Received: 10 June 2019 | Revised: 8 October 2019 | Accepted: 4 November 2019

DOI: 10.1111/1755-0998.13116

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

PIME: A package for discovery of novel differences among microbial communities

Luiz Fernando W. Roesch[1] | Priscila T. Dobbler[1] | Victor S. Pylro[2] | Bryan Kolaczkowski[3] | Jennifer C. Drew[3] | Eric W. Triplett[3]

[1]Interdisciplinary Research Center on Biotechnology-CIP-Biotec, Universidade Federal do Pampa, São Gabriel, Brazil

[2]Microbial Ecology and Bioinformatics Laboratory, Department of Biology, Universidade Federal de Lavras, Lavras, Brazil

[3]Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL, USA

Correspondence
Eric W. Triplett, Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL, USA.
Email: EWT@ufl.edu

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; NIH, Grant/Award Number: 5R21AI120195-02; JDRF, Grant/Award Number: 1-INO-2018-637-A-N

Abstract

The data used for profiling microbial communities is usually sparse with some microbes having high abundance in a few samples and being nearly absent in others. However, current bioinformatics tools able to deal with this sparsity are lacking. PIME (Prevalence Interval for Microbiome Evaluation) was designed to remove those taxa that may be high in relative abundance in just a few samples but have a low prevalence overall. The reliability and robustness of PIME were compared against existing methods and tested using 16S rRNA independent data sets. PIME filters microbial taxa not shared in a per treatment prevalence interval started at 5% prevalence with increasing increments of 5% at each filtering step. For each prevalence interval, hundreds of decision trees were calculated to predict the likelihood of detecting differences in treatments. The best prevalence-filtered data set was user-selected by choosing the prevalence interval that kept a large portion of the 16S rRNA sequences in the data set while also showing the lowest error rate. To obtain the likelihood of introducing type I error while building prevalence-filtered data sets, an error detection step based was also included. A PIME reanalysis of published data sets uncovered other expected microbial associations than previously reported, which may be masked when only relative abundance was considered.

KEYWORDS

16S-rRNA, core microbial taxa, microbial biomarkers, microbial prevalence, next generation sequencing, taxa filtering

1 | INTRODUCTION

Sequencing of amplified genetic markers (amplicon survey), e.g., the 16S rRNA gene, is traditionally used for testing hypotheses on microbial community composition. The major challenge for using the data obtained by these surveys is their interpretation for the discovery of the drivers of microbial diversity. Excluding microbiomes from simple ecosystems (e.g., habitats with extreme temperature or pH), amplicon surveys usually identify a large number of taxa (also called operational taxonomy units (OTUs) or amplicon sequence variants (ASVs) not shared among all samples (also called low prevalent taxa) (Sze & Schloss, 2016). Often prefiltering steps in the data analysis eliminate many of these taxa with low prevalence. Those steps include, but are not limited to, the exclusion of sequences found only once in a sample. These are the so-called singletons (Edgar, 2013; Edgar & Flyvbjerg, 2015). According to Tedersoo et al. (2010), singletons are artefactual and account for the greatest source of bias in next generation sequencing. Also, very low abundant reads might be the result of a low level of contaminants from commercial kits (Eisenhofer et al., 2019; Salter et al., 2014).

Another prefiltering approach involves the exclusion of microbial taxa of low prevalence across all samples. The prevalence of

microbes in the human microbiome is characterized by variable distribution patterns (Kraal, Abubucker, Kota, Fischbach, & Mitreva, 2014) with prominent abundance of some strains in some subjects while nearly absent in others. While the presence of microbes with low prevalence across all samples could be the focus of research for future experimental study (Kraal et al., 2014), the identification of microbial taxa present in the majority of the subjects, also known as the core microbiome, has been one of the primary goals of the Human Microbiome Project (Human Microbiome Project Consortium, 2012; Huse, Ye, Zhou, & Fodor, 2012). The microbial core can be used as standard to identify significant variations that might be associated with disease states or other treatments.

Many tools such as PHYLOSEQ (McMurdie & Holmes, 2013), QIIME (Caporaso et al., 2010), UPARSE (Edgar, 2013), MG-RAST (Meyer et al., 2008), MOTHUR (Schloss et al., 2009), and MICROBIOMEANALYST (Dhariwal et al., 2017) have been developed to contrast experimental factors in microbiome studies. The choice of a given analysis package is usually based on the user's questions of interest, level of experience in bioinformatics, and on the available resources at the user's host institution (Pollock, Glendinning, Wisedchanwet, & Watson, 2018). Nevertheless, most approaches embedded in these packages rarely consider microbial prevalence within treatments.

Here, we propose a new workflow based on the core microbiome concept that is designed to identify and remove the within group variation found in amplicon surveys (16S rRNA data sets) by capturing only biological differences at high sample prevalence levels. In an experiment comparing two treatments (e.g., healthy vs. diseased subjects) one core for each treatment will be calculated and relevant microbial taxa responsible for differences within microbial cores will be detected. That is, we are asking the question of the extent to which core microbiomes differ. To implement this concept, we developed an R package called PIME (Prevalence Interval for Microbiome Evaluation). PIME is a tool specifically designed to work with data sets with high variation among samples. PIME removes low abundance taxa in each treatment or group keeping only those taxa that are shared at some level of prevalence. It calculates prevalence levels in 5% intervals from 5% to 95%. For each prevalence level a list of the most relevant taxa responsible for differences between or among treatments is provided. We also implemented an error detection step based on randomizations which calculates the likelihood of false predictions (i.e., existence of distinct groups when there are not) throughout the data set filtration process.

2 | MATERIALS AND METHODS

2.1 | Program description

2.1.1 | Bioinformatics workflow

The bioinformatics workflow described here is embedded in an R package called PIME (Prevalence Intervals for Microbiome Evaluation) available at: https://github.com/microEcology/PIME. PIME identifies statistically significant bacterial community differences considering the proportion of samples hosting a specific microbial community in a given time period. For the purpose of this work, prevalence was defined as the proportion of samples in a specific group (e.g., treatment or any other factor the user wants to compare) that share taxa, irrespective of the relative abundance, at the time of sampling. For example, a prevalence cutoff of 50% means that the taxa selected at this prevalence interval are found in 50% of the samples. PIME's strategy is based on four fundamental steps depicted in Figure 1 that we describe below.

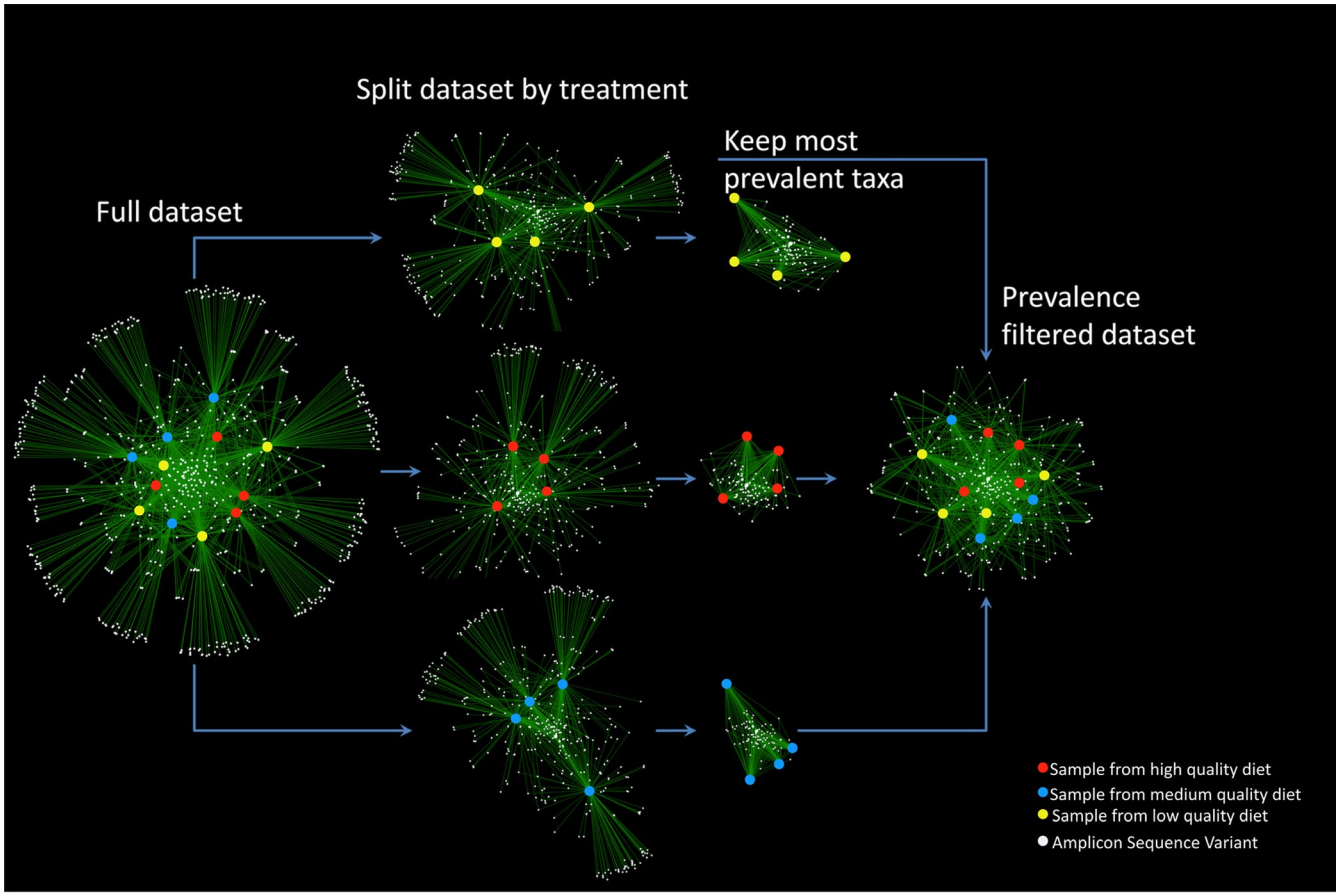
Prediction of differences in full data set

PIME takes a PHYLOSEQ object (McMurdie & Holmes, 2013) as input. PHYLOSEQ enables handling many data formats. PIME then builds hundreds of randomized decision trees, where each gives a vote for the prediction of the target variable, using a supervised nonparametric machine learning algorithm and combines them into a single model to predict the likelihood of detecting any user defined treatments or variables as a source of sample variation (Breiman, 2001). The model performance is indicated by the out-of-bag (OOB) estimate of the error rate calculated by training the algorithm on a subset of samples and tested on the remaining samples. Values can vary between 0 and 1, where 0 and 1 indicate the model has 100% or 0% accuracy, respectively. This overall measurement of accuracy can be interpreted as an estimate of error obtained when the model is applied to new observations. Higher OOB error indicates low accuracy of the model in predicting differences among the categorical variables tested.

This first PIME step is implemented in a function called *pime.oob.error*. This function is run using the data set without any filtering proposed by PIME. After obtaining the OOB error rate, the user decides whether PIME is adequate for the data set. For example, an OOB error near zero indicates the prevalence filtering with PIME is not necessary, as the model accuracy is already reasonably good. On the other hand, if OOB error rate is greater than zero, filtering the data set using PIME might improve the model accuracy. In this case only, the user can proceed and execute the function *pime.split.by.variable*. This step is defined below.

Split the data set by predictor variable and compute prevalence intervals

The full data set is split according to the tested categorical variables (e.g., treatment and/or any other factor) defined by the user in the metadata file. Each variable will be used to define data subsets. Those per variable subsets are filtered using different prevalence levels from 5% to 95% with increments of 5% for each level (see Figure 1 for a simplified schema illustrating this filtering step). Prevalence levels (usually high prevalence levels e.g., 90%) where samples have zero sequences are not calculated. After removal of taxa that do not match the prevalence criteria, the subsets are merged to compose a new filtered data set (one per prevalence interval) for subsequent downstream analysis. This step is implemented in these two functions: *pime.split.by.variable* and *pime.prevalence*. The *pime.split.*



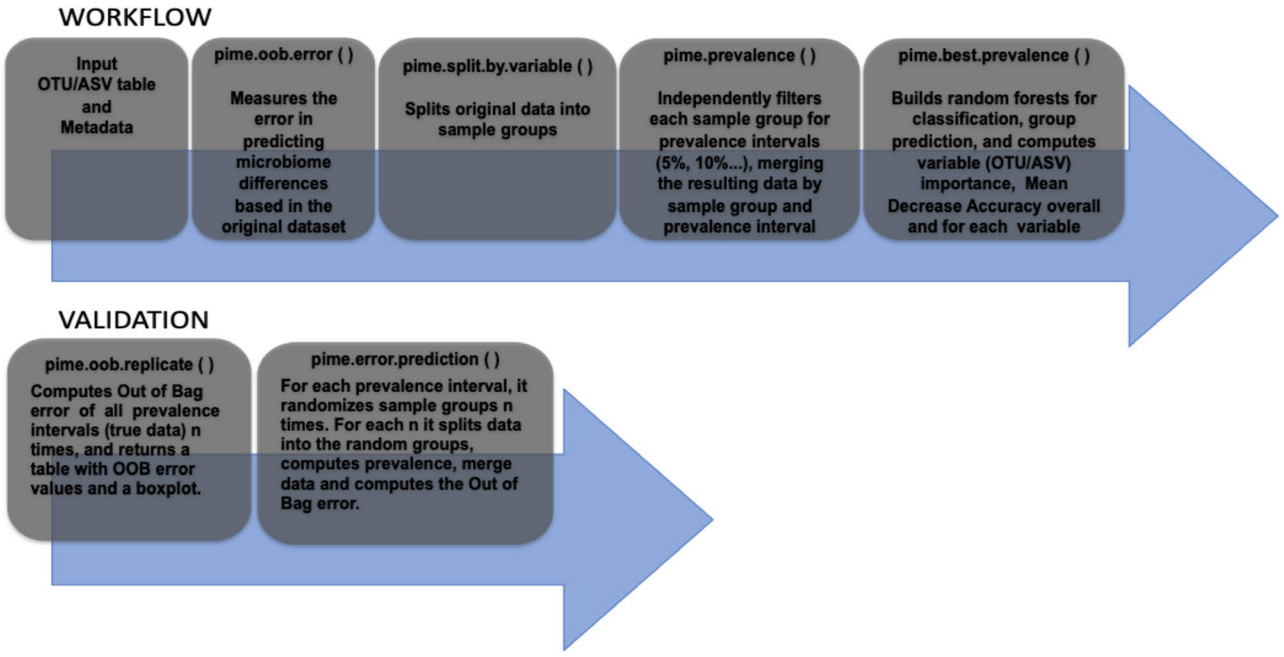


FIGURE 1 Empirical representation of steps used in PIME. Top panel. Bipartite network illustrating PIME method with a subset of 12 saliva microbiome samples. Each sample (red, yellow and blue circles) is connected to an ASV (white circles) through edges (green). ASVs observed in more than one sample are connected by at least two edges and are displayed at the center of the network. ASVs present in only one sample are connected by a single edge and are displayed at the border of the network. The first step applied by PIME is to split the full data set according to the treatments defined by the user. Within this example red, yellow and blue circles depict three different treatments. At each of the three new groups the low prevalent ASVs are removed. Finally, the subsets are merged to compose a new filtered data set used in the downstream analysis. Bottom panel. Step-by-step representation of PIME's workflow and validation. [Colour figure can be viewed at wileyonlinelibrary.com]

by.variable function uses the original data set as input and its output is used as input for *pime.prevalence*. The function *pime.prevalence* keeps, for each treatment group, every OTU/ASV according to the following equation:

$$N_0/N_s > P_i == \text{True}$$

where N_0 is the number of OTUs/ASVs counts with Sum > 0, N_s is the number of samples and P_i is the prevalence interval P_i = 0.05, ..., P_{max} = 0.95.

Computation of OOB error on each prevalence interval and importance of each taxa in the differentiation of microbial communities

Next, random forest analyses (Breiman, 2001) are used to determine the level of prevalence that provides the best model to predict differences in the communities, while still including as many taxa as possible in the analysis. After prevalence filtering, performed according to the equation above, the OOB error rate and the number of remaining taxa and sequences are calculated for each prevalence level. The results are provided in a table that allows the user to determine the optimal prevalence interval with high accuracy. This step is implemented in a function called: *pime.best.prevalence*. Within the same function, the contribution of each taxa to the mean decrease in classification accuracy is calculated using the same random forests algorithm. High values of mean decrease accuracy indicate the importance of taxa to differentiate two or more microbial communities. The user can access the importance of taxa in each of the prevalence intervals.

Validation

To obtain the likelihood of producing a type I error where PIME predicts the presence of distinct groups where no groups exist, an error detection step is included. Consider the scenario in which the null hypothesis of "no difference between groups" is false. If we randomly shuffle the labels that identify the sample groups and run the test again, the expected outcome is that the randomized data set will have a small chance to present distinct groups. Running the test multiple times with the random data set is expected to produce a high OOB error rate in most cases. This error detection test is implemented in these two functions: *pime.error.prediction* and *pime.oob.replicate*. The first function randomizes the samples labels into arbitrary groupings using 100 random permutations. For each randomized prevalence filtered data set, the OOB error rate is calculated to determine whether differences in the original groups occur by chance. The second function performs the random forest analyses and computes the OOB error for 100 replications in each prevalence interval without randomizing the sample labels. The biological difference among samples is expected to be greater than the differences generated randomly. Thus, the greatest fraction of randomizations should generate high error rates. On the other hand, no improvement in accuracy is expected within the randomized data set.

2.2 | Empirical validation

The PIME workflow was compared against other existing filtering methods and by using empirical tests with 16S rRNA data sets. The performance of PIME was compared against filtering methods based on overall prevalence, low abundance, and low variance. Also, four 16S rRNA data sets were analyzed using PIME to illustrate its usefulness. These included an assessment of: (a) the association between diet and saliva microbiome composition (unpublished original research); (b) the gut microbiome in subjects at high genetic risk for type 1 diabetes (Davis-Richardson et al., 2014); (c) the vaginal microbiome in pregnant women randomized to receive milk with or without probiotic bacterial strains (Avershina et al., 2017); and (d) the saliva microbome compared to the microbiome of the left antecubital fossa of healthy individuals (Human Microbiome Project Consortium, 2012).

2.2.1 | Comparison with other existing filtering methods

Comparisons were performed using a data set composed by 16S rRNA sequences from microbes extracted from saliva of 125 undergraduate and graduate students from the University of Florida (accessible through BioProject ID PRJNA504439). The following filtering tests were performed: (a) filtering the data set such that the taxa kept in the data set must be present in at least 20% of the subjects; (b) filtering the data set by abundance to include only those taxa with at least five sequences; and (c) filtering by low variance such that all taxa in the data set have variance higher than 20%. Filtered data sets were compared against the prevalence interval of 65% as calculated by PIME as the best prevalence interval where the OOB error was zero. A record of this analysis containing a step-by-step R-code and results is provided in Supporting Information S1.

2.3 | Performance evaluation with 16S rRNA data sets

A novel and four published data sets were analyzed with PIME. These data sets covered a broad range of habitats including human and environmental samples. These are used to show that PIME does give predicted results in those cases where we expect to see no differences in the treatment, such as in the studies of Avershina et al. (2017) and Davis-Richardson et al. (2014) and in those cases where we expect large differences between the treatments, such as the saliva microbiome described in this paper and the comparison between saliva and the left antecubital fossa from the human microbiome project (Human Microbiome Project Consortium, 2012). The novel data set used in this work comprised of 16S rRNA gene sequences from saliva samples obtained from 125 undergraduate and graduate students from the University of Florida. The study

assessed the subject's diet as a factor influencing the saliva microbiome. This study was approved by the University of Florida's Institutional Review Board and assigned number IRB201602134. Approximately 224 undergraduate and graduate students taking three courses were invited to anonymously participate in this study as volunteers. A study coordinator was chosen to collect samples and code the samples so that those who did the analysis were unaware of the identity of the volunteers. To assess the diet, the subjects also completed the KIDMED survey (Serra-Majem et al., 2004). The sampling collection, DNA extraction and library preparation are described below.

2.3.1 | Sampling collection, DNA extraction and library preparation

Of the 224 students invited, 125 volunteers obtained the saliva sample collection and provided 2 ml of saliva. The samples were taken from each subject using the GeneFiX Saliva DNA Collection device. The collection kit allows immediate stabilization of the DNA. Total DNA was extracted using the GeneFix Saliva-prep-2 kit (Cell Projects Ltd) following the manufacturer's protocol. DNA samples were stored at −20°C until use.

To assess the diet, the subjects also completed the KIDMED survey (Serra-Majem et al., 2004). The KIDMED Index is based on a series of 16 questions which measures the degree to which a subject adheres to the Mediterranean diet. The KIDMED index has been validated with nutritional data (Serra-Majem, Ribas, García, Pérez-Rodrigo, & Aranceta, 2003) and was much simpler to implement than a diet diary or a serum-based nutrition analysis. Participant's age and gender were also obtained.

The 16S rRNA library preparation as well as the PCR reactions, primers and thermocycling conditions were performed as described previously (Davis-Richardson et al., 2014) and sequenced with Illumina MiSeq: 2 × 300 cycles run. The raw fastq files were used to build a table of exact amplicon sequence variants (ASVs) with DADA2 version 1.8 (Callahan et al., 2016). Taxonomy was assigned to each ASV using the SILVA ribosomal RNA gene database version v132 (Quast et al., 2012). A detailed R script containing the code used to generate the ASV table is provided in the Supporting Information S2. Downstream analyses were carried out after the normalization of the number of sequences in all samples as recommended by Lemos, Fulthorpe, Triplett, and Roesch (2011). The rarefied data set comprised of 24,900 sequences per sample.

2.3.2 | Description of the previously published data sets

The first previously published data set used here was described by Davis-Richardson et al. (2014) and comprised of partial 16S rDNA sequences from faecal samples of 76 subjects born between 1996 and 2007 at the Turku University Hospital in southwestern Finland. All subjects were at high genetic risk for type 1 diabetes. The cohort was retroactively selected to create an age-matched genotype-controlled set of subjects for the investigation of the microbiome as an environmental factor influencing the development of type 1 diabetes. The raw Fastq files were obtained and sequences were processed using DADA2 version 1.8 (Callahan et al., 2016), as described above. Cases were defined as subjects who developed at least two persistent islet cell autoantibody (ICA), IAA, GADA, or IA-2A. Controls were defined as subjects with no detectable islet autoantibodies. Samples from subjects older than one year and post seroconversion were removed.

The second published data set used here has been previously described by Avershina et al. (2017). This data set is comprised of amplified and sequenced 16S rRNA genes from vaginal swab samples collected from a cohort of 256 pregnant women. These subjects were randomized to receive a daily dose of fermented milk containing probiotic bacterial strains, or milk without probiotics. The corresponding author kindly provided an OTU table with 3,000 sequences per sample and the accompanying metadata. This table was used in all downstream bioinformatics and statistical analysis. Only those samples collected at the 36th week of gestation were used in these analyses.

The third previously published data set comprised of 16S rRNA gene sequences from the V1–V3 hypervariable region downloaded from the NIH Human Microbiome Project (https://www.hmpdacc.org/HMQCP/#data). The final OTU table processed by QIIME (Caporaso et al., 2010) using an OTU-clustering strategy and accompanying metadata were obtained and loaded into the R environment. After removing singletons, only saliva and left antecubital fossa samples were kept. The filtered data set comprised of 113 saliva samples and 59 left antecubital fossa samples. All were rarefied to 2,000 sequences per sample.

The fourth published data set comprised of 16S rRNA sequences from soils in a well controlled microcosm system designed to investigate the individual and interactive effects of moisture and temperature (Lupatini et al., 2019). Specifically, we compared three moisture regimes at 10°C using only DNA samples rarefied at 7,100 sequences. A record of all statistical analyses comparing the data sets with and without using PIME including the R-code are included as Supporting Information S3.

3 | RESULTS

3.1 | Performance of PIME compared against other filtering methods

The performance of PIME was compared with other filtering methods (Figures 2 and 3). After quality filtering the saliva data set, a total of 4,981,638 high-quality sequences, 400 bp long, were obtained from all subjects. An average 44,258 (*SD* = 27,743) sequences per sample were obtained. The data set was rarefied to 24,900 sequences per sample in all analyses commensurate with the lowest number of

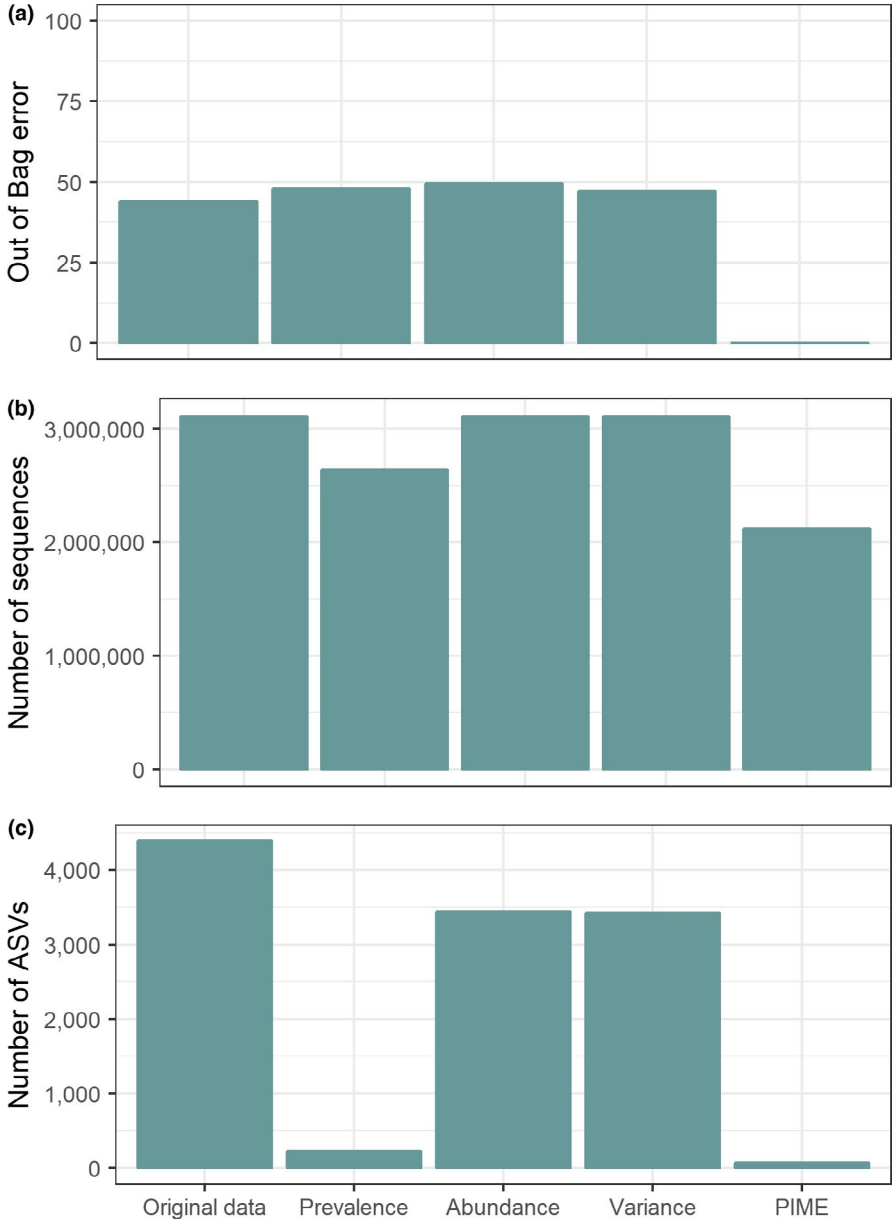


FIGURE 2 Performance of PIME compared to other filtering methods. (a) Out of Bag error rate (OOB error rate); (b) total number of sequences; (c) Total number of ASVs. Prevalence = filter by overall taxa prevalence in at least 20% of the subjects; Abundance = filter by abundance of at least five sequences; Variance = filter by variance higher than 20%. PIME = filter by prevalence interval of 65%. Data was generated by using the saliva data set and the step-by-step analysis can be found into the Supporting Information 1. [Colour figure can be viewed at wileyonlinelibrary.com]

sequences found in any one sample. The coverage of Good (1953) ranged from 0.97 to 1.00 indicating this number of sequences was sufficient to accurately reflect the microbial diversity in these samples given the low complexity of saliva samples. The optimal prevalence interval calculated by PIME was 65%. This prevalence interval was used to compare the performance of PIME against the other filtering methods. The original data set, without any filtering, presented 4,555 ASVs and a total of 3,112,500 sequences after rarefaction. Both filtering methods, prevalence overall and PIME, excluded the highest proportion of ASVs and sequences while filtering by abundance or variance excluded only 22% of ASVs and kept 99.9% of the sequences. Nevertheless, the overall prevalence kept 84% of the sequences while PIME kept 68% of the total number of sequences. Without using the PIME filtering the OOB error obtained while attempting to classify the salivary microbiome according to the three diet categories was 44%. This shows that the overall prevalence without using PIME model had low accuracy in predicting diet according to the microbiota. However, the PIME model had an OOB error of 0% (accuracy of 100%). The analysis of the taxonomic composition at phylum level after filtering the saliva data set with PIME and other filtering methods are presented in Figure 3. PIME did not skew the phylum distribution but, as expected, removed low prevalent ASVs from particular phyla (e.g., Actinobacteria and Proteobacteria).

3.2 | PIME application and effectiveness

Different data sets were used to validate the PIME workflow. PIME computed the OOB error rate from random forests, the number of taxa, and the number of remaining sequences for each prevalence

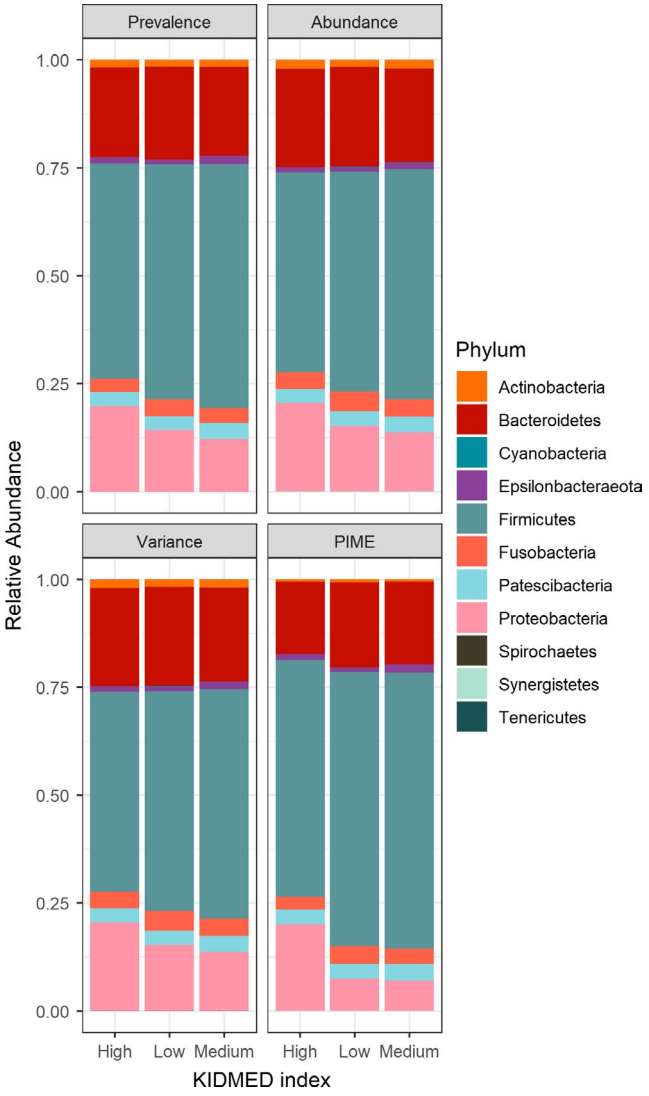


FIGURE 3 Changes in taxonomic composition at phylum level after filtering the saliva data set with PIME and other commonly used filtering methods. [Colour figure can be viewed at wileyonlinelibrary.com]

interval from the diet-saliva data set (Figure 4). Stringent criteria for definition of prevalence led to greater improvement in accuracy for predicting diet based on the salivary microbiota. The prevalence interval of 65% provided the best separation of microbial communities (OOB error = zero) while still including the majority of the sequences in the analysis. This prevalence interval was chosen for further analysis, but other intervals of prevalence can also be tested. For instance, the prevalence interval of 25% had OOB error of 7.2%. This indicates that the model is 92.8% accurate, which is a reasonably good model and keeps 88% of the sequences. Those ASVs that contributed to separating the core microbiomes among the diet categories (high, medium, or low diet categories) at 65% prevalence are provided by PIME (Table 1). The table indicates the ability of each variable to classify the microbes according to the three diet categories. The ASVs are ordered as most- to least-important. The more the accuracy of the random forest decreases due to the exclusion of a single ASV, the more important that ASV is, and therefore variables with a large mean decrease in accuracy are more important for classification of the ASVs according to diet. The mean decreased accuracy of the unfiltered data set presented extremely low values. Negative values or close to zero indicate that the variable does not have a role in the prediction. In other words, the variable is not important to differentiate groups (Table 1). On the other hand, after PIME filtering, the mean decrease accuracy values increased indicating a true contribution of each ASV to classify diet differences among the core microbiomes. Altogether, the results indicated that after PIME filtering differences in the saliva microbiome was partially explained by diet rather than by random distribution patterns. The traditional approach, not accounting for microbial prevalence, was unable to distinguish these differences.

Following this first test, 16S rRNA data from stools of 76 children at high genetic risk for type 1 diabetes (Davis-Richardson et al., 2014) were tested for prevalence differences in those samples from children who remained healthy versus those that became autoimmune. PIME computed the OOB error rate from random forests, the number of taxa, and the number of remaining sequences for each prevalence interval from this data set described (Figure 4). PIME also calculated prevalence intervals up to 70%. None of the sequences had a prevalence level higher than 70%. As expected, the OOB error rate decreased with higher prevalence intervals. At 60% prevalence the OOB error was zero and the number of remaining sequences was 1,165,304. The importance of each ASV in finding core microbiome differences between cases and controls subjects at 60% prevalence was also determined by PIME (Table 2). Accuracy was improved by applying PIME filtering to the data set compared to the unfiltered data set. Previously, Davis-Richardson et al. (2014) discovered that the relative abundance of *Bacteroides* was significantly higher in autoimmune versus control subjects. The presence of *Bacteroides* as an important taxa associated with autoimmune subjects was confirmed by PIME and other amplicon sequence variants (ASVs) belonging to *Veillonella* genus were also found associated with autoimmune subjects.

In the third data set tested, taxa were equally likely to be detected in the probiotic and placebo groups (Avershina et al., 2017). PIME prevalence filtering did not capture any difference between treatments (Figure 4). As the vaginal environment is dominated by *Lactobacillus*, a severe drop in the number of sequences at 5% prevalence interval was observed. The OOB error rate of the overall model obtained by random forest analysis suggests that irrespective of the prevalence interval no distinction between probiotic consumption and placebo exists (Supporting Information 3). Those results confirm the authors' previous findings and demonstrate that our approach is not prone to type I errors (finding false positive results).

PIME tested the association between saliva microbiome and the left antecubital fossa using a data set from the Human Microbiome Project (Human Microbiome Project Consortium, 2012). These two distinctive human microbial habitats were

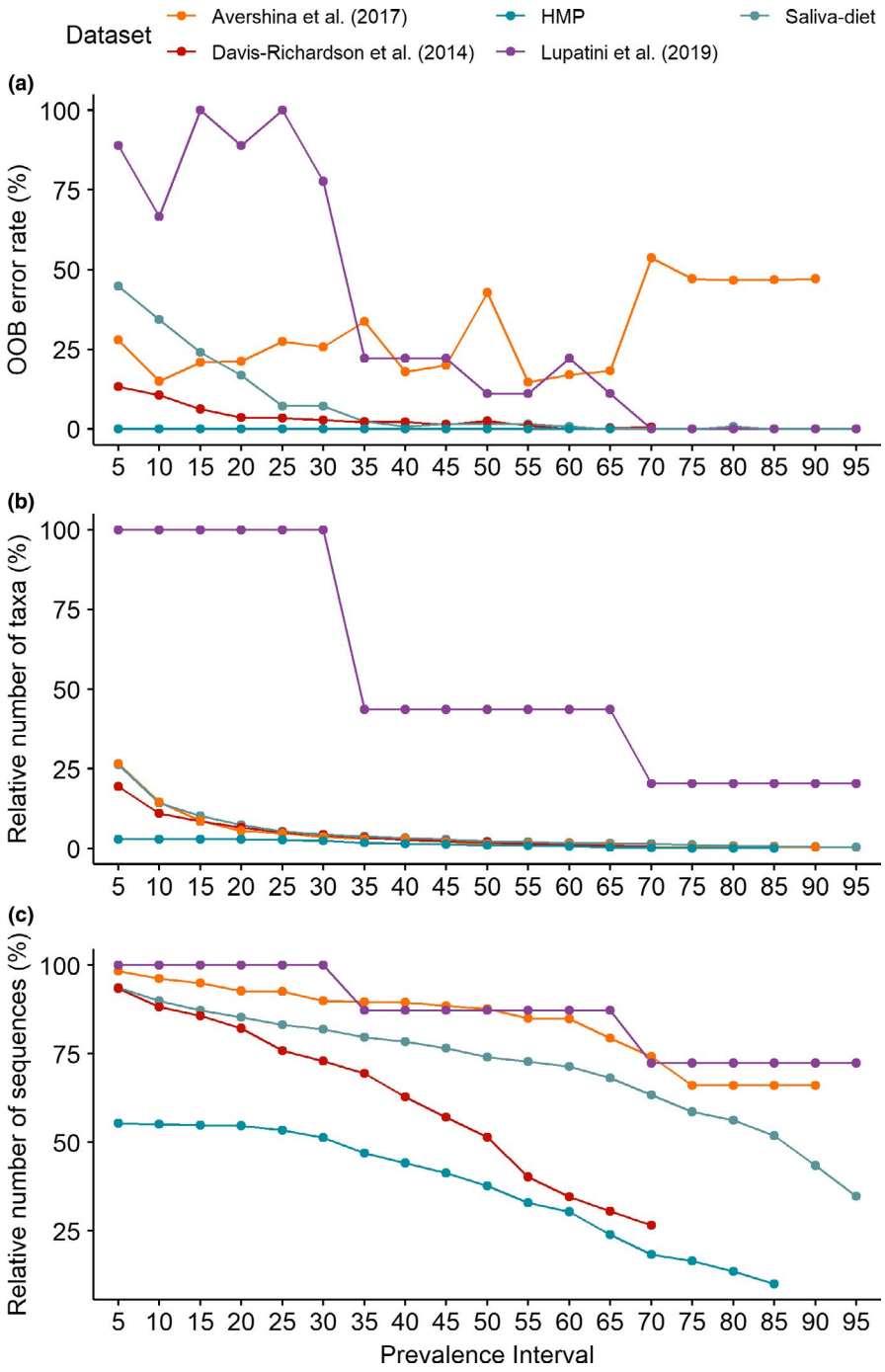


FIGURE 4 (a) Computations of the out-of-bag error rate from random forests, (b) percentage of remaining taxa and (c) percentage of remaining sequences for each prevalence interval from five different 16S rDNA data sets. [Colour figure can be viewed at wileyonlinelibrary.com]

selected as they were expected to harbour very different communities. As predicted, PIME showed that the microbial habitats tested were very distinct. The OOB error rate was 0.005 within the original data set and zero at all prevalence intervals applied (Figure 4 and Supporting Information 3) indicating the prevalence filtering does not increase the differentiation between these very different microbial habitats.

Finally, to show PIME can also be applied to any environmental survey and not only to human data sets, a database comprised of 16S rRNA sequences from soils was tested (Lupatini et al., 2019). The soil 16S rRNA sequencing was designed to investigate the effects of moisture and temperature under the microbial community (Figure 4 and Table 3). As expected, the OOB error rate decreased with higher prevalence intervals. After PIME filtering, the OOB error rate was approximately 88%. The best prevalence interval for this data set was 70% where the OOB error rate was zero. Also after PIME filtering, the mean decrease accuracy values (a measure of importance of a particular OTU/ASV to explain the model) increased indicating a true contribution of each ASV to detect moisture regime differences among the core microbiomes (Table 3). The mean decrease accuracy values were obtained from the calculations of OTU contribution to prevalence.

TABLE 1 Importance of ASVs measured by mean decrease accuracy and the confusion matrix prior and after PIME to differentiate the three diet categories (High, Low and Medium) from the diet-saliva data set

Mean decrease accuracy				
High	**Low**	**Medium**	**Over all classes**	**Closest microbial relative at genus level**
Unfiltered data set				
0.0088	0.0002	0.0041	0.0037	*Neisseria*
0.0013	0.0046	0.0003	0.0017	*Parvimonas*
−0.0002	0.0019	0.0022	0.0016	*Veillonella*
0.0009	0.0025	0.0013	0.0015	*Prevotella*
0.0003	0.0041	0.0006	0.0015	*Parvimonas*
0.0004	0.0022	0.0014	0.0015	*Porphyromonas*
0.0021	0.0014	0.0009	0.0013	*Actinobacillus*
0.0014	0.0021	0.0007	0.0012	*Haemophilus*
0.0024	0.0005	0.0010	0.0011	*Alloprevotella*
0.0007	0.0004	0.0016	0.0011	*Alloprevotella*
Data set filtered by PIME				
0.0200	0.0287	0.0883	0.0573	*Haemophilus*
0.0490	0.0060	0.0776	0.0504	*Haemophilus*
0.0167	0.0153	0.0567	0.0364	*Prevotella_7*
0.0317	0.0455	0.0318	0.0349	*Gemella*
0.0279	0.0345	0.0260	0.0287	*Prevotella*
0.0236	0.0312	0.0278	0.0276	*Haemophilus*
0.0087	0.0046	0.0483	0.0273	*Capnocytophaga*
0.0200	0.0653	0.0027	0.0239	*Selenomonas_3*
0.0145	0.0250	0.0250	0.0228	*Granulicatella*
0.0188	0.0206	0.0233	0.0216	*Rothia*

	Confusion matrix prior PIME			
	High	**Low**	**Medium**	**Classification error**
High	0	1	23	1.0000
Low	0	4	33	0.8918
Medium	0	2	62	0.0312
	Confusion matrix after PIME			
	High	**Low**	**Medium**	**Classification error**
High	24	0	0	0.0000
Low	0	37	0	0.0000
Medium	0	0	64	0.0000

Note: Showing only the first 10 hits. A complete table with the 30 most important ASVs is provided in the Supporting Information 3.

Once the best prevalence-filtered data set is determined, the logical next step is to use the same algorithm (random forests) to find the most important OTUs/ASVs responsible for the differences related to a given condition. PIME performs this analysis as described above. However, PIME users might also take advantage of third-party software to further analyze the filtered data set. To demonstrate this capability, the salivary microbiome from high and low KIDMED diet scores were compared using the DESeq2 algorithm (Love, Huber, & Anders, 2014). This example is provided in the Supporting Information 3.

3.3 | Likelihood of introducing type I error while building prevalence-filtered data sets

PIME includes an error detection step and the results are presented here (Figure 5). The biological difference among samples is expected to be greater than the differences generated randomly. Thus, as the prevalence interval increases, the OOB error should decrease. As expected, the OOB error rate of samples with true biologically relevant differences (Figure 5a,b,d,e) decreased (or remained constant in low noise data sets: Figure 5c) with the

TABLE 2 Importance of the ASVs measured by mean decrease accuracy and the confusion matrix prior and after PIME from the data set described by Davis-Richardson et al. (2014)

Mean decrease accuracy			
Controls	Cases	Over all classes	Closest microbial relative at genus level
Unfiltered data set			
0.0043	0.0043	0.0043	*Bacteroides*
0.0028	0.0061	0.0040	*Bacteroides*
0.0036	0.0039	0.0038	*Bacteroides*
0.0032	0.0045	0.0037	*Bacteroides*
0.0033	0.0041	0.0035	*Bacteroides*
0.0036	0.0033	0.0035	*Bacteroides*
0.0028	0.0047	0.0035	*Bacteroides*
0.0031	0.0035	0.0032	*Bacteroides*
0.0025	0.0041	0.0031	*Bacteroides*
0.0031	0.0033	0.0031	*Bacteroides*
Data set filtered by PIME			
0.0588	0.0148	0.0422	*Bacteroides*
0.0470	0.0119	0.0339	*Veillonella*
0.0386	0.0118	0.0284	*Bacteroides*
0.0372	0.0073	0.0260	*Veillonella*
0.0375	0.0063	0.0257	*Bacteroides*
0.0366	0.0070	0.0255	*Bacteroides*
0.0356	0.0074	0.0251	*Bacteroides*
0.0372	0.0045	0.0251	*Veillonella*
0.0366	0.0058	0.0250	*Veillonella*
0.0346	0.0076	0.0245	*Bacteroides*
	Confusion matrix prior PIME		
	Controls	Cases	Classification error
Controls	210	14	0.0625
Cases	54	79	0.4060
	Confusion matrix after PIME		
	Controls	Cases	Classification error
Controls	224	0	0.0000
Cases	0	133	0.0000

Note: Showing only the first 10 hits. A complete table with the 30 most important ASVs is provided in the Supporting Information 3.

increase in the prevalence interval. On the other hand, random sampling produced OOB error rate always higher than those obtained based on the original data set. In data sets with no expected biologically relevant differences (Figure 5c), the OOB error did not decrease with higher prevalence intervals. In those cases, the randomized data sets produced higher OOB error rates. Thus, the signal to noise ratio increases with the prevalence intervals generating low OOB error rate values while no improvements in accuracy are observed within the randomized data sets. This error detection analysis showed that no bias was introduced while building prevalence-filtered data sets confirming this workflow is not prone to type I errors.

4 | DISCUSSION

Prevalence is a key epidemiological concept where the number of people affected by a disease are counted with respect to the entire population (Noordzij, Dekker, Zoccali, & Jager, 2010; Ward, 2013). PIME was designed based on this concept. Here, the importance of a microbial community found in a single sample is less than if the same community is present in the majority of samples. Under that rationale, a workflow was designed to compare the prevalent populations between treatments. Prevalence has been used in the last as a filter of an entire data set but never as a means to distinguish treatments. Prevalence differences between treatments are masked when only relative abundance is considered.

Challenges in microbiome data include the presence of many taxa represented spacely in the data set. This often results in large variation in distribution patters (also known as overdispersion). Hence, microbes that are prominent in some subjects/samples and nearly absent in others (Kraal et al., 2014; Li, 2015). The current major challenge for using this information is how to convert it into rational biological conclusions providing control for error rates of false discoveries. Many tools are available to contrast experimental factors but they usually only take into account the microbial abundance and/or presence/absence. Thus, PIME overcomes those challenges by determining per treatment microbial prevalence in the analysis. This approach greatly improves the results by removing a substantial amount between-sample variation within groups that are represented by organisms that are rare in the population. PIME keeps only microbes found in many of the subjects of a population. The prevalence frequency can be chosen by the user after considering the OOB error rate. Thus, PIME can lead to a greater understanding of pathogenesis and the identification of potential probiotic treatments and prevention strategies that are masked by traditional analyses.

The definition of the best prevalence cutoff and the importance of taxa to discriminate treatments are performed by the random forests algorithm through the PIME workflow. This machine-learning algorithm has no formal distributional assumptions and can manage skewed and multimodal data as well as categorical data. It can also manage situations in which the number of predictor variables (OTUs/ASVs) greatly exceeds the number of observations (Cutler et al., 2007). PIME is simple and accurate compared to other machine learning methods (Statnikov et al., 2013) and is applicable for classification of binary and multicategory experiments. Classification by these means is very accurate even with the default parameters, (Statnikov et al., 2013; Zhou & Gallins, 2019). This ensures a more broad and practical use of PIME. Other methods that are not based on machine learning are able to identify taxa that are indicative of a given condition. This is often called differential abundance analysis. They usually compute *p*-values, adjusted *p*-values, false

TABLE 3 Importance of the OTUs measured by mean decrease accuracy and the confusion matrix prior and after PIME from the data set described by Lupatini et al. (2019)

Mean decrease accuracy				
Moisture regimes (%)				
8	16	23	Over all classes	Closest microbial relative at genus level
Unfiltered data set				
0.002	0.0000	0.0040	0.0020	*Reyranella*
0.002	0.0000	0.0020	0.0020	*Streptomyces*
0.002	0.0020	0.0000	0.0020	Unclassified Genus of Class OPB35
0.002	0.0020	0.0000	0.0020	*Candidatus Nostocoida*
0.000	0.0040	0.0020	0.0020	Unclassified Genus of Order *Armatimonadales*
0.000	0.0020	0.0040	0.0020	Unclassified Genus of Class BD7-11
0.000	0.0020	0.0040	0.0018	*Acidothermus*
0.002	0.0040	0.0000	0.0018	Unclassified Genus of Order *Rickettsiales*
0.004	0.0020	0.0000	0.0017	*Bacillus*
0.002	0.0000	0.0020	0.0017	*Jatrophihabitans*
Data set filtered by PIME				
0.004	0.006	0.002	0.0053	Unclassified Genus of Phylum *Armatimonadetes*
0.004	0.000	0.006	0.0037	Unclassified Genus of Family *Nitrosomonadaceae*
0.000	0.006	0.004	0.0035	Unclassified Genus of Family GR-WP33-30
0.002	0.001	0.004	0.0033	*Pirellula*
0.002	0.003	0.002	0.0033	*Chthonomonas*
0.003	0.000	0.004	0.0033	Pir4_lineage
0.004	0.004	0.000	0.0032	Unclassified Genus of Family TX1A-55
0.002	0.006	0.001	0.0030	*Gemmatimonas*
0.004	0.000	0.006	0.0030	*Sphingomonas*
0.000	0.003	0.004	0.0030	Unclassified Genus of Family *Anaerolineaceae*

	Confusion matrix prior PIME			
	8%	16%	23%	Classification error
8%	2	0	1	1.0000
16%	1	2	0	1.0000
23%	2	0	1	0.6667
	Confusion matrix after PIME			
	8%	**16%**	**23%**	**Classification error**
8%	3	0	0	0.0000
16%	0	3	0	0.0000
23%	0	0	3	0.0000

Note: Showing only the first 10 hits. A complete table with the 30 most important ASVs is provided in the Supporting Information 3.

discovery rates and effect sizes and are based on microbial abundances. Random forest analysis does not perform this traditional statistical inference. The importance of OTU/ASVs to differentiate treatments or ecological conditions may be used for the purpose of prediction. This machine-learning algorithm also provides an indication of the performance (OOB error rate) for comparisons of two or more microbial communities without OTU/ASV selection. This information is key into the workflow of PIME as the estimate of

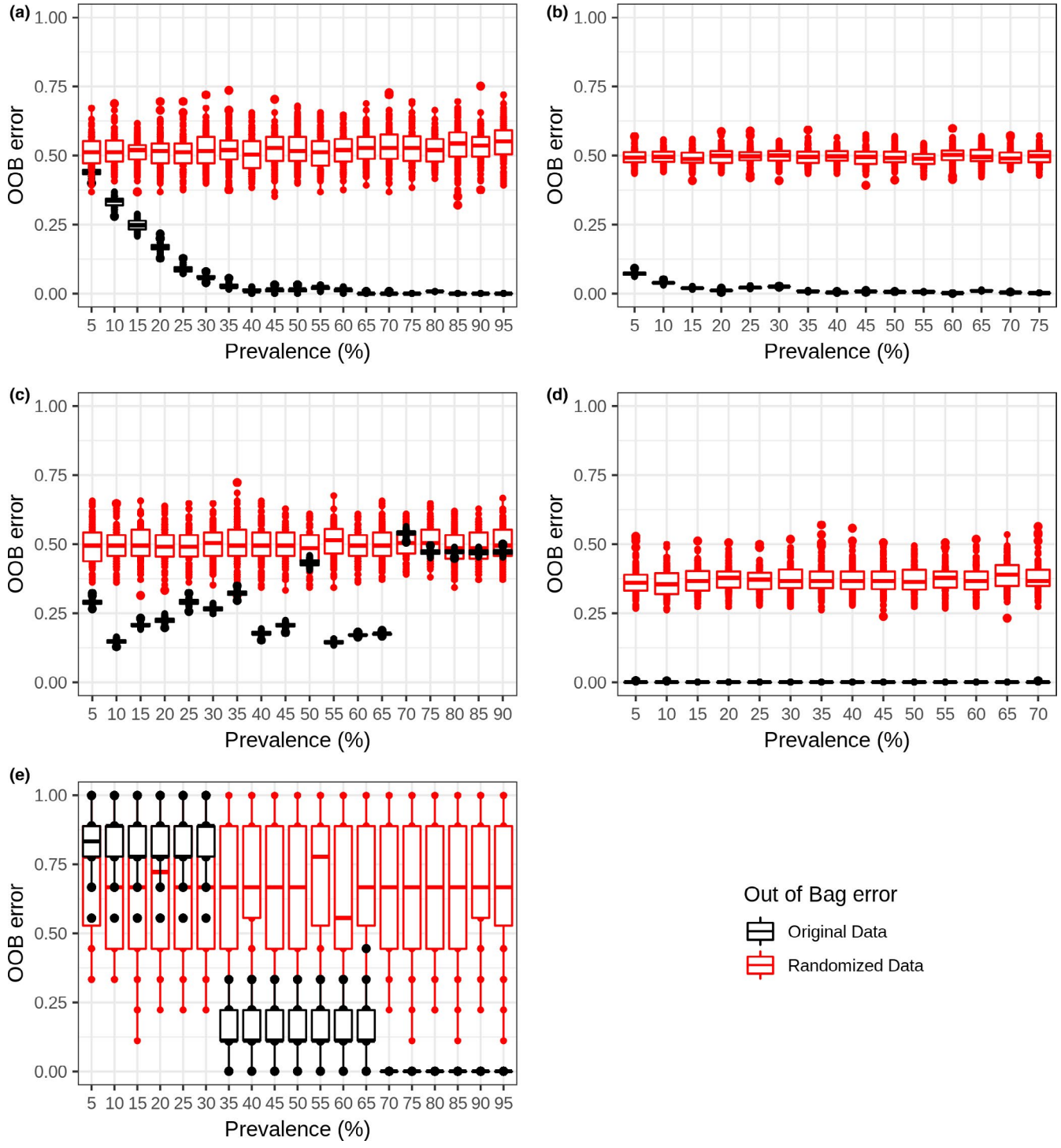


FIGURE 5 Boxplot depicting the PIME error detection step. Red boxes represent the OOB error rate obtained by randomly shuffling the labels into arbitrary groupings using 100 random permutations and running pime.error.prediction function at each randomization for each prevalence interval. Black boxes represent the OOB error rate against the 100 replications in each prevalence interval against the original sampling labels obtained by running pime.oob.replicate function. (a) Original data set from salivary microbiome samples. (b) Data from the gut microbial of 76 children at high genetic risk for type 1 diabetes. (c) Data from the vaginal microbiome of pregnant women randomized to receive milk with or without probiotic bacterial strains. (d) Data from the microbiome of saliva and left antecubital fossa of healthy individuals. (e) Data from the soil microcosm system designed to investigate the individual and interactive effects of moisture and temperature. Boxes span the first to third quartiles; the horizontal line inside the boxes represents the median. Whiskers extending vertically from the boxes indicate variability outside the upper and lower quartiles, and the circles indicate outliers. [Colour figure can be viewed at wileyonlinelibrary.com]

error is used to define the best prevalence cutoff for filtering low prevalent OTUs/ASVs. Taking these considerations into account, PIME is not comparable to other methods designed to perform microbial differential abundance analysis. However, any other tool can be applied after obtaining the prevalence-filtered data set by PIME. For example, the filtered data set can be analyzed using the DESeq2 algorithm (Love et al., 2014) to identify taxa that differ between the saliva microbiome from High and Low KIDMED index (Supporting Information 3).

Several tools designed to support microbiome statistical data analysis include data filtering as one of the first steps. The most commonly used filtering includes the exclusion of low count features (low abundance) using a minimum, yet arbitrary, relative abundance. Such features are very unlikely to be significant in the comparative analysis and likely have low overall prevalence. Arguably filtering those low abundance taxa can reduce the data sparsity issue, improving statistical power. However, when PIME performance is compared with other filtering methods, PIME outperformed all of those other approaches reducing the error rate and detecting microbial community differences where none were seen by other methods. This was illustrated by implementing PIME and other methods to a variety of 16S rRNA data sets. Within all of our tests, PIME confirmed previous findings and improved the results.

PIME does have some limitations. As PIME relies strongly on group prevalence, it is sensitive to the quality of sample groups. Poorly categorized groups comprised of subjects/samples with very different microbial composition may affect the prevalence calculations. Therefore, PIME might not be as effective in suggesting a good prevalence interval for filtering where groups are simply not different (Figure 5c). For data sets with a very large number of samples, PIME might not find a clear prevalence interval for data filtering as prevalence of a given taxon will probably decline as the number of samples increases. With an increasing number of samples, the chance of sampling different "cores" or subpopulations is also increased. In addition, when there is large heterogeneity within sample groups, coupled with high data sparsity, prevalence computation might not be successful. Also, although this does not affect prediction errors, random forest models are sensitive to multicollinear variables when informing variable importance. Colinear variables might have inaccurate importance values. For example, if the first chosen variable provides little information, the model may be less accurate. Nevertheless, we have shown using a variety of data sets that PIME can be useful in many circumstances to unveil differences in community structure that are not detected by other methods and it not subject to type 1 errors.

ACKNOWLEDGEMENTS

L.F.W.R. and P.C.T.D. were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), respectively. The work was supported by grants to E.W.T. from NIH (5R21AI120195-02) and JDRF (1-INO-2018-637-A-N). This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil (CAPES), Finance Code 001.

AUTHOR CONTRIBUTIONS

L.F.W.R. conceived the project, supervised the code writing, analyzed the data and wrote the paper; P.T.D. wrote the R code, analyzed the data and wrote the paper; V.S.P. supervised the data analysis, assisted with the experiment design and wrote the paper; B.K. supervised the code writing and wrote the paper; J.C.D. wrote the paper; E.W.T. supervised the data analysis, assisted with the experiment design and wrote the paper.

ORCID

Luiz Fernando W. Roesch https://orcid.org/0000-0003-1450-8828
Priscila T. Dobbler https://orcid.org/0000-0002-7681-5463
Victor S. Pylro https://orcid.org/0000-0003-2154-9150
Eric W. Triplett https://orcid.org/0000-0002-1845-4866

DATA AVAILABILITY STATEMENT

The R package, installation instructions and a step-by-step example on how to use PIME are freely available at: https://github.com/microEcology/PIME. The original 16S rRNA gene sequences generated in this work have been deposited in NCBI's Short Raw Archive and are accessible through BioProject ID PRJNA504439.

REFERENCES

Avershina, E., Slangsvold, S., Simpson, M. R., Storrø, O., Johnsen, R., Øien, T., & Rudi, K. (2017). Diversity of vaginal microbiota increases by the time of labor onset. *Scientific Reports*, *7*(1), https://doi.org/10.1038/s41598-017-17972-0

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Davis-Richardson, A. G., Ardissone, A. N., Dias, R., Simell, V., Leonard, M. T., Kemppainen, K. M., … Triplett, E. W. (2014). Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Frontiers in Microbiology*, *5*, 678. https://doi.org/10.3389/fmicb.2014.00678

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., & Xia, J. (2017). MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research*, *45*, W180–W188. https://doi.org/10.1093/nar/gkx295

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, *31*(21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2019). Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends in Microbiology*, *27*(2), 105–117. https://doi.org/10.1016/j.tim.2018.11.003

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3–4), 237–264. https://doi.org/10.1093/biomet/40.3-4.237

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Huse, S. M., Ye, Y., Zhou, Y., & Fodor, A. A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS ONE*, *7*(6), e34242. https://doi.org/10.1371/journal.pone.0034242

Kraal, L., Abubucker, S., Kota, K., Fischbach, M. A., & Mitreva, M. (2014). The prevalence of species and strains in the human microbiome: A resource for experimental efforts. *PLoS ONE*, *9*(5), e97279. https://doi.org/10.1371/journal.pone.0097279

Lemos, L. N., Fulthorpe, R. R., Triplett, E. W., & Roesch, L. F. W. (2011). Rethinking microbial diversity analysis in the high throughput sequencing era. *Journal of Microbiological Methods*, *86*(1), 42–51. https://doi.org/10.1016/j.mimet.2011.03.014

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, *2*(1), 73–94. https://doi.org/10.1146/annurev-statistics-010814-020351

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), https://doi.org/10.1186/s13059-014-0550-8

Lupatini, M., Suleiman, A. K. A., Jacques, R. J. S., Lemos, L. N., Pylro, V. S., Van Veen, J. A., … Roesch, L. F. W. (2019). Moisture is more important than temperature for assembly of both potentially active and whole prokaryotic communities in subtropical grassland. *Microbial Ecology*, *77*(2), 460–470. https://doi.org/10.1007/s00248-018-1310-1

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., … Edwards, R. A. (2008). The metagenomics RAST server – A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*(1), 386. https://doi.org/10.1186/1471-2105-9-386

Noordzij, M., Dekker, F. W., Zoccali, C., & Jager, K. J. (2010). Measures of disease frequency: Prevalence and incidence. *Nephron Clinical Practice*, *115*(1), c17–c20. https://doi.org/10.1159/000286345

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The madness of microbiome: Attempting to find consensus "Best Practice" for 16S microbiome studies. *Applied and Environmental Microbiology*, *84*(7), e02627-17. https://doi.org/10.1128/AEM.02627-17

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., … Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Serra-Majem, L. L., Ribas, L., García, A., Pérez-Rodrigo, C., & Aranceta, J. (2003). Nutrient adequacy and Mediterranean Diet in Spanish school children and adolescents. *European Journal of Clinical Nutrition*, *57*(S1), S35–S39. https://doi.org/10.1038/sj.ejcn.1601812

Serra-Majem, L., Ribas, L., Ngo, J., Ortega, R. M., García, A., Pérez-Rodrigo, C., & Aranceta, J. (2004). Food, youth and the Mediterranean diet in Spain. Development of KIDMED, Mediterranean Diet Quality Index in children and adolescents. *Public Health Nutrition*, *7*(07), 931–935. https://doi.org/10.1079/PHN2004556

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., … Alekseyenko, A. V. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, *1*(1), 11. https://doi.org/10.1186/2049-2618-1-11

Sze, M. A., & Schloss, P. D. (2016). Looking for a signal in the noise: Revisiting obesity and the microbiome. *Mbio*, *7*(4), 1–9. https://doi.org/10.1128/mBio.01018-16

Tedersoo, L., Nilsson, R. H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., … Kõljalg, U. (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, *188*(1), 291–301. https://doi.org/10.1111/j.1469-8137.2010.03373.x

Ward, M. M. (2013). Estimating disease prevalence and incidence using administrative data: Some assembly required. *The Journal of Rheumatology*, *40*(8), 1241–1243. https://doi.org/10.3899/jrheum.130675

Zhou, Y.-H., & Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in Genetics*, *10*, 579. https://doi.org/10.3389/fgene.2019.00579

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Roesch LFW, Dobbler PT, Pylro VS, Kolaczkowski B, Drew JC, Triplett EW. PIME: A package for discovery of novel differences among microbial communities. *Mol Ecol Resour*. 2020;20:415–428. https://doi.org/10.1111/1755-0998.13116