

# GSR-DB: a manually curated and optimized taxonomical database for 16S rRNA amplicon analysis

Leidy-Alejandra G. Molano,<sup>1</sup> Sara Vega-Abellaneda,<sup>1</sup> Chaysavanh Manichanh<sup>1,2</sup>

**AUTHOR AFFILIATIONS** See affiliation list on p. 14.

**ABSTRACT** Amplicon-based 16S ribosomal RNA sequencing remains a widely used method to profile microbial communities, especially in low biomass samples, due to its cost-effectiveness and low-complexity approach. Reference databases are a mainstay for taxonomic assignments, which typically rely on popular databases such as SILVA, Greengenes, Genome Taxonomy Database (GTDB), or Ribosomal Database Project (RDP). However, the inconsistency of the nomenclature across databases and the presence of shortcomings in the annotation of these databases are limiting the resolution of the analysis. To overcome these limitations, we created the GSR database (Greengenes, SILVA, and RDP database), an integrated and manually curated database for bacterial and archaeal 16S amplicon taxonomy analysis. Unlike previous integration approaches, this database creation pipeline includes a taxonomy unification step to ensure consistency in taxonomical annotations. The database was validated with three mock communities, two real data sets, and a 10-fold cross-validation method and compared with existing 16S databases such as Greengenes, Greengenes 2, GTDB, ITGDB, SILVA, RDP, and MetaSquare. Results showed that the GSR database enhances taxonomical annotations of 16S sequences, outperforming current 16S databases at the species level, based on the evaluation of the mock communities. This was confirmed by the 10-fold cross-validation, except for Greengenes 2. The GSR database is available for full-length 16S sequences and the most commonly used hypervariable regions: V4, V1–V3, V3–V4, and V3–V5.

**IMPORTANCE** Taxonomic assignments of microorganisms have long been hindered by inconsistent nomenclature and annotation issues in existing databases like SILVA, Greengenes, Greengenes2, Genome Taxonomy Database, or Ribosomal Database Project. To overcome these issues, we created Greengenes-SILVA-RDP database (GSR-DB), accurate and comprehensive taxonomic annotations of 16S amplicon data. Unlike previous approaches, our innovative pipeline includes a unique taxonomy unification step, ensuring consistent and reliable annotations. Our evaluation analyses showed that GSR-DB outperforms existing databases in providing species-level resolution, especially based on mock-community analysis, making it a game-changer for microbiome studies. Moreover, GSR-DB is designed to be accessible to researchers with limited computational resources, making it a powerful tool for scientists across the board. Available for full-length 16S sequences and commonly used hypervariable regions, including V4, V1–V3, V3–V4, and V3–V5, GSR-DB is a go-to database for robust and accurate microbial taxonomy analysis.

**KEYWORDS** database, 16S rRNA, microbiome, taxonomy

1 6S rRNA gene amplicon-based (16S) sequencing is a widespread method that has profoundly impacted the human microbiome characterization, revealing important insights into the complex interactions between microorganisms and their hosts. This

**Editor** Nicola Segata, Universita degli Studi di Trento, Trento, Italy

Address correspondence to Chaysavanh Manichanh, cmanicha@gmail.com.

Leidy-Alejandra G. Molano and Sara Vega-Abellaneda contributed equally to this article. Author order was determined by tossing a coin.

The authors declare no conflict of interest.

See the funding table on p. 15.

**Received** 8 September 2023

**Accepted** 27 November 2023

**Published** 8 January 2024

Copyright © 2024 Molano et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

approach has allowed us to associate altered microbial profiles with diseases, including gut-associated conditions, inflammatory bowel disease, metabolic diseases, and colorectal cancer (1, 2).

16S analysis involves various upstream steps, including quality control, read trimming, and taxonomic classification. Previous studies have reported the impact of bioinformatic pipelines in the microbial profiling of biological samples, highlighting the importance of reference databases for taxonomic prediction (3). Currently, the most widely used databases are Greengenes (4), Genome Taxonomy Database (GTDB) (5), SILVA (6), and Ribosomal Database Project (RDP) (7). However, discrepancies between these databases have been acknowledged. Robeson et al. (8) found that Greengenes, SILVA, and GTDB presented sequence similarities but were taxonomically different, leading to a low proportion of taxonomic labeling shared among databases at all ranks below the domain level. Moreover, outlier sequences were found in the length distribution across databases, probably corresponding to partial or untrimmed 16S sequences, which are recommended to be discarded to avoid biases in the analysis. Additionally, SILVA and Greengenes exhibited an immense amount of unannotated or unknown labeled sequences at genus and species level (~80%), which might introduce taxonomic noise during assignment (8).

To overcome these limitations and enhance classification performance, we created the GSR database for bacterial and archaeal 16S-based taxonomic profiling by integrating and manually curating the Greengenes, SILVA, and RDP databases. The taxonomic nomenclature of the GSR database has been unified to guarantee the coherence of annotations. Its performance has been compared with Greengenes, Greengenes2 (9), GTDB, SILVA, and RDP databases and other existing integrated databases, including ITGDB (10) and MetaSquare (11). The GSR database is available for full-length 16S sequences and the most commonly used hypervariable regions: V4, V1–V3, V3–V4, and V3–V5. It can be downloaded from the link <https://manichanh.vhir.org/gsrdb/>.

## MATERIALS AND METHODS

### Creation of GSR database

#### *Creation of the GSR full-16S database*

A full-length 16S database, the GSR-DB (Greengenes-SILVA-RDP database), was created by merging three already existing databases: Greengenes (version 13\_8, 99%) (4), SILVA (version 138, 99%) (6), and RDP (train set no. 18) (7). A data set with vaginal-related species was also included to ensure species detection for vaginal samples. The number of original entries of the Greengenes, SILVA, and RDP was 203,452, 436,681, and 21,194, respectively (more information regarding the construction of these databases can be found in the supplemental material). Before the integration, taxonomy filtering and formatting were performed on each original database. Only Bacteria and Archaea kingdoms were retained from the databases, excluding Eukaryota and Virus kingdoms in the SILVA database. Additionally, a manual curation process was applied to ensure the removal of potential redundancies for the subsequent merging of the original databases. After this process, the percentage of retained entries was 10.05%, 17.08%, and 95.08% for Greengenes, SILVA, and RDP, respectively. The vaginal data set was created with the 16S NCBI sequences proposed by Fettweis et al. (12) to create a vaginal reference database. Sequences and nomenclature corresponding to GenBank IDs provided in the study were retrieved from the NCBI. Sequences without exact species names or those corresponding to non-16S sequences were excluded. Once all original databases were preprocessed correctly, they were merged using the forthcoming algorithm.

#### *Manual curation of the GSR-DB*

The curation process for the GSR-DB included several steps to ensure the quality and accuracy of the data. These steps involved manual identification and removal of

patterns associated with unknown species (entries unannotated or with unknown labels, such as “uncultured,” “unidentified,” and “candidate”). Additionally, sequences that only provide information at the kingdom and species levels were discarded, particularly if they refer to rare bacteria from non-characterized environments (e.g., k\_Bacteria,..., s\_bacterium\_Te63R). Lastly, taxonomic nomenclature was carefully reviewed during the integration of databases, using the Python module ETE toolkit (version 3.0) (13) to retrieve synonyms from the NCBI database. The NCBI taxonomy database (14) was chosen as the reference for taxonomic annotation as it enables the identification of synonyms for all the taxonomic annotations in the databases and provides a standardized nomenclature. This procedure is capable of ensuring consistency and identifying misannotated organisms. One specific example mentioned is the identification of misannotated entries from the SILVA database, where certain entries labeled as bacteria are actually eukaryotic species, such as the annotation d\_Bacteria; p\_Proteobacteria; c\_Gammaproteobacteria; o\_Burkholderiales; f\_Comamonadaceae; g\_Paucibacter; s\_Cenchrus\_americanus, which is a plant species. This suggests that thorough steps were taken to ensure the accuracy and reliability of taxonomic information in the GSR-DB.

### Merging algorithm

The algorithm used to merge the Greengenes, SILVA, RDP, and vaginal processed databases was based on the integration algorithms proposed by Hsieh et al. (10). The algorithm took two databases as inputs and integrated them as follows (Fig. 1A). First, one database was assigned as the reference database (R) and the other as the candidate database (C). Then, for each entry in the candidate database, it checked whether the candidate taxon ( $T_C$ ) was already present in the reference database. The candidate entry (sequence and taxon) was added to the data set if not present. If  $T_C$  was present, the algorithm compared the candidate sequence ( $S_C$ ) to all the sequences in the reference data set ( $S_R$ ) with the same nomenclature as  $T_C$ . No integration was performed if  $S_C$  was identical or present as a substring in any of the  $S_R$ . On the other hand, if  $S_C$  was not found in the reference data set, the candidate entry (taxon and sequence) was added to the data set. The RDP data set was chosen as the first reference data set for its taxonomic consistency, then the remaining data sets were added in the following order: SILVA, Greengenes, and vaginal (Fig. 1B). The resulting data set is the GSR full-16S database.

### Creation of 16S variable region databases

#### Variable region extraction

The full-length GSR 16S database was used to create region-specific databases containing the most commonly used hypervariable regions in 16S analysis (V4, V3–V4, V1–V3, and V3–V5). The sequences for each region were extracted from the full-length GSR 16S database using the extract-reads function implemented in the QIIME 2 feature-classifier plugin (15) and the corresponding primers (3). QIIME 2 RESCRIPt plugin (8) was also used to dereplicate the resulting databases to remove redundant entries. These steps were also performed on the databases [Greengenes, Greengenes 2 (version 2022.10), GTDB (version 207.0), ITGDB, SILVA, and RDP) used in the validation analysis step.

#### Clustering

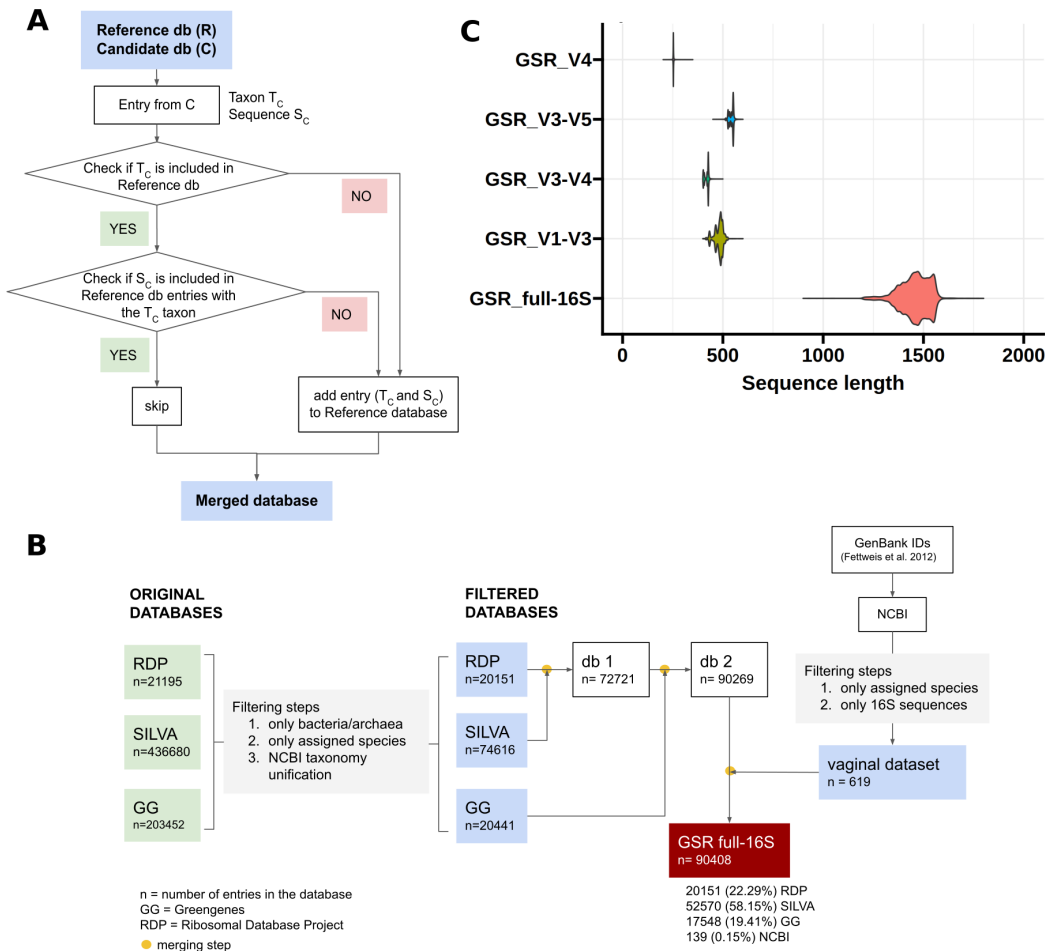
Upon extracting the variable region sequences from the full-length 16S database, we encountered numerous identical sequences that do not correspond to the same species. The region-specific GSR databases underwent clustering via CD-HIT (16, 17) at a 100% identity threshold to host unique sequences and improve species detection. Taxonomic designations of these clustered sequences were merged into a unified taxonomic name, as shown in the subsequent example.

Nomenclatures to be clustered:

1. *k\_Bacteria*; *p\_Firmicutes*; *c\_Bacilli*; *o\_Lactobacillales*; *f\_Lactobacillaceae*; *g\_Limosilactobacillus*; and *s\_Limosilactobacillus\_fermentum*.
2. *k\_Bacteria*; *p\_Firmicutes*; *c\_Bacilli*; *o\_Lactobacillales*; *f\_Lactobacillaceae*; *g\_Limosilactobacillus*; and *s\_Limosilactobacillus\_oris*.
3. *k\_Bacteria*; *p\_Firmicutes*; *c\_Bacilli*; *o\_Lactobacillales*; *f\_Lactobacillaceae*; *g\_Lactobacillus*; and *s\_Lactobacillus\_crispatus*.

Clustered nomenclature: *k\_Bacteria*; *p\_Firmicutes*; *c\_Bacilli*; *o\_Lactobacillales*; *f\_Lactobacillaceae*; *g\_Lactobacillus-Limosilactobacillus*; and *s\_Lactobacillus\_crispatus:Limosilactobacillus\_fermentum-oris*.

Clustered rank nomenclatures containing more than 10 taxonomic names were substituted with "Unknown" to prevent wordy nomenclatures, as we perceive them to lack informativeness. We intend to provide information on these sequence taxonomies on the GSR web server in the future.



**FIG 1** Creation of the GSR database. (A) Merging algorithm to create the GSR database. The algorithm takes as an input a Reference database (R) and a Candidate database (C). Entries from the Candidate database are susceptible to being added to the Reference database after being evaluated. (B) Database merging workflow to obtain the GSR full-length 16S database. It has a final size of 90,408 entries with the following source composition: 22.29% RDP, 58.15% SILVA, 19.41% Greengenes, and 0.15% NCBI. Merging steps were performed using the merging algorithm described in panel A. (C) Sequence length distribution of the GSR databases. Databases for the variable regions are clustered at 100% identity.

## Phylogeny construction

To build the phylogenetic tree, we used the pre-trained model for WoL marker genes and ASV data of DEPP software. This framework allows the positioning of the GSR-DB sequences onto the WoL species backbone tree via a convolutional neural network. The resulting tree is made available in both Newick and QIIME2 formats (Fig. S1).

## *In silico* mock community data sets

To assess the performance of the newly built database, three different mock communities (mockrobiota, vagimock, and gutmock) were constructed *in silico*. The vagimock and gutmock data sets simulate the relative abundance and species of biological samples from two different body sites. They were built from our GSR database with species commonly found in the human vagina and gut. The mockrobiota data sets were constructed using sequences obtained from the mockrobiota repository, a public resource for microbiome Bioinformatics benchmarking. From this repository, we recovered full-length 16S sequences provided only by data sets 3, 4, 5, and 12–23 (18).

Each *in silico* mock community data set contained five samples, with given microbial abundance profiles, and taxonomic and sequence information. The taxonomic information of the sequences was unified using the `ncbi_taxonomy` python module from the ETE toolkit (version 3.0) (13). Each mock community has a different level of complexity, which is crucial to reveal possible database issues (3). The composition of the mock communities at the species level can be found in Table S1.

## Validation

### *Validation data sets*

To evaluate the classification performance of the full-length and region-specific databases, we adapted the sequence length of the mock communities accordingly. For regions shorter than 460 nt (V3–V4 and V4), MiSeq Illumina reads were simulated using ART (19), using paired-end and single-end for V3–V4 and V4 regions, respectively. The corresponding parameters for each region were: V3–V4 `-ss 'MSv1' -amp -na -nf 0 -l 250 -c 1 -rs 123 --minQ 25 -p -m 450`; V4 `-ss 'MSv1' -amp -na -nf 0 -L 200 -c 1 -rs 123 --minQ 25`. Subsequently, the DADA2 module (20), implemented in QIIME2, was used to denoise the reads and construct representative sequences (rep-seqs). To recover the expected nomenclature of rep-seqs, rep-seqs were mapped to their corresponding community sequences using a convolution method, as performed in the TAX CREdit framework (15). For regions larger than 460 nt (full-16S, V1–V3, and V3–V5), sequences were directly treated as rep-seqs, due to software restrictions in simulating MiSeq (or PacBio) reads larger than 250 nt.

### *Classifier training and taxonomy assignment*

It is known that some classifiers are strongly affected by parameter configurations. Therefore, different parameters for classifier training and taxonomy assignment steps were tested to find the optimal configuration. The sequences and taxonomy of each database were used to train the multinomial naive Bayes classifier implemented in `q2-feature-classifier` QIIME2 module (15). During this training, the `n-gram-range` parameter was tested with the values [6,6] and [7,7] (default), as its developers have already reported these ranges as optimal. Then, these classifiers were used to perform the taxonomy assignment of rep-seqs for each region. During the taxonomy assignment, the confidence threshold for limiting taxonomic depth was tested with the values “disable,” 0.5, 0.7 (default), 0.9, and 0.98. Evaluating two `n-gram-range` values and five confidence thresholds generated 10 different taxonomic profiling for each database.

TABLE 1 Example of a multi-class confusion matrix for  $T_i = Lactobacillus iners^a$

		Expected taxon (E)		
		<i>Lactobacillus iners</i>	<i>Lactobacillus jensenii</i>	<i>Lactobacillus crispatus</i>
Assigned taxonomy (A)	<i>Lactobacillus iners</i>	TP		FP
	<i>Lactobacillus jensenii</i>			
	<i>Lactobacillus crispatus</i>	FN		TN

<sup>a</sup>TPs were all the *Lactobacillus iners* classified as *Lactobacillus iners*. FPs were all taxa classified as *Lactobacillus iners* that were not actual *Lactobacillus iners*. FNs were actual *Lactobacillus iners* not classified as *Lactobacillus iners*. TNs are other taxonomies different from *Lactobacillus iners* correctly classified as non-*Lactobacillus iners*.

**Parameter comparison**

To compare the performance of the 10 possible parameter configurations for each database, we calculated the average F1 scores across mock communities for each taxonomic level. The configuration with higher scores was retained for subsequent benchmarking of the databases.

**Database benchmarking**

The performance of the GSR database was compared with widely used databases, such as Greengenes, Greengenes 2, GTDB, SILVA, and RDP, but also with other available databases, such as ITGDB. Two independent approaches were used to assess the performances: the multi-class confusion matrix and the Bray-Curtis distances. The multi-class confusion matrix was used to evaluate the performance of a machine learning classification (e.g., naive Bayes classifier) by comparing the expected sequence taxonomy versus the classified (Table 1). This confusion matrix allowed us to obtain validation metrics such as accuracy, precision, recall, and F1 score by using the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\underline{s} = \sum_{i=1}^n a_i s_i \tag{4}$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

The four metrics were measured at each taxonomic level as follows: a match was called when two taxonomic names (ID) were identical between the expected (E) and the assigned (A) name or, in the case of assignments with clustered databases, a match was called when one name was included in the other one (for instance,  $A = Amylolactobacillus amylophilus$ -*Lactobacillus iners*;  $E = Lactobacillus iners$ ). For each taxonomic ID ( $T_i$ ), (i) TP was considered when  $T_i$  matched both A (assigned taxonomic ID) and E (expected taxonomic ID). (ii) FP was defined when  $T_i$  matches A but not E. (iii) FN was defined when  $T_i$  matches E but not A. (iv) TN was defined when neither A nor E matches  $T_i$ .

Finally, validation metrics for all taxonomic IDs were integrated using a weighted mean, taking the corresponding expected abundance as weight, using the following equation:

$$\underline{s} = \sum_{i=1}^{n_i} a_i s_i \tag{5}$$

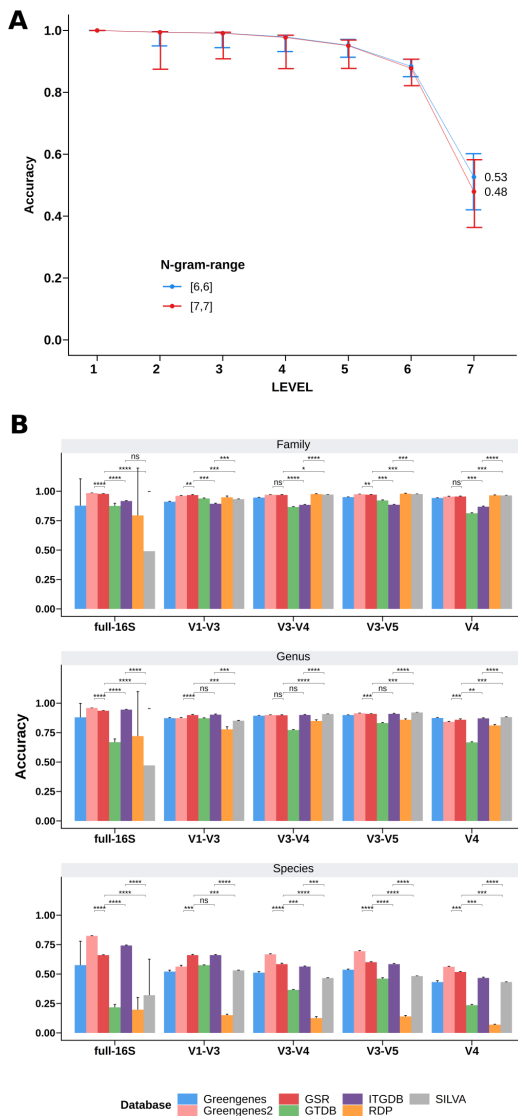
where  $\underline{s}$  = weighted mean score of the validation metric (precision, recall, F1 score or accuracy) for all taxonomies of a mock community,  $s_i$  = score of the validation metric for

taxonomic ID  $T_i$ ,  $a_i$  = expected relative abundance of taxonomic ID  $T_i$  (weight),  $n$  = total number of taxonomic IDs included in the mock community.

Bray-Curtis distances were calculated between the expected and assigned composition for each sample in R (version 4.2.1) using the vegan package (version 2.6-4).

To discover significant differences in performance metrics, F1 scores and Bray-Curtis distances were compared among the GSR, Greengenes 2, ITGDB, and SILVA databases using the Wilcoxon test ( $P$ -values adjusted by the Benjamini-Hochberg method).

Additionally, since different databases might use different taxonomic nomenclature, in order to consider synonyms of scientific names as correct matches, taxonomy unification (ETE toolkit v.3.0) was applied to each taxonomic classification before comparisons.



**FIG 2** Database evaluation with 10-fold cross-validation. (A) Average n-gram parameter performance in 10-fold cross-validation. N-gram-range with value [6,6] shows a better performance in the 10-fold cross-validation data sets. (B) Tenfold cross-validation results. Average accuracy of classification at family, genus, and species level. Error bars are the standard deviations. Wilcoxon test was conducted between accuracy scores of GSR, Greengenes2, ITGDB, and SILVA databases. ns, not significant; \* $P_{adj} < 0.05$ ; \*\* $P_{adj} < 0.01$ ; and \*\*\* $P_{adj} < 0.001$ .

### Tenfold cross-validation

We conducted a 10-fold cross-validation to validate the results obtained with the mock communities. Train and test data sets for each database were built using the scikit-learn Python module (v0.24.1). These training data sets were used to train a naive Bayes classifier in QIIME2, setting the n-gram-range parameter to [6,6], as it yielded globally better classification results (Fig. 2A). These classifiers were used to assign the taxonomic nomenclature to the test data sets, with QIIME2 using the default parameters. Accuracy was assessed following the methodology outlined by Edgar (21).

### Gut and vaginal microbial data sets

To further validate our database performance, we performed a case study using actual biological data from human gut samples (22) and human vaginal samples (23). These data sets contained V4 amplicon sequences. Taxonomic assignments were performed in both data sets using QIIME2 naive Bayes feature classifier, with the following V4 databases as reference: Greengenes, Greengenes 2, GSR, GTDB, ITGDB, MetaSquare, SILVA, and RDP. The n-gram-range parameter was set to [7, 7] and the confidence threshold to “disable,” as these were the parameters found to perform best in the validation step.

### Computational benchmarking

Furthermore, we also tested the computational cost of obtaining a taxonomic profile with the QIIME2 naive Bayes classifier with each of the V4 reference databases employed in this case study. We measured the time and memory consumption of the classifier training and the taxonomic assignment processes. Time was measured with the Python built-in time module, and memory consumption was tracked using the memory\_profiler module. These analyses were run on a computer with an Intel Xeon Gold 6238 processor with 44 CPUs and 187 GB of RAM, and Ubuntu 18.04.4. Classifier training was run with default settings. Taxonomy assignment was performed by setting the confidence threshold to “disable” and using 10 threads.

## RESULTS

### GSR database

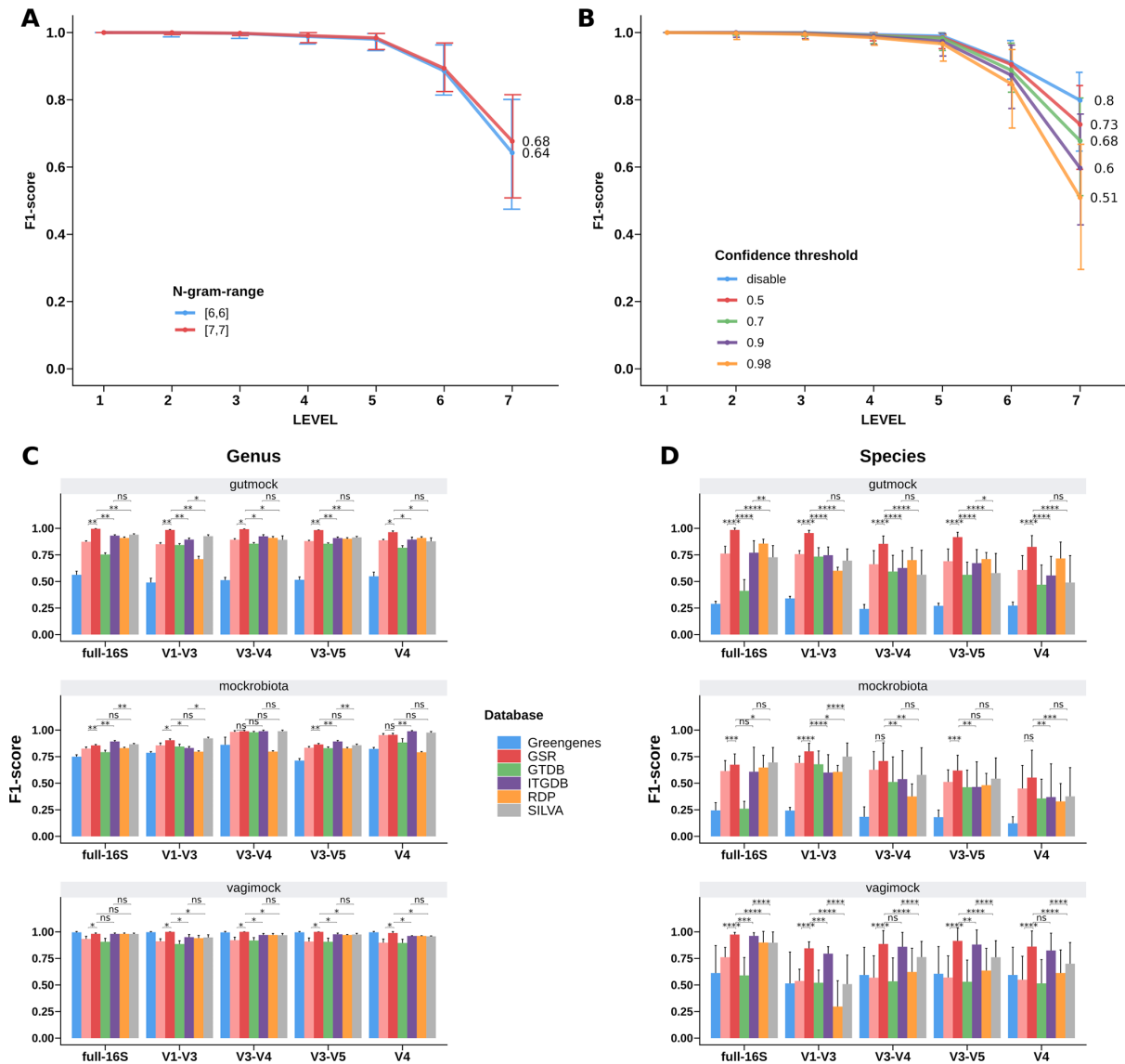
To optimize the prokaryotic taxonomic assignment, we created the GSR database by integrating and manually curating Greengenes (v13\_8, 99%), SILVA (v138, 99%), and RDP (train set no. 18) data sets (Fig. 1A and B). The integrated full-length 16S GSR database has a total size of 90,408 sequences, with the following source composition: 22.29% RDP, 58.15% SILVA, 19.41% Greengenes, and 0.15% NCBI (vaginal-specific sequences). The source composition and total size of the variable region databases are shown in Table 2. The V1–V3 and V4 databases are those with fewer available sequences. The sequence length distributions of the GSR databases are presented in Fig. 1C.

TABLE 2 Source composition of GSR databases<sup>a</sup>

Database	Region	Cluster	Source database				Total
			RDP	SILVA	Greengenes	NCBI	
GSR	V1–V3	100%	6,707 (24.96%)	14,595 (54.31%)	5,521 (20.54%)	51 (0.19%)	26,874
	V3–V4	100%	15,401 (30.79%)	22,621 (45.23%)	11,949 (23.89%)	45 (0.09%)	50,016
	V3–V5	100%	16,659 (29.43%)	26,239 (46.35%)	13,655 (24.12%)	53 (0.09%)	56,606
	V4	100%	12,670 (32.65%)	16,186 (41.72%)	9,916 (25.56%)	29 (0.07%)	38,801
	Full-16S	None	20,151 (22.29%)	52,570 (58.15%)	17,548 (19.41%)	139 (0.15%)	90,408

<sup>a</sup>Number of entries in each GSR database that were recovered from each source database.





**FIG 3** Database evaluation with mock communities. (A and B) Benchmarking of n-gram-range (A) and confidence threshold (B). The median F1 score is shown at each taxonomic level across all tested databases, regions, and validation data sets. Error bars represent the interquartile range ( $N = 900$ ). Full data for the family, genus, and species level are available in Tables S2 to S4. (C and D) Database benchmarking at genus (C) and species (D) levels using validation metrics. The mean F1 score across the five metagenomic samples is shown for each evaluated region and data set. Error bars are the standard deviation. Database benchmarking results at the family level are available in Fig. S2. Precision and recall metrics are available in Tables S5 to S7 for family, genus, and species levels, respectively. Wilcoxon test was conducted between F1 scores of Greengenes2, GSR, ITGDB, and SILVA databases. ns, not significant;  $*P_{adj} < 0.05$ ;  $**P_{adj} < 0.01$ ; and  $***P_{adj} < 0.001$ .

### QIIME2 parameters impact taxonomic assignment performance

The 16S rRNA analysis pipeline of QIIME2 (24) includes training a naive Bayes classifier with a reference database and a subsequent taxonomic assignment of the rep-seqs (15). In these two steps, we tested different values of n-gram-range and confidence threshold parameters for each database (full-length 16S and specific 16S regions) as it is known to affect the classifier’s performance. Fig 3A and B summarizes the performance of the aforementioned parameters across tested databases and regions. Two n-gram-range values were tested: [6,6] and [7,7]. The Wilcoxon test showed that [7,7] performed better than [6,6] ( $P < 0.0001$ ) (Fig. 3A). Confidence threshold values show significant differences in F1 score (Fig. 3B,  $P < 0.0001$  for all comparisons in a pairwise manner), precision, and

recall at both genus and species levels (Tables S2 to S4). Setting the confidence threshold to “disable” provided the best classification results at the species level, suggesting that setting a confidence threshold for the QIIME2 classifier notably restricts the predictions at the species level without improving the predictions at higher levels. Therefore, the n-gram-range of [7,7] and “disable” confidence threshold were further used to benchmark the GSR database with other already existing databases.

### GSR outperforms most existing databases across all tested regions

To assess the performance of the newly created database, we benchmarked the GSR database with the other existing databases (Greengenes, Greengenes2, SILVA, RDP, ITGDB, and MetaSquare), using two different approaches: validation metrics (F1 score shown in Fig. 3C and D; Fig. S2, precision and recall shown in Tables S5 to S7) and Bray-Curtis distances (Fig. S3; Table S8). In order to increase the robustness of the results, we defined the combination of the F1 score and the Bray-Curtis distance as the validation scores. The database with the best validation scores will achieve the highest F1 score and the shortest Bray-Curtis distance.

At the family level, the Greengenes2, GSR, ITGDB, and SILVA databases reached, in most cases, the best validation scores across regions in all the validation data sets (Fig. S2 and S3A; Tables S5 and S8). At genus level (Fig. 3C; Fig. S3B; Tables S6 and S8), GSR achieved significantly better validation scores across almost all regions in the gutmock data set, followed by Greengenes2, ITGDB, and SILVA databases. In the mockrobiota data set, ITGDB and SILVA achieved the best validation scores across regions, sharing similar values with GSR and for V1–V3 and V3–V4. Finally and most importantly, at the species level (Fig. 3D; Fig. S3C; Tables S7 and S8), except for the full-16S region where ITGDB had the best validation score, the GSR database presented the best scores for almost all the regions and validation data sets. Overall, these results indicate that whereas the database performance remains relatively stable up to the family level, substantial differences were observed at the genus and species level, with GSR showing the best performance results among the tested databases.

The Greengenes database performed worst in all tested environments, except for the vagimock data set at the genus level, for which it performed similarly to the other databases. On the other hand, the RDP, GTDB, and SILVA databases yielded better results across all environments and regions. Previous studies have already pointed out the increased accuracy of SILVA and RDP databases in comparison to Greengenes (3), mainly due to the fact that, in the last few years, SILVA and RDP have been updated more frequently than Greengenes. The better performance of Greengenes in identifying genus-level classifications within the vagimock data set could be attributed to the low complexity of this mock community. It has been observed that database limitations may not be as apparent when analyzing mock communities with limited characteristics (3).

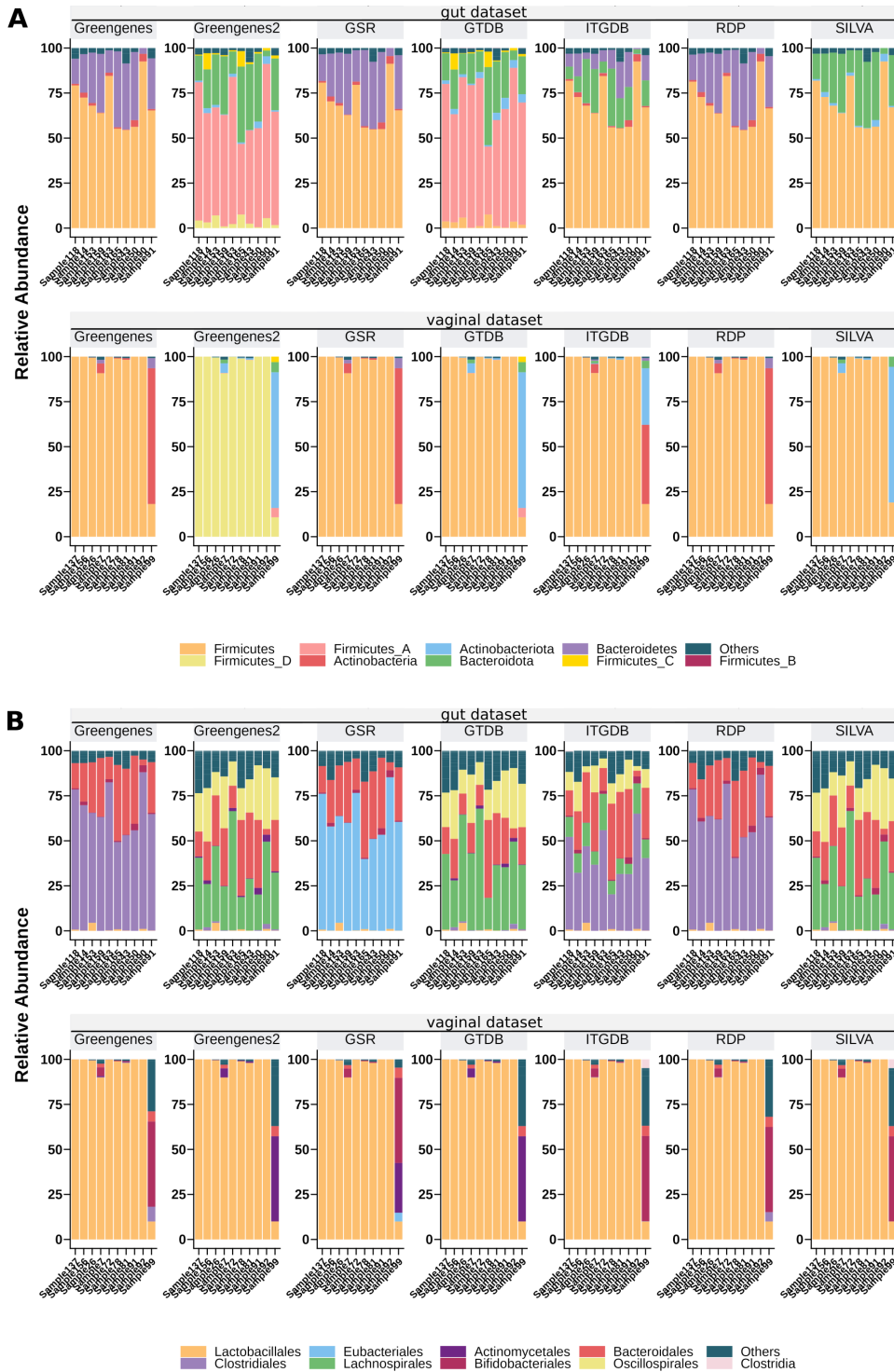
Results from the 10-fold cross-validation showed that Greengenes2, GSR, and ITGDB databases presented significantly better performance than the other databases in almost all levels and regions, which is consistent with the results obtained in the mock community validation (Fig. 2). At the species level, Greengenes 2 outperformed GSR, with the exception of the region V1–V3.

### Case study: vaginal and gut data sets

In methodological benchmarking studies, it is crucial to contextualize the benchmarking outcomes using actual biological data. Therefore, 10 vaginal and 10 gut microbiome samples were analyzed from Vargas et al. (23) and Yáñez et al. (22) data sets, containing 2,089 and 31,885 V4 representative sequences, respectively. These data sets allowed us to assess the consistency of taxonomic nomenclature among our newly built database and other existing databases, including Greengenes, Greengenes2, GTDB, ITGDB, RDP, and SILVA. Additionally, the analysis of real data sets allowed us to compare the computational cost of taxonomy profiling among the aforementioned reference databases.

**GSR annotation enhances taxonomic nomenclature consistency**

Each database uses different synonym terms for the same NCBI taxonomy ID, as shown in Fig. 4. For instance, in Fig. 4A, Greengenes, GSR, and RDP use exclusively the term *Bacteroidetes* for NCBI:txid976, SILVA, Greengenes2, and GTDB use the synonym *Bacteroidota*, and ITGDB uses both of them. Similarly, for NCBI:txid201174,



**FIG 4** Relative abundance of gut and vaginal samples at phylum (A) and order (B) levels. Only relevant taxa are displayed. The remaining taxa are included in the label “Others.” Family level is available in Fig. S4.

TABLE 3 Benchmarking results for the classifier training step<sup>a</sup>

Database	Elapsed time	Classifier size (MB)	Memory usage peak (GB)	Memory usage mean (GB)
RDP	0:01:23	20.85	4.46	2.6
GSR	0:02:28	25.61	6.3	4.06
Greengenes	0:02:51	28.14	3.52	2.43
ITGDB	0:06:38	44.92	14.88	8.79
GTDB	0:06:42	49.0	14.6	9.24
Greengenes2	0:11:11	47.77	15.2	9.53
SILVA	0:40:07	106.55	23.54	16.75
Metasquare	1 day 21:27:39	589.99	175.92	125.98

<sup>a</sup>A QIIME2 naive Bayes classifier was trained with each one of the reference databases using the default parameters.

Greengenes, GSR, and RDP use the term *Actinobacteria*, SILVA, Greengenes2, and GTDB use the synonym *Actinobacteriota*, and ITGDB uses both of them. In addition, GTDB and Greengenes2 split the phylum *Firmicutes* into several clusters, namely *Firmicutes*, *Firmicutes\_A*, *Firmicutes\_B*, *Firmicutes\_C*, and *Firmicutes\_D*. At the order level, another example can be found in Fig. 4B. For NCBI:txid186802, Greengenes and RDP use the term *Clostridiales*, and GSR uses the synonym *Eubacteriales*. SILVA, Greengenes2, GTDB, and ITGDB use several non-NCBI terms such as *Clostridia*, *Lachnospirales*, and *Oscillospirales*. Moreover, ITGDB also uses the accepted term *Clostridiales*. Finally, other taxonomy inconsistencies can be found at the family level (Fig. S4). For NCBI:txid216572, whereas Greengenes uses the term *Ruminococcaceae* and GSR uses its synonym *Oscillospiraceae*, SILVA, GTDB, and ITGDB use both aforementioned terms, and SILVA and ITGDB also use the synonym *Hungateiclostridiaceae*. Taken together, these results indicate that Greengenes, RDP, and GSR databases have robust taxonomic nomenclatures, using exclusive terms for one NCBI taxonomy id. In contrast, SILVA, Greengenes2, GTDB, and ITGDB databases use several terms to refer to the same taxon, some of which are non-NCBI terms.

### Computational benchmarking

The benchmarking was performed in two steps in which the databases were involved: classifier training and taxonomic assignment. A naive Bayes classifier was trained in QIIME2 using the V4 region of each reference database. Training time and memory usage for each classifier are shown in Table 3. The most computationally efficient classifier training was obtained using the RDP, GSR, or Greengenes databases. These three classifiers were trained within 3 minutes and required less than 7 GB of RAM. ITGDB, Greengenes2, and GTDB classifiers show an increased computational cost, doubling the time and memory usage of the aforementioned ones. The SILVA classifier required a significantly higher amount of computational resources, taking up to 40 minutes and 25 GB of RAM to be trained. The MetaSquare (version 1.0.2) classifier was the most computationally expensive to train, being time-consuming and memory-intensive.

The trained classifiers were then used to perform a taxonomy profiling of the intestinal and vaginal data sets, with the confidence threshold set to “disable” and multithreading used with 10 threads. The benchmarking results for this step are presented in Table 4. Resource consumption resembled the pattern seen in the classifier training step. Greengenes classifier was the most computationally inexpensive, followed by GSR and RDP, which are still affordable. ITGDB and GTDB almost double the required resources, and SILVA and Greengenes2 were the most resource consuming. MetaSquare classifier was also tested, but its taxonomy assignment could not be completed due to a lack of computational resources.

TABLE 4 Benchmarking results for the taxonomy assignment step<sup>a</sup>

Data set	Database	Elapsed time	Memory usage peak (GB)	Memory usage mean (GB)
Gut	Greengenes	0:00:17	7.09	3.11
	RDP	0:00:46	18.21	6.97
	GSR	0:01:08	18.55	5.11
	ITGDB	0:01:28	33.62	11.74
	Greengenes2	0:01:31	49.46	16.72
	GTDB	0:01:44	37.47	13.46
Vagina	SILVA	0:02:17	48.93	16.36
	Greengenes	0:00:09	5.64	2.55
	GSR	0:00:29	14.85	4.33
	RDP	0:00:29	12.56	4.28
	Greengenes2	0:00:53	42.86	9.2
	ITGDB	0:01:00	26.59	6.83
	GTDB	0:01:07	30.12	7.83
	SILVA	0:01:30	37.86	9.28

<sup>a</sup>Gut and vaginal datasets were taxonomically profiled using previously trained classifiers of each reference database. MetaSquare classifier was also tested but no results were obtained due to a lack of computational resources (>187 GB RAM).

## DISCUSSION

In this study, we generated a new 16S database for prokaryotic and archaea organisms: the GSR database. The performance of the GSR database was assessed in conjunction with six other existing 16S databases: Greengenes, Greengenes2, GTDB, ITGDB, SILVA, and RDP. Our attempt to evaluate the MetaSquare database was hampered due to its extremely high demand for computational resources compared with other existing databases (Table 3). We believe these requirements are unreasonable and impractical. Therefore, we discarded MetaSquare for subsequent analysis and cannot rationally advise its use.

Before database comparisons, we first explored the best parameter configuration for each of the five databases, as previous studies have pointed out the impact of n-gram-range and confidence threshold parameters on classifier performances (15). The n-gram-range value of [7,7] performed better than [6,6], whereas the confidence threshold value of “disable” significantly outperformed at the species level. These results suggest that confidence threshold value plays an essential role in the taxonomic resolution and should be consistently reported in studies. Based on our results, we recommend using the n-gram-range of [7,7] and confidence threshold “disable” values in microbial profiling studies that utilise the GSR database.

Regarding database performance, GSR outperformed GTDB, SILVA, RDP, and Greengenes databases in almost all tested environments and regions. ITGDB database presented a comparable performance to the GSR database, performing better in the mockrobiota data set. However, the ITGDB database has some significant shortcomings, not detected in the GSR database, such as taxonomical discrepancies and lower computational efficiency. Based on the most unbiased experimental evaluation, GSR was only outperformed by Greengenes 2, with the exception of the region V1–V3.

The case study performed with gut and vaginal sample data sets exposed the consequences of not unifying the taxonomy when merging databases with different taxonomic annotations. ITGDB database presented multiple cases of taxonomical inconsistencies (Fig. 4), where several synonym terms were used to refer to the same taxonomic clade. A similar behavior is also noticeable in SILVA but to a lesser extent. The lack of a consistent taxonomy might severely interfere with microbial taxonomy analyses, impacting diversity metrics or differential composition analyses. In this regard, it is worth noting that the GSR database does not suffer from taxonomic consistency issues and can provide more reliable and robust results. Furthermore, this case study

revealed that the computational resources used by QIIME2 differ depending on which reference database is employed. ITGDB database made QIIME2 consume twice as many computational resources as GSR (Tables 3 and 4), making GSR a more suitable alternative for obtaining high-resolution taxonomy profiles at lower computing costs.

Despite the described results, several limitations need to be considered. First, the lack of testing on non-human samples, such as soil and water samples, raises concerns about the generalizability of the database to different contexts. Without this information, we cannot fully understand how well our database will perform in these environments. Second, the GSR database only contains sequences from known species, excluding unclassified organisms or organisms labeled as uncultured. While this may be detrimental for the analysis of environments containing a large amount of unknown or uncultured species (8), we demonstrated that it improves species detection in well-described environments, such as human body sites. Additionally, the utilization of a single classification software (QIIME2 naive Bayes) precludes the ability to extrapolate the performance of our database to alternative classification methods, as the use of different software may yield different results. Finally, another limitation is the restricted testing conducted in human-like environments. Although gut and vagina samples have been examined in this study, the database usefulness could be more comprehensively evaluated by extending the analysis to other human environments, such as skin and saliva. Overall, the GSR database demonstrates potential, but it is crucial to acknowledge and address the aforementioned factors to obtain a thorough understanding of its applications and potential drawbacks.

While 16S amplicon-based sequencing has limitations, its low cost and simplified methodology still make it a valuable tool for analyzing the microbiome composition, especially for low-biomass samples. The vast amount of data generated during the last decade can not only help to answer pressing questions about microbiome-disease relationships in larger epidemiological studies but also can be used along with shotgun metagenomic sequencing data to explore new clinical applications (25). Therefore, the GSR database offers several advantages for microbial taxonomic classification using 16S sequencing. It integrates three of the main reference databases, ensuring a comprehensive and accurate taxonomic annotation. The taxonomy consistency allows for reliable analysis, which is crucial for the robustness of microbiome studies. GSR database also demonstrates an improved performance with microbial communities containing mainly known species, enhancing its utility in various applications. Finally, its usage is not computationally expensive, making it accessible to researchers with limited computational resources. Overall, these features make the GSR database a valuable resource for the scientific community to further investigate microbial communities.

Upcoming versions of GSR-DB will prioritize keeping integrated databases current with the latest versions and consistently updating the taxonomy to align with the most recent NCBI taxonomy release. Additionally, we have the intention to expand the web server's functionalities, allowing users to navigate through the database.

## ACKNOWLEDGMENTS

This work was supported by the Instituto de Salud Carlos III/FEDER (PI20/00130). Funding for open access charge was provided by Instituto de Salud Carlos III/FEDER.

L.-A.G.M. and S.V.-A. were involved in conceptualization, formal analysis, methodology, software, validation, and writing of the original draft. C.M. conceptualized and supervised the study, acquired funding, and reviewed and edited the manuscript.

## AUTHOR AFFILIATIONS

<sup>1</sup>Microbiome Lab, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron, Barcelona, Spain

<sup>2</sup>Medicine Department, Universitat Autònoma de Barcelona, Barcelona, Spain

## AUTHOR ORCID*s*

Leidy-Alejandra G. Molano <http://orcid.org/0000-0003-0069-6098>

Sara Vega-Abellaneda <http://orcid.org/0009-0007-0051-8554>

Chaysavanh Manichanh <http://orcid.org/0000-0002-2287-4003>

## FUNDING

Funder	Grant(s)	Author(s)
MEC   Instituto de Salud Carlos III (ISCIII)	PI20/00130	Leidy-Alejandra Molano Sara Vega-Abellaneda

## AUTHOR CONTRIBUTIONS

Leidy-Alejandra G. Molano, Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing | Sara Vega-Abellaneda, Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing.

## DATA AVAILABILITY

The database and the validation data are available at the following link: <https://manichanh.vhir.org/gsrdb/>.

## ADDITIONAL FILES

The following material is available [online](#).

### Supplemental Material

**Fig. S1 (mSystems00950-23-s0001.tiff).** Visualization of the phylogenetic tree for GSR-DB, rendered using Empress.

**Fig. S2 (mSystems00950-23-s0002.tiff).** Database benchmarking at the family level using validation metrics.

**Fig. S3 (mSystems00950-23-s0003.tiff).** Database benchmarking using Bray-Curtis distances between expected and observed composition at family, genus, and species levels.

**Fig. S4 (mSystems00950-23-s0004.tiff).** Relative abundance of gut and vaginal samples at the family level.

**Supplemental Material (mSystems00950-23-s0005.docx).** Databases.

**Supplemental Legends (mSystems00950-23-s0006.docx).** Legends for supplemental tables and figures.

**Supplemental Tables (mSystems00950-23-s0007.pdf).** Tables S1 to S8.

## REFERENCES

- Gurung M, Li Z, You H, Rodrigues R, Jump DB, Morgun A, Shulzhenko N. 2020. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 51:102590. <https://doi.org/10.1016/j.ebiom.2019.11.051>
- Lavelle A, Nancey S, Reimund J-M, Laharie D, Marteau P, Treton X, Allez M, Roblin X, Malamut G, Oeuuvray C, Rolhion N, Dray X, Rainteau D, Lamaziere A, Gauliard E, Kirchgessner J, Beaugerie L, Seksik P, Peyrin-Biroulet L, Sokol H. 2022. Fecal microbiota and bile acids in IBD patients undergoing screening for colorectal cancer. *Gut Microbes* 14:2078620. <https://doi.org/10.1080/19490976.2022.2078620>
- Abellan-Schneyder I, Machado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K. 2021. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *mSphere* 6:e01202-20. <https://doi.org/10.1128/mSphere.01202-20>
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072. <https://doi.org/10.1128/AEM.03006-05>
- Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 50:D785–D794. <https://doi.org/10.1093/nar/gkab776>
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–D596. <https://doi.org/10.1093/nar/gks1219>
- Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267. <https://doi.org/10.1128/AEM.00062-07>

8. Robeson MS, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, Bokulich NA. 2021. RESCRIPt: reproducible sequence taxonomy reference database management. *PLOS Comput Biol* 17:e1009581. <https://doi.org/10.1371/journal.pcbi.1009581>
9. McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, Morton JT, Nicolaou G, Parks DH, Karst SM, Albertsen M, Hugenholtz P, DeSantis T, Song SJ, Bartko A, Havulinna AS, Jousilahti P, Cheng S, Inouye M, Niiranen T, Jain M, Salomaa V, Lahti L, Mirarab S, Knight R. 2023. Greengenes2 unifies microbial data in a single reference tree. *Nat Biotechnol*. <https://doi.org/10.1038/s41587-023-02026-w>
10. Hsieh Y-P, Hung Y-M, Tsai M-H, Lai L-C, Chuang EY. 2022. 16S-ITGDB: an integrated database for improving species classification of prokaryotic 16S ribosomal RNA sequences. *Front Bioinform* 2:905489. <https://doi.org/10.3389/fbinf.2022.905489>
11. Liao C-C, Fu P-Y, Huang C-W, Chuang C-H, Yen Y, Lin C-Y, Chen S-H. 2022. MetaSquare: an integrated metadatabase of 16S rRNA gene amplicon for microbiome taxonomic classification. *Bioinformatics* 38:2930–2931. <https://doi.org/10.1093/bioinformatics/btac184>
12. Fettweis JM, Serrano MG, Sheth NU, Mayer CM, Glascock AL, Brooks JP, Jefferson KK, Vaginal Microbiome Consortium (additional members), Buck GA. 2012. Species-level classification of the vaginal microbiome. *BMC Genomics* 13:S17. <https://doi.org/10.1186/1471-2164-13-S8-S17>
13. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 33:1635–1638. <https://doi.org/10.1093/molbev/msw046>
14. Schoch CL, Ciuffo S, Domrachev M, Hottton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020:baaa062. <https://doi.org/10.1093/database/baaa062>
15. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90. <https://doi.org/10.1186/s40168-018-0470-z>
16. Li W, Godzik A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
17. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
18. Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF, Dutton RJ, Turnbaugh PJ, Knight R, Caporaso JG. 2016. Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 1:e00062-16. <https://doi.org/10.1128/mSystems.00062-16>
19. Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* 28:593–594. <https://doi.org/10.1093/bioinformatics/btr708>
20. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>
21. Edgar RC. 2018. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* 6:e4652. <https://doi.org/10.7717/peerj.4652>
22. Yáñez F, Soler Z, Oliero M, Xie Z, Oyarzun I, Serrano-Gómez G, Manichanh C. 2021. Integrating dietary data into microbiome studies: a step forward for nutri-metacoims. *Nutrients* 13:2978. <https://doi.org/10.3390/nu13092978>
23. Vargas M, Yáñez F, Elias A, Bernabeu A, Goya M, Xie Z, Farrás A, Sánchez O, Soler Z, Blasquez C, Valle L, Olivella A, Muñoz B, Brik M, Carreras E, Manichanh C. 2022. Cervical pessary and cerclage placement for preterm birth prevention and cervicovaginal microbiome changes. *Acta Obstet Gynecol Scand* 101:1403–1413. <https://doi.org/10.1111/aogs.14460>
24. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:852–857. <https://doi.org/10.1038/s41587-019-0209-9>
25. Liang H, Jo J-H, Zhang Z, MacGibeny MA, Han J, Proctor DM, Taylor ME, Che Y, Juneau P, Apolo AB, McCulloch JA, Davar D, Zarour HM, Dzutsev AK, Brownell I, Trinchieri G, Gulley JL, Kong HH. 2022. Predicting cancer immunotherapy response from gut microbiomes using machine learning models. *Oncotarget* 13:876–889. <https://doi.org/10.18632/oncotarget.28252>