}i\zeta_j} N_i(w) + N_j(w)}{\hat{a}_{\hat{w}i\tilde{E}j} N_i(w) + N_j(w)}$$

$$AJ_{asym}(i, j) = \frac{\hat{a}_{\hat{w}i\zeta_j} N_i(w)}{\hat{a}_{\hat{w}i} N_i(w)}$$

Hierarchical clustering

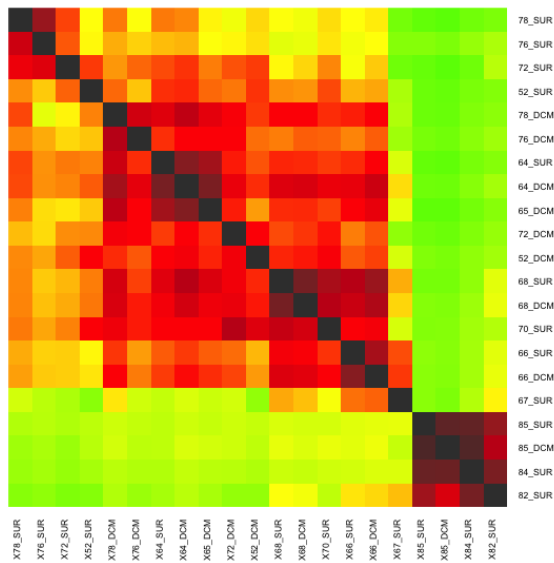
Heatmap



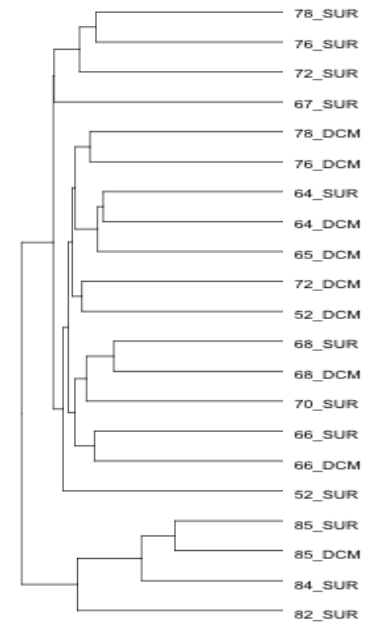
Asymmetry can show information about diversity

Results

Two things to compare with Commet



Heatmap
absolute similarity

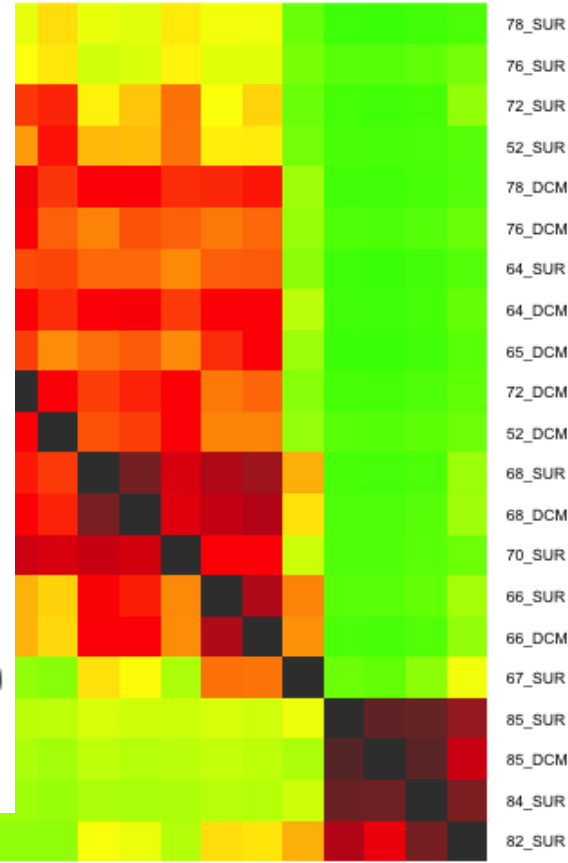
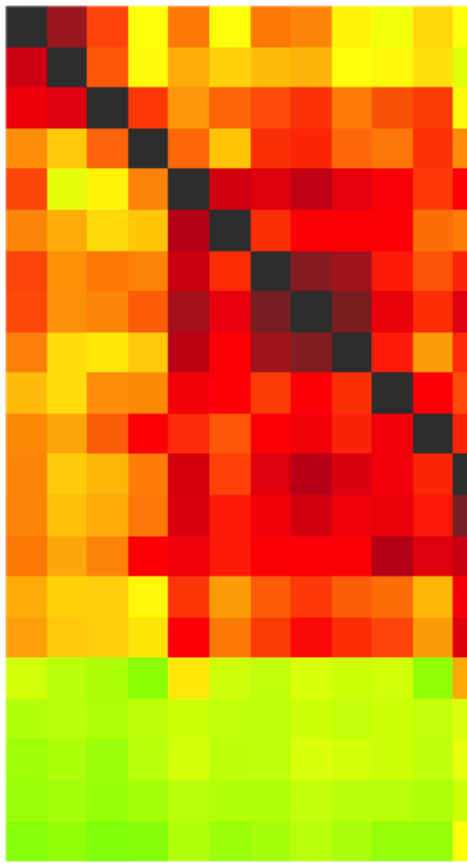
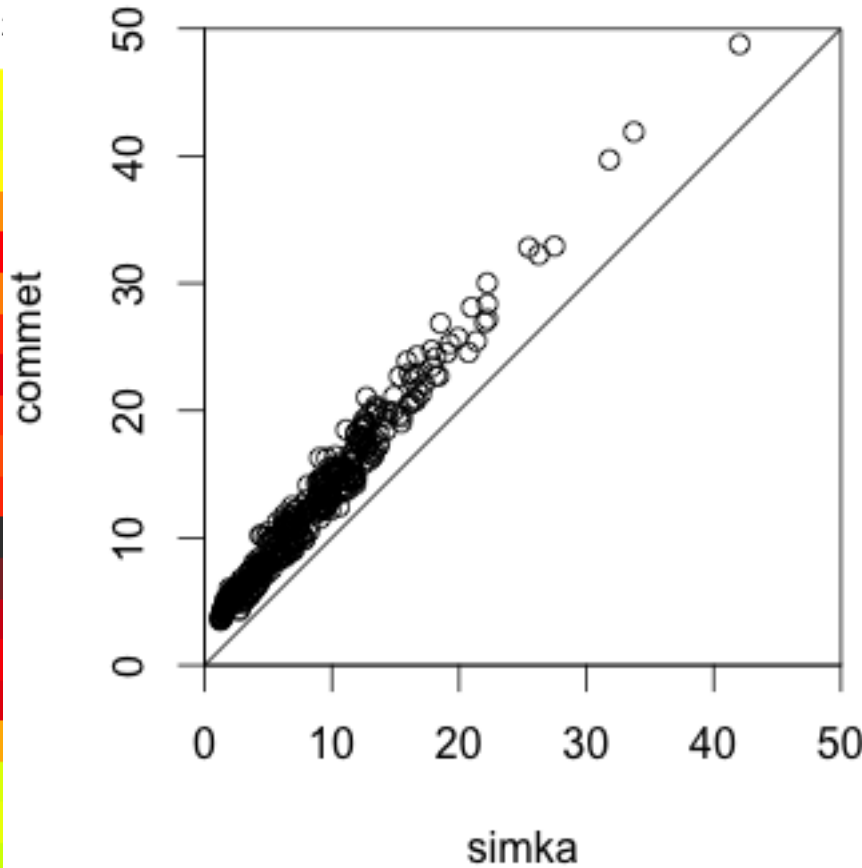
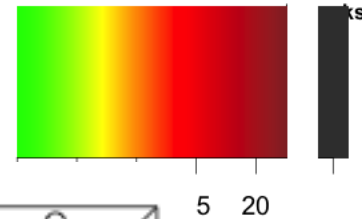
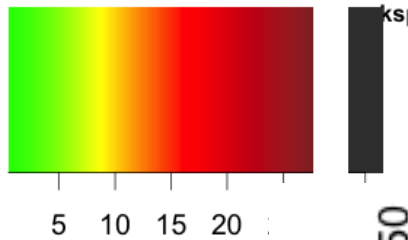


Hierarchical clustering
relative similarity

On 21 datasets from Tara Oceans

Commet (k=33)

Simka (k=31)



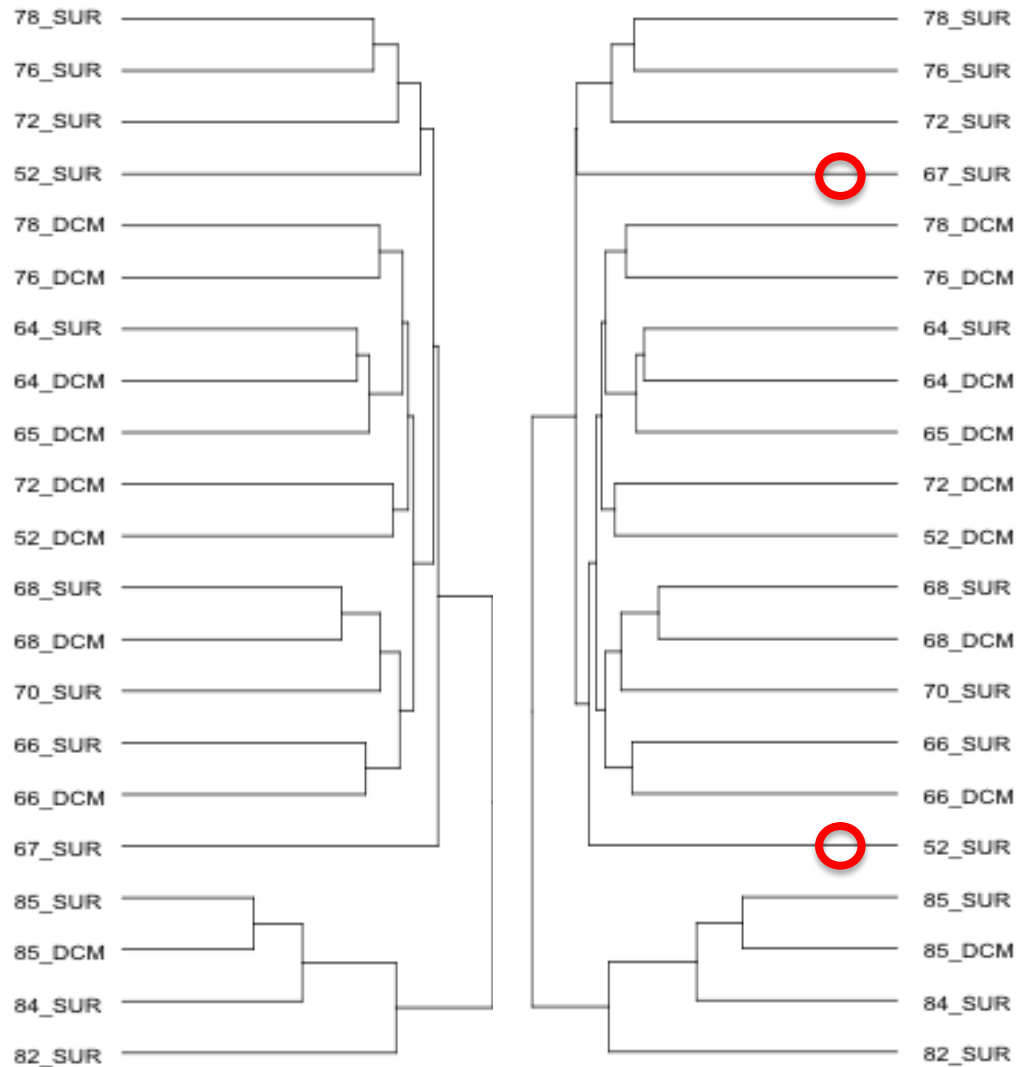
X78_SUR
X76_SUR
X72_SUR
X52_SUR
X78_DCM
X76_DCM
X64_SUR
X64_DCM
X65_DCM
X72_DCM
X52_DCM
X68_SUR
X68_DCM
X70_SUR
X66_SUR
X66_DCM
X67_SUR
X85_SUR
X85_DCM
X84_SUR
X82_SUR

82_SUR
X78_SUR
X76_SUR
X72_SUR
X52_SUR
X78_DCM
X76_DCM
X64_SUR
X64_DCM
X65_DCM
X72_DCM
X52_DCM
X68_SUR
X68_DCM
X70_SUR
X66_SUR
X66_DCM
X67_SUR
X85_SUR
X85_DCM
X84_SUR
X82_SUR

On 21 datasets from Tara Oceans

Commet (k=33)

Simka (k=31)



Other measures provided

- Transform the **abundance vectors** to get new measures
 - From **abundance** to **presence/absence**



Jaccard similarity:

$$J_{sym}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad J_{asym}(A, B) = \frac{|A \cap B|}{|A|}$$

- Filtering potentially erroneous kmers
 - Discard kmer if **specific** ($D=1$) and **unique** ($N_i=1$)

	A	B	C
ACGATC	0	1	0



	A	B	C
ACGATC	0	1	1



	A	B	C
ACGATC	0	0	8



Conclusion

- Simka
 - New pairwise similarity functions based on **shared kmers**
 - Provides measures based on **abundance** and **presence/absence**
 - Results close to read-based methods
- Fast and low memory thanks to the GATB library
 - Execution time (21 Tara samples, 2.1G reads, 210GB data)
 - Commet (state of the art): **few weeks**
 - Simka
 - On cluster: **3h**
 - On standard computer: 10h

GATB

www.gatb.inria.fr

Simka

www.gatb.fr/software/simka

Perspectives

- Add similarity measures well used in ecology (Ex: Bray Curtis)

	A	B	C
species	0	4	2

→

	A	B	C
kmer	0	4	2

- Bootstrapping
 - Test the robustness of Simka
 - Are there differences between using 100M or 10M reads?
 - Provides similarity matrix with confidence intervals
 - $\text{Sim}(A, B) = 23\% \pm 0.6\%$
 - Add confidence levels to dendrogram
- Simka Potara
 - For cluster and cloud
 - MapReduce model
 - Process huge amount of data
 - **130 Tara samples** (13G reads, 1.3TB data) compared in **< 1 day**

Acknowledgements



PHD Supervisor

Claire Lemaitre

Pierre Peterlongo

Dominique Lavenier

GATB devs

Erwan Drezen

Guillaume Rizk

Commet

Nicolas Maillet

Partners



INRA

Stéphane Robin

Sophie Schbath

Mahendra Mariadassou

Julien Chiquet

Julie Aubert

Sarah Ouadah

Emilie Lebarbier

CEA

Olivier Jaillon

Jean-Marc Aury

Eric Pelletier

Thomas Vannier

Founded

ANR Hydrogen