BMC Bioinformatics

RESEARCH

Open Access

SpeciateIT and vSpeciateDB: novel, fast, and accurate per sequence 16S rRNA gene taxonomic classification of vaginal microbiota

Johanna B. Holm[1], Pawel Gajer[1] and Jacques Ravel[1*]

*Correspondence:
jravel@som.umaryland.edu

[1] Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Abstract

Background: Clustering of sequences into operational taxonomic units (OTUs) and denoising methods are a mainstream stopgap to taxonomically classifying large numbers of 16S rRNA gene sequences. Environment-specific reference databases generally yield optimal taxonomic assignment.

Results: We developed SpeciateIT, a novel taxonomic classification tool which rapidly and accurately classifies individual amplicon sequences (https://github.com/Ravel-Laboratory/speciateIT). We also present vSpeciateDB, a custom reference database for the taxonomic classification of 16S rRNA gene amplicon sequences from vaginal microbiota. We show that SpeciateIT requires minimal computational resources relative to other algorithms and, when combined with vSpeciateDB, affords accurate species level classification in an environment-specific manner.

Conclusions: Herein, two resources with new and practical importance are described. The novel classification algorithm, SpeciateIT, is based on 7th order Markov chain models and allows for fast and accurate per-sequence taxonomic assignments (as little as 10 min for 10^7 sequences). vSpeciateDB, a meticulously tailored reference database, stands as a vital and pragmatic contribution. Its significance lies in the superiority of this environment-specific database to provide more species-resolution over its universal counterparts.

Keywords: Amplicon sequencing, Taxonomic classification, Vaginal microbiota, 16S rRNA gene

Background

High-throughput next generation sequencing has revolutionized the field of metataxonomics by producing millions of sequences at an affordable cost, increasing the depth at which microbial communities are characterized. However, large sequence datasets have led to new challenges such as high computational costs associated with data analyses and accurate taxonomic classification. Bioinformaticists have developed novel sequence clustering algorithms which either produce, for a given similarity threshold, groups of sequences known as operational taxonomic units (OTUs) [1–3], or reduce sequencing

errors and minimize noise in the data [4–7]. These approaches have proven useful, though noise reduction/error correction may artificially remove or produce diversity [6] and the process of OTU clustering simply shifts the computational cost from taxonomic classification assignment to clustering and is not without problems. Most significantly, the transitive taxonomic assignments obtained from an OTU representative sequence are often flawed as 10–30% of sequences within the OTU, if processed separately, are assigned different taxonomy, thus challenging the ecological value of an OTU [8]. Further, output from clustering-based analyses are dataset specific and when data are added to a study, clustering must be run again and at increasing computational cost.

To alleviate these issues, we developed SpeciateIT, an algorithm capable of fast, accurate individual sequence taxonomic classification. Using a model guide tree and 7th order Markov chain models to represent bacterial species trained on taxonomy-adjusted amplicon specific regions sequences, SpeciateIT requires little computational resources, and can quickly process large sequence datasets. Additionally, environment-specific reference databases improve species-level classification accuracy and precision by reducing misclassification to species irrelevant to the environment and increasing study reproducibility and generalizability to other studies [9, 10]. For these reasons, we have also developed vSpeciateDB, a set of custom databases of reference sequences for classifying vaginal microbiota. SpeciateIT models for the vaginal microbiota correctly classified to the species level 99.9% of training set 16S rRNA gene V1-V3, V3-V4, and V4 regions sequences. This is a major improvement over the RDP Naïve Bayesian Classifier [11], which is capable of 99% classification accuracy of known sequences and 76% of novel sequences (not part of the training data) to the genus-level [12, 13].

Methods

SpeciateIT algorithm

The core model building algorithm in SpeciateIT produces higher order Markov chain models for groups of phylogenetically related sequences. These groups were organized in a model tree reflecting the species lineages (buildModelTree). For each node of the model tree (except the root) a fasta file of all reference sequences corresponding to the node's subtree was created and used to build Markov chain models (buildMC). Node-specific classification error thresholds were estimated to produce confidence in taxonomic assignments.. Node-specific classification error thresholds were estimated to produce confidence in taxonomic assignment. The offset coefficient defines the threshold adjustment for sequence classification, where the threshold is set at the maximum posterior probability of a sibling species' sequences with respect to the reference species' model, reduced by a specified value (e.g., 0.7, see Fig. S1). This reduction allows for the possibility of assigning novel sequences that might not match the reference species exactly but are still considered close enough for classification. (est_err_thlds). Classification of a query sequence begins at the top of the model tree. The model producing the highest posterior probability is chosen, and the assignment is given to the sequence given that the posterior probability is greater than the classification error threshold for that model. Model comparisons at the next and lower taxonomic levels commence until either a terminal node (species-level) classification is reached, or the classification error

threshold criterion is not met. All code presented here is available at https://github.com/ravel-lab/speciateIT and is included in SpeciateIT.Rmd.

vSpeciateDB curation

All steps, tests, and validations were performed on a 2021 Macbook Pro with an Apple M1 Max processor and 64G RAM.

Environment-specific reference databases increase classification accuracy and study reproducibility [9, 10]. All code for reference database curation can be found in speciateIT.Rmd. To build the vagina-specific vSpeciateDB reference database, we extracted sequences from the GTDB [14] small subunit rRNA gene sequence dataset (https://data.gtdb.ecogenomic.org/releases/release214/214.1/genomic_files_all/ssu_all_r214.tar.gz, referred herein as GTDB-SSU v214.1) for species found in the vaginal microbiome [15, 16] as per VIRGO2 (virgo.igs.umaryland.edu, and species list available on https://github.com/ravel-lab/speciateIT) resulting in 308,611 16S rRNA gene sequences from 14 bacterial phyla including 16 classes, 36 orders, 77 families, 497 genera, and 2224 species (Table 1). Sequences were truncated to the V1-V4 region using tagcleaner.pl [17] with the 27F and 806R primer sequences allowing for 9 and 17 mismatches, respectively. Using mothur v.1.48.0 [2], truncated sequences were dereplicated after filtering those with ambiguous bases and lengths < 250 bp or > 1000 bp. RDP-formatted lineages were re-formatted (reversed and tab-delimited), and a taxonomy file was created connecting sequence IDs to species annotation.

Production of 16S rRNA gene sequence region-specific datasets

The final de-replicated V1-V4 dataset was used to produce datasets for the 16S rRNA gene amplicon V1-V3, V3-V4, and V4 regions using tagcleaner.pl (version 0.16) [17] and the 319F, 515F and 534R primers allowing for 9, 3, and 5 mismatches, respectively. V1-V3 and V3-V4 sequences were screened to remove sequences < 400 and > 500 bp long. V4 sequences were required to be 240–260 bp. Each dataset was then dereplicated using the unique.seqs command from mothur v.1.48.0 [2]. SpeciateIT models were constructed for each dataset and training set evaluation identified incorrectly classified sequences, which were subsequently removed (Fig. S2). Most incorrect classifications at this stage originated from species over-represented in the reference database including *Escherichia coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Salmonella enterica*, *Acinetobacter*

Table 1 Summary of sequence information comprising each SpeciateIT 16S rRNA gene sequence region-specific database

	Vaginal subset of GTDB-SSU v214.1	V1–V3	V3–V4	V4
Sequences	308,611	4502	2584	1735
Species	2224	1322	1272	1165
Genera	497	406	415	411
Families	77	74	73	69
Orders	36	35	35	33
Classes	16	15	15	14
Phyla	14	13	13	12

baumannii, and *Streptococcus pneumoniae, S. epidermidis,* and *S. agalactiae*. Therefore, for these species and others with more than 50 sequences, three sequences with the highest posterior probabilities from training set evaluations were maintained. To further ensure quality of training data, pairwise alignments of all sequences from each species were performed with Biostrings v2.70.2 [18], and sequences with less than 90% identity to all other sequences in the species were removed (Fig. S3). Next, multiple sequence alignments for each database were produced with MAFFT v7.394 [19], and used to build phylograms with FastTree 2.1.10 [20]. Semi-supervised clustering using VI-cut [21] was performed and VI-cut clusters were evaluated for species purity and species indistinguishable by the targeted variable region(s). Within a region, if a cluster contained multiple species annotations, these annotations were merged and captured in the region's concatenation map or "cat map". All species annotations for models were replaced with the first alphanumeric species in the concatenation. When a species' annotation was present in multiple nearby clusters (difference in cluster numbers ≤ 2), all species' annotations within and between the clusters were concatenated. When species' annotations were present in distant clusters, sequences in the smaller clusters were removed. The resulting vagina-specific datasets are collectively referred to as vSpeciateDB.

Testing classification accuracy of known and novel sequences

To estimate the capability of models to classify novel sequences, ten-fold cross-validations were performed on each vSpeciateDB (pecan_cv5.pl, available on GitHub). All curated sequences for a targeted region were included in the training set. Each dataset was randomly spit into training (90% of sequences) and test (10%) sets. SpeciateIT models were built from the training set and training set evaluation were performed for construction of error thresholds. Subsequently, the test set was classified.

Comparing classifications for multiple classifiers

Sequences from the GTDB-SSU v214.1 datasets that were truncated to the V1-V3, V3-V4, and V4 regions and excluded from vSpeciateDB were used as independent query sets to compare classification between (1) SpeciateIT trained with vSpeciateDB, (2) RDP Naïve Bayesian Classifier stand-alone Bioconda version 2.13 (default settings), and the DADA2 implementation of RDP Naïve Bayesian Classifier trained with (3) SILVA v138.1 or (4) GTDB r86 reference sets. For the DADA2 implementations, both the assignTaxonomy and addSpecies functions were employed. Classifications were compared to the taxonomy of the GTDB-SSU v214.1 dataset to determine correctness.

Speed of classification

Random test sets of 10^1–10^7 sequences were produced from the reference sequences used to build models. Classification for each set was performed on 2021 Macbook Pro with an Apple M1 Max processor using 1 core. For RDP classification, the RDP Naïve Bayesian Classifier (stand-alone Bioconda version 2.13, default settings) was used (rdp_classifier classify -f allrank). For SpeciateIT classification, models and error thresholds for variable region-specific 16S rRNA gene sequences being classified were employed. Time of classification was measured using the time bash utility.

Results

Estimates of classification accuracy for novel sequences were obtained using tenfold cross validation. To ensure confidence in assignments, SpeciateIT imposes model-specific classification error thresholds: when the posterior probability of a query sequence does not exceed this threshold, the query sequence is classified as the next highest taxonomic level at which this threshold requirement is met. In the case of novel taxa, SpeciateIT is expected to assign higher-level classifications. In tenfold cross-validation testing, 98.7, 97.6, and 97.2% of sequences from "known" species (a species with at least 1 sequence present in the training dataset) were correctly assigned with > 90% of assignments made to the genus or species levels (Fig. 1A). For sequences from "novel" species (those with no sequences present in the training set), 60–70% were correctly assigned to their respective taxonomic categories, with the accuracy varying depending on the targeted region. This highlights the efficacy of SpeciateIT in accurately classifying bacterial taxa using higher order Markov chain models.

An essential aspect of SpeciateIT is its provision of posterior probabilities for query sequences. In the context of classification using Markov chain models, posterior probabilities represent the likelihood or confidence that a given sequence belongs to a particular category or class. These probabilities are calculated based on the observed sequence

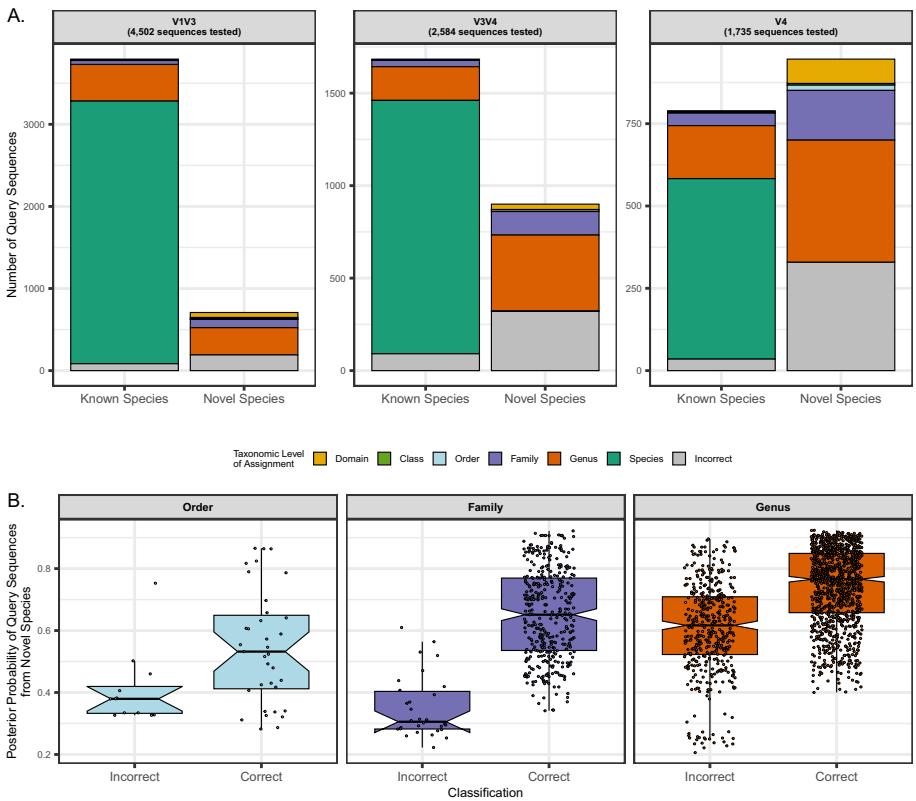

Fig. 1 **A** Ten-fold cross validation of the vSpeciateDB V1V3, V3V4, and V4 models demonstrated exceptional classification of sequences from "Known Species" with at least 1 sequence present in models. Most sequences from "Novel Species" were correctly classified at some taxonomic level. **B** The posterior probabilities of query sequences from "Novel Species" tended to be higher for correct classifications relative to incorrect classifications

data and the parameters of the Markov chain model. When a query sequence has a lower posterior probability, it suggests that the observed sequence data is less consistent with the model's parameters. This can indicate that the sequence deviates more from the typical patterns captured by the model, potentially suggesting a poorer match between the sequence and the model. However, it's important to note that a lower posterior probability does not necessarily mean that the classification result is incorrect or that the sequence is not related to the modeled categories. It simply suggests lower confidence in the classification result. In some cases, a sequence with a lower posterior probability may still be correctly classified, especially if the model captures only part of the variability present in the data. Regarding sequences from novel species (those absent from the training set), cross-validation results illustrate that the posterior probabilities from correct genus-, family-, or order-level assignments tend to be greater than incorrect classifications (Fig. 1B).

To compare classification of vaginal microbiota using SpeciateIT with vagina-specific vSpeciateDB to other popular classifiers and reference sets (RDP Naïve Bayesian Classifier stand-alone Bioconda version 2.13, default settings; DADA2 implementation of RDP Naïve Bayesian Classifier trained with vSpeciateDB, SILVA v138.1 and GTDB r86 reference sets), we classified independent sequences from GTDB (not included in the production of vSpeciateDB) truncated to each variable region and included those from the 100 most abundant species detected in the vaginal microbiota [22]. SpeciateIT with vSpeciateDB provided more species-level assignments than other methods including the DADA2 implementation of the RDP classifier which provided species level assignments, when possible (function: addSpecies) (Fig. 2).

The speed of SpeciateIT is incomparable because of its novel model tree-based approach which directs query sequence classification from the top of the tree (Root) to the branch or node of its final classification (Fig. S4). Classification speed was measured on a 2021 Macbook Pro with an Apple M1 Max processor and 64G RAM using each amplicon reference training set sampled to 10^1–10^7 sequences and processed on one core. We compared the speed of SpeciateIT classification to the RDP Naïve Bayesian Classifier (stand-alone Bioconda version 2.13, default settings). SpeciateIT classified 1 million sequences in 3, 2, and 1 min for the V1V3, V3V4, and V4 classifiers, respectively (Fig. 3). Speed is dependent on the number of models read for each classifier (the V4 classifier represents fewer species and therefore contains fewer models). Comparatively,

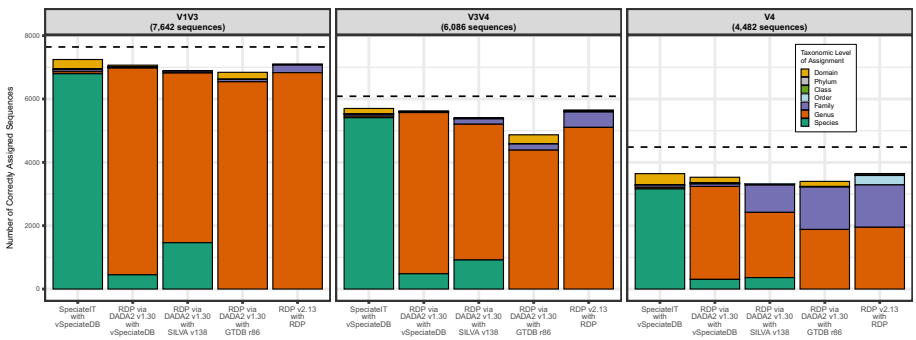

Fig. 2 SpeciateIT outperforms other classification methods by providing correct species-level assignments. Dashed lines indicate total number of sequences tested

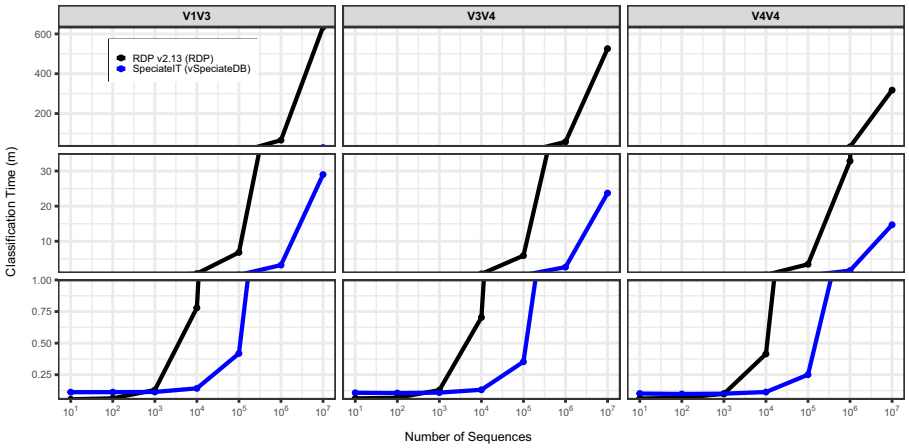

Fig. 3 SpeciateIT is faster than the RDP classifier when datasets are greater than 1000 sequences

the RDP Classifier classified 1 million V1-V3, V3-V4, and V4 sequences in 66, 57, and 32 min, respectively.

The performance of any classifier is entirely dependent on the quality of the sequence training set used to build it. Currently, SpeciateIT models have been built from full length 16S rRNA gene sequences curated from the Genome Taxonomy Database (GTDB) for the taxonomy-adjusted V1-V3, V3-V4, and V4 amplicon sequence regions for vaginal microbiota, and are publically available (https://github.com/ravel-lab/speciateIT). The full-length database comprises 2224 species, 497 genera, 77 families, 36 orders, 16 classes, and 14 phyla.

One recent change in the field of vaginal microbiota is the expansion of *Gardnerella vaginalis* to multiple species. Eleven species are represented in the genus *Bifidobacterium* in the GTDB SSU rRNA reference sequence set from which vSpeciateDB sequences originated. We chose to maintain the *Gardnerella* annotation for these species because of the vast clinical context surrounding *Gardnerella*. *G. vaginalis* C was not included in the final training sets because no reference sequences contained the V3 or V4 regions. *Gardnerella vaginalis A* and *Gardnerella vaginalis F* were distinct from other *Gardnerella* species in both the V2 and V4 regions (Fig. S5a). It was not possible to confidently distinguish other *Gardnerella* species at any targeted region. To maintain simplicity, one *Gardnerella* model ("*G. vaginalis*") represents GTDB species: *G. leopoldii, G. piotii, G. swidsinskii, G. vaginalis and G. vaginalis* A, B, C, D, E, F, and H combined. Of other prevalent species in the vaginal microbiota, *Lactobacillus iners*, *L. jensenii*, *L. mulieris,* and "*Ca.* Lachnocurva vaginae" were distinct in vSpeciateDB while *L. gasseri* and *L. paragasseri* were not distinguishable at any region and are referred to as only *L. gasseri*. Notably, *L. crispatus* and *L. acidophilus* were indistinguishable at the V4 region (Fig. S5b). Because *L. crispatus* is arguably more prevalent in the vaginal microbiota, these models are referred to as *L. crispatus.*

Lastly, the VAginaL community state typE Nearest CentroId classifier (VALENCIA) uses reference centroids representing microbiota compositions for each CST. The taxonomic annotations used in building the reference centroids are integral to correct CST classification. Because vSpeciateDB-based taxonomic assignments differ from those

used in the current version of VALENCIA reference centroids, we have produced reference centroids based on vSpeciateDB taxonomy and compatible with the VALENCIA algorithm for CST assignment (Fig. S6).

Conclusions

We anticipate vSpeciateDB will grow as more vaginal species are characterized, and more 16S rRNA gene variable region vSpeciateDB models will be produced. Importantly, taxonomy is adjusted in each V1-V3, V3-V4 and V4 database to reflect the loss of taxonomic information associated with sequence truncation, a known problem when using amplicon sequences [23]. Furthermore, the steps for vSpeciateDB curation are provided and can be used as a foundation upon which other environment-specific reference databases can be curated.

Abbreviation

OTUs Operational taxonomic unit

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05930-3.

Supplementary file 1

Acknowledgements

The authors acknowledge the contributions of the anonymous reviewers.

Author contributions

JR and PG conceived and developed the SpeciateIT algorithm. JH and PG developed the database curation methods. JH performed all tests and produced all figures. All authors contributed to the writing of the manuscript.

Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award numbers F32-AI136400 (JH) and K01-AI163413 (JH), U19-AI158930 (JR), U19-AI084044 (JR), and R01-AI116799 (JR, Co-I).

Availability of data and materials

The classification algorithm script, reference models, and VALENCIA reference version 2 centroids are available via https://github.com/Ravel-Laboratory/speciateIT.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

JR is co-founder of LUCA Biologics, a biotechnology company focusing on translating microbiome research into live biotherapeutics drugs for women's health. All other authors declare that they have no competing interests.

Received: 8 May 2024 Accepted: 16 September 2024
Published online: 27 September 2024

References

1. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol. 2005;71(3):1501–6.
2. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.

3. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Meth. 2010;7:335–6.
4. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Meth. 2016;13:581–3.
5. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv. 2016;081257.
6. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech XuZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems. 2017;2:e00191–116.
7. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high- throughput marker gene sequences. ISME J. 2014;9:968–79.
8. Nguyen N-P, Warnow T, Pop M, White B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. NPJ Biofilms Microbiomes. 2016;2:2567–8.
9. Darzi Y, Falony G, Vieira-Silva S, Raes J. Towards biome-specific analysis of meta-omics data. ISME J. 2016;10(5):1025–8.
10. Lobanov V, Gobet A, Joyce A. Ecosystem-specific microbiota and microbiome databases in the era of big data. Environ Microbiome. 2022;17(1):37.
11. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
12. Lan Y, Wang Q, Cole JR, Rosen GL. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PLoS ONE. 2012;7(3):e32491.
13. Dong CVQ. Evaluation of the RDP classifier accuracy using 16S rRNA gene variable regions. Metagenomics. 2012;1:104303.
14. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996–1004.
15. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50(D1):D20–6.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
17. Schmieder R, Lim YW, Rohwer F, Edwards R. TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinf. 2010;11:341.
18. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R Package Version. 2019;2:10–18129.
19. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
20. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e9490.
21. White JR, Navlakha S, Nagarajan N, Ghodsi MR, Kingsford C, Pop M. Alignment and clustering of phylogenetic markers–implications for microbial diversity studies. BMC Bioinf. 2010;11:152.
22. France M, Ma B, Gajer P, Brown S, Humphrys MS, Holm JB, Brotman RM, Ravel J. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. Microbiome. 2020;8(166):1–15.
23. Martinez-Porchas M, Villalpando-Canchola E, Vargas-Albores F. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. Heliyon. 2016;2(9):e00170.

Publisher's Note