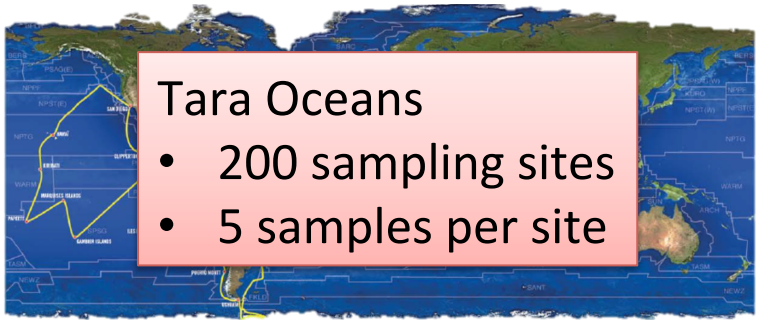
Simka

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

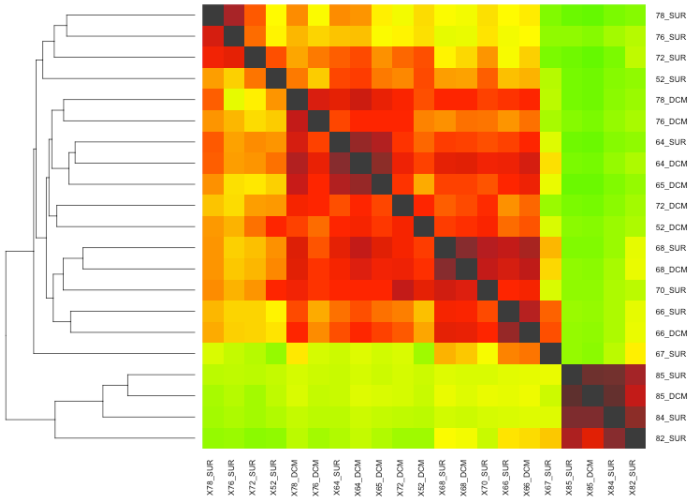
Gaëtan BENOIT
PHD student - ANR Hydrogen

GenScale bioinformatics research group
IRISA/INRIA – Rennes - FRANCE
30/06/2015

Context



N metagenomic samples



Similarity heatmap N^2

Comparative metagenomic

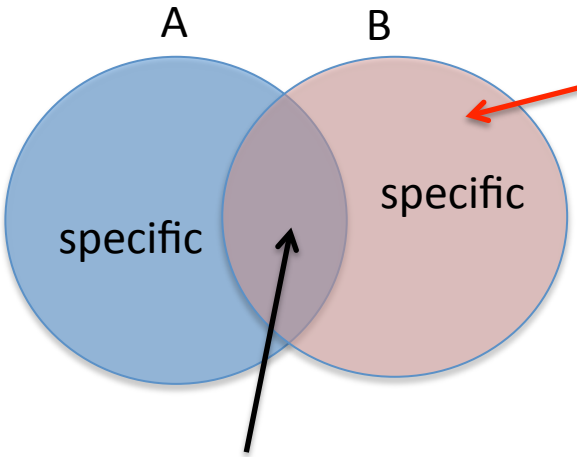
fasta fasta … fasta

N datasets MetaG

Tara Oceans
- 1000 datasets
- > 100 millions reads each

Similarity between 2 samples

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



100M reads

✖ Blast-like approach required **months** of computation

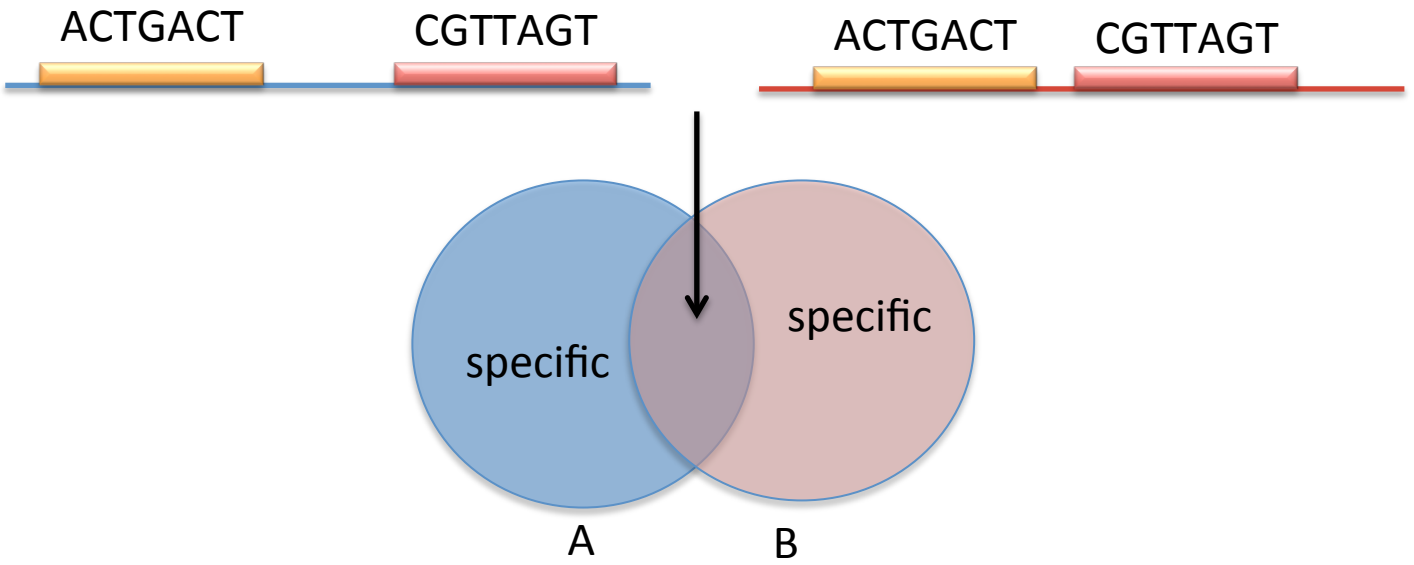
Intersection = number of **similar reads**

Similar: alignment score > 90%

State of the art

Commet (Maillet *et al.* 2014)

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



- Computes one intersection in **few hours**
  - We have fast methods for **indexing** and **querying** kmers

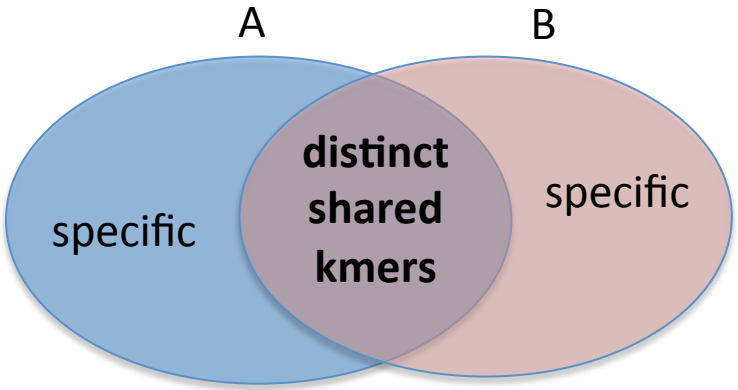
Does not scale on large metagenomic projects : N(N-1) comparisons

Scaling on large metagenomic projects

Simka

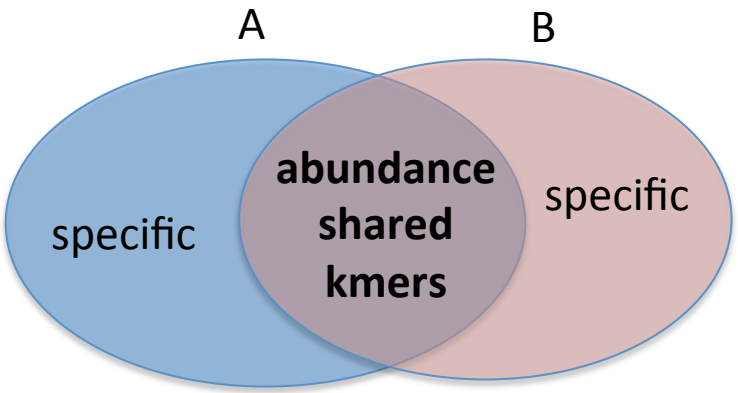
- Long kmers are biologically significant (k > 30)
- A dataset is view as a **set of its kmers**
- New pairwize similarity measures based on **shared kmers**

Presence / absence of kmers



Jaccard similarity

Abundance of kmers



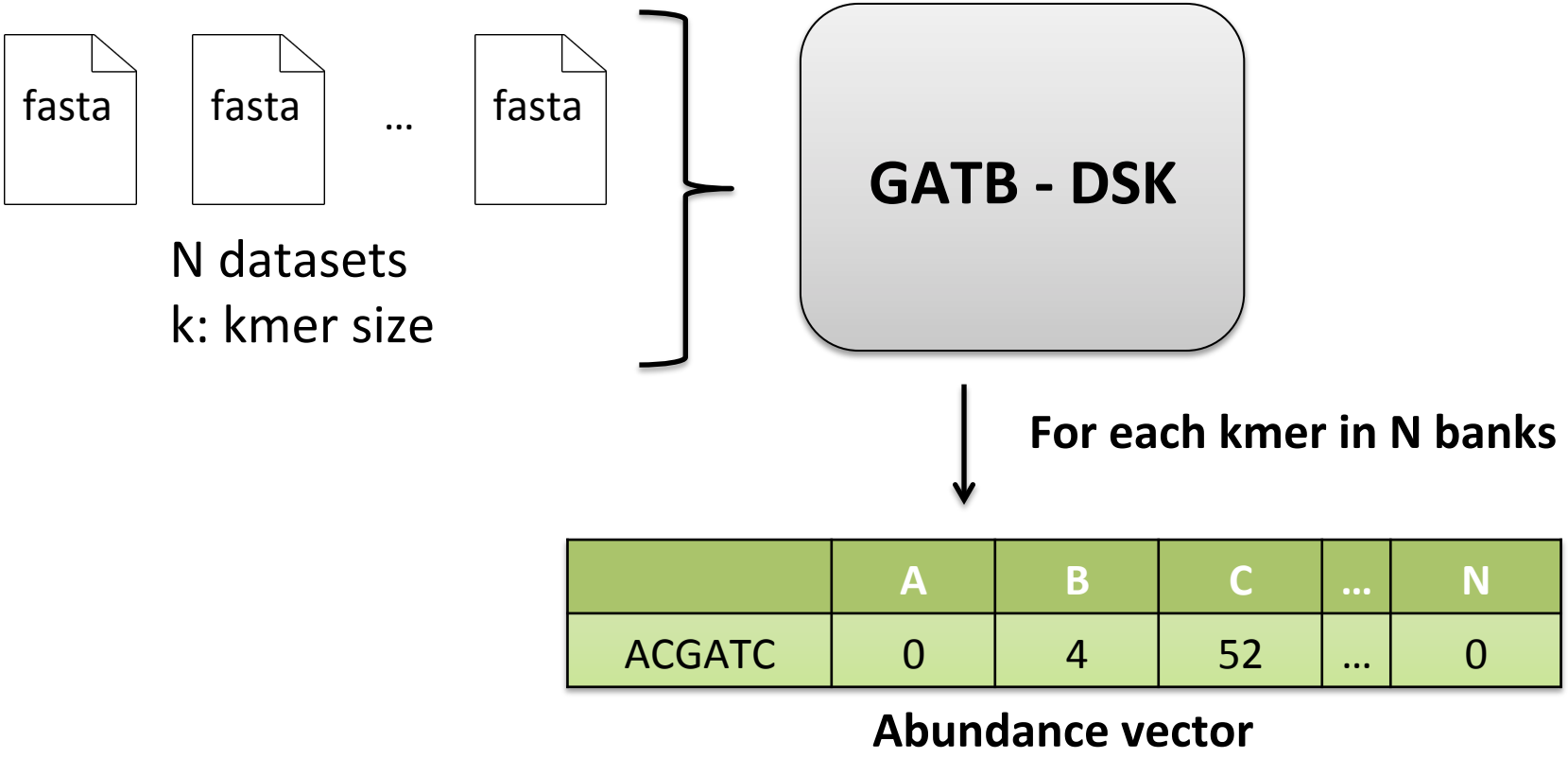
Abundance-based Jaccard similarity

Required kmer counting stage

Always N(N-1) comparisons

Multi-dataset kmer counting

- Based on KMC2 algorithm (Deorowicz *et al.* 2015)
  - Very fast kmer counting algorithm with low resources
- Count the kmers of N datasets **simultaneously** (Erwan Drezen)



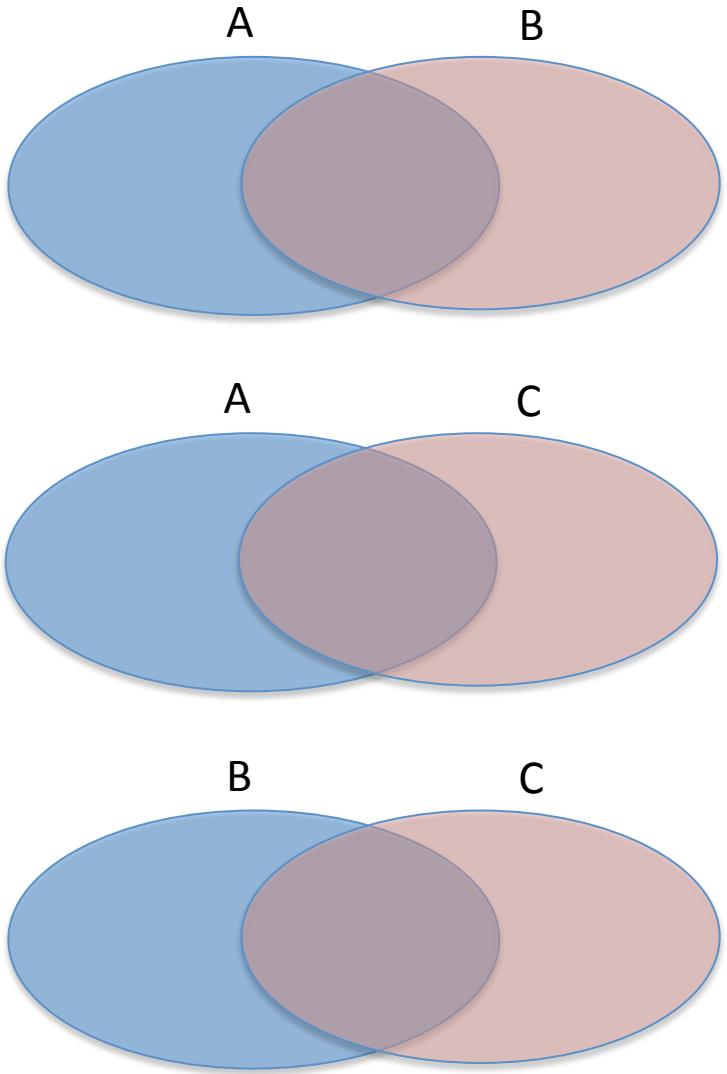
Simka algorithm

Example for 3 datasets

	A	B	C
ACGATC	4	2	8

Kmer's abundances

- Check if pairs of datasets share the current kmer
- Update statistics of sets and intersections

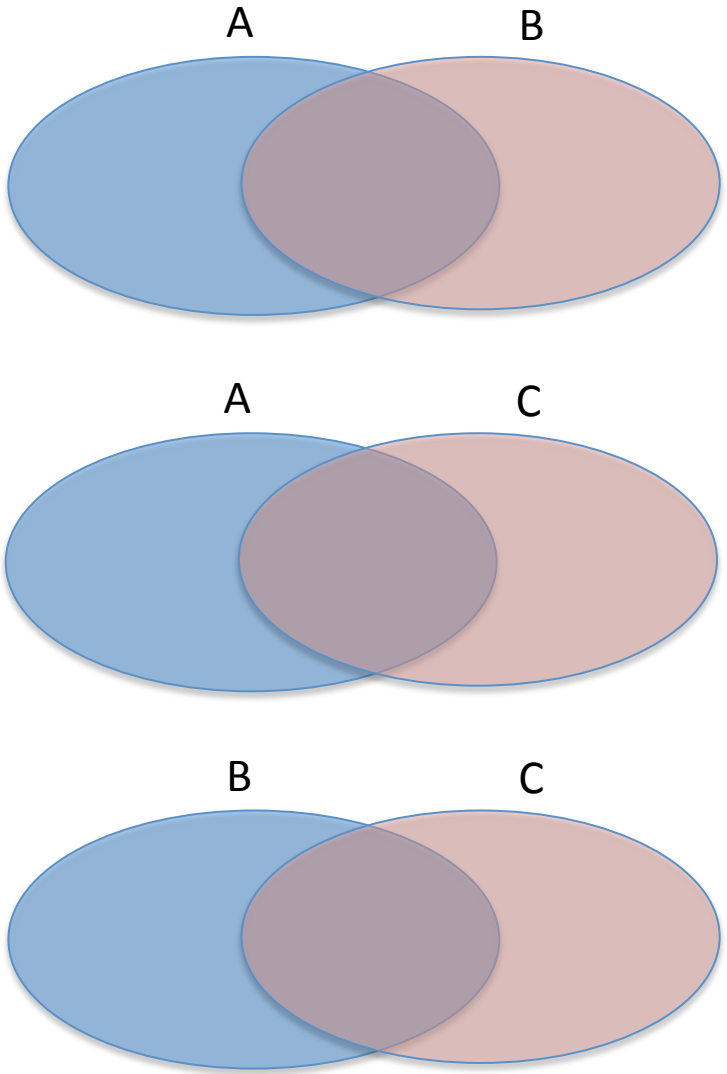


Simka algorithm

Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8

Abundance-based Jaccard similarity

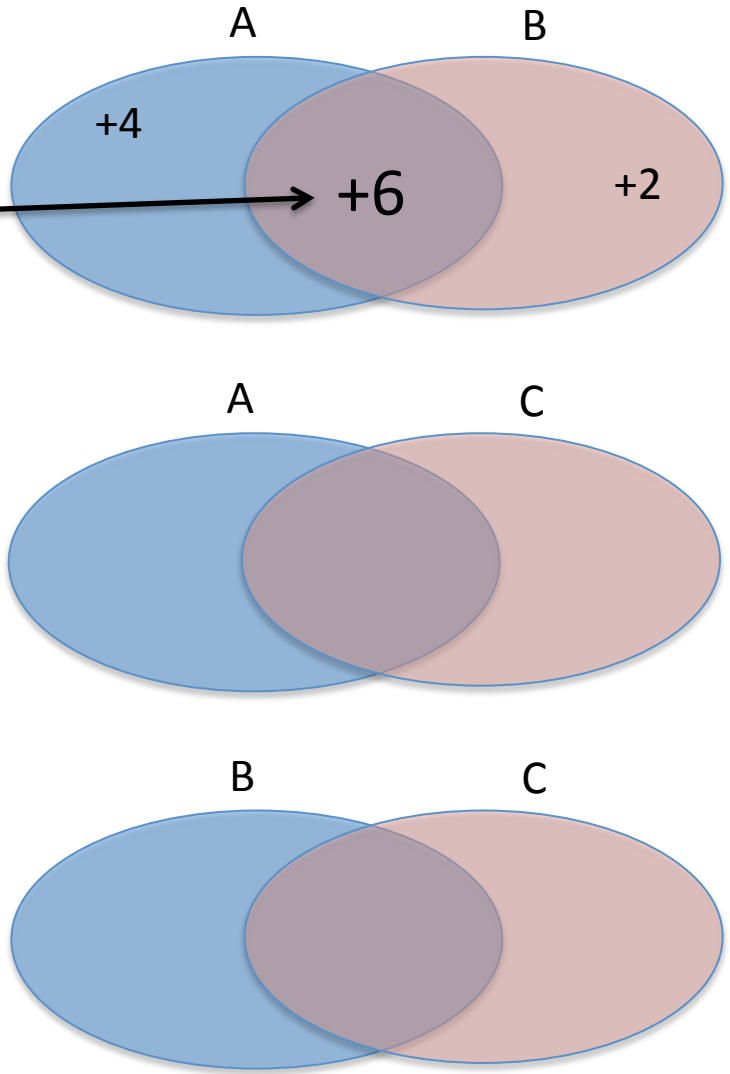


Simka algorithm

Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8

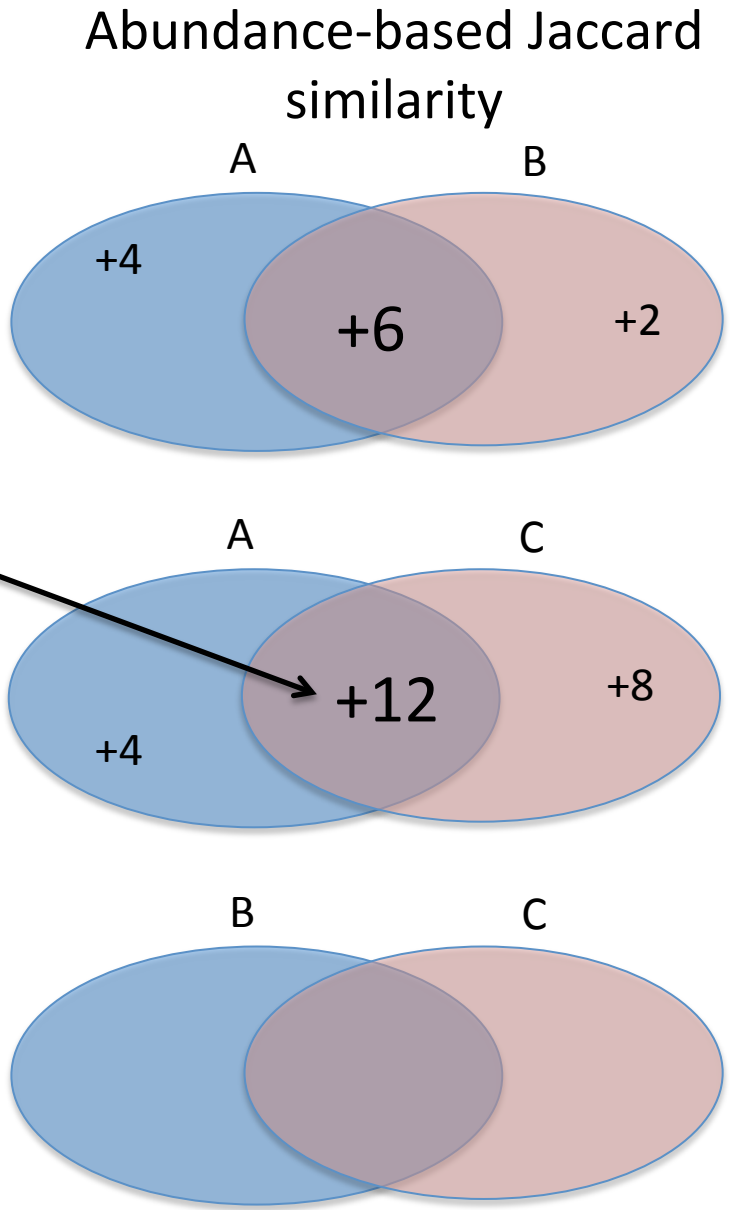
Abundance-based Jaccard similarity



Simka algorithm

Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8

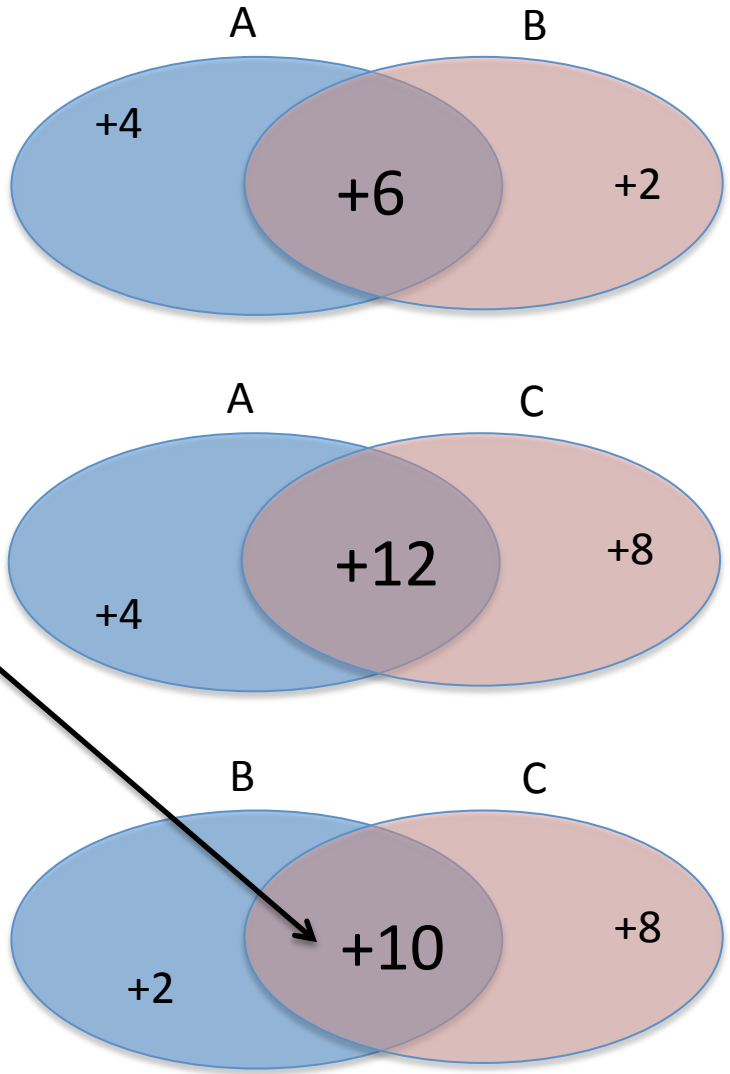


Simka algorithm

Kmer shared by A, B and C

	A	B	C
ACGATC	4	2	8

Abundance-based Jaccard similarity

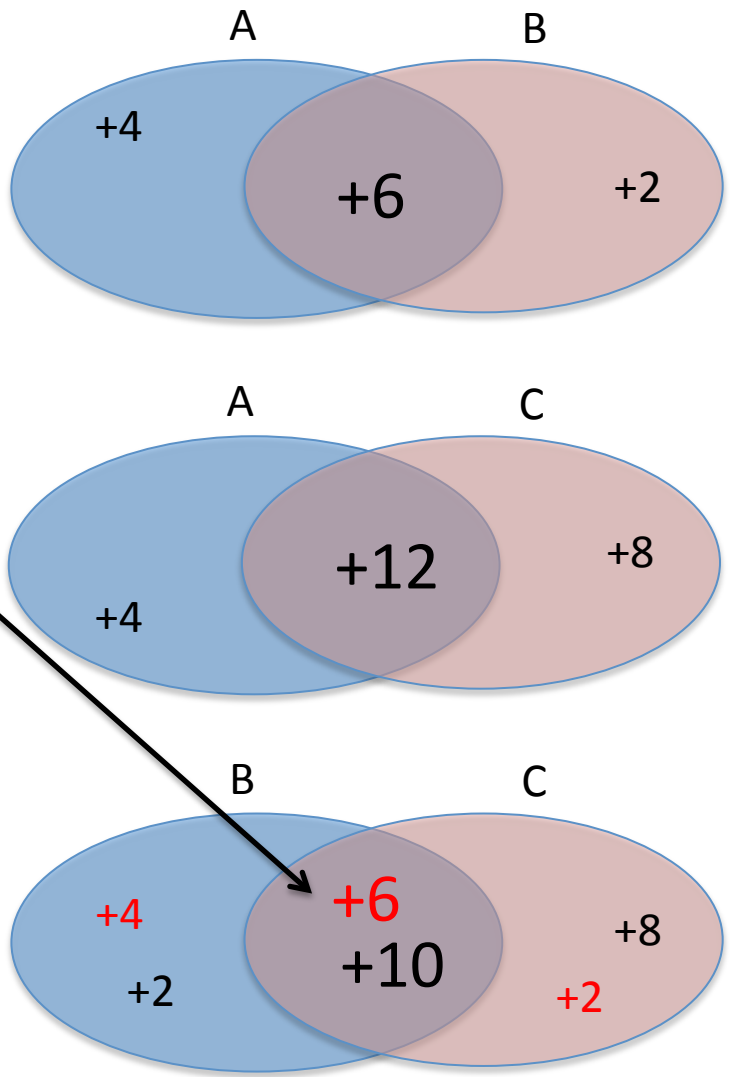


Simka algorithm

Kmer shared by B and C

	A	B	C
ACGATC	0	4	2

Abundance-based Jaccard similarity

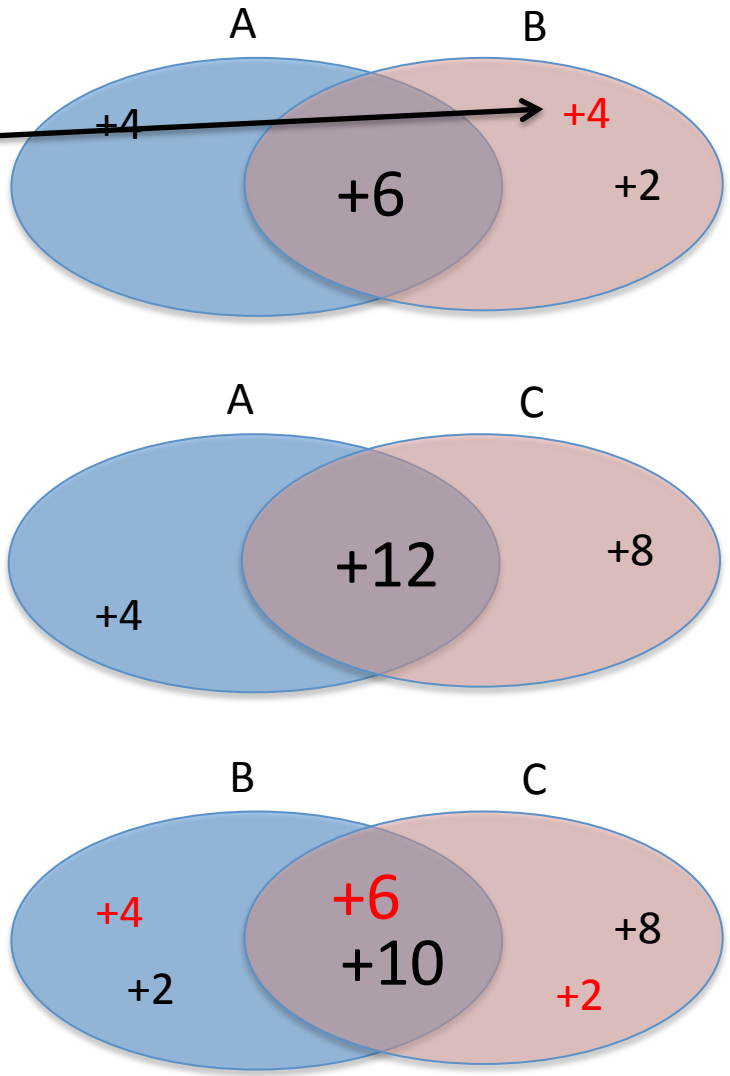


Simka algorithm

Kmer shared by B and C

	A	B	C
ACGATC	0	4	2

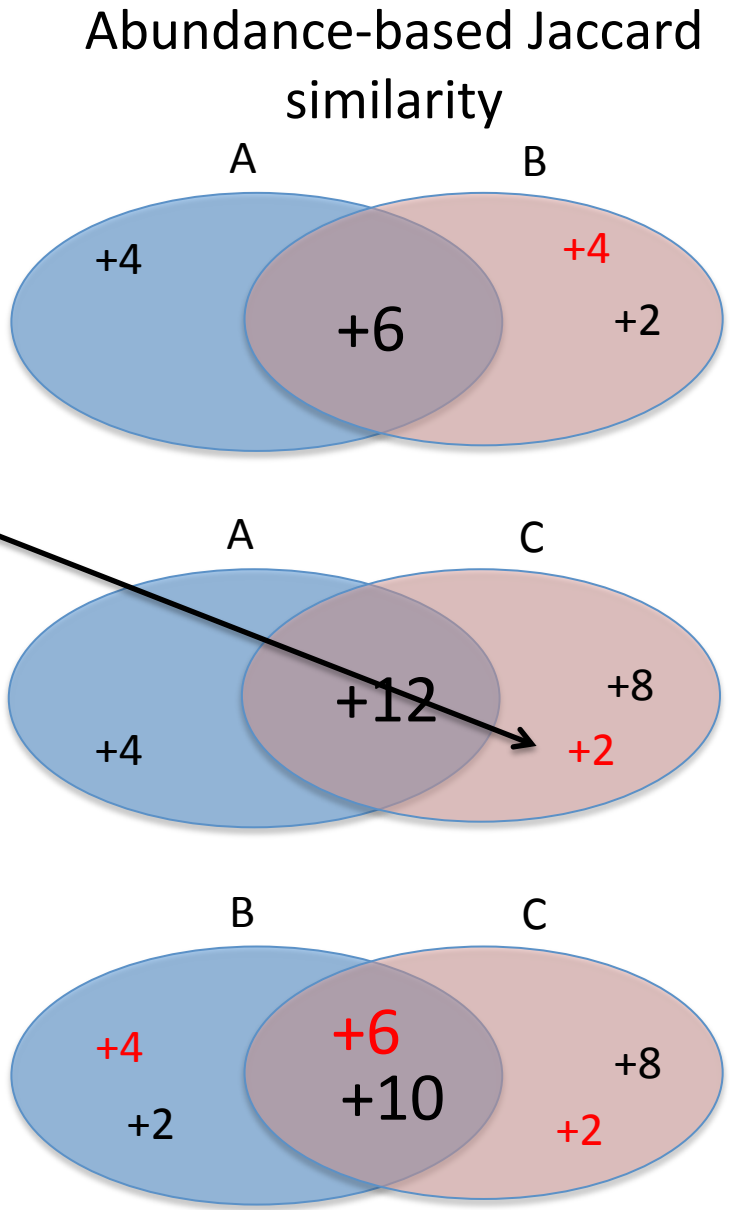
Abundance-based Jaccard similarity



Simka algorithm

Kmer shared by B and C

	A	B	C
ACGATC	0	4	2

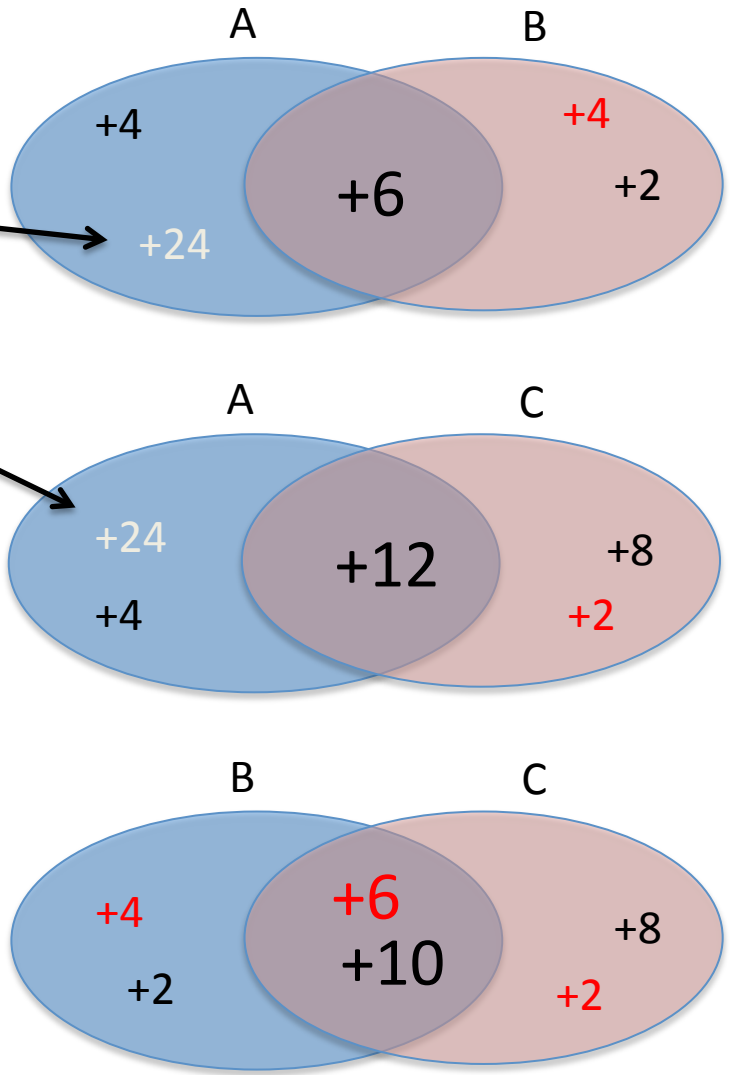


Simka algorithm

Kmer specific to A

	A	B	C
ACGATC	24	0	0

Abundance-based Jaccard similarity



Simka algorithm

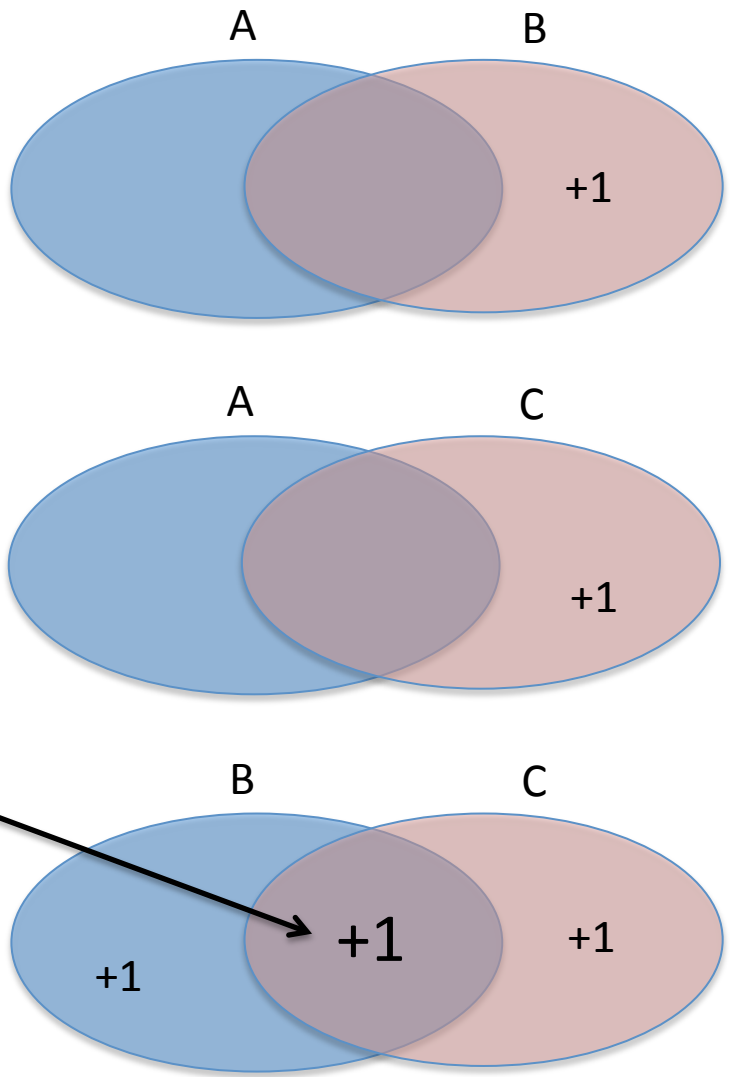
Presence/absence of kmers

Kmer shared by A, B and C

	A	B	C
ACGATC	0	22	4

To boolean vector

	A	B	C
ACGATC	0	1	1



Simka algorithm

- Execution time (21 Tara samples, 3G reads, 400 GB)
  - Commet
    - On cluster (one node)
      - A few weeks
  - Simka :
    - On cluster (one node): 4h
      - Counting kmers: 75%
      - Computing stats $O(N^2)$: 25%
    - On standard computer: 10h
      - 4 GB memory, 4 cores

Similarity measures

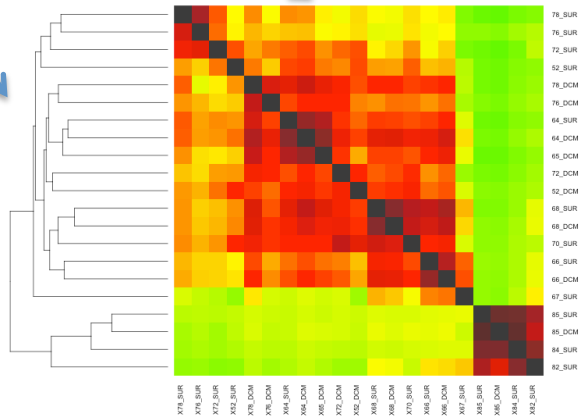
- Presence / absence of kmers

$$DKS_{asym}(A,B) = \frac{DistinctKmers(A \cap B)}{DistinctKmers(A)}$$

$$DKS_{sym}(A,B) = \frac{DistinctKmers(A \cap B)}{DistinctKmers(A \cup B)}$$

Heatmap

Clustering



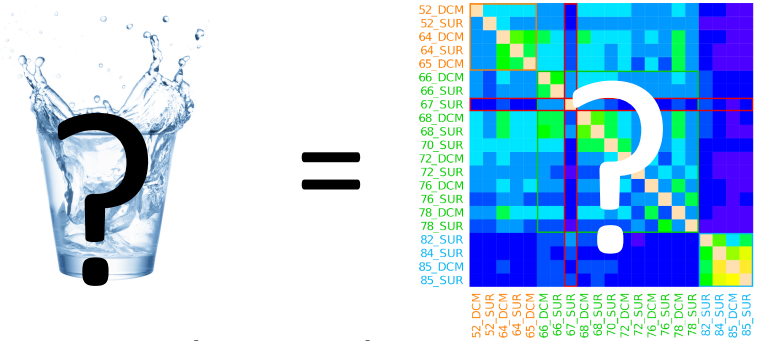
- Abundance of kmers

$$AKS_{asym}(A,B) = \frac{\sum_{w \in A \cap B} N_A(w)}{\sum_{w \in A} N_A(w)}$$

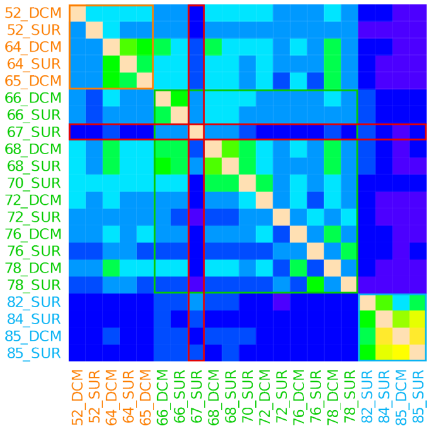
$$AKS_{sym}(A,B) = \frac{\sum_{w \in A \cap B} N_A(w) + N_B(w)}{\sum_{w \in A \cup B} N_A(w) + N_B(w)}$$

Validation

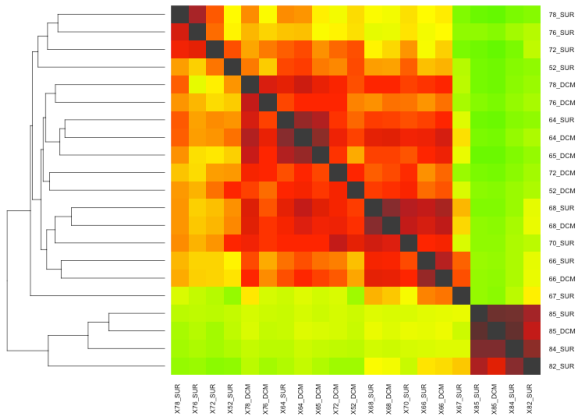
- How to validate Simka ?



- We can compare Simka results to Commet ones
  - Two things to compare in heatmaps:



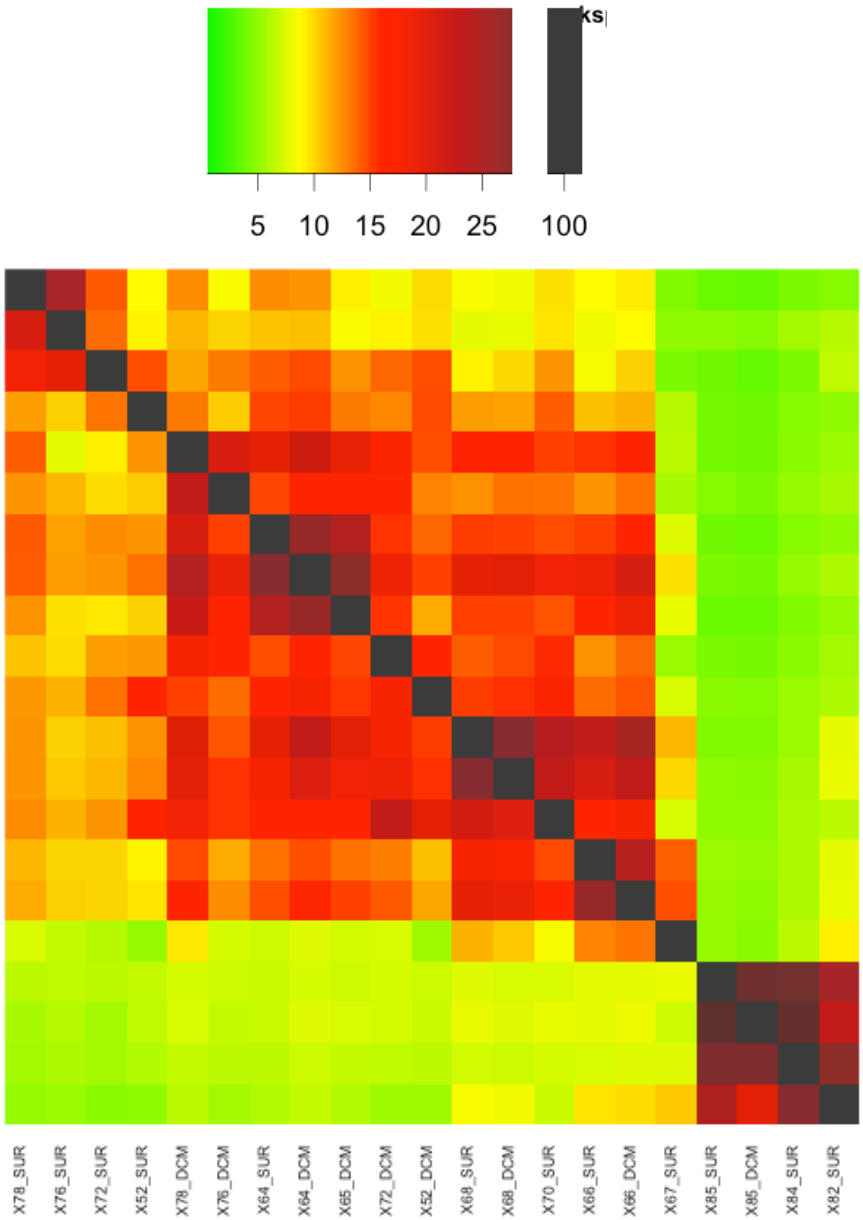
Colors
(absolute similarity)



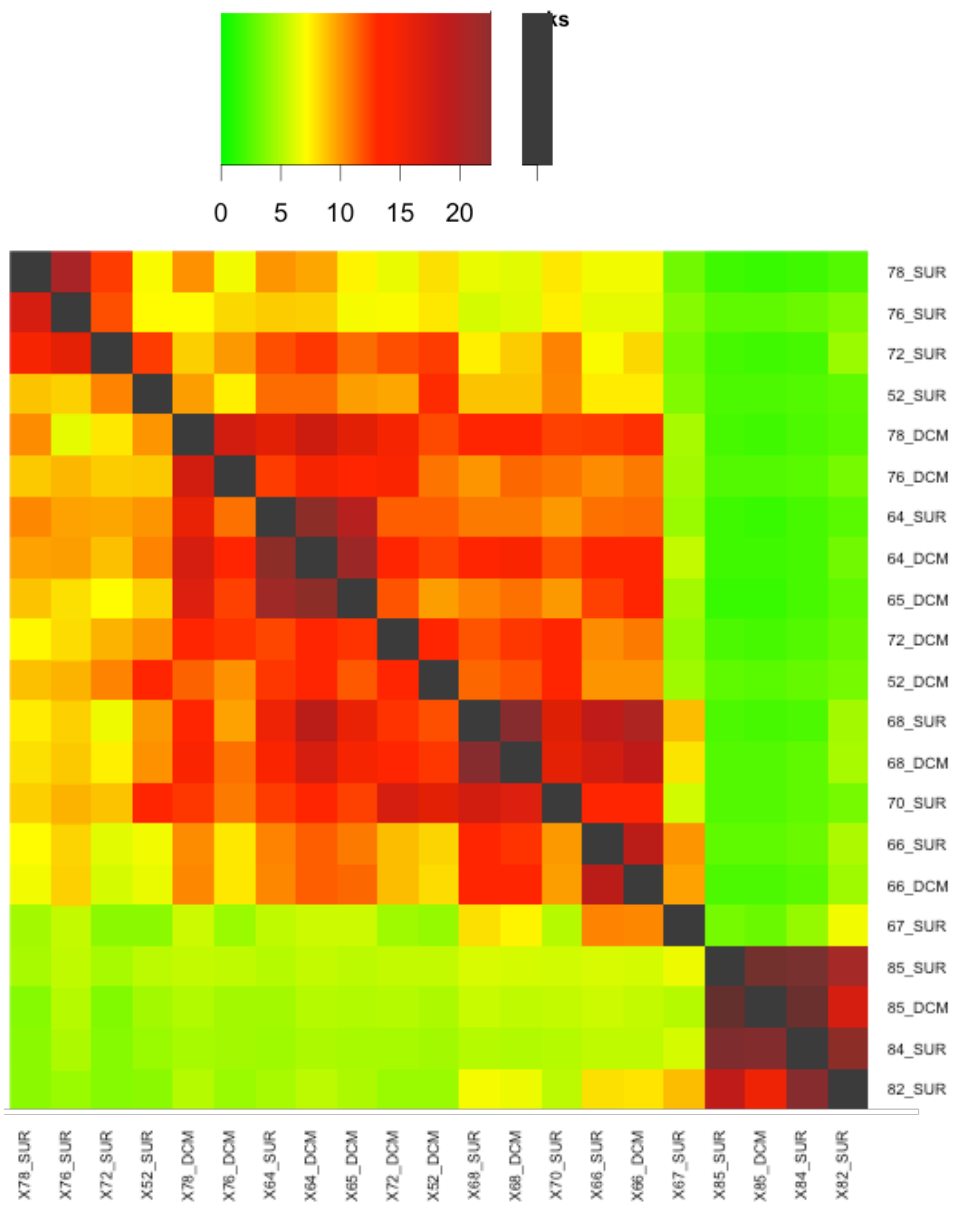
Clustering
(relative similarity)

Absolute similarity

Commet (k=33)



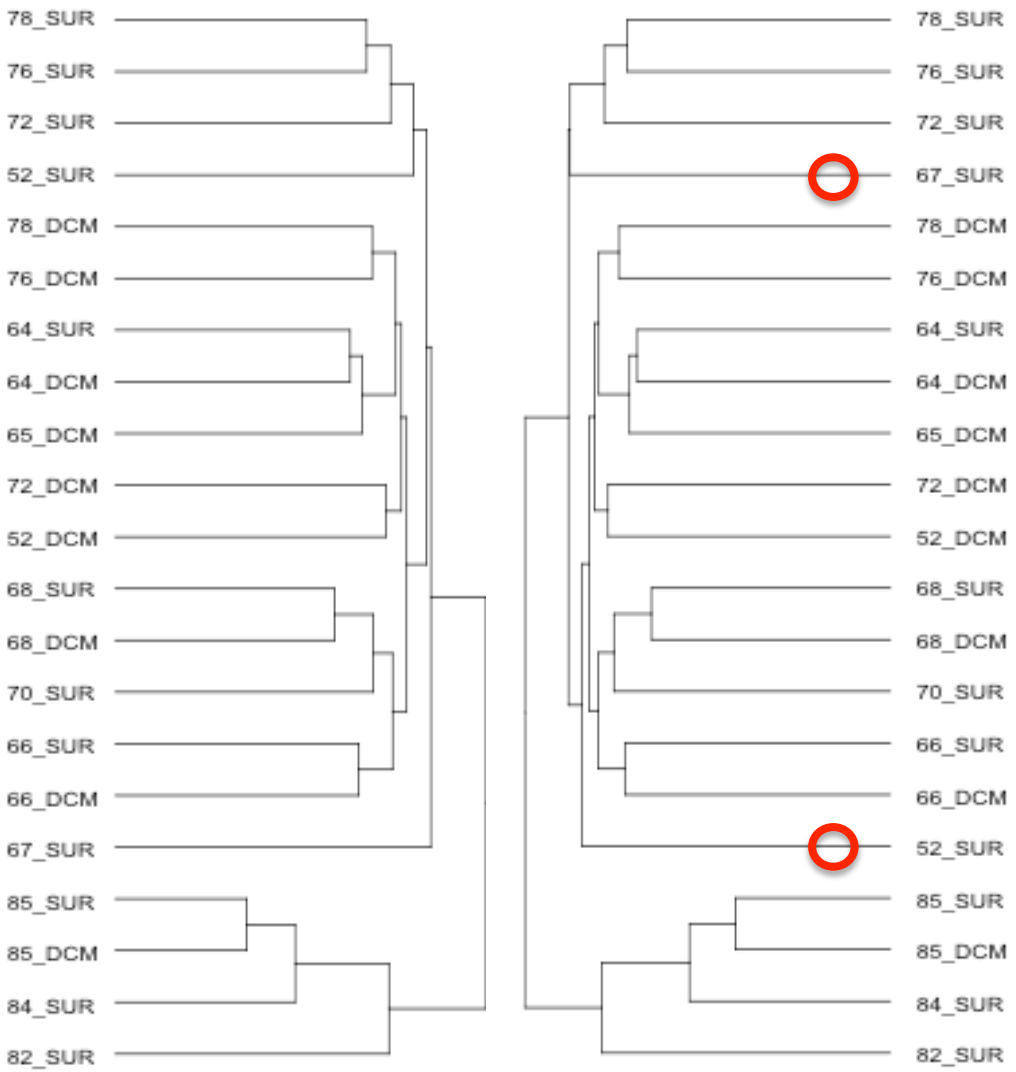
Simka abundance (k=31)



Clustering

Commet (k=33)

Simka abundance (k=31)



Conclusion

- Simka
  - New similarity functions based on **shared kmers**
    - Based on **abundance** and **presence/absence** of species
    - Results close to read-based methods
  - Fast and low memory thanks to the GATB library
    - Execution time (21 Tara samples, 3G reads, 400 GB)
      - Commet (state of the art): few weeks
      - Simka
        - On cluster: 4h
        - On standard computer: 10h

GATB: https://gatb.inria.fr/

Perspectives

- Selecting discriminative kmers
- Add similarity measures well used in ecology (ex: Bray Curtis)
- Bootstrapping
  - Compute tens of similarity matrix with datasets randomly subsampled (~10% of the reads)
  - Test robustness of Simka
    - Provides similarity matrix with confidence intervals
  - Add confidence levels to dendrogram

Perspectives

- Selecting discriminative kmers
- Add similarity measures well used in ecology (ex: Bray Curtis)
- Bootstrapping
  - Compute tens of similarity matrix with datasets randomly subsampled (~10% of the reads)
  - Test robustness of Simka
    - Provides similarity matrix with confidence intervals
  - Add confidence levels to dendrogram

Acknowledgements

Claire LEMAITRE
Pierre PETERLONGO
Dominique LAVENIER
-
ANR Hydrogen