Identification of bacterial strains using Oxford Nanopore sequencing

Gregoire Siekaniec

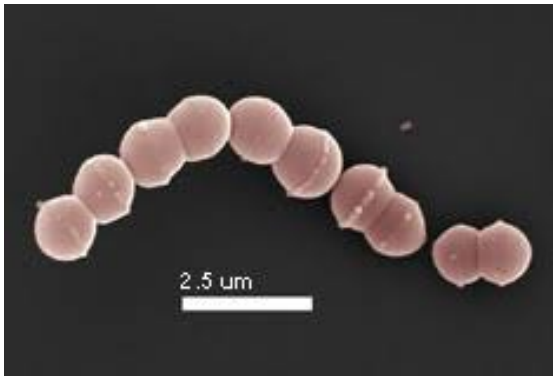
Inria Rennes, GenScale Team
INRAE, UMR STLO, MicroBio Team
Supervisors : Jacques Nicolas, Eric Guedon and Emeline Roux

Strain identification

1. Bacterial strains of the same species have different phenotypes (e.g. commensal/pathogens *E. coli*)
2. How to differentiate strains ?
   - On petri dishes, no differences;
   - With conventional molecular technics (PCR on 16S or housekeeping genes), it is often not very discriminating
   - today, 3rd generation sequencing offers lower cost, increased flow rates... allowing to use the whole-genome information.

My study model

- *Streptococcus thermophilus* (recent species with low genetic divergence)*



- MinION

* Christine Delorme,Safety assessment of dairy microorganisms: Streptococcus thermophilus, International Journal of Food Microbiology,Volume 126, Issue 3,2008,Pages 274-277, ISSN 0168-1605,https://doi.org/10.1016/j.ijfoodmicro.2007.08.014.

Objectives / Challenges

Objectives

To be able to identify different bacterial strains in a mixture with Nanopore long reads sequencing technology.

Challenges

Index all bacterial genomes & managed the sequence errors

- Start with strains from a single species (*S. thermophilus*)
- Robust to errors and fast
- New strains identification tool
  ORI (Oxford nanopore Reads Identification)

Nanopore sequencing: reads errors

Errors rate

Average percent error rate in the *Streptococcus thermophilus* strains sequences.

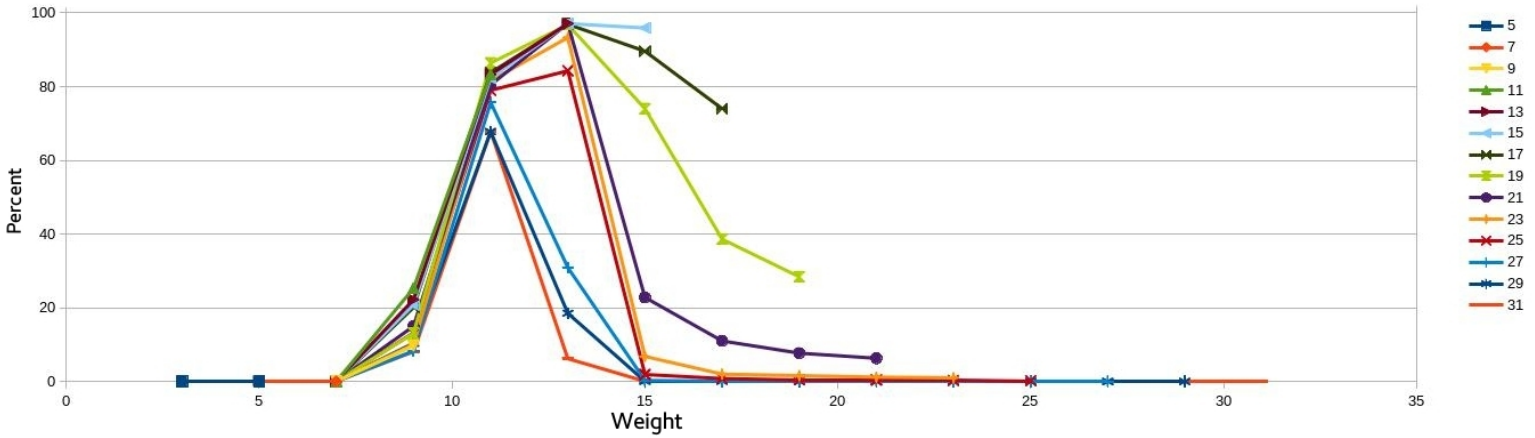
Errors	**Mismatchs**	**Deletions**	**Insertions**	**Total**
All sequences	2.43 %	2.93%	2.74%	8.1%
With filters	2.34%	2.84%	2.69%	7.87%

Retain sequences with quality ≥ 9 and size ≥ 2000.

How to deal with errors ?

Spaced seeds* (introduce don't care positions in kmers)

- example of spaced seed: 110011
- ATTCGA $\rightarrow$ AT- -GA $\rightarrow$ ATGA (qgram)
- best seed for the identification: 111111001111111 (length = 15, weight = 13,designed with iedera**)



* Leimeister, C.A.; Boden, M.; Horwege, S.; Lindner, S.; Morgenstern, B. Fast alignment-free sequence comparison using spaced-word frequencies. Bioinformatics 2014, 30, 1991–1999. doi:10.1093/bioinformatics/btu177.

** Noe L., Best hits of 11110110111: model-free selection and parameter-free sensitivity calculation of spaced seeds, Algorithms for Molecular Biology, 12(1). 2017 <http://doi.org/10.1186/s13015-017-0092-1>

How to store genomes ?

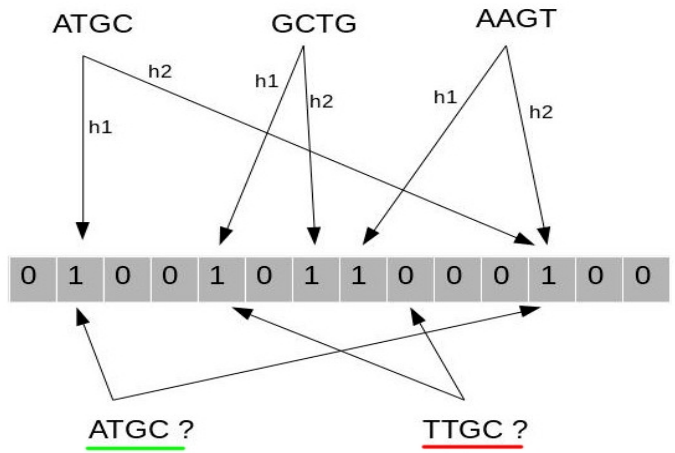
Compression

Many reference genomes have to be stored : need for compression

Some examples of existing techniques for indexing a genome:

1. Suffix tree
2. Burrows–Wheeler transform
3. **Bloom filter** (Burton Howard Bloom in 1970)

- + really compact
- + no false negative
- - false positives (can be minimized)

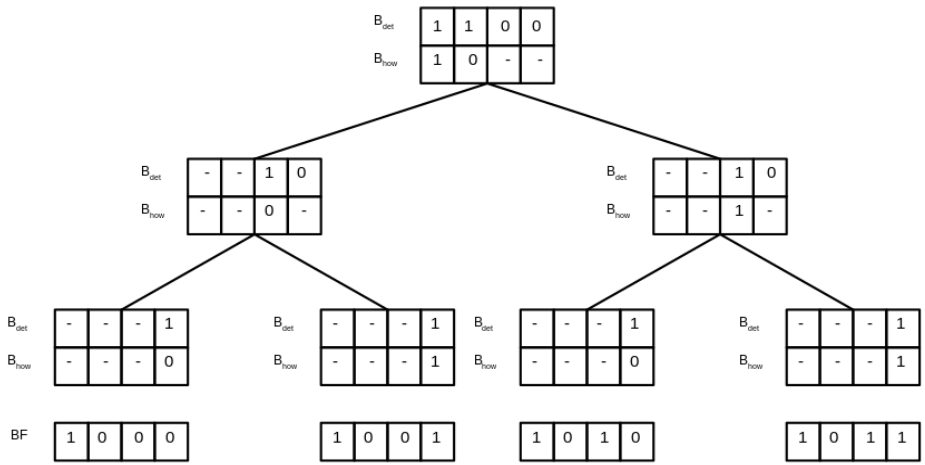


Bloom filter is not enough

Need to differentiate genomes:

- index based on differences
- staying compact
- allowing errors to be taken into account (spaced seeds)

Index based on the Bloom filter tree topology from HOWdeSBT* (modified with the help of Téo Lemane):



* Robert S Harris and Paul Medvedev, Improved representation of sequence bloom trees, Bioinformatics, btz662.

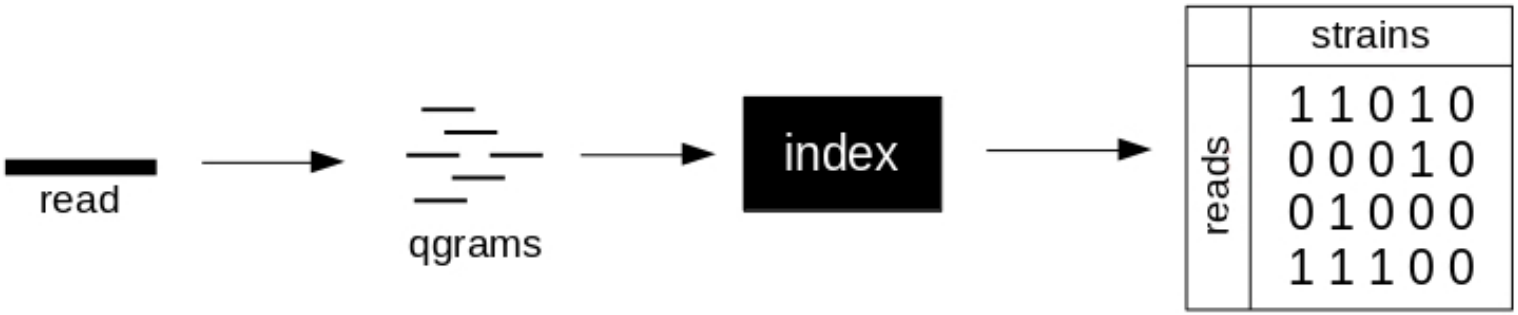
Index in practice

1. 77 *S.thermophilus* genomes + 1 *S. macedonicus* + 1 *L. delbrueckii* subsp. *bulgaricus*:
   - fasta: 143 MB
   - Kraken 2: 18.1 MB
   - ORI : 23 MB
2. 3662 genomes: strains from the Lactobacillales order:
   - fasta.gz: 115 GB
   - Kraken 2: 3.2 GB
   - ORI : 1.4 GB

Scale to all Lactobacillales
Smaller than the smallest state-of-the-art index

Read identification with ORI

Read processing pipeline :



Then find a minimal subset of strains allowing to fully explain the reads.

- Don't use reads found in too many strains (core genome).
- Only the best strains are used (according to the proportion of qgrams matching them in the read).
- Instance of the Set Cover optimization problem, solved using ASP (Answer Set Programming)

Experimental design

Goal

The goal of these experiments was to compare ORI and the most widely used identification method: Kraken 2.

1. Identification of reads from *S. thermophilus* strains.
2. We use the 79 strains index for both methods.
3. 180 identification experiments: 4 parameters with 5 replicates each time.
   - reads number: 1000, 4000, 16000
   - strains number: 4, 6
   - proximity of the strains : Close, moderately close, distant
   - distribution of strains : uniform or dominant and sub-dominant

How to compare Kraken 2 and ORI results ?

Validation of identification methods:

- Matthews correlation coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

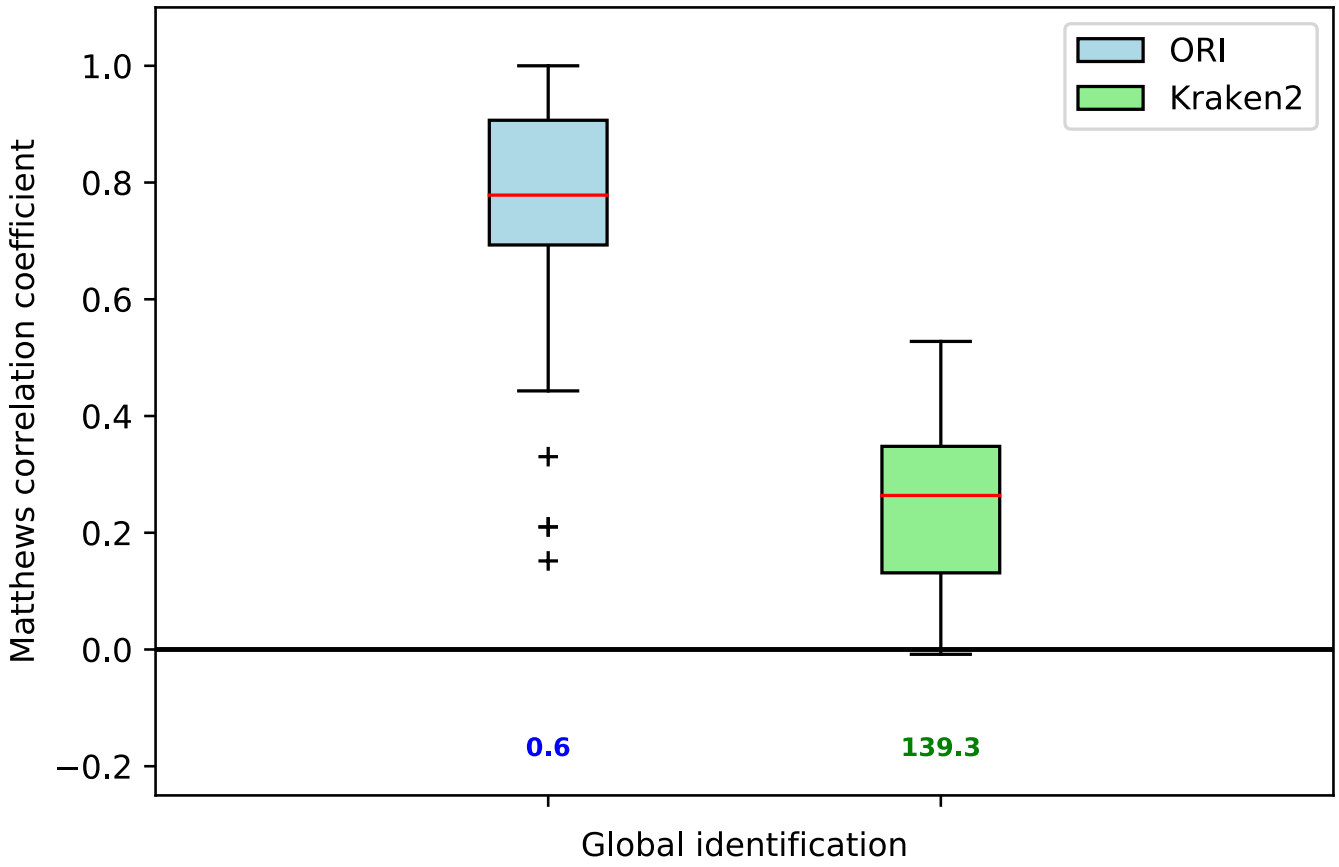
with TP = number of true positives, FP = number of false positives, TN = number of true negatives and FN = number of false negatives.

- Average sum of Hamming distances between predicted and real strains

$$H(G1, G2) = \frac{\text{Number of positions where filters differ}}{\text{Total number of positions}}$$

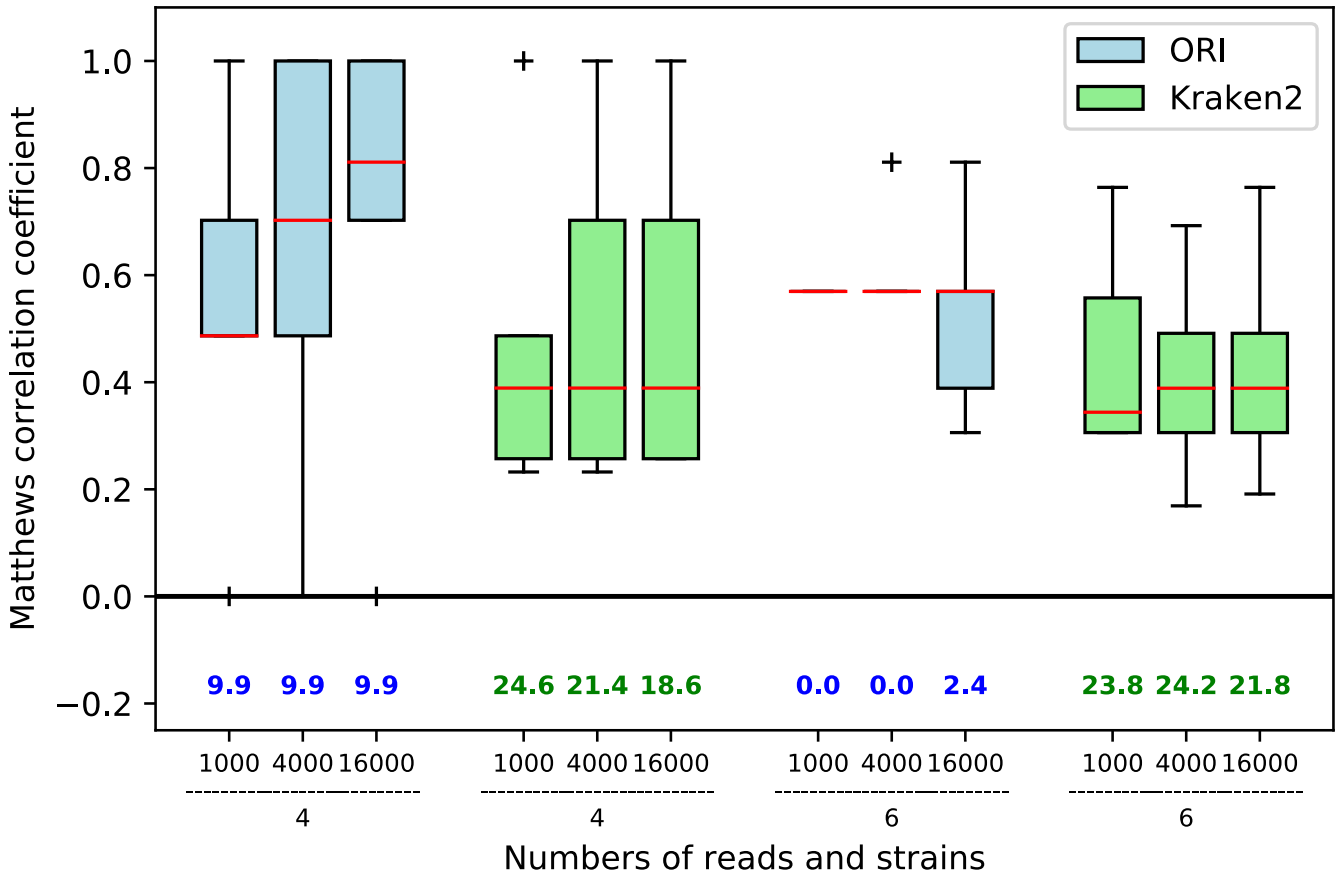
Global identification results: ORI vs Kraken 2

90 experiments containing an equal number of reads for each strain .



Sub-dominant strains identification: ORI vs Kraken 2

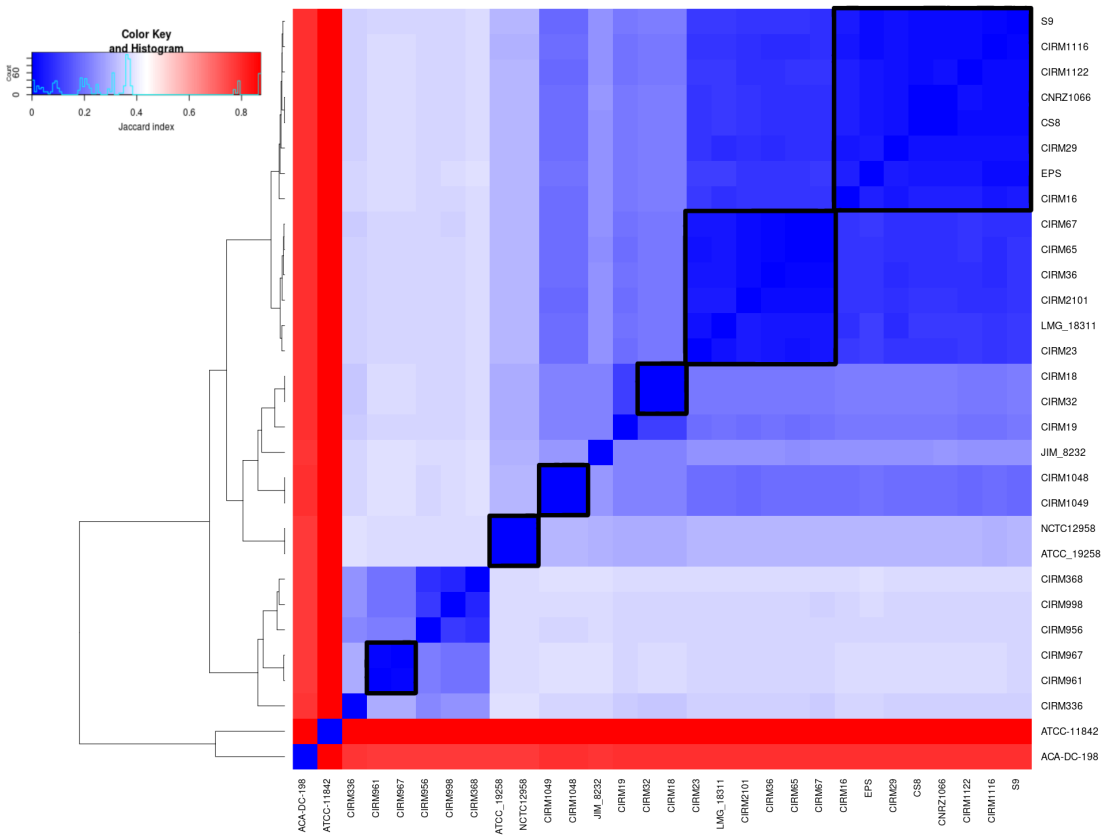
Experiments containing dominant and sub-dominant strains.



Distance between strains

What is a strain ?

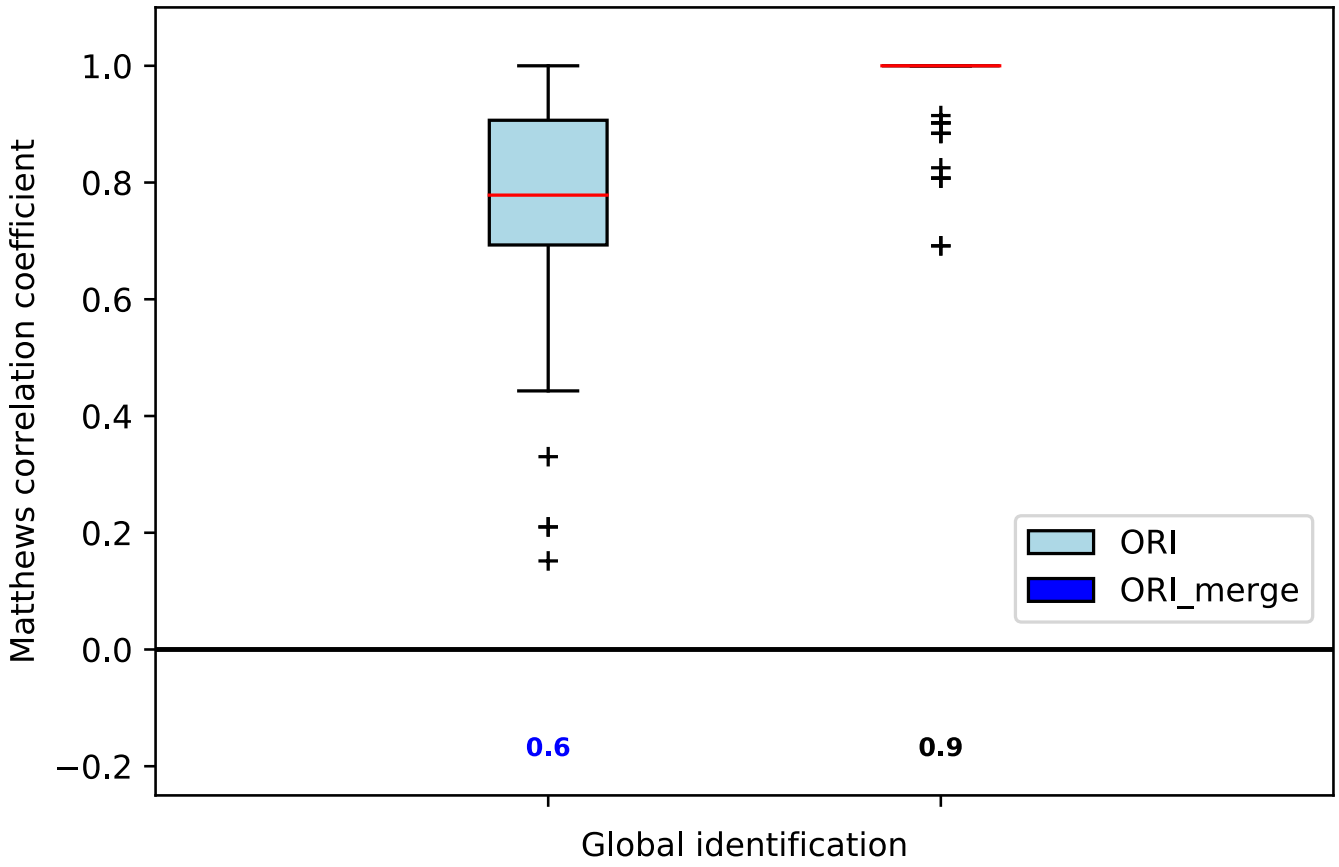
Distinction between strains and mutants is not clear.



Very closely related strains ($H < 2e^{-4}$) have been grouped together.

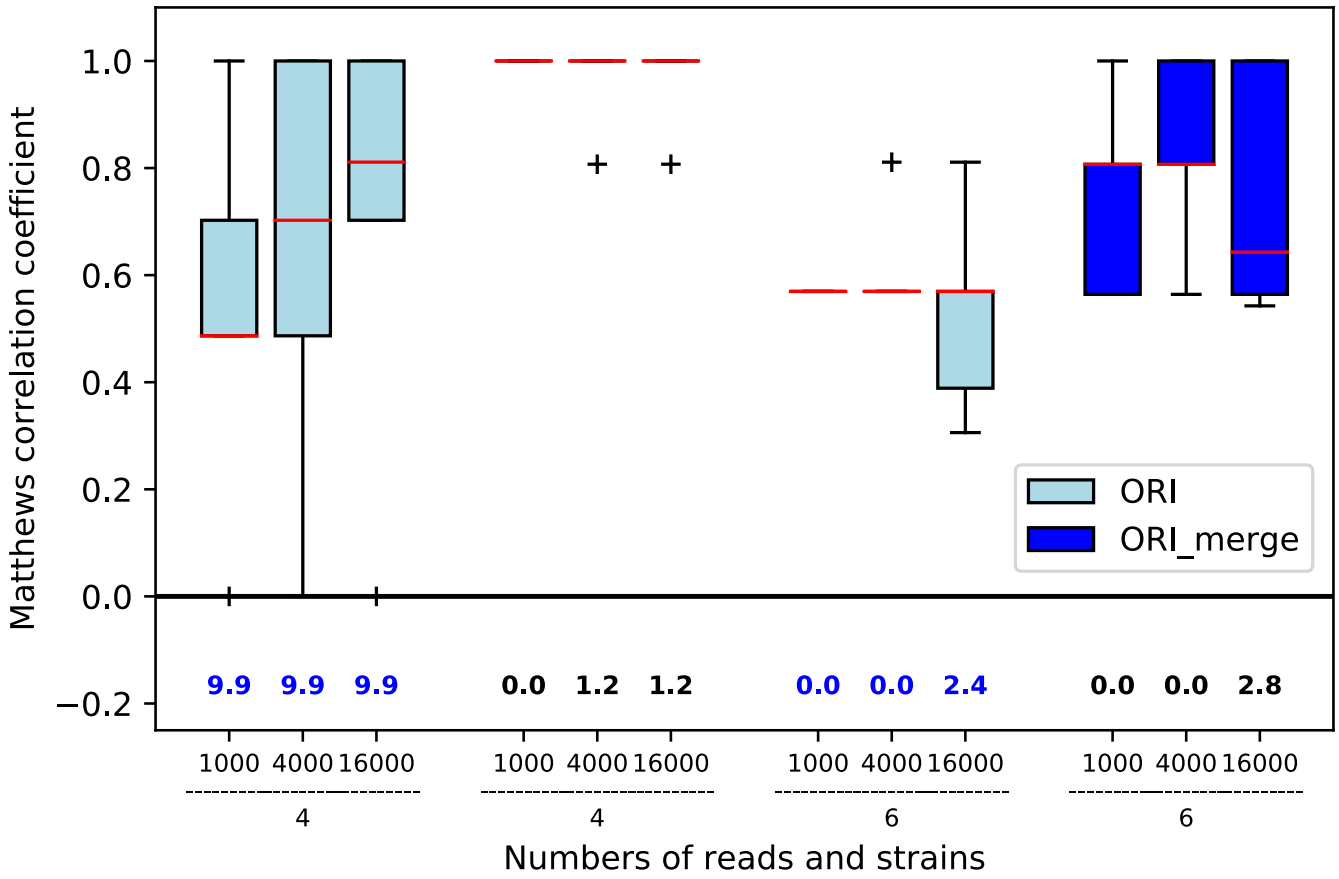
Identification results: ORI vs ORI merge

90 experiments containing an equal number of reads for each strain .



Sub-dominant strains identification: ORI vs ORI merge

Experiments containing dominant and sub-dominant strains.



Conclusion & perspectives

Conclusion:

- ORI better than Kraken2 to identify bacterial strains.
- Merging very close strains seems to be a good idea: very few strains within a cluster have a known gene specific to it.

Perspectives:

- Strain abundance.
- Increase our database to the whole order of Lactobacillales.

The end

Thanks for listening !