Supplementary Information

Environmental sample collection

Water was collected from Upper Mystic Lake, (Winchester, MA, ~ 42 26.155N, 71 08. 961W) on Aug, 13, 2008 using a peristatic pump and plastic Tygon tubing. Tubing was lowered to a point ~1 m from the bottom, running the pump in reverse to prevent water from entering the tubing until the appropriate depth was reached. Water from depth was allowed to flow through the tubing for 5 minutes before 14 mls were collected into a 15 ml sterile falcon tube and immediately placed on dry ice. The first sample was taken from 22m depth and subsequent samples were taken every meter until 3m, then at 1.5m and the surface. Samples were transported on dry ice and stored at -80 °C until processing (about 1 year later).

DNA extraction

DNA was extracted as previously described (Blackburn 2010). Briefly, DNA was extracted with a modified version of the Qiagen DNeasy Blood and Tissue Kit (Qiagen, USA). Water was filtered through Swinnex filter holders onto 0.22 µm filters (Millipore, Billerica, MA). Filters were sterilely cut and added to a 2 ml screw cap tube containing 0.25 g of 0.1 mm zirconium/glass beads (MoBio Laboratories, Inc., Carlsbad, CA). 180 µl of lysis buffer consisting of 20 mM Tris HCl, 2 mM EDTA, and 1.2% Triton-X100 (pH 8.0) was added and samples were placed on a Mini Beadbeater-1 (BioSpec Products, Inc., Bartlesville, OK) for 1 minute at maximum speed. 180 µl of lysis buffer with 40 mg/ml lysozyme was added and the sample was incubated at 37 °C for 1 hr with shaking (450 rpm). 50 µl of proteinase K was added along with 400 µl of AL buffer (Qiagen DNeasy kit) without ethanol. Samples were mixed and incubated at 56 °C for 45 min. followed by a 5 min. incubation at 95 °C. Samples were centrifuged and the liquid was transferred to a sterile 1.5 ml tube. 400 µl of 100% ethanol was added and the liquid was added to the Qiagen DNeasy column. DNA was washed on the column following the Qiagen DNeasy protocol, including 500 µl wash with AW1 and AW2 and a final elution in 100 µl AE.

Mock community template preparation

To make the clone library, 16S rRNA sequences were amplified with Phusion polymerase (New England Biolabs, Ipswich, MA) and 27F and 1492R primers (Lane 1991). PCR products were cloned into the pCR Blunt II plasmid with the Zero Blunt TOPO PCR cloning kit (Invitrogen, Carlsbad, CA) and sequenced in at least one direction with Sanger sequencing (Genewiz, South Plainfield, NJ). Plasmids were purified using the plasmid DNA isolation reagent system (Carolina Biological Supply Co., Burlington, NC) and digested with restriction enzyme NotI (New England Biolabs, Ipswich, MA) to linearize the plasmid.

Illumina library preparation

Real-time PCR reactions were done first to normalize template concentrations and avoid cycling any templates past mid-log phase. PCR reactions for Illumina libraries were carried out as follows: 0.5 units of Phusion with 1 x High Fidelity buffer, 200 μM of each dNTP, 0.3 μM of PE16S_V4_U515_F and PE16S_V4_E786_R first step primers and approximately 40 ng of mixed DNA template were added for each 25 μl reaction. Additionally, 5 X SYBR Green I nucleic acid stain (Molecular Probes, Eugene, OR) was added for real-time PCR. Samples were cycled with the following conditions: denaturation at 98 °C for 30 sec annealing at 52 °C for 30 sec and extension at 72 °C for 30 sec. 14 cycles was mid-log for all samples and was subsequently used as the number of cycles for the first step PCR. The first step PCR reaction was cycled as four 25 μl reactions for each sample. PCR reactions were pooled and cleaned with Agencourt AMPure XP- PCR purification (Beckman Coulter, Brea, CA) according to the manufacture's protocol.

Illumina specific adaptors were added during a second step amplification. The conditions for the second step PCR were similar to the first step, although 4 μl of the purified first step reaction was used as a template and 0.4 μM of each PE-III-PCR-F and the barcoded reverse primer was used with 9 cycles. Samples were cycled as four 25 μl reactions and cleaned with Agencourt AMPure XP- PCR purification system. The nine libraries were sequenced in groups of three across three lanes (two flow cells) on both the Illumina GA II and HiSeq at the Biomicro

Center (MIT, Cambridge, MA) with 93 other samples per lane.

Calculation of error rate per sample

Raw data from reads with an exact match to one of the nine barcodes used for this experiment were used for comparing error rates across flow cells and lanes. This was necessary because these samples were multiplexed into lanes containing up to 93 additional unrelated samples. The raw, unfiltered fastq files were converted into a fasta file using a custom perl script. Blast was used to map the raw sequences to the mock community members, where the mock community database was trimmed to the amplified region between, but not including, the forward and reverse primer site. Raw sequences were only considered if the query and subject start and stop positions corresponded to the full length of the Illumina forward read. The perfect match, and single and double base mismatches, taken from the blast output, were calculated as a percent of the total that map to the full length Illumina sequence (Fig. S6). Sequences with less than 100% query or subject coverage were not considered in this calculation.

Commands used during processing

The following commands were used during processing.

Closed-reference clustering with QIIME (Shell):

```
#! /bin/sh
#$ -S /bin/bash
# -cwd

source /etc/profile.d/modules.sh
module load qiime-default
module load mothur
#fasta file name in QIIME format from first string after command
FASTAFILE=$1
#output directory as second string after command
OUTPUT=$2
```

```
#reference fasta file (latest greengenes OTUS)
REFERENCEFA~/greengenes/gg_12_10_otus/rep_set/97_otus.fasta
#reference taxonomies
REFERENCETAX=~/greengenes/gg_12_10_otus/taxonomy/97_otu_taxonomy.txt
PARAMS~/bin/methods_scripts/closed_ref_params.txt

echo "Start time"
date +"%m-%d-%y"
date +"%T"

pick_reference_otus_through_otu_table.py -o ${OUTPUT} -i ${FASTAFILE} -r
${REFERENCEFA} -t ${REFERENCETAX} -p ${PARAMS}

pick_rep_set.py --input ./${OUTPUT}/uclust_ref_picked_otus/*_otus.txt --
rep_set_picking_method most_abundant --fasta_file ${FAST
AFILE} -o ./${OUTPUT}/uclust_ref_picked_otus/otus_rep_set.fa

echo "End time"
date +"%m-%d-%y"
date +"%T"
```

Closed-reference QIIME parameters:

```
pick_otus:otu_picking_method uclust_ref
pick_otus:refseqs_fp  /greengenes/gg_12_10_otus/rep_set/97_otus.fasta
pick_otus:enable_rev_strand_match True
pick_otus:suppress_new_clusters      True
```

Open-reference clustering with QIIME (shell)

```
#! /bin/sh
#$ -S /bin/bash
```

```
# -cwd

source /etc/profile.d/modules.sh
module load qiime-default
module load mothur
#fasta file name in QIIME format
FASTAFILE=$1
#output folder (unique)
OUTPUT=$2
#reference fasta file (latest greengenes OTUS)
REFERENCEFA=/data/spacocha/Qiime_dir/greengenes/gg_12_10_otus/rep_set/97
_otus.fasta
#reference taxonomies
REFERENCETAX=/data/spacocha/Qiime_dir/greengenes/gg_12_10_otus/taxonomy
/97_otu_taxonomy.txt
PARAMS=/home/spacocha/bin/methods_scripts/open_ref_params.txt

echo "Start time"
date +"%m-%d-%y"
date +"%T"


pick_reference_otus_through_otu_table.py -o ${OUTPUT} -i ${FASTAFILE} -r
${REFERENCEFA} -t ${REFERENCETAX} -p ${PARAMS}

pick_rep_set.py --input ./${OUTPUT}/uclust_ref_picked_otus/*_otus.txt --
rep_set_picking_method most_abundant --fasta_file ${FAST
AFILE} -o ./${OUTPUT}/uclust_ref_picked_otus/otus_rep_set.fa

echo "End time"
date +"%m-%d-%y"
```

```
date +"%T"
```

Open-reference QIIME parameters:

```
pick_otus:otu_picking_method uclust_ref
pick_otus:refseqs_fp greengenes/gg_12_10_otus/rep_set/97_otus.fasta
pick_otus:enable_rev_strand_match True
pick_otus:suppress_new_clusters      False
```

De novo USEARCH (shell)

```
#! /bin/sh
#$ -S /bin/bash
# -cwd

#fastafile
FASTAFILE=$1
#matfile
MATFILE=$2

echo "Start time"
date +"%m-%d-%y"
date +"%T"

perl ~/bin/fasta2uchime_mat.pl ${MATFILE} ${FASTAFILE} > ${FASTAFILE}.ab
~/bin/usearch6.0.307_i86linux32 -cluster_fast ${FASTAFILE}.ab -id 0.97 -uc
${FASTAFILE}.uc
perl ~/bin/UC2list2.pl ${FASTAFILE}.uc > ${FASTAFILE}.list
perl ~/bin/list2mat.pl ${MATFILE} ${FASTAFILE}.list eco > ${FASTAFILE}.list.mat
perl ~/bin/fasta2filter_from_mat.pl  ${UNIQUE}.list.mat ${FASTAFILE} >
${FASTAFILE}.list.mat.fa
```

```
echo "End time"
date +"%m-%d-%y"
date +"%T"
```

Mothur command (batch)

```
unique.seqs(fasta=unique.uchime.remove.tocluster.fa)
align.seqs(fasta=unique.uchime.remove.tocluster.unique.fa,
reference=/data/spacocha/tmp/silva.bacteria.fasta)
screen.seqs(fasta=unique.uchime.remove.tocluster.unique.align,
name=unique.uchime.remove.tocluster.names, start=13862,
end=15958,minlength=76)
filter.seqs(fasta=unique.uchime.remove.tocluster.unique.good.align, vertical=T,
trump=.)
unique.seqs(fasta=unique.uchime.remove.tocluster.unique.good.filter.fasta,
name=unique.uchime.remove.tocluster.good.names)
system(cp unique.uchime.remove.tocluster.unique.good.filter.unique.names
final.names)
system(cp unique.uchime.remove.tocluster.unique.good.filter.names final.names)
dist.seqs(fasta=final.fasta, cutoff=0.15)
cluster(column=final.dist, name=final.names)
```

Generation of principal component analysis plots

Principal component analysis was done on the final OTU by library matrices for each clustering algorithm using QIIME beta_diversity_through_plots.py. The lowest number of sequences in a library was determined using QIIME's per_library_stats.py and input into beta_diversity_through_plots.py (-e). Trees of the representative samples were made with FastTree.

Simulated mock community data with varying error rates across libraries

To determine the impact of different error rates across libraries on distribution-based clustering performance, simulated mock community was generated using the template sequences for each members added across libraries. The total number of sequences generated was proportional to measured concentration and resulted in the creation of 748,463 total *in silico* reads. The geometric mean (R version 2.12.1; rgeom) was used to create error rates of both 0.9 and 0.8 to simulate high and low quality sequencing runs, respectively. The constant error rate dataset used in Table S5 was 0.9 for all libraries while the variable error rate dataset was 0.90 for 6 libraries and 0.08 for 3 of the libraries.

The geometric mean was used to determine which of the simulated reads would contain errors and how many errors it would contain. This was implemented in R (version 2.12.1) with rgeom using the total read count needed for each sequence and the error rate. For example, if a template was supposed to have 10 reads with an error rate of 0.8, the results would look similar to the following:

```
> rgeom(10,0.8)
 [1] 1 0 0 0 0 0 1 4 0 0
```

Where two sequences would have one bp different, one would have four mismatches and seven sequences would have no errors.

After determining how many errors to generate for each read, the position of the errors was also determined in R using the hypergeometric mean (rhyper). The distribution results in either 0 or 1 and depends on the input probability. Starting at the 3' ending position, the hypergenometric mean was used to determine whether to alter the base to another random base (1=alter, 0=evaluate next base). The probability of having an error decreased toward the 5' end to mimic sequence quality being poor at the 3' end. This was repeated until the required number of errors was generated.

Two datasets were generated in this manner. One set had a constant error rate across all libraries, and another had three libraries with a higher error rate. The dataset was clustered using the distribution based clustering algorithm as normal and the results are presented in Table S5.

Supplementary References

Blackburn MC (2010). Development of New Tools and Applications for High-Throughput Sequencing of Microbiomes in Environmental or Clinical Samples. Master of Science in Chemical Engineering thesis, Massachusetts Institute of Technology, Cambridge, MA.

Lane DJ (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniuqes in Bacterial Systemantics*. Wiley & Sons: Chinchester. pp 115-175.

Supplementary Figures

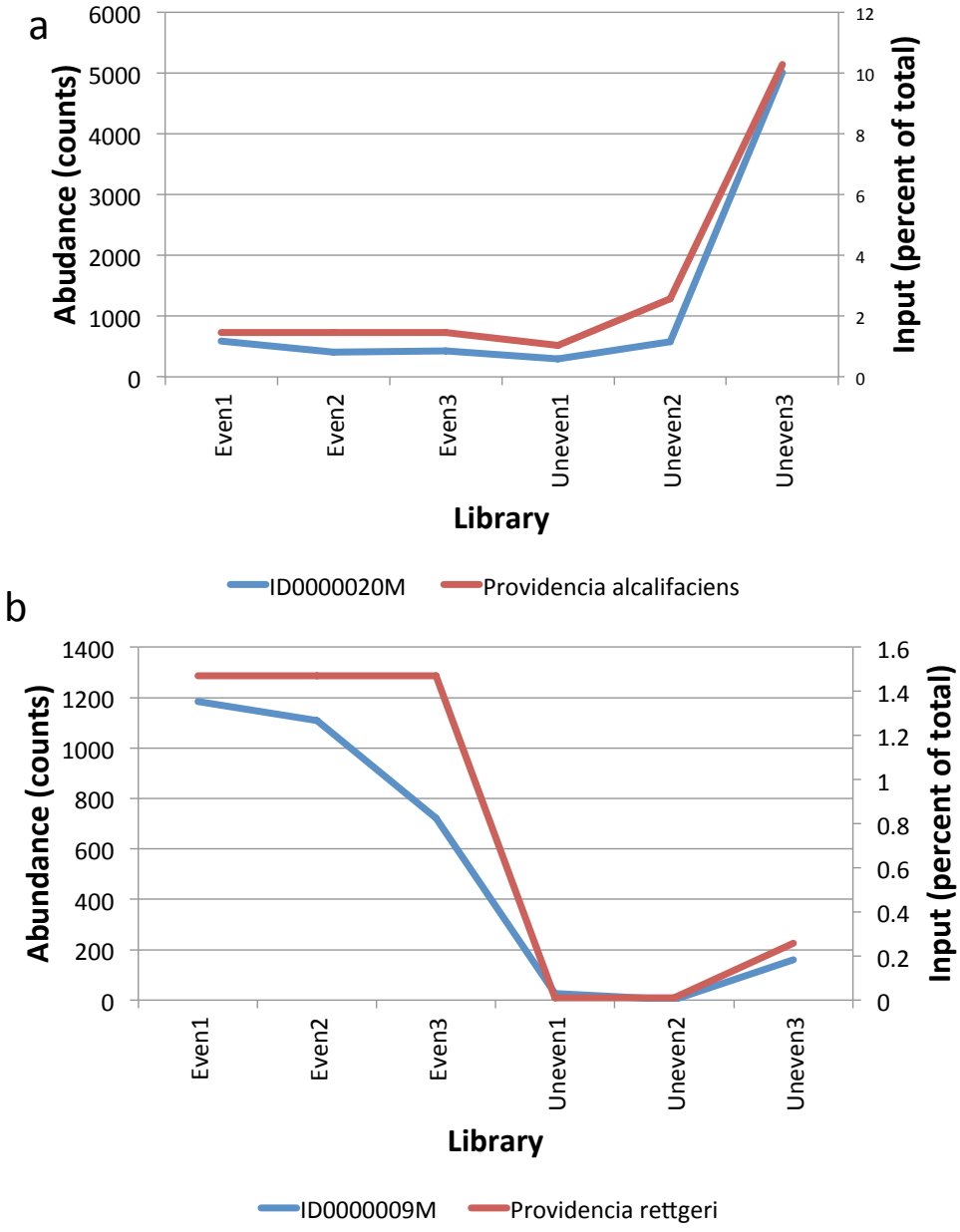


Figure S1. Unique distributions across the mock libraries help to identify a mislabeled sequence in the reference dataset. a. The distribution of a sequence matching the reference sequence labeled *Providencia rettgeri* and the input distribution of *Providencia alcalifaciens*. This sequence also matched others strains labeled *Providencia alcalifaciens* in NCBI's nr database*.* It was changed to *Providencia alcalifaciens.* b.) The distribution of another sequence which corresponds to the correct input of *Providencia rettgeri.* This other sequence also hits many other *Providencia rettgeri* strain in NCBI's nr database. This sequence was included in the analysis as the reference sequence for *Providencia rettgeri.*

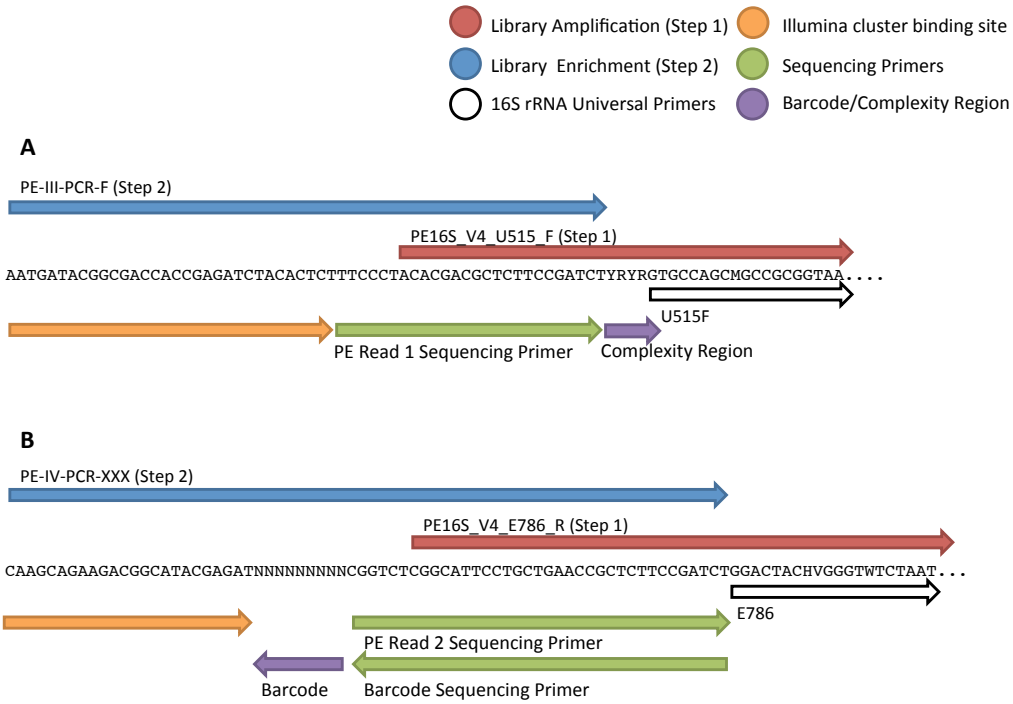


Figure S2. 16S rDNA library construct from two-step PCR. a.) 5' end of the Illumina library construct, including both first and second step forward primer sequences and sequencing primers. b.) 3' end of Illumina library construct including barcoded region and first step and second step reverse primers.

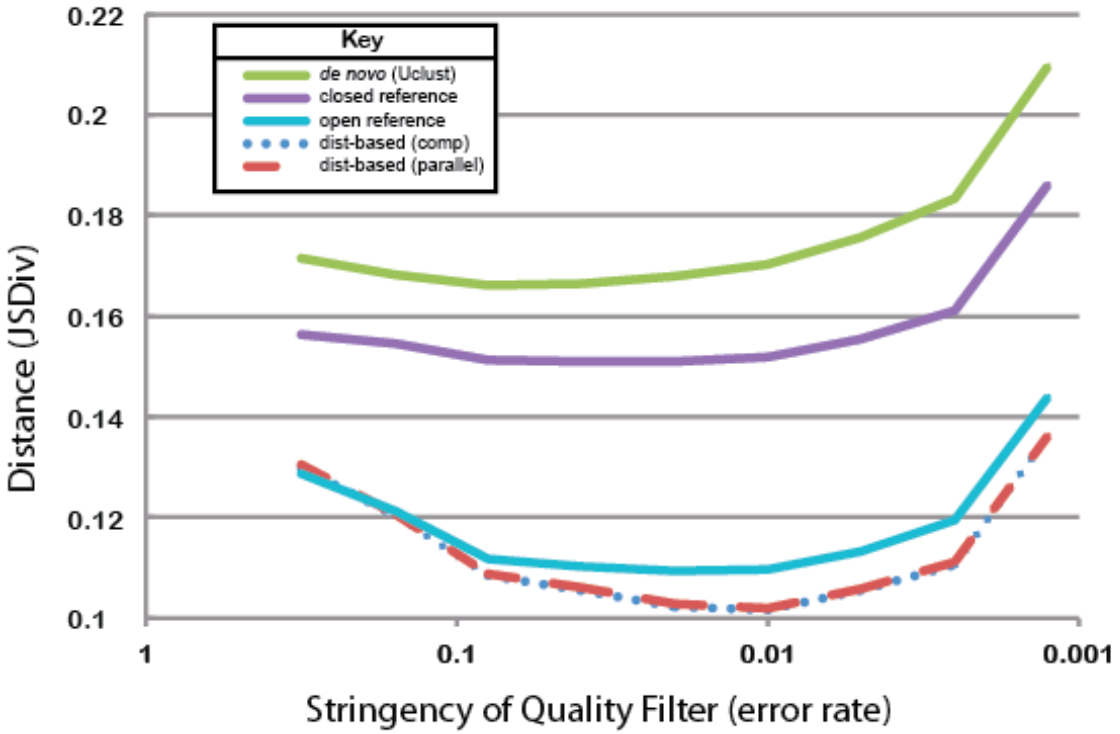


Figure S3. Distribution-based clustering results in a distribution of sequences that is most similar to the input distribution. Additionally, an intermediate amount of quality filtering results in a better representation of the input community for all clustering methods. The Jensen-Shannon divergence (JSDiv) is used as a measure of distance between the input concentration and resulting OTU counts after applying each clustering method at different levels of quality filtering. At the highest error rates, incorrect OTUs add to the distance from the true distribution. At the lowest error rates, the small number of reads kept creates the large distance values. Both parallel and complete distribution-based clustering methods result in OTUs that are most similar to the true distribution at intermediate levels of quality filter stringency.

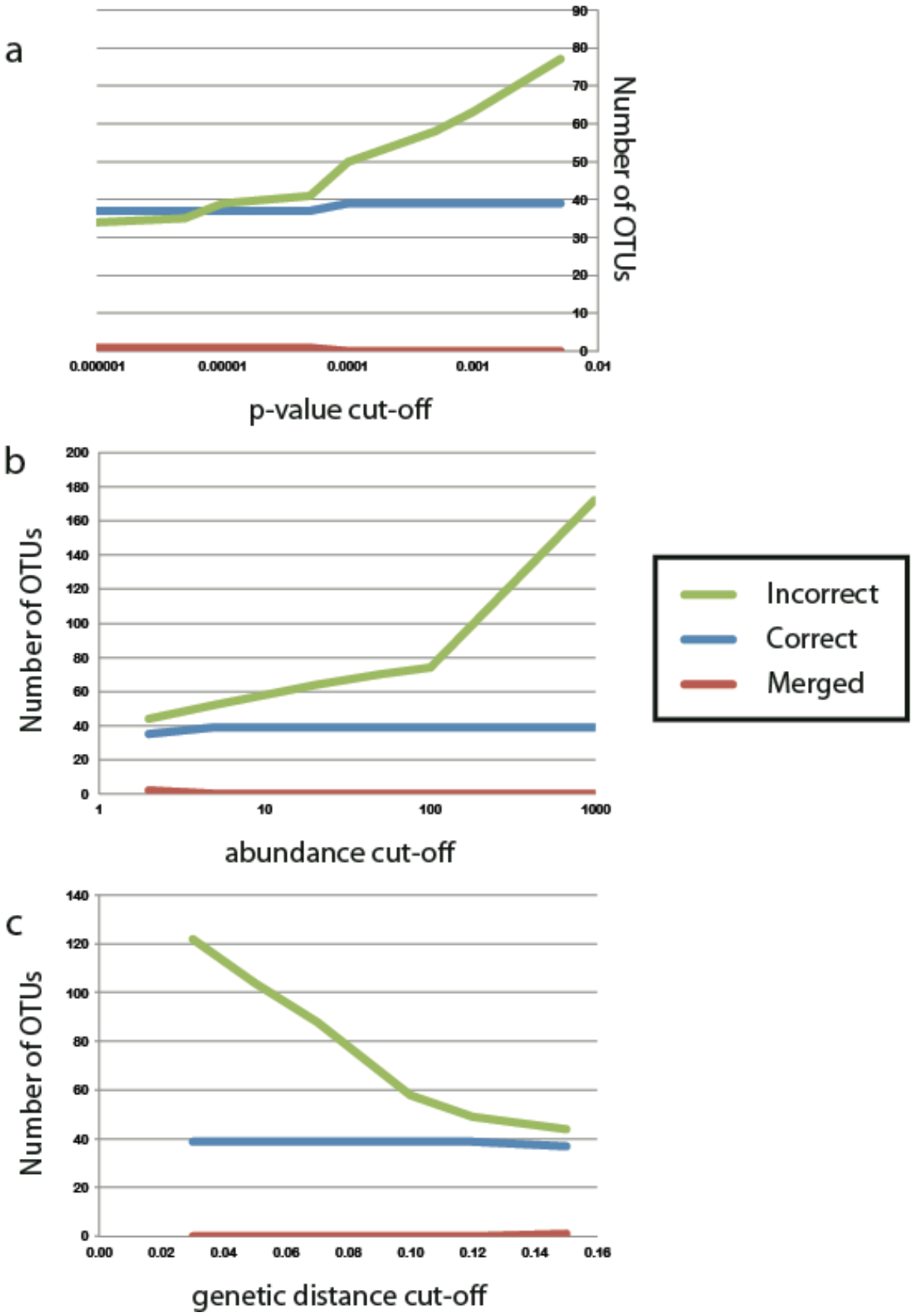
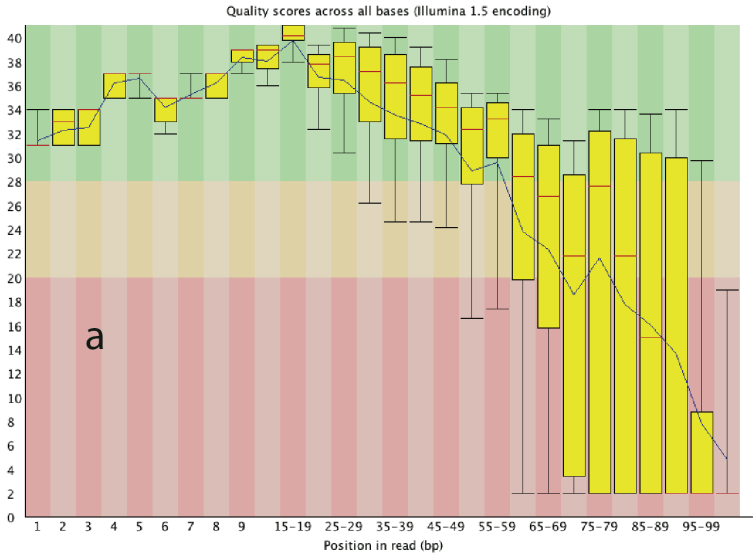
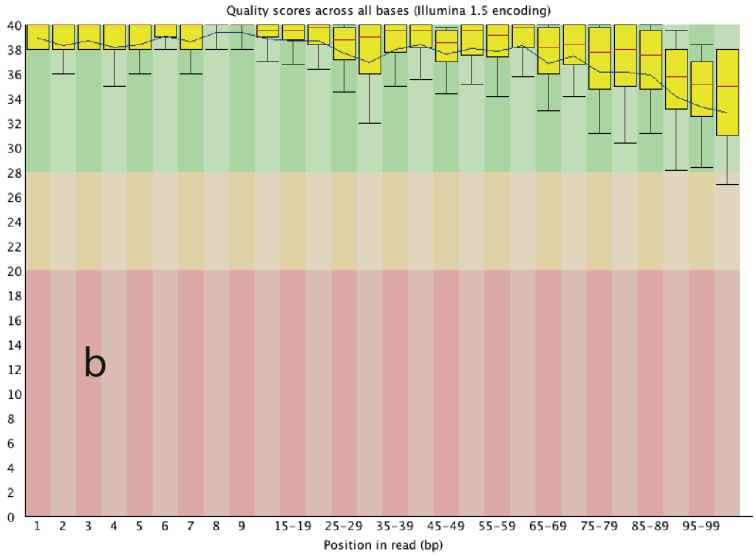

a
90
80
70
60
50
40
30
20
10
0
Number of OTUs
0.000001
0.00001
0.0001
0.001
0.01
p-value cut-off
b
200
180
160
140
120
100
80
60
40
20
0
Number of OTUs
1
10
100
1000
abundance cut-off
Incorrect
Correct
Merged
c
140
120
100
80
60
40
20
0
Number of OTUs
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
genetic distance cut-off

Figure S4. Sensitivity of the resulting OTUs to changes in the distribution-based algorithm parameters. a.) Increasing the significance cut-off value of the chi-sq test creates more incorrect OTUs whereas lower p-value cut-offs tend to merge sequences. X- axis in plotted in log scale b.) Decreasing the abundance criteria merges true input sequences with similar distributions, but increasing the cut-off to 10 mainly detects sequencing errors. X-axis is plotted in log scale c.) Lower genetic similarity cutoffs generate more incorrect OTUs, whereas at high genetic cut-off values, some mock community sequences with similar distributions are merged. "Correct" are the number of OTUs containing a single exact match to an input sequence. "Incorrect" are the number of OTUs that do not have any sequences exactly matching the input community. "Merged" are the number of OTUs that contain more than one sequence matching an input sequence.


Quality scores across all bases (Illumina 1.5 encoding)
a
Position in read (bp)


Quality scores across all bases (Illumina 1.5 encoding)
b
Position in read (bp)

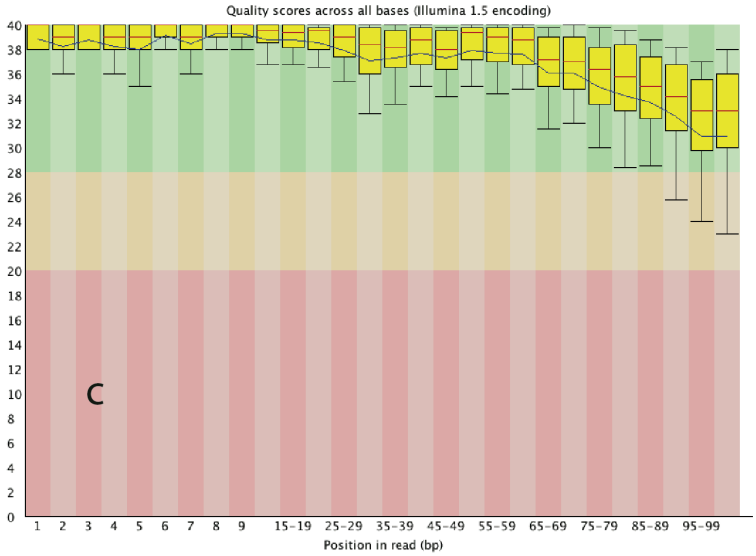

Quality scores across all bases (Illumina 1.5 encoding)
c
Position in read (bp)

Figure S5. Per base quality scores for the three lanes of Illumina. The quality of one set of samples was substantially worse than the others. (a) Flow 1, Lane 1, samples com4-com6 (b) Flow 2, Lane 1, samples com1-com3 (c) Flow 2, Lane 2, samples com7-com9.

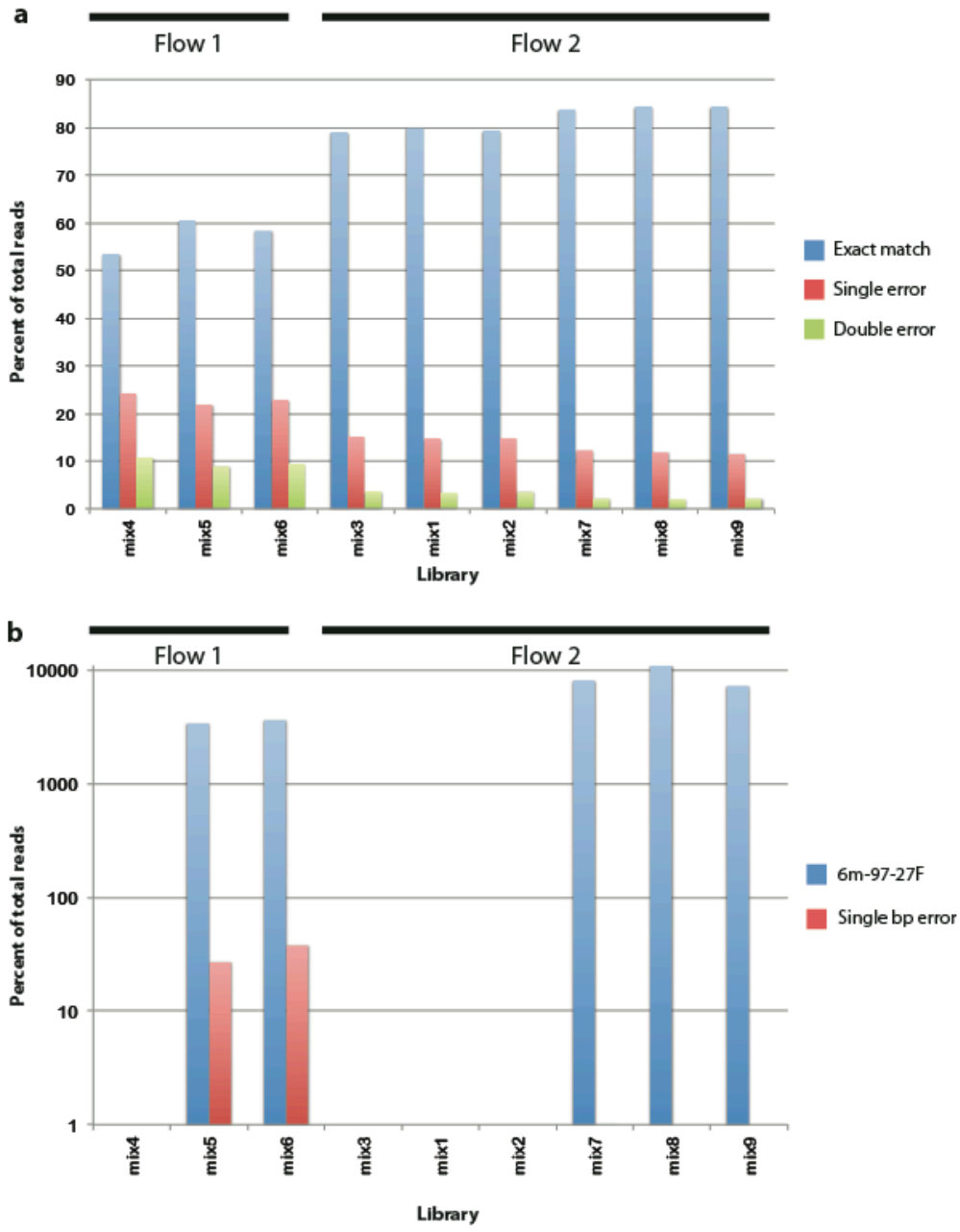
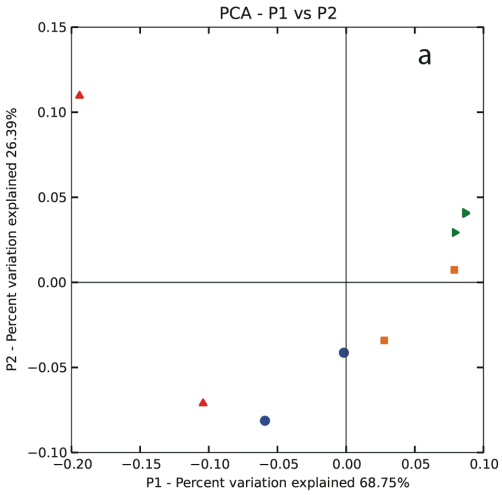
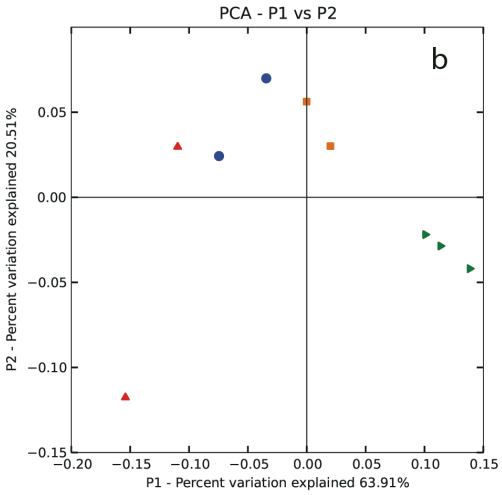
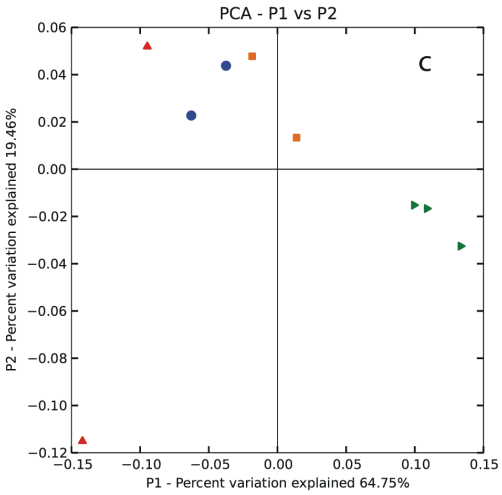


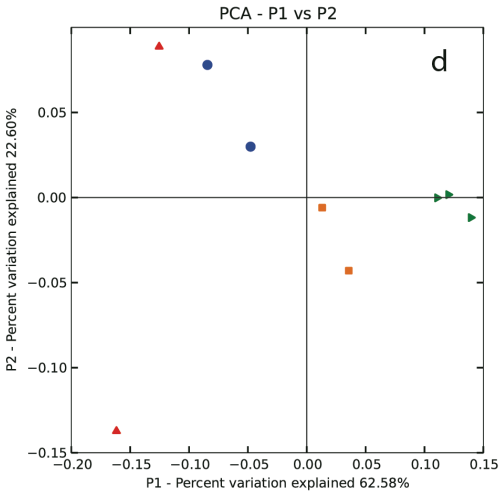
Figure S6. Error rates were higher on Flow cell 1 (Flow1) than Flow cell 2 (Flow2) causing non-random distribution of erroneous sequences across samples. (a) Exact match, single base mismatches (single error) and double base mismatches (double

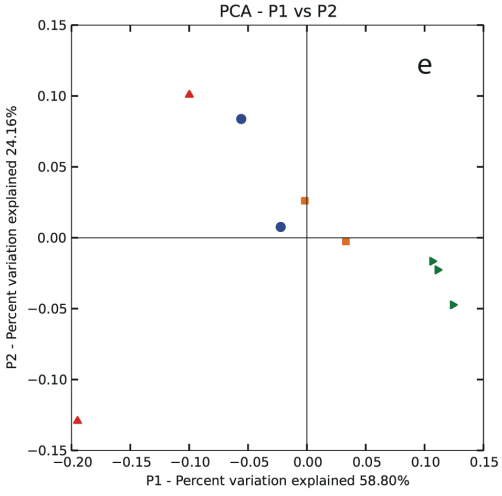
error) as a percent of the total number of raw (not quality filtered) sequences that blast to the entire 76 bp of any mock community member. Samples are labeled with the flow cell number (Flow1 or Flow2) corresponding to two different Illumina runs. Additionally, the corresponding sample name (mix1-9) is labeled on the X-axis. (b) The distribution of the true sequence (6m-94-27F) and a sequence with a single bp error sequence across samples after quality filtering and clustering. Although the single bp error sequence was generated from the true sequence, it does not have the same distribution across samples because of the difference in error rates across flow cells. Y-axis is log scale.


PCA - P1 vs P2
a
P2 - Percent variation explained 26.39%
P1 - Percent variation explained 68.75%


PCA - P1 vs P2
b
P2 - Percent variation explained 20.51%
P1 - Percent variation explained 63.91%


PCA - P1 vs P2
c
P2 - Percent variation explained 19.46%
P1 - Percent variation explained 64.75%


PCA - P1 vs P2
d
P2 - Percent variation explained 22.60%
P1 - Percent variation explained 62.58%


PCA - P1 vs P2
e
P2 - Percent variation explained 24.16%
P1 - Percent variation explained 58.80%

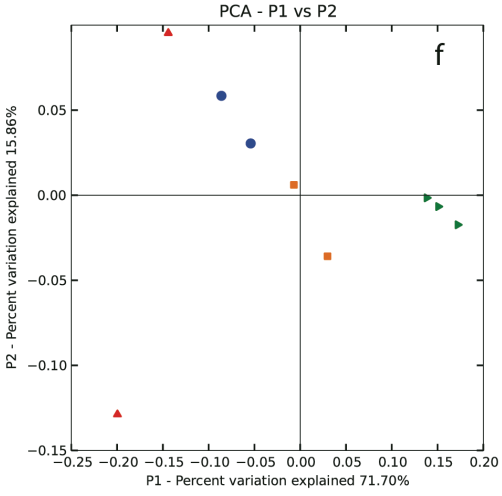

PCA - P1 vs P2
f
P2 - Percent variation explained 15.86%
P1 - Percent variation explained 71.70%

Figure S7. Principal components analysis of mock community libraries com1-com9. The primary (P1) and secondary (P2) components are plotted for the (a) true input community and for each clustering method: (b) distribution-based clustering, complete; (c) distribution-based clustering, parallel; (d) *de novo,* usearch; (e) open-reference clustering; (f) closed-reference clustering. Samples are colored according to the total number of input sequences: 1-10 input sequences, red triangle; 11-20, blue circle; 21-30 orange square; 31-40 green triangle.

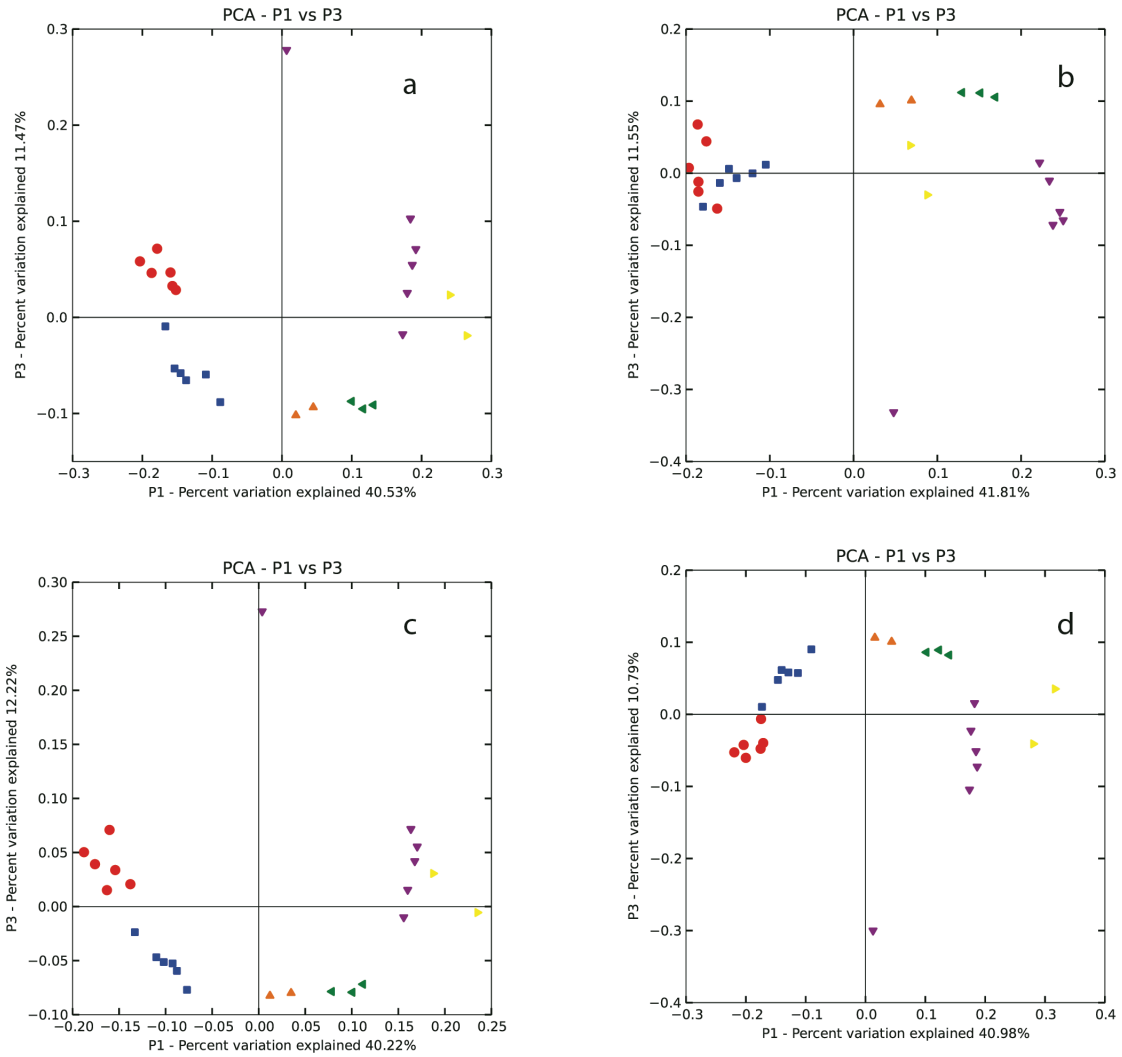


Figure S8. Principal components analysis of environmental samples from a stratified lake is similar across clustering algorithms. The primary (P1) and tertiary (P3) components are plotted for (a) distribution-based clustering, (b) *de novo* (USEARCH), (c) closed-reference and (d) open-reference clustering. Samples are colored according to depth: surface to 5 meters (m) depth, red circles; 6-10 m, blue square; 11-12 orange triangle; 13-15 m, green triangle; 16-22 m, purple triangle; surface and end blank, yellow triangle.

Supplementary Tables

Table S1. Mock community template concentrations and primer mismatches

Name	Set no.	Added to libs.[1]	Concentration (pg/ul)	Notes
21m-94-27F	1	com1-com9	22.67	NA
6m-05-27F	1	com1-com9	4.9	NA
6m-16-27F	1	com1-com9	1.988	NA
6m-10-27F	1	com1-com9	26.36	NA
21m-66-27F	1	com1-com9	38.07	2 mismatches 8 and 9 bp from 3' end of F primer
21m-32-27F	2	com2-com9	60.4875	NA
6m-09-27F	2	com2-com9	7.5625	NA
6m-44-27F	2	com2-com9	6.95	1 mismatch 1 bp from 3' end of F primer
6m-06-27F	2	com2-com9	17.6875	NA
21m-41-27F	2	com2-com9	39.5625	An additional 21.575 pg/ul of 21m-41 was added to com7, com8 and com9 as a mislabeled template.
6m-80-27F	3	com3-com9	8.0735	1 mismatch 13 bp from 3' end of the R primer
21m-90-27F	3	com3-com9	1.77	1 bp mismatch 9 bps from 3' end of F primer
6m-70-27F	3	com3-com9	1.99875	NA
6m-89-27F	3	com3-com9	28.85	NA
21m-02-27F	4	com4-com9	1.6525	2 bp mismatch 8 and 9 bp from 3' end of F primer
6m-22-27F	4	com4-com9	19	NA
6m-69-27F	4	com4-com9	8.625	NA
6m-50-27F	4	com4-com9	47.825	NA
21m-83-27F	4	com4-com9	56.05	1 mismatch 14 bp from 3' end R primer
6m-94-27F	5	com5-com9	59.775	NA
21m-25-27F	5	com5-com9	19.5	NA
21m-29-27F	5	com5-com9	38.7875	NA
21m-05-27F	5	com5-com9	10.3375	NA
6m-86-27F	6	com6-com9	0.60475	NA
21m-87-27F	6	com6-com9	36.2	1 mismatch 7 bp from 3' end of R primer
6m-65-27F	6	com6-com9	40.3625	1 mismatch 1 bp from 3' end F primer

21m-61-27F	6	com6-com9	94.7125	4 bps mismatch 12, 8, 7 and 2 bp from 3' end of F primer
6m-04-27F	7	com7-com9	3.657	NA
21m-54-27F	7	com7-com9	30	NA
6m-20-27F	7	com7-com9	17.175	NA
6m-40-27F	7	com7-com9	26.8625	NA
21m-08-27F	8	com8-com9	0.5205	2 mismatch 8 and 9 bp from 3' end of F primer
6m-81-27F	8	com8-com9	40.2875	NA
6m-13-27F	8	com8-com9	26.1625	NA
6m-52-27F	8	com8-com9	1.2065	NA
6m-75-27F	9	com9	16.675	NA
6m-82-27F	9	com9	22.675	NA
21m-68-27F	9	com9	13.875	NA
6m-19-27F	9	com9	8.85	NA
6m-87-27F	9	com9	0.8125	NA

[1] Samples were added to libraries in sequential order, starting with com1 and ending with com9. If a set was added to com1, it was also added to all subsequent libraries com2 through com9.

Table S2. Expected relative amounts of each DNA template per library

Name	Conc. (pg/ul)	com1	com2	com3	com4	com5	com6	com7	com8	com9
21m-94-27F	22.67	2.41E-01	1.11E-01	9.61E-02	6.62E-02	5.09E-02	4.78E-02	3.52E-02	3.24E-02	3.02E-02
6m-05-27F	4.9	5.21E-02	2.41E-02	2.08E-02	1.43E-02	1.10E-02	1.03E-02	7.60E-03	7.01E-03	6.54E-03
6m-16-27F	1.988	2.12E-02	9.77E-03	8.43E-03	5.81E-03	4.47E-03	4.19E-03	3.08E-03	2.84E-03	2.65E-03
6m-10-27F	26.36	2.80E-01	1.30E-01	1.12E-01	7.70E-02	5.92E-02	5.55E-02	4.09E-02	3.77E-02	3.52E-02
21m-66-27F	38.07	4.05E-01	1.87E-01	1.61E-01	1.11E-01	8.55E-02	8.02E-02	5.90E-02	5.44E-02	5.08E-02
21m-32-27F	48.39	0.00E+00	2.38E-01	2.05E-01	1.41E-01	1.09E-01	1.02E-01	7.50E-02	6.92E-02	6.45E-02
6m-09-27F	6.05	0.00E+00	2.97E-02	2.56E-02	1.77E-02	1.36E-02	1.27E-02	9.38E-03	8.65E-03	8.07E-03
6m-44-27F	5.56	0.00E+00	2.73E-02	2.36E-02	1.62E-02	1.25E-02	1.17E-02	8.62E-03	7.95E-03	7.42E-03
6m-06-27F	14.15	0.00E+00	6.96E-02	6.00E-02	4.13E-02	3.18E-02	2.98E-02	2.19E-02	2.02E-02	1.89E-02
21m-41-27F	35.2475	0.00E+00	1.73E-01	1.49E-01	1.03E-01	7.92E-02	7.43E-02	5.47E-02	5.04E-02	4.70E-02
6m-80-27F	6.4588	0.00E+00	0.00E+00	2.74E-02	1.89E-02	1.45E-02	1.36E-02	1.00E-02	9.24E-03	8.62E-03
21m-90-27F	1.416	0.00E+00	0.00E+00	6.00E-03	4.13E-03	3.18E-03	2.98E-03	2.20E-03	2.02E-03	1.89E-03
6m-70-27F	1.599	0.00E+00	0.00E+00	6.78E-03	4.67E-03	3.59E-03	3.37E-03	2.48E-03	2.29E-03	2.13E-03
6m-89-27F	23.08	0.00E+00	0.00E+00	9.78E-02	6.74E-02	5.18E-02	4.86E-02	3.58E-02	3.30E-02	3.08E-02
21m-02-27F	1.322	0.00E+00	0.00E+00	0.00E+00	3.86E-03	2.97E-03	2.79E-03	2.05E-03	1.89E-03	1.76E-03
6m-22-27F	15.2	0.00E+00	0.00E+00	0.00E+00	4.44E-02	3.41E-02	3.20E-02	2.36E-02	2.17E-02	2.03E-02
6m-69-27F	6.9	0.00E+00	0.00E+00	0.00E+00	2.01E-02	1.55E-02	1.45E-02	1.07E-02	9.87E-03	9.20E-03
6m-50-27F	38.26	0.00E+00	0.00E+00	0.00E+00	1.12E-01	8.59E-02	8.06E-02	5.93E-02	5.47E-02	5.10E-02
21m-83-27F	44.84	0.00E+00	0.00E+00	0.00E+00	1.31E-01	1.01E-01	9.45E-02	6.95E-02	6.41E-02	5.98E-02
6m-94-27F	47.82	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.07E-01	1.01E-01	7.42E-02	6.84E-02	6.38E-02
21m-25-27F	15.6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.50E-02	3.29E-02	2.42E-02	2.23E-02	2.08E-02
21m-29-27F	31.03	0.00E+00	0.00E+00	0.00E+00	0.00E+00	6.97E-02	6.54E-02	4.81E-02	4.44E-02	4.14E-02
21m-05-27F	8.27	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.86E-02	1.74E-02	1.28E-02	1.18E-02	1.10E-02
6m-86-27F	0.4838	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.02E-03	7.50E-04	6.92E-04	6.45E-04
21m-87-27F	28.96	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	6.10E-02	4.49E-02	4.14E-02	3.86E-02
6m-65-27F	32.29	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	5.01E-02	4.62E-02	4.31E-02
21m-61-27F	75.77	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.18E-01	1.08E-01	1.01E-01
6m-04-27F	2.9256	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	4.54E-03	4.18E-03	3.90E-03
21m-54-27F	24	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.72E-02	3.43E-02	3.20E-02
6m-20-27F	13.74	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.13E-02	1.96E-02	1.83E-02
6m-40-27F	21.49	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.33E-02	3.07E-02	2.87E-02
21m-08-27F	0.4164	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	5.95E-04	5.55E-04
6m-81-27F	32.23	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	4.61E-02	4.30E-02
6m-13-27F	20.93	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.99E-02	2.79E-02
6m-52-27F	0.9652	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.38E-03	1.29E-03
6m-75-27F	13.34	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.78E-02
6m-82-27F	18.14	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.42E-02
21m-68-27F	11.1	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.48E-02
6m-19-27F	7.08	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	9.44E-03
6m-87-27F	0.65	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	8.67E-04

Table S3. Barcode sequences and sequencing outline

Sample ID	Barcode Sequence	Description	Flow cell No.	Lane No.	Diversity
com1	CGAATAT	E8, plate 63umP2	1	1	
com2	AAGGAAC	E9, plate 63umP2	1	1	
com3	GATTGAA	E10, plate 63umP2	1	1	
com4	CCGCACC	H1, plate 63umP1	2	1	
com5	ATGCCAG	H2, plate 63umP1	2	1	
com6	TCGAACA	H3, plate 63umP1	2	1	
com7	GTACGTT	H10, plate 63umP3	1	2	
com8	AGTAGAT	H11, plate 63umP3	1	2	
com9	TCATTAA	H12, plate 63umP3	1	2	

Table S4. Correlation of OTUs from various clustering methods with matching Sanger environmental clone sequence

Sanger Clone	USEARCH correlation	Open-reference correlation	Closed-reference correlation	DBC correlation
21m-02-27F	0.999831867	0.988437304	NA	NA
21m-03-27F	0.999999465	0.999768514	0.999768514	0.999953024
21m-04-27F	0.982567028	0.999991003	0.999991003	0.999968253
21m-05-27F	0.99998637	0.999952053	NA	0.999967953
21m-08-27F	0.998120162	0.996556495	NA	0.999521037
21m-09-27F	0.999922215	0.999952171	0.999952171	0.999858585
21m-11-27F	0.997205609	0.994006215	NA	0.947095652
21m-13-27F	0.999971592	0.999971592	0.999971592	0.999973636
21m-14-27F	NA	NA	NA	NA
21m-22-27F	0.999999929	0.999995556	0.999995763	0.999993992
21m-24-27F	0.999936853	0.999945906	NA	0.999948108
21m-29-27F	0.999999379	0.999996912	0.999996912	0.999995608
21m-30-27F	0.999044469	0.999185914	NA	0.99998369
21m-31-27F	0.999999836	0.999997084	0.999997084	0.999997146
21m-32-27F	0.999995376	0.999458687	0.999458687	0.999969546
21m-36-27F	0.999999722	0.999964949	NA	0.999955969
21m-40-27F	0.999999988	0.999998671	0.999998649	0.99999804
21m-41-27F	NA	0.997646314	0.997646314	0.999758171
21m-45-27F	0.999999156	NA	NA	0.99995629
21m-48-27F	0.976110722	1	0.99750752	0.976621605
21m-49-27F	0.999982722	NA	NA	0.999980964
21m-52-27F	NA	NA	NA	0.999698894
21m-60-27F	0.999999952	0.998506324	0.998504326	0.999990881
21m-63-27F*	**0.880625476**	**0.825360924**	**0.822861822**	0.990520344
21m-65-27F	0.986325434	0.999990787	0.999990787	0.999895439
21m-66-27F	NA	NA	NA	NA
21m-67-27F	0.999999957	0.99999697	0.999996992	0.999966109
21m-68-27F	0.99999981	0.99998327	NA	0.999983149
21m-70-27F	0.999999359	0.999996268	0.999996268	0.999991529
21m-71-27F	0.99985004	0.999995199	0.999995199	0.999771713
21m-72-27F	0.99999655	0.996681307	NA	0.999996955
21m-76-27F	0.999999988	0.999998354	0.999998363	0.999996232
21m-81-27F	0.999999928	0.999987794	0.999986911	0.999978358
21m-82-27F	NA	0.999832758	0.999832758	0.999819939
21m-83-27F	0.999579834	0.999993462	0.999993667	0.99995589
21m-84-27F	0.999999816	0.999986865	0.999986865	0.999983076
21m-85-27F	**0.850149712**	0.999982378	0.999981951	0.999993487

21m-86-27F	NA	0.99682474	0.99682474	0.999970848
21m-87-27F	0.999999946	0.999459162	0.999459488	0.999991363
21m-91-27F	0.999999143	0.999733801	0.999734327	0.999997782
21m-92-27F	0.999999993	0.99999712	0.999997098	0.999997873
21m-94-27F	0.999965615	0.999878505	0.999876502	0.999574867
6m-02-27F	0.948124255	0.999886989	0.999886989	0.999973135
6m-04-27F	0.9999997	0.999993308	0.999993308	0.999993803
6m-05-27F	0.999999809	0.999973188	0.999973188	0.999969356
6m-06-27F	0.999999945	0.99999637	0.999996606	0.999994691
6m-09-27F	0.999999932	0.999996944	0.999996945	0.99999668
6m-10-27F	0.999700326	0.999044474	0.999044474	0.999762791
6m-13-27F	0.999999965	0.999990493	0.999990493	0.999983703
6m-14-27F	0.999999946	0.999995685	0.999995721	0.999993595
6m-15-27F	0.962483452	0.999985389	0.9999855	0.999964317
6m-16-27F	0.99999992	0.999903105	NA	0.999982484
6m-17-27F	0.999991884	0.999662611	0.999662787	0.999992086
6m-19-27F	0.999999916	0.999959879	0.999959486	0.999940495
6m-22-27F	0.97744411	0.998862541	0.998893941	0.999137677
6m-27-27F	0.999999868	0.955022707	0.955057044	0.99998669
6m-28-27F	0.99874705	0.99855622	NA	0.998620463
6m-29-27F	0.999979883	0.999973291	0.999973291	0.999948297
6m-30-27F	0.9999976	0.999981418	0.999981418	0.999970075
6m-33-27F	0.99999998	0.999990445	0.999990445	0.999990895
6m-34-27F	0.999980535	0.999999831	0.999999125	0.999981344
6m-37-27F	0.999999308	0.999885815	0.999885815	0.99989468
6m-39-27F	0.903636547	0.99999425	0.999994191	0.999996968
6m-40-27F	0.999999319	0.999995966	0.999995966	0.999992703
6m-41-27F	0.999713983	0.999604776	0.999604776	0.999953389
6m-43-27F	0.99996984	0.999970768	0.999970768	0.999972493
6m-44-27F	**0.881376523**	0.999989511	0.999989511	0.999989323
6m-50-27F	NA	0.999867733	0.999867733	0.999823275
6m-51-27F	0.999999964	0.999966258	NA	0.999981791
6m-53-27F	0.999999969	0.999970441	0.999971235	0.999986506
6m-56-27F	**0.829520279**	0.999994788	0.999994778	0.999974311
6m-58-27F	0.999999702	0.999985723	0.999985723	0.999985641
6m-59-27F	0.999999212	0.999929601	0.99992735	0.999918721
6m-63-27F	0.99998016	0.999983384	0.999983384	0.999920585
6m-64-27F	0.999999372	**0.789783151**	**0.789950812**	0.999990593
6m-65-27F	0.999999759	0.999915811	0.999915811	0.99995473
6m-66-27F	**0.877146275**	**0.845708551**	**0.845780711**	0.999983645
6m-70-27F	0.999999199	0.999989507	0.999989513	0.999999429
6m-74-27F	0.999994474	0.997338006	NA	0.999741203

6m-75-27F	0.99998019	0.999686327	0.999686327	0.996771341
6m-77-27F	0.999998362	0.999985919	0.999986018	0.999986266
6m-79-27F	0.998191658	0.999811502	0.999811502	0.999399639
6m-81-27F	0.999999983	0.999999171	0.999999173	0.999997648
6m-83-27F	NA	0.999699594	0.999699594	0.99972764
6m-84-27F	0.999999864	0.99999772	NA	0.999997007
6m-85-27F	0.99999995	0.999989398	0.999989355	0.999996201
6m-87-27F	0.999999702	NA	NA	0.999977672
6m-91-27F	0.99999992	NA	NA	0.999987023
6m-94-27F	0.999931904	0.999929312	0.999929312	0.999935375

* Clone names and the corresponding correlations below 0.9 are in bold

Table S5. Correct and incorrect OTUs predicted by distribution-based clustering on simulated data

Dataset	**Correct OTUs**	**Incorrect OTUs**
Constant error rate[1]	40	3
Variable error rate[2]	40	157

[1] Error rate generated from a geometric distribution was 0.9 for 9 libraries
[2] Error rate was 0.8 across 3 libraries and 0.9 across 6 libraries