Improving microbial taxonomic profiling with long read technologies

Jean Mainguy, Adrien Castinel, Olivier Bouchez, Sylvie Combes, Carole Iampietro, Christine Gaspin, Denis Milan, Cécile Donnadieu, Claire Hoede, Géraldine Pascal

Improving microbial taxonomic profiling with long read technologies

Jean MAINGUY[1], Adrien CASTINEL[2], Olivier BOUCHEZ[2], Sylvie COMBES[3], Carole IAMPIETRO[2], Christine GASPIN[1], Denis MILAN[2,3], Cécile DONNADIEU[2], Claire HOEDE[1] and Géraldine PASCAL[3]

1 INRAE, UR875 MIAT PF Bioinfo GenoToul F-31326 Castanet Tolosan, France
2 INRAE, US 1426, GeT-PlaGe, Genotoul, F-31326, Castanet Tolosan, France
3 GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

Limits of short-read metabarcoding

Metabarcoding main steps:

Usually, a short read PCR amplicon marker (250-500pb) is targeted. In most cases, a small part of the 16S rRNA gene is used.

Problem: **Barcodes** of closely related organisms are often too similar and thus **cannot discriminate** them.

Long read technologies in metabarcoding

PacBio Technologies & Oxford Nanopore Technologies

A longer barcode may include more discriminative information, thus more accurate taxon identification.

Some studies have targeted the full 16S rRNA gene and show encouraging results [1][2].

Problem: **16S** rRNA genes are **present in several copies** and might not be conserved within a given organism.

Our strategy

Are they **better targets** for metabarcoding for long read technologies ?

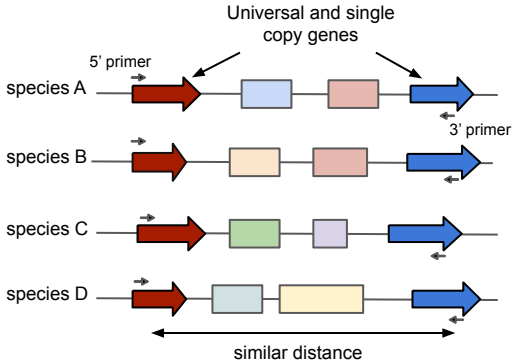
A suitable target should be:

- Universal with conserved regions to design primers
- Diverse and long enough to identify species
- In single copy to have a better estimation of the relative abundance

Our goal is to identify **genomic regions** bounded by two **universal** and **single copy** genes.

Objectives:

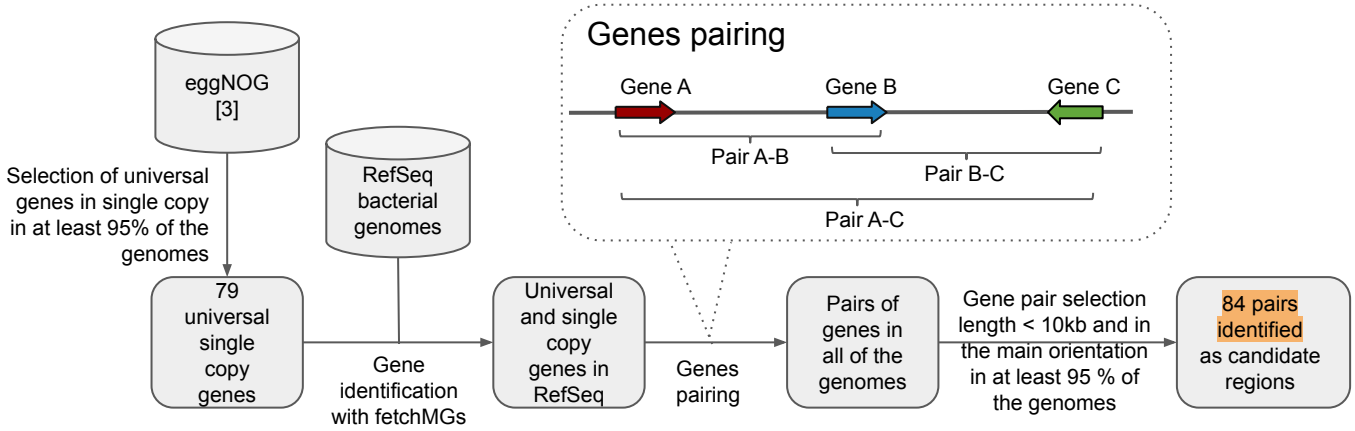
- develop a pipeline to explore databanks and to identify candidate regions
- method to compare candidate regions based on taxonomic resolution
- primers design



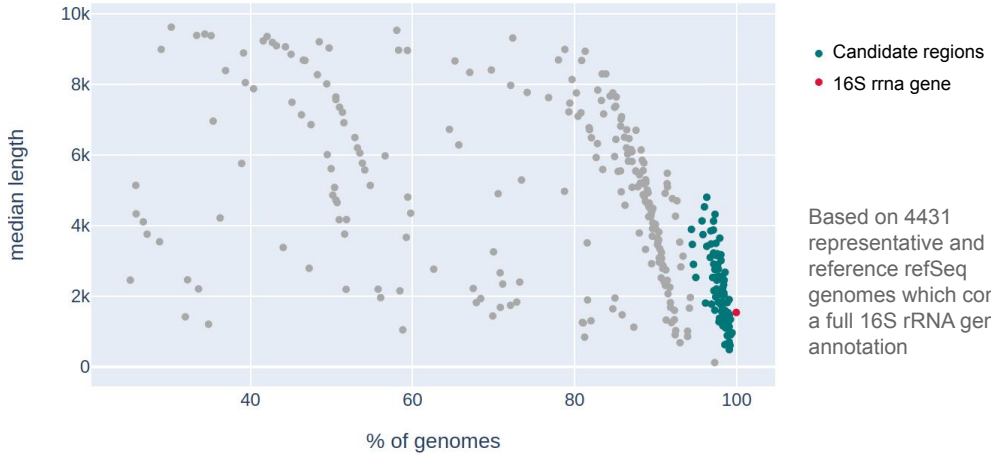
A pipeline to explore databanks and to identify candidate regions

The identification is based on universal single copy genes.

Each possible pair of genes is then checked across refSeq genomes and the ones with consistent length and orientation are selected.



84 candidate regions



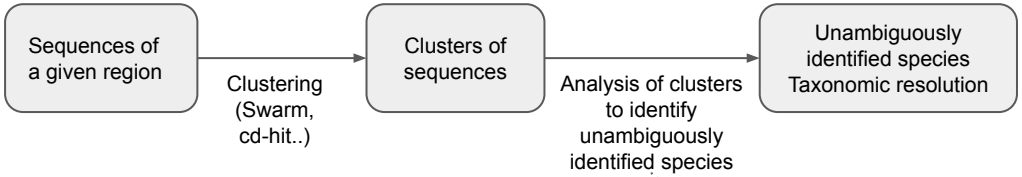
Based on 4431 representative and reference refSeq genomes which contain a full 16S rRNA gene annotation

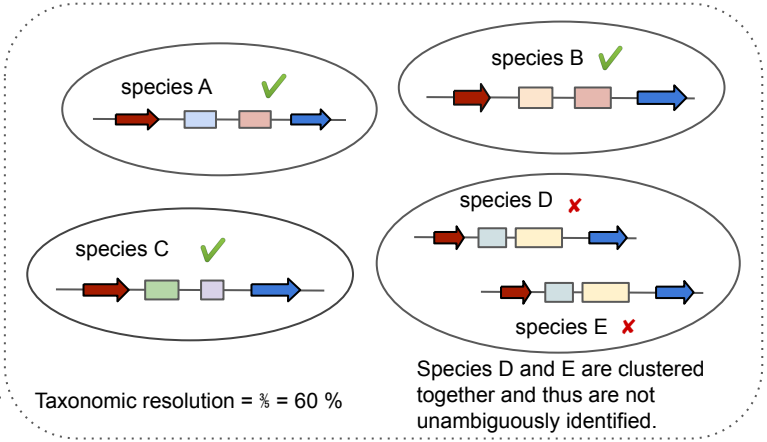
We identified **84 genomic regions** which were found all across the Bacteria kingdom in single copy.

How can we measure the resolution power of these regions to select the best ones?

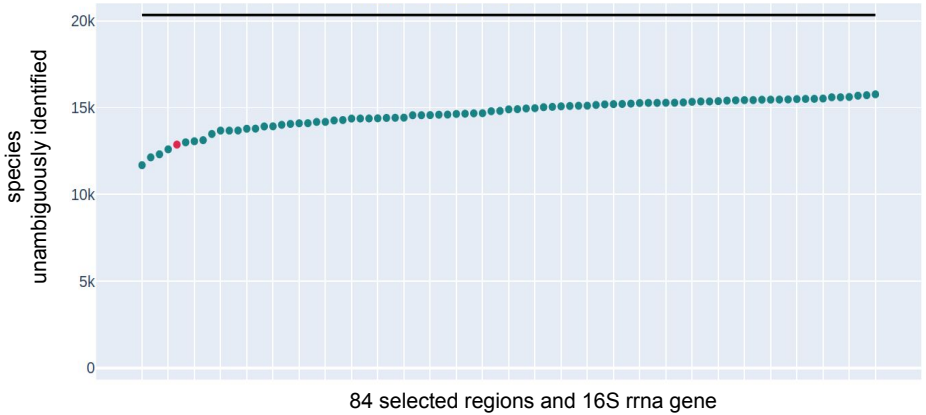
Method to compare candidate regions based on taxonomic resolution

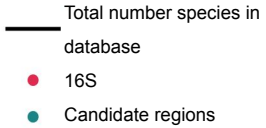
What is the taxonomic resolution of a region?
It is the proportion of species that is unambiguously identified by the region.
A species is **unambiguously identified** when its target sequence is different enough from the other species.





Taxonomic resolution of the 84 regions and of the 16S rRNA gene





Based on a selection of 20 337 RefSeq bacterial genomes.
One genome per species which contains a full 16S rRNA gene annotation.

Our regions allow to unambiguously **identify more species** than the **16S** rRNA gene.
However, important parts of the species are still not correctly identified by any of the regions.

Primer design challenge

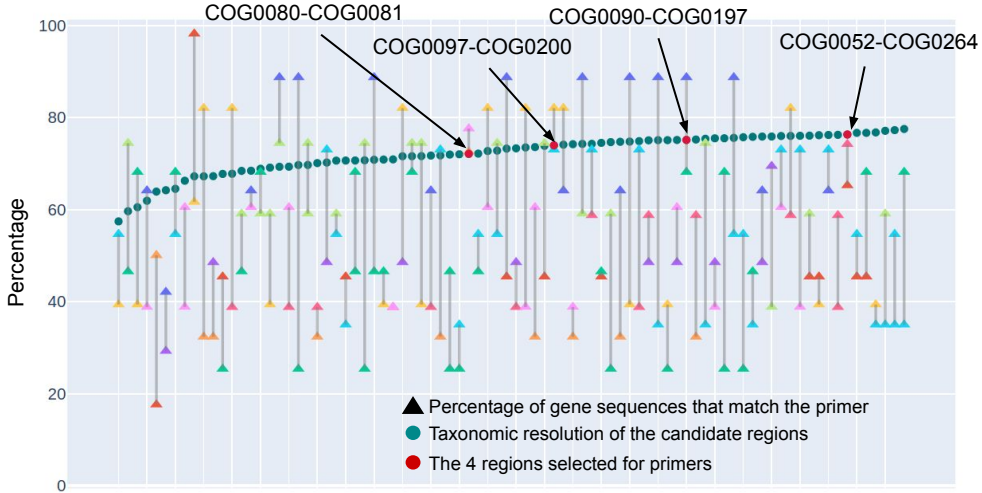
A primer pair is needed to amplify the target region by PCR.

- The pair should be able to amplify the target region in a maximum of species → **taxonomic coverage** of the primers
- It should only amplify the target region → **specificity** of the primers

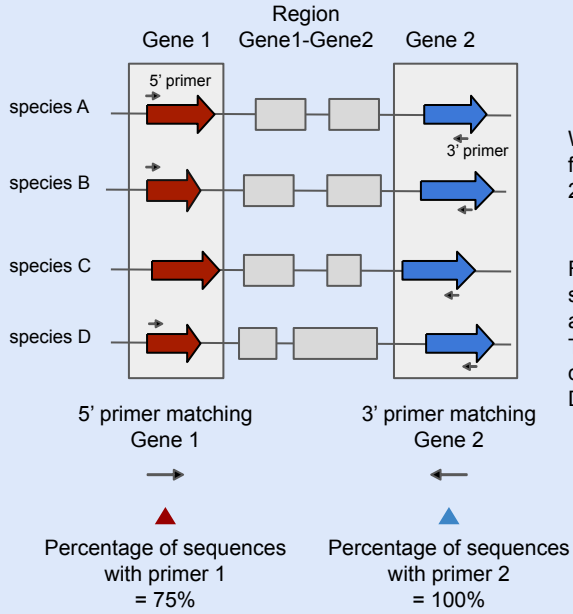
To be able to amplify a maximum number of taxa, we use **degenerate primers** which have several possible nucleotides at some position of their sequences.

Constraint for a good amplification:
→ melting temperature, %GC, GC clamp in 3', inter binding, size of the primers

Primer quality and taxonomic resolution of candidate region



Primer design method



We design primers for Gene 1 and Gene 2 individually.

First, gene sequences are aligned with ClutalO. Then, primers are designed with the Degeprime[4] tool.

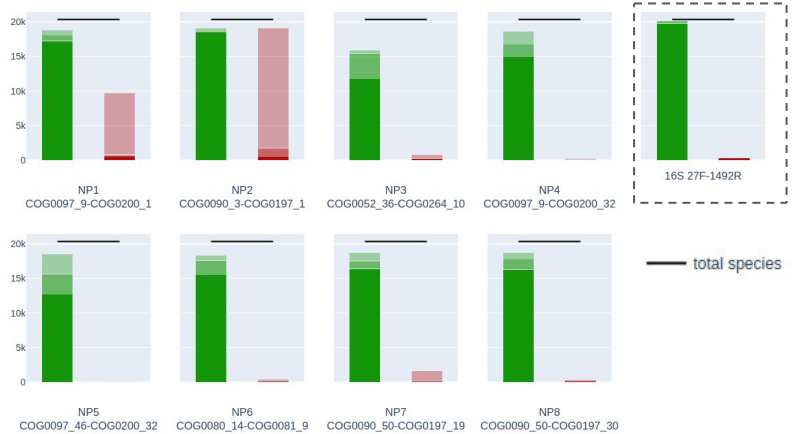
Not all universal genes identified are **suitable for primer design**.
Their sequences vary too much across bacterial taxa.

Primers selection:
We selected 8 primer pairs (in 4 regions).
Our selection is based on:

- the primer ability (*in silico*) to match a maximum of sequences
- the taxonomic resolution of the candidate region
- how well primers follow criteria for a good amplification (in wet lab).

In silico evaluation of selected primers with the ecoPCR [5] tool

Taxonomic coverage and unspecific amplification of the selected primers



Species with expected amplification

- 0
- 1
- 2

Species with unspecific amplification

- 0
- 1
- 2

Number of mismatches between the primer sequence and the DNA matrix. PCR amplification can occur even with few mismatches in the sequence.

There is a great influence of the number of mismatches, especially iwith NP1 and NP2 → *In vitro* PCR and sequencing steps will help us know to which extent primers are going to amplify, despite mismatches.

Conclusion and perspectives

We have developed a method to identify universal single copy regions in bacterial genomes which could be used in long read metabarcoding.

Applied to bacterial genomes:

- Our method enables the identification of 84 regions.
- Primers have been designed to amplify these candidate regions and 8 pairs of them are currently being tested *in vitro*.

Designing universal primers on coding sequences and finding a challenger to the 16S rRNA gene marker are very challenging tasks. But all remains to be concluded from the *in vitro* results.

In parallel, *in silico* investigations continue because our method is not limited to Bacteria but can be used to establish new potential target regions within any prokaryotic taxonomic levels.

References

[1] Johnson, Jethro S., et al. "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis." *Nature communications* 10.1 (2019): 1-11.

[2] Santos, Andres, et al. "Computational methods for 16S metabarcoding studies using Nanopore sequencing data." *Computational and Structural Biotechnology Journal* 18 (2020): 296-305.

[3] Huerta-Cepas, Jaime, et al. "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." *Nucleic acids research* 47.D1 (2019): D309-D314.

[4] Hugerth, Luisa W., et al. "DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies." *Applied and environmental microbiology* 80.16 (2014): 5116-5123.

[5] Ficetola, Gentile Francesco, et al. "An in silico approach for the evaluation of DNA barcodes." *BMC genomics* 11.1 (2010): 434.