TAG.ME: Taxonomic Assignment of Genetic Markers for Ecology

Douglas E. V. Pires[1], Francislon S. Oliveira[1,2], Felipe B. Correa[1,2], Daniel K. Morais[1] and Gabriel R. Fernandes[1*]

1 Instituto René Rachou - Fiocruz Minas, Belo Horizonte, Minas Gerais, Brazil; **2** Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Bioinformatica, Belo Horizonte, Minas Gerais, Brazil.

1. Abstract

Here we present TAG.ME, a novel alignment-independent and amplicon-specific method for taxonomic assignment based on genetic markers (http://gabrielrfernandes.github.io/tagme/). TAG.ME uses a multilevel supervised learning approach to efficiently assign taxonomy based on user-defined genetic marker genes. We applied our method to assess gut and soil sample diversity and showed it outperforms alternative approaches, identifying a larger proportion of species.

2. Main

The study of microbial community structure and variability has revolutionized our understanding of human health and disease, the discovery of novel species and their role in environments across the globe[1,2] .

Sequencing of amplified genetic markers, including the 16S ribosomal RNA gene, have been extensively used to characterize microbial community composition[3]. The cost-effectiveness of amplicon-based microbiome studies may explain its popularity and the establishment of specialized databases cataloging taxonomic information of marker genes, including the Ribosomal Database Project[4], GreenGenes[5] and SILVA[6] has been supporting the proposal and wide usage of such approaches.

The organization of sequences into Operational Taxonomic Units (OTUs) (*i.e.*, groups of sequences that correspond to a given level of dissimilarity, usually 3%) traditionally precedes the taxonomic classification[7]. However, the OTU approach ignores specific variation by summarizing them into arbitrary identity thresholds[8]. The Dada2[9] and Deblur[10] approaches consider Amplicon Sequence Variants (ASVs) as a representation of each organism in the sequenced microbiome. In this context, alignment-free methods are an interesting alternative for taxonomic classification of ASVs, preventing the introduction of biases from sequence identity thresholds. Machine learning methods, such as the RDP Classifier[11], are widely used for taxonomic assignment. For instance, Dada2 uses a Naïve Bayesian algorithm to classify the sequences up to genus level, although species identification relies on exact sequence matches against a reference database, which makes the procedure slow and insensible to fine-scale variations. Supervised learning techniques can leverage on the wealth of genetic marker sequencing information available on the literature, being capable of generalization (predicting unseen data) as well as providing scalable approaches. In this work, we propose TAG.ME, a scalable alignment-independent and amplicon-specific method for taxonomic assignment based on genetic markers. TAG.ME uses a multilevel supervised learning approach to assign taxonomy achieving not only a better assignment performance but also using considerably less computational resources. TAG.ME is available as an open-source R package (https://gabrielrfernandes.github.io/tagme/, Supplementary Software) that accompanies complementary analysis workflows and tutorials making TAG.ME fully integrated with other well-established pipelines, such Dada2's.

To assess TAG.ME's performance, we compare our approach with four well-established methods - Utax (From Usearch package)[12], RDP Naïve Bayesian classifier[11], QIIME[7] and Dada2[9] - on a large-scale test, using the RDP database sequences as assignment target. Figure 1 depicts the performance of each method on assigning the different taxonomic levels of the RDP database (release 11.5) from Domain to Species. We show TAG.ME performed as well as or better than alternative approaches for all levels. It is noteworthy that Dada2's performance on Species level, despite being slightly superior, occurred because the method finds exact matches in the reference database: the RDP dataset.

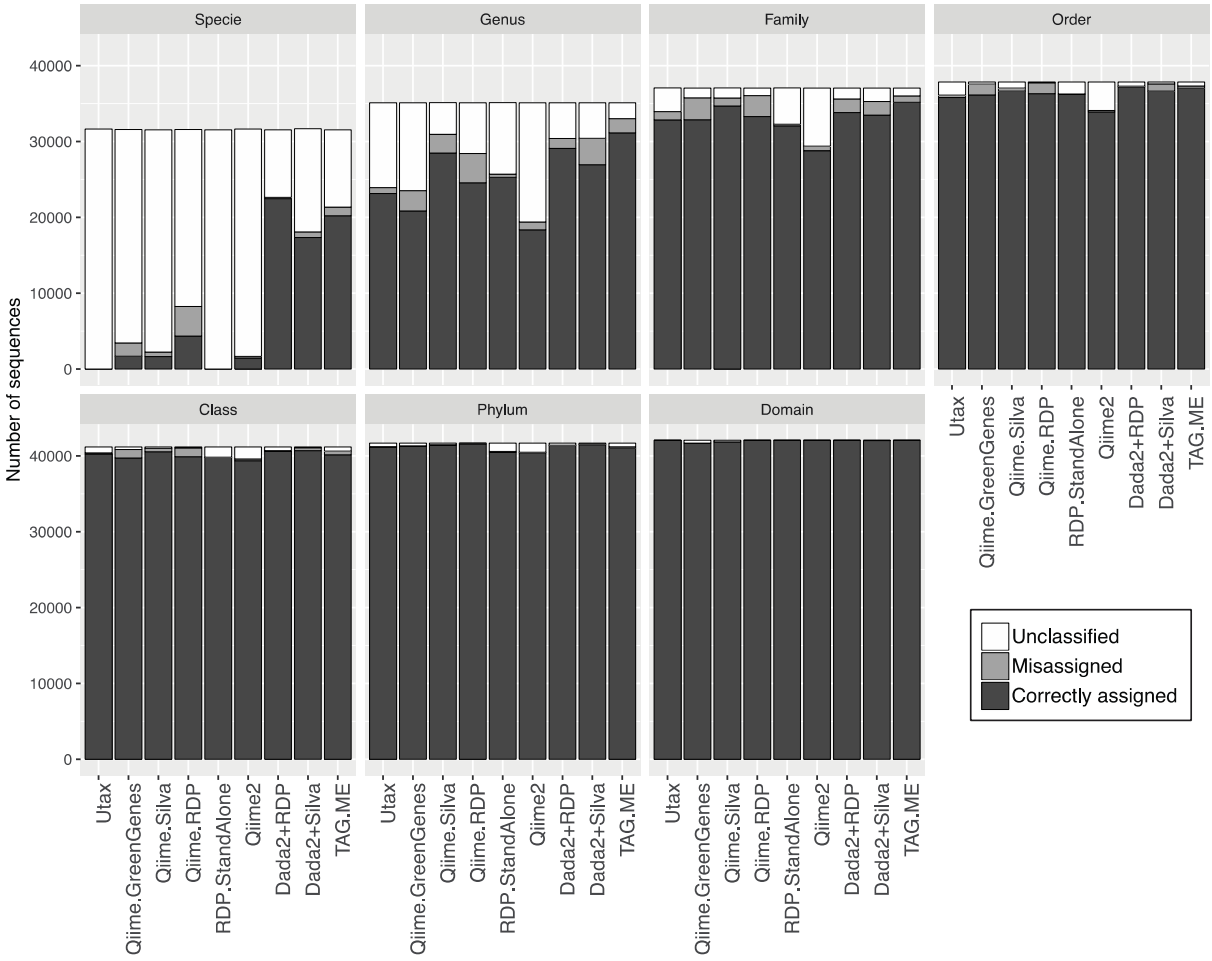


Figure 1. Comparison of TAG.ME taxonomic assignment with similar methods.

TAG.ME performance was then further compared with Dada2 on two real-world scenarios involving gut[13] and root samples[14]. In these cases, both Dada2+RDP and Dada2+Silva versions

were compared. Figure 2A depicts the percentage of assignments per taxonomic level for gut samples, while Figure 2B shows the results for the root samples. TAG.ME capabilities become increasingly evident as more specific taxonomic levels are inferred. Our method significantly outperforms Dada2 on Species level, while being superior or equivalent on other levels. For gut microbiome samples, TAG.ME assigned almost twice as many taxa in comparison with Dada2 on Species level. The performance difference for the root sample was even higher.

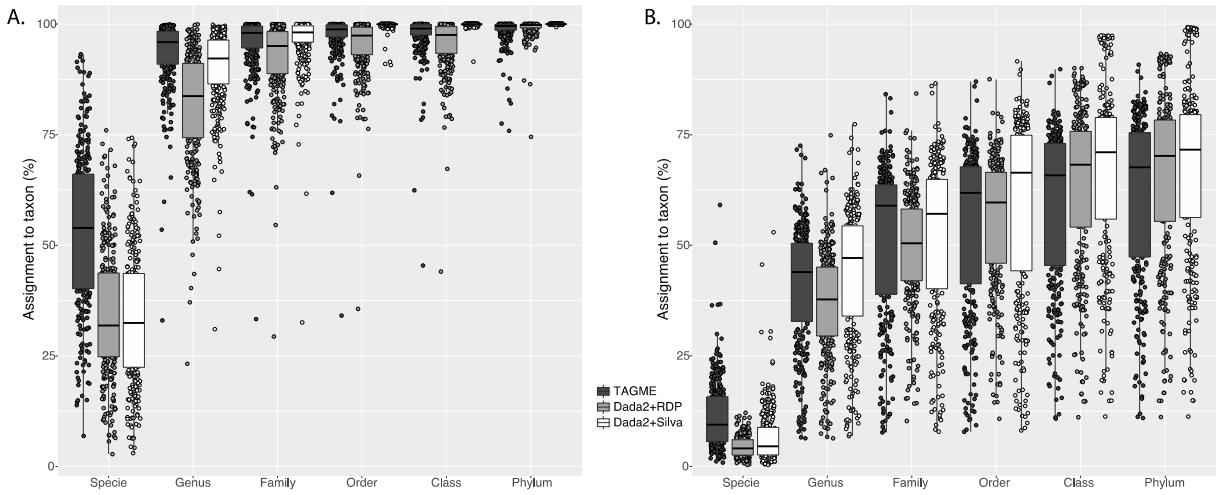


Figure 2. Comparative performance of TAG.ME and Dada2 on human gut (A) and root samples (B). TAG.ME manages to assign a considerably larger proportion of species while maintaining a compatible performance on other taxonomic levels.

It is interesting to notice, however, that 16S rRNA databases rely on the NCBI submitter's classification. We set out to evaluate in greater detail the entries TAG.ME misclassifies. Interestingly, a significant proportion of classification errors were either due to inconsistencies on the database or unclassified/uncharacterized entries, showing not only the robustness of our method but also its potential applicability to identify spurious or erroneous entries in public databases.

We show TAG.ME is an efficient and effective method for taxonomy assignment, capable of classifying well until species level, outperforming well-established methods. TAG.ME can be

executed in a low-spec laptop with as little as 8GB of RAM memory, outperforming Dada2 in both running time and memory usage.

We believe it will enable researchers to assess bacterial community structure and composition better, leading to novel insights regarding their role in nature and how to modulate these communities to our benefit.

3. Methods

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

4. Competing financial interest

The authors declare no competing financial interests.

5. Author Contributions

D.E.V.P designed the experiment, supervised the research and wrote the paper. F.S.O and F.B.C implemented de algorithm, performed the analysis. D.K.M performed the analysis. G.R.F designed the experiment, implemented de algorithm, performed the analysis and wrote the paper.

6. References

1.Gilbert, J. A., Jansson, J. K. & Knight, R. T*BMC biology* **12,** 69 (2014).

2.Cho, I. & Blaser, M. J. T*Nature Reviews Genetics* **13,** 260–270 (2012).

3.Kuczynski, J. *et al. Nature Reviews Genetics* **13,** 47–58 (2011).

4.Cole, J. R. *et al. Nucleic Acids Research* **42,** NaN–NaN (2014).

5.DeSantis, T. Z. *et al. Applied and environmental microbiology* **72,** 5069–5072 (2006).

6.Quast, C. *et al. Nucleic Acids Research* **41,** NaN–NaN (2013).

7.Caporaso, J. G. *et al. Nature Methods* **7,** 335–336 (2010).

8.Callahan, B. J., McMurdie, P. J. & Holmes, S. P. *The ISME journal* **11,** 2639–2643 (2017).

9.Callahan, B. J. *et al. Nature Methods* **13,** 581–583 (2016).

10.Amir, A. *et al. mSystems* **2,** e00191–16 (2017).

11.Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. *Applied and environmental microbiology* **73,** 5261–5267 (2007).

12.Edgar, R. C. *Bioinformatics* **26,** 2460–2461 (2010).

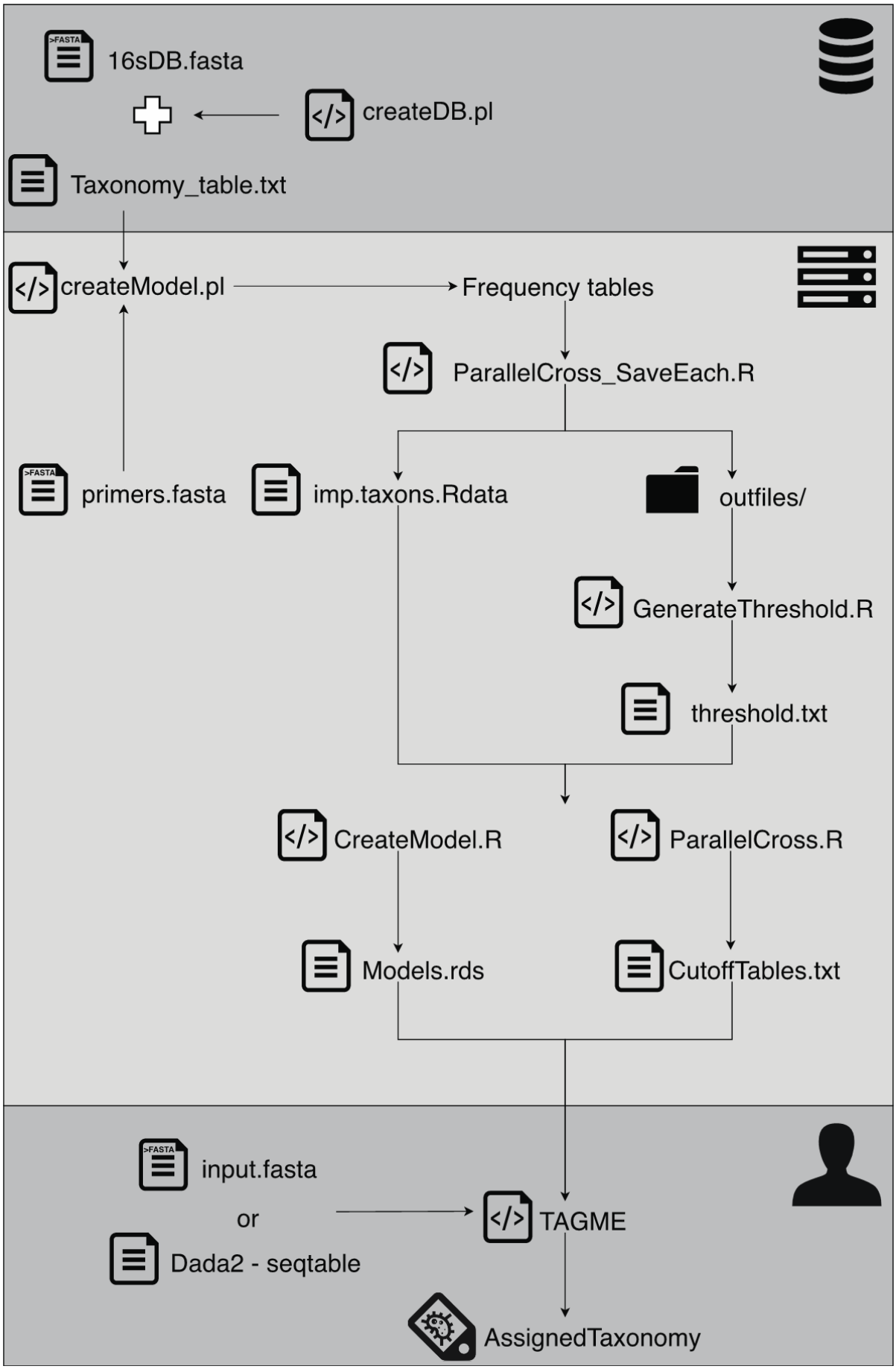
13.Franco-de-Moraes, A. C. *et al. Diabetology & metabolic syndrome* **9,** 62 (2017).

14.Castrillo, G. *et al. Nature* **543,** 513–518 (2017).

7. Online Methods

TAG.ME pipeline

Database preparation. TAG.ME is fully customizable, allowing users to tailor predictive models based on personalized databases. The available precomputed models were developed using the SILVA database, version 132. Figure S1 of Supplementary Materials depict the complete TAG.ME pipeline. Users can build their training databases from FASTA files containing the Marker Genes sequences of interest and a Taxonomic description per sequence. After database creation, amplicons are generated based on user-selected primers, followed by a dereplication[15] step and Last Common Ancestor (LCA) calculation to obtain a non-redundant set. K-mer frequencies are computed and used as evidence for training classification models. K=4 was the best performing word length and was therefore used in this work. A frequency table is calculated and used during supervised learning. More details on database preparation procedures are available at http://gabrielrfernandes.github.io/tagme/.


16sDB.fasta
createDB.pl
Taxonomy_table.txt
createModel.pl
Frequency tables
ParallelCross_SaveEach.R
primers.fasta
imp.taxons.Rdata
outfiles/
GenerateThreshold.R
threshold.txt
CreateModel.R
ParallelCross.R
Models.rds
CutoffTables.txt
input.fasta
or
Dada2 - seqtable
TAGME
AssignedTaxonomy

Supplementary Figure 1. TAG.ME workflow to create the predictive models

Supervised Learning. TAG.ME uses the Random Forest ensemble classification algorithm to train its amplicon-specific predictive models. Other models can be trained upon request via the website https://gabrielrfernandes.github.io/tagme, or locally via the package of scripts available for download. For each taxonomic level, a predictive model is trained, where each taxon represents a class. During training of each taxonomic level, 40% of classes are randomly filtered out of the training process. This composes a true negative set. The remaining set is further divided in training (60% of the remaining) and testing (40% of the remaining), making sure all classes have representatives in each of the sets. This procedure is repeated 20 times and average performance metrics are derived. The optimal number of features used to build the models and class representativity are also calculated.

Prediction assignment and sensitivity/specificity tradeoffs. As a multi-level classification system, TAG.ME employs a bottom-up procedure for taxonomy assignment as follows. Starting from the deepest level (Species), it assesses the class distribution generated by the Random Forest classifier and compares the ratio between the two top-ranked classes:

$$\text{Score} = \log2(P(\alpha)/P(\beta))\times P(\alpha),$$

Where, $P(\alpha)$ is the probability for the top-ranked class prediction, α, and $P(\beta)$ is the probability for the second-ranked class, β. The chosen value for the Score influences sensitivity and specificity tradeoffs (Supplementary Table 1), which can be set by the user. By default, TAG.ME taxonomic assignment is set to deliver a specificity of at least 80%. If the Score is higher than the threshold, a taxon is assigned. Otherwise, TAG.ME raises the taxonomy to the next taxonomic level and repeats the process.

Supplementary Table 1. Specificity and Sensibility of the genus model for the amplicon 515F-806R according to the Score. The highlighted row represents the specificity of 80%, resulting in a score threshold of 0.6 for classification.

Score	Sensitivity	Specificity	PPV	NPV
0.1	0.984190121	0.478892283	0.647279624	0.968919479
0.2	0.975631533	0.615909899	0.712889907	0.962764896
0.3	0.968872155	0.688126456	0.752922255	0.95751319
0.4	0.962632004	0.738157799	0.78332854	0.952578742
0.5	0.957010261	0.772579475	0.805662335	0.948029983
0.6	0.951183055	0.801418229	0.82535705	0.943306867
0.7	0.945568549	0.823850696	0.84135012	0.938727465
0.8	0.940311248	0.841457149	0.854332673	0.93445229
0.9	0.93458178	0.857628928	0.86662213	0.929799674
1	0.929271376	0.869908879	0.87617084	0.925466246
1.1	0.923407099	0.882199264	0.885968938	0.920764379
1.2	0.917888018	0.891610508	0.893598425	0.916311407
1.3	0.911905677	0.900797096	0.901195948	0.911545814
1.4	0.905877168	0.908937326	0.908050057	0.906783251

Benchmark. The rRNA 16S FASTA sequences and taxonomic lineage for each entry were downloaded from the Ribosomal Database Project website[4]. The sequences were filtered to remove entries without the full lineage description, keeping only those with information up to Species level. The dataset was used as the reference for the amplicon generation, dereplication, and LCA identification, as described in Database preparation. The non-redundant amplicon sequences were submitted to the taxonomic classification using popular pipelines with their default parameters: Utax[12], Qiime with three different databases[7], RDP Classifier[11], Qiime2, and Dada2[9]. The unclassified or inaccurate classifications were excluded from the reference dataset before comparison. Correct assignments were considered when the prediction exactly matches the reference classification, and were considered a misassignments otherwise. The Unclassified sequences represent entries that have a precise taxonomic description in the reference set, but could not be identified by the classification tool. The benchmark dataset was used to evaluate the running time as well. While Dada2 took 1228 minutes to classify up to species level, TAG.ME took 6.5 minutes in a 40-core CPU with 64Gb of RAM.

Test data sets. Dada2 classification into species-level relies on exact sequence matches against the database. The benchmark dataset was obtained from RDP, which is the same database used for Dada2 Species assignment, leading to a better accuracy than TAG.ME in that taxon. Two published

datasets were used to evaluate the performance of sequenced data. The dataset available under the accession PRJEB19103[13] represents the human gut environment, and accession PRJEB15671[14] represents the root ecosystem. The amplicon sequence variants (ASVs) were identified using the Dada2 workflow. Each ASV was classified by Dada2 using the training set from RDP and SILVA. TAG.ME used the 515F-806R model (V4 region) to perform the taxonomic assignment.

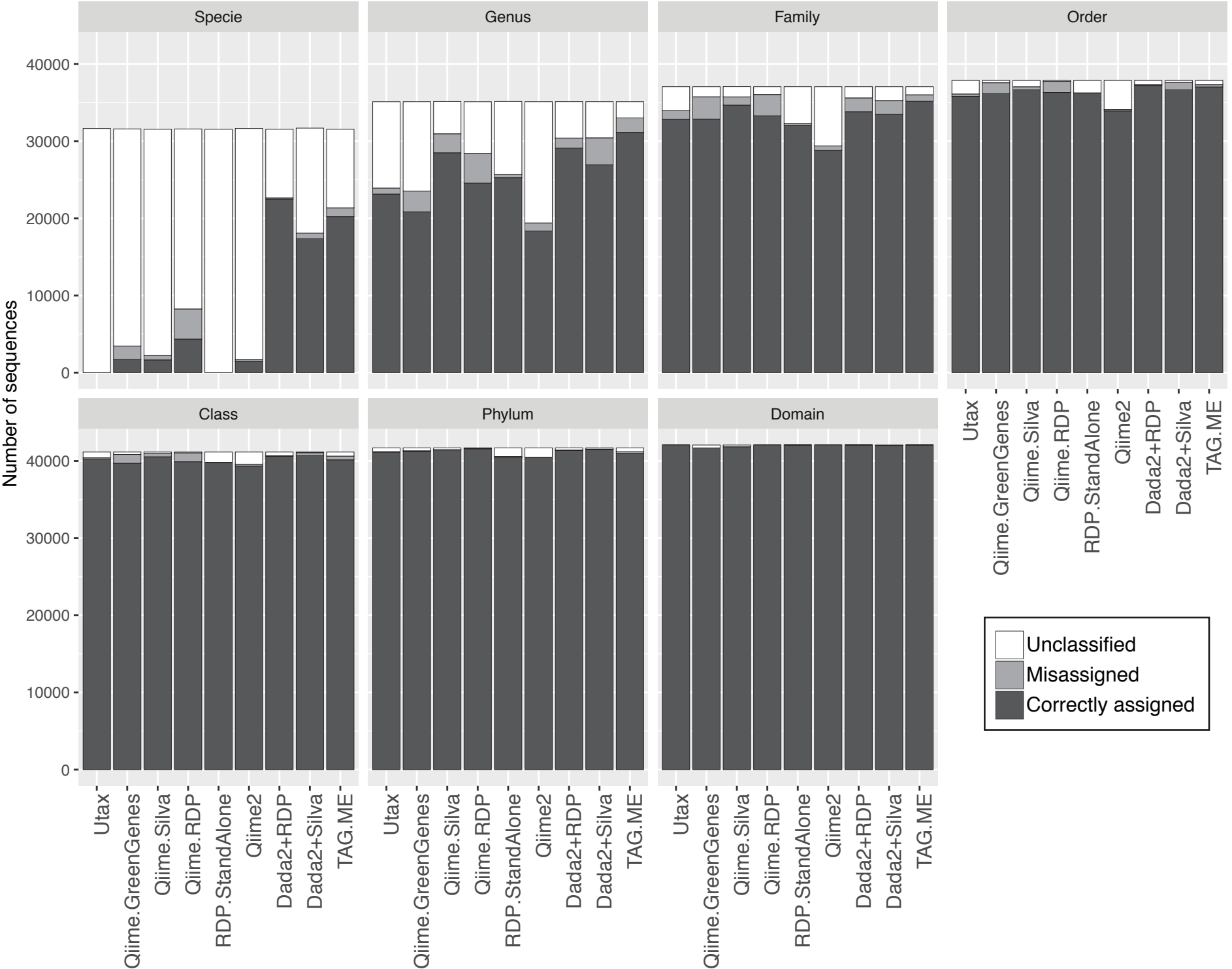
Code availability. TAG.ME is freely available as a R package via GitHub (https://gabrielrfernandes.github.io/tagme). Tutorials, as well as a help guides, were developed and are available via a user-friendly web interface. TAG.ME is fully integrated with other well established pipelines, such Dada2's and Phyloseq[16].

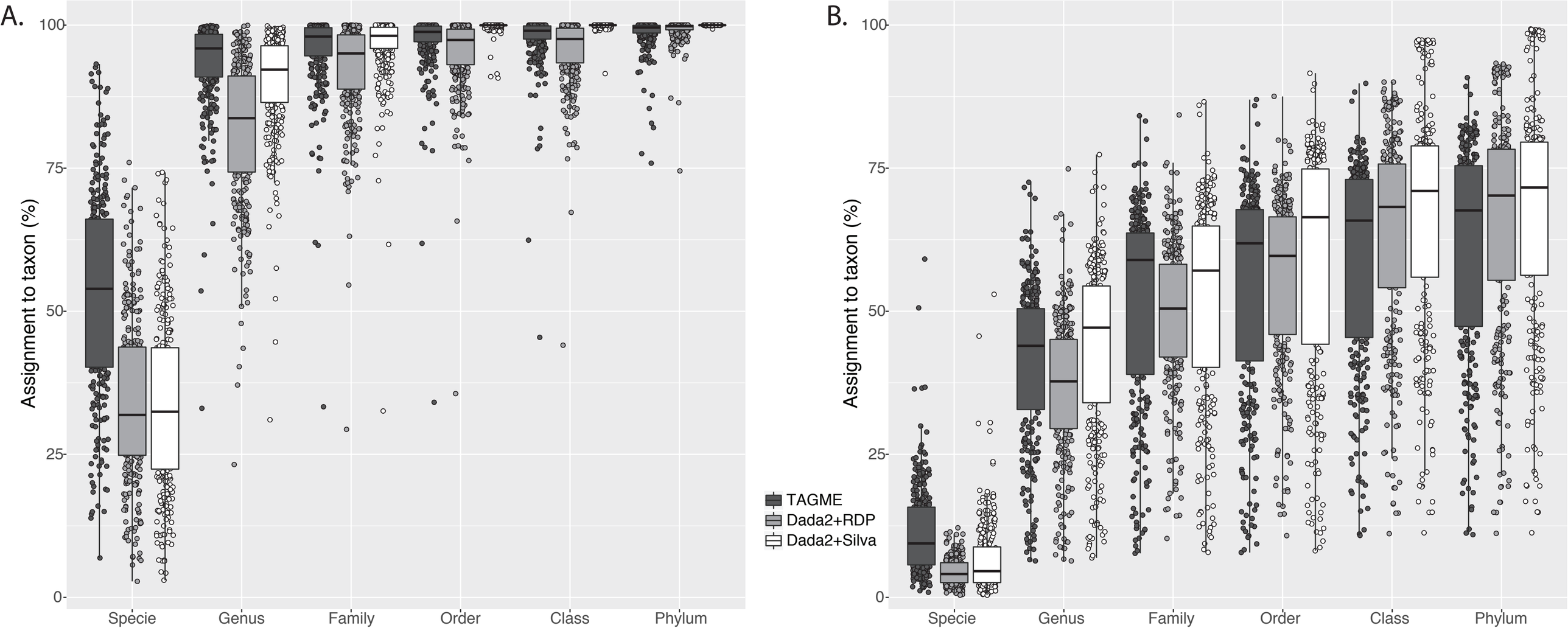
Limitations. Since TAG.ME builds amplicon specific models, it is required that the input sequences cover the whole amplicon generated by the pair of primers. Due to that, it is highly suggested that the data shall be provided by Illumina sequencing, since it has fixed number of cycles reflecting on exact sequence length. The usage of this approach is a trend, and it is a requirement also for Dada2 analysis, further integrating the two workflows.

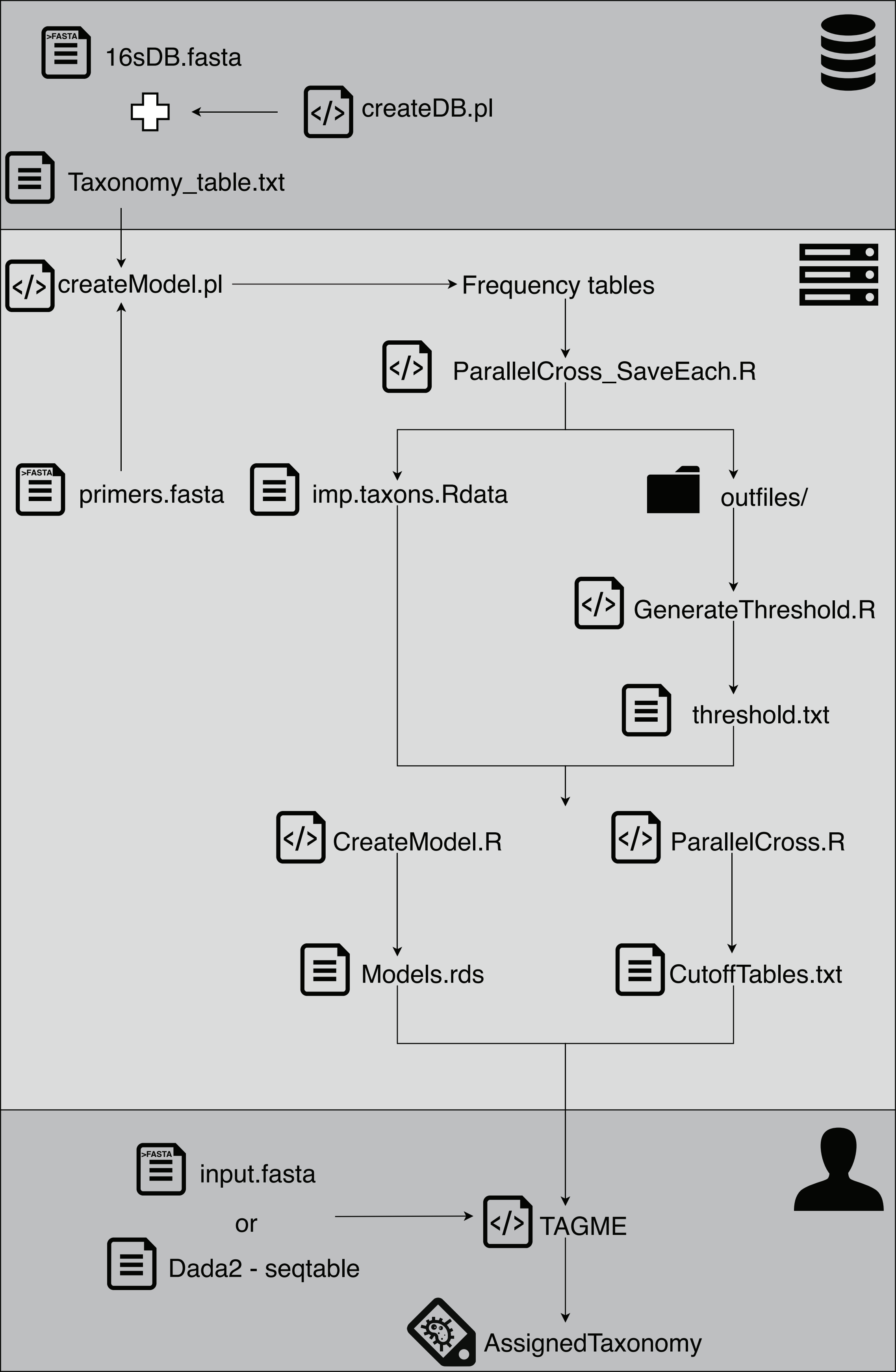
References.

15.Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. *PeerJ* **4,** e2584 (2016).

16.McMurdie, P. J. & Holmes, S. *PloS one* **8,** e61217 (2013).


Specie
Genus
Family
Order
Class
Phylum
Domain
Number of sequences
0
10000
20000
30000
40000
Utax
Qiime.GreenGenes
Qiime.Silva
Qiime.RDP
RDP.StandAlone
Qiime2
Dada2+RDP
Dada2+Silva
TAG.ME
Unclassified
Misassigned
Correctly assigned


A.
B.
Assignment to taxon (%)
100
75
50
25
0
Specie
Genus
Family
Order
Class
Phylum
TAGME
Dada2+RDP
Dada2+Silva


16sDB.fasta
createDB.pl
Taxonomy_table.txt
createModel.pl
Frequency tables
ParallelCross_SaveEach.R
primers.fasta
imp.taxons.Rdata
outfiles/
GenerateThreshold.R
threshold.txt
CreateModel.R
ParallelCross.R
Models.rds
CutoffTables.txt
input.fasta
or
Dada2 - seqtable
TAGME
AssignedTaxonomy