http://dx.doi.org/10.7124/bc.000AFD

Assessing microbial genome representation across various reference databases: A comprehensive evaluation

G. Boldirev[1], N. Sharma[2], V. Munteanu[3], A. Bhavatharini[2], D. Koslicki[4],
A. Zelikovsky[1], S. Mangul[2]

[1] Georgia State University
Atlanta, GA, USA, 30302,
[2] University of Southern California
3470, Trousdale Parkway, Los Angeles, CA, USA, 90089
[3] Technical University of Moldova
168, Stefan cel Mare Blvd., Chisinau, Republic of Moldova, MD-2004
[4] Pennsylvania State University
201, Old Main, University Park, PA, USA, 16802
grigore.boldirev@gmail.com

Background/Aim. Metagenomics research can provide significant insights into the composition, diversity and functions of mixed microbial communities found in various environments. To identify bacterial species, reads from samples are mapped to references that are found in bacterial reference databases. Multiple references may be assigned the same taxonomic identifiers yet these references may contain different genomic information. This project was designed to uncover and correct inconsistencies in bacterial reference databases by comparing species names and genomic representation for the three most commonly used bacterial reference databases (PATRIC, RefSeq and Ensembl). Our first study "Improving the usability and comprehensiveness of microbial databases" [1] considered the concordance of the databases based solely on species names. We extended that research to compare not only the species names but also bacterial genomes and to estimate their similarity. **Methods.** NCBI's taxonomic identifiers were utilized to assess the agreement of reference databases at the species rank. We proceeded with comparing the genomes for the species that are present in at least two databases. Same species across two databases were identified by finding the same taxID in two databases. Comparison of genomic representation across databases was performed using the BLAST tool. After finding the exact same strain, all the contigs from one database were aligned to all contigs from another. This analysis was extended to all overlapping species where strain information was available. **Results.** A comparative assessment of species names shared across three bacterial databases (Ensembl, RefSeq, PATRIC) indicated that 14.27% of species are present in all three databases. 49.71% of bacterial species are found only in two databases among which PATRIC and RefSeq share 44.97% of the species and 4.72% are common between PATRIC and Esembl. 36% of species are found only in one database where 30% are exclusively found in PATRIC, 4.7% in RefSeq and 1.39% in Ensembl. To compare genomic representation, we visualized the gathered data on all observed alignment cases showing that quality of reference genomes can be improved through consolidation of contigs. **Conclusions.** The lack of species and genus overlap not only undermines the accuracy of metagenomic analysis but also emphasizes the critical need for a standardized integration of existing databases. Our analysis will not only enhance the identification and characterization of microbial life but also improve the comparability and rigor of metagenomic research.

Keywords: metagenomics, bacterial reference databases, taxonomic discrepancies.

REFERENCES

1. *Loeffler C et al.* Improving the usability and comprehensiveness of microbial databases [published correction appears in *BMC Biol.* 2020; **18**(1):92]. *BMC Biol.* 2020; **18**(1):37.