Microb Ecol (2013) 65:709–719
DOI 10.1007/s00248-012-0145-4

METHODS

Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates

Nastassia V. Patin · Victor Kunin · Ulrika Lidström · Matthew N. Ashby

Received: 22 May 2012 / Accepted: 26 November 2012 / Published online: 12 December 2012

Abstract Next-generation sequencing has increased the coverage of microbial diversity surveys by orders of magnitude, but differentiating artifacts from rare environmental sequences remains a challenge. Clustering 16S rRNA sequences into operational taxonomic units (OTUs) organizes sequence data into groups of 97 % identity, helping to reduce data volumes and avoid analyzing sequencing artifacts by grouping them with real sequences. Here, we analyze sequence abundance distributions across environmental samples and show that 16S rRNA sequences of >99 % identity can represent functionally distinct microorganisms, rendering OTU clustering problematic when the goal is an accurate analysis of organism distribution. Strict postsequencing quality control (QC) filters eliminated the most prevalent artifacts without clustering. Further experiments proved that DNA polymerase errors in polymerase chain reaction (PCR) generate a significant number of substitution errors, most of which pass QC filters. Based on our findings, we recommend minimizing the number of PCR cycles in DNA library preparation and applying strict postsequencing QC filters to reduce the most prevalent artifacts while maintaining a high level of accuracy in diversity estimates. We further recommend correlating rare and abundant sequences across environmental samples, rather than clustering into OTUs, to identify remaining sequence artifacts without losing the resolution afforded by high-throughput sequencing.

Introduction

Microbes represent the overwhelming majority of life on Earth, but the inability to culture most taxa in the laboratory means they remain understudied and difficult to quantify [1]. The use of DNA sequencing to investigate 16S rRNA gene diversity has produced continuous discoveries of new microbial taxa in environments as diverse as the deep sea and the human microbiome [1–3]. New high-throughput sequencing technologies have vastly expanded the amount of taxonomic data that can be acquired from a single environmental sample [3–5], and this has facilitated studies showing that microbial diversity has previously been underestimated by orders of magnitude [3, 6]. However, the production of massive amounts of sequence data accentuates the need to accurately distinguish sequences containing errors (artifacts) from sequences making up true diversity [3, 7–9]. Two crucial components of sequence data analysis are discussed in this study: using environmental distribution patterns rather than sequence similarity to group sequences into taxonomic entities and minimizing sources of potential error.

Errors can arise at different stages of the DNA library preparation and sequencing process. The polymerase chain reaction (PCR) used to prepare DNA for pyrosequencing can be a source of both point substitutions as well as chimeras, which are formed when partially extended sequences act as primers in the next round of amplification [10, 11]. Chimeric sequences can be effectively identified in 454 amplicon data sets using existing programs [10, 12, 13]. Experimental

Electronic supplementary material The online version of this article (doi:10.1007/s00248-012-0145-4) contains supplementary material, which is available to authorized users.

N. V. Patin · M. N. Ashby
Romberg Tiburon Center for Environmental Studies,
San Francisco State University, 3152 Paradise Drive,
Tiburon, CA 94920, USA

V. Kunin · U. Lidström · M. N. Ashby
Taxon Biosciences, Inc., 3150 Paradise Drive,
Tiburon, CA 94920, USA

Present Address:
N. V. Patin (✉)
Center for Marine Biotechnology and Biomedicine, Scripps
Institution of Oceanography, University of California, San Diego,
La Jolla, CA 92093-0204, USA
e-mail: nvpatin@ucsd.edu

conditions like primer choice, number of cycles, and the specific region of amplification can also affect sequencing results and subsequent ecological analyses [11, 14, 15]. One of the most important sources of error is misincorporation of bases by the DNA polymerase enzymes used to synthesize new molecules. Although the fidelity of these enzymes is extremely high, PCR by its nature occurs on a massive scale and mistakes early in the amplification cycle can multiply exponentially [16, 17]. Further, the high-throughput sequencing process itself is an important source of artifacts. Pyrosequencing acts by flowing each base sequentially through a plate of microscopic wells, each of which contains a bead attached to millions of single-stranded DNA molecules encoding the sequence of interest. The incorporation of a base into the complementary strand produces an enzyme cascade leading to a light reaction, which is registered and recorded [4]. Errors have been found to occur most frequently in homopolymers (single-base repeats) because light intensities do not necessarily increase linearly with the number of consecutive bases [7, 9]. Several studies have focused on improving post-sequencing bioinformatic analysis methods to minimize the danger of including artifacts in environmental data sets [18–22]. Regardless of the source, sequencing errors produce artifacts in the data that lead to uncertainty about microbial community richness and composition.

The prevalence of artifacts in large sequence data sets has led to concerns that microbial diversity and richness estimates have been greatly overestimated [8, 23]. The standard method of analyzing community sequence information is to cluster 16S rRNA gene sequences with 97 % or higher identity into operational taxonomic units (OTUs) [3, 8, 18, 24], which also serves to minimize such overestimation. In the absence of a universally accepted species concept for Bacteria, this method of taxonomic classification is considered sufficient to generate estimates of community richness and evenness by using highly similar gene sequences as proxies for phylogenetic distance. While this approach does eliminate many potential sequencing errors, it is done at the expense of sample resolution and may combine organisms with distinct biological properties and capabilities into a single OTU. Since clustering can amalgamate several species and obscure true abundance levels, it may subsequently affect community evenness along with richness estimates. Despite the recent improvements in bioinformatic filtering and clustering algorithms [10, 20, 22, 25], the use of OTUs as a basis for ecological interpretation remains inherently flawed.

While the 16S rRNA gene is a useful marker for taxonomic divisions among Bacteria and Archaea, it is not a definitive proxy for biological identity. Horizontal gene transfer can confer important cellular functions faster than the evolutionary rate of the 16S rRNA gene by transferring entire gene clusters in a single recombination event [26]. For example, the K-12 and O157:H7 strains of *E. coli* share two identical copies of the V5–V6 region of the 16S rRNA gene, but O157:H7 exhibits an extreme pathogenicity that has made it a worldwide public health threat, absent in K-12 and thought to have originated from horizontal gene transfer [27]. Considering this important difference between strains with identical 16S rRNA genes, even slight variations in the variable regions potentially represent significant genetic deviation. Grouping sequences that differ by up to 3 % into single taxonomic units, therefore, creates a risk of overlooking important microbial diversity. However, next-generation sequencing errors remain problematic, with the inverse risk of treating artifacts as true strains [7, 8]. One way of overcoming this risk is to use individual sequence distributions in a set of environmental samples. A correlation between distribution patterns of two similar sequences is likely to indicate either a sequencing error or a different 16S operon from the same genome. While the latter possibility must be considered, it does not detract from the inverse risk, namely, that two sequences with variant distributions have a high probability of being unique phylotypes. Such an analysis can be further supported by an understanding of how artifacts arise and a method of minimizing them during the sequencing process.

We investigated the hypothesis that OTU clustering can merge distinct organisms into the same taxonomic group, thereby presenting an underestimate of true microbial diversity. We further hypothesized that organisms have unique distribution patterns across environmental samples, and these patterns can be used to distinguish sequencing errors from phylogenetic diversity. In order to perform this study, we used sequence data from natural assemblages because synthetic communities would lack the element of natural distribution. In addition, we examined sources of errors in 16S rRNA gene sequence data sets using Roche 454 GS-FLX Titanium technology. We used a single-template sequence to focus specifically on point substitutions, insertions, and deletions, and thus, did not examine the occurrence of PCR chimeras. Based on our findings, we propose ways to minimize such errors through careful sample preparation and stringent sequence quality filtering.

Methods

OTU Clustering Experiment

Sample Information and DNA Library Preparation

The test samples were comprised of 18 piston cores taken from marine sediments just below the sea floor in the Green Canyon Block 607/608 area of the Gulf of Mexico in August 2006 [28, 29]. DNA was isolated from each sample using a bead beating protocol with phenol:chloroform:isoamyl alcohol (Supplementary Online Material), from which the V5–V8 region of the 16S rRNA gene was amplified and sequenced using multiplex 454 Titanium technology and sample-specific barcodes [25], as follows:

PCR cycle titration was used in a diagnostic step to determine the lowest number of cycles needed to amplify sufficient template material, followed by two rounds of amplification to minimize any bias in adaptor modification. All reagents were obtained from Invitrogen (Carlsbad, CA, USA) except for the primers and milliQ water. The universal primers TX9 (GGATTAGAWACCCBGGTAGTC) [30] and 1391R (CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) [31] were used to amplify the V5–V8 variable region (Fig. 1). Microfuge tubes, PCR tubes, 10× HiFi Buffer, bovine serum albumin (BSA), $MgSO_4$, and milliQ water were UV-sterilized on the day of each experiment. Thermocycling in all reactions proceeded with 94 °C for 1 min, followed by varying numbers of cycles of 94 °C for 30 s, 55 °C for 30 s, and 68 °C for 1 min, and final extension at 68 °C for 1 min. For the cycle titration, each 20-μl reaction consisted of the following components: 2 μl 10× HiFi Buffer, 1 μl forward primer TX9 (8 μM), 1 μl reverse primer 1391R (8 μM), 1 μl dNTPs (4 mM), 0.8 μl $MgSO_4$ (50 mM), 0.8 μl BSA (10 mg/ml), 0.08 μl Platinum *Taq* HiFi, 9.32 μl sterile milliQ water, 4 μl template DNA (1 ng/μl). Four reactions were run for each sample, each for a different number of cycles to encompass 12, 18, 24, and 30 rounds of amplification. According to the titration results, evaluated on a diagnostic 0.9 % agarose gel, the minimum number of PCR cycles necessary to generate sufficient product for the subsequent step was performed for the first stage of amplification. Each sample was evaluated on an individual basis because the minimum cycle number can vary based on quality of the DNA, presence of PCR inhibitors, and so forth. This groundtruthing step, therefore, served to relativize the quality and quantity of the PCR products. Amplification reactions were run in four replicates of 50 μl for each sample, with the following components: 5 μl 10× HiFi Buffer, 2.5 μl forward primer TX9 (8 μM), 2.5 μl reverse primer 1391R (8 μM), 2.5 μl dNTPs (4 mM), 2 μl $MgSO_4$ (50 mM), 2 μl BSA (10 mg/ml), 0.2 μl Platinum *Taq* HiFi, 23.3 μl sterile milliQ water, and 10 μl template DNA (1 ng/μl). This was followed by 0.9 % agarose gel purification and recovery with a gel extraction kit according to the manufacturer's instructions (Qiagen, Valencia, CA, USA). A second, 7-cycle PCR was employed to add the 454 GS-FLX Titanium A- and B-Adaptors and sample-specific barcodes onto the PCR products using fusion primers. Each reaction was run in four replicates of 100 μl each, with the following components: 10 μl 10× HiFi Buffer, 5 μl dNTPs (4 mM), 4 μl $MgSO_4$ (50 mM), 4 μl BSA (10 mg/ml), 0.8 μl reverse primer 1391R (with B-adaptor; 100 μM), 0.8 μl forward primer TX9 (with A-adaptor and sample-specific barcode; 100 μM), 0.4 μl Platinum *Taq* HiFi, 55 μl sterile milliQ water, and 20 μl template DNA (0.35 ng/μl). The second PCR product was purified on a 8 % polyacrylamide gel electrophoresis–Tris–borate–ethylenediaminetetraacetic acid (EDTA) (PAGE-TBE) gel and isolated by the following steps: excision of the roughly 700-bp amplicon (including the barcode and adaptors), fragmentation of the excised gel, diffusion of the DNA from the gel, and recovery of DNA with a PCR purification kit according to the manufacturer's instructions (Qiagen, Valencia, CA, USA). Barcoded PAGE-purified amplicons were merged into a sequencing library and processed according to the Roche 454 Titanium chemistry protocol using the 454 Life Sciences Lib-L kit and shotgun processing pipeline [4].

Custom Quality Filtering

After passing the 454 signal processing filters (GS-FLX Software v. 2.0.01.12), all sequences were trimmed to remove the TX9 primer sequence at the 5′ end and at an assigned point corresponding to a location within the conserved region following the V6 region at the 3′ end. Thus, the reads that comprised the final data set spanned the V5–V6 region of the 16S rRNA gene and were approximately 250 bp long. Post-trimming, all sequences were subjected to in-house quality control (QC) filters that relied on 454 quality scores, or Q scores. The Q scores are based on a log scale, so that any nucleotide with a Q score of 20 (i.e., Q20) has a 1 % chance of error, Q30 has a 0.1 % chance, and so on [4]. The in-house QC filters required every base to have a Q score of 20 or higher. In addition to the individual base minimum score, 90 % of the bases in the trimmed sequence were required to be Q25 or higher.

Statistical Analysis

No clustering was performed so that each unique sequence was treated as a distinct organism. From this data set, 20 sequences that differed by at most two base pairs, the

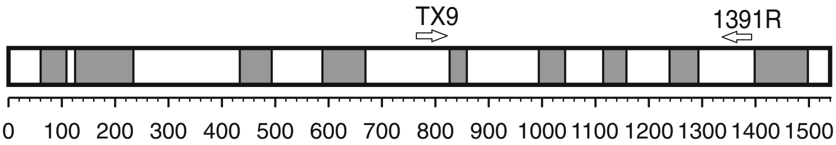

Fig. 1 The 16S rRNA gene. The variable V5 and V6 regions were flanked by primers TX9 and 1391R in all PCRs performed during library preparation, amplifying a region of approximately 600 bp. All sequences were trimmed, postsequencing, at the end of TX9 and a set trim point within the conserved region following V6, producing reads with an approximate length of 250 bp

equivalent of >99 % identity, were chosen for analysis of abundance distribution patterns. The abundance data for each sequence were relativized across samples to correspond to the number of times each sequence was observed per 10,000 sequences examined (Table 1). The most abundant sequence over all 18 sampling locations was used as a reference for all subsequent sequences, and the resulting R^2 values are shown in the final column of Table 1. Three of these sequences were chosen for further statistical analysis (SOM).

Error Rate Experiment

Sequence Choice

The V5–V6 region of the 16S rRNA gene from *Actinoallomurus GMKU370* (*Streptosporanginiae* spp.) was used as the sequence template (SOM). This sequence contains 15 homopolymers, where a homopolymer is defined as three or more consecutive identical bases. Seven of these elements consist of three bases, seven have four, and one has five bases. Thus, this sequence is considered less prone to errors than those with more and longer homopolymers.

Plasmid Preparation, 454 Library Preparation, and Sequencing

The source plasmids were prepared from a natural population, potentially containing variant operons of the 16S rRNA gene within one genome. Cloning was, therefore, performed as the first step in order to isolate a single-template sequence (SOM).

Plasmid DNA was divided into six aliquots, and thereafter, each aliquot was treated as a separate sample to replicate the results in one sequencing run. The V5–V8 region of the 16S rRNA gene was amplified and purified as described for the OTU clustering experiment (SOM). All samples underwent 16 PCR cycles in the first amplification, and 7 cycles in the second amplification. The DNA library was prepared for sequencing in accordance with the Roche 454 protocol. Barcoded sequences were mixed in different concentrations to produce varying amounts of sequence data from each sample (Table 2), and the library went through four 454 Titanium runs.

Data Analysis

All sequence data were filtered as previously described in “Custom Quality Filtering”, and the sequences from all six plasmid samples were compiled into two separate data sets. One set consisted of sequences rejected by the in-house quality filters and one set contained all sequences that passed the filters.

Each data set was manually analyzed to identify contaminations using the alignment program Seaview [32]. Sequences that appeared to differ significantly from the parent sequence were checked using Basic Local Alignment Search Tool (BLAST) [33] and removed if they were identified as contaminants. The number and type of errors for each data set were then manually counted, with errors classified either as miscalls, insertions, or deletions. Whether or not the error occurred in a homopolymer was also noted.

PCR Cycle Experiment

Sequence Choice

Two 16S rRNA gene fragments were chosen to test the effect of varying numbers of PCR cycles on sequencing errors. Both sequences originated from an environmental soil sample and were prepared as plasmid inserts as described for the preceding “Error Rate Experiment”. The plasmids were designated pPT143 and pPT42C, respectively (SOM).

DNA Preparation: PCR Cycle Variation

As with pPT343, chemically competent *E. coli* cells were transformed with two different plasmids containing Amp^R, pPT42C, and pPT143 and plated on LB-Amp. Three single colonies were picked from each plasmid sample and each colony was grown overnight in a separate tube of 1-ml LB-Amp. One culture was chosen for each plasmid and DNA was isolated using a DNA miniprep kit according to the manufacturer's instructions (Qiagen, Valencia, CA, USA). The DNA underwent the same library preparation as previously described. However, to test the effect of PCR cycle number on miscalls, the number of cycles in the first round of amplification was varied so that each plasmid sample was divided into subsamples, each undergoing a different number of cycles (Table 3). All samples received 7 cycles in the second amplification. The volumes and DNA concentrations used in the PCR reactions were calculated so that the post-PCR concentrations of the amplicon products were equivalent for all subsamples in each plasmid sample. The remainder of the library preparation protocol was identical for all subsamples.

DNA Preparation: Zero PCR Cycle Experiment

In order to obtain enough amplicon DNA without generating PCR errors, samples for this experiment were prepared separately. Plasmid DNA was linearized using the restriction enzyme *Xba*1 according to the manufacturer's protocol (New England BioLabs, Ipswich, MA, USA). The linearized plasmid DNA was gel purified on a 0.9 % agarose gel

Table 1 Twenty sequences selected from a 454 data set of 16S rRNA gene abundances in 18 marine sediment samples, in decreasing order of total sequence abundance and relativized to abundance per 10,000 sequences. All sequences are 99 % identical to Sequence 1 and have lengths of 257±1 bp. The *R*-squared value for the abundance distribution plotted against Sequence 1 is shown in the last column

Sequence ID	Sample ID																		R^2
	A	B	C	D	E	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	62.1	112.4	89.5	15.2	27.7	547.4	43.5	67.9	39.6	21.5	75.0	100.5	0.5	92.4	3.2	35.7	0.6	3.8	1
2	0	74.6	109.9	39.6	82.7	7.3	20.2	102.2	174.5	21.9	59.8	0.1	0.7	15.6	0.6	0	0.3	0	0.00
3	0	6.2	3.9	17.0	19.0	31.1	37.9	111.0	38.9	0.2	9.6	0	2.4	4.8	1.7	0	0	0	0.03
4	0	45.6	20.0	13.2	41.2	1.7	4.4	8.2	28.8	14.9	5.9	0.1	0	3.0	0.1	0.1	0	0	0.00
5	0.1	15.0	19.5	0.1	6.4	15.6	3.6	15.0	0.8	0.4	2.9	2.6	0.3	51.2	0.1	10.8	0.1	0	0.10
6	1.1	10.4	4.5	0	0	9.5	2.4	6.3	0	0.1	4.6	4.2	0	75.3	0	12.5	0.1	0.3	0.02
7	0.1	2.0	1.2	0	0	36.0	1.1	3.5	0.3	0.2	13.5	0.2	0	8.9	0	0.7	0	1.0	0.86
8	3.9	4.3	4.2	0.9	1.3	30.3	2.9	2.5	3.1	0.7	4.6	3.3	0	2.3	0.2	0.9	0	0.1	0.98
9	2.5	1.5	2.6	0	0	5.7	0.5	5.5	0	0.1	0.5	8.7	0	29.8	0	1.8	0	0	0.05
10	0	26.4	0.8	0	0	24.2	0	0	0	0	0	0	0	0.5	0	0	0	0	0.53
11	0	5.6	8.9	0.1	0.8	0	1.1	21.7	4.8	0.3	0.1	0	0	0.1	0	0	0	0	0
12	0	11.3	1.3	0	0	1.7	6.0	2.6	0.1	0	1	0.1	0	2.9	0	0	0	0	0.03
13	2.4	1.0	2.0	0	0	0.4	0.1	1.4	0	0	0.1	1.4	0	3.1	0.1	3.4	0	0.6	0
14	0	5.3	0	0	0	7.2	0	0	0.1	0	0.8	0	0	0	0	0	0	0	0.71
15	0	0	0	0	0	0	0	0	0	0	0.3	0	0	6.5	0	3.5	0	0	0
16	0	2.7	0.9	0.4	2.5	0	0.3	0.5	1.4	0.4	0.3	0.1	0	0.1	0	0	0	0	0
17	0	0.1	0	0	0	0.3	0	0	0.1	0	0.4	0.1	0	1.6	0	4.1	0	1.1	0
18	0	0.8	0.1	0	0	1.0	0.1	0	0	0	3.1	0	0	0.4	0	0.1	0	0	0.11
19	0	0	0	0	0	5.6	0	0	0	0	0	0	0	0	0	0	0	0	0.92
20	0.3	0.6	0.3	0	0	0.7	0.3	0	0	0	0.5	0.2	0	1.7	0	0.4	0	0	0.15

Table 2 The amount of DNA for each sample library. The number of sequences that passed both 454 and in-house QC filters is shown in the third column, including contaminant sequences that were removed before the subsequent analysis

Sample	DNA (ng)	Sequences
pPT343-1	9	15,544
pPT343-2	3	4,747
pPT343-3	3	5,136
pPT343-4	1	966
pPT343-5	1	1,633
pPT343-6	9	11,499

following the Qiagen gel purification protocol. The purified product was used as the template for a PCR, in which 454 A- and B-adaptors were attached to the plasmid insert using fusion primers. The primers were designed such that the PCR product could subsequently be ligated into pUC19 previously double-digested with the restriction enzymes *Eco*RI and *Sma*I (Eurofins MWG Operon, Huntsville, AL, USA) (SOM).

Following gel purification on a 0.9 % agarose gel, the PCR product was incubated with Klenow fragment, dGTP, dCTP, and BSA at 25 °C for 30 min, followed by heat inactivation at 75 °C for 20 min (NEB, Ipswich, MA, USA). This reaction used the 3′→5′ exonuclease activity of the Klenow fragment to create an overhang suitable for ligation with *Eco*RI, while the presence of dGTP and dCTP prevented indiscriminate chewback. The DNA was ethanol precipitated and resuspended in T(8) buffer before ligation into the double-digested pUC19 using a Quick Ligation Kit according to the manufacturer's specifications (NEB, Ipswich, MA, USA). Newly ligated plasmids were immediately transformed into chemically competent *E. coli* DH10B cells. After recovery in SOC media at 37 °C for 1 h, cells were spread onto LB-Amp plates containing X-gal. White colonies that appeared after incubation at 37 °C were picked, resuspended in 3 ml 2XYT media, and grown up overnight. Plasmid DNA was isolated from the overnight cultures using a Qiagen miniprep kit according to the manufacturer's instructions.

Table 3 PCR cycle experiment. Each plasmid sample was divided into subsamples that underwent different numbers of PCR cycles. Some cycle experiments were repeated, and in these cases, the average number of miscalls per base from all repetitions was used in the analysis

Plasmid	PCR cycles	Replications
pPT42C	0	1
pPT42C	18	1
pPT42C	24	1
pPT42C	30	1
pPT42C	36	1
pPT143	0	1
pPT143	16	2
pPT143	18	1
pPT143	20	2
pPT143	24	3
pPT143	30	3
pPT143	36	1

A standard double-digest with *Eco*RI and *Hin*dIII was performed on pPT42C to isolate the modified insert. However, because the pPT143 sequence contained an *Eco*RI restriction site (GAATTC), a partial restriction digest was used to isolate the insert. Plasmid DNA was cut first with *Hin*dIII, ethanol precipitated, and subsequently digested with *Eco*RI for 45 min at 37 °C at a ratio of 1 μg DNA to 1 U *Eco*RI enzyme (NEB, Ipswich, MA, USA). The digested products from both plasmids were ethanol precipitated and gel purified, and the bands corresponding to the insert size were excised and purified. The final DNA product went directly into 454 sequencing preparation, beginning with emulsion PCR, without further modification.

Data Analysis

A BLAST database was used to tally the number of indels and substitutions per 100,000 bases in each data set (SOM). Substitution values were plotted against PCR cycle number and compared with the expected values based on the reported error rate of Platinum *Taq* HiFi (Invitrogen, Carlsbad, CA, USA; personal correspondence) (Fig. 5).

Results and Discussion

Testing OTU Clustering Effects on Sample Resolution

Highly similar 16S rRNA gene sequences may represent closely related strains or variant operons within the same genome. Similar abundance distributions (high correlation) can be derived from 16S rRNA gene sequence variants within the same genome, related strains with similar ecological distributions, or sequencing error. In contrast, highly divergent distributions (as signified by low correlation) can only be derived from related strains with distinct ecological distributions.

To test if clustering can merge distinct organisms, we investigated abundance distributions of highly similar (>99 % identity) sequences across 18 marine sediment samples (Table 1). Notably, 14 of the 19 sequences have correlation coefficient values less than or equal to 0.15 when plotted against a reference sequence (Sequence 1). This supports the hypothesis that many sequences within an OTU represent distinct organisms. This is particularly clear when abundance distribution patterns are plotted graphically (Fig. 2). Sequences that would normally be clustered into one OTU show highly variable abundance levels among samples (Fig. 2a, c). A Spearman's Rank Correlation test confirmed the absence of any correlation between these two abundance profiles (Spearman's rho=0.229, P=0.362), while the "parent" sequence and its presumed artifact (Fig. 2a, b) had a highly significant linear correlation (Spearman's rho=0.995, P<0.05). The variability between nearly identical sequences highlights the danger of clustering similar sequences with potentially important biological differences.

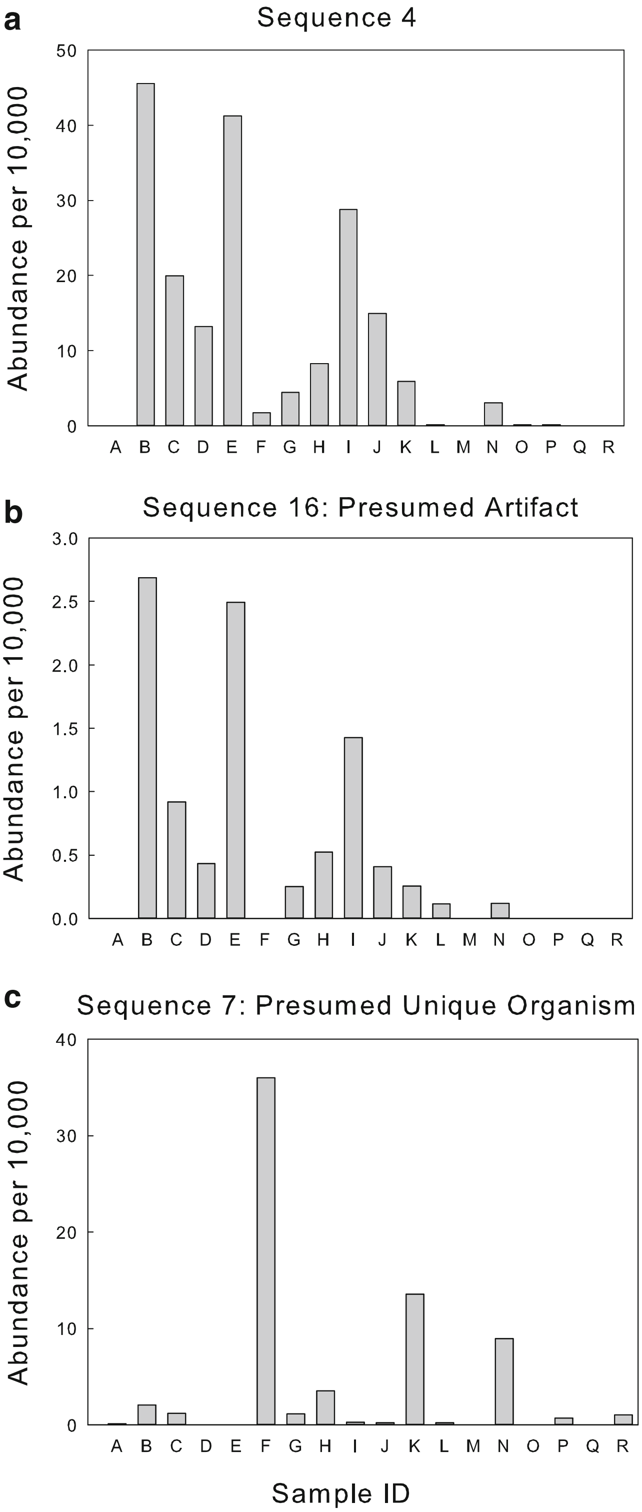

Fig. 2 Abundance distributions of three unique 16S rRNA gene sequences with >99 % sequence identity across 18 environmental samples. Each *bar* represents the abundance in a distinct sample, relativized to number of sequences per 10,000

Table 4 A breakdown of the noncontaminant sequences that passed the 454 filters and the subsets that either passed the in-house QCs (described in "Custom Quality Filtering") or were rejected. "Artifacts" refers to all sequences that differed from the template as a result of sequencing errors, including miscalls, insertions, and deletions. "Template" refers to sequences with 100 % identity to the 16S rRNA gene sequence used in the experiment. The total number of reads in each data set is the combination of artifact sequences and template sequences

	Artifacts	Template	Total reads
Passed 454 filters	14,991	96,015	111,006
Rejected by QCs	10,996	62,531	73,527
Passed QCs	3,995	33,484	37,479

Environmental distribution patterns of similar strains can provide empirical evidence for biochemical differences that would be overlooked if they were clustered together based on sequence identity. Such pairwise comparisons provide a depth of analysis unattainable by sequence comparison alone. Even at the strictest OTU level of 1 % divergence, diversity can be artificially reduced as distinct sequences are grouped together. The standard clustering level of 3 % divergence has the potential to massively amplify this problem. Avoiding OTU clustering, therefore, provides a depth of resolution that would be hidden had these sequences been grouped into one taxonomic unit. However, it also increases the likelihood that sequence artifacts will be classified as distinct organisms. Strict quality filters were, therefore, designed to reduce this probability and their efficacy was tested on a known sequence template.

Table 5 Total errors in the sequence data before and after quality control filters were applied. Note that miscalls were not filtered out to the same extent as insertions or deletions

	Pre-QC filters	Post-QC filters
Miscalls	3,806	3,187
Insertions	21,199	172
Deletions	9,376	300
Total errors	33,870	4,289

454 Error Rate Analysis

To assess how stringent quality filters affect 16S rRNA gene sequence data, a known template was sequenced using 454 GS-FLX Titanium technology. A total of 133,396 reads without Ns that passed the instrument's filters were analyzed from four pyrosequencing runs. Contaminant sequences were removed from the data set and excluded from all subsequent analyses. Of the noncontaminant reads, 70 % did not pass postinstrument, custom-designed QC filters (Table 4). The 73,527 rejected sequences included 62,531 correct template sequences as well as 6,623 artifacts containing homopolymer errors. The remaining artifacts contained other errors, including miscalls, insertions, and deletions. The level of rejection ultimately compromised the number of correct template sequences in the post-QC data set, suggesting that the filtering process may be too strict in its exclusion of valid data. However, out of the reads that passed the quality filters, 85 % were identical to the parent sequence, while 15 % of the reads contained errors and consisted of 850 unique artifact sequences seen at very low levels. Furthermore, all artifacts were at least 98 % identical to the template, corresponding to a single OTU.

The vast majority of artifacts occurred once or twice, a level of abundance that would be ignored in most environmental data sets. The most abundant artifact was seen only 63 times, at a ratio of less than 1:500 relative to the parent sequence. The tendencies for environmental sequences to produce low 454 Q

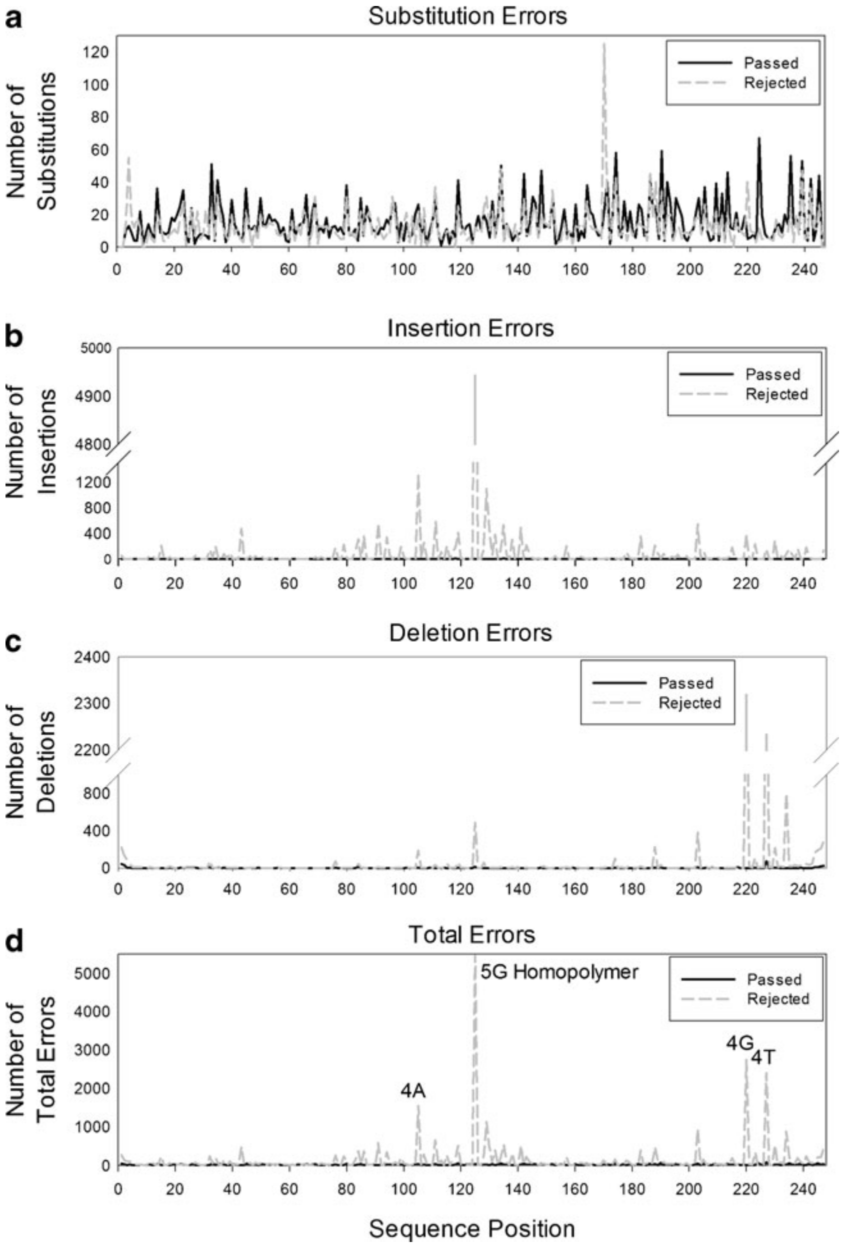

Fig. 3 Distribution of errors along the sequence. Each *graph* presents a different error type and includes both data sets: the sequences that passed the QC filters (*solid line*) and those that did not (*dashed line*). The biggest spike in total errors occurs at a 5 G homopolymer and consists mostly of insertions

scores are highly variable due to sequence composition, including number and length of homopolymers [7], so this ratio cannot be broadly applied to all samples. Nevertheless, it is a useful basis for comparison with other sequences of similar nucleotide composition. Moreover, the number of individual errors decreased by an order of magnitude, from a total of 33,870 in the rejected sequence set to 4,289 in the sequences that passed the QCs (Table 5). These results show that the in-house QC filtering process significantly improved the quality of the sequence data and prevented an overwhelming majority of artifacts.

While the quality filters substantially reduced the proportion of reads containing errors, the effect varied by error type. The drastic reduction of insertion and deletion errors, or indels, was not matched in the miscall numbers (Table 5). Most errors in the rejected data set were indels occurring at homopolymers within the sequence. The single most common error was an extension of a five-base homopolymer (Fig. 3). This single-base overcall accounted for 23 % of all insertions and 13 % of all errors in this set. Post-QCs, this insertion was seen only four times out of 172 total insertion errors in the data set. In contrast, miscalls did not decrease to the same extent as the indels (Fig. 3a, Table 5). As a result, errors in this category made up a much higher proportion of the errors after the filters were applied. Miscalls in the rejected sequences, with many more reads and nearly tenfold the number of total errors, comprised fewer than 10 % of total errors. By contrast, in the sequences that passed the QCs, miscalls made up almost 90 % of all errors (Fig. 4).

Miscalls are much less common than insertions or deletions, a natural result of the sequencing biochemistry [34]. Their predominance in the post-QC data set suggests that miscalls do not produce low 454 Q scores and are more likely to pass through filters that rely on those scores. One possible explanation for the high Q scores is that substitution errors exist in the amplicon library before sequencing has begun.

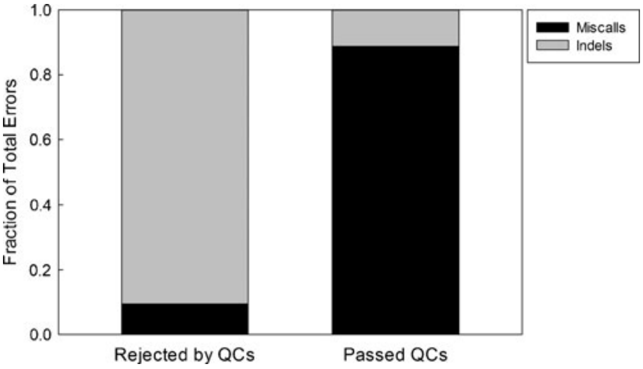

Fig. 4 The percentage of indels and miscalls in the sequences that did not pass the QC filters (*Rejected by QCs*) and those that did (*Passed QCs*). Indels compose over 90 % of the errors in the rejected data set, with miscalls making up the remainder. The inverse is true for the sequences that passed the QCs, with 88 % of all errors composed of miscalls

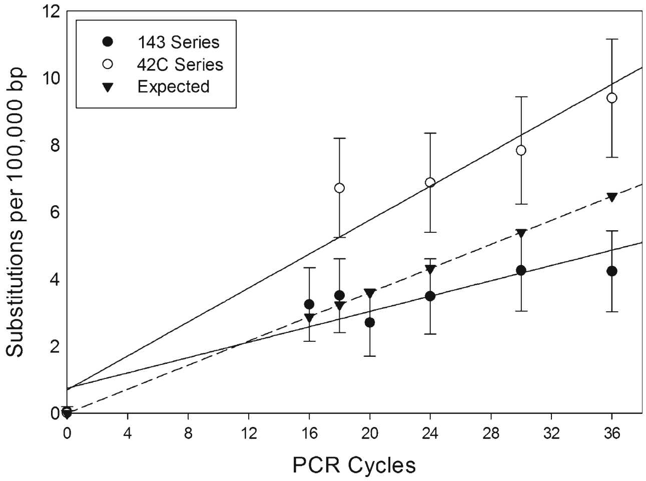

Fig. 5 The number of substitutions per 100,000 bases in post-QC sequences as a function of increasing PCR cycles. Experimental data series are labeled by their respective plasmid names (*42C* and *143*) and represented by *solid lines*, while the reported error rate for Platinum *Taq* HiFi polymerase enzyme (*Expected*) is represented by a *dashed line*. *Error bars* denote the standard deviation of substitutions seen in 1,000 iterations of subsampling 1,000 sequences in the data set

PCR Error Rate Analysis

The relatively high Q scores of miscall errors that precluded their removal with QC filters suggested that they had originated during PCR amplification of genomic DNA through *Taq* polymerase errors and were, in fact, substitution errors, not miscalls. If this notion was correct, then the substitution error rate would be dependent on the number of cycles used during PCR amplification. To test the effect of PCR cycles on substitution errors, two samples of a known template sequence were prepared for sequencing in parallel. Each template sample was subdivided into subsamples, and each subsample underwent a different number of PCR cycles during the library preparation. Substitution errors, measured per 100,000 base pairs per cycle, increased linearly with an increase in PCR cycles (Fig. 5). Experimental rates of increase approached the reported error rate for the DNA polymerase enzyme mixture used in the experiments (Table 6). The two sequences demonstrated

Table 6 Substitution error rates for the DNA polymerase used in the experiment, as reported by Invitrogen (Carlsbad, CA, USA) and experimental rates for each sequence sample. Experimental *R*-squared values are shown in the third column

	Substitutions/bp/cycle	R^2
Taq	1.8E-06	n/a
42C	2.5E-06	0.94
143	1.2E-06	0.84

slightly different slopes, suggesting that the precise rate of substitution is potentially dependent on nucleotide composition, subtle differences in PCR conditions, or other factors independent of 454 sequencing protocol. Nevertheless, the extremely high correlation between the reported *Taq* error rate and the experimental error values supports the hypothesis that substitution errors seen in 454 sequencing data are highly dependent on PCR and only minimally affected by the sequencing process. Experimental rates did have an initial value slightly above zero, potentially due to additional miscalls produced by the 454 sequencing process. This idea is supported by the fact that some sequences containing miscall errors were filtered out by QC filters dependent on 454 Q scores (Table 5). These data indicate that most of the miscall errors in 454 sequencing data can be attributed to the PCR performed during the amplicon library preparation, and lower numbers of PCR cycles can significantly improve data quality. Filtering out artifacts containing substitution errors will be more challenging than for those with indels resulting from 454 pyrosequencing and may not be possible with filters that rely exclusively on 454 Q scores. Environmental data and metadata should be used whenever possible to distinguish closely related strains from artifacts.

Next-generation sequencing technology continues to increase in output and decrease in cost. The increasingly popular Illumina platform can produce several million reads at 1/100th the cost of 454 technology. Although it is limited by read length, this, too, can be improved with paired-end protocols [35, 36]. The low cost makes it possible to resequence environmental samples with relative ease, providing another potential way to confirm the presence of artifacts, although rare sequences may not appear in all data sets. Furthermore, metagenomic shotgun sequencing can avoid the amplification biases and errors associated with PCR altogether. Nevertheless, microbial community surveys will likely continue to rely on PCR in the near future, and methodologies for improving data quality and avoiding artifacts will continue to play an important role in the field of microbial ecology.

Conclusions

While the vast majority of single-base differences are hidden when sequences are grouped into OTUs, resolution between different sequences is lost as a consequence of clustering. Evidence given here demonstrates the high chance of obscuring true diversity when similar sequences are consolidated. In this study, an attempt was made to strike a balance between minimizing the number of sequencing errors and maintaining the deep resolution afforded by high-throughput technology. Clearly, a critical component of producing high-quality data is in the library preparation. Minimizing the number of PCR cycles can prevent substitution errors that will pass through QC filters screening for low Q scores. Furthermore, using sequence data from across environmental samples can aid in identifying artifacts by their high correlation to abundant sequences. Thoughtful preparation of the amplicon library and strict quality filtering can help retain accurate estimates of community composition and diversity without consolidating reads based solely on identity.

Acknowledgments We thank Stewart Scherer for his bioinformatics assistance and helpful comments.

References

1. Pace N (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740
2. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA, Progra NCS (2009) Topographical and temporal diversity of the human skin microbiome. Science 324(5931):1190–1192
3. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103(32):12115–12120
4. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, de Winter A, Drake J, Du L, Fierro JM, Forte R, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Hutchison SK, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lee WL, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Reifler M, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Willoughby DA, Yu PG, Begley RF, Rothberg JM (2006) Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005). Nature 441(7089):120–120
5. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1(4):283–290
6. Galand PE, Casamayor EO, Kirchman DL, Potvin M, Lovejoy C (2009) Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. ISME J 3(7):860–869
7. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8(7):R143. doi:10.1186/gb-2007-8-7-r143
8. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12(1):118–123
9. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6(9):639–U627
10. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12 (38). doi:10.1186/1471-2105-12-38
11. von Wintzingerode F, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21(3):213–229

12. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methe B, DeSantis TZ, Petrosino JF, Knight R, Birren BW, Consortium HM (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21 (3):494–504
13. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27(16):2194–2200
14. Engelbrektson A, Kunin V, Wrighton K, Zvenigorodsky N, Chen F, Ochman H, Hugenholtz P (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J 4:642–647
15. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L, Li HZ, Bushman FD (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. Bmc Microbiology 10(206). doi:10.1186/1471-2180-10-206
16. Cline J, Braman J, Hogrefe H (1996) PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 24(18):3546–3551
17. Eckert KA, Kunkel TA (1991) DNA polymerase fidelity and the polymerase chain reaction. Genome Res 1(1):17–24
18. Huse SH, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12(7):1889–1898
19. Lundin D, Severin I, Logue JB, Ostman O, Andersson AF, Lindstrom ES (2012) Which sequencing depth is sufficient to describe patterns in bacterial alpha- and beta-diversity? Environ Microbiol Rep 4 (3):367–372
20. Schloss PD (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. Plos Comput Biol 6(7):e1000844. doi:10.1371/journal.pcbi.1000844
21. Sipos M, Jeraldo P, Chia N, Qu AI, Dhillon AS, Konkel ME, Nelson KE, White BA, Goldenfeld N (2010) Robust computational analysis of rRNA hypervariable tag datasets. Plos One 5(12): e15220. doi:10.1371/journal.pone.0015220
22. Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS (2009) Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. Appl Environ Microbiol 75(16):5227–5236
23. Reeder J, Knight R (2009) The ‘rare biosphere’: a reality check. Nat Methods 6(9):636–637
24. Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44 (4):846–849
25. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Tumbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336
26. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405 (6784):299–304
27. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Postal G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin JY, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7 (vol 409, pg 529, 2001). Nature 410 (6825):240–240
28. Dembicki Jr. H, Samuels BM (2007) Identification, characterization, and groundtruthing of deepwater thermogenic hydrocarbon macroseepage utilizing high-resolution AUV geophysical data. In: Conference OT (ed) Offshore technology conference, Houston, TX, USA.
29. Dembicki Jr. H, Samuels BM (2008) Improving the detection and analysis of seafloor macro-seeps: an example from the Marco Polo Field, Gulf of Mexico, USA. In: International petroleum technology conference, Kuala Lumpur, Malaysia.
30. Ashby MN, Rine J, Mongodin EF, Nelson KE, Dimster-Denk D (2007) Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. Appl Environ Microbiol 73(14):4532–4542
31. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428 (6978):37–43
32. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27(2):221–224. doi:10.1093/molbev/msp259
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
34. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18(5):763–770
35. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl Environ Microbiol 77(11):3846–3852
36. Degnan PH, Ochman H (2012) Illumina-based analysis of microbial community diversity. ISME J 6(1):183–194