I2SysBio INSTITUTE FOR INTEGRATIVE SYSTEMS BIOLOGY

VNIVERSITAT DE VALÈNCIA

16S rRNA GENE DATABASE ENRICHMENT STRATEGY TO IMPROVE CLASSIFICATION OF NOVEL BACTERIAL TAXA IN NUDIBRANCH MICROBIOME

Dafne Porcel Sanchis[1], Samuel Piquer-Esteban[1], Vicente Arnau[1], Wladimiro Diaz-Villanueva[1], Maria Dzunkova[1,2,3]

1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2. Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
3. Department of Energy Joint Genome Institute, Berkeley, CA, USA

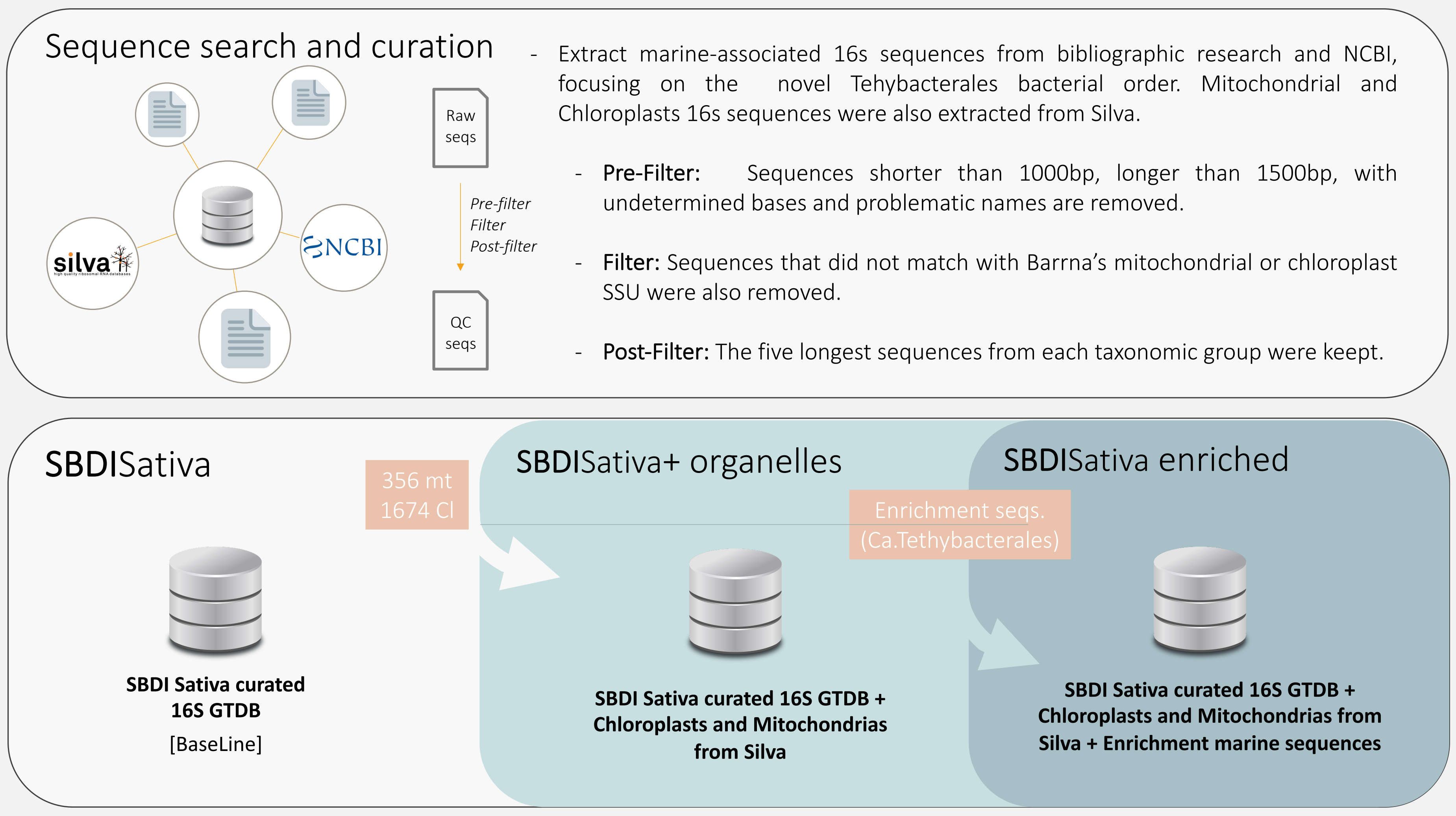
INTRODUCTION

Soft-bodied marine animals, such as marine sponges and nudibranchs, use bioactive molecules to protect themselves from their predators. Their microbiomes are seen as a possible source of new bioactive compounds. Sequencing of the 16S rRNA gene amplicons represents the first insights into the presence of novel bacterial groups that might be further investigated for their biosynthetic potential. Nudibranch microbiome still contains a large portion of unknown bacteria. Characterization of the nudibranch microbiome is challenging due to under representation of its symbiotic bacteria in conventional databases. In our previous single-cell genomics study of nudibranch microbiome, we detected a new member of Candidatus Tethybacteriales, an uncultured order of endosymbiotic microbes recently discovered in marine sponges which is not yet included in conventional databases. For that reason, we focuse on enriching a reference database of 16S rRNA gene sequences with previously known sequences of the order Ca. Tethybacterales.

METHOLOGY

To construct our marine enrichment 16S rRNA database gene sequences we take as a reference the baseline SBDI Sativa curated 16S GTDB database (SBDI; 2021)[1]. This database is based on the GTDB[2], which is currently considered taxonomically highly rigorous and is created by genome clustering at the species level. Furthermore, SBDI Sativa has already passed an extensive quality control process. However, this database lacked sequences from Mitochondia and Chloroplast that must be added for the taxonomic classification in posterior filtering steps. Those sequences are obtained from the SILVA SSU 138.1 database. The next step in this database construction was to add marine sequences of interest, focusing on Ca. Tethybacterales order. In our previous study we discovered Ca. Doriopsillibacter californiensis, a new member of the Ca. Tethybacterales order, which is not yet respresented in the databases. We used the two main recent publications related to this taxon[3,4] for enriching the database. Also, other novel marine related bacteria were added to the enrichment dataset. All those sequences (organelles, Tehybacterales and other marine related bacteria) were subjected to quality control similarly to the SBDI Sativa curation process (Figure 1). After this quality control, only Ca. Thethybacterales were preserved to enrich the database.

Those three different versions obtained in the enrichment process (SBDI Sativa, SBDI Sativa with organelles and SBDI Sativa with organelles and enriched with 16S gene sequences from Tehybacterales order), as well as other two conventional databases (GTDB r202 –modified– and SILVA SSU r138 from DESCIPHER package[5]) were used to classify a dataset of 83 nudibranch microbiome samples using the IDTAXA[6] classifier of the DECIPHER package. IDTAXA is a taxonomic classifier that employs artificial intelligence. The different database versions obtained were used to create their respective training set models used afterwards by the IDTAXA classifier.



Sequence search and curation

- Extract marine-associated 16s sequences from bibliographic research and NCBI, focusing on the novel Tehybacterales bacterial order. Mitochondrial and Chloroplasts 16s sequences were also extracted from Silva.
  - **Pre-Filter:** Sequences shorter than 1000bp, longer than 1500bp, with undetermined bases and problematic names are removed.
  - **Filter:** Sequences that did not match with Barrna's mitochondrial or chloroplast SSU were also removed.
  - **Post-Filter:** The five longest sequences from each taxonomic group were keept.

SBDISativa

356 mt
1674 Cl

SBDI Sativa curated 16S GTDB
[BaseLine]

SBDISativa+ organelles

Enrichment seqs.
(Ca.Tethybacterales)

SBDI Sativa curated 16S GTDB + Chloroplasts and Mitochondrias from Silva

SBDISativa enriched

SBDI Sativa curated 16S GTDB + Chloroplasts and Mitochondrias from Silva + Enrichment marine sequences

Figure1. Representation of 16S gene sequence databse curation and construction process. SBDI Sativa curated 16S GTDB database (SBDI; 2021) was used as benchmark. Chloroplasts and Mitochondia sequences extracted from SILVA SSU 138.1 were added for posterior filterings as well as marine related 16S sequences.

RESULTS

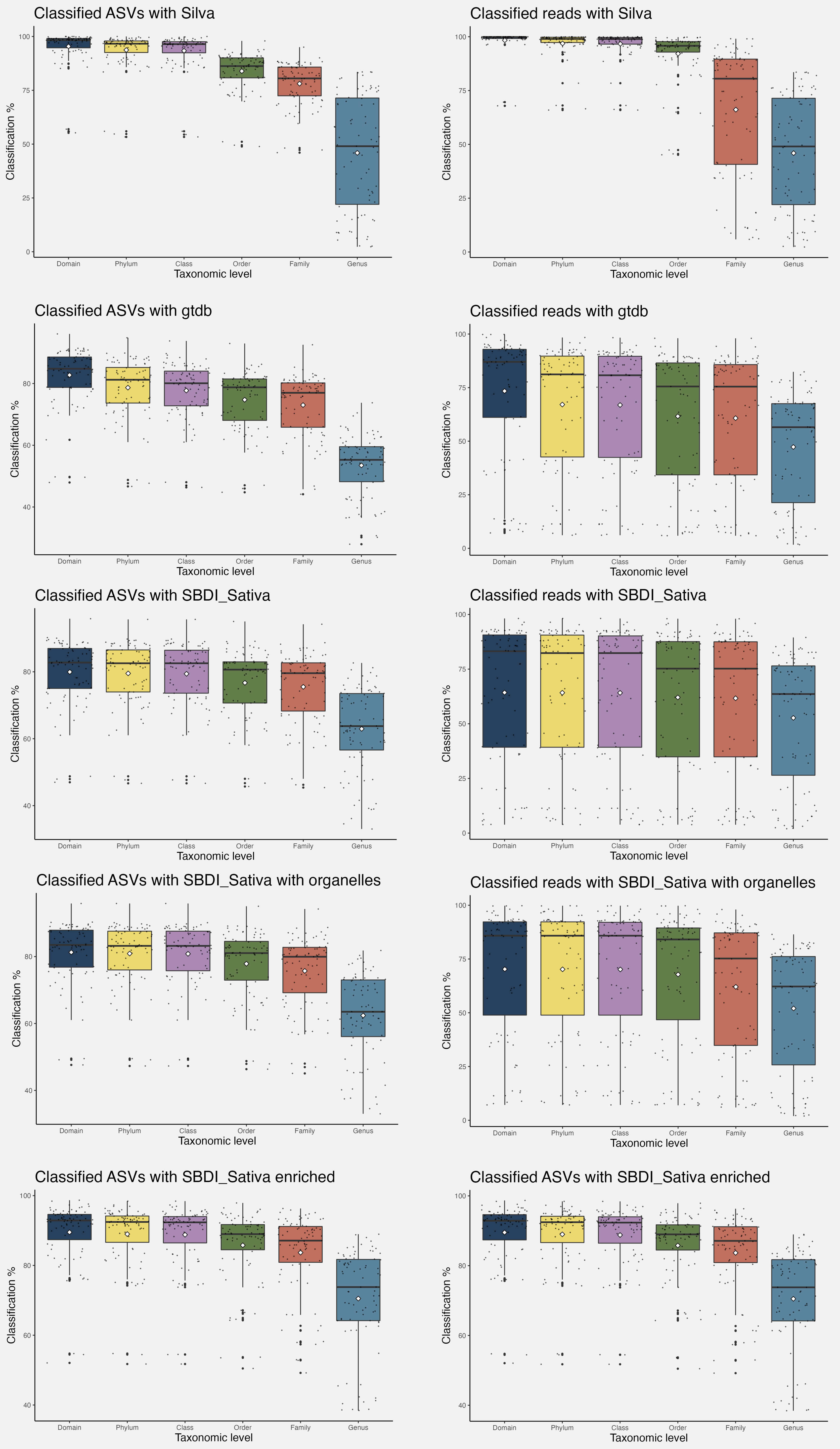


Figure2. Percentage of classification capacity of each database per sample for ASVs and reads and for each taxonomic level.

Results shown in **Fig.2** corroborate the use of GTBD as benchmark for the enrichment approach. Comparing SBDI Sativa enriched with organelles sequences and SILVA SSU r138, the first one reveals higher classification power at lower levels such as genus and family.

The enriched database manages to improve the reads classification capacity at the genus and family taxonomic levels with respect to the conventional databases. However, the SILVA database manages to classify slightly more sequences at higher taxonomic levels (domain-order) and also shows higher percentages of classified ASVs. This was expected since the SILVA database has the most breadth of sequences.

	Silva				GTDB				SBDI Sativa				SBDI Sativa with organelles				SBDI Sativa enriched			
	Mean ASV	Sd ASV	Mean read	Sd. Read	Mean ASV	Sd ASV	Mean read	Sd. Read	Mean ASV	Sd ASV	Mean read	Sd. Read	Mean ASV	Sd ASV	Mean read	Sd. Read	Mean ASV	Sd ASV	Mean read	Sd. Read
Domain	95.4	8.4	98.4	5.9	82.7	9.2	73.4	28.2	80.0	9.6	64.3	31.2	81.3	9.4	70.3	29.4	89.5	9.2	92.0	13.6
Phylum	93.7	8.7	96.9	6.6	78.7	9.4	67.1	29.0	79.5	9.6	64.2	31.2	80.9	9.4	70.2	29.4	89.0	9.2	91.9	13.6
Class	93.4	8.7	96.7	6.7	77.7	9.3	66.9	28.8	79.4	9.7	64.1	31.1	80.8	9.4	70.1	29.3	88.8	9.2	91.8	13.6
Order	83.9	9.2	92.1	11.4	74.7	9.6	61.6	30.4	76.7	10.2	62.1	31.8	77.9	10.1	67.8	30.6	85.8	10.2	89.6	16.3
Family	78.1	10.0	66.1	29.6	73.0	9.9	60.7	30.0	75.5	10.5	61.7	31.5	75.7	10.5	62.0	30.6	83.7	11.4	83.7	22.3
Genus	45.9	26.3	45.9	26.3	53.5	9.9	47.3	25.6	62.9	12.6	52.7	29.2	62.4	12.5	52.0	28.8	70.5	14.4	73.4	23.9

Table 1.. Mean value of the classification percentage of each database for ASVs and reads for each taxonomic level .

CONCLUSIONS

Despite Silva SSUr138 manages to improve classification at higher levels, our enrichment database approach has shown higher classification power than other conventional databases at genus level, even whithout the Ca. Thethybacterales order 16s sequences. It proves the efficacy of database enrichment strategies for characterization of unknown micribiomes.

REFERENCES

1. Lundin, Daniel; Andersson, Anders (2021): SBDI Sativa curated 16S GTDB database. SciLifeLab. Dataset.
2. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019 Nov 15;36(6):1925–7.
3. Taylor, J. A., Palladino, G., Wemheuer, B., Steinert, G., Sipkema, D., Williams, T. J., & Thomas, T. (2021). Phylogeny resolved, metabolism revealed: functional radiation within a widespread and divergent clade of sponge symbionts. The ISME Journal, 15(2), 503-519.
4. Waterworth, S. C., Parker-Nance, S., Kwan, J. C., & Dorrington, R. A. (2021). Comparative genomics provides insight into the function of broad-host range sponge symbionts. Mbio, 12(5), e01577-21.
5. Wright E. S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. R Journal, 8(1): 352-359.
6. A Murali et al. (2018) "IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences." Microbiome,