Fast and accurate pathogen identification with NGS 16S amplicon analysis in the cloud

Dimitrios Arnellos[1], Martin Vondracek[2], Fredrik Dyrkell[1]

[1]1928 Diagnostics AB – Gothenburg (Sweden), [2]Division of Clinical Microbiology, Department of Laboratory Medicine, Karolinska Institutet, Karolinska University Hospital – Stockholm (Sweden)

Background

Next generation sequencing (NGS) of the 16S rRNA gene has emerged as a widely used method for bacterial identification, due to being fast and accurate, as well as avoiding a culturing step. In addition this facilitates analysis of an extended range of clinically relevant bacteria that are slow-growing, fastidious or difficult to grow. The cloud-based 1928 platform analyses 16S amplicon NGS sequences (FASTQ files) to classify genomic data in order to identify bacteria. The platform allows a fully automated analysis workflow, supporting major sequencing platforms. The aim of this project was to compare 1928's amplicon pipeline to the DADA2 open source software, evaluate performance, and analyze clinical samples from Karolinska University Hospital to compare two sets of primer regions.

Methods

A new pipeline was developed, where reads were cleaned of adapter sequences using Cutadapt (3.4) and subsequently VSEARCH (v2.7.0) used for merging pair-end reads (Illumina only), filtering out chimeras, filter on amplicon length and finally dereplication. A custom developed denoising algorithm was implemented, based on methods from Deblur and USEARCH-UNOISE3 but adapted to also support both Illumina and IonTorrent data. Finally, identified ASVs were used to create abundance estimations from the original amplicon sequences, and taxonomic classification with BLAST (v2.6.0) was done against SILVA (v138) with an identity threshold of 0.99.

Results

Two publicly available mock community datasets were analysed, in addition to 5 clinical samples.

Table 1

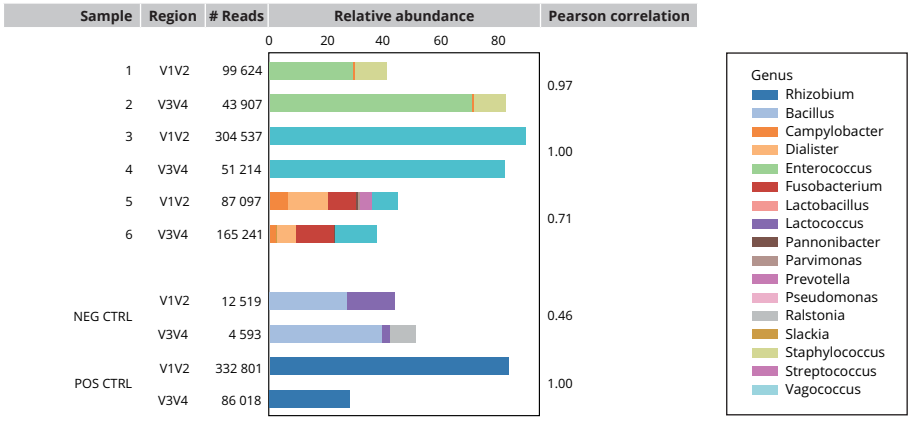
Mock community sample	16S region/ Platform	Method	Precision %	Recall %	Runtime (s)	Peak memory (GB)
Kozich et al. (Appl Environ Microbial 2013)	V4 Illumina	1928	89.5	85.0	117	1.481
		DADA2	87.5	70.0	244	3.273
Kleiner et al. (Nat Commun 2017)	V3V4 Illumina	1928	63.2	52.5	317	1.550
		DADA2	61.5	33.3	706	3.873

DADA2 results with default settings against the SILVA database (v138) using the Naive Bayes classifier. Sequential analysis of each sample was performed on a Google Cloud n1-standard-8 instance, using GNU time (1.7) for time and resource analysis. Precision and recall determined from species level taxonomic classification.

Conclusions

The 1928 platform is a fast and accurate way to perform 16S amplicon analysis, with a good tradeoff between precision and recall, supporting multiple sequencing platforms including multiple 16S regions for analysis.

Image 1



Conflict of interest statement

D.A. and F.D. are employees of 1928 Diagnostics AB.

Contact

dimitrios.arnellos@1928diagnostics.com