Optimization of 16S amplicon analysis using mock communities: implications for estimating community diversity

Andrew Krohn[1,2], Bo Stevens[1], Adam Robbins-Pianka[4], Matthew Belus[5], Gerard J. Allan[1,2], Catherine Gehring[1,3]

[1]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ
[2]NAU Environmental Genetics and Genomics Laboratory, Flagstaff, AZ
[3]Merriam-Powell Center for Environmental Research, Flagstaff, AZ
[4]Department of Computer Science, University of Colorado Boulder, Boulder, CO
[5]Anschutz Medical Campus, University of Colorado Denver, Aurora, CO

Corresponding Author:
Andrew Krohn[1,2]

Email address: alk224@nau.edu

Abstract:

Diversity of complex microbial communities can be rapidly assessed by community amplicon sequencing of marker genes (*e.g.*, 16S), often yielding many thousands of DNA sequences per sample. However, analysis of community amplicon sequencing data requires multiple computational steps which affect the outcome of a final data set. Here we use mock communities to describe the effects of parameter adjustments for raw sequence quality filtering, picking operational taxonomic units (OTUs), taxonomic assignment, and OTU table filtering as implemented in QIIME 1.9.1. We demonstrate a workflow optimization based upon this exploration which we also apply to environmental samples. We found that quality filtering of raw data and filtering of OTU tables had large effects on observed OTU diversity. While all taxonomy assigners performed with similar accuracy, an appropriate choice of similarity threshold for defining OTUs depended on the method used for OTU picking. Our "default" analysis in QIIME overestimated mock community diversity by at least a factor of ten, compared to the optimized analysis which correctly characterized the taxonomic composition of the mock communities while still overestimating OTU diversity by about a factor of two. Though observed relative abundances of mock community member taxa were approximately correct, most were still represented by multiple OTUs. Low-frequency OTUs conspecific to constituent mock community taxa were characterized by multiple substitution and indel errors and the presence of a low quality base call resulting in sequence truncation during quality filtering. Low quality base calls were observed at "G" positions most of the time, and were also associated with a preceding "TTT" trinucleotide motif. Environmental diversity estimates were reduced by about 40% from 2508 to 1533 OTUs when comparing output from the default and optimized workflows. We attribute this reduction in observed diversity to the removal of erroneous sequences from the data set. Our results indicate that both strict quality filtering of raw sequencing data and careful filtering of raw OTU tables are important steps for accurate estimation of microbial community diversity.

Introduction:

Over the past decade, community amplicon sequencing has become the preferred method for profiling diversity in microbial communities. Briefly, the technique uses the polymerase chain reaction (PCR) to amplify a pool of PCR products from an environmental sample to be

resolved by high throughput DNA sequencing. Similar sequences are binned together into operational taxonomic units (OTUs) which are compared against a database to obtain taxonomic classifications. Amplicon sequencing is flexible in that a community can be profiled for different genes which may represent markers better suited for certain microbial constituents (*e.g.*, 16S for prokaryotes, ITS for fungi), while profiling with functional genes can offer a better understanding of community traits (*e.g.*, Bentzon-Tilia *et al.*, 2015). While communities were originally profiled on 454 pyrosequencing instruments (Sogin *et al.*, 2006), amplicon sequencing has been adapted to newer instrumentation including sequencers from Illumina (Caporaso *et al.*, 2012) and Pacific Biosciences (Fichot & Norman, 2013). Illumina sequencing is currently the most popular option due to several factors including cost, throughput, instrument availability, and the existence of multiple protocols for amplification and sequencing of marker gene pools on this platform (Caporaso *et al.*, 2012; Bokulich & Mills, 2013; Kozich *et al.*, 2013; Fadrosh *et al.*, 2014).

Accurate determination of community diversity and taxonomic content are often primary aims of community amplicon sequencing projects. Systematic errors experienced during sample preparation such as PCR and sequencing errors can contribute to overestimates of diversity (Kunin, 2010). Additionally, signal cross-talk during index sequence cycles on Illumina sequencers can lead a researcher to falsely conclude that an organism is present in a sample (Kircher, Sawyer & Meyer, 2012; Nelson *et al.*, 2014). In the face of such potential complications, careful analysis is warranted to ensure that diversity estimates are not inflated and that data are properly filtered to avoid Type II errors. Several comprehensive tools exist for processing such data including mothur (Schloss *et al.*, 2009), QIIME (Caporaso *et al.*, 2010a), and UPARSE (Edgar, 2013). Many stand-alone tools are also available for performing specific bioinformatic tasks which may or may not be implemented in QIIME, mothur or UPARSE. It may be beneficial in some cases to perform separate bioinformatic steps with different software packages in order to obtain the most accurate community representation for a given ecosystem. However, it is up to the individual researcher to have a comprehensive understanding of the production and processing of amplicon sequencing data in order to make the best decisions during data processing.

Automated quality filtering is among the first steps performed in any sequencing project and is a necessity for managing modern DNA sequencing data sets. To achieve the status of

"finished," genome sequencing projects require consensus base quality scores where the likelihood of an incorrect base call is less than 1 in 100,000 (q50), whereas assemblies using unfiltered data are considered "standard draft" and are expected to contain errors (Chain & Grafham, 2009). The default parameters in QIIME 1.9.1 require a minimum quality score of q4 as recommended by Bokulich *et al*. (2013), and such data should be similarly treated as "draft" data. More reads are retained for downstream analysis, but a low quality score requirement also introduces an unknown degree of sequencing error as base quality scores may vary widely across a single sequencing run. Thus, data generated on runs with higher average error rates are more likely to overestimate alpha diversity if quality scores are not strictly controlled. While inconsistent qualities from sequencing runs can be effectively controlled via quality filtering, default quality filtering in QIIME retains reads that may be variably trimmed to a range of 75-100% of the original sequence length. Because the quality of different sequences may decrease nonuniformly across a sequencing run, variable read lengths may also contribute to an inflated estimate of OTU richness if reads are not dereplicated or sorted by size prior to clustering.

Quality-filtered amplicon sequencing data are clustered into OTU definitions, a computational process for which numerous programs are available. CD-HIT (Fu *et al*., 2012), UCLUST (Edgar, 2010), BLAST (Altschul, 1990), and Swarm (Mahé *et al*., 2014) are popular options that are all available in QIIME. Reference-based analysis techniques, such as BLAST, are known to incur biases according to the choice of reference database (Nelson *et al*., 2014), but can easily be parallelized for more efficient computation. UCLUST can utilize a reference database, perform database-independent *de novo* clustering, or, as with the open-reference strategy currently implemented in QIIME, a combination of both methods (Navas-Molina *et al*., 2013). Pure *de novo* analysis is preferred by many as the approach least likely to impose a bias on the final outcome. One popular option for *de novo* OTU clustering is CD-HIT, but as this program cannot be parallelized it can be time-prohibitive when used with larger data sets. Swarm, another *de novo* OTU clustering program, allows for portions of the *de novo* clustering process to be parallelized, thus eliminating database-specific effects while also optimizing computational requirements. All OTU picking programs require the researcher to choose a similarity or distance threshold beyond which two sequences must be considered as separate OTUs. If present at this stage, PCR or sequencing errors may contribute to OTU inflation to an unknown degree. In addition to ensuring the data is properly filtered, one can also utilize a

conservative clustering threshold in order to avoid overestimation of community diversity (*e.g.*, ≤97%; Kunin *et al*., 2010).

Taxonomic assignment, achieved through comparison of OTU definition sequences to a reference database, can also be performed in a variety of ways. Popular methods include BLAST, UCLUST, and RDP (Wang *et al*., 2007), and each are available in QIIME. In 2008, Liu *et al*. reported that RDP provided the most accurate taxonomic assignments. Presently, other techniques continue to be utilized by various amplicon sequencing analysis pipelines (*e.g.*, Giongo *et al*., 2010; Gweon *et al*., 2015), revealing a lack of consensus among researchers. Considering that improved taxonomic accuracies may be observed when sequences obtained for study organisms are more similar to those populating the reference database, it seems plausible that the relative success of each algorithm can be context-dependent. For environmental data sets, accuracies of taxonomic assignments are estimated by means of a confidence value relevant to the utilized technique (*e.g.*, e-value for BLAST). Careful assessment of taxonomic accuracies can only be done when the sequence content of a given sample can be anticipated. This can be achieved with synthetic mock communities created *in silico* by extracting sequences from a database (*e.g.*, Bellemain *et al*., 2010) or using genomic mock communities that combine DNA extracts from cultured organisms. Neither scenario is likely to provide an outcome that is directly comparable to the natural complexities of environmental communities, yet both can offer a test of accuracy for taxonomic assignment methods.

Once quality filtered sequences have been clustered and taxonomically classified, they are compiled into an OTU table with count data for each observation. As OTUs defined from erroneous sequences may persist even to this point in the analysis, the resulting OTU table must be filtered prior to conducting diversity analyses, and the filtering approach can have a profound effect on the final result (Bokulich *et al*., 2013). Although Bokulich *et al*. (2013) suggested the inclusion of mock communities on sequencing runs to assess the overall run quality and improve diversity assessments, they also provide a general recommendation to quality filter the final table by removing OTUs that represent less than 0.005% of the total read abundance. This has proven to be a useful guideline for numerous studies in which mock communities were not included. However, this practice ignores the independence of each sample and will treat samples differently according to sequencing depth such that low read count samples will be more severely filtered than samples with higher read counts.

Considering samples independently, Kircher, Sawyer & Meyer (2012) observed an indexing inaccuracy rate of 0.3%, citing cluster mixing during sequencing as a mechanism by which single-indexed Illumina sequences are likely attributed incorrectly to a particular sample. For certain applications, their result argues that such data must be filtered at this level in order to avoid Type II errors. Another common practice is to remove singleton OTUs (by sample or by table) under the assumption that such OTUs represent errors generated during sequencing (see Dickie, 2010). However, errors introduced during early PCR cycles may be faithfully replicated many times so as to appear as valid OTUs, causing overestimation of OTU richness even after singleton filtering (Nguyen *et al.*, 2015). As an alternative, Nguyen *et al.* (2015) suggest the removal of low-count or low-proportion OTUs by sample at a threshold informed by mock community data. Mock communities used in this way may also identify certain sequence motifs prone to error, which may help to identify whether novel OTUs observed in environmental data should be considered suspect. Unfortunately, such controls are not available for many data sets and artificial communities may not perform similarly to environmental communities during sample prep and analysis. Because samples are amplified independently, PCR errors are likely to be present in the form of private OTUs observed only in a single sample, so removal of unshared OTUs may be another effective precaution against overestimation of diversity due to sequencing error.

As these examples illustrate, proper filtering of an OTU table is not a straightforward task. The sequence misattribution rate reported by Kircher, Sawyer & Meyer (2012) is orders of magnitude above the filtering threshold of 0.005% recommended by Bokulich *et al*. (2013), though their recommendation was to filter across the entire OTU table. Since many amplicon sequencing studies report relatively few taxa present above 0.3% per sample, filtering by sample at this threshold (Kircher threshold) will exclude many valid taxa. The presence of misattributed sequences may also diminish the efficacy of private OTU removal to eliminate PCR errors, though dual-indexing of samples should reduce or eliminate sequence misattribution events (Kircher, Sawyer & Meyer, 2012). Singleton filtering, however applied, is unlikely to be thorough enough to remove errors that are either replicated during the PCR process, or represent systematic errors from the sequencing process. For single- or dual-indexed Illumina data, filtering at 0.005% across the entire table (Bokulich threshold) may represent a viable

compromise between confident assignment of sequences to samples and the stringency that one imposes on filtering the final table.

In this study we used simple genomic mock communities and an environmental data set to describe the effects of parameter adjustments for methods implemented in QIIME 1.9.1 (Caporaso *et al*., 2010a) on sequence quality filtering, OTU picking, taxonomic assignment, and OTU table filtering. We hypothesized that observed OTU diversity is dramatically inflated due to the presence of PCR and/or sequencing artifacts, and that such effects should be observed in simple genomic mock communities. Using five mock communities consisting of 4-8 taxa each, we developed a modified protocol for the analysis of 16S community amplicon sequencing data, and demonstrate the method on an environmental data set. By carefully controlling each of the steps that we investigated, we were able to describe mock community compositions more correctly than with a default workflow.

Materials and Methods:

Mock communities

DNA was extracted from axenic cultures of *Pseudomonas aeruginosa* (Proteobacteria), *Proteus vulgaris* (Proteobacteria), *Klebsiella pneumoniae* (Proteobacteria), *Escherichia coli* (Proteobacteria), *Bacillus megaterium* (Firmicutes), *Lactococcus lactis* (Firmicutes), *Staphylococcus aureus* (Firmicutes), and *Micrococcus luteus* (Actinobacteria) using a PowerSoil DNA Extraction Kit (MoBio Laboratories, Carlsbad, CA). DNA was quantified by PicoGreen (Life Technologies, Carlsbad, CA) fluorescence, and normalized to approximately 0.75 ng/µL. Five mock communities containing different ratios of bacterial taxa were constructed from the extracted DNA. Community 0 contained equal volumes of DNA from each taxon; Community 1a contained 8% *M. luteus*, 42% *B. megaterium*, 42% *L. lactis*, and 8% *S. aureus*; Community 1b contained 42% *M. luteus*, 8% *B. megaterium*, 8% *L. lactis*, and 42% *S. aureus*; Community 2a contained 8% *E. coli*, 8% *K. pneumoniae*, 42% *P. vulgaris*, and 42% *P. aeruginosa*; Community 2b contained 42% *E. coli*, 42% *K. pneumoniae*, 8% *P. vulgaris*, and 8% *P. aeruginosa*. Final concentrations for each mock community were determined to be ~ 0.75 ng/µL (Table S1: mock community construction). Expected compositions of mock communities were corrected for genome size and copy number against the CBS Genome Atlas Database (Hallin & Ussery, 2004).

Environmental samples

Environmental samples with an expected environmental contrast were collected from the Northern Arizona University Pinyon Pine Common Garden near Sunset Crater National Monument, AZ. During garden installation in October 2009, soil samples were collected from holes dug to plant seedlings ("pre-tree" treatment). Soil core samples were taken from the same seedlings in December 2010 ("post-tree" treatment). The top 2 centimeters (cm) of soil were brushed aside prior to taking cores. A 2.5 cm diameter metal corer was placed 2 cm from the seedling base and driven to a depth of 10 cm. Samples were kept on ice in the field and stored at -20 ^{o}C until DNA extraction. DNA was extracted from homogenized soil cores using a PowerSoil DNA Extraction Kit. Only samples which produced a clean ribosomal PCR product were included in this study, resulting in unequal sample sizes between pre-tree (n = 13) and post-tree (n = 28) groups. A random number generator was used to select a subset of post-tree samples (n = 13) for comparisons of data with equal sample sizes. Samples were normalized to c. 1 ng/µL prior to PCR amplification for library construction.

The environmental samples presented here are meant only to allow a demonstration of the effects of a mock community-based workflow optimization on actual data. Though we expect the presence of a tree to create additional niche space which would increase observed diversity, no background soil control samples were collected. Observed differences, though likely to be real, could be influenced in part or in total by interannual environmental variations. Additionally, pre-tree and post-tree samples were collected during different months of the year, so seasonal differences could also contribute to the outcome.

Library construction and sequencing

Amplicons were produced in a two-step protocol as suggested by Berry *et al*. (2011). Briefly, samples were amplified in triplicate PCR reactions for the 16S v4 region using the universal prokaryotic primers 515F and 806R (Bates *et al*., 2011). First round reactions were performed in triplicate in 384 well plates. The 8 µL volumes contained the following: 1 µM each primer (Eurofins MWG Operon, LLC), 200 µM each dNTP (Phenix Research, Candler, NC), 0.01 U/µL Phusion Hot Start II DNA Polymerase (Life Technologies), 1X HF Phusion Buffer (Life Technologies), 3 mM $MgCl_2$, 6% glycerol, and 1 µL normalized template DNA. Cycling

conditions were: 2 minutes at 95°C followed by 20 cycles of 30 seconds at 95°C, 30 seconds at 55°C, 4 minutes at 60°C. Triplicate reactions for each sample were pooled by combining 4 µL from each, and 2 µL was used to check for results on a 1% agarose gel. The remainder was diluted 10-fold and used as template in a second PCR reaction in which 12 base Golay indexed tails (Caporaso *et al*., 2012) were added. Second round reaction conditions were identical to the first round except only one reaction was conducted per sample and only 15 total cycles were performed. Indexed PCR products were purified using carboxylated magnetic beads as described in Rohland & Reich (2012), quantified by PicoGreen fluorescence, and an equal mass of each sample was combined into a final sample pool. The pool was purified and concentrated, and subsequently quantified by quantitative PCR against Illumina DNA Standards (Kapa Biosystems, Wilmington, MA). Sequencing was carried out on a MiSeq Desktop Sequencer (Illumina Inc, San Diego, CA) running in paired end 2x150 mode.

Data processing and statistical analysis

All bioinformatics were carried out on a Mac Pro (Apple, Inc.) running Ubuntu Linux 14.04 LTS (Canonical Ltd.) or the monsoon high-performance computing cluster at Northern Arizona University (https://nau.edu/hpc/) running CentOS 6.6 (The CentOS Project). Figures were generated in Veusz v1.24 (http://home.gna.org/veusz/) or Geneious v8.1 (Biomatters Ltd.). As contaminating PhiX Control sequence can complicate sequencing projects (Mukherjee *et al*., 2015), we calculated the amount of PhiX Control among our demultiplexed data and removed it prior to sample processing. This task was performed with the `akutils phix_filtering` command in akutils v1.2 (Krohn, 2016; https://github.com/alk224/akutils-v1.2) which maps raw data against the Enterobacteria phage phiX174 sensu lato complete genome sequence (NC_001422.1) using Smalt 0.7.6 (http://www.sanger.ac.uk/resources/software/smalt/).

Paired end reads were joined using the `akutils join_paired_reads` command in akutils which employs `fastq-join` from ea-utils (Aronesty, 2011). Demultiplexing and quality filtering of raw, joined data (mean length = 253 bp) was carried out in QIIME with the `split_libraries_fastq.py` script using default parameters, or with more strict requirements of a minimum quality threshold of q20 (q = 19), allowing 0-3 low-quality base calls (r = 1-3), and requiring at least 95% of each read to be high quality (p = 0.95). Chimeras were removed using vsearch 1.1.1 (Rognes *et al*., 2015; https://github.com/torognes/vsearch)

against the Gold reference database (http://drive5.com/uchime/gold.fa). OTU picking and taxonomy assignments were performed using the `akutils pick_otus` command in akutils. After manual inspection of sequence divergence among congeneric mock community members, sequences were de-replicated on the first 100 bases using the prefix_suffix OTU picker in QIIME. OTU picking was performed with multiple similarity thresholds using common OTU picking algorithms (CD-HIT, UCLUST and BLAST at 97%, 95%, 92%, 90%, 85%, and Swarm at *d*1, *d*2, *d*3, *d*4, *d*5). BLAST was used only for closed reference analysis, UCLUST for open reference analysis, and CD-HIT and Swarm for *de novo* analyses. Taxonomy was assigned using BLAST, RDP, and UCLUST using default settings available in QIIME 1.9.1. Reference-based OTU picking steps and taxonomy assignments were conducted against the Greengenes 97% database (McDonald *et al*., 2012) which had been formatted to include only the v4 region using the `akutils format_database` command in akutils. Sequence alignments and phylogenetic trees were produced using the `akutils align_and_tree` command in akutils which aligns sequences using PyNAST (Caporaso *et al*., 2010b) and generates phylogenies with FastTree (Price, Dehal & Arkin, 2009). Diversity analyses were conducted using the `akutils core_diversity` command in akutils.

In order to facilitate assessment of optimal workflow steps, we first sought to establish a method of filtering the final OTU tables by eliminating OTUs resulting from mixed clusters. To this end, we processed the mock and environmental data sets through a default QIIME workflow (see below) to assess taxonomic components, and compared methods for filtering OTU tables, which would remove contaminating taxa. An ideal filtering method should remove erroneous OTUs that arise either from sequencing error or cluster mixing. Table filtering was carried out using either the Kircher threshold (0.3% by sample; Kircher, Sawyer & Meyer, 2012), the Bokulich threshold (0.005% by table; Bokulich *et al*., 2013), singletons removed by table (mc2), or singletons removed by sample (n2). Private OTUs were assumed to be errors and were also removed in the n2 tables. Filtered OTU tables were grouped according to filtering method, and differences in the amount of OTUs classified as contaminating taxa was assessed by one-way ANOVA. Tukey's HSD test was used to determine which groups were statistically distinct.

An optimal workflow was chosen by assessing diversity estimates and taxonomic identities assigned to mock community data. The optimal OTU picking algorithm was determined as the method that yielded the correct diversity result over the broadest range of

similarity thresholds. Taxonomic accuracy was determined by seeding the Greengenes database with the expected sequences from the mock community constituent taxa prior to analysis, and inspecting the results. OTU tables from the optimal workflow across the accurate range of similarity thresholds were filtered at each of the four thresholds described above. Our "default QIIME workflow" was identical to the optimal workflow with the following changes: the `split_libraries_fastq.py` command was performed with default settings; OTU picking was performed with the `pick_open_reference_otus.py` command; taxonomic assignment was performed with UCLUST; OTU tables were filtered with the Bokulich threshold. Results from the optimal workflow were compared to the result obtained from our default workflow. Environmental data was then processed using the best workflow determined from this process and compared to the default result.

Diversity analyses for mock community data were calculated on OTU tables that had been rarefied to 10,000 reads, or 5,000 reads for environmental data. Comparison of observed mock community composition to the *a priori* expectation (Table S1) was conducted with Spearman's rank correlation using species-level assignments. Comparison of observed OTU diversity between environmental sample groupings was performed with nonparametric t-tests. A random subset of post-tree samples from the environmental data (n = 13) was selected to determine if unequal sample sizes were contributing to observed OTU diversity. Distance matrices were calculated from environmental data for weighted UniFrac distance (Lozupone & Knight, 2005). Tests of differences of total beta diversity were carried out on distance matrices using PERMANOVA (Anderson, 2001), and differences in multivariate dispersion were detected with PERMDISP (Anderson, Ellingsen & McArdle, 2006).

Representative sequences for the optimized mock community result were extracted from the output data. When multiple OTU definition sequences represented the same taxonomic identity, they were aligned with Mafft v7.123b (Katoh & Standley, 2013) using the L-INS-i setting. The lower abundance OTU for each multi-OTU taxon was assumed to be erroneous and base differences compared to the major OTU were characterized. Trinucleotide motifs preceding each base difference and terminal truncation position were tabulated. Because 2x150 sequencing data does not fully overlap for 515F-806R amplicons (mean length = 253 bp), terminal base and preceding trimers were considered in the context of the second read. Environmental data processed through the optimal workflow was also investigated for terminal truncation positions

and preceding trinucleotide motifs. Because we have no reliable reference sequence for many environmental OTUs, we investigated only OTUs that shared a taxonomic designation with at least one other OTU, and had been truncated by more than 3 bases during quality filtering. For mock and environmental data, motif and terminal base representations were tested against the assumption of random occurrence with Chi-square tests.

Results:

The sequencing run clustered at 1119 k/mm^2 (+/- 70) and resulted in 17.96 million total reads passing filter, an overall error rate of 0.36%, and 91% of reads exceeded q30. PhiX Sequencing Control v3 sequences (Illumina, Inc.) constituted 8.31% of the total run (percent aligned). Once demultiplexed, mock community data contained 4.35% PhiX (103,070/2,371,510 reads) while the environmental data contained 4.10% PhiX (259,366/6,332,586 reads). Raw sequencing data for the samples used in this study and a QIIME-formatted mapping file are publicly available in the QIITA database (https://qiita.ucsd.edu/) under study ID number 10479.

Under default QIIME assessment, the mock community data showed substantial OTU inflation; where there should have been just 8 OTUs, there were 127 (Table S2, default mock analysis). When the environmental data set was processed through the same workflow, 73 OTUs were classified at the family level as Sphingomonadaceae. Together, these OTUs made up 5.3% of environmental sequences, and Sphingomonadaceae was the most abundant classification observed at the family level (Table S3, default environmental analysis). Three OTUs representing about 0.13% of the mock community data set were also classified as Sphingomonadaceae, a designation which should be absent from the mock data. This result led us to surmise that sequences from the environmental data set were contaminating the mock communities during sequencing. Such sample cross-talk presumably arises from the cluster mixing effect described by Kircher, Sawyer & Meyer (2012) where the index read from a flowcell cluster is spuriously attributed to a neighboring cluster. The mock data also contained 3 OTUs classified as Planococcaceae (<0.03%) and 1 OTU classified as Methylobacteriaceae (<0.01%), again corresponding with OTUs observed within the environmental data. Sphingomonadaceae sequences were observed across all five mock communities, whereas Planococcaceae was only associated with communities 0, 1a, and 1b, suggesting that cluster-mixing events may occur non-randomly. Methylobacteriaceae was present as just a single read

among community 1b. Three mock community OTUs were observed at low levels in communities from which they should be absent, indicating additional cluster-mixing within the mock community data.

As the most prevalent non-target taxon observed among the mock community data, we sought to establish a method for filtering OTU tables that would eliminate the presence of contaminating Sphingomonadaceae reads. OTU tables generated for the mock communities by each of the OTU picking, taxonomy assignment and table filtering methods were compared for the presence of Sphingomonadaceae contaminants. Considering filtering method (mc2, n2, Kircher threshold, or Bokulich threshold) as the predictive variable, we found strong differences among them in removing non-target OTUs ($F_{3,239} = 89.301$, $p < 0.0001$). The least severe filtering method (mc2) retained the most Sphingomonadaceae OTUs (2.50 +/- 1.21) followed by n2 (2.45 +/- 1.21), and Bokulich threshold (1.85 +/- 0.86). Only the Kircher threshold completely removed Sphingomonadaceae contamination from the mock community OTU tables effectively (Tukey's HSD, $p < 0.05$).

Default quality filtering and OTU picking in QIIME resulted in overestimation of mock community diversity regardless of how the final OTU table was filtered (Figure 1a-d; Figure S1: Default mock community rarefactions). Diversity estimates were inflated up to 35 times when singletons were removed by table, compared to nearly 3.5 times when filtering with the Kircher threshold. Despite the reduction of OTU inflation by an order of magnitude, these results indicate that revisions to initial processing steps may yield improved results. We therefore sought to establish an optimized workflow that would produce the correct number of OTUs for an input of known constituents. Using data that had been filtered according to strict standards during the `split_libraries_fastq.py` step in QIIME ($q = 19$, $r = 0$, $p = 0.95$), a correct result was achieved for each of the OTU picking algorithms tested. However, each algorithm differed in which similarity threshold was required for the optimal result (Table 1). Closed reference picking with BLAST overestimated diversity above a similarity threshold of 92%. Open reference picking with UCLUST overestimated diversity at every threshold except 95% similarity. *De novo* picking using CD-HIT at thresholds below 92% and Swarm resolutions below *d*4 underestimated diversity. Swarm yielded the correct result over the broadest range of tested similarity thresholds (*d*1-*d*4), and offers other attractive features that made it stand out among the tested OTU pickers (*e.g.*, *de novo* picking, multi-threaded analysis). Thus, Swarm was chosen as

the optimal OTU picking method for the remainder of the study. We chose *d*4 similarity as the optimal threshold as it was the most conservative setting to yield a correct result.

Taxonomic accuracy for Swarm-picked OTUs (*d*4) was assessed for the different taxonomy assigners using default parameters in QIIME 1.9.1. To control for reference database bias, we added representative sequences from each of the correct OTUs to our Greengenes reference with a unique identifier. We observed that BLAST returned the representative sequence 100% of the time, while RDP and UCLUST never found the exact match (Table 2). Even though RDP and UCLUST did not find optimal sequences, assignments were correct, though less specific in taxonomic depth. BLAST yielded similar results when the representative sequences were not present in the database (Table 2). While BLAST offers the advantage of obtaining the best sequence match when available in the database, RDP and UCLUST both offer an advantage in substantially reducing computational time while providing reasonable accuracy for most applications. For the analysis presented here, we chose BLAST as the optimal taxonomy assigner for its superior accuracy.

A perfect result for analysis of our mock communities requires stringent quality filtering of the raw data. Default quality filtering in QIIME 1.9.1 was established according to Bokulich *et al*. (2013). This imposes a minimum Phred quality score of 4 ($q = 3$), truncates sequences after three bases are observed below this threshold ($r = 3$), and retains truncated reads that represent a minimum of 75% of the original sequence length ($p = 0.75$). In contrast, we performed strict quality filtering using $q = 19$, $r = 0$, and $p = 0.95$. This more stringent filtering protocol ensures that data used for analysis are of much higher quality with approximately uniform read lengths. An important consequence of such stringency is that much of the raw data is discarded. Of the 2,373,247 raw mock community sequences, default quality filtering retained 2,020,542 reads (85.1%), whereas stringent parameters retained just 657,544 reads (27.7%). Holding constant $q = 19$ and $p = 0.95$, we found that increasing r during quality filtering had a profound effect on the amount of data retained (Figure 2a). Allowing $r = 1$ resulted in an increase of data retention from approximately 27% (r = 0) to over 56%. When $r = 2$ and $r = 3$, increases in data retention showed diminishing returns, with 70% and 75% of the data retained, respectively. However, we also found that allowing $r > 0$ will generally cause an inaccurate estimate of the number of OTUs, depending on the criteria used for OTU picking (Figure 2b). With the *d*1 resolution, increasing r will create a proportional inflation in the number of OTUs determined by Swarm. At

434 *d*2 resolution, allowing $r = 1$ still correctly described our simple mock community whereas
435 allowing $r = 2$ or $r = 3$ caused diversity to be overestimated. At resolutions *d*3 and *d*4, allowing r
436 > 0 caused underestimates of diversity. This suggests that the best result is obtained with the
437 most stringent quality filter, which we selected for our optimal workflow ($q = 19$). Similar results
438 using more data may be possible by allowing a small amount of errors (*e.g.*, $r = 1$) and picking
439 OTUs with a more conservative similarity threshold (*e.g.*, Swarm at *d*2 resolution).
440 The Kircher threshold was effective at removing contaminating OTUs in our mock
441 community data thus yielding a near-perfect result (Figure 3). However, we anticipated that such
442 filtering could be too stringent for environmental analysis given the low per-sample OTU
443 frequencies commonly reported (*e.g.*, Sogin *et al.*, 2006). We compared the expected mock
444 community results to those observed with either default settings, or optimized settings for quality
445 filtering, OTU picking and taxonomy assignment, using each of the final OTU table filtering
446 methods we tested. For all comparisons, Spearman's rank correlation yielded significant p-values
447 (<0.001), so we present only correlation values and 95% confidence intervals here. When
448 comparing the default analysis to the expected outcome, Spearman's r showed a negative
449 correlation ($r = -0.3494$; CI = [−0.4280, −0.2655]). Optimized results exhibited strong positive
450 correlations regardless of filtering threshold used. Lower values for Spearman's r occured when
451 diversity was overestimated and when contaminants were present. Correlation with the expected
452 outcome improved as filtering stringency increased with every filtering method producing a
453 dramatic improvement over the default workflow (mc2: $r = 0.8075$, CI = [0.7663, 0.8420]; n2: r
454 = 0.8702, CI = [0.8344, 0.8987]; Bokulich threshold: $r = 0.9135$, CI = [0.8841, 0.9357]; Kircher
455 threshold: $r = 0.9646$, CI = [0.9495, 0.9752]). The Bokulich threshold was chosen as our optimal
456 OTU table filtering method because it yielded the best correlation without being overly strict.
457 Output for the environmental data using either the default or optimized workflow were
458 examined for basic diversity statistics. Default analysis identified 2508 OTUs classified into 388
459 taxonomic assignments (OTUs per taxon: mean = 6.46, median = 2; Figure S3: Default
460 environmental rarefactions). The optimized analysis identified 1533 OTUs classified into 328
461 taxonomic assignments (OTUs per taxon: mean = 4.67, median = 2; Figure S4: Optimized
462 environmental rarefactions). By treatment, OTU diversity was reduced about twofold when
463 assessed via the optimized workflow and compared to the default results (Figure 4a-4b). In the
464 default analysis, pre-tree soils hosted 978.30 +/- 128.42 OTUs while post-tree soils had 1138.35

+/- 86.34 OTUs (nonparametric T-test = 4.578, p < 0.001). In the optimized analysis, pre-tree soils contained 543.28 +/- 79.32 compared to 674.95 +/- 50.21 OTUs in post-tree soils (nonparametric T-test = 6.277, p < 0.001). Differences in beta diversity were observed between treatments for each workflow using weighted UniFrac distance matrices (Figure 4c-4d; default PERMANOVA = 8.181, p < 0.001; optimized PERMANOVA = 9.355, p < 0.001). We also noticed an increase in multivariate dispersion in the optimized workflow, though the differences were not found to be significant in either case (default PERMDISP = 1.086, $p = 0.294$; optimized PERMDISP = 2.160, $p = 0.158$). When data was processed with equivalent sample sizes, the same patterns were observed for both alpha diversity (pre-tree = 545.76 +/- 78.84, post-tree = 679.00 +/- 47.86; nonparametric T-test = 5.004, p < 0.001) and beta diversity (PERMANOVA = 6.585, p < 0.001), though statistical power was slightly reduced, and multivariate dispersion increased (PERMDISP = 3.248, $p = 0.071$), consistent with a reduction in sample size.

Of the 17 OTUs observed in the optimized mock result, the nine extra OTUs therein were composed of three contaminants and six spurious OTUs representing sequence variants of the target taxa. All extra OTUs were present at low levels ranging from 0.003% to 0.17% per sample (Table S4). That sequence counts of contaminant OTUs were observed in all samples, but only for select taxa, strongly suggests that cluster mixing events occur non-randomly during Illumina sequencing. Species-level mock community observations from the optimized workflow describe the eight constituent taxa at approximately the correct proportions. However, six of the eight taxa were represented by two OTUs each. The main OTU for each taxon was present as 6.30% to 19.09% of the total community while the rates of lower frequency OTUs ranged from 0.01% to 0.05%. Manual inspection of conspecific OTU sequence alignments revealed multiple substitution and indel positions within the first 100 bases which prevented these sequences from dereplicating into the correct sequence during our workflow (Table S5: OTU sequence alignments by taxonomy). Additionally, these sequence variants were shorter than the main constituent sequence by at least seven bases, indicating that they derive from inherently lower quality reads. Inspection of trinucleotide motifs preceding each substitution or indel position did not reveal any pattern relating to the observed errors (Table S6: Error-associated sequence motifs). Consistent with the results of Schirmer *et al*. (2015), we observed a higher rate of errors among A or C bases than G or T errors (error ratio = 1.67). Since A and C or G and T bases share fluorescence excitation wavelengths during Illumina 4-channel sequencing-by-synthesis

(SBS), this result suggests that some of the errors we observed were indeed the result of systematic errors during sequencing, although this study was not designed to distinguish between such errors and those generated during PCR. Examining the terminal trinucleotide motif immediately preceding truncation positions (Table S7: Terminal-associated sequence motifs) we observed "TTT" 83% of the time (X^2_{63} = 271.333, p < 0.0001). Additionally, the correct base at the truncation position was "G" 83% of the time (X^2_3 = 11.33, p = 0.0101). An example alignment for the two OTUs representing *B. megaterium* is presented in Figure 5a, illustrating the "TTT" motif preceding a "G" truncation position (reverse complimented).

Truncation positions and preceding trimers were also characterized for environmental data, resulting in 34 "suspect" OTUs (Table S8: Environmental terminal errors). Of these, 27 OTUs had been truncated at a "G" position (79.41%; X^2_3 = 54.235, p < 0.0001), and just 10 possible trimers were represented preceding the truncation position. The motifs "TTT" and "TTC" were substantially overrepresented, being observed 14 (41.18%) and 7 (20.59%) times, respectively (X^2_{63} = 474.235, p < 0.0001). An example alignment for 5 OTUs classified to the family level as Sphingomonadaceae is presented in Figure 5b, and includes one such suspect OTU with a "TTT" motif preceding a "G" truncation position (reverse complimented).

Discussion:

We have shown that amplicon sequencing data from Illumina MiSeq instruments should be stringently filtered in order to provide the most accurate estimates of diversity. Kunin *et al.* (2010) found that diversity was grossly overestimated for their mock community data until a quality threshold of q27 was implemented. Similarly, Nelson *et al.* (2014) observed high overestimation of mock community diversity (25-125 times expected) unless the data was carefully controlled. Our optimal workflow still overestimated the OTU diversity of our simple mock communities by a factor of about 2. This slight overestimation is a dramatic improvement over that obtained by default processing, and our optimized protocol yielded a reasonable characterization of taxonomic content for mock communities (Table S4) and environmental data (Table S9) alike.

Schirmer *et al*. (2015) observed that error rates as reported by the sequencer according to the PhiX Control v3 do not accurately reflect those of amplicon sequences. Their conclusion that actual error rates were higher than those indicated by q-scores reported by the MiSeq has

important implications for the use of Illumina sequencing in estimating microbial diversity. It is possible that newer imaging strategies (*e.g.*, 2-channel SBS chemistry used by Illumina NextSeq and MiniSeq instruments) will provide improved parity between the estimated and actual error rates, but this will require careful testing to determine empirically. Of the non-target OTUs present in our optimized mock community result, one third were contaminants arising from cluster mixing events during sequencing and two thirds were sequence variants of the constituent OTUs which may have arisen during PCR, sequencing, or a combination of the two. Cluster mixing can be controlled by dual-indexing of samples (Kircher, Sawyer & Meyer, 2012), but errors arising during PCR or sequencing represent systematic errors inherent to the procedure of amplicon sequencing which are difficult, if not impossible, to completely eliminate irrespective of indexing strategy. Even though dual-indexing offers a clear advantage over single indexing with regard to sample attribution, single-indexed protocols (*e.g.*, Caporaso *et al*., 2012) remain popular and are widely used. Such data still yields valuable information and should not be discounted, as long as researchers are aware of the limitations. Dual-indexed designs should be encouraged for new research projects (*e.g.*, Kozich *et al*., 2013; Fadrosh *et al*., 2014).

We echo the recommendation by others (*e.g.*, Bokulich *et al*., 2013; Schirmer *et al*., 2015) to include control mock community samples to guide data analysis. PhiX Control v3 is still needed to improve sequence diversity for the purpose of cluster map generation (https://support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/hiseq-phix-control-v3-technical-note.pdf), but an alternative reference sequence could be used with onboard mock communities to more directly estimate error profiles for community amplicon sequencing data. PhiX sequence itself likely contributes little (if at all) to inflation of diversity estimates, and is easily quantified and removed. Though such an effect is direct evidence of cluster mixing, the rate of PhiX infiltration is likely much higher than the rate of sample mixing because PhiX Control is unindexed, producing no fluorescent signal during indexing cycles. Spurious OTUs defined from contaminating PhiX sequence may be more prevalent amid sequence data which was accompanied by higher concentrations of PhiX Control v3 during sequencing.

Though this study was not designed for careful investigation of errors generated during amplicon sequencing projects, we were able to observe that certain bases and motifs were more frequently associated with low-quality base calls than should be expected by chance. The

presence of a “TTT” or “TTC” motif immediately preceding a “G” position near the end of a sequence (near the start of the second read) was most frequently associated with an erroneous or suspect OTU (Table S7). Indeed, mock community diversity was inflated on account of this effect, but determining the source of such error requires more careful investigation than is possible here. In addition to the terminal truncation observations, we note that all other observed errors in the mock community sequences occurred within the first 100 bp of sequence, specific to the non-overlapping region of the first sequencing read (Figure 5a). It is likely that the errors we observed here would have occurred less frequently had we used fully-overlapping reads for this study. Importantly, the motif-specific patterns we observed were consistent between the mock and environmental data sets (Figure 5; Table S7; Table S8).

Estimates of alpha diversity are more sensitive than beta diversity calculations to the effects of cluster mixing and systematic errors. Input sequence quality had the most profound effect on alpha diversity estimates. Increasing the number of allowed low-quality reads (*r* parameter in `split_libraries_fastq.py`) increases the amount of data available for processing, but also changes observed diversity. For this reason, we suggest that diversity estimates should be performed only with data that has been stringently filtered for quality. Because errors in amplicon sequencing data may follow sequence-specific patterns (Schirmer *et al*., 2015; this study), spurious OTUs may provide artificial support to the statistical separation of experimental treatments when derived from OTUs driving such differences. Alternatively, spurious OTUs arising from taxa which are not differentially represented among treatments could provide artificial noise, making it more difficult to detect real differences. In either scenario, careful quality filtering can diminish such effects.

Given the vast number of studies that have already utilized Illumina sequencing for community amplicon profiling, it seems likely that estimates of alpha diversity for a wide variety of environments could be inflated due to uncontrolled error rates. Here we observed this effect with a QIIME-based workflow, though QIIME is just one of a variety of tools used in data analysis for such work. That errors may arise systematically during PCR or sequencing implies that a similar effect is likely to be observed regardless of which analysis pipeline is used to assess the data. We also made use of a high-fidelity polymerase (Phusion Hot Start II) in contrast to many studies which continue to utilize Taq polymerase, with which PCR-derived errors will be more prevalent. Lower fidelity will promote more PCR-derived errors, and those generated

during early cycles will be highly perpetuated, an effect which would be more problematic under high-cycling conditions. Because errors may follow sequence-specific patterns, some diversity estimates may be particularly inflated for certain taxa, which can further affect studies using taxonomic content to predict community function (*e.g.* Langille *et al.*, 2013). The use of phylogenetic metrics (*e.g.*, phylogenetic diversity for alpha diversity, UniFrac for beta diversity) during data analysis will likely diminish the effects of complications associated with systematically-inflated OTU diversity. Though the quality-filtering recommendations outlined by Bokulich *et al.*, (2013) have subsequently provided valuable guidance to numerous researchers, careful consideration of the results presented here and elsewhere (Kunin *et al.*, 2010; Schirmer *et al.*, 2015) will improve upon our collective interpretation of microbial diversity across environments.

Conclusions:

In this study, we observed that each of the various workflow components tested (quality filtering, OTU picking, taxonomic assignment, and OTU table filtering) affect the outcome of an amplicon sequencing project. Though high quality output can be achieved through a variety of means, in this study the optimal result was achieved with a specific set of steps. We outline them here as a general recommendation for processing community amplicon data generated on MiSeq instruments through QIIME 1.9.1 (Caporaso *et al.*, 2010a). Analysis parameters can and should be adjusted as necessary for individual data sets. The optimal workflow as performed in this study was as follows (optimized steps in bold):

1. Remove PhiX Control v3 contamination with Smalt
2. Align read pairs with fastq-join
3. **Strict quality filter in QIIME (*q* = 19, *r* = 0, *p* = 0.95)**
4. Chimera filtering with vsearch
5. Sequence dereplication with prefix/suffix OTU picker
6. **Pick OTUs with Swarm (*d*4 resolution, adjust as necessary)**
7. **Assign taxonomy with BLAST (default settings)**
8. **Filter output table at the Bokulich threshold**

Our results were consistent with the hypothesis that mock community diversity would be inflated due to the presence of PCR or sequencing errors in the data. By imposing more rigorous quality filtering of raw sequencing data, much of this error is removed. The effects of remaining errors can be minimized by utilizing a conservative similarity or distance threshold during OTU picking. By characterizing mock communities at multiple thresholds, one can identify a sufficiently conservative similarity or distance value (*d*4 in our case) which should offer improved confidence when measuring environmental diversity. If mock communities are unavailable, we advocate the use of a workflow based upon the above optimization. For studies utilizing an alternative locus, we suggest adjusting the clustering threshold based on the length of the amplicon (*e.g.*, more conservative clustering for longer amplicons) until mock communities can be employed to determine a more informed threshold.

Acknowledgements:

The authors would like to thank Linda Fitchett-Hewitt for providing axenic bacterial cultures, Dreux Patch for propagating them and extracting DNA, and the NAU Environmental Genetics and Genomics Laboratory.

References:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–10.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*, 32-46.

Anderson, M.J., Ellingsen, K.E., McArdle, B.H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters, 9*, 683–693.

Aronesty, E. (2011). *ea-utils*: Command-line tools for processing biological sequencing data; http://code.google.com/p/ea-utils

Bates, S. T., Berg-Lyons, D., Caporaso, J. G., Walters, W. A., Knight, R., & Fierer, N. (2011). Examining the global distribution of dominant archaeal populations in soil. *The ISME Journal*, *5*, 908–17.

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiology*, *10*, 189.

Bentzon-Tilia, M., Traving, S. J., Mantikci, M., Knudsen-Leerbeck, H., Hansen, J. L., Markager, S., & Riemann, L. (2015). Significant N2 fixation by heterotrophs, photoheterotrophs and heterocystous cyanobacteria in two temperate estuaries. *The ISME Journal*, *9*, 273–285.

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011). Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Applied and Environmental Microbiology*, *77*, 612–612.

Bokulich, N. A., & Mills, D. A. (2013). Improved selection of internal transcribed spacer-specific primers enables quantitative, ultra-high-throughput profiling of fungal communities. Applied and Environmental Microbiology, 79(8), 2519–26.

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, *10*, 57–9.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, P., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., & Knight, R. (2010a). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335–336.

Caporaso, J. G., Bittinger, K., Bushman, F. D., Desantis, T. Z., Andersen, G. L., & Knight, R. (2010b). PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics*, *26*, 266–267.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., 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.

Chain, P., & Grafham, D. (2009). Genome project standards in a new era of sequencing. *Science*, *326*, 1–5.

Dickie, I. A. (2010). Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist*, *188*, 916-918.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*, 2460–1.

Edgar, R. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods, 10(10), 996–998.

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*, 6.

Fichot, E. B., & Norman, R. S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome, 1(1), 10.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*, 3150–3152.

Giongo, A., Crabb, D. B., Davis-Richardson, A. G., Chauliac, D., Mobberley, J. M., Gano, K. A., Mukherjee, N., Casella, G., Roesch, L. F. W., Walts, B., Riva, A., King, G., & Triplett, E. W. (2010). PANGEA: pipeline for analysis of next generation amplicons. *The ISME Journal*, *4*, 852–61.

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., Griffiths, R. I., & Schonrogge, K. (2015). PIPITS: An automated pipeline for analyses of fungal ITS sequences from the Illumina sequencing platform. *Methods in Ecology and Evolution, 6*, 973-980.

Hallin, P. F., & Ussery, D. W. (2004). CBS Genome Atlas database: A dynamic storage for bioinformatic results and sequence data. *Bioinformatics, 20*, 3682–3686.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution, 30*, 772-780.reference

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research, 40*, e3.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology, 79*, 5112-5120.

Krohn, A. (2016). akutils-v1.2: Facilitating analyses of microbial communities through QIIME. Zenodo. 10.5281/zenodo.56764

Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology, 12*, 118–123.

Langille, M. G. I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*, 814-821.

Liu, Z., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, *36*, e120.

Lozupone, C., & Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, *71*, 8228–8235.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*:e593.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*, 610–8.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., & Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences*, *10*, 1–4.

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G., Knight, R. (2013). Advancing our understanding of the human microbiome using QIIME. *Methods in Enzymology*, *531*, 371-444.

773 Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L., & Graf, J. (2014). Analysis,
774 optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PloS*
775 *One*, *9*, e94249.
776
777 Nguyen, N. H., Smith, D., Peay, K., & Kennedy, P. (2015). Parsing ecological signal from noise
778 in next generation amplicon sequencing. *New Phytologist*, *205*, 1389-1393.
779
780 Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). Fasttree: Computing large minimum evolution
781 trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, *26*, 1641–
782 1650.
783
784 Rognes, T., Mahé, F., Flouri, T., Quince, C., & Nichols, B. (2015). vsearch: VSEARCH version
785 1.0.16. Zenodo. 10.5281/zenodo.15524
786
787 Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for
788 multiplexed target capture. *Genome Research*, *22*, 939–46.
789
790 Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into
791 biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic*
792 *Acids Research*, *43*, 1–16.
793
794 Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber,
795 C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported
796 software for describing and comparing microbial communities. Applied and Environmental
797 Microbiology, 75(23), 7537–7541.
798
799 Sogin, M. L., Morrison, H. G., Huber, J. a, Mark Welch, D., Huse, S. M., Neal, P. R., … Herndl,
800 G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere".
801 Proceedings of the National Academy of Sciences of the United States of America, 103(32),
802 12115–20.
803

804 Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid
805 assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental*
806 *Microbiology*, *73*, 5261–7.