OCToPUS: a comprehensive pipeline for 16S rRNA metagenetic data analysis.

Mohamed Mysara[1], **Natalie Leys**[1], **and Pieter Monsieurs**[1]

[1] *Unit of Microbiology, Belgian Nuclear Research Centre (SCK-CEN), Mol, Belgium.*

Abstract

Bacteria plays a vital role in industry, from the production of traditional foods such as yoghurt, cheese, and vinegar; to the production substances such as drugs and vitamins. They also affect the human health, being involved in the immune responses and food digestion and contributing to the pathogenicity of various diseases. Several important aspects of bacteria in various environments are poorly understood, making the identification and monitoring microbial communities of utter importance to the industry and patient care. Recent developments in new high-throughput sequencing technologies have revolutionized molecular biology, including the study of microorganisms without the need for culturing them in the lab, an approach often referred to as **metagenomics**. Such **metagenomics** applications allow the simultaneous high-throughput analysis of genetic material of most of the microbes present in a given sample, without the need for culturing the bacteria first. Although this approach has nowadays been adopted in many project, it is far from straightforward necessitating various **bioinformatics** optimization.

In this work, we aimed at optimizing and standardizing the analysis of 16S metagenomics with a final aim to develop a pipeline to starts from the raw sequencing reads and deliver their microbial composition. Therefore, different tools were developed to deal with chimera (Mysara, Saeys, *et al.*, 2015) and sequencing errors (Mysara, Leys, *et al.*, 2015; Mysara *et al.*, 2016), each of them found to outperform the other existing state-of-the-art tools. Additionally, a new method was introduced to bring closer correspondence between the number of microorganisms detected and the actual diversity within the samples (Mysara, Vandamme, *et al.*, 2017). A one stop-shop software, named OCToPUS, assembles these various algorithms, thereby leading to a highly accurate assessment of microbial diversity starting from the raw sequencing reads (Mysara, Njima, *et al.*, 2017).

These tailored algorithms have already been successfully applied to assess complex microbial communities in a wide range of environments such as deep subsurface geological clay formations (**HADES, Boom clay, Belgium; and Mont Terri**, Opalinus clay Swiss, Moors *et al.* 2013), the human gut microbiome after radiotherapy treatment, and **cooling water circuits of nuclear reactors** (Props *et al.*, 2016), as such revealing microbial diversity at **unseen depth**.

References

Moors H, Cherkouk A., Mysara M, Bleyen N., Boven P, Selenska-Pobell S, et al (2013).BN Experiment: Intermediate results of the microbiological analyses, Belgium: Mont Terri consortium. 68 p(Technical Note)

Mysara, M., Saeys, Y., *et al.* (2015) 'CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies.', *Applied and environmental microbiology*, 81(5), pp. 1573–84. doi: 10.1128/AEM.02896-14.

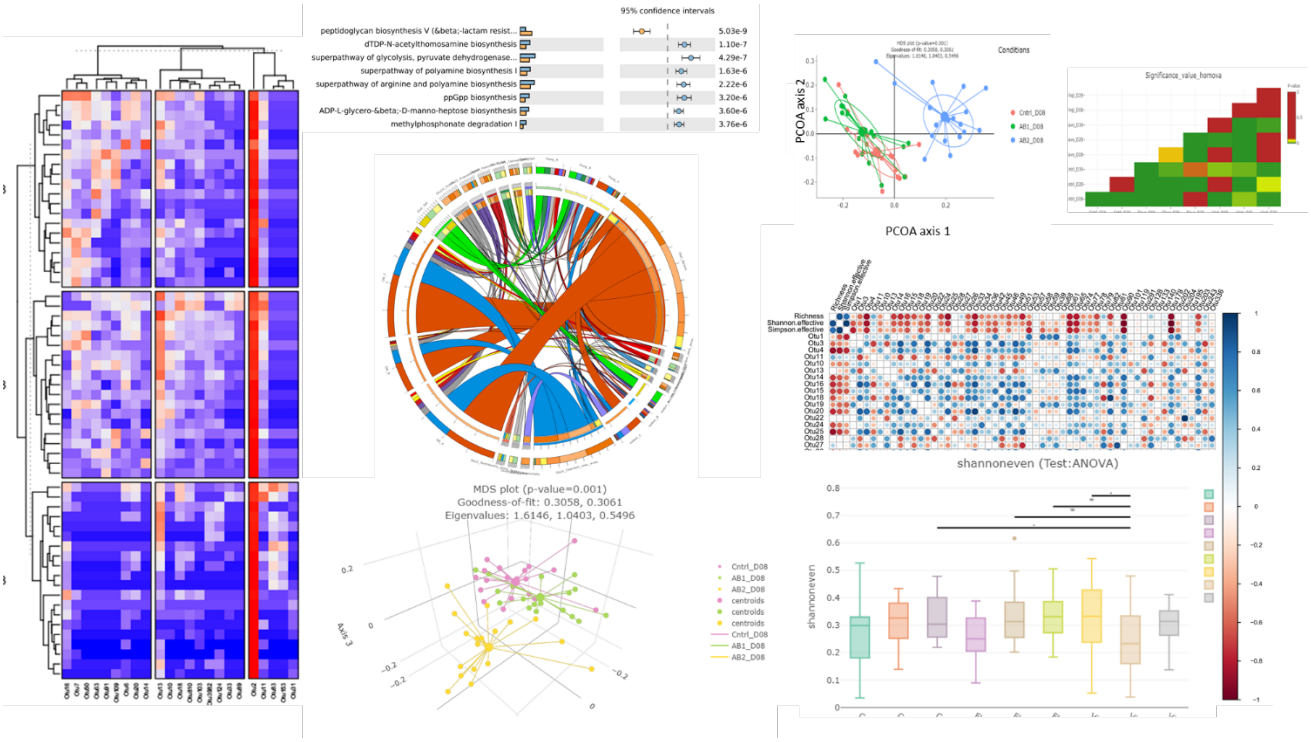
Mysara, M., Leys, N., *et al.* (2015) 'NoDe: a fast error-correction algorithm for pyrosequencing amplicon reads', *BMC Bioinformatics*, 16(1), p. 88. doi: 10.1186/s12859-015-0520-5.

Mysara, M. *et al.* (2016) 'IPED: a highly efficient denoising tool for Illumina MiSeq Paired-end 16S rRNA gene amplicon sequencing data', *BMC Bioinformatics*, 17(1), p. 192. doi: 10.1186/s12859-016-1061-2.

Mysara, M., Njima, M., *et al.* (2017) 'From reads to operational taxonomic units: an ensemble processing pipeline for MiSeq amplicon sequencing data', *GigaScience*. GigaScience Database. doi: 10.5524/100265.

Mysara, M., Vandamme, P., *et al.* (2017) 'Reconciliation between Operational Taxonomic Units and Species Boundaries', *FEMS Microbiology Ecology*. doi: 10.1093/femsec/fix029.

Props, R. *et al.* (2016) 'Measuring the biodiversity of microbial communities by flow cytometry', *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.12607.


95% confidence intervals
peptidoglycan biosynthesis V (β-lactam resist...
dTDP-N-acetylthomosamine biosynthesis
superpathway of glycolysis, pyruvate dehydrogenase...
superpathway of polyamine biosynthesis I
superpathway of arginine and polyamine biosynthesis
ppGpp biosynthesis
ADP-L-glycero-β-D-manno-heptose biosynthesis
methylphosphonate degradation I
5.03e-9
1.10e-7
4.29e-7
1.63e-6
2.22e-6
3.20e-6
3.60e-6
3.76e-6
Conditions
Cntrl_D08
AB1_D08
AB2_D08
PCOA axis 1
PCOA axis 2
Significance_value_homova
MDS plot (p-value=0.001)
Goodness-of-fit: 0.3058, 0.3061
Eigenvalues: 1.6146, 1.0403, 0.5496
shannoneven (Test:ANOVA)
shannoneven