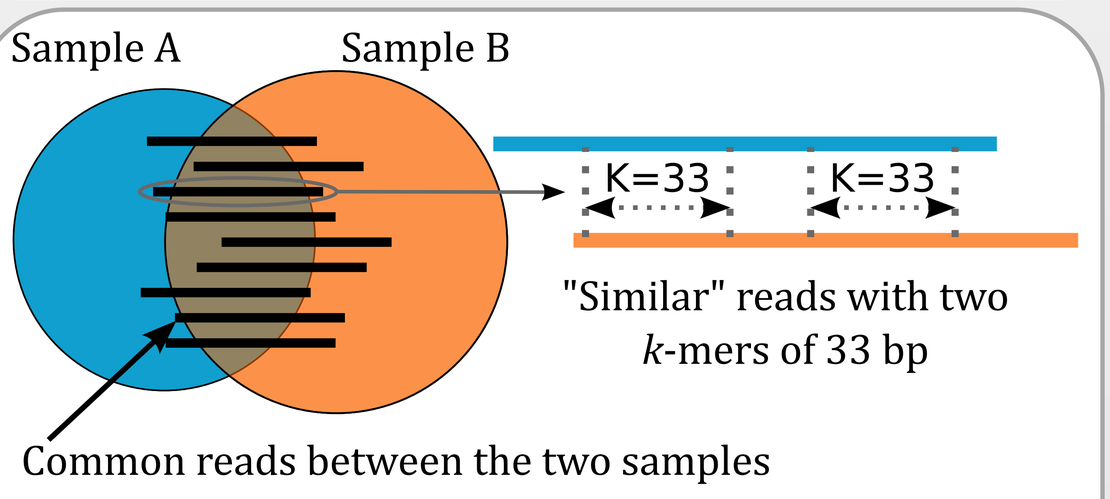
Compareads: comparing huge metagenomic experiments

Nicolas Maillet*[1], Claire Lemaitre[1], Rayan Chikhi[2], Dominique Lavenier[1], Pierre Peterlongo*[1]

[1]INRIA Rennes - Bretagne Atlantique / IRISA, EPI GenScale, Rennes, France
[2]ENS Cachan/IRISA, EPI GenScale, Rennes, France
*Corresponding authors

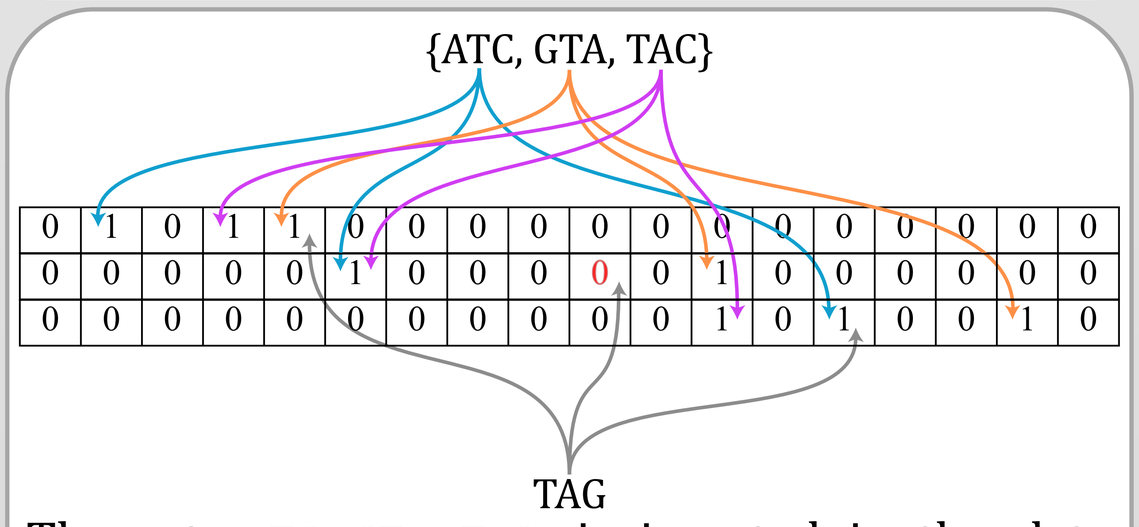
Metagenomics studies overall genomic information of multiple organisms coming from the same biotope. The information is generally provided by High Throughput Sequencers. Nowadays, metagenomic sample analyses are mainly achieved by comparing them with *a priori* knowledge stored in data banks.
This work introduces Compareads, a *de novo* comparative metagenomic approach that returns the similar reads between two datasets. Thanks to a new data structure, its time and memory features make Compareads able to retrieve biological information while scaling to huge datasets.

How to define "similarity"?



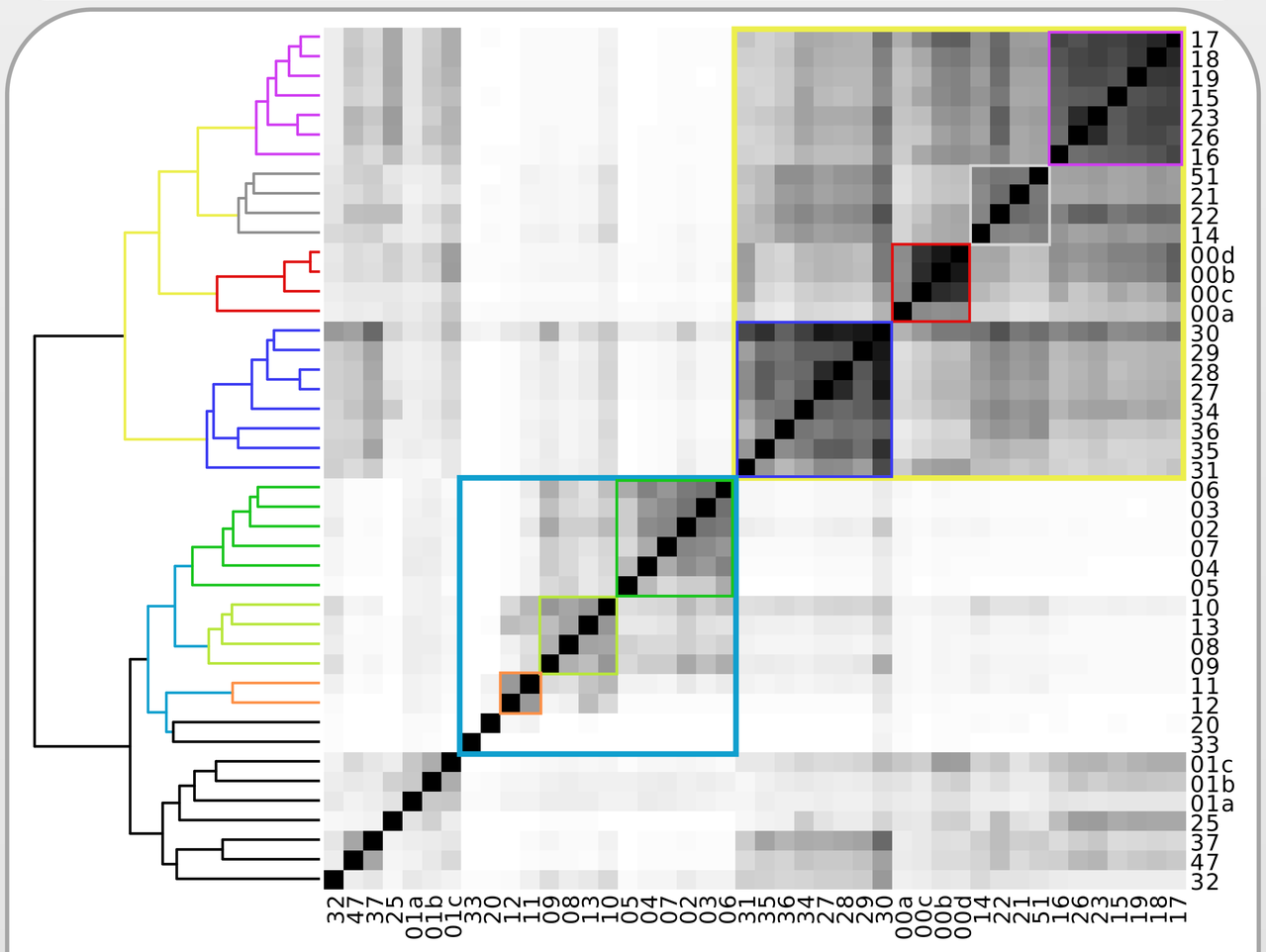
In order to perform efficiently the intersection, Compareads uses a rough but efficient notion of "*similar sequences*": two sequences are said similar if and only if they share at least **t** non overlapping words of length **k** (*k*-mers).

Modified Bloom Filter



The set {ATC, GTA, TAC} is inserted in the data structure using 3 hash functions. A query is said to be present in the set if the 3 hashes in the bit-arrays are filled with 1. The element TAG is not in the set {ATC, GTA, TAC}, because it hashes to at least one bit-array position containing 0.

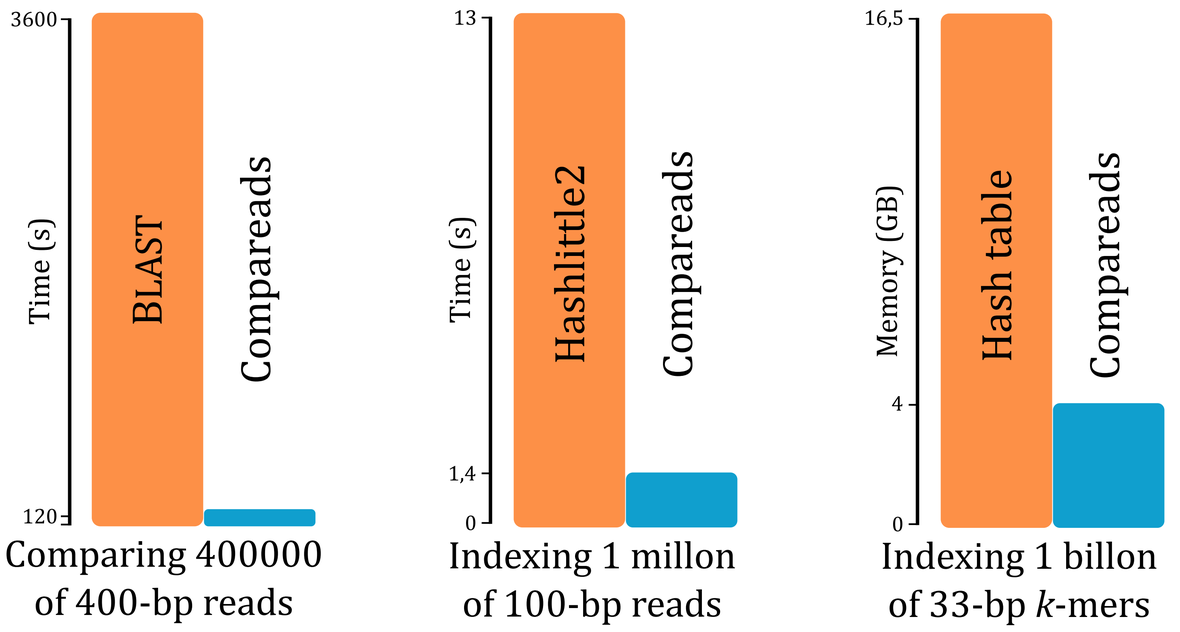
Results on Global Ocean Sampling[1]



Hierarchical clustering based on pairwise intersections between all samples of the Global Ocean Sampling (The Sorcerer II expedition[1]).

1. Rusch et al. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. Plos Biol (2007) vol. 5 (3) pp. e77

Performance comparisons with other methods



Using our new data structure, Compareads efficiently performs *de novo* intensive comparisons of huge metagenomic datasets generated by High Throughput Sequencers. This approach enables to retrieve and classify differences in genomic content between metagenomic samples. For this kind of comparison, our approach is much faster than alternative ones such as BLAST and thus enables to scale to huge datasets. Compareads is released under the CeCILL license and can be freely downloaded from:

http://alcovna.genouest.org/compareads