From Reads to OTUs
Improved Algorithms for Preprocessing Amplicon Sequencing data

Mohamed Mysara[1,2,3], Natalie Leys[1], Jeroen Raes[2,3,4] and Pieter Monsieurs[1]
[1] Unit of Microbiology, Belgian Nuclear Research Centre (SCK-CEN), Mol, Belgium.
[2] Department of Bioscience Engineering, Vrije Universiteit Brussel, Brussels, Belgium.
[3] VIB Center for the Biology of Disease, VIB, Leuven, Belgium.
[4] Department of Microbiology and Immunology, REGA institute, KU Leuven, Belgium.

A major breakthrough in microbial ecology has been realized by clonal amplification of the 16S or 18S rRNA gene for the assessment of microbial diversity in a specific environment, thereby omitting the time-consuming and challenging culturing approach. This approach has been accelerated via the introduction of high-throughput sequencing technologies, leading to a dramatic increase of marker gene sequencing studies for the assessment of microbial communities.

In the most straightforward approach, the reads from different samples are pre-processed and clustered based on their sequence similarity to each other, commonly named operational taxonomic units (OTUs) approach. Several algorithms and pipelines were proposed to address raw data pre-processing originating from different sequencing platforms, including 454 GS-FLX, 454 GS-FLX +, MiSeq, and PacBio. In order to assess the suitability of each of the technologies, we used different mock communities, i.e. samples with a known composition (varying between 16 and 60 species), either produced in-house or obtained from publically available samples. Regardless of the sequencing platform used, all technologies suffer from the presence of erroneous sequences: (i) chimera, i.e. artificial (non-biological) sequences mainly introduced by the PCR reaction during sample preparation, and (ii) sequencing errors produced by the sequencing platform itself. For both types of sequencing errors, we developed novel preprocessing algorithms to remove or correct erroneous reads.

First, a machine learning method called CATCh (Combining Algorithms to Track Chimeras) is developed which is able to integrate the output of existing chimera detection tools into a new more powerful method. Via a comparative study with other chimera detection tools, CATCh was shown to outperform all other tools, thereby predicting up to 9% more chimera than could be obtained with the best individual tool (Fig 1). **Second**, NoDe (Noise Detector) was introduced as an algorithm to correct existing 454 pyrosequencing errors, thereby decreasing the number of reads and nucleotides that are disregarded by the current state-of-the-art denoising algorithms (Fig 2). NoDe was benchmarked against state-of-the-art denoising algorithms, thereby outperforming all other existing denoising tools in reduction of the error rate (reduction of 75%), and decrease in computational costs (15 times faster than the best individual tool). **Third**, as the 454 pyrosequencing platform is in many microbial diversity assessments replaced with the more cost-effective Illumina MiSeq technology, the IPED (Illumina Paired End Denoiser) algorithm was developed to handle error correction in Illumina MiSeq sequencing data as the first tool in the field. It uses an artificial intelligence-based classifier trained to identify Illumina's error and remove them, reducing the error rate by 73% (Fig 2).

The combined effect of improved algorithms for chimera removal and error correction had a positive effect on the clustering of reads in operational taxonomic units, with an almost perfect correlation between the number of clusters and the number of species present in the mock communities. Indeed, when applying our improved pipeline containing CATCh and NoDe on a 454 pyrosequencing mock dataset, our pipeline could reduce the number of OTUs to 28 (i.e. close 18, the correct number of species present in the 454 pyrosequencing mock community). In contrast, running the straightforward pipeline without our algorithms included would inflate the number of OTUs to 98. Similarly, when tested on Illumina MiSeq sequencing data obtained for a mock community, using a pipeline integrating CATCh and IPED, the number of OTUs returned was 33 (i.e. close to the real number of 21 species present in the Illumina MiSeq mock community), while a much higher number of 86 OTUs was obtained using the default mothur pipeline. Our algorithms are freely available, via our website: http://science.sckcen.be/en/Institutes/EHS/MCB/MIC/Bioinformatics/ and can easily be integrated into other 16S rRNA data analysis pipelines (e.g. mothur).

Reference

- *Mysara M., Leys N., Raes J., Monsieurs P.- NoDe: a fast error-correction algorithm for pyrosequencing amplicon reads.- In: BMC Bioinformatics, 16:88(2015), p. 1-15.- ISSN 1471-2105*
- *Mysara M., Saeys Y., Leys N., Raes J., Monsieurs P.- CATCh, an Ensemble Classifier for Chimera Detection in 16S rRNA Sequencing Studies.- In: Applied and Environmental Microbiology, 81:5(2015), p. 1573-1584.- ISSN 0099-2240*
- *Mysara M., J. Raes, N. Leys, P. Monsieurs, 2015,* IPED: A highly efficient denoising tool for Illumina paired-end 16S rRNA amplicon sequencing data, PLOS Computational Biology, **submitted**.

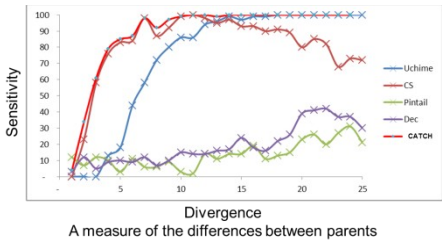

Figure 2 Plot indicating the effect of having 5% indels on the sensitivity of different tools.

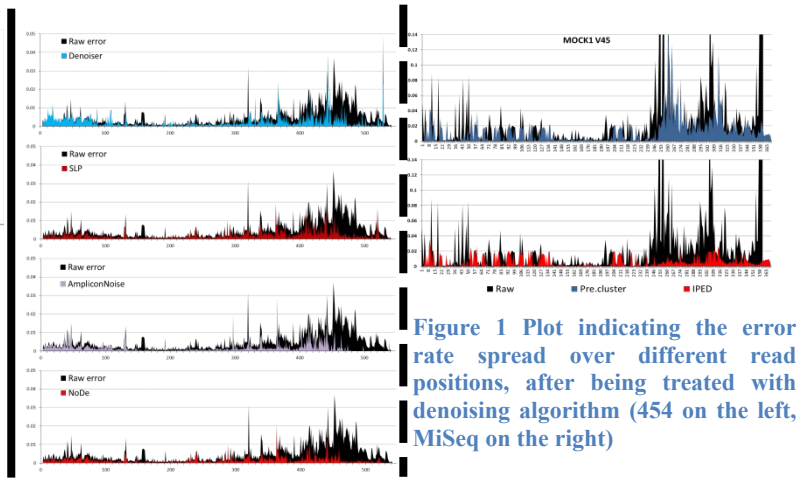

Figure 1 Plot indicating the error rate spread over different read positions, after being treated with denoising algorithm (454 on the left, MiSeq on the right)