IMPACT OF 16S DATABASE CURATION IN DOWNSTREAM MICROBIOME ANALYSES

M. Soverini[a*] and A. Castagnetti[a]

[a]Wellmicro, via Antonio Canova 30, 40138 Bologna, Italy

*matteo.soverini@wellmicro.it

Background:

Microbial identification is pivotal in microbial community analysis. With the advent of Next-generation sequencing techniques[1], it has been necessary to use increasingly refined and complete databases to uniquely classify each taxonomic unit obtained. Despite this relevance, there are only a few public 16S databases currently available for microbial identification[2,3,4], all containing inconspicuous taxonomical information, thus underpowering further downstream analyses on microbiome data.

Methods and Results:

Here we compared the performance of the available databases with a novel database of our own creation and curation. This database, namely the WellMicro database (WMdb), was created by merging all the unambiguous reference 16S reads publicly available on multiple databases. To obtain an updated, correct, and concordant taxonomic lineage for each sequence a custom-made pipeline for complete taxonomic assignation was designed and applied. All the reads that did not have a complete lineage from kingdom to species or other ambiguous taxon classification were discarded at the end of the process. The remaining conflicts in the database were solved through manual curation, obtaining a total of 212,476 reference reads encompassing 12,843 different Archaeal and bacterial species. Microbial taxa identified using this approach were used in downstream machine-learning analyses, allowing us to successfully highlight bacterial consortia and markers linked to specific pathologies at the species level (data not shown).

Conclusions and Significance:

Taken together, the results show that database curation is a crucial process to obtain a solid taxonomic assignment even at lower phylogenetic levels such as family, genus and species using 16S rRNA databases. The large number of sequences with unmatched or ambiguous taxonomy contained in publicly available databases such as Silva or Greengenes limits the possibility of resolution at a lower taxonomic level, impairing the classification performance and compromising the opportunity to explore deeper the microbial communities' data obtained from 16S surveys.

Keywords:

Gut microbiome, Database curation, Amplicon sequencing

References:

1. Wetterstrand, K. A. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). *National Human Genome Research Institute* [online]
2. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069-72. doi: 10.1128/AEM.03006-05
3. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan;41(Database issue):D590-6. doi: 10.1093/nar/gks1219. Epub 2012 Nov 28.
4. https://www.ncbi.nlm.nih.gov/nuccore?term=33175%5BBioProject%5D+OR+33317%5BBioProject%5D (visited 11/11/2022)

Thematic Area:

- Frontiers in Microbiome Research