PICRUSt2-MPGA: an update to the reference database used for functional prediction within PICRUSt2

Robyn J. Wright [1], Morgan. G.I. Langille [1]

[1] Department of Pharmacology, Faculty of Medicine, Dalhousie University, Halifax, Canada

*Corresponding author: robyn.wright@dal.ca

Abstract

Summary: PICRUSt2 is a bioinformatic tool that predicts microbial functions in amplicon sequencing data using a database of annotated reference genomes. We have constructed an updated database for PICRUSt2 that has substantially increased the number of bacterial (19,493 to 26,868) and archaeal (406 to 1,002) genomes as well as the number of functional annotations present. The previous PICRUSt2 database relied on many timely and computationally intensive manual processes that made it difficult to update. To make PICRUSt2 great again (MPGA), we constructed a new streamlined process to allow regular upgrades to the PICRUSt2 database on an ongoing basis. Additionally, we have shown that this updated database contains genomes that more closely match study sequences from a range of different environments. The genomes contained in the database therefore better represent these environments and this leads to an improvement in the predicted functional annotations obtained from PICRUSt2.

Availability: PICRUSt2 source code is freely available at https://github.com/picrust/picrust2 and at https://anaconda.org/bioconda/picrust2. The PICRUSt2-MPGA database comes pre-installed with PICRUSt2 from version 2.6.0 onwards. Step-by-step instructions for making the updated database are at https://github.com/picrust/picrust2/wiki/Updating-the-PICRUSt2-database. All code used for the analyses and figures in this manuscript is at https://github.com/R-Wright-1/picrust2_mpga_application_note.

Funding and acknowledgements

This work was supported by an NSERC Discovery Grant (2016-05039) and was enabled in part by the Digital Research Alliance of Canada (alliancecan.ca).

We thank Dr. Gavin Douglas for invaluable insight and assistance with troubleshooting and several PICRUSt2 users for their time testing and giving feedback on an early version of the PICRUSt2 update.

1. Introduction

PICRUSt2 (**P**hylogenetic **I**nvestigation of **C**ommunities by **R**econstruction of **U**nobserved **St**ates) is a widely used bioinformatic tool that allows the prediction of microbial community functions based on amplicon sequencing data. The first version of PICRUSt (PICRUSt1) was published in 2013 [1] and the second version (PICRUSt2) was published in 2020 [2]. In PICRUSt2, 16S rRNA gene sequences collapsed to amplicon sequence variants (ASVs) are placed into a reference tree and multiple-sequence alignment and then hidden-state prediction is used to infer the gene content of study sequences based on the gene content of the taxa used to construct the reference tree. Information on their predicted 16S rRNA gene copy number and their abundance in the original samples is used to weight the predicted genome annotations to generate a predicted metagenome.

One of the largest limitations to PICRUSt2 currently is that the default database uses functional annotations acquired from the Integrated Microbial Genomes database [3] in 2017. The number of genomes and functions have continued to expand during this time and timely updates to PICRUSt2 are essential for its accuracy and functional comprehensiveness. However, the previous approach for updating the default database and functions was difficult due to a reference tree needing to be re-built with newly added genomes. To **m**ake **P**ICRUSt2 **g**reat **a**gain (MPGA), we have constructed a new database for PICRUSt2, PICRUSt2-MPGA, that uses Genome Taxonomy Database (GTDB) r214 genomes [4–7] and the bacterial and archaeal phylogenetic trees released with GTDB. We have annotated these genomes using Eggnog [8, 9], and as the GTDB genomes are readily available for download, PICRUSt2 users can easily add annotations to this updated database.

2. PICRUSt2-MPGA Database creation

All genomes (*n*=402,709) from the Genome Taxonomy Database (GTDB) r214.1 were downloaded from the GTDB repository. Only representative genomes that were present in the GTDB bacterial or archaeal phylogenetic trees, had a high quality 16S rRNA gene and that were ≤ 10% contamination and ≥ 90% completion were used, giving *n*=27,870 total genomes (*n*=26,868 and *n*=1,002 genomes for bacteria and archaea, respectively). Full details of the methods used for 16S rRNA gene identification and processing are in Supplementary Information section 2.1. This has increased the number of taxa present at every phylogenetic rank from phylum to species (Fig. 1), with approximately 3-fold or 4-fold more species present in the PICRUSt2-MPGA database for bacteria or archaea, respectively, compared with the previous PICRUSt2 database (hereafter referred to as PICRUSt2-oldIMG). Included bacterial genomes have a median completeness of 99.35% and contamination of 0.66% while archaeal genomes have a median completeness of 99.03% and contamination of 0.67%.

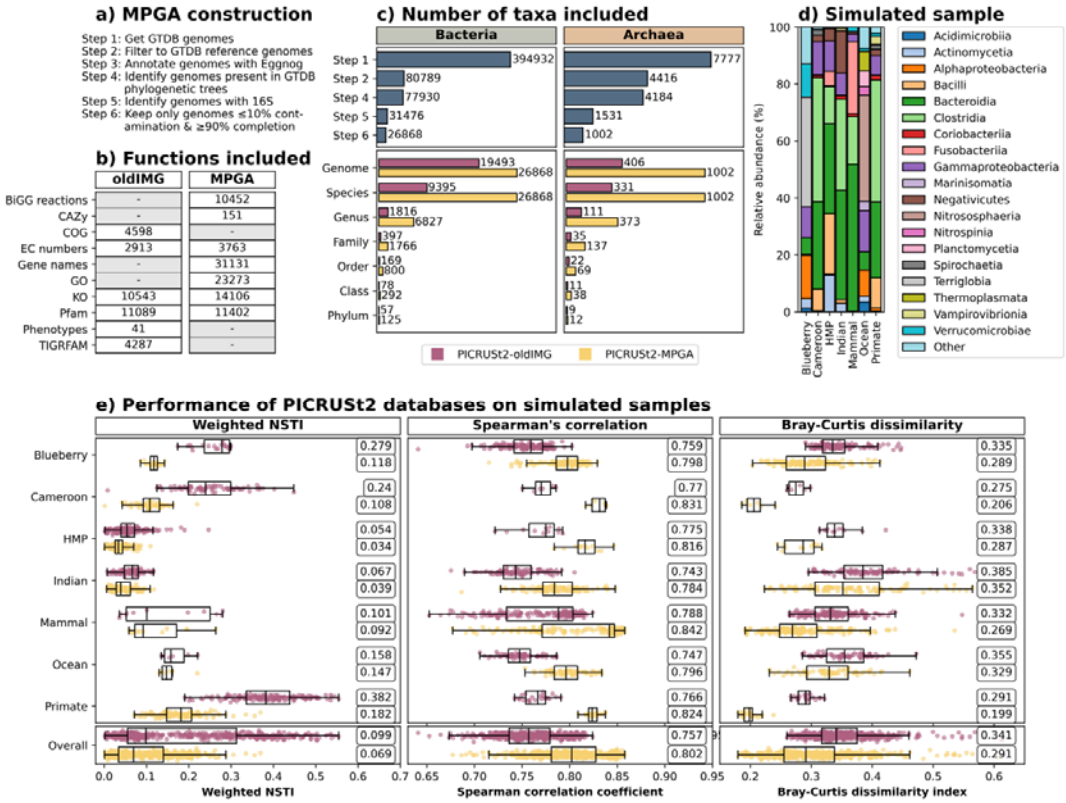


Figure 1. Comparison of the PICRUSt2-oldIMG and PICRUSt2-MPGA databases showing: (**a**) the steps in the construction of the PICRUSt2-MPGA database; (**b**) the number of functions annotated within different frameworks for the default and updated databases (note that not all frameworks were included in both databases); (**c**) the number of taxa included for each step of the PICRUSt2-MPGA construction (top) and for each phylogenetic rank (bottom) for bacteria and archaea; (**d**) composition at the class level for the simulated samples (the mean relative abundance is shown for each dataset); and (**e**) the performance of the PICRUSt2-oldIMG and PICRUSt2-MPGA databases on the simulated samples from each dataset and overall (bottom). Individual points are shown for each sample with points being coloured pink for the PICRUSt2-oldIMG database and yellow for the PICRUSt2-MPGA database. Boxplots represent the median, upper and lower quartiles and whiskers show the range of the data (1.5 times the Interquartile Range) and values in boxes are medians.

Annotation of all representative genomes was carried out using Eggnog v 2.1.12 with Prodigal v2.6.3 [10] for gene prediction and the Eggnog diamond database v 5.0.2. This gave Biochemical, Genetic and Genomic (BiGG) model reactions [11], Carbohydrate Active Enzymes (CAZy) [12], Enzyme Commission (EC) number, Gene Ontology (GO) [13, 14], Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs (KO) [15], Protein family (Pfam) [16], and gene name annotations for all genomes. The default PICRUSt2 database contained annotations for KO, EC, Pfam, Clusters of Orthologous Genes (COG) [17, 18], Phenotype [3] and TIGRFAM [19] annotations. By default, only the KO and EC annotations are used. In the PICRUSt2-MPGA database, we have 1.3-fold more annotations than the PICRUSt2-oldIMG

database for both KOs (14,106 *versus* 10,543) and EC numbers (3,763 *versus* 2,913; Fig. 1). See supplementary information section 2.2 for further details.

3. Placement of sequences into phylogenetic trees

With the PICRUSt2-oldIMG database, study sequences are placed in a reference phylogenetic tree constructed using 16S rRNA gene sequences that contains both bacterial and archaeal sequences. In our PICRUSt2-MPGA database, we now use separate (unrooted) trees for bacteria and archaea constructed using multiple marker genes (from GTDB), and we therefore initially place sequences in both trees and perform 16S copy number prediction and Nearest Sequenced Taxon Index (NSTI) calculations for both trees. We then compare the NSTI obtained for each and choose the most appropriate (lowest NSTI) for functional predictions. These updates to the code used for running PICRUSt2 are described in Supplementary Information section 3.1. To verify that this tree placement is appropriate, we used the same 16S datasets used in Douglas *et al.* [2]: (i) Blueberry *n*=22, bulk soil and blueberry rhizosphere samples[20, 21]; (ii) Cameroon, *n*=57, stool samples from Cameroonian individuals [22, 23]; (iii) HMP, *n*=137, samples spanning the human body from the Human Microbiome Project [24]; (iv) Indian, *n*=91, stool samples from Indian individuals [25]; (v) Mammal, *n*=8, mammalian stool samples [26]; (vi) Primate, *n*=77, non-human primate stool samples [27]; and (vii) Ocean, *n*=6, ocean samples [28]. We took the processed feature tables and representative sequences used by Douglas *et al.* and classified these taxonomically using the naïve bayes Scikit-Learn classifier trained on full-length 16S rRNA gene sequences from the Greengenes2 database [29] within QIIME2 (v2024.5) [30]. We ran PICRUSt2 using both the PICRUSt2-oldIMG and the PICRUSt2-MPGA databases for each of the datasets.

The domain obtained from the taxonomic classifications and the domain of the tree giving the lowest NSTI value for each sequence agreed for 21,543 of 21,552 (99.9%) total sequences within these datasets. See Supplementary Information section 3.2 for details on the nine ASVs that were potentially placed into the wrong tree. Comparing the NSTIs obtained with the PICRUSt2-MPGA *versus* the PICRUSt2-oldIMG database on a per sequence basis and on a per sample basis (*i.e.*, the weighted NSTIs that account for sequence abundance within samples), we find that the NSTIs have decreased in almost all cases (Supplementary Information section 3.3).

4. Verification of the functional predictions obtained with the updated database

The functional predictions obtained with the PICRUSt2-MPGA database were correlated with those obtained with the PICRUSt2-oldIMG database (Supplementary

Information section 4.1). In Douglas *et al.* [REF], Spearman correlation coefficients were calculated between PICRUSt2-obtained KO predictions and HUMAnN2-obtained KO annotations of the paired MGS samples. Although we did not expect this to be a reliable standard for a database constructed at a different time using different databases for annotation, we calculated Spearman correlation coefficients and Bray-Curtis dissimilarity indices between the PICRUSt2-predictions for both databases and the MGS HUMAnN2 annotations for comparability with the previous study (Supplementary Information section 4.2).

We instead constructed mock samples using Eggnog-annotated genomes with simulated abundances. Briefly, we matched the taxa in the seven datasets described above to GTDB genomes that were not present in the PICRUSt2-MPGA database and used these to construct mock communities that matched the abundances within the seven real datasets. Full details of this mock sample construction can be found in Supplementary Information section 4.3. The mock samples that we constructed comprised a range of taxa typical of the environment that they were designed to simulate (Fig. 1), for example, the Blueberry soil samples contained genomes from the Terriglobia class, while the Ocean water samples contained genomes from the Nitrososphaeria class. We ran PICRUSt2 with both the PICRUSt2-oldIMG and the PICRUSt2-MPGA databases using the mock representative sequences and feature tables for each dataset. Comparing the NSTIs obtained with the PICRUSt2-MPGA database *versus* the PICRUSt2-oldIMG database on a per sample basis (*i.e.*, the weighted NSTIs that account for sequence abundance within samples), we find that the NSTIs have decreased in all cases (Fig. 1). The NSTIs have decreased from a median of 0.099 (range 0.054 [HMP] – 0.382 [Primate]) with PICRUSt2-oldIMG to 0.069 (range 0.034 [HMP] – 0.182 [Primate]) with PICRUSt2-MPGA. While the median NSTI per sequence (Supplementary Information section 4.4) was slightly lower with the PICRUSt2-oldIMG database for the Cameroon and Indian datasets, the median weighted NSTIs were always lower with the PICRUSt2-MPGA database, suggesting that it is rarer sequences that have higher NSTIs.

We compared the functional predictions obtained using PICRUSt2 with either the PICRUSt2-oldIMG or the PICRUSt2-MPGA database with the gold-standard and calculated both Spearman correlation coefficients and Bray-Curtis dissimilarity indices. Median Spearman correlation coefficients are higher for the PICRUSt2-MPGA (median 0.802; range 0.784 [Indian] – 0.842 [Mammal]) *versus* the PICRUSt2-oldIMG (median 0.757; range 0.743 [Indian] – 0.788 [Mammal]) database for both KOs (Fig. 1) and EC numbers (Supplementary Information section 4.4). Bray-Curtis dissimilarity indices were also lower for the PICRUSt2-MPGA (median 0.291; 0.352 [Indian] – range 0.199 [Primate]) *versus* the PICRUSt2-oldIMG (median 0.341; 0.385 [Indian] – range 0.275 [Cameroon]) database for both KOs (Fig. 1) and EC numbers (Supplementary Information section 4.4). KOs had both higher Spearman correlation coefficients and lower Bray-Curtis dissimilarity indices between the mock and PICRUSt2-predictions than EC numbers, and the lowest

correlations and highest Bray-Curtis dissimilarity indices were found consistently with the Indian dataset.

6. Conclusions

We have shown that the updated PICRUSt2-MPGA database using GTDB genomes performs as expected and outperforms the previous PICRUSt2-oldIMG database in almost all cases. This provides PICRUSt2 users with a database that contains more genomes and functions and provides a framework for the PICRUSt2 database to be updated more frequently in the future.

References

1. **Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, *et al.*** Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;31:1814–821.

2. **Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, *et al.*** PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* 2020;1–5.

3. **Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, *et al.*** IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res*;40. Epub ahead of print January 2012. DOI: 10.1093/nar/gkr1044.

4. **Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, *et al.*** GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785–D794.

5. **Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, *et al.*** A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 2020;1–8.

6. **Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, *et al.*** A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;36:996.

7. **Rinke C, Chuvochina M, Mussig AJ, Chaumeil PA, Davín AA, *et al.*** A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology 2021 6:7* 2021;6:946–959.

8. **Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, *et al.*** EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.

9. **Cantalapiedra CP, Hern□andez-Plaza A, Letunic I, Bork P, Huerta-Cepas J**. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 2021;38:5825–5829.

10. **Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, *et al.*** *Prodigal: prokaryotic gene recognition and translation initiation site identification*. http://www.biomedcentral.com/1471-2105/11/119 (2010).

11. **Schellenberger J, Park JO, Conrad TM, Palsson BØ**. *BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions*. http://www.biomedcentral.com/1471-2105/11/213 (2010).

12. **Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, *et al.*** The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res*;37. Epub ahead of print 2009. DOI: 10.1093/nar/gkn663.

13. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, *et al.*** *Gene Ontology: tool for the unification of biology The Gene Ontology Consortium**. http://www.flybase.bio.indiana.edu (2000).

14. **Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, *et al.*** The Gene Ontology knowledgebase in 2023. *Genetics*;224. Epub ahead of print 1 May 2023. DOI: 10.1093/genetics/iyad031.

15. **Kanehisa M, Goto S**. KEGG; Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;28:27–30.

16. **Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, *et al.*** Pfam: The protein families database. *Nucleic Acids Research*;42. Epub ahead of print 1 January 2014. DOI: 10.1093/nar/gkt1223.

17. **Tatusov RL, Koonin E V, Lipman DJ**. *A Genomic Perspective on Protein Families*. https://www.science.org.

18. **Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D, *et al.*** COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021;49:D274–D281.

19. **Li W, O'Neill KR, Haft DH, Dicuccio M, Chetvernin V, *et al.*** RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;49:D1020–D1028.

20. **Yurgel SN, Douglas GM, Dusault A, Percival D, Langille MGI**. Dissecting community structure in wild blueberry root and soil microbiome. *Front Microbiol*;9. Epub ahead of print 6 June 2018. DOI: 10.3389/fmicb.2018.01187.

21. **Yurgel SN, Douglas GM, Comeau AM, Mammoliti M, Dusault A, *et al.*** Variation in bacterial and eukaryotic communities associated with natural and managed wild blueberry habitats. *Phytobiomes J* 2017;1:102–113.

22. **Morton ER, Lynch J, Froment A, Lafosse S, Heyer E, *et al.*** Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by Entamoeba and Subsistence. *PLoS Genet*;11. Epub ahead of print 1 November 2015. DOI: 10.1371/journal.pgen.1005658.

23. **Lokmer A, Cian A, Froment A, Gantois N, Viscogliosi E, *et al.*** Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. *PLoS One*;14. Epub ahead of print 1 February 2019. DOI: 10.1371/journal.pone.0211139.

24. **Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, *et al.*** Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–214.

25. **Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, *et al.*** The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience*;8. Epub ahead of print 30 January 2019. DOI: 10.1093/gigascience/giz004.

26. **Finlayson-Trick ECL, Getz LJ, Slaine PD, Thornbury M, Lamoureux E, *et al.*** Taxonomic differences of gut microbiomes drive cellulolytic enzymatic potential within hind-gut fermenting mammals. *PLoS One*;12. Epub ahead of print 1 December 2017. DOI: 10.1371/journal.pone.0189404.

27. **Amato KR, G. Sanders J, Song SJ, Nute M, Metcalf JL, *et al.*** Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *ISME Journal* 2019;13:576–587.

28. **Gillies LE, Thrash JC, deRada S, Rabalais NN, Mason OU**. Archaeal enrichment in the hypoxic zone in the northern Gulf of Mexico. *Environ Microbiol* 2015;17:3847–3856.

29. **McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, *et al.*** Greengenes2 unifies microbial data in a single reference tree. *Nat Biotechnol* 2024;42:715–718.

30. **Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, *et al.*** Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;37:852–857.

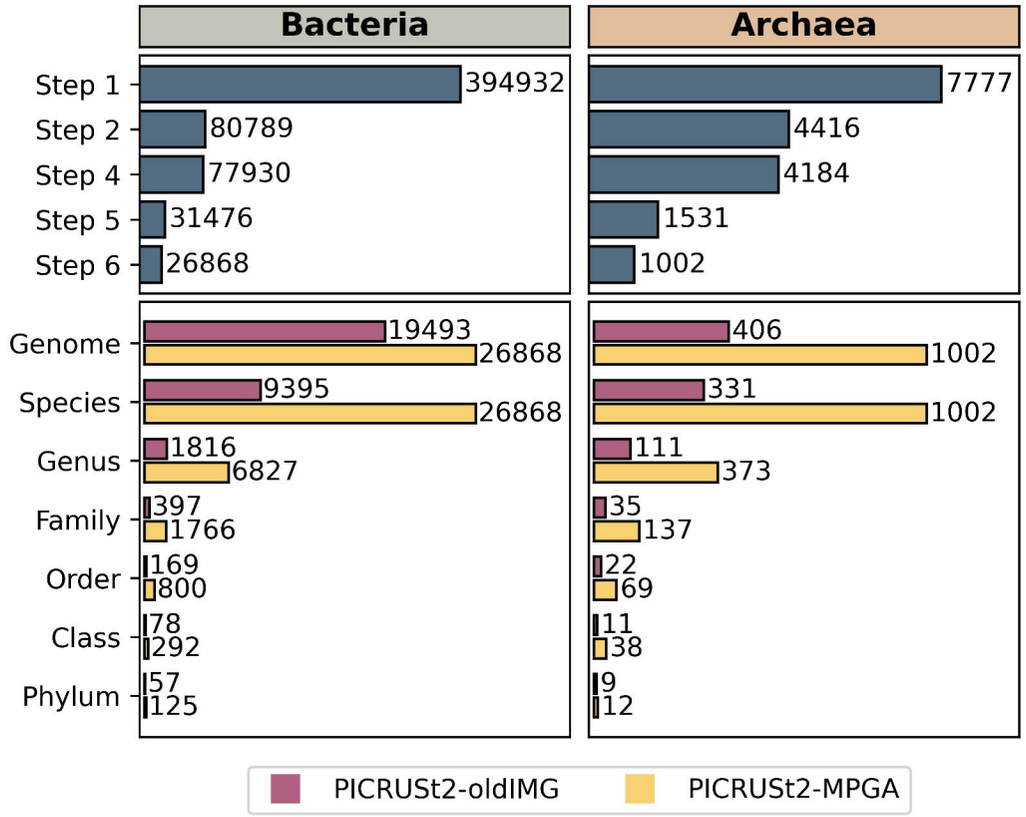
a) MPGA construction

Step 1: Get GTDB genomes
Step 2: Filter to GTDB reference genomes
Step 3: Annotate genomes with Eggnog
Step 4: Identify genomes present in GTDB phylogenetic trees
Step 5: Identify genomes with 16S
Step 6: Keep only genomes ≤10% contamination & ≥90% completion

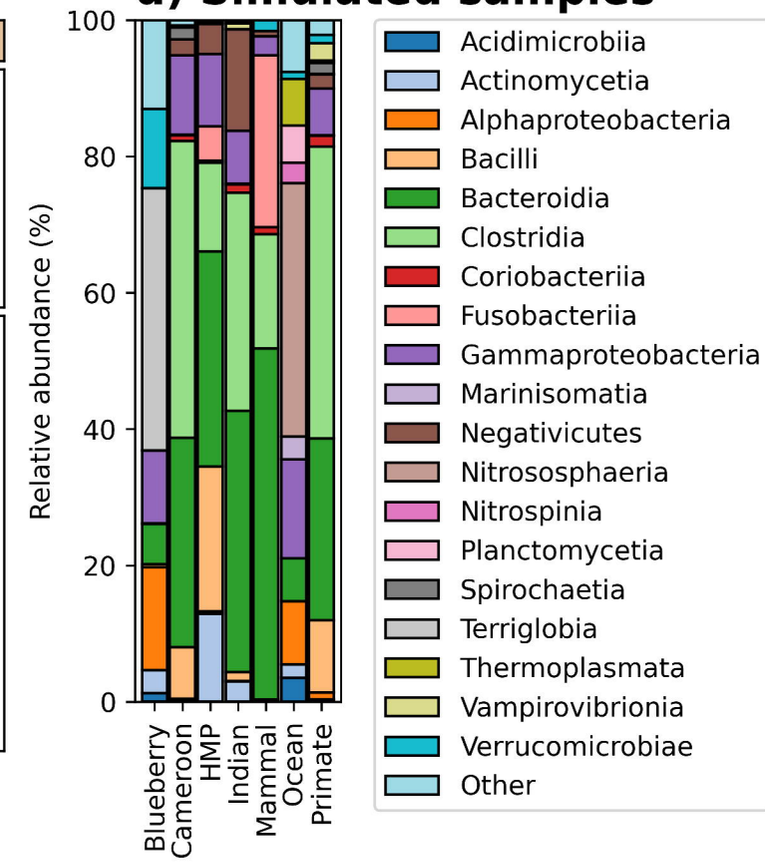
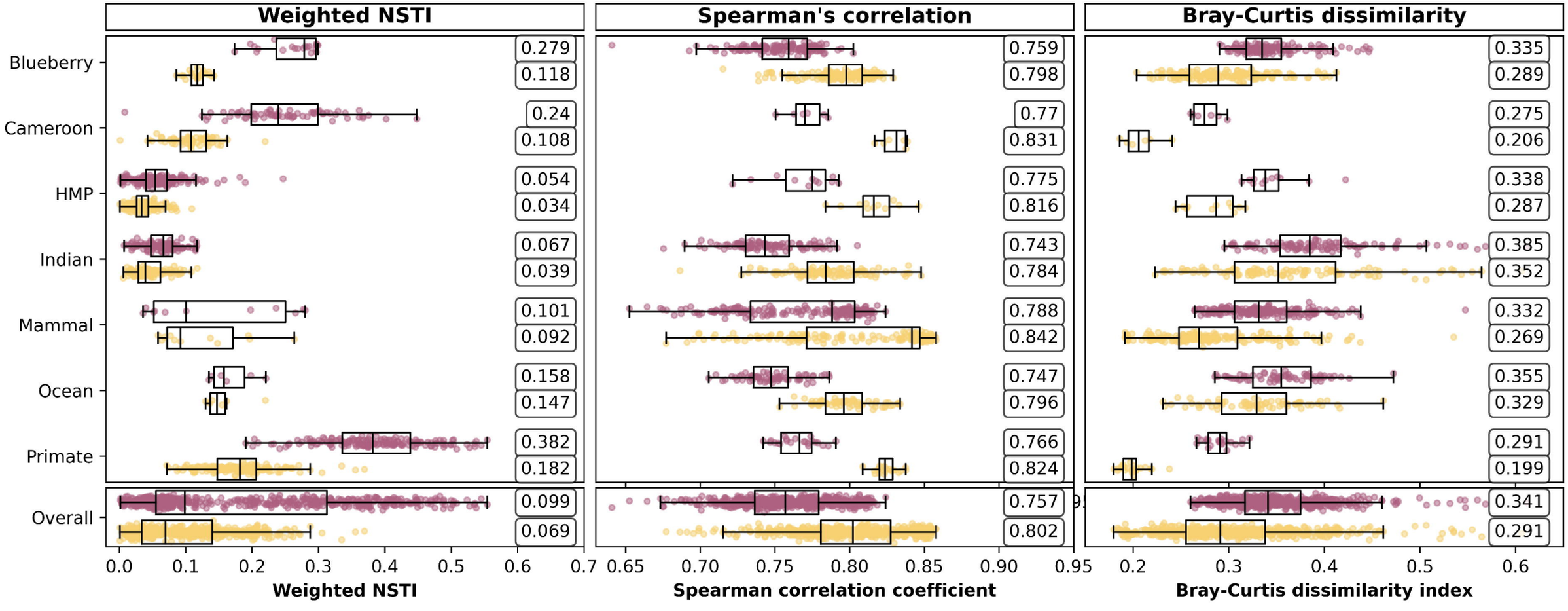
b) Functions included

	oldIMG	MPGA
BiGG reactions	-	10452
CAZy	-	151
COG	4598	-
EC numbers	2913	3763
Gene names	-	31131
GO	-	23273
KO	10543	14106
Pfam	11089	11402
Phenotypes	41	-
TIGRFAM	4287	-

c) Number of taxa included



d) Simulated samples



e) Performance of PICRUSt2 databases on simulated samples