# Minimizing spurious features in 16S rRNA gene amplicon sequencing

Jing Wang, Qianpeng Zhang, Guojun Wu, Chenhong Zhang, Menghui Zhang* and Liping Zhao*

State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, Ministry of Education Key Laboratory of Systems Biomedicine, and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, P. R. China

*Address correspondence to Liping Zhao, lpzhao3517@126.com; Menghui Zhang, mhzhang@sjtu.edu.cn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26872v1 | CC BY 4.0 Open Access | rec: 19 Apr 2018, publ: 19 Apr 2018

## Abstract

The 16S rRNA gene amplicon sequencing is a widely used high-throughput method for the taxonomic inference in microbial communities. Many data analysis pipelines have been developed to enhance the accuracy in reflecting the real taxonomy, in order to better guide the downstream identification, isolation and mechanistic studies. Though rigorous quality filtration steps were adopted in these pipelines, with well-designed mock and simulated data sets, we found that there were still a widely divergent number of spurious features due to the "pseudo sequences" artificially generated during the PCR and sequencing process. These pseudo sequences were in low abundances, and were unreliable determined through a weighted re-sampling test. To minimize their influences on the characterization of taxonomy, we proposed an approach that contains two steps, an abundance filtering (AF) step and the subsequent AF-based OTU picking and remapping (AOR) step, which can efficiently decrease the spurious OTUs, sequences or oligotyping features, and improve Matthew's Correlation Coefficient (MCC) values in OTU clustering. The approach can be easily integrated with the popularly-used 16S rRNA sequencing data analysis pipelines, to make the number of OTUs, alpha and beta diversities from divergent pipelines more consistent with the real structure of microbial communities.


## Introduction

It is well known that the 97% similarity of 16S rRNA genes, which corresponds to the 70% DNA-DNA hybridization of whole genomes, is the primary criterion in molecular microbiology to define prokaryotic species (Stackebrandt & Goebel, 1994; Rosselló-Mora & Amann, 2001). Therefore clustering 16S rRNA gene amplicon sequences into OTUs with 97% similarity threshold has been extensively applied to reflect the phylogenetic delineation of microbial organisms at roughly the species level (Schloss & Handelsman, 2005; Goodrich et al., 2014). Although new OTU delineation

algorithms with flexible similarity threshold have been introduced (Kopylova et al., 2016), and other methods were also proved to have better sub-OTU resolution (Eren et al., 2013, 2015; Callahan et al., 2016; Amir et al., 2017), the principle of 16S rRNA gene amplicon sequencing data analysis remains the same that the characterization of features, such as OTUs, sequences or other units in microbial community samples should represent the real bacterial diversity in the community, and lead to the correct identification and isolation of functionally important bacteria for mechanistic studies. That is, spurious features should be minimized to avoid tracking down non-existent organisms.

However, when fed the same sequencing data set at the same 97% similarity cutoff, different delineation methods often produce widely divergent spurious OTUs (Bonder et al., 2012; Chen et al., 2013; Westcott & Schloss, 2015). For example, using the same dataset containing 43 known species, the number of OTUs varied from 133 to 4,397 among 10 different methods, overestimating by up to 2 orders of magnitude (Chen et al., 2013). The disparity among these methods has long been considered as a consequence of the distinct algorithms and parameters used (Westcott & Schloss, 2015; Schmidt, Matias Rodrigues & von Mering, 2015). However, by directly performing OTU delineation on high-quality sequences from 16S rRNA gene database, the number of OTUs obtained becomes more consistent and less overestimated (a median overestimation of three times compared to 33 times in actual sequencing data) (Chen et al., 2013). These results imply that the erroneous sequence introduced during actual PCR and sequencing process is the primarily influence worsening 16S gene amplicon sequencing analysis.

Substantial efforts have been made to improve the quality score-based filtration (Joshi & Fass, 2011; Bokulich et al., 2012; Edgar, 2013; Schirmer et al., 2015; Puente-Sánchez, Aguirre & Parro, 2016). Recent methods such as DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017) and UNOISE

(Edgar, 2016) apply denoising algorithm to provide putative error-free sequences. Another fine scale method, MED (Eren et al., 2015), performs oligotyping analysis (Eren et al., 2013) on informative nucleotide positions to ignore other noises. In this study, we constructed a series of simplified mock communities using clones of 16S rRNA genes sharing >3% dissimilarity in the V3-V4 region, in which case the difference is large enough that all clones should be correctly identified. We first used these data sets to evaluate several quality filtration pipelines to test if a combination of stringent methods could minimize the effect of pseudo sequences on OTU-based methods, including average linkage (AL) (Schloss & Westcott, 2011), UCLUST (Edgar, 2010), UPARSE (Edgar, 2013) and Swarm (Mahé et al., 2015). Afterwards, the non-OTU-based analysis methods, DADA2, Deblur and MED were applied to see whether and how they can overcome the influences from the pseudo sequences. Finally, we developed an approach containing abundance filtering (AF) and subsequent AF-based OTU picking and remapping (AOR) steps to minimize the spurious features from sequencing errors. The efficiency of the approach is further validated with more complex simulated or real-world communities.

# Materials and Methods

## Construction of mock communities

A total of 22 16S rRNA gene clones were chosen to construct 7 mock communities, each with varying clone compositions (Table S1). Each community had 3 replicates in the same sequencing run (run1, a total of 21 samples). Four communities were sequenced in 2 additional runs (run2 and run3, a total of 12 samples each).

## Sequencing procedures

Hypervariable region V3-V4 amplicons from the 16S rRNA gene were sequenced by Illumina MiSeq, as described in http://res.illumina.com/documents/products/appnotes/16s-metagenomic-library-prep-guide.pdf, with the following modifications. Platinum Pfx DNA polymerase (C11708021, Invitrogen, USA) was used for two steps during the amplification. PCR cycles for the amplicon PCR (amplification of the 16S rRNA V3-V4 region) were reduced to 21 to diminish PCR bias. The primers used were as follows: S-D-Bact-0341-b-S-17, 5'-CCTACGGGNGGCWGCAG-3' and S-D-Bact-0785-a-A-21, 5'-GACTACHVGGGTATCTAATCC-3' (Klindworth et al., 2013). The amplicons were sequenced using 2*300 bp paired-end sequencing.

**Quality control methods**

Quality control of raw sequences was performed using UPARSE (Edgar, 2013) with USEARCH v8.0.1623, mothur (Schloss et al., 2009) v1.35.0, moira (Puente-Sánchez, Aguirre & Parro, 2016) v1.1.0 or a workflow (Schirmer et al., 2015) including quality trimming (Sickle (Joshi & Fass, 2011) v1.33), error correction (BayesHammer (Nikolenko, Korobeynikov & Alekseyev, 2013) with SPAdes v3.5.0) and read overlapping (PANDAseq (Masella et al., 2012) v2.8) (aliased as S+BH+P). Overlaps with ≥50 bp lengths were required for each sequence pair, resulting in ≥400 bp merged sequences, and no ambiguous bases were allowed. USEARCH further filtered out sequences with ≥0.5 expected errors. The PCR primers were then truncated from the QC sequences using the "search_pcr" command in USEARCH.

**Obtaining simulated datasets**

To achieve more complex data sets for testing, we used Grinder (Angly et al., 2012) to simulate the sequencing reads based on 87 randomly picked OTU references from Greengenes with <97% similarity

to each other (Table S2). The distribution of Illumina sequencing errors was simulated with a fourth degree polynomial model (Korbel et al., 2009) as follows:

$$3\times10^{-3}+3.3\times10^{-10}i^{4} \quad (1)$$

wherein $i$ indicates the position alongside the sequences, the coefficients were adjusted to fit the profile of 300 bp sequencing platform.

Among the errors, the ratio of substitutions vs. insertions/deletions was set as 9:1. The portion of chimeras was designed as 10%, with the distribution of bimeras, trimeras quadrameras was 314:38:1 (Quince et al., 2011). A total of 99 samples were simulated, each had 15,000 paired-end reads with 2*300 bp length. The abundances of the 87 references were shuffled across samples based on power law distribution.

**Obtaining real datasets**

PWS data: a published data set containing 110 human fecal samples collected from children diagnosed with Prader–Willi syndrome or simple obesity during dietary intervention. The V3-V4 hypervariable region was sequenced on an Illumina MiSeq machine using 2*300 bp paired-end sequencing (Zhang et al., 2015). Sequences are available at http://www.ncbi.nlm.nih.gov/bioproject/PRJNA306596.

Ultra data: a downloaded data set including microbial communities from host-associated and free-living environments, sequencing the V4 region with 150 bp single-end (Caporaso et al., 2012). Sequences are available at https://qiita.ucsd.edu/study/description/1684.

Water data: a downloaded data set that were collected from drinking water systems in the Netherlands, spanning the V4 region with a 2*200 bp read length (Roeselers et al., 2015). Sequences are available at the European Nucleotide Archive under accession number PRJEB7435.

River data: a downloaded data set containing water samples along the midstream of the Danube River, applying the V3-V4 region for 2*250 bp sequencing (Savio et al., 2015). The raw sequencing data were submitted to the NCBI Sequence Read Archive under accession number SRP045083.

**Preparation of qualified sequences for downstream analysis**

Sequence merging, error correction and quality control (QC) were performed using moira v1.1.0. The PCR primers were truncated from the QC sequences afterwards. The sequence lengths were restricted to >100 bp for V4 amplicons and >400 bp for V3-V4 amplicons. The QC sequences were de-replicated into unique sequences and aligned to the SILVA bacteria reference database (Quast et al., 2013) with the “align.seqs” command in mothur. The alignment space was optimized by removing the sequences that failed to align correctly. This optimization is to ensure that all the remaining sequences overlapped at the same region of the SILVA reference alignment. The sequences were then divided by samples and checked for chimeras using abundant sequences as references with the UCHIME (Edgar et al., 2011) *de novo* algorithm. Non-chimeric sequences were classified according to the mothur-formatted version of the RDP classifier training set v9 (Cole et al., 2014), and non-bacterial sequences were further filtered out. The final qualified sequences were rarefied to an even number per sample to avoid the bias of unbalanced sequencing effort (10,000 per sample for Mock, Simulated, PWS and Ultra data, 20,000 per sample for Water data, and 1,000 per sample for River data). The size or abundance of a qualified unique sequence was defined as the number of duplicates after rarefaction.

DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017) implemented in QIIME2 (Caporaso et al., 2010) require the quality score (Phred Q score) as part of the inputs. Therefore a different process using QIIME2 framework and command line interface was performed. First of all, PCR primers were truncated with Cutadapt (Martin, 2011). Then in DADA2, forward and reverse reads were respectively truncated to the first 270 and 200 bp high-quality region. For Deblur, paired-end reads were merged to have a length between 400 and 500 bp with VSEARCH (Rognes et al., 2016), followed by the "quality-filter" command (Bokulich et al., 2012) called in QIIME2 with default parameters. The final sequences were truncated to obtain the first 400 bp region by Deblur itself.

**OTU delineation**

UPARSE: qualified unique sequences were sorted by decreasing abundance, and singletons were discarded. Non-chimeric OTU representative sequences were selected afterwards with a 97% similarity threshold. The OTU table was finalized by mapping qualified sequences to the obtained OTUs with the USEARCH (Edgar, 2010) global alignment algorithm.

Average linkage (AL): qualified sequences were pre-clustered with up to one difference per 100 bp length. OTUs were then delineated by >97% similarity with an average neighbor algorithm by mothur.

UCLUST: qualified sequences were clustered into *de novo* OTUs by >97% similarity using UCLUST within the QIIME pipeline.

Swarm: qualified sequences were grouped together as an OTU with 1-base-difference connections. Large OTUs with multiple abundant cores were broken down. Nearby low-abundance sequences were connected through fastidious option. The boundary of each OTU is flexible depending on the distribution of sequences. There is no fixed similarity threshold.

UPARSE, UCLUST and Swarm chose the most abundant sequence in each OTU as representative sequence, whereas AL chose the sequence with the smallest maximum distance to the other sequences within the same OTU.

**Abundance filtering (AF) of unique sequences**

Weighted bootstrap resampling was performed 1,000 times with replacement using the original abundance of the unique sequences ($Abund_{real}$) as weights. The confidence intervals were adjusted according to Meyer *et al.* (Meyer et al., 2016). The estimated abundance of each unique sequence ($Abund_{adj}$) was calculated as follows:

$$Abund_{adj} = 2 \times Abund_{real} - Abund_{boot} \quad (2)$$

where $Abund_{boot}$ indicates the mean of the bootstrapped abundances obtained from the corresponding 1,000 replicates.

The 99% confidence interval for each unique sequence could then be obtained as follows:

$$CI_{99\%} = [Abund_{adj} - (Abund_{boot} - Abund_{0.5}); Abund_{adj} + (Abund_{99.5} - Abund_{boot})] \quad (3)$$

where the $Abund_{0.5}$ and $Abund_{99.5}$ values represent the $0.5^{th}$ and $99.5^{th}$ percentiles of the 1,000 replicates. A unique sequence was considered as unreliable when its lower bound of $CI_{99\%}$ dropped below zero, then was filtered out.

The custom R script (resample_uniques_ci.r) used to perform this bootstrapping approach is available in the supplementary information.

**AF-based OTU picking and remapping (AOR)**

We propose an AOR approach to modify the current OTU delineation pipelines, as follows (Fig. 1):

(i) Filter out unreliable sequences determined with AF. This step can be performed using the "sortbysize" command within USEARCH or the "split.abundance" command within mothur.

(ii) Input the remaining sequences into the initial OTU delineation step.

(iii) For the OTU delineation methods that depend on similarity threshold, remap the filtered sequences in (i) to the obtained OTUs if they match the same similarity threshold with global alignment methods. This step can be performed using USEARCH global alignment, the "align.seqs" command within mothur or the "pick_closed_reference_otus.py" pipeline within QIIME.

**OTU clustering quality assessment**

Matthew's correlation coefficient (MCC) (Matthews, 1975) was calculated following the description of Schloss *et al.* (Westcott & Schloss, 2015). We counted the number of sequence pairs that had ≥97% similarity and were in the same OTUs as true positives (TPs), those that had <97% similarity and were in different OTUs as true negatives (TNs), those that had ≥97% similarity and were in different OTUs as false negatives (FNs) and those that had <97% similarity and were in the same OTU as false positives (FPs). The MCC was then calculated as follows:

$$MCC=\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

MCC is not applicable to Swarm as the OTU boundary is not based on the 97% similarity threshold.

**Software**

This work used QIIME (v1.9.1, v2-2018.2) (Caporaso et al., 2010), mothur (v1.35.0) (Schloss et al., 2009), USEARCH (v8.0.1623) (Edgar, 2010), DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017), MED (in Oligotyping Pipeline v2.1) (Eren et al., 2015), Swarm (v2.2.2) (Mahé et al., 2015), Grinder (v0.5.4) (Angly et al., 2012), moira (v1.1.0) (Puente-Sánchez, Aguirre & Parro, 2016), Sickle (v1.33) (Joshi & Fass, 2011), BayesHammer (Nikolenko, Korobeynikov & Alekseyev, 2013), (in SPAdes v3.5.0), PANDAseq (v2.8) (Masella et al., 2012) and R (v3.2.0) (R Core Development Team & R Core Team, 2015). The Mantel test (Mantel, 1967) was performed using the vegan (v2.3-4) (Oksanen et al., 2016) package in R. Parallel computing was performed with GNU Parallel (Tange, 2011).

# Results

**"Pseudo sequences" with significant sequencing errors remain in the dataset after quality filtration**

On average, 15,080 (12,780-17,460) (median, minimum-maximum), 16,770 (13,060-18,520) and 32,510 (26,240-35,050) sequences per sample were achieved from three independent MiSeq runs on Mock data. Four quality control pipelines were individually applied, including UPARSE (with USEARCH v8.0.1623), mothur v1.35.0, moira (Puente-Sánchez, Aguirre & Parro, 2016) v1.1.0, and a combination of Sickle (Joshi & Fass, 2011) v1.33, BayesHammer (Nikolenko, Korobeynikov & Alekseyev, 2013) (in SPAdes v3.5.0) and PANDAseq (Masella et al., 2012) v2.8 (aliased as S+BH+P, introduced by Schirmer *et al.* (Schirmer et al., 2015)). After further truncation of PCR primers, the retained quality controlled (QC) sequences were aligned to mock references using global alignment with the "align.seqs" command in mothur. We did not choose the "seq.error" command in mothur because it tended to align sequences to multiple templates to achieve lower error rates, although only a small part were actual chimeras. As reported by the Illumina Sequencing Analysis Viewer on the sequencing platform, the raw sequences of the three runs yielded 2.5%, 2.1% and 3.9% errors during the sequencing procedure. These error rates were reduced to less than 0.5% after applying quality controls (Table S3). Chimeras, contaminants and other errors were not filtered out yet at this step.

For the same mock community, the absolute quantities of the QC sequences varied largely among the different sequencing runs and filtration methods (Table S4), although the error distributions of the qualified sequences were similar among the four methods. S+BH+P was the least robust and obtained the fewest QC sequences. Moira maintained the highest number of sequences with a moderate error rate, due to its denoising algorithm; therefore, it was chosen as the uniform quality control method in this study. We did not follow the default pipelines of QIIME, UPARSE or mothur to process the sequences because different quality filtration pipelines had inconsistent "qualified sequences." The purpose of this study was to focus on the data analysis step, and thus, we hypothesized that it would be better to begin with the same baseline data.

The QC sequences obtained by moira were further examined for chimeras, non-bacterial sequences and sequences that started or ended at the wrong position. Sequences that failed to align to mock references but showed high similarity (>97%) to species in the SILVA bacteria database were defined as contaminants and discarded. Retained sequences were rarefied to the same number per sample and de-replicated into qualified unique sequences. As a result, the error rates were further reduced to less than 0.2% (Table S3), and all errors with known sources were eliminated. These qualified unique sequences were used as the input data for downstream analyses.

We then performed global alignment with the "align.seqs" command in mothur to compare the qualified sequences with the mock references. The three sequencing runs contained 75.7%, 60.9% and 51.4% of qualified sequences that were 100% identical to the mock references. In addition, 99.9%, 99.5% and 99.4% of the corresponding qualified sequences shared 97% or higher identity with the closest mock reference. However, up to 0.6% of qualified sequences had more than 3% errors, and some showed less than 90% identity to the closest mock reference. This small amount of pseudo sequences contributed to 229, 615 and 744 unique sequences in each sequencing run.

These pseudo sequences had a relatively lower abundance (Fig. 2a). In general, a lower relative abundance was associated with a higher number of different unique sequences, forming an L-shaped distribution curve (Fig. 2b). More than 90% of the unique sequences had a relative abundance <0.01%.

**Extra number of spurious OTUs is primarily derived from the pseudo sequences**

The V3-V4 regions of the 22 reference clones used to construct the mock communities shared <97% similarity (see Table S1 for detailed sequence contents). These communities were designed to ensure that UCLUST (with QIIME v1.9.1), average linkage (AL, with mothur v1.35.0) and UPARSE (with USEARCH v8.0.1623) and Swarm v2.2.2 would not cluster any two of the mock references together. Therefore, within this mock data set, one outcome OTU should be expected for one species. This design makes the guaranty that, in this case, the inconsistent algorithms and parameters would not perturb the downstream results.

However, none of the OTU delineation methods could provide expected results on actual sequencing data (Table 1). We defined three kinds of OTU as "perfect" (representative sequence was 100% identical to mock references), "good" (97%≤identity<100%) or "spurious" (identity<97%). All methods got 22 "perfect" OTUs, showing one-to-one correspondence with 22 "real" species. However, UPARSE, UCLUST, AL and Swarm also obtained 1 (0-1), 308 (154-326), 308 (155-328) and 456 (204-486) spurious OTUs, respectively. The overestimation of OTU numbers were mainly from spurious OTUs that representing non-existent species.

We then traced the unique sequences back to their assigned OTU types (Fig. 3). The dots in the diagram are used to represent the unique sequences, and the ellipses and links indicate how the unique sequences were clustered into the OTUs. The majority of unique sequences were clustered with their corresponding species (green and blue clusters), while a few low-abundance unique sequences whose similarity <97% to the references formed "error clouds." As shown, a OTU delineation algorithm needs to create extra spurious OTUs (red clusters) to fully cover these pseudo sequences if they are distant enough to form an independent "error" cluster (Fig. 3e). Even worse, they could be clustered with other perfect and good sequences to form non-perfect OTUs (Fig. 3f) or, conversely, good sequences could

be trapped to become spurious OTUs (Fig. 3g). UPARSE discarded singletons (unique sequences without replicates) and stringently checked for potential chimeras once more during OTU delineation, thereby distinctly reducing the number of retained low-identity pseudo unique sequences. However, discarding only singletons was not sufficient, as the non-singleton pseudo unique sequences remained and became sources of spurious OTUs.

**Abundance filtering (AF) approach minimizes the spurious OTUs**

The results from the mock data demonstrate that the unique sequences with relatively low abundances are the major sources of pseudo sequences and spurious OTUs. Assuming that the errors occur randomly, the sequences with more errors are less likely to have replicates with exactly the same errors by chance, *i.e.*, sequences with more errors are expected to have relatively low abundances. We propose AF and the subsequent AF-based OTU picking and remapping (AOR) approach to modify the current analysis pipelines.

The determination of unreliable sequences is critical in AF. Among the three replicated sequencing runs of mock communities, which contained 22,844, 26,814 and 33,109 unique sequences, only 5,126 unique sequences were consistently detected. Considering the robustness and reproducibility of the sequencing data, a threshold should be able to separate the unreliable sequences that fail to consistently appear in technical replicates. We applied a bootstrapping strategy to estimate the uncertainty level of the unique sequences in microbial communities. The 99% confidence interval of bootstrapped abundances and the corresponding coefficient of variation (CV, calculated as the bootstrapped standard error of each sequence divided by its observed abundance) were then estimated (Fig. 4).

The 99% confidence interval of sequences touched zero when their abundances were ≤6 (Fig. 4a-c). This means that although these sequences appeared in the original sequencing data, they were detected by chance and may not occur when the same communities are sequenced again. When the threshold for the unique sequences in the three sequencing runs was set at 7, their corresponding relative abundances were 0.003%, 0.005% and 0.005% (Table S5). For sequences with abundances below the threshold, the corresponding CV values were >50% (Fig. 4d-f), indicating that among the replicated sequencing runs, the abundances of these sequences vary substantially. The unreliability of these low abundant sequences implied they were below the detection limit of the current sequencing technology.

AOR can be summarized as a mixed *de novo*/reference-based approach. Unreliable sequences are filtered out by AF. Then *de novo* clustering is performed on the remaining sequences. Finally a reference-based clustering method remaps all sequences onto the OTUs obtained during the *de novo* step. After AF step, all 97%-similarity-threshold-based methods combined with the AOR step provided 22 OTUs with one-to-one correspondence to the real species in mock communities (Fig. S1a-c, Table 1); less than 1% of total sequences were eventually discarded (Fig. S1d-f), and MCC values had already achieved 0.99 (Fig. S1f-i). Swarm does not apply a fixed similarity threshold during OTU delineation, thus the remapping procedure and the calculation of MCC values are not available. Nevertheless, filtering out low-abundance sequences dramatically improved the accuracy of Swarm's OTU results, with only one spurious OTU left in one of the three sequencing runs. These results suggest that the abundance threshold determined by the above statistical strategy was qualified to detect most of the pseudo sequences and maintain the desired OTUs belonging to the expected real species. Indeed, the AOR approach improved the quality of OTU delineation.

**AF also improves the accuracy of Non-OTU-based methods**

DADA2 and Deblur perform denoising procedure to obtain high-quality unique sequences, while MED focuses on a subset of informative nucleotide positions along the sequences to ignore the random noises. These methods were reported to provide better resolution of microbial communities than OTU-based methods (Eren et al., 2015; Callahan et al., 2016; Amir et al., 2017). In this study, we tested these methods with our Mock data sets to see if they can be affected by the pseudo sequences as well (Table 2). The differences between reference sequences would also be large enough to be correctly identified by the three non-OTU-based methods.

DADA2 obtained 41 (41-42) perfect unique sequences that were 100% identical to the Mock references. The number was higher than 22 actual species because after paired-end merging, DADA2 still maintained single-end sequences that failed to be merged but were identical to references. In addition, 1 (1-3) spurious sequence was observed. After setting an abundance threshold at 7 in AF step, all spurious sequences in run1 were discarded, yet run2 and run3 still obtained one spurious sequences whose abundance was 12.

By default in each sample, Deblur discards the unique sequences whose abundance less than 2. Afterwards, it further discards the unique sequences whose total abundance is less than 10 across all samples. With this default behavior, exactly 22 perfect sequences were identified. Once these low-abundance sequences were maintained, additionally 2 (2-7) good sequences as well as 1 (1-3) spurious sequences were observed. When 7 was set to replace the default abundance threshold, none of these good or spurious sequences existed anymore.

MED obtained 306 (150-313) spurious oligotyping features with default settings. By filtering out the sequences whose abundance was less than 7, the output good and spurious features were reduced to 1 (0-1).

**AF and AOR is effective in more complex Simulated data sets**

We further applied a series of Simulated data to increase the complexity while still being aware of the actual composition. A total of 99 samples containing the same 87 reference species with variant compositions were simulated, each produced 10,000 qualified sequences. Abundance threshold was set to 7 based on bootstrapping strategy.

Similar to the results in mock data, the OTU-based methods UCLUST, AL and Swarm obtained at least one magnitude more spurious OTUs (1577, 1566 and 2079) than actual 87 references. By implementing AOR or AF approach, the number of spurious OTUs could be reduced to 206, 202 and 218 (Table 3). UPARSE were not significantly affected by AOR approach in this data.

AF also decreased the number of spurious sequences for DADA2 and Deblur. The most dramatic improvement was observed in MED results. A total of 7312 oligotyping features were identified with all sequences. Since this number exceeded the hard limit of the maximum number of open file descriptors (1024) on our computer server, we could not obtain detailed results by MED. However, by simply filtering out low-abundance sequences, only 185 spurious features were retained (Table 4).

**The AOR approach produces consistent alpha and beta diversity in real data sets**

We used four published real data sets to further evaluate our AOR approach on the three 97%-similarity based OTU delineation methods, UCLUST, AL and UPARSE. The four real datasets, PWS, Ultra, River and Water, contain 248,654, 25,544, 45,834 and 147,778 qualified unique sequences, respectively. Although it is not possible to obtain the sequencing error information for the real datasets, similar CV values and confidence interval distributions of the unique sequences were observed in all four datasets (Fig. S2). Incorporating the AOR approach with different pipelines and changing the relative abundance thresholds allowed us to obtain a series of OTU delineation results for each dataset (Fig. S3). All results showed dramatic decreases at the beginning and maintained slow descending tendencies as more sequences were set aside from the *de novo* OTU delineation step. Different methods implementing distinct algorithms showed divergent behaviors; however, they all obtained similar numbers of OTUs after identifying the unreliable sequences, whose abundances were no more than 6 (0.0006% in relative abundance), 7 (0.003%), 6 (0.0004%) and 6 (0.007%) for PWS, Ultra, River and Water, respectively (Fig. S3, Table. S5). At these levels, at least 95% of the qualified sequences could be remapped to pre-defined OTUs, except for the River dataset, which remapped 85% of the sequences. The MCC values were also higher than the original values (Fig. S4).

In the alpha diversity comparison, the number of OTUs and Chao1 (Chao et al., 2000), Shannon (Shannon, 1948) and Simpson (Simpson, 1949) indices of each sample were calculated (Fig. S5). The first two indices directly reflect the richness of the sample, and the latter two reflect the overall diversity. Because of the great disparities in total OTU numbers, significant differences occurred between the original results and the AOR results with respect to the estimation of OTU numbers and Chao1 indices. However, Shannon and Simpson indices were not significantly reduced by AOR, indicating that the overall diversities of communities are not underestimated using the AOR approach.

Moreover, in the original results, different OTU delineation methods provided significantly divergent alpha diversities (multiple Wilcoxon test, FDR-adjusted $p<0.01$). After integration with AOR, the divergences among the methods were no longer significant because the OTU delineation was no longer affected by sequencing errors, and the different methods were all able to reflect the same real community composition.

Four types of beta diversity distance matrices, namely, the Euclidean (EU), Bray-Curtis (BC) (Bray & Curtis, 1957), weighted normalized UniFrac (WU) and unweighted UniFrac (UU) (Lozupone et al., 2011) distances, were measured, and the results obtained by different methods were compared by the Mantel test (Mantel, 1967). AOR showed an improvement in beta diversity consistency among the different OTU delineation methods (Fig. S6), validating that our AOR approach not only simply decreases the number of OTUs but also provides much more consistent profiling of compositions approaching the real communities, which would no longer be affected by the choice of OTU delineation method.

## Discussion

We developed AF and AOR approach to minimize the spurious features produced by either OTU-based or non-OTU-based algorithms from low-quality “pseudo sequences” introduced by errors that are resistant to current quality filtration processes. These pseudo sequences, which had >3% divergence from the reference sequences, remained after current mainstream pipelines implementing error correction, denoising and stringent filtration of chimeric sequences, contaminants and non-bacterial contents. Although the overall abundance of these pseudo sequences was low (<1% of the total qualified sequences passing quality filtration), introducing them into analysis increased the total

number of features to 10 times higher than expected and enlarged the divergence of the alpha and beta diversity analyses among the different methods. By filtering out these pseudo sequences, our AF and AOR approach further diminished unexpected spurious features both in mock and simulated communities. When incorporated in OTU-based methods, AOR approach also provided higher MCC values of clustering quality and resulted in more consistent alpha and beta diversities among the different methods with real data sets (see supplementary).

Lower-abundance and lower-quality sequences were observed to surround higher-abundance, biologically real sequences, forming "error clouds" (Bokulich et al., 2012; Edgar, 2013). Various researchers have developed different approaches to remove these pseudo sequences. Chen *et al.* discarded all sequences whose abundance was <100 in 454 sequencing data despite their accuracies (Chen et al., 2013), which resulted in the loss of many low-abundance but high-quality sequences. Bokulich *et al.* removed lower-abundance OTUs with a relative abundance <0.005% before downstream analyses (Bokulich et al., 2012); however, this strategy led to the risk of abandoning good sequences trapped in these OTUs. Edgar set aside singletons during OTU delineation by UPARSE to prevent them from becoming the centroids of OTUs and then remapped them to the defined OTUs (Edgar, 2013). This strategy improves the accuracy of OTU delineation, but the results in our study indicate that singletons are not the only source of pseudo sequences. The unreliability of low-abundance sequences has been noticed by new non-OTU-based methods as well. By default, Deblur requires the putative error-free sequences to have an abundance no less than 2 in each sample and no less than 10 across all samples (Amir et al., 2017). MED recommended a filtration based on the count of the most abundant sequence in each oligotyping feature. The threshold was set to the average sequence number per sample divided by 1,000 in the first oligotyping paper (Eren et al., 2013), then was changed to the total number of sequences divided by 10,000 in the MED paper (Eren et al., 2015).

However, such methods just take into account the sensitivity-vs.-error trade-offs but lack of the basic detection limit concept of metrology. They may provide ideal number of OTUs in some cases, but they are difficult to reproduce or generalize when sequencing data is generated from varied choices of primers, sequencing lengths and depths (Tremblay et al., 2015).

Although microbiologists have raised the concerns about low-abundance sequences and the so called "rare biosphere" for a long time (Huse et al., 2010; Kunin et al., 2010), the concept "detection limit" has not been introduced into this area ever before. In this study, the low abundance threshold was set based on the concept that real sequences should consistently appear in repeated observations (Zhou et al., 2011). We performed a weighted bootstrap resampling strategy based on the observed abundance distribution to estimate the occurrence and abundance of each unique sequence in replicated sequencing runs. This approach makes use of the lower detection limit of the sequencing protocol by indicating that the rare sequences below the threshold are statistically unreliable because they cannot be consistently detected across observations. thus limiting robustness and reproducibility. Based on the basic detection limit concept of trace and metrologic analysis (Analytical Methods Committee, 1987), the sequences below detection limit are actually not detectable. It is out of confidence to make any conclusions based on their stochastic occurrences and abundances. If very rare species are of interest, a deeper sequencing depth is required to ensure that they are covered with confidence.

Our AOR approach takes advantages of both *de novo* and reference-based OTU delineation methods. By performing *de novo* clustering on the reliable sequences only, the resulting OTUs are ensured to represent the real species in the query communities. This approach outperforms the pre-clustered OTU references based on large databases such as Greengenes (DeSantis et al., 2006) or SILVA (Quast et al., 2013), as some novel species may not yet have been collected by them. The subsequent remapping step

ensures that the remaining low-abundance sequences can be maximally rescued back once they adhere to similarity criteria of the obtained OTUs rather than being arbitrarily abandoned due to their relatively low abundance.

A simple universal threshold for removing unreliable sequences, as the ones provided by previous publications, is admirable in application. In our study, the relative abundance threshold varied from 0.0006% to 0.005% depending on the total number of sequences, which implies that relative abundance level is not an ideal criterion. However, for most data sets in this study, the absolute count of unreliable sequences were no more than 6, which suggests that an absolute count threshold might be set as ≥7. Meanwhile, we still recommend to use our bootstrap re-sampling script to find out the exact criterion with statistical confidence in sequences denoising.

Our approach can reduce the risk of observing distorted microbial community structures with spurious features representing non-existent species. It can be easily integrated with the current mainstream pipelines and may be of potential use in various microbiome-wide association studies.

## Authors' contributions statement

JW, LZ and MZ designed the study; JW, QZ and CZ performed the experiments; MZ provided the analytical tools; and JW, GW, LZ and MZ wrote the paper.

## Funding

This work was supported by grants from the National Science and Technology Major Project of China (2012ZX10005001-009), the National Natural Science Foundation of China (31330005, 30730005, 81401141 and 20875061), and the Science and Technology Commission of Shanghai Municipality (14YF1402200).

## Accession codes

The mock communities and PWS datasets supporting the conclusions of this article are available in the NCBI Short Read Archive repository under BioProject PRJNA306596 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA306596).

## Competing financial interests

The authors have declared that no competing interests exist.

## References

Amir A., McDonald D., Navas-Molina JA., Kopylova E., Morton JT., Zech Xu Z., Kightley EP., Thompson LR., Hyde ER., Gonzalez A., Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* 2.

Analytical Methods Committee. 1987. Recommendations for the definition, estimation and use of the detection limit. *The Analyst* 112:199–204. DOI: 10.1039/an9871200199.

Angly FE., Willner D., Rohwer F., Hugenholtz P., Tyson GW. 2012. Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research* 40:e94. DOI: 10.1093/nar/gks251.

Bokulich NA., Subramanian S., Faith JJ., Gevers D., Gordon JI., Knight R., Mills DA., Caporaso JG. 2012. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* 10:57–59. DOI: 10.1038/nmeth.2276.

Bonder MJ., Abeln S., Zaura E., Brandt BW. 2012. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28:2891–2897. DOI: 10.1093/bioinformatics/bts552.

Bray RJ., Curtis JT. 1957. An ordination of the upland forest communities of southern Winsconin. *Ecological Monographs* 27:325–349.

Callahan BJ., McMurdie PJ., Rosen MJ., Han AW., Johnson AJA., Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13:581.

Caporaso JG., Kuczynski J., Stombaugh J., Bittinger K., Bushman FD., Costello EK., Fierer N., Pena AG., Goodrich JK., Gordon JI., Huttley GA., Kelley ST., Knights D., Koenig JE., Ley RE., Lozupone CA., McDonald D., Muegge BD., Pirrung M., Reeder J., Sevinsky JR., Turnbaugh PJ., Walters WA., Widmann J., Yatsunenko T., Zaneveld J., Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 7:335–336.

Caporaso JG., Lauber CL., Walters W a., Berg-Lyons D., Huntley J., Fierer N., Owens SM., Betley J., Fraser L., Bauer M., Gormley N., Gilbert J a., Smith G., Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* 6:1621–1624. DOI: 10.1038/ismej.2012.8.

Chao A., Hwang W., Chen Y., Kuo C. 2000. Estimating the Number of Shared Species. *Statistica sinica* 10:227–246.

Chen W., Zhang CK., Cheng Y., Zhang S., Zhao H. 2013. A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLoS ONE* 8:e70837. DOI: 10.1371/journal.pone.0070837.

Cole JR., Wang Q., Fish JA., Chai B., McGarrell DM., Sun Y., Brown CT., Porras-Alfaro A., Kuske CR., Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42:D633–D642. DOI: 10.1093/nar/gkt1244.

DeSantis TZ., Hugenholtz P., Larsen N., Rojas M., Brodie EL., Keller K., Huber T., Dalevi D., Hu P., Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* 72:5069–5072. DOI: 10.1128/AEM.03006-05.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461. DOI: 10.1093/bioinformatics/btq461.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10:996–998. DOI: 10.1038/nmeth.2604.

Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*.

Edgar RC., Haas BJ., Clemente JC., Quince C., Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200. DOI: 10.1093/bioinformatics/btr381.

Eren AM., Maignien L., Sul WJ., Murphy LG., Grim SL., Morrison HG., Sogin ML. 2013. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution* 4:1111–1119. DOI: 10.1111/2041-210X.12114.

570 Eren AM., Morrison HG., Lescault PJ., Reveillaud J., Vineis JH., Sogin ML. 2015. Minimum entropy
571 decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker
572 gene sequences. *ISME Journal* 9:968–979. DOI: 10.1038/ismej.2014.195.

573 Goodrich JK., Di Rienzi SC., Poole AC., Koren O., Walters WA., Caporaso JG., Knight R., Ley RE.
574 2014. Conducting a microbiome study. *Cell* 158:250–262. DOI: 10.1016/j.cell.2014.06.037.

575 Huse SM., Welch DM., Morrison HG., Sogin ML. 2010. Ironing out the wrinkles in the rare biosphere
576 through improved OTU clustering. *Environmental Microbiology* 12:1889–1898. DOI:
577 10.1111/j.1462-2920.2010.02193.x.

578 Joshi N., Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files
579 (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle.*:2011.

580 Klindworth A., Pruesse E., Schweer T., Peplies J., Quast C., Horn M., Glöckner FO. 2013. Evaluation
581 of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-
582 based diversity studies. *Nucleic Acids Research* 41:e1. DOI: 10.1093/nar/gks808.

583 Kopylova E., Navas-Molina JA., Mercier C., Xu ZZ., Mahé F., He Y., Zhou H-W., Rognes T., Caporaso
584 JG., Knight R. 2016. Open-Source Sequence Clustering Methods Improve the State Of the Art.
585 *mSystems* 1:e00003-15. DOI: 10.1128/mSystems.00003-15.

586 Korbel JO., Abyzov A., Mu XJ., Carriero N., Cayting P., Zhang Z., Snyder M., Gerstein MB. 2009.
587 PEMer: a computational framework with simulation-based error models for inferring genomic
588 structural variants from massive paired-end sequencing data. *Genome biology* 10:R23. DOI:
589 10.1186/gb-2009-10-2-r23.

590 Kunin V., Engelbrektson A., Ochman H., Hugenholtz P. 2010. Wrinkles in the rare biosphere:
591 Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental*
592 *Microbiology* 12:118–123. DOI: 10.1111/j.1462-2920.2009.02051.x.

593 Lozupone C., Lladser ME., Knights D., Stombaugh J., Knight R. 2011. UniFrac: an effective distance
594 metric for microbial community comparison. *ISME J* 5:169–172. DOI: 10.1038/ismej.2010.133.

595 Mahé F., Rognes T., Quince C., de Vargas C., Dunthorn M. 2015. Swarm v2: highly-scalable and high-
596 resolution amplicon clustering. *PeerJ* 3:e1420. DOI: 10.7717/peerj.1420.

597 Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer*
598 *Research* 27:209–220.

599 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
600 *EMBnet.journal* 17:10. DOI: 10.14806/ej.17.1.200.

601 Masella AP., Bartram AK., Truszkowski JM., Brown DG., Neufeld JD. 2012. PANDAseq: paired-end
602 assembler for illumina sequences. *BMC Bioinformatics* 13:1–7. DOI: 10.1186/1471-2105-13-31.

603 Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage
604 lysozyme. *BBA - Protein Structure* 405:442–451. DOI: 10.1016/0005-2795(75)90109-9.

Meyer JS., Ingersoll CG., Mcdonald LL., Boyce MS., Meyer JS., Ingersoll CG., Mcdonald LL., Boyce MS. 2016. Estimating Uncertainty in Population Growth Rates : Jackknife vs . Bootstrap Techniques. *Ecology* 67:1156–1166. DOI: 10.2307/1938671.

Nikolenko SI., Korobeynikov AI., Alekseyev MA. 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* 14:S7. DOI: 10.1186/1471-2164-14-S1-S7.

Oksanen J., Blanchet FG., Kindt R., Legendre P., Minchin PR., O'Hara RB., Simpson GL., Solymos P., Stevens MHH., Wagner H. 2016. vegan: Community Ecology Package.

Puente-Sánchez F., Aguirre J., Parro V. 2016. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic acids research* 44:e40. DOI: 10.1093/nar/gkv1113.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Gloeckner., Gloeckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* 41:D590–D596. DOI: doi: 0.1093/ nar/gks1219.

Quince C., Lanzen A., Davenport RJ., Turnbaugh PJ. 2011. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics* 12:38. DOI: 10.1186/1471-2105-12-38.

R Core Development Team., R Core Team. 2015. R: a language and environment for statistical computing. *Document freely available on the internet at: http://www. r-project. org*. DOI: 10.1017/CBO9781107415324.004.

Roeselers G., Coolen J., van der Wielen PWJJ., Jaspers MC., Atsma A., de Graaf B., Schuren F. 2015. Microbial biogeography of drinking water: Patterns in phylogenetic diversity across space and time. *Environmental Microbiology* 17:2505–2514. DOI: 10.1111/1462-2920.12739.

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. DOI: 10.7717/peerj.2584.

Rosselló-Mora R., Amann R. 2001. The species concept for prokaryotes. *FEMS Microbiology Reviews* 25:39–67. DOI: 10.1111/j.1574-6976.2001.tb00571.x.

Savio D., Sinclair L., Ijaz UZ., Parajka J., Reischer GH., Stadler P., Blaschke AP., Blöschl G., Mach RL., Kirschner AKT., Farnleitner AH., Eiler A. 2015. Bacterial diversity along a 2600 km river continuum. *Environmental Microbiology* 17:4994–5007. DOI: 10.1111/1462-2920.12886.

Schirmer M., Ijaz UZ., D'Amore R., Hall N., Sloan WT., Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* 43:1–16. DOI: 10.1093/nar/gku1341.

Schloss PD., Handelsman J. 2005. Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology* 71:1501–1506. DOI: 10.1128/AEM.71.3.1501.

639 Schloss PD., Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-
640 based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental*
641 *Microbiology* 77:3219–3226. DOI: 10.1128/AEM.02810-10.

642 Schloss PD., Westcott SL., Ryabin T., Hall JR., Hartmann M., Hollister EB., Lesniewski RA., Oakley
643 BB., Parks DH., Robinson CJ., Sahl JW., Stres B., Thallinger GG., Van Horn DJ., Weber CF.
644 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software
645 for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*
646 75:7537–7541. DOI: 10.1128/AEM.01541-09.

647 Schmidt TSB., Matias Rodrigues JF., von Mering C. 2015. Limits to robustness and reproducibility in
648 the demarcation of operational taxonomic units. *Environmental Microbiology* 17:1689–1706.
649 DOI: 10.1111/1462-2920.12610.

650 Shannon CE. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27:379–
651 423.

652 Simpson EH. 1949. Measurement of Diversity. *Nature* 163:688.

653 Stackebrandt E., Goebel BM. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S
654 rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal*
655 *of Systematic Bacteriology* 44:846–849. DOI: 10.1099/00207713-44-4-846.

656 Tange O. 2011. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine* 36:42–
657 47.

658 Tremblay J., Singh K., Fern A., Kirton ES., He S., Woyke T., Lee J., Chen F., Dangl JL., Tringe SG.
659 2015. Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology* 6:1–
660 15. DOI: 10.3389/fmicb.2015.00771.

661 Westcott SL., Schloss PD. 2015. De novo clustering methods outperform reference-based methods for
662 assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487. DOI:
663 10.7717/peerj.1487.

664 Zhang C., Yin A., Li H., Wang R., Wu G., Shen J., Zhang M., Wang L., Hou Y., Ouyang H., Zhang Y.,
665 Zheng Y., Wang J., Lv X., Wang Y., Zhang F., Zeng B., Li W., Yan F., Zhao Y., Pang X., Zhang X.,
666 Fu H., Chen F., Zhao N., Hamaker BR., Bridgewater LC., Weinkove D., Clement K., Dore J.,
667 Holmes E., Xiao H., Zhao G., Yang S., Bork P., Nicholson JK., Wei H., Tang H., Zhang X., Zhao
668 L. 2015. Dietary Modulation of Gut Microbiota Contributes to Alleviation of Both Genetic and
669 Simple Obesity in Children. *EBioMedicine* 2:968–984. DOI: 10.1016/j.ebiom.2015.07.007.

670 Zhou J., Wu L., Deng Y., Zhi X., Jiang YH., Tu Q., Xie J., Nostrand JD Van., He Z., Yang Y., Van
671 Nostrand JD., He Z., Yang Y. 2011. Reproducibility and quantitation of amplicon sequencing-based
672 detection. *ISME J* 5:1303–1313. DOI: 10.1038/ismej.2011.11.

# Figures

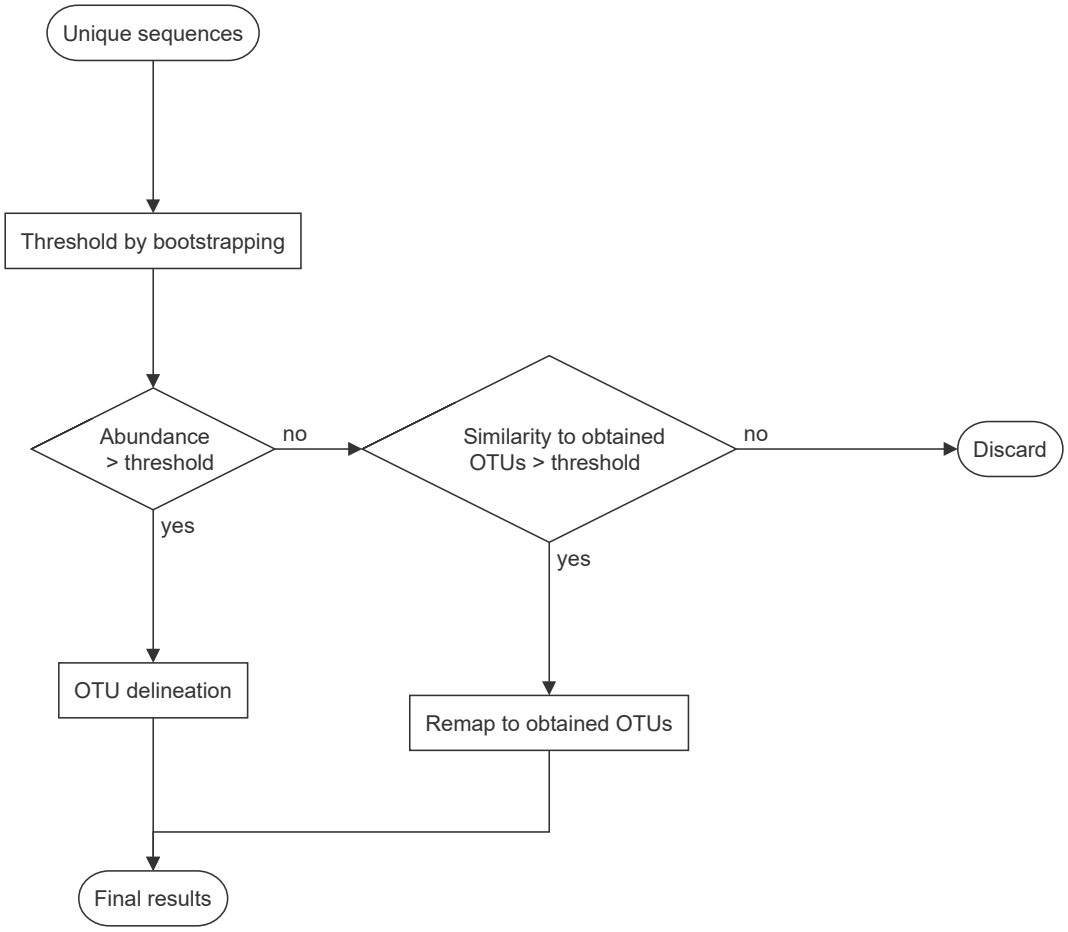
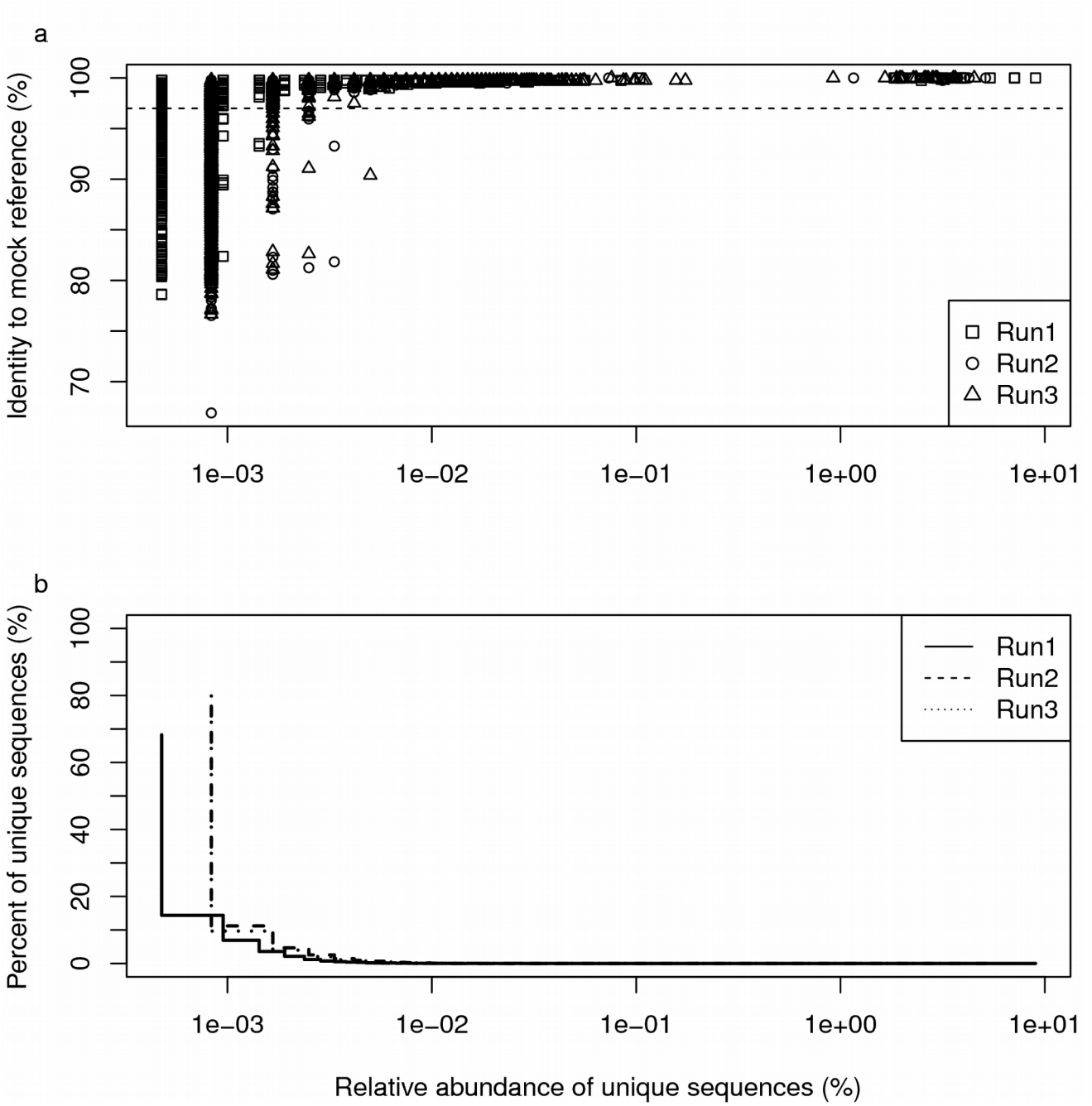
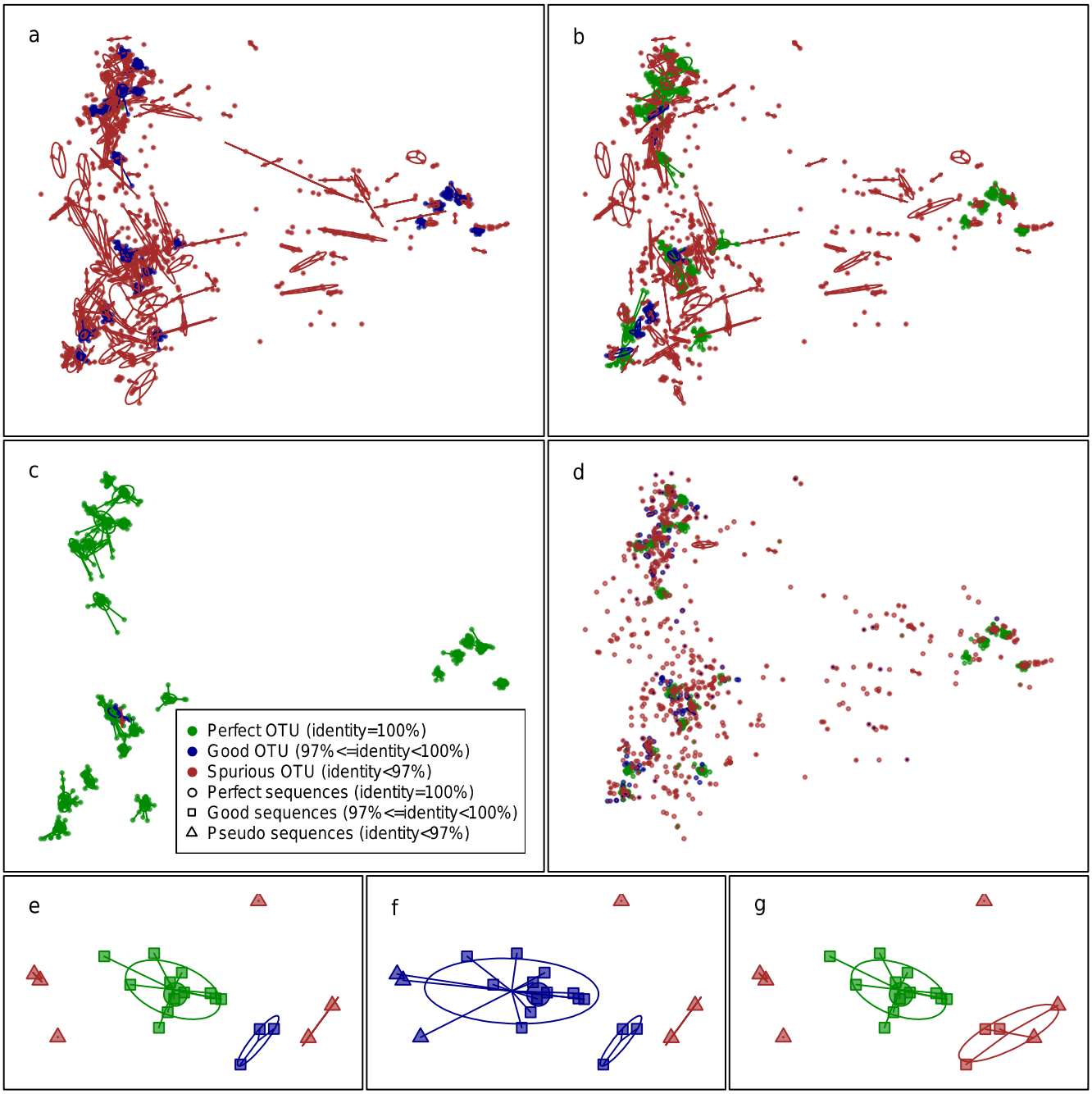



Figure 1 **Abundance filtering (AF) based OTU picking and remapping (AOR) approach.** The unique sequences were separated to reliable and unreliable ones based on their abundances in AF step. At AOR step, reliable sequences were used in OTU delineation, then unreliable sequences were remapped back to the obtained OTUs if they match the similarity threshold.

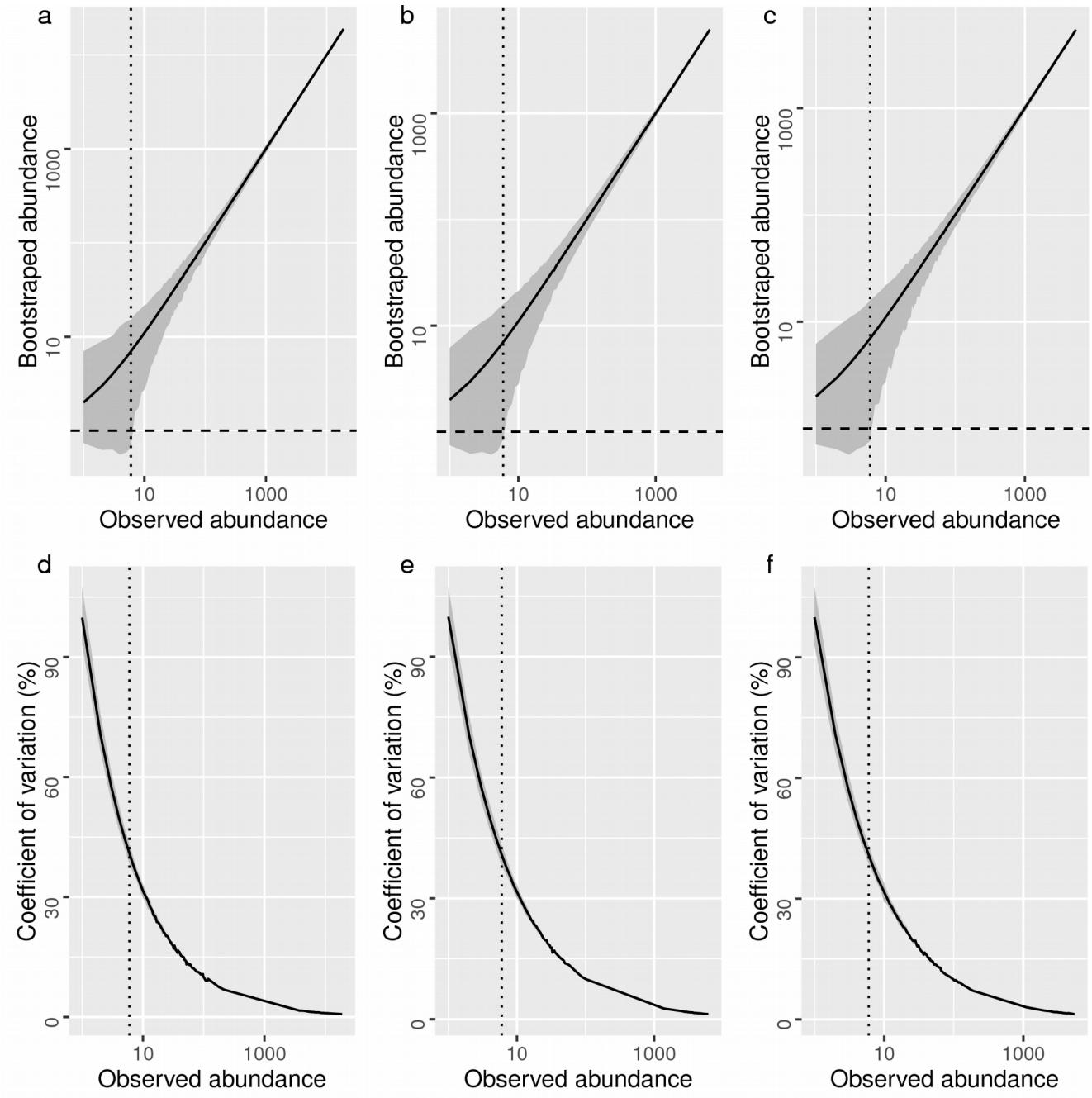



680 Figure 2 **Distribution of qualified unique sequences in mock communities.** (a) Similarity of
681 qualified unique sequences to the closest mock references. All qualified unique sequences with >3%
682 errors had lower relative abundances. (b) Distribution of qualified unique sequences according to their
683 relative abundance. The majority of unique sequences had a low relative abundance.

684




686 Figure 3 **Effect of low-abundance pseudo sequences on OTU delineation.** Dots represent the unique
687 sequences belonging to the 22 species. The ellipses and lines indicate how the unique sequences were
688 clustered into OTUs. The dot shape indicates the accuracy of each unique sequence. The color indicates
689 the type of OTU that each unique sequence was assigned to. (a) UCLUST, (b) AL, (c) UPARSE, (d)
690 Swarm. The existing pseudo sequences resulted in a large number of spurious OTUs around real
691 species. Diverse algorithms and parameters treated these pseudo sequences differently so that they
692 could (e) form spurious OTUs by themselves, (f) be clustered with perfect and good sequences to make
693 consequent OTUs not identical to real species or (g) attract good sequences to form spurious OTUs.




695 Figure 4 **Statistical characterization of unique sequences in the mock data.** (a-c) The 99%
696 confidence intervals of bootstrapped abundances. The distribution of bootstrapped abundances included
697 zero when the abundance was low. (d-f) The coefficient of variation values decreased quickly along
698 with the abundance of the sequence. Dashed vertical lines show the abundance thresholds for OTU
699 delineation.

700

Table 1. **The AOR approach can overcome the overestimation of OTU number in mock communities constructed by 22 16S rRNA gene clones**.

| Method | Species | Original result | | | With AOR or AF | | |
|---|---|---|---|---|---|---|---|
| | | Perfect | Good | Spurious | Perfect | Good | Spurious |
| UPARSE | 22 | 22, 22, 22 | 0, 0, 0 | 1, 0, 1 | 22, 22, 22 | 0, 0, 0 | 0, 0, 0 |
| UCLUST_denovo | | 22, 22, 22 | 31, 21, 44 | 154, 308, 326 | 22, 22, 22 | 0, 0, 0 | 0, 0, 0 |
| mothur_AL | | 22, 22, 22 | 9, 10, 10 | 155, 308, 328 | 22, 22, 22 | 0, 0, 0 | 0, 0, 0 |
| Swarm | | 22, 22, 22 | 487, 709, 816 | 204, 456, 486 | 22, 22, 22 | 4, 5, 1 | 0, 1, 0 |

Results from the three sequencing runs are separated by comma.

Table 2. **The AF approach is also efficient on non-OTU-based methods in mock communities.**

| Method | Species | run1 | | | run2 | | | run3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Perfect | Good | Spurious | Perfect | Good | Spurious | Perfect | Good | Spurious |
| DADA2 | 22 | 42 | 3 | 3 | 41 | 0 | 1 | 41 | 0 | 1 |
| DADA2 (abundance >=7) | | 42 | 3 | 0 | 40 | 0 | 1 | 40 | 0 | 1 |
| Deblur (abundance >=10) | | 22 | 0 | 0 | 22 | 0 | 0 | 22 | 0 | 0 |
| Deblur (abundance >= 7) | | 22 | 0 | 0 | 22 | 0 | 0 | 22 | 0 | 0 |
| Deblur (all reads) | | 22 | 7 | 3 | 22 | 2 | 1 | 22 | 2 | 1 |
| MED (abundance >=7) | | 22 | 87 | 0 | 22 | 14 | 1 | 22 | 32 | 1 |
| MED (all reads) | | 22 | 1112 | 150 | 22 | 883 | 306 | 22 | 1061 | 313 |

706 Table 3. **OTU-based analysis in Simulated data.**

| References | | Original result | | | With AOR or AF | | |
|---|---|---|---|---|---|---|---|
| | | Perfect | Good | Spurious | Perfect | Good | Spurious |
| UPARSE | 87 | 76 | 2 | 3 | 81 | 2 | 6 |
| UCLUST_denovo | | 83 | 6 | 1577 | 83 | 1 | 206 |
| mothur_AL | | 82 | 240 | 1566 | 60 | 85 | 202 |
| Swarm | | 81 | 2481 | 2079 | 80 | 472 | 218 |

707

708 Table 4. **Non-OTU-based analysis in Simulated data.**

| Method | References | Perfect | Good | Spurious |
|---|---|---|---|---|
| DADA2 | | 107 | 196 | 155 |
| DADA2 (abundance >=7) | | 103 | 183 | 137 |
| Deblur (abundance >=10) | | 100 | 133 | 166 |
| Deblur (abundance >= 7) | 87 | 105 | 191 | 235 |
| Deblur (all reads) | | 154 | 648 | 725 |
| MED (abundance >=7) | | 83 | 554 | 185 |
| MED (all reads) | | Exceeds the hard limit (7312 in total) | | |

709