

1 **PIME: a package for discovery of novel differences among microbial communities**

2
3 Luiz Fernando W. Roesch¹, Priscila Thiago Dobbler¹, Victor Satler Pylro², Bryan Kolaczkowski³,
4 Jennifer C. Drew³ and Eric W. Triplett³

5
6 ¹Interdisciplinary Research Center on Biotechnology-CIP-Biotec, Universidade Federal do Pampa, São
7 Gabriel, Rio Grande do Sul, Brazil.

8 ²Microbial Ecology and Bioinformatics Lab., Department of Biology, Universidade Federal de Lavras, MG,
9 Brazil.

10 ³Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of
11 Florida, Gainesville, FL, USA.

12
13 **Abstract**

14
15 Massive sequencing of genetic markers, such as the 16S rRNA gene for prokaryotes, allows the
16 comparative analysis of diversity and abundance of whole microbial communities. However, the
17 data used for profiling microbial communities is usually low in signal and high in noise
18 preventing the identification of real differences among treatments. PIME (Prevalence Interval for
19 Microbiome Evaluation) fills this gap by removing those taxa that may be high in relative
20 abundance in just a few samples but have a low prevalence overall. The reliability and
21 robustness of PIME were compare against the existing methods and verified by a number of
22 approaches using 16S rRNA independent datasets. To remove the noise, PIME filters microbial
23 taxa not shared in a per treatment prevalence interval starting at 5% with increments of 5% at
24 each filtering step. For each prevalence interval, hundreds of decision trees are calculated to
25 predict the likelihood of detecting differences in treatments. The best prevalence-filtered dataset
26 is user-selected by choosing the prevalence interval that keeps the majority of the 16S rRNA
27 reads in the dataset and shows the lowest error rate. To obtain the likelihood of introducing bias
28 while building prevalence-filtered datasets, an error detection step based in random
29 permutations is also included. A reanalysis of previews published datasets with PIME
30 uncovered previously missed microbial associations improving the ability to detect important
31 organisms, which may be masked when only relative abundance is considered.

32
33

1 **INTRODUCTION**

2

3 Sequencing of amplified genetic markers (metataxonomics), e.g. the 16S rRNA gene, is
4 traditionally used for testing hypotheses based on microbial community composition. Taxonomic
5 differences among treatments or outcomes in microbiome surveys have changed our
6 understanding of the role played by microorganisms in the environment, plant and animal hosts,
7 including humans. The major challenge for using this information is their interpretation for the
8 discovery of the drivers of microbial diversity, the main taxa related to a given factor and the
9 reduction in false discovery rates. Generally, as microbiome studies present a large number of
10 taxa that are low in prevalence in many of the samples (1), these approaches frequently include
11 a variety of pre-filtering steps. Those steps include, but are not limited, to the exclusion of
12 sequences and/or taxonomic unities with low abundance, low variation, or low presence across
13 all samples. Moreover, removing arguably uninformative information, pre-filtering is also
14 advantageous because low abundance features in metataxonomic surveys might be also due to
15 sequencing errors or low level of contaminants from commercial kits (2, 3).

16 Besides filtering low abundance sequences, a frequent approach involves the exclusion
17 of microbial taxa under low prevalence across all samples. The prevalence of microbes in the
18 human microbiome is characterized by variable distribution patterns (4) with prominent
19 abundance of some strains in some subjects and nearly absence in others. While this unusual
20 distribution might be focus of research for future experimental study (4), identify microbial
21 taxonomic unities present in the majority of the subjects, also known as microbial core, has
22 been one of the primary goals of the Human Microbiome Project (5, 6). The central objective of
23 obtaining a healthy core microbiome is to use it to identify significant deviations from normality
24 that might be associated with disease states, for example.

25 Many tools such as DADA2 (7), Phyloseq (8), Qiime (9), UPARSE (10), MG-RAST (11),
26 mothur (12), MicrobiomeAnalyst (13) among others, have been developed to contrast
27 experimental factors in microbiome studies. The choice of a given analyses package is usually
28 based on the user's level of experience in bioinformatics and on the available resources at the
29 user's host institution (14), but unfortunately, the most used approaches embedded in these
30 packages rarely consider microbial prevalence.

31 Based on the microbial core concept, here we propose a new workflow designed to
32 identify and remove the within group variation found in metataxonomic surveys (16S rRNA
33 datasets) by capturing only biological differences at high sample prevalence levels. That means
34 in an experiment comparing two treatments (e.g. health against diseased subjects) one core for

1 each treatment will be calculated and relevant microbial taxa responsible for differences within
2 microbial cores will be detected. To implement this concept, we developed an R package called
3 PIME (Prevalence Interval for Microbiome Evaluation). PIME is a tool specifically designed to
4 work with datasets presenting high variations among samples. It removes per group microbial
5 taxa to keep only those taxa that are shared at some level of prevalence, using a machine
6 learning algorithm. For each prevalence level a list with the most relevant taxa responsible for
7 differences between or among groups is provided. To obtain the likelihood of introducing bias
8 while building prevalence-filtered datasets, an error detection step based on randomizations is
9 also included.

10

11

12 **PROGRAM DESCRIPTION AND METHODS**

13 **Bioinformatics Workflow**

14 The bioinformatics workflow described here is embedded into an R package called PIME
15 (Prevalence Intervals for Microbiome Evaluation) available at:
16 <https://github.com/microEcology/PIME>. PIME identifies statistically significant bacterial
17 community differences taking into account the proportion of samples hosting a specific microbial
18 community in a given time period. For the purpose of this work, prevalence was defined as the
19 proportion of individuals in a specific group who share taxa, irrespective of the abundance, at
20 the time of sampling. That is, a prevalence cutoff of 50% means that the taxa selected at this
21 prevalence interval are found in 50% of subjects. PIME's strategy is based on four fundamental
22 steps depicted in Figure 1 and explained below:

23

24 *I) Prediction of differences on full dataset:*

25 PIME takes a phyloseq object (8) as input, builds hundreds of decision trees using a
26 supervised machine learning algorithm and combines them into a single model to predict the
27 likelihood of detecting any user predefined factor (e.g. difference between treatments) as source
28 of sample variation (15). The model performance is indicated by the out-of-bag (OOB) estimate
29 of the error rate calculated by training the algorithm on a subset of samples and tested on the
30 remaining samples. Values can vary from 0 to 1, where zero indicates the model has 100%
31 accuracy and 1 that the model has zero accuracy. This overall measurement of accuracy can be
32 interpreted as an estimate of error obtained when the model is applied to new observations.
33 Higher OOB error indicates low accuracy of the model in predicting differences among the

1 variables tested. In this case, PIME might be used as an alternative to reduce noise by
2 removing microbial taxa with low prevalence among samples. This might help to improve the
3 model accuracy. This first step using the full dataset is implemented in a function called
4 *pime.oob.error*. This function is run using the dataset without any filtering proposed by PIME.
5 After obtaining the OOB error rate, the user should decide whether or not running PIME is
6 adequate to the dataset. For instance, an OOB error close to zero indicates the prevalence
7 filtering with PIME is not necessary, as the model accuracy is already reasonably good. On the
8 other hand, if OOB error rate is greater than zero, filtering the dataset using PIME might
9 improve the model accuracy. The user might then run the next function called
10 *pime.split.by.variable*, which is described below.

11

12 *II) Split the dataset by predictor variable and compute prevalence intervals:*

13 The full dataset is split according to treatments (or variables) defined by the user in the
14 metadata file. PIME can deal with two or more variables. Each variable will be used to define
15 data subsets. Those per variable subsets will be filtered using different prevalence levels from
16 5% up to 95% with increments of 5% for each level (see Figure 1 for a simplified schema
17 illustrating this filtering step). Prevalence levels (usually high prevalence levels – e.g. 90%)
18 where samples have zero counts are not calculated. After removal of taxa that did not match the
19 prevalence criteria, the subsets are merged to compose a new filtered dataset (one per
20 prevalence interval) used in the downstream analysis. This step is implemented in two functions
21 called: *pime.split.by.variable* and *pime.prevalence*. The *pime.split.by.variable* function uses the
22 original dataset as input and its output is used as input to the *pime.prevalence*. The function
23 *pime.prevalence* keeps, for each treatment group, every OTU/ASV according to the following
24 logical equation:

25

26
$$N_0/N_s > P_i \implies \text{True}$$

27

28 Where: N_0 is the number of OTUs/ASVs with $\text{Sum}>0$, N_s is the number of samples and
29 P_i is the prevalence interval $P_i=0.05, \dots, P_{\max}$.

30

1 *III) Computation of OOB error on each prevalence interval and importance of each taxa*
2 *to differentiate microbial communities*

3 At this step Random Forests analysis (15) is used to determine the level of prevalence
4 that provides the best model to predict differences in the communities while still including as
5 many taxa as possible in the analysis. The approach uses multiple learning algorithms to run
6 classifications based on decision trees. After prevalence filtering, for each prevalence interval
7 the OOB error rate, the number of remaining taxa and sequences is calculated. The results are
8 provided in a table that can be used to decide the best prevalence interval that provides a
9 classification model with reasonably good accuracy. This step is implemented in a function
10 called: *pime.best.prevalence*. Within the same function, the contribution of each taxa to the
11 mean decrease in classification accuracy is calculated. High values of mean decrease accuracy
12 indicate the importance of taxa to differentiate two or more microbial communities. The user can
13 access the importance of taxa in each of the prevalence interval calculated.

14

15 *IV) Validation*

16 To obtain the likelihood of introducing bias while building prevalence-filtered datasets, an
17 error detection step is also included under the following rationale: Consider the scenario in
18 which the null hypothesis of “no difference between groups” is false. If we randomly shuffle the
19 labels that identify the sample groups and run the test again the expected outcome is that the
20 randomized dataset will have a small chance to present distinct groups. Running the test
21 multiple times with the random dataset would produce a high OOB error rate in most cases.
22 This error detection test is implemented in two functions called *pime.error.prediction* and
23 *pime.oob.replicate*. The first function randomizes the samples labels into arbitrary groupings
24 using 100 random permutations. For each randomized prevalence filtered dataset, the OOB
25 error rate is calculated to determine whether differences in the original groups occur by chance.
26 The second function performs the Random Forest analyses and computes the OOB error for
27 100 replications in each prevalence interval without randomizing the sample labels. The
28 biological difference among samples is expected to be greater than the differences generated
29 randomly. Thus, the greatest fraction of randomizations should generate high error rates. On the
30 other hand, no improvement in accuracy is expected within the randomized dataset.

31

1

2 **Empirical Validation**

3 The PIME workflow was compared against other existing filtering methods and by using
4 empirical tests with 16S rDNA datasets. The performance of PIME was compared against
5 filtering methods based on overall prevalence, low abundance and low variance. Also, four 16S
6 rDNA datasets were analyzed using PIME to illustrate its usefulness. These include an
7 assessment of: a) the association between diet and saliva microbiome composition
8 (unpublished original research); b) the gut microbiome in subjects at high genetic risk for type 1
9 diabetes (16); c) the vaginal microbiome in pregnant women randomized to receive milk with or
10 without probiotic bacterial strains (17); and d) the saliva microbiome compared the left
11 antecubital fossa of healthy individuals (Human Microbiome Consortium, 2012).

12 The 16S rRNA gene sequences generated in this work have been deposited in NCBI's
13 Short Raw Archive and are accessible through BioProject ID PRJNA504439.

14

15

16 *Comparison with other existing filtering methods*

17 Comparisons were performed using a dataset composed by 16S rRNA sequences from
18 microbes extracted from saliva of 125 undergraduate and graduate students from the University
19 of Florida (accessible through BioProject ID PRJNA504439). The following filtering tests were
20 performed: a) filtering the dataset by overall prevalence. To be kept, taxa must be present in at
21 least 20% of the subjects; b) filtering the dataset by abundance. To be kept, taxa must have at
22 least 5 sequences; c) filtering by low variance. To be kept taxa must have variance higher than
23 20%. Filtered datasets were compared against the prevalence interval of 65% as calculated by
24 PIME as the best prevalence interval where the OOB error was zero. A record of this analysis
25 containing a step-by-step R-code and results is provided in the Supplementary File S1.

26

27

28 **Performance evaluation with 16S rRNA datasets**

29 A novel and three published datasets were analyzed with PIME. The novel dataset used
30 in this work comprised of 16S rRNA gene sequences from saliva samples obtained from 125

1 undergraduate and graduate students from the University of Florida. The study assessed the
2 subject's diet as a factor influencing the saliva microbiome. This study was approved by the
3 University of Florida's Institutional Review Board and assigned number IRB201602134.
4 Approximately 224 undergraduate and graduate students taking three courses were invited to
5 anonymously participate in this study as volunteers. A study coordinator was chosen to collect
6 samples and code the samples so that those who did the analysis were unaware of the identity
7 of the volunteers. To assess the diet, the subjects also completed the KIDMED survey (18). The
8 sampling collection, DNA extraction and library preparation are described below.

9

10

11 *Sampling collection, DNA extraction and library preparation*

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

17 To assess the diet, the subjects also completed the KIDMED survey (18). The KIDMED
18 Index is based on a series of 16 questions, which measures the degree to which a subject
19 adheres to the Mediterranean diet. The KIDMED index has been validated with nutritional data
20 (19) and was much simpler to implement than a diet diary or a serum-based nutrition analysis.
21 Participant's age and gender were also obtained.

22 The 16S rRNA library preparation was performed as described previously (16) and
23 sequenced with Illumina MiSeq: 2x300 cycles run. The raw fastq files were used to build a table
24 of exact amplicon sequence variants (ASVs) with DADA2 version 1.8 (7). Taxonomy was
25 assigned to each ASV using the SILVA ribosomal RNA gene database version v132 (20). A
26 detailed R script containing the code used to generate the ASV table is provided in the
27 Supplementary File S2. Downstream analyses were carried out using a rarefied dataset of
28 24,900 sequences as previously recommended by Lemos et al. (21).

29

30

31 *Description of the previously published datasets*

1 The first previously published dataset used here was described by Davis-Richardson et
2 al. (16) and comprised of partial 16S rDNA sequences from fecal samples of 76 subjects born
3 between 1996 and 2007 at the Turku University Hospital in southwestern Finland. All subjects
4 were at high genetic risk for type 1 diabetes. The cohort was retroactively selected to create an
5 age-matched genotype-controlled set of subjects for the investigation of the microbiome as an
6 environmental factor influencing the development of Type-1 diabetes. The raw Fastq files were
7 obtained and sequences were processed using DADA2 version 1.8 (7), as described above.
8 Cases were defined as subjects who developed at least two persistent islet cell autoantibody
9 (ICA), IAA, GADA, or IA-2A. Controls were defined as subjects with no detectable islet
10 autoantibodies. Samples from subjects older than one year and post seroconversion were
11 removed.

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

20 The third previously published dataset comprised of 16S rRNA gene sequences from the
21 V1-V3 hypervariable region downloaded from the NIH Human Microbiome Project
22 (<https://www.hmpdacc.org/HMQCP/#data>). The final OTU table processed by Qiime (9) using an
23 OTU-clustering strategy and accompanying metadata were obtained and loaded into the R
24 environment. After removing singletons, only saliva and left antecubital fossa samples were
25 kept. The final dataset comprised of 113 saliva samples and 59 left antecubital fossa samples
26 all rarefied at 2,000 sequences pre sample. A record of all statistical analyses comparing the
27 datasets with and without using PIME including the R-code and results is included in
28 Supplementary File S3.

29

30

31 **Performance of PIME compared against other filtering methods**

1 The results comparing the performance of PIME with other filtering methods are
2 presented in Figure 2. After quality filtering the saliva dataset, a total of 4,981,638 high-quality
3 paired sequences, 400 bp long, were obtained from all subjects. An average 44,258 reads per
4 sample were obtained. The dataset was rarefied to 24,900 reads per sample in all analyses
5 commensurate with the lowest number of reads found in any one sample. This number of reads
6 was sufficient to accurately reflect the microbial diversity in these samples given the low
7 complexity of saliva samples. The best prevalence interval calculated by PIME was at 65%. This
8 prevalence interval was used to compare the performance of PIME against the other filtering
9 methods. The original dataset, without any filtering, presented 4,555 taxa and a total of
10 3,112,500 sequences after rarefaction. Both prevalence overall and PIME excluded the highest
11 proportion of ASVs and sequences while filtering by abundance or variance excluded only 78%
12 of ASVs and kept 99.9% of the reads. Nevertheless, the overall prevalence kept 84% of the
13 sequences while PIME kept 68% of the total number of sequences. Without using the PIME
14 filtering the OOB error obtained while attempting to classify the salivary microbe according to
15 the three diet categories was 44% indicating the model had low accuracy in predicting diet
16 according to the microbiota. Overall prevalence, abundance and variance filtering also
17 presented low accuracy in classifying diet according to the microbiota however, after PIME
18 filtering the accuracy of the model increased to 100%.

19

20

21 **PIME application and effectiveness**

22 Different datasets were used to validate the PIME workflow. The computations of the
23 OOB error rate from random forests, the number of taxa and the number of remaining
24 sequences for each prevalence interval from the diet-saliva dataset are presented in Table 1.
25 Stringent criteria for definition of prevalence lead to greater improvement in accuracy for
26 predicting diet based on the salivary microbiota. The prevalence interval of 65% provided the
27 best separation of microbial communities (OOB error = zero) while still including the majority of
28 the sequences in the analysis. This prevalence interval was chosen for further analysis, but
29 other intervals of prevalence can also be tested. For instance, the prevalence interval of 25%
30 had OOB error of 7.2%. This indicates that the model is 92.8% accurate, which is a reasonably
31 good model and keep 88% of sequences. The importance of each ASV in finding microbiome

1 differences among diet categories (high, medium, or low diet categories) for the prevalence
2 interval of 65% is presented in Table 2. The table indicates the ability of each variable to classify
3 the microbes according to the three diet categories. The ASVs are ordered as most- to least-
4 important. The more the accuracy of the random forest decreases due to the exclusion of a
5 single ASV, the more important that variable is, and therefore variables with a large mean
6 decrease in accuracy are more important for classification of the ASVs according to diet. The
7 mean decrease accuracy of the unfiltered dataset presented negative values, which are a clear
8 warning sign the model might be overfitting noise (Table 2). On the other hand, after PIME
9 filtering, the mean decrease accuracy values were all positive indicating a true contribution of
10 each ASV to classify diet according to the microbiota. Altogether, the results indicated that after
11 PIME filtering differences in the saliva microbiome was partially explained by diet rather than by
12 random distribution patterns. The traditional approach, not accounting for microbial prevalence,
13 was unable to distinguish these differences.

14 Following this first test, 16S rDNA data from stool of 76 children at high genetic risk for
15 type 1 diabetes (16) were tested for prevalence differences in those samples from children who
16 remained healthy versus those that became autoimmune. The computations of the OOB error
17 rate from random forests, the number of taxa and the number of remaining sequences for each
18 prevalence interval from the dataset described by Davis-Richardson *et al.* (2014) are presented
19 in Table 3. PIME was able to calculate prevalence interval up to 70%. At prevalence intervals
20 higher than 70% samples had zero counts and prevalence was not calculated. As expected, the
21 OOB error rate decreased with higher prevalence intervals. At 60% prevalence interval the OOB
22 error was zero and the number of remaining sequences was 1,165,304. The importance of each
23 ASV in finding microbiome differences among cases and controls subjects under risk for T1
24 diabetes for the prevalence interval of 60% is presented in Table 4. Comparing the results
25 obtained by the unfiltered dataset with the PIME filtered dataset we observe an improvement in
26 accuracy. Previously, Davis-Richardson *et al.* (2014) discovered that the relative abundance of
27 *Bacteroides* was significantly higher in autoimmune vs. control subjects. The higher abundance
28 of *Bacteroides* was confirmed by PIME and other Amplicon Sequence Variants (ASVs)
29 belonging to *Bifidobacterium* genus were also found associated with autoimmune subjects.

30 In the third dataset tested, taxa were equally likely to be detected in the probiotic and
31 placebo groups (17). Prevalence testing by PIME also does not capture any difference between

1 treatments (Table 5). As the vaginal environment is dominated by *Lactobacillus*, a severe drop
2 in the number of sequences at 5% prevalence interval was observed. The OOB error rate of the
3 overall model obtained by Random Forest analysis suggests that irrespective of the prevalence
4 interval no distinction between probiotic consumption and placebo exists (Supplementary File
5 3). Those results confirm the author's previous findings and demonstrate our approach is not
6 prone to type I errors (finding false positive results).

7 Finally, PIME was tested using the association between saliva microbiome and the left
8 antecubital fossa, a dataset from the Human Microbiome Project (22). These two distinctive
9 human microbial habitats were selected as they are expected to harbor very different
10 communities. As predicted, PIME showed that the microbial habitats tested are very distinct.
11 The OOB error rate was 0.005 within the original dataset and zero at all prevalence intervals
12 applied (Table 6) indicating the prevalence filtering does not increase the differentiation between
13 these very different microbial habitats.

14

15

16 **Likelihood of introducing bias while building prevalence-filtered datasets**

17 The results obtained by the PIME error detection step are presented in Figure 3. The
18 biological difference among samples is expected to be greater than the differences generated
19 randomly. This way, as the prevalence interval increases the OOB error might decrease. As
20 expected, the OOB error rate of samples with true biological relevant differences (Figures 3A,
21 3B and 3D) decreased (or remained constant in low noise datasets – Figure 3D) with the
22 increase in the prevalence interval definition. On the other hand, random sampling produced
23 OOB error rate always higher than those obtained based on the original dataset. In datasets
24 with no expected biological relevant differences (Figure 3C), the OOB error did not decrease
25 with higher prevalence interval definitions and the randomized dataset produced higher OOB
26 error rates. Thus, the signal to noise ratio increases with the prevalence intervals generating low
27 OOB error rate values while no improvements in accuracy are observed within the randomized
28 datasets. This error detection analysis showed that no bias was introduced while building
29 prevalence-filtered datasets confirming this workflow is not prone to type I errors.

30

31

1 **CONCLUSIONS**

2

3 Prevalence is a key epidemiological concept involving the counting of the number of
4 people affected by a disease (23, 24). PIME was designed based on this concept. Here we
5 argue the importance of a microbial community found in a single sample is smaller than if the
6 same community is present in the majority of samples. Under such rationale, we designed a
7 workflow capable of improving the ability to detect important organisms as it considers the
8 extent to which an organism is present across a given population, which may be masked when
9 only relative abundance is considered. Challenges in microbiome data include sparsity
10 (presence of many zeros) and large variance in distribution patters (also known as over-
11 dispersion) with prominent abundance of some microbes in some subjects/samples and nearly
12 absence in others (4, 25). The current major challenge for using this information is indubitably
13 how to convert it into rational biological conclusions providing control for error rates of false
14 discoveries. Many tools have been successfully developed aiming to contrast experimental
15 factors but they usually only take into account the microbial abundance and/or
16 presence/absence. Thus, PIME is expected surpass those challenges including the concept of a
17 per treatment microbial prevalence in the analysis. This approach greatly improves the results
18 by removing interpersonal variation within groups (unique microbes found in a single
19 subject/sample) and keeping only microbes found in most of the subjects of a population (likely
20 to be associated with a experimental variable). This approach reveals microbes important in a
21 disease that may be overlooked by traditional methods leading to a greater understanding of
22 pathogenesis and the identification of potential probiotic treatment and prevention strategies.

23 Several tools designed to support microbiome statistical data analysis include data
24 filtering as one of the first steps. The most commonly used filtering includes the exclusion of low
25 count features (low abundance) using a minimum, yet arbitrary, cutoff, low variance (assuming
26 that features under low variance are very unlikely to be significant in the comparative analysis)
27 and low overall prevalence. Arguably filtering those uninformative taxa can improve the data
28 sparsity issue, improving statistical power. Here we compare the performance of PIME with
29 these other filtering methods. PIME outperformed all of those other approaches reducing the
30 error rate and detecting microbial community differences where none were seen by other
31 methods. To illustrate the application and the value of PIME, it was also implemented in a

1 variety of 16S rRNA datasets. Within all of our tests we confirmed previous findings and
2 improved the results.

3 During the course of analysis and tests we also detected some potential limitations of
4 PIME. As PIME relies strongly on group prevalence, it is sensitive to the quality of sample
5 groups. Poorly categorized groups made up of subjects/samples with very different microbial
6 composition might affect the prevalence computations and therefore PIME might not be as
7 effective in suggesting a good prevalence interval for filtering. For datasets with very large
8 number of samples, PIME might not find a clear prevalence interval for data filtering. With
9 increased number of samples, the chance of sampling different “cores” or subpopulations is also
10 increased. In addition, when there is large heterogeneity within sample groups, coupled with
11 high data sparsity, prevalence computation might not be successful. Another possible limitation
12 of PIME arises from random forests method, wrapped in pime.oob.error and
13 pime.best.prevalence functions. Random forests models are sensitive to multicollinear variables
14 when informing variable importance, though it doesn’t affect prediction errors. Colinear variables
15 might have inaccurate importance values as the difference is explained by the, randomly, first
16 chosen variable and little information is added to the model after this.

17

18

19

20 **Acknowledgements**

21 L.F.W.R. and P.C.T.D. were supported by Conselho Nacional de Desenvolvimento Científico e
22 Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
23 (CAPES), respectively. The work was supported by grants to E.W.T. from NIH (5R21AI120195-
24 02) and JDRF (1-INO-2018-637-A-N).

25

26 **Availability and implementation**

27 The R package, installation instructions and a step-by-step example on how to use PIME are
28 freely available at: <https://github.com/microEcology/PIME>.

29

1 **Table 1.** Computations of the out-of-bag error rate from random forests, number of taxa and
2 number of remaining sequences for each prevalence interval from the diet-saliva dataset.

Prevalence Interval	OOB error rate (%)	Number of OTUs	Number of seqs.
0.05	44.8	1,158	2,915,670
0.10	34.4	627	2,797,858
0.15	24	448	2,715,217
0.20	16.8	327	2,653,319
0.25	7.2	235	2,587,433
0.30	7.2	196	2,547,536
0.35	2.4	169	2,479,055
0.40	0.8	151	2,438,026
0.45	1.6	130	2,383,250
0.50	1.6	104	2,302,147
0.55	1.6	91	2,264,871
0.60	0.8	84	2,222,431
0.65	0	76	2,120,215
0.70	0	66	1,972,958
0.75	0	53	1,824,764
0.80	0.8	43	1,747,231
0.85	0	34	1,612,451
0.90	0	26	1,351,690
0.95	0	17	1,078,200

3
4

1 **Table 2.** Importance of ASVs measured by mean decrease accuracy to differentiate the three
2 diet categories (High, Low and Medium) from the diet-saliva dataset.

Mean Decrease Accuracy				
High	Low	Medium	Over all classes	Genus
Unfiltered Dataset				
0.0001	0.0002	-0.0004	-0.0002	<i>Veillonella</i>
0.0005	-0.0001	0.0004	0.0002	<i>Streptococcus</i>
-0.0007	-0.0002	-0.0009	-0.0007	<i>Prevotella_7</i>
0.0006	-0.0013	0.0005	-0.0001	<i>Haemophilus</i>
-0.0003	-0.0008	-0.0008	-0.0007	<i>Veillonella</i>
0.0011	-0.0017	0.0018	0.0007	<i>Veillonella</i>
-0.0003	-0.0004	-0.0006	-0.0005	<i>Veillonella</i>
-0.0018	-0.0016	0.0007	-0.0005	<i>Unclassified Genera</i>
0.0001	-0.0006	-0.0009	-0.0006	<i>Neisseria</i>
0.0001	0.0002	-0.0004	-0.0002	<i>Veillonella</i>
Dataset filtered by PIME				
0.0200	0.0287	0.0883	0.0573	<i>Veillonella</i>
0.0490	0.0060	0.0776	0.0504	<i>Neisseria</i>
0.0167	0.0153	0.0567	0.0364	<i>Neisseria</i>
0.0317	0.0455	0.0318	0.0349	<i>Prevotella_7</i>
0.0279	0.0345	0.0260	0.0287	<i>Alloprevotella</i>
0.0236	0.0312	0.0278	0.0276	<i>Prevotella_7</i>
0.0087	0.0046	0.0483	0.0273	<i>Haemophilus</i>
0.0200	0.0653	0.0027	0.0239	<i>Porphyromonas</i>
0.0145	0.0250	0.0250	0.0228	<i>Megasphaera</i>
0.0188	0.0206	0.0233	0.0216	<i>Selenomonas_3</i>

3 Showing only the first 10 hits. A complete table with the 30 most important ASVs is provided in the
4 Supplementary File 3.

5
6

1 **Table 3.** Computations of the out-of-bag error rate from random forests, number of taxa
2 and number of remaining sequences for each prevalence interval from the dataset
3 described by Davis-Richardson *et al.* (2014)

4

Prevalence Interval	OOB error rate (%)	Number of OTUs	Number of seqs.
0.05	13.17	979	3,152,779
0.10	10.64	556	2,975,527
0.15	6.16	431	2,891,465
0.20	3.64	335	2,773,432
0.25	3.36	258	2,556,564
0.30	2.8	219	2,457,867
0.35	2.24	175	2,339,966
0.40	2.24	143	2,119,062
0.45	1.4	115	1,922,701
0.50	2.52	99	1,733,801
0.55	1.12	74	1,357,557
0.60	0	62	1,165,304
0.65	0.28	49	1,026,879
0.70	0.56	39	897,367

5

6

7

8

9

1 **Table 4.** Importance of ASVs measured by mean decrease accuracy to differentiate the cases
2 and controls at high genetic risk for type 1 diabetes from the dataset described by Davis-
3 Richardson *et al.* (2014).

Mean Decrease Accuracy			
Controls	Cases	Over all classes	Genus
Unfiltered Dataset			
0.0028	0.0061	0.0040	<i>Bacteroides</i>
0.0031	0.0033	0.0031	<i>Bacteroides</i>
0.0011	0.0036	0.0020	<i>Bacteroides</i>
0.0037	0.0017	0.0030	<i>Bacteroides</i>
0.0006	0.0012	0.0008	<i>Bacteroides</i>
0.0012	0.0003	0.0008	<i>Bacteroides</i>
0.0025	0.0020	0.0023	<i>Bacteroides</i>
0.0021	0.0015	0.0019	<i>Bacteroides</i>
0.0028	0.0047	0.0035	<i>Bacteroides</i>
0.0005	0.0008	0.0006	<i>Bifidobacterium</i>
Dataset filtered by PIME			
0.0588	0.0148	0.0422	<i>Bacteroides</i>
0.0470	0.0119	0.0339	<i>Bacteroides</i>
0.0386	0.0118	0.0284	<i>Bifidobacterium</i>
0.0372	0.0073	0.0260	<i>Bifidobacterium</i>
0.0375	0.0063	0.0257	<i>Bacteroides</i>
0.0366	0.0070	0.0255	<i>Bacteroides</i>
0.0356	0.0074	0.0251	<i>Bacteroides</i>
0.0372	0.0045	0.0251	<i>Bacteroides</i>
0.0366	0.0058	0.0250	<i>Bacteroides</i>
0.0346	0.0076	0.0245	<i>Bacteroides</i>

4 Showing only the first 10 hits. A complete table with the 30 most important ASVs is provided in the
5 Supplementary File 3.

6

7

1 **Table 5.** Computations of the out-of-bag error rate from random forests, number of taxa
2 and number of remaining sequences for each prevalence interval from the dataset
3 described by Avershina *et al.*, (2017).

Prevalence Interval	OOB error rate (%)	Number of OTUs	Number of seqs.
0.05	28.06	123	988,309
0.10	14.93	68	967,461
0.15	20.90	40	954,379
0.20	21.19	26	931,177
0.25	27.46	22	929,563
0.30	25.67	17	903,330
0.35	33.73	14	899,992
0.40	17.91	14	898,656
0.45	20.00	10	889,244
0.50	42.69	7	881,104
0.55	14.63	7	853,874
0.60	17.01	6	853,217
0.65	18.21	5	797,394
0.70	53.73	3	745,832
0.75	47.16	2	663,347
0.80	46.57	2	663,347
0.85	46.87	2	663,347
0.90	47.16	2	663,347

4
5
6

1 **Table 6.** Computations of the out-of-bag error rate from random forests, number of taxa
2 and number of remaining sequences for each prevalence interval from the Human
3 Microbiome Project dataset.

Prevalence Interval	OOB error rate (%)	Number of OTUs	Number of seqs.
0.05	0	509	190,455
0.10	0	509	189,296
0.15	0	509	188,555
0.20	0	506	188,189
0.25	0	473	183,574
0.30	0	421	176,217
0.35	0	328	161,346
0.40	0	274	151,525
0.45	0	234	142,112
0.50	0	185	129,369
0.55	0	143	112,978
0.60	0	118	104,474
0.65	0	74	82,052
0.70	0	48	62,817
0.75	0	34	56,457
0.80	0	23	46,475
0.85	0	13	34,459

4
5
6
7
8
9

1 **Figures Captions**

2

3 **FIGURE 1.** Empirical representation of steps used in PIME. Top panel. Simplified schema
4 illustrating PIME method with a subset of 12 saliva's microbiome samples. Each sample (red,
5 yellow and blue circles) is connected to an ASV (white circles) through edges (green). ASVs
6 observed in more than one sample are connected by at least two edges and are displayed at
7 the center of the network. ASVs present in only one sample are connected by a single edge and
8 are displayed at the border of the network. The first step applied by PIME is to split the full
9 dataset according to the treatments defined by the user. Within this example red, yellow and
10 blue circles depict three different treatments. At each of the three new groups the low prevalent
11 ASVs are removed. Finally, the subsets are merged to compose a new filtered dataset used in
12 the downstream analysis. Bottom panel. Step-by-step representation of PIME workflow and
13 validation.

14

15 **FIGURE 2.** Performance of PIME compared to other filtering methods. A) Out of Bag error rate
16 (OOB error rate); B) total number of sequences; C) Total number of ASVs. Prevalence = filter by
17 overall taxa prevalence in at least 20% of the subjects; Abundance = filter by abundance of at
18 least 5 sequences; Variance = filter by variance higher than 20%. PIME = filter by prevalence
19 interval of 65%.

20

21 **FIGURE 3.** Boxplot depicting the PIME error detection step. Red boxes represent the OOB error
22 rate obtained by randomly shuffling the labels into arbitrary groupings using 100 random
23 permutations and running `pime.error.prediction` function at each randomization for each
24 prevalence interval. Black boxes represent the OOB error rate against the 100 replications in
25 each prevalence interval against the original sampling labels obtained by running
26 `pime.oob.replicate` function. (A) Original dataset from salivary microbiome samples. (B) Data
27 from the gut microbial of 76 children at high genetic risk for type 1 diabetes. (C) Data from the
28 vaginal microbiome of pregnant women randomized to receive milk with or without probiotic
29 bacterial strains. (D) Data from the microbiome of saliva and left antecubital fossa of healthy
30 individuals. Boxes span the first to third quartiles; the horizontal line inside the boxes represents

1 the median. Whiskers extending vertically from the boxes indicate variability outside the upper
2 and lower quartiles, and the circles indicate outliers.

3

4

5 **Table captions**

6 **Table 1.** Computations of the out-of-bag error rate from random forests, number of taxa and
7 number of remaining sequences for each prevalence interval from the diet-saliva dataset.

8

9 **Table 2.** Importance of ASVs measured by mean decrease accuracy to differentiate the three
10 diet categories (High, Low and Medium) from the diet-saliva dataset.

11

12 **Table 3.** Computations of the out-of-bag error rate from random forests, number of taxa and
13 number of remaining sequences for each prevalence interval from the dataset described by
14 Davis-Richardson *et al.* (2014).

15

16 **Table 4.** Importance of ASVs measured by mean decrease accuracy to differentiate the cases
17 and controls at high genetic risk for type 1 diabetes from the dataset described by Davis-
18 Richardson *et al.* (2014).

19

20 **Table 5.** Computations of the out-of-bag error rate from random forests, number of taxa and
21 number of remaining sequences for each prevalence interval from the dataset described by
22 (17).

23

24 **Table 6.** Computations of the out-of-bag error rate from random forests, number of taxa and
25 number of remaining sequences for each prevalence interval from the Human Microbiome
26 Project dataset.

27

28

29 **Supplementary information**

30

31 **Supplementary File S1.** Comparison of PIME with other existing filtering methods.

32 **Supplementary File S2.** The pipeline used to assign 16S rRNA sequences to ASVs.

33 **Supplementary File S3.** Detailed and reproducible description of PIME data analysis.

34

35

36 **References**

37 1. Sze, M.A. and Schloss, P.D. (2016) Looking for a Signal in the Noise: Revisiting Obesity and
38 the Microbiome. *mBio*, 7.

1 2. Salter,S.J., Cox,M.J., Turek,E.M., Calus,S.T., Cookson,W.O., Moffatt,M.F., Turner,P.,
2 Parkhill,J., Loman,N.J. and Walker,A.W. (2014) Reagent and laboratory contamination
3 can critically impact sequence-based microbiome analyses. *BMC Biol.*, **12**, 87.

4 3. Eisenhofer,R., Minich,J.J., Marotz,C., Cooper,A., Knight,R. and Weyrich,L.S. (2019)
5 Contamination in Low Microbial Biomass Microbiome Studies: Issues and
6 Recommendations. *Trends Microbiol.*, **27**, 105–117.

7 4. Kraal,L., Abubucker,S., Kota,K., Fischbach,M.A. and Mitreva,M. (2014) The Prevalence of
8 Species and Strains in the Human Microbiome: A Resource for Experimental Efforts.
9 *PLoS ONE*, **9**, e97279.

10 5. Consortium,H.M.P. (2012) Structure, function and diversity of the healthy human microbiome.
11 *Nature*, **486**, 207–14.

12 6. Huse,S.M., Ye,Y., Zhou,Y. and Fodor,A.A. (2012) A Core Human Microbiome as Viewed
13 through 16S rRNA Sequence Clusters. *PLoS ONE*, **7**, e34242.

14 7. Callahan,B.J., McMurdie,P.J., Rosen,M.J., Han,A.W., Johnson,A.J.A. and Holmes,S.P.
15 (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat.*
16 *Methods*, **13**, 581–583.

17 8. McMurdie,P.J. and Holmes,S. (2013) phyloseq: An R Package for Reproducible Interactive
18 Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, **8**, e61217.

19 9. Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K.,
20 Fierer,N., Peña,A.G., Goodrich,J.K., Gordon,J.I., et al. (2010) QIIME allows analysis of
21 high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

22 10. Edgar,R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon
23 reads. *Nat. Methods*, **10**, 996–998.

24 11. Meyer,F., Paarmann,D., D’Souza,M., Olson,R., Glass,E., Kubal,M., Paczian,T.,
25 Rodriguez,A., Stevens,R., Wilke,A., et al. (2008) The metagenomics RAST server – a
26 public resource for the automatic phylogenetic and functional analysis of metagenomes.
27 *BMC Bioinformatics*, **9**, 386.

28 12. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B.,
29 Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J., et al. (2009) Introducing
30 mothur: Open-Source, Platform-Independent, Community-Supported Software for
31 Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.*, **75**, 7537–
32 7541.

33 13. Dhariwal,A., Chong,J., Habib,S., King,I.L., Agellon,L.B. and Xia,J. (2017)
34 MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-
35 analysis of microbiome data. *Nucleic Acids Res.*, 10.1093/nar/gkx295.

36 14. Pollock,J., Glendinning,L., Wisedchanwet,T. and Watson,M. (2018) The Madness of
37 Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies.
38 *Appl. Environ. Microbiol.*, **84**.

39 15. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

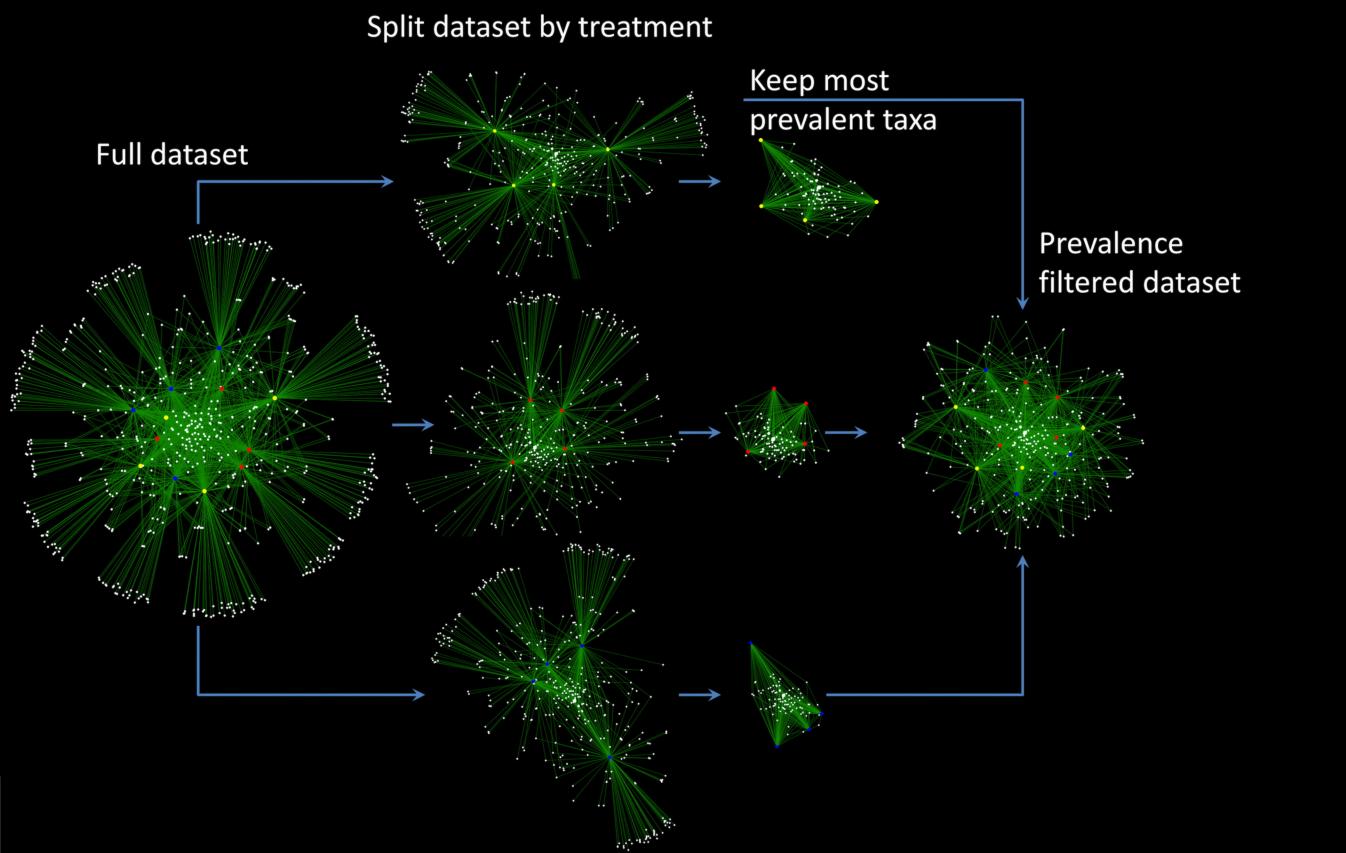
40 16. Davis-Richardson,A.G., Ardisson,A.N., Dias,R., Simell,V., Leonard,M.T.,
41 Kemppainen,K.M., Drew,J.C., Schatz,D., Atkinson,M.A., Kolaczkowski,B., et al. (2014)
42 Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at
43 high risk for type 1 diabetes. *Front. Microbiol.*, **5**.

44 17. Avershina,E., Slangsvold,S., Simpson,M.R., Storrø,O., Johnsen,R., Øien,T. and Rudi,K.
45 (2017) Diversity of vaginal microbiota increases by the time of labor onset. *Sci. Rep.*, **7**.

46 18. Serra-Majem,L., Ribas,L., Ngo,J., Ortega,R.M., García,A., Pérez-Rodrigo,C. and
47 Aranceta,J. (2004) Food, youth and the Mediterranean diet in Spain. Development of
48 KIDMED, Mediterranean Diet Quality Index in children and adolescents. *Public Health*
49 *Nutr.*, **7**.

1 19. Serra-Majem,L., Ribas,L., García,A., Pérez-Rodrigo,C. and Aranceta,J. (2003) Nutrient
2 adequacy and Mediterranean Diet in Spanish school children and adolescents. *Eur. J.*
3 *Clin. Nutr.*, **57**, S35–S39.
4 20. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and
5 Glöckner,F.O. (2012) The SILVA ribosomal RNA gene database project: improved data
6 processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
7 21. Lemos,L.N., Fulthorpe,R.R., Triplett,E.W. and Roesch,L.F.W. (2011) Rethinking microbial
8 diversity analysis in the high throughput sequencing era. *J. Microbiol. Methods*, **86**, 42–
9 51.
10 22. The Human Microbiome Project Consortium (2012) Structure, function and diversity of the
11 healthy human microbiome. *Nature*. 486:207–214.
12 23. Noordzij,M., Dekker,F.W., Zoccali,C. and Jager,K.J. (2010) Measures of Disease
13 Frequency: Prevalence and Incidence. *Nephron Clin. Pract.*, **115**, c17–c20.
14 24. Ward,M.M. (2013) Estimating Disease Prevalence and Incidence Using Administrative Data:
15 Some Assembly Required. *J. Rheumatol.*, **40**, 1241–1243.
16 25. Li,H. (2015) Microbiome, Metagenomics, and High-Dimensional Compositional Data
17 Analysis. *Annu. Rev. Stat. Its Appl.*, **2**, 73–94.

18



WORKFLOW

Input
OTU/ASV table
and
Metadata

`pime.oob.error()`
Measures the error in predicting microbiome differences based in the original dataset

`pime.split.by.variable()`
Splits original data into sample groups

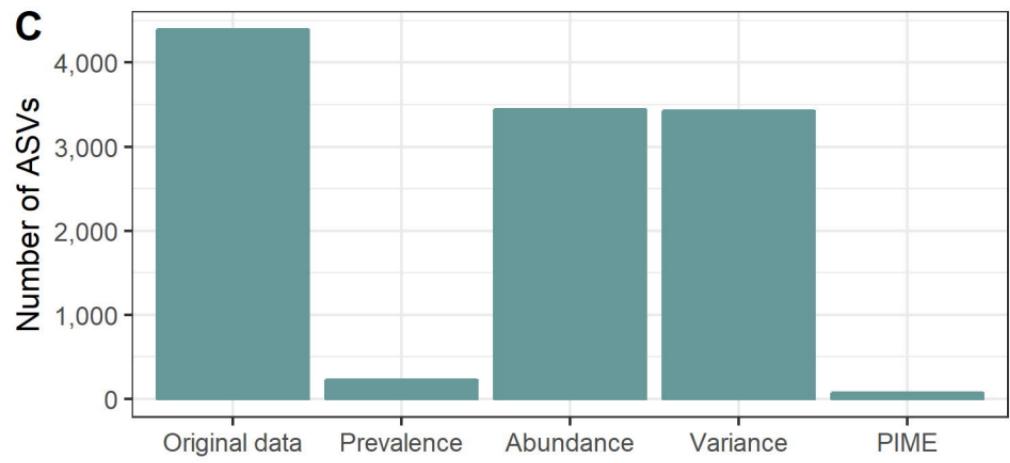
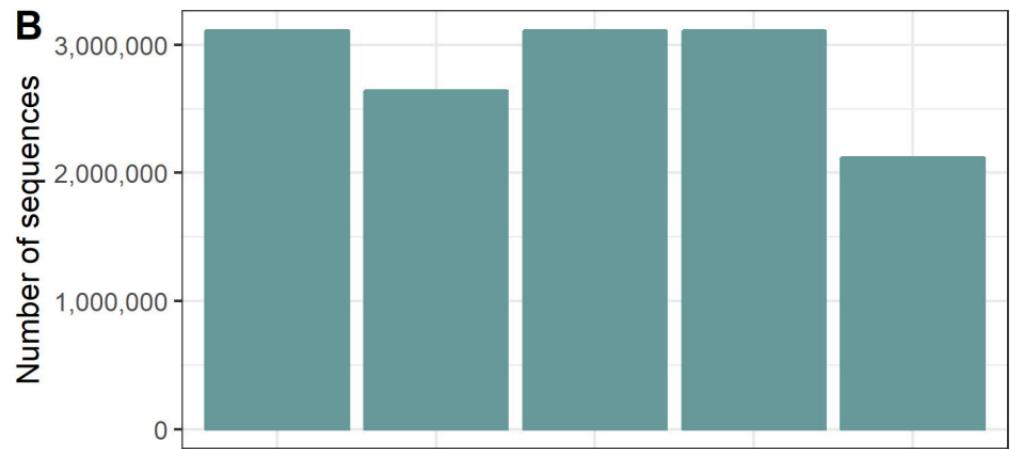
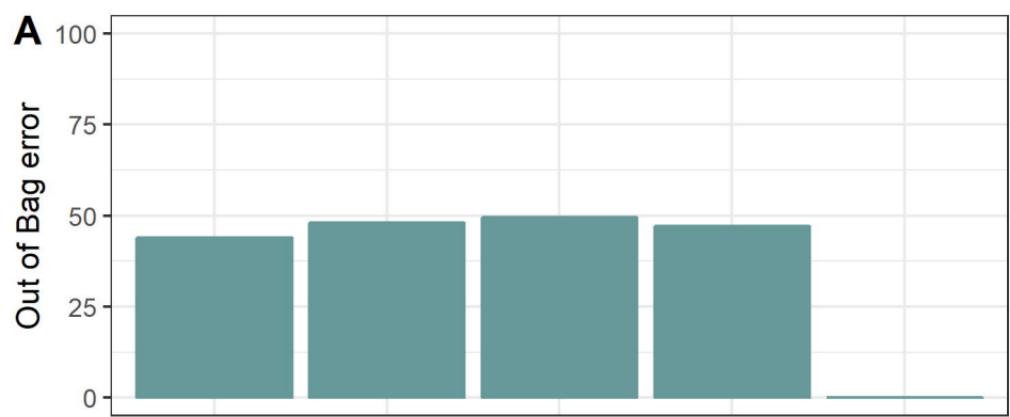
`pime.prevalence()`
Independently filters each sample group for prevalence intervals (5%, 10%,...), merging the resulting data by sample group and prevalence interval

`pime.best.prevalence()`
Builds random forests for classification, group prediction, and computes variable (OTU/ASV) importance, Mean Decrease Accuracy overall and for each variable

VALIDATION

`pime.oob.replicate()`
Computes Out of Bag error of all prevalence intervals (true data) n times, and returns a table with OOB error values and a boxplot.

`pime.error.prediction()`
For each prevalence interval, it randomizes sample groups n times. For each n it splits data into the random groups, computes prevalence, merge data and computes the Out of Bag error.



Out of Bag errors  Original Data  Randomized Data

