COMPARISON OF COMPUTATIONAL METHODS FOR ANALYSING 16S RRNA SEQUENCING DATA PRODUCED WITH THE USE OF SHORT AND LONG READS

M.M. Shtilkind

Al-Farabi Kazakh National University, Republic of Kazakhstan, 050040, Almaty, Al-Farabi Avenue 71, e-mail: maxx.xxonsh@gmail.com

This research is driven by the importance of identifying species in microbial community for use in medical testing. Conventional methods including culturing, biochemical assays are often limited to the culturable organisms and do not provide information about relative abundance of different microbial taxa within a community. Comparative analysis of metagenomic programs method was used to identify the most appropriate option for long-read 16S rRNA data.

Since the development of high-throughput sequencing technologies, novel methods of identifying species include taxonomic profiling of 16S rRNA. 16S rRNA gene contains variable regions (V1-V9) that are specific to different taxonomic groups. Illumina short-read sequencing technology provides read lengths 2*251 base pairs (bp) for paired-end sequencing, while fraction of 16S rRNA gene is around 1500 bp, thus reads can successfully cover 2 of 9 regions (300-400 bp), but not the whole gene.

The Oxford Nanopore Technology (ONT) representing long-read sequencing or so-called 3rd generation sequencing is much more applicable for 16S rRNA profiling, as it can provide reads with length 1500 bp (exactly in line with the length of 16S rRNA gene). Extended read length provides high alignment quality making the downstream taxonomy classification and relative abundance calculation much easier and relevant.

The samples from endometrium of uterus were studied in the first part of the research. Endometrial microbiome composition and balance can play an important role in the success of in vitro fertilization (IVF).

Taxonomical classification was tested with NCBI, Silva databases. NCBI database consists of full genomes of Bacteria, Archaea and Eukaryote, Silva contains curated and annotated sequences of only small subunit ribosomal RNA of these domains. Among the enormous collection of metagenome classifiers there were selected three particular programs applying completely different algorithms: Emu, PathoScope and Kraken2.

The comparative process was divided in three phases. In the first one, pipelines were run with three programs and 11 samples using Silva and NCBI databases, the results were compared both quantitively and visually. In the second phase, the results were compared to the short-read taxonomical classification, the discrepancies were analyzed and questions were put together for further investigation. During the third phase pipelines with best performed programs were run on sample with acknowledged microbial diversity and following metrics were assessed and visualized: True positive, False positive, Precision, Recall, F1-score, L1-norm, L2-norm.

Results: After the first phase it was revealed that Kraken2, Emu, PathoScope produce very similar results, while MetaScope discrepancies exceeded substantially 10% threshold for 70% of the reads. Additionally, Kraken2 annotated significantly more genera upon the threshold than other programs indicating its low specificity. Consequently, these two programs were excluded from further analysis. The second phase brought to the conclusion that short-read sequencing reports reveal only one prevailing genus in the community, while long-read establish 3 of them, thus it was assumed that short-read classification is much less sensitive. In the third phase it was determined that Emu and PathoScope have least deviation from reference value of abundances (ground truth) and they successfully annotate all the species introduced in the sample.

Therefore, the main outcome of the research is the recommendation to use Emu and PathoScope programs for the purpose of identifying species using long-read data.