Metabarcoding on the deep seafloor : optimizing multigene approaches for large-scale biodiversity assessments

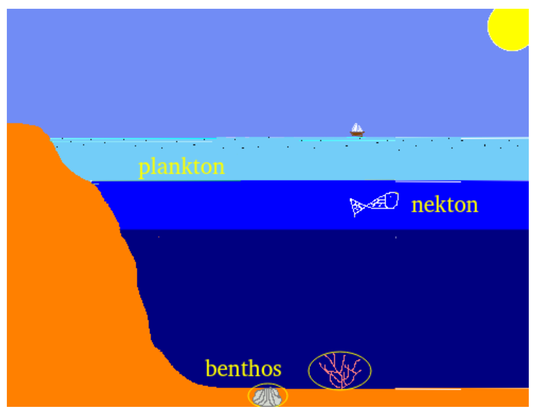
Caroline Dussart [1,3], Miriam Brandt [2], Sophie Arnaud-Haond [2], Laure Quintric [1]

1. Assessing biodiversity in the deep sea

The deep sea :

- the largest and most poorly known biome on Earth
- despite its inaccessibility, it is threatened by human activities

Improved baseline knowledge, and environmental impact assessment protocols, are needed.



→ **Diversity inventories** of the benthos
→ **Protocols** : sampling, sequencing, **bioinformatics**

2. A metabarcoding approach

Barcode genes are genes that can be used for taxonomical assignment of a sample :

- **informative** : much inter-species variation
- **conserved** : little intra-species variation

Metabarcoding : use barcode genes to assess the biodiversity in an environmental sample

Advice : use several barcodes to get a more precise overview of the samples
e.g. 18S, COI

3. The standard bioinformatic process

Cleaning of the reads

- adapter removal
- merging of paired-end reads
- chimera removal

Clustering of reads into OTUs

- Grouping similar reads allows to discard some sequencing errors by keeping only the most abundant sequence as representative of the cluster

Taxonomical assignment

- Comparison of the sequences against databases

4. How well does this workflow apply to the benthos?

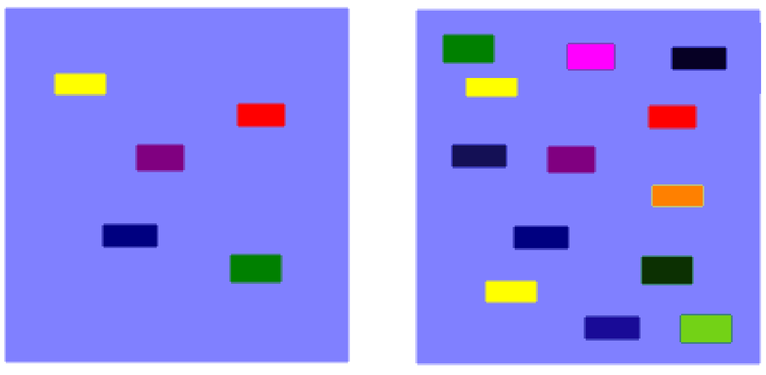
→ to a poorly known biote?
→ to eukaryotes?

5. Assessing the efficiency with control samples

Five control samples were designed, containing DNA of ten common marine species at various quantities. They were analyzed with FROGS [1], a pipeline to perform metabarcoding analyses.

Results:

- Artefactual sample richness, due to sequencing errors
- Inaccurate taxonomical assignments : the OTUs are often assigned to a close species



What we wanted... *... and what we got*

→ **How can we get a more accurate overview of the samples content?**

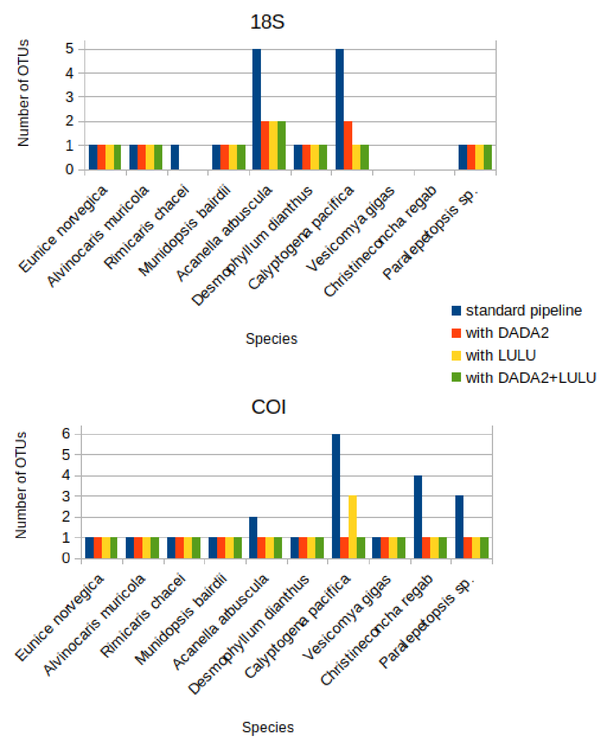
6. Correcting the errors

6a. Correcting the reads

DADA2 [2] compares all reads A and B. Based on their similarity and respective abundances, a Poisson model is used to determine whether they are the same sequence.

The resulting OTUs are:

- far less numerous
- better assigned



Results on a sample containing 10 DNAs, i.e. 10 OTUs (one per species) were expected

6b. Correcting the OTUs

LULU [3] key assumption : two sequence-similar OTUs that have a high rate of co-occurrence among the samples are one OTU.

Results:

- less OTUs
- no loss of species

→ **The corrected pipeline describes more precisely the samples**
→ **The OTUs are still too numerous**, but more stringent parameters cause the loss of information
→ **A single barcode doesn't guarantee a full overview of the diversity**

7. Conclusions

- **Statistical methods are essential to correct the effects of sequencing errors**
- **We remain with OTUs not assigned to the exact species**
  Can't be solved by bioinformatics (problem of primers)
- **It is desirable to lead both morphological and bioinformatic analyses as complementary approaches** when studying a new kind of ecosystem

8. Perspectives

- tests on RNA and size-selected DNA (lacking extracellular DNA fragments)
- add MACSE [4] to remove numts (nuclear copies of the mitochondrial gene of interest)
- use the new pipeline on "Pourquoi pas les Abysses?" data

9. References

[1] F. Escudié, L. Auer, M. Bernard, M. Mariadassou, L. Cauquil, K. Vidal, S. Maman, G. Hernandez-Raquet, S. Combes, G. Pascal; FROGS: Find, Rapidly, OTUs with Galaxy Solution, *Bioinformatics*, Volume 34, Issue 8, 15 April 2018, Pages 1287–1294, https://doi.org/10.1093/bioinformatics/btx791.
[2] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. Dada2 : High resolution sample inference from illumina amplicon data. *Nature Methods*, July 2016.
[3] T. G. Frøslev, R. Kjøller, H. H. Bruun, R. Ejrnæs, A. K. Brunbjerg, C. Pietroni, and A. J. Hansen. Algorithm for post-clustering curation of dna amplicon data yields reliable biodiversity estimates. *Nature Communications*, October 2017.
[4] V. Ranwez, S. Harispe, F. Delsuc, and E. J. P. Douzery. Macse : Multiple alignment of coding sequences accounting for frameshifts and stop codons. *Plos One*, September 2011

[1]Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER)
Cellule Bioinformatique, Brest, France
[2]MARine Biodiversity, Exploitation and Conservation (MARBEC), Sète, France
[3]Master Bioinformatique, Normandie Université, UNIROUEN