SEG 2021
XLII CONGRESS
14-18 June

Benchmarking Conventional Reference Genome Databases in Human Gut Metagenomics

Samuel Piquer-Esteban[a,b], Wladimiro Diaz-Villanueva[a], Vicente Arnau[a], Andrés Moya[a,b,c]

a. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
b. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
c. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain

INTRODUCTION

One of the key elements of any metagenomic classifier is its database of reference microbial genomes. Among them, for shotgun metagenomics, the most popular reference databases are the BLAST nucleotide collection (NT Database) and the Reference Sequence Database (RefSeq Database) for high-quality nucleotide sequences (1,2). Despite their importance, there are no available comparisons in terms of classification performance between them.

In the present work we have carried out a comparative study of the most popular conventional reference genome databases in terms of classification capacity and database performance in a case of special interest such as the human gut microbiota.

MATERIALS AND METHODS

Conventional databases. We downloaded the NT and RefSeq Databases in January 2021. In the case of the RefSeq Database, we downloaded all available genomes at assembly levels Complete Genome (CG) and Chromosome (Chr) for Archaea, Bacteria, Fungi, Protozoa and Viral organisms.

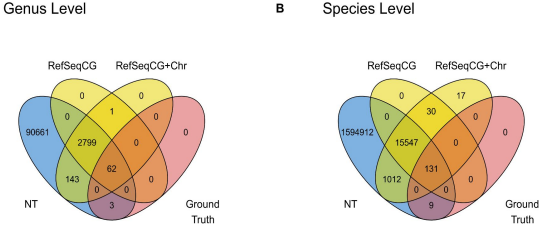
Samples. We downloaded two publicly available studies of human gut shotgun metagenomic samples, 57 samples from Cameroon (PRJEB27005) and 60 samples from Germany (PRJEB27928, control samples only). Additionally, we downloaded 10 gastrointestinal tract simulated Illumina HiSeq metagenome samples from "2ndCAMI Toy Human Microbiome Project Dataset" (https://data.cami-challenge.org/).

Quality Control.Real samples were processed with fastp (3). Human and phix-virus contamination removal was done with Bowtie2 (4) and samtools (5).

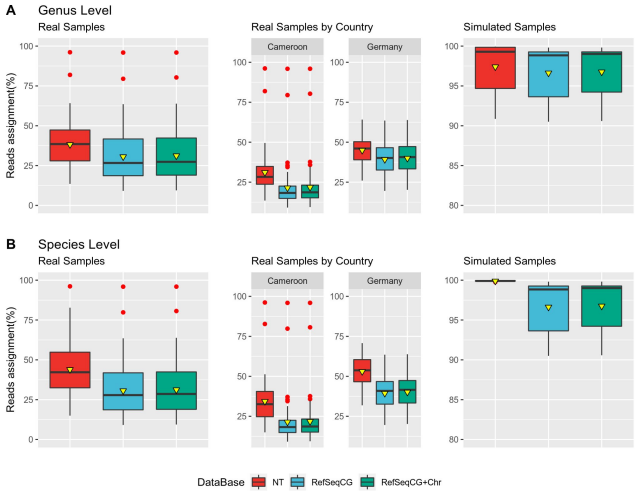
Metagenomic classification. Samples were classified using Kraken2 (6) and Bracken (7).

Performance Analysis. We analyzed classification capacity and various performance metrics including Precision (1), Recall (1), F1 Score, Bray-Curtis Dissimilarity and L2 Distance.

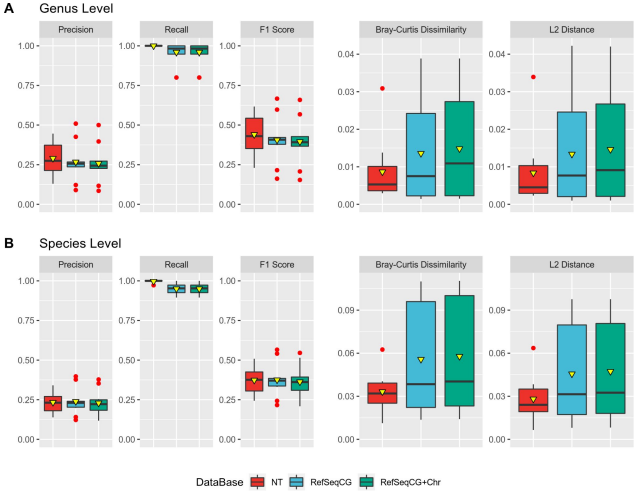
DATABASES SIMILARITIES WITH SIMULATED SAMPLES



CLASSIFICATION CAPACITY



PERFORMANCE METRICS



CONCLUSIONS

Among the conventional databases, the NT Database offers the best results in terms of read assignment and performance at the examined levels.

In the case of the RefSeq Databases, including the Chr assembly level does not substantially improve the results in terms of classification capacity. Furthermore, this leads to a decrease in performance as the size of the database increases, without adding new genera and species relevant to the samples.

REFERENCES

1. Ye et al. (2019). Cell 178(4), 779-794.
2. Breitwieser et al. (2019). Briefings in Bioinformatics 20(4), 1125-1139.
3. Chen et al. (2018). Bioinformatics 34(17), i884-i890.
4. Langmead & Salzberg (2012). Nature methods 9(4), 357-359.
5. Li et al. (2009). Bioinformatics 25(16), 2078-2079.
6. Wood et al. (2019). Genome biology 20(1), 257.
7. Lu et al. (2017). PeerJ Computer Science 3, e104.

ACKNOWLEDGEMENTS & FUNDING

This work was supported by MICINN-PID2019-105969GB-I00 and GV-Prometeo/2018/133.

CONTACT

sapies@alumni.uv.es