Title
High-resolution phylogenetic microbial community profiling

Journal
The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 10(8)

ISSN
1751-7362

Authors
Singer, Esther
Bushnell, Brian
Coleman-Derr, Devin
et al.

Publication Date
2016-08-01

DOI
10.1038/ismej.2015.249

Peer reviewed

The ISME Journal (2016) 10, 2020–2032

www.nature.com/ismej

ORIGINAL ARTICLE

High-resolution phylogenetic microbial community profiling

Esther Singer[1], Brian Bushnell[1], Devin Coleman-Derr[1,2], Brett Bowman[3], Robert M Bowers[1], Asaf Levy[1], Esther A Gies[4], Jan-Fang Cheng[1], Alex Copeland[1], Hans-Peter Klenk[5], Steven J Hallam[4], Philip Hugenholtz[6], Susannah G Tringe[1] and Tanja Woyke[1]

[1]*US Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA;* [2]*USDA-ARS, Albany, CA, USA;* [3]*Pacific Biosciences, Menlo Park, CA, USA;* [4]*University of British Columbia, Vancouver, BC, Canada;* [5]*Newcastle University, School of Biology, Newcastle upon Tyne, UK and* [6]*Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences and Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia*

Over the past decade, high-throughput short-read 16S rRNA gene amplicon sequencing has eclipsed clone-dependent long-read Sanger sequencing for microbial community profiling. The transition to new technologies has provided more quantitative information at the expense of taxonomic resolution with implications for inferring metabolic traits in various ecosystems. We applied single-molecule real-time sequencing for microbial community profiling, generating full-length 16S rRNA gene sequences at high throughput, which we propose to name PhyloTags. We benchmarked and validated this approach using a defined microbial community. When further applied to samples from the water column of meromictic Sakinaw Lake, we show that while community structures at the phylum level are comparable between PhyloTags and Illumina V4 16S rRNA gene sequences (iTags), variance increases with community complexity at greater water depths. PhyloTags moreover allowed less ambiguous classification. Last, a platform-independent comparison of PhyloTags and *in silico* generated partial 16S rRNA gene sequences demonstrated significant differences in community structure and phylogenetic resolution across multiple taxonomic levels, including a severe underestimation in the abundance of specific microbial genera involved in nitrogen and methane cycling across the Lake's water column. Thus, PhyloTags provide a reliable adjunct or alternative to cost-effective iTags, enabling more accurate phylogenetic resolution of microbial communities and predictions on their metabolic potential.
The ISME Journal (2016) **10,** 2020–2032; doi:10.1038/ismej.2015.249; published online 9 February 2016

Introduction

Enabled by the advent of polymerase chain reaction (PCR) in 1983, the small subunit (SSU or 16S) ribosomal RNA gene has become the most widely used marker for performing phylogenetic analyses allowing the classification of novel bacterial and archaeal taxa. In addition to providing taxonomic information, cultivation-independent 16S rRNA gene profiling has transformed the study of microbial ecology and human health, enabling quantitative insights into microbial community diversity in natural and engineered ecosystems including our own bodies (e.g. Giovannoni *et al.*, 1990; Muyzer *et al.*, 1993; Janssen, 2006; Turnbaugh *et al.*, 2007; Bolhuis and Stal, 2011; Kembel *et al.*, 2012; Yatsunenko *et al.*, 2012). Expanding exponentially over the past three decades, the public 16S rRNA gene databases have however been faced with the challenge of accurately placing sequences into a given reference tree. This challenge is particularly prominent for environmental 16S rRNA gene sequences, which are marked by high numbers of novel taxa without cultivated representatives. Massive individual and institutional efforts have been made to standardize classification of environmental 16S rRNA sequences through dedicated database development and custom analysis tools (Giovannoni *et al.*, 1990; Muyzer et al., 1993; Desantis *et al.*, 2006; Janssen, 2006; Turnbaugh *et al.*, 2007; Bolhuis and Stal, 2011; Pagani *et al.*, 2011; Pruitt *et al.*, 2011; Kembel *et al.*, 2012; Pruesse *et al.*, 2012; Quast *et al.*, 2012; Yatsunenko *et al.*, 2012; Fish, 2013). Despite these improvements, reference sequences with low read accuracy, chimeric sequences and partial rRNA gene sequences with reduced phylogenetic resolution generated on short-read sequencing platforms such as 454 and Illumina remain problematic,

Correspondence: T Woyke or SG Tringe, US Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.
E-mail: twoyke@lbl.gov (TW) or sgtringe@lbl.gov (SGT)
Received 15 July 2015; revised 24 November 2015; accepted 30 November 2015; published online 9 February 2016

resulting in incorrect or less accurate classification of environmental sequences. Although read lengths on these platforms continue to improve, only full-length (FL) or near FL 16S rRNA sequences have been proven adequate for tree construction necessary for precise phylogenetic placement (Kim *et al*., 2011; Yarza *et al*., 2014). This reality poses a serious analytic challenge given that the majority of contemporary 16S rRNA sequence information emanates from short-read sequencing platforms (Tringe and Hugenholtz, 2008; Yarza *et al*., 2014).

Environmental 16S rRNA gene profiles were first performed using Sanger sequencing, which could provide accurate, near FL sequences. However, this process remains costly and at low throughput, involving the cloning of PCR products before paired-end sequencing. Thus, Sanger-based profiles generally involved relatively few samples with sequence information for less than tens to hundreds of clones per sample (Giovannoni *et al*., 1990). Today, microbial community profiles generated on the Sanger platform are scarce and unlikely to capture complete community diversity as estimated from richness analyses, rendering them adjunctive to short-read sequencing data sets (Youssef *et al*., 2009). The first commercially available next-generation sequencer, the Roche/454 FLX pyrosequencer, offered high-throughput technology at roughly 1/10th the cost of Sanger sequencing. To adopt this technology for microbial community profiling, Sogin *et al*. (2006) PCR-amplified the V6 variable region of the bacterial 16S rRNA gene and generated ~118 000 '16S pyrotags' averaging 100 bp read length in a single run, orders of magnitude more sequences than any previous Sanger study (Sogin *et al*., 2006). The use of barcodes enabled multiplexing of different samples within a single run further increasing the statistical power of the 454 platform (Parameswaran *et al*., 2007; Hamady *et al*., 2008). Lazarevic *et al*. (2009) ported this sequencing paradigm to the Illumina platform (Illumina, Inc., San Diego, CA, USA) by amplification and sequencing of the V5 loop region, providing even greater depth of coverage and a reduced price point. Currently, the most common approach for microbial community profiling uses V4, V3–V4 or V4–V5 primers on Illumina platforms to generate the so-called Illumina V4 16S rRNA gene sequences (iTags) averaging ~250–430 bp read length (Caporaso *et al*., 2012; Takahashi *et al*., 2014; Parada *et al*., 2015). Indeed, most 16S rRNA gene sequences in GenBank were generated on Illumina platforms because of their economy of scale (>10 million reads in a single MiSeq run) and high base-calling accuracy (Lazarevic *et al*., 2009; Claesson *et al*., 2010). Despite the ease and quantitative power of short-read amplicon sequencing, the representation of microbial community diversity at different taxonomic levels based on partial 16S rRNA gene sequences has been received with skepticism, as the specific combination of primer choice, read length, environmental source, reference database and assignment method influence both taxon abundance estimates and placement precision on the tree of life (Soergel *et al*., 2012; Yarza *et al*., 2014). Optimal primer selection for short-read sequencing requires comparisons with other data sets and recruitment to FL 16S rRNA gene sequences to assign accurate taxonomy to incomplete sequences (Liu *et al*., 2007; Walters *et al*., 2011; Soergel *et al*., 2012). Pacific Biosciences (PacBio) has recently developed a long-read sequencing technology, which for the first time in sequencing history has the capacity to cost-effectively sequence FL 16S rRNA genes at comparatively high throughput. A resurgence of FL sequences used as 'gold standards' has the potential to yet again transform microbial community studies, increasing the accuracy of taxonomic assignments for known and novel branches in the tree of life on previously unobtainable scales.

Here we directly address current limitations associated with partial 16S rRNA gene sequencing, through the application of PacBio's long-read, single-molecule real-time (SMRT) sequencing technology for high-resolution phylogenetic microbial community profiling. As PacBio sequencing performance has improved in recent years, its average read lengths now exceed 8 kb at ~87% read accuracy (Koren and Phillippy, 2015). In theory, such read lengths should provide high-quality sequences for 1.5 kb 16S rRNA gene amplicons via circular consensus sequencing, yet this method has only been used for a few environmental surveys (Babauta *et al*., 2014; Mosher *et al*., 2014). To test and validate this approach, we generated PacBio shotgun sequences as well as PacBio FL (PhyloTags) and iTags from a defined mock community of 23 cultivated bacterial strains (Supplementary Table 1). We then used this same approach to assess the microbial diversity of Sakinaw Lake on the Sunshine coast of British Columbia, Canada, a meromictic lake rich in candidate phyla.

Materials and methods

DNA extraction

The mock community was made up of 23 bacterial and 3 archaeal species as described in Supplementary Table 1. DNA from *Escherichia coli*, *Salmonella bongori*, *Salmonella enterica*, *Clostridium perfringens*, *Clostridium thermocellum* and *Streptococcus pyogenes* was purchased from the American Type-Culture Collection (ATCC, Manassas, VA, USA). DNA from *Fervidobacterium pennivorans*, *Thermobacillus composti* and *Corynebacterium glutamicum* was extracted using phenol–chloroform extraction, as described in Moore and Dennis (2002). DNA from *Desulfosporosinus acidiphilus*, *Desulfosporosinus meridiei*, *Desulfotomaculum gibsoniae*, *Echinicola vietnamensis*, *Frateuria aurantia*, *Natronococcus occultus*, *Olsenella uli* and *Terriglobus roseus* was isolated using the Jetflex Genomic DNA Purification Kit (Genomed GmbH, Loehne, Germany). DNA from *Hirschia baltica* was extracted using the Blood and Cell Culture DNA Maxi Kit (Qiagen, Valencia, CA, USA). DNA from *Meiothermus silvanus*, *Nocardiopsis dassonvillei* and *Segniliparus rotundus* was extracted using the Qiagen Genomic 500 DNA Kit (Qiagen, Hilden, Germany). DNA from *Pseudomonas stutzeri* was isolated using the Wizard Genomic DNA Purification Kit (Promega Corp., Madison, WI, USA). DNA from *Coraliomargarita akajimensis*, *Halovivax ruber* and *Spirochaeta smaragdinae* was extracted using the Masterpure Gram-Positive DNA Purification Kit (Epicentre, Madison, WI, USA). All DNA extracts were quantified using the PicoGreen assay and the Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA, USA) (Supplementary Figure 1). Each sample was quantified in quadruplicate. Samples were pooled at varying ratios to generate the mock community (Supplementary Figure 1). Environmental DNA was collected from Sakinaw Lake, British Columbia, Canada (49° 40.968′ N, 124° 00.119 ′W), at 30 m–80 m depth intervals on 6 June 2013, and at 120 m on 5 January 2010. Water was filtered onto 0.22 μM Sterivex filters (Mo Bio Laboratories Inc., Carlsbad, CA, USA), DNA was extracted as described previously (Wright *et al*., 2009) and quantified using the PicoGreen assay (Invitrogen).

Shotgun sequencing and processing of mock community DNA

Shotgun sequences of the mock community were generated using one SMRT cell on the PacBio RSII platform (Pacific Biosciences, Menlo Park, CA, USA). Quality filtering and error correction of PacBio sequences was performed using hgap self-correction by mapping all reads against each other. This resulted in 23 848 quality-filtered reads with average read length of 1472 bp used for analysis of the mock community. Reads were mapped against genomes downloaded from IMG (Markowitz, 2006) using BBMap (http://sourceforge.net/projects/bbmap/). Read counts were normalized to the chromosome size of reference genomes.

Primers, 16S rRNA gene amplification and sequencing procedures

For universal amplification of the V4 region of the 16S rRNA gene (V4 iTags), we used forward primer 515 F (5′-GTGCCAGCMGCCGCGGTAA-3′) and reverse primer 806 R (5′-GGACTACHVGGGTT CTAAT-3′) containing a variable 12 bp barcode sequence (Caporaso *et al*., 2012). Primers used for FL 16S rRNA gene amplification include primers 27 F (5′-AGRGTTYGATYMTGGCTCAG-3′) (Stackebrandt and Goodfellow, 1991) and 1492 R (5′-RGYTACCTTG TTACGACTT-3′). DNA amplicon generation of the V4 region and the FL 16S rRNA gene was performed using the KAPA SYBR FAST qPCR Kit (20 replication cycles) (Kapa Biosystems, Boston, MA, USA). Pooled amplicons were purified with the Agencourt AMPure XP purification system (Beckman Coulter, Brea, CA, USA) and analyzed with an Agilent bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, USA) to confirm appropriate amplicon size. Both iTag and PacBio sequencing were performed according to JGI's standard procedures: iTag amplicons were diluted to 10 nM, quantified by quantitative PCR and sequenced on the Illumina MiSeq platform (reagent kit v.3; Illumina Inc., Carlsbad, CA, USA). Mock Community PacBio libraries were constructed from five PCR technical replicate products using the PacBio SMRTbell Template Prep Kit (Pacific Biosciences) with a target insert size of 2 kbp. PacBio libraries of the Sakinaw Lake depth samples were constructed using the PacBio DNA Library Prep Kit 2.0 (Pacific Biosciences; 250 bp–< 3 kbp). All PacBio libraries were sequenced on the PacBio RS II platform using P4C2 chemistry. Sequence volumes obtained are listed in Supplementary Table 2. In this study, we discarded the data generated from the archaeal DNA because we used universal bacterial primers to generate 16S rRNA gene sequence amplicons.

Processing, clustering and classification of amplicon reads

iTag sequences were analyzed using the JGI iTag analysis pipeline (iTagger v.1.1) (Tremblay *et al*., 2015). Classification of clusters was achieved by alignment to the SILVA database (Ref 119, 8 December 2014). Mock community iTag sequences were grouped into 35 operational taxonomic unit (OTU) clusters with ⩾10 reads per cluster after quality screening (1 680 879 reads). OTUs originating from reagent contaminants made up 0.14% of the total community. The Sakinaw Lake sample returned 366 185 iTag sequences, which were binned into 2230 OTU clusters using a 97% cutoff.

PacBio 16S rRNA gene sequences were filtered using the JGI SMRT Portal 'reads of insert' protocol with predicted accuracy > 99%, corresponding to Q20. Filtering, chimera detection and clustering was performed using a set of MOTHUR tools (align.seqs, summary.seqs, screen.seqs, chimera.uchime using SILVA Gold as reference database, remove.seqs, filter.seqs, unique.seqs, pre.cluster, dist.seqs, cluster, align.seqs, filter.seqs, dist.seqs) (Schloss *et al*., 2009) (Figure 1). Chimeras were additionally removed by filtering reads ⩽1340 and ⩾1640 bp based on read length analysis using reformat.sh in BBMap (http://sourceforge.net/projects/bbmap/) (Supplementary Figure 3). Each step in the workflow was first optimized using a synthetic data set generated using randomreads.sh in BBMap (http://sourceforge.net/projects/bbmap/). Synthetic reads were made from copies of the 16S rRNA gene sequences from our selected 23 mock community genomes of variable read length (1.4–1.8 kbp) and variable average quality scores (Q10–Q27). Edits (insertions, deletions and/or substitutions) were assigned based on the quality scores of the reads,

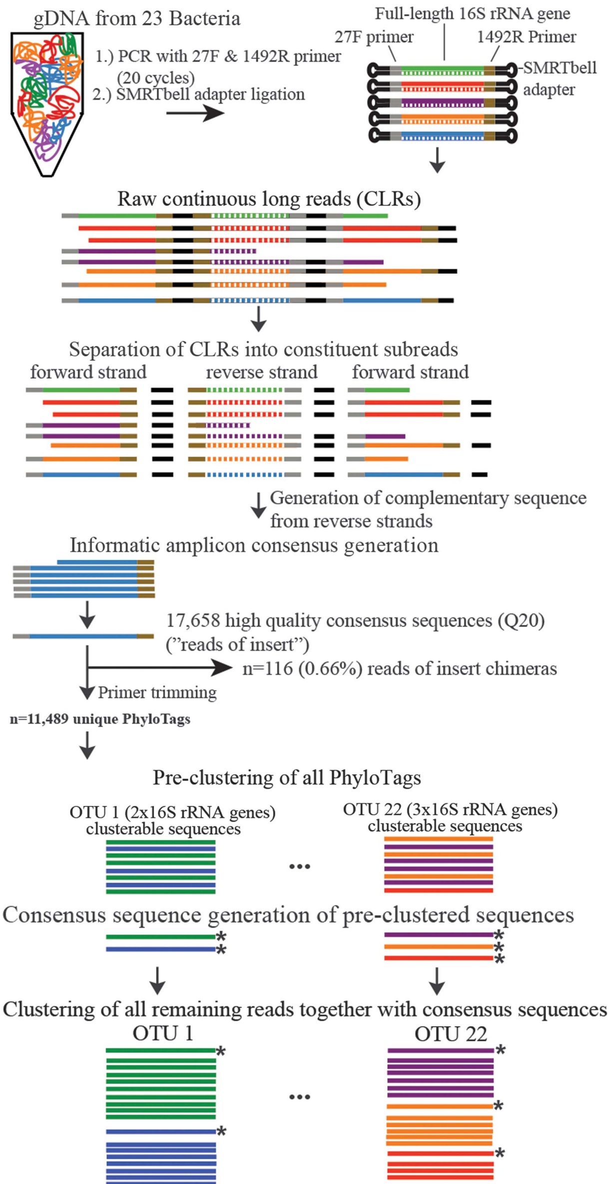


Figure 1 Workflow of the PhyloTag sequence generation and cluster analysis pipeline with results from one of the five mock community replicate data sets (more details in Supplementary Figure 1). A simulated FL 16S rRNA gene read data set generated from the 23 genomes of the selected bacterial species was used to optimize the clustering steps in the pipeline (see Materials and methods section). Detailed processing steps for iTag and shotgun sequences are illustrated in Supplementary Figure 1.

mimicking the PacBio error model. Clusters with <3 reads were discarded. Using this workflow (https://github.com/PacificBiosciences/rDnaTools), all quality-filtered reads from the simulated FL 16S rRNA gene sequences were mapped and rendered 28 OTU clusters. The same parameters were used to cluster all FL 16S rRNA gene sequences from biological sources. Sequence throughput for each sample and corresponding OTU numbers are listed in Supplementary Table 2. Mock community 16S rRNA gene abundances were normalized using copy number information from respective reference genomes.

For platform comparisons, PacBio FL and Illumina V4 16S rRNA gene sequences from Sakinaw Lake were classified according to the latest non-redundant small subunit SILVA NR Ref 119 database using the RDP-classifier (Wang *et al.*, 2007; Quast *et al.*, 2012). Taxonomic classifications were reported as unambiguous if confidence thresholds were $\geqslant 0.5$.

The differences in the mock community structures recovered from PacBio and Illumina sequencing were evaluated using Spearman's ranks correlation coefficient analysis. Coefficients were calculated for each pairwise comparison in R (Racine, 2012). Principal coordinate analyses for comparison of Sakinaw Lake depth (PhyloTags vs V4 iTags) and mock community samples (PhyloTags, PacBio shotgun and V4 iTags) were performed in R using the Bray–Curtis dissimilarity index. Data sets were subsampled by rarefaction to 6000 reads in the Sakinaw Lake sample analysis and to 2000 reads in the mock community analysis. Raw and processed sequence data is publically available on the JGI Genome portal page (http://genome.jgi.doe.gov/PhyloTag.html).

Community comparisons and phylogenetic reconstruction
Sequences were filtered and manipulated using a variety of tools available in the BBMap package (http://sourceforge.net/projects/bbmap/): for platform-independent community comparisons, V4 16S rRNA regions were retrieved by aligning V4 primer sequences (515 F, 806 R) to PhyloTag sequences (msa.sh) and selecting intervening sequences (cutprimers.sh). V4 sequences were screened for a length of 232 ± 60 bp (3 s.e.m. V4 iTag length) (reformat.sh), resulting in 195 036 sequences present in both FL and V4 sequence pools (filterbyname.sh). V4 sequences were mapped against PhyloTags using bbmap.sh (flag 'ambiguous = all'). Ambiguous matches were defined by a mapping quality of <4 (indicating a $\leqslant 50\%$ chance of correct assignment). Pairwise sequence alignments of V4 and FL sequences and subsequent data formatting were conducted using BBMap (idmatrix.sh, matrixtocolumns.sh). Sequence pairs above various % identities reported in the Table inset in Supplementary Figure 7 entail those that were exclusively present in the FL or V4 sequences, respectively. For community comparison at various taxonomic levels in Table 2, both FL and V4 sequences were clustered at 90%, 93%, 95%, 97% and 98% identity thresholds by aligning against the non-redundant small subunit SILVA NR Ref 119 database using the pick_open_reference.py workflow in QIIME (v.1.9.0). Statistical significance of differences between community structures according to clustered FL and V4 sequences was evaluated in QIIME (beta_significance.py).

Results

Different sequencing technologies for microbial community profiling exhibit clear platform-specific advantages and disadvantages. Primary benefits of next-generation sequencing over Sanger sequencing include the high-throughput nature and the cloning-free process, with Illumina providing the lowest cost per base (Table 1). Sanger and PacBio both allow (near) FL 16S rRNA gene sequencing, with PacBio being orders of magnitude more cost-effective, providing increased phylogenetic resolution in community analysis. Using a mock and a lake community, our following in-depth analysis further resolves the strengths and potential weaknesses of the PacBio platform for community analysis.

Mock community analysis

Using the mock community reference genomes, we generated a simulated PacBio 16S rRNA gene sequence data set, which directed the optimization of the sequence processing pipelines for the PacBio 16S rRNA gene sequences described in this study (Figure 1, Supplementary Figure 2 and Materials and methods). Figure 1 shows the workflow implemented for the generation of PhyloTags (defined as FL 16S rRNA gene sequences generated using the SMRT technology). Consensus sequences were generated from raw continuous long reads to correct most sequencing errors resulting in 'reads of insert' with 99% accuracy, and median 99% sequence identity. PhyloTag OTUs were defined by alignment against the SILVA Gold database in a preclustering step. Preclustered PhyloTags were grouped into consensus sequences for each individual 16S rRNA gene copy within OTU clusters. These consensus sequences were then used to map remaining reads back to cognate OTUs. OTUs were defined at 97% identity.

Genomic DNA obtained from the 23 bacterial mock community members (see Materials and methods section) was pooled and FL 16S rRNA genes as well as the V4 hypervariable regions were PCR-amplified. Amplicons were sequenced using the PacBio SMRT RSII system to generate PhyloTags and on the Illumina MiSeq platform producing V4 'iTags'. To test data reproducibility, five technical PhyloTag replicates were generated (Materials and methods and Supplementary Figure 3). Owing to their negligible bias, PacBio shotgun sequences provided a baseline of the relative abundances of each mock community member (Supplementary Figure 4). They also provided higher accuracy and reproducibility as compared with DNA molarity (Supplementary Figures 1 and 4). All five mock community PhyloTag data sets yielded similar percentages of high-quality PhyloTags and were successfully grouped into 22 OTU clusters with the standard method of grouping any two sequences that shared >97% 16S rRNA gene identity (Figure 1). The two *Salmonella* spp. were 97.4% identical on the basis of their FL 16S rRNA gene sequences. The single best PhyloTag in each cluster as chosen by its quality scores was on average 99.5% identical to the reference 16S rRNA gene sequences of the mock genomes, whereas iTag consensus sequences showed 99.9% identity. Relative abundance patterns

Table 2 (**a**) Significance test of Sakinaw Lake community structure differences between PhyloTags and *in silico* generated V4 sequences at various taxonomic levels and (**b**) percentage of PhyloTags and *in silico* generated V4 sequences classified at various taxonomic levels

(a)

% Identity clusters	*Unweighted unifrac*	P-*value*
FL vs V4 16S rRNA gene sequences total community		
90	0.30	<0.001
93	0.21	<0.001
95	0.30	<0.001
97	0.16	<0.001
98	0.20	<0.001

(b)

Taxonomic level	*% FL sequences classified*	*% V4 sequences classified*
Phylum	94.6	82.9
Class	92.9	80.7
Family	88.8	71.5
Genus	85.8	62.2
Species	74.5	49.4

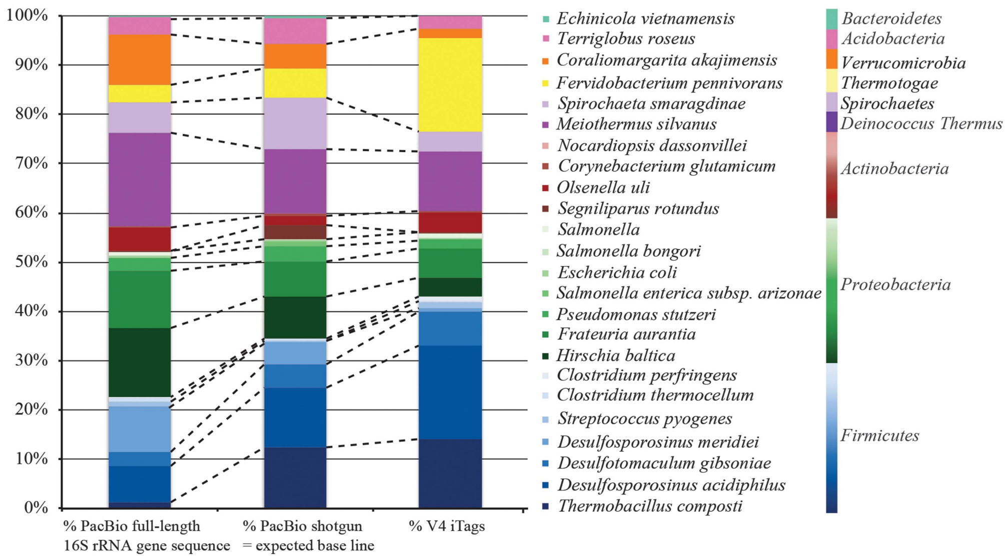
Abbreviation: FL, full length.

Table 1 Platform-dependent properties in Sanger-, Illumina- and PacBio-based community profiling

	Sanger	*Illumina (MiSeq)*	*PacBio RSII*
Cloning required	Yes	No[a]	No[a]
Average sequence time	~3 h/96-well plate	8 h	2 h/SMRT cell
Commonly used primers/amplicons	4aF/27 F, 1392 R → near full-length[a]	Various → up to 500 bp	4aF/27 F, 1492 R → full-length[a]
Amplification during sequencing	No[a]	Yes	No[a]
Average data output	~0.1 Mb per 96-well plate	8 Gb per Flowcell[a]	0.3 Gb per SMRT Cell[a]
Approximate cost per Mb	~US$2000.00[b]	US$0.11[a]	US$2.50

Abbreviations: F, forward; PacBio, Pacific Biosciences; R, reverse; SMRT, single-molecule real-time.
[a]Clear platform advantages.
[b]Cost is burdened.

a



b

	PhyloTags 2	PhyloTags 3	PhyloTags 4	PhyloTags 5	V4 iTags	PacBio shotgun
PhyloTags 1	0.98 (4.63e-06)	0.97 (4.95e-06)	0.99 (4.47e-06)	0.99 (4.52e-06)	0.80 (1.43e-05)	0.78 (3.54e-05)
PhyloTags 2	1.00 (3.95e-06)	0.96 (5.05e-06)	0.99 (4.48e-06)	1.00 (4.02e-06)	0.77 (5.69e-05)	0.76 (0.11e-05)
PhyloTags 3		1.00 (3.95e-06)	0.97 (4.98e-06)	0.97 (5.04e-06)	0.77 (6.65e-05)	0.77 (7.29e-05)
PhyloTags 4			1.00 (3.95e-06)	0.99 (4.98e-06)	0.77 (5.04e-06)	0.76 (8.37e-05)
PhyloTags 5				1.00 (3.95e-06)	0.78 (4.91e-05)	0.76 (9.90e-05)
V4 iTags					1.00 (3.95e-06)	0.89 (2.56e-06)
PacBio shotgun						1.00 (3.95e-06)

c

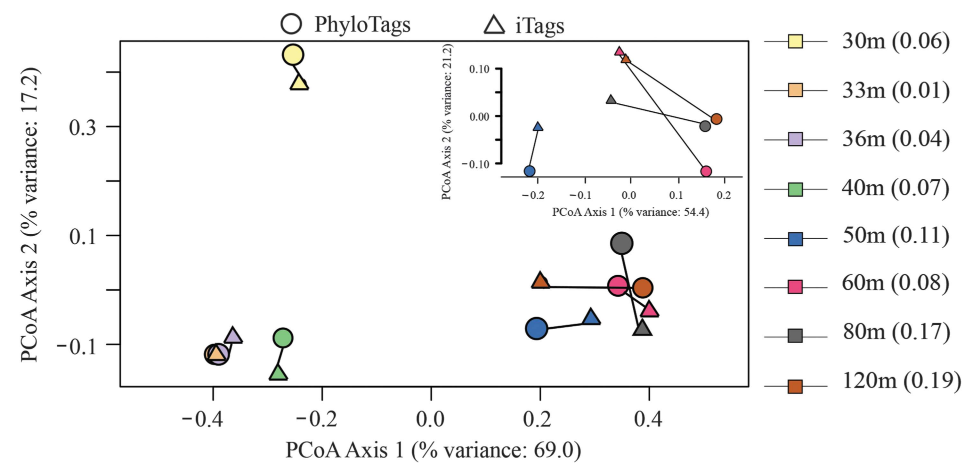


Figure 2 Analysis of the mock community profiles. (**a**) Abundance profiles of the mock community as represented by PhyloTags (pooled from all five replicates), PacBio shotgun sequences and V4 iTags. *Nocardiopsis dassonvillei*, which was added at very low relative abundance, was exclusively detected in the PacBio shotgun data set. Additional contaminant OTUs were found only in the V4 iTags (Supplementary Table 3). (**b**) Spearman's rank correlation coefficients and corresponding *P*-values were calculated to evaluate the strength of relationships between various sequence data sets. (**c**) Principal coordinate analysis (PCoA) of microbial community structures at various Sakinaw Lake depths according to PhyloTags and iTags. Mean PCoA distances between iTag and PhyloTag pairs of the same depth are stated within parentheses in the legend. The inset shows depths 50 –120 m reanalyzed for higher resolution.

on the phylum level as revealed by respective sequencing platforms are shown in Figure 2a (and Supplementary Figure 5). Shotgun sequences are expected to be the most accurate assessment of community structure because of the lack of amplification bias and were hence used as reference for the amplicon data sets. Spearman's rank correlation analysis was performed on the read abundance of the mock community strains. The five PhyloTag technical replicates for the mock community show significant congruence based on community composition and OTU clustering (Figure 2b). All data sets shared a correlation coefficient of at least 0.84 with significant *P*-values and thus do not considerably deviate from one another (Figure 2b). Comparison of species representation according to %GC showed no obvious bias across sequencing platforms (Supplementary Figure 6). The slightly higher correlation between V4 iTags and PacBio shotgun data suggests that the short tag data set is overall less PCR-/primer-biased, at least for the mock sample, providing a more accurate community profile. However, some discrepancies in the V4 iTag data set are noteworthy, for example, the relatively high abundance of *Fervidobacterium pennivorans* and the lack of *Nocardiopsis dassonvillei*. DNA from *N. dassonvillei* was added at 0.01% (±22.74%) molarity and appeared only in the PacBio shotgun data set, with a relative abundance of 0.0016%. The absence of this species from the amplicon data is likely due to specific PCR bias. Last, the V4 iTag data set contained various contaminant sequences, which comprised about 0.05% of all sequences that were not observed in PhyloTags (Supplementary Table 3).

Sakinaw Lake community analysis

To evaluate the performance of PhyloTag sequencing for environmental surveys, we applied PhyloTag and iTag sequencing to capture the microbial diversity of Sakinaw Lake. Sakinaw Lake is a meromictic lake rich in candidate phyla that partition along defined redox gradients in the water column (Gies *et al.*, 2014). Such candidate phyla are challenging to accurately classify due to a paucity of phylogenetic references in public databases. Geographically isolated meromictic lakes have consistently been shown to provide a natural enrichment in candidate phyla in the redox transition zone and monimolimnion. Indeed, Sakinaw Lake has been recognized for its extraordinarily high richness and diversity in bacterial, as well as archaeal candidate phyla (Rinke *et al.*, 2013; Gies *et al.*, 2014). By definition, candidate phyla have no cultivated representatives and their phylogenetic placement has largely relied on 16S rRNA gene sequencing data alone (Hugenholtz *et al.*, 1998). The accurate placement of novel lineages within candidate phyla is hence an essential step towards extending phylogenetic databases. We generated PhyloTag and V4 iTag libraries for Sakinaw Lake communities from eight depths spanning the water column redox gradient. As bacterial universal primers were used for the amplification of the FL 16S rRNA gene, we focused our comparative analysis exclusively on the occurrence and abundance patterns of bacterial taxa by assignment to the SILVA database.

Interestingly, 0.2–4.1% of V4 iTags were taxonomically unresolved at the phylum level, whereas all PhyloTags were classified into distinct bacterial phyla (data not shown). Overall, comparisons between PhyloTag and V4 iTags at various depths in Sakinaw Lake indicated that microbial community composition profiles between 30 and 40 m depth intervals where bacterial candidate phyla are less prevalent are in good agreement (Figure 2c and Supplementary Figure 7). At these depths, a few phyla are dominating the microbial communities and the % variance in community composition is larger between these samples than between 50 and 120 m depth intervals. Separate principal coordinates analysis of 50–120 m depth intervals where bacterial candidate phyla are much more prevalent reveal pronounced differences in community composition profiles between PhyloTags and iTags at relatively high % variance (Figure 2c, inset).

Phylogenetic resolution analysis

To evaluate discrepancies in community profiles based on amplicon length rather than sequencing technology and/or primer choice, we compared PhyloTags and *in silico* generated partial V4 16S rRNA gene sequences extracted from the PacBio FL sequences. First, a randomly subsampled set of 1818 unclustered PhyloTags spanning the Sakinaw Lake water column and their corresponding extracted V4 regions were used in an all-against-all pairwise identity comparison. In multiple instances, the same sequence pairs exhibited varying percent identities when FL and V4 sequences were compared (Supplementary Figure 8; examples are depicted by the dashed lines). The number of pairs within various percent identity thresholds provides an overview of these discrepancies, which are caused by non-homogeneous distribution of mutations across the 16S rRNA gene (Figure 3). This non-homogeneous distribution varied across different phylogenetic groups and hence leads to both over- and underestimation of community diversity. Although this comparison does not allow conclusion of the impact of clustering on the actual diversity within the microbial community, it reveals cluster patterns directly resulting from gene lengths considered.

We next compared taxonomic classifications of unclustered PhyloTags and *in silico* generated V4 16S rRNA gene sequences, according to the SILVA database. Phylogenetic assignments at various taxonomic levels were evaluated for ~195 000 unclustered PhyloTags (84.0% of total sequences) and their corresponding V4 regions retrieved from all Sakinaw Lake depth samples

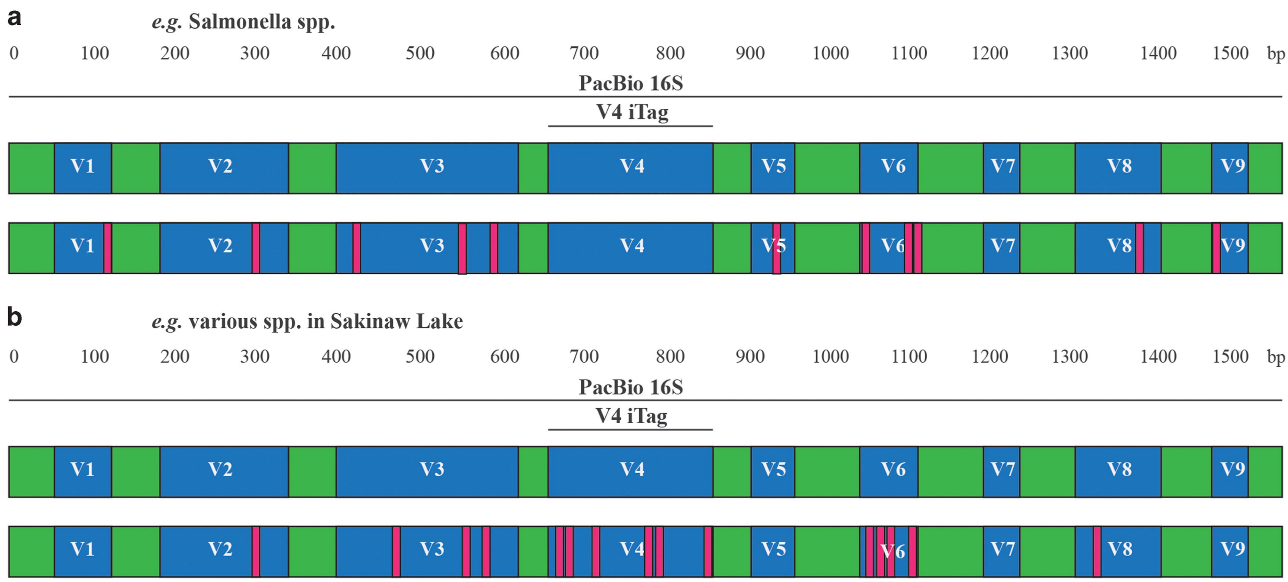


Figure 3 Conceptual representation of the 16S rRNA gene sequence with conserved (green) and hypervariable (blue) regions. Pink strips represent the abundance of mutations in respective variable regions. Target selection in amplicon sequencing determines community fingerprints. (**a**) The 16S rRNA gene variability is not homogeneous across taxonomic groups and subregions of the FL sequence. For example, *Salmonella* spp. are 97.4% identical across the FL 16S rRNA gene sequence, but are 100% identical across the V4 region. (**b**) In other instances, exclusively considering the hypervariable V4 region may lead to an overestimation of community diversity because mutations may accumulate here more than across the entire 16S rRNA gene.

combined. Generally, taxonomic classification of V4 16S rRNA gene sequences was more often either impossible or incorrect, significantly altering community profiles across all taxonomic levels (Table 2). Mapping *in silico* generated V4 sequences back to their original FL sequences (via sequence alignment using BBMap) resulted in 34 345 (17.6%) V4 sequences with ambiguous matches to FL sequences. These ambiguous matches are also linked to more frequent ambiguous classifications of the V4 than the FL sequences. Discrepancies in % classified sequences ranged from 11.7% at the phylum level to 25.1% at the species level (Table 2b). Although the relative classification differences at the sequence level do not directly translate to differences in community representation, they impact the subsequent clustering steps (Figure 1 and Supplementary Figure 2), which may result in community structure differences as seen in Figure 2c and previously discussed in (Liu *et al.*, 2007).

For instance, we compared the relative number of FL and *in silico* generated V4 sequences per phylum that were either discarded because of low confidence values (<0.5) according to RDP-classifier or exhibited phylum level classification discrepancies. The 33 283 (17.1%) V4 and 10 507 (5.4%) FL sequences were classified ambiguously at the phylum level. In all, 68.4% of the V4 sequences that could not be phylogenetically placed were classified at the phylum level according to their FL sequences (12.0% of total sequences). Interestingly, there are several phyla for which a partial 16S rRNA gene analysis resulted in a higher proportion of misclassifications and/or ambiguous match results. For instance, more than 40% of all sequences from three out of five candidate phyla were ambiguously classified according to the V4 data, that is, candidate phylum KB1 would not have been reported with confidence, whereas 66.7% of Atribacteria (OP9) and 42.4% of Parcubacteria would have been missed (Figure 4a). Other phyla with large discrepancies in classification results between FL and V4 16S rRNA gene sequences include Nitrospirae (96.7% low confidence values; 0.02% misclassified), Firmicutes (79.7% low confidence value; 0.04% misclassified), Armatimonadetes (52.8% low confidence value; 2.8% misclassified), Acidobacteria (41.0% low confidence value; 0.5% misclassified), Deferribacteres (39.6% low confidence value), Verrucomicrobia (32.6% low confidence value; 1.9% misclassified) and Fibrobacteres (24.6% low confidence value) (Figure 4a, Supplementary Figure 9 and Supplementary Table 5). In comparison, Supplementary Table 4a shows that 52.7% of the ambiguously classified FL sequences form clusters with one to two sequences and are hence likely the result of sequencing error. The remaining 47.3% were grouped into 213 sequence clusters and returned RDP-classifications with closest hits to various phyla, including Proteobacteria (17.4%), Verrucomicrobia (7.5%), Chloroflexi (7.0%), Acidobacteria (5.6%) and three candidate phyla: Parcubacteria (9.9%), candidate division KB1 (2.8%) and Saccharibacteria (1.4%) (encompassing such with confidence values <0.5) (Supplementary Table 4b). Members with high sequence similarity to our sequences are either currently missing from the SILVA database or could in fact constitute new candidate phyla. According to their corresponding V4 sequences, 608 of the un- or

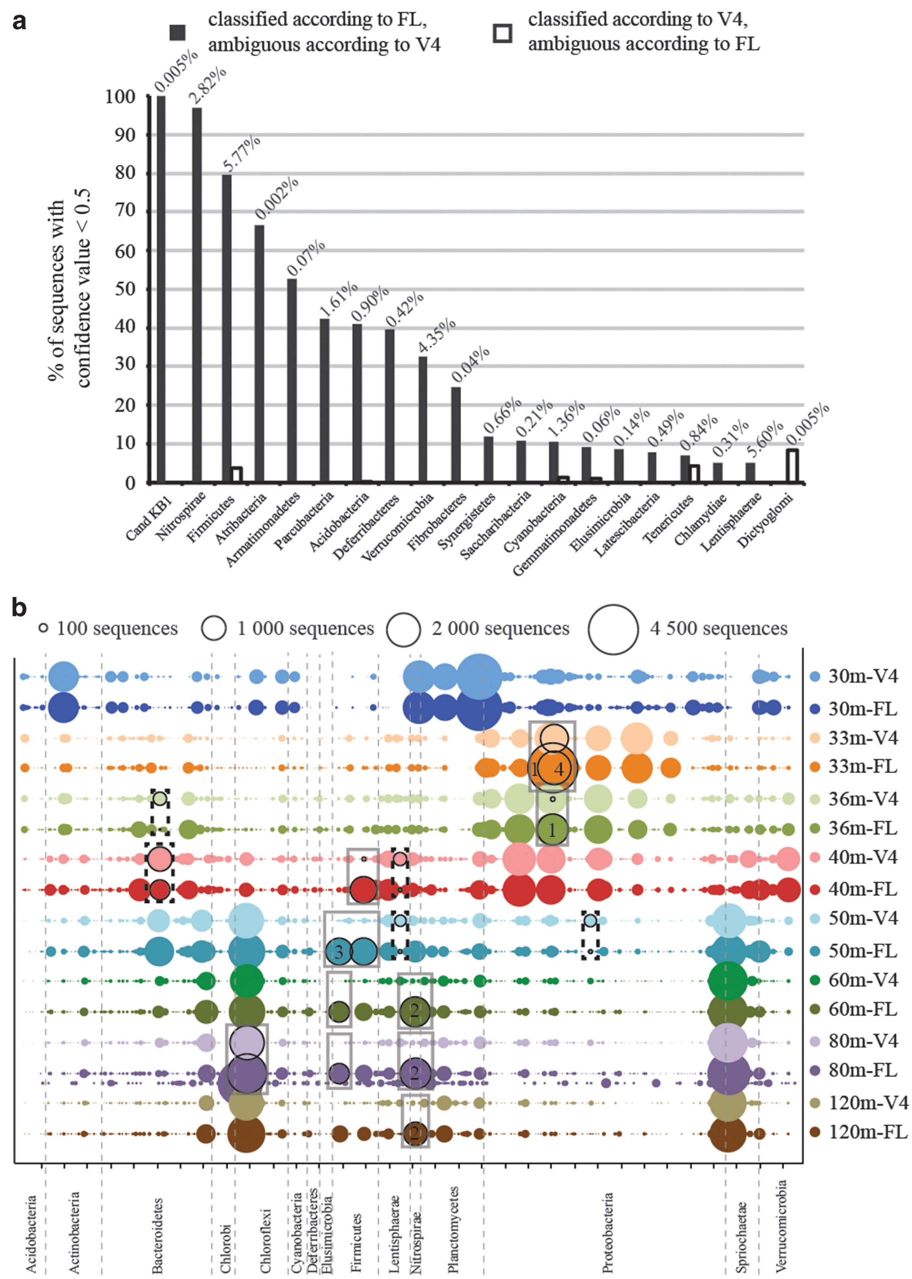


Figure 4 Community composition analysis of Sakinaw Lake depth profile at phylum and genus level represented by PhyloTag and *in silico* generated V4 sequences. (**a**) Percentage of FL PhyloTag sequences by phylum with ambiguous classifications according to their *in silico* generated V4 region and *vice versa*. Phyla with ambiguous sequences (V4: ⩾5.0% of their total; FL: ⩾1.0% of their total) are reported in this figure. Relative sequence abundance of phyla in the total community based on the number of sequences is stated above bars. (**b**) Community composition analysis of Sakinaw lake depth profile at genus level and arranged by phylum (with >1% relative abundance). Color pairs denote samples of the same depth represented by FL and V4 sequences. Bubble sizes indicate read abundance of individual genera. Several OTUs showing largest discrepancy between V4 and FL abundances are highlighted by boxes (solid gray: more FL > V4; dotted black: V4 > FL). Numbered boxes around bordered bubbles represent genera *Methylocaldum* (1), *uncultivated* genus within the *Nitrospiraceae* (2), *Bacillus* (3) and *Methylotenera* (4). Biological importance of these selected genera is discussed in the text. Examples of other genera with >1000 more FL than V4 sequences and >200 more V4 than FL sequences are depicted by bordered bubbles and boxes. Ecological significance of these genera in Sakinaw Lake was difficult to predict, for example, owing to the lack of reference genomes.

misclassified FL sequences (5.8% of ambiguously classified FL sequences) were primarily classified into Dictyoglomi (8.3% low confidence value), Tenericutes (4.4% low confidence value; 0.6% misclassified), Firmicutes (3.7% low confidence value; 1.0% misclassified) and Cyanobacteria (1.4% low confidence value; 2.0% misclassified) (Figure 4a).

Discrepancies between community profiles represented by FL and V4 sequences, respectively, are also apparent at the genus level (Figure 4b). Genera strongly under-represented in the V4 sequence data include important players in the biogeochemical cycling of methane between 33 and 45 m depth intervals (Gies *et al*., 2014). These genera encompass *Methylocaldum* (4510 FL and 27 V4 sequences at 33 m; 1745 FL and 35 V4 sequences at 36 m; 7314 FL and 274 V4 sequences total) and *Methylotenera* (2021 FL and 1409 V4 sequences at 33 m depth; 1150 FL and 803 V4 sequences at 36 m depth; 4331 FL and 3099 V4 sequences total). Methylotenera is a group of methylotrophs, which appears to be one of the dominant players in maintaining the balance of C_1 compounds at Sakinaw Lake according to relative sequence abundance (Kalyuzhnaya *et al*., 2012). Furthermore, sequence abundance comparison indicates that *Methylocaldum* together with *Methylobacter* could be the two dominant obligate methanotrophic genera in the sulfate methane transition zone, which occurs between 33 and 45 m (Gies *et al*., 2014). *Methylobacter* is a genus of methanotrophs representing a subset of unique obligately methylotrophic bacteria that use methane as their primary carbon and energy source (Bowman *et al*., 1993). *Methylocaldum* belongs to a group of type X methanotrophs, with members capable of using methane as well as methanol (Pimenov *et al*., 2010). Methane concentrations were determined to be highest between 33 and 45 m, while O_2 concentrations decrease below 33 m (Gies *et al*., 2014). This depth interval hence represents an optimal habitat for (micro-)aerophilic methane oxidizers (Gies *et al*., 2014).

In addition to methane cycling, identification of *Nitrosomonas* (149 FL sequences and 0 V4 sequences at 30 m depth; 173 FL and 0 V4 sequences total) and an uncultivated genus within the *Nitrospiraceae* (882 FL and 0 V4 sequences at 50 m depth; 1634 FL and 0 V4 sequences at 60 m depth; 1737 FL and 0 V4 sequences at 80 m depth; 1006 FL and 0 V4 sequences at 120 m depth; 5260 FL and 1 V4 sequence total) provides a potential link to the nitrogen cycle, yet V4 data largely missed identification of these nitrifier groups. Members of the genus *Nitrosomonas* oxidize ammonia into nitrite as their basis for energy metabolism and fix CO_2 to obtain carbon (Schmid *et al*., 2000). Its predominant presence at 30 m is likely due to its need for oxygen, however, avoidance of light (Theodore and Wardle, 2012), which is granted at that depth in Sakinaw Lake (Gies *et al*., 2014). *Nitrobacter* and *Nitrospira* are capable of performing the second step of nitrification (Nogueira and Melo, 2006). *Nitrospira* was found in both FL and V4 sequences at 30 m depth (67 FL and 67 V4 sequences), completing the nitrification process. Other genera exhibiting similar or larger sequence abundance discrepancies between FL and V4 data sets were affiliated with candidate phyla (Parcubacteria, Omnitrophica, Aminicenantes), Chloroflexi, Bacteroidetes, Planctomycetes and Tenericutes (Figure 4b). Genus-level representatives of these phyla in the current databases are either uncultivated and/or the metabolic potential of reference organisms have not been previously associated with important biogeochemical cycles.

Genera that were under-represented in the FL sequences according to the V4 sequences are dominated by groups of organisms without genomes in the database or any other function prediction. Examples with largest sequence abundance differences are uncultivated Bacteroidetes bacterium (21 and 718 FL sequences, 320 and 1046 V4 sequences at 36 and 40 m, respectively), uncultivated Lentisphaerae bacterium (27 and 28 FL, 316 and 234 V4 sequences at 40 and 50 m, respectively) and Smithella (27 FL and 268 V4 sequences at 50 m). Although the lack of ecological data for the genera under-represented in the PhyloTags does not allow us to make inferences about their functional properties and/or ecological roles, the significantly higher sequence discrepancy between FL and *in silico* generated V4 sequences suggests a more impactful ecological misinterpretations of the community profile, if only V4 sequences would be considered.

Discussion

We here demonstrate that PhyloTags do not need technical replication and strongly correlate with shotgun metagenome sequences. PhyloTags overall showed comparable results to traditional iTag sequences for the relatively simple mock community, as well as the more complex environmental sample with PCR and/or primer bias being likely the chief driver for differences in community profiles between platforms. A comparison between FL and *in silico* generated partial amplicon data in the environmental sample, however, showed that multiple phyla were completely missed by short-read sequences, community structure was significantly shifted at the genus level and that several dominant microbial genera across the water column of Sakinaw Lake could only be resolved via PhyloTags.

The 16S rRNA gene surveys have been radically changing our view of microbial evolution and diversity. FL 16S rRNA gene sequences are known to be more effective than partial gene sequences in inferring phylogenetic affiliation among and between microbial community members (Liu *et al*., 2007; Walters *et al*., 2011; Soergel *et al*., 2012). Hence, near FL sequences generated on the Sanger sequencing platform have remained the gold standard for a long time. However, while Sanger sequencing is associated with the trouble and expense of low-throughput cloning into host cells, PacBio has recently been offering a cost-effective, high-throughput alternative that produces long reads (2–15 kb), which can be used to generate FL 16S rRNA gene sequences.

Few 16S rRNA gene sequence studies have taken advantage of the long reads that the PacBio platform offers. Although recently Babauta *et al*. (2014) sequenced the V1–V3 region of a microbial mat community to successfully track composition changes during enrichment for microelectrode interactions, Mosher *et al.* (2014) concluded that 16S rRNA gene sequences >1400 bp allowed enhanced phylogenetic and taxonomic resolution to the species level in environmental samples compared with the 454 platform. Our study complements these efforts by evaluating pros and cons for various types of community analyses, including known simple and unknown complex communities with phyla abundantly and minimally represented in the database. It is the first benchmark study using FL 16S rRNA gene sequences generated on the PacBio platform and provides a comprehensive comparison between current iTag and emerging PhyloTag 16S rRNA sequencing paradigms, highlighting the impact of both short- and long-read sequencing platforms on microbial community profile interpretations. Our benchmarked 16S rRNA gene sequence analysis pipeline for use with SMRT sequencing technology was consistently reproducible. Although composition analysis of the mock community exhibited a marginally higher correlation between shotgun data and iTags, analysis of environmental samples indicated superior phylogenetic resolution of PhyloTags. We attribute the slightly higher correlation between iTags and shotgun sequence data to lower primer/PCR bias in the V4 primers and the resulting shorter amplicons, as compared with the FL amplification products. Moreover, the mock community was composed of few, mostly distantly related organisms, which are well represented in the 16S rRNA gene databases. Therefore, accurate taxonomic placement was not problematic for either FL or partial 16S rRNA gene sequences. The resolving power of PhyloTags in our data sets was more apparent in samples with complex microbial communities and when reference sequences in the database were scarce. Misclassifications and inability to classify sequences due to read length alone impaired interpretation of community function inferred from community diversity information at different taxonomic levels. From species to phylum, ~12–25% more FL than V4 sequences were unambiguously classified. Thus, FL sequences provide a more complete picture of community composition needed to accurately link microbial players with important biogeochemical cycles within the given ecosystem. Indeed, FL sequences enabled the identification of abundant genera known to participate in methane and nitrogen cycling in Sakinaw Lake, which were under-represented in the V4 sequences.

Since the generation of PhyloTags does not require amplification during the sequencing step, sequencing platform-specific bias is predicted to be generally reduced compared with other platforms. PhyloTag sequencing also offers the highest contig accuracy without discrimination against GC-rich or -poor regions, which further reduces bias in amplicon-based profiling (Quail *et al.*, 2012). The raw error rate in PacBio sequences is $\leqslant 15\%$ and dominated by indels, which are more difficult to correct than substitutions (B Bushnell, personal communication). For this study, shorter reads were used representing the consensus of many passes over the same molecule. These consensus reads had an error rate around 0.5% relative to the original genomic sequence. This is adequate to confidently assign OTUs at the species level using a 97% identity threshold, as two reads with 0.5% error from the same sequence will retain 99% identity. However, differentiation between strains or quantification of the 16S rRNA copy number of an organism remains difficult at this point. PhyloTag error rates can be further reduced in a number of ways: first, by selecting an inter-read consensus after cluster generation. This requires new algorithm development, as the consensus program we tested did not produce adequate results (typically yielding chimeras between different 16S rRNA copies). Second, longer movies (capturing image information of the SMRT cell) will allow more passes over a molecule, increasing the intra-read consensus quality. Third, PacBio chemistry, software and calibration improvements will directly result in more accurate sequences. Finally, structural modeling of the folded RNA may aid in differentiating between genetic variation and sequencing error, allowing better error correcting or filtering of high-error-rate reads. PacBio has been directing efforts towards improving their technology considering exactly these parameters (Supplementary Figure 10), so that approaching the quality of Sanger amplicon sequencing appears realistic over time.

Although the use of V4 iTags for microbial community profiling has multiple advantages including cost-efficiency (lowest cost per base at 0.11$/Mb), high-throughput multiplexing, the possibility of using universal primers that target archaeal and bacterial taxa simultaneously and the opportunity to get a deep insight into the rare biosphere, these do come at the expense of taxonomic resolution. Accurately extending the microbial 16S rRNA gene catalogue will be challenging if only partial 16S rRNA gene sequences are considered and evaluated as short-read sequences can potentially lead to both inflation of diversity and missing diversity, for example, with respect to new candidate ranks at various taxonomic levels. Moreover, comparison between data sets generated with different primers may lead to classification discrepancies, which limit the accuracy of microbial community profiling. This limitation can be mitigated if FL 16S rRNA gene sequencing at high throughput as alternative to Sanger sequencing becomes the new standard, or at minimum complementary to Illumina 16S rRNA gene surveys. Using PhyloTags to assess microbial community diversity in environmental samples allows us to fill important gaps in the tree of life while improving classification and microbial community profiling accuracy with important implications for inferred metabolic potential and biogeochemical roles of uncultivated microorganisms in natural and human engineered ecosystems.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank the JGI production team for assistance in sequencing and Doina Ciobanu for the generation of the mock community. This work was performed under the auspices of the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility with funding support from Contract No. DE-AC02–05CH11231, the Natural Sciences and Engineering Research Council (NSERC) of Canada, Canada Foundation of Innovation (CFI), the Tula Foundation funded Centre for Microbial Diversity and Evolution (CMDE) and the Canadian Institute for Advanced Research (CIFAR) through grants awarded to SJH. EAG was supported by a 4-year fellowship (4YF) from the University of British Columbia.

References

Babauta JT, Atci E, Ha PT, Lindemann SR, Ewing T, Call DR *et al.* (2014). Localized electron transfer rates and microelectrode-based enrichment of microbial communities within a phototrophic microbial mat. *Front Microbiol* **5**: 1–2.

Bolhuis H, Stal LJ. (2011). Analysis of bacterial and archaeal diversity in coastal microbial mats using massive parallel 16S rRNA gene tag sequencing. *ISME J* **5**: 1701–1712.

Bowman JP, Sly LI, Nichols DS, Hayward AS. (1993). Revised taxonomy of the methanotrophs: description of *Methylobacter* gen. nov., emendation of Methylococcus, validation of *Methylosinus* and Methylocystis species, and a proposal that the family Methtylococcaceae includes only the group I methanotrophs. *Int J Syst Bacteriol* **43**: 735–753.

BBMap. Available at: http://bbmap.sourceforge.net/ (accessed October 2015).

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–1624.

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP *et al.* (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* **38**: e200–e200.

Desantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM *et al.* (2013). FunGene: the functional gene pipeline and repository. *Front Microbiol* **4**: 291.

Gies EA, Konwar KM, Beatty JT, Hallam SJ. (2014). Illuminating microbial dark matter in meromictic Sakinaw Lake. *Appl Environ Microbiol* **80**: 6807–6818.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60–63.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235–237.

Hugenholtz P, Goebel BM, Pace NR. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**: 4765–4774.

Janssen PH. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72**: 1719–1728.

Kalyuzhnaya MG, Beck DAC, Vorobev A, Smalley N, Kunkel DD, Lidstrom ME *et al.* (2012). Novel methylotrophic isolates from lake sediment, description of *Methylotenera versatilis* sp. nov. and emended description of the genus *Methylotenera*. *Int J Syst Evol Microbiol* **62**: 106–111.

Kembel SW, Wu M, Eisen JA, Green JL. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* **8**: e1002743.

Kim M, Morrison M, Yu Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* **84**: 81–87.

Koren S, Phillippy AM. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* **23**: 110–120.

Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Østerås M *et al.* (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**: 266–271.

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35**: e120–e120.

Markowitz VM. (2006). The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34**: D344–D348.

Moore D, Dennis D. (2002). Purification and concentration of DNA from aqueous solutions. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Stahl K (eds), *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc: New York, NY, USA, pp 2.1.1–2.1.3.

Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J *et al.* (2014). Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *J Microbiol Methods* **104**: 59–60.

Muyzer G, de Waal EC, Uitterlinden AG. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**: 695–700.

Nogueira R, Melo LF. (2006). Competition between *Nitrospira* spp. and *Nitrobacter* spp. in nitrite-oxidizing bioreactors. *Biotechnol Bioeng* **95**: 169–175.

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B *et al.* (2011). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**: D571–D579.

Parada A, Needham DM, Fuhrman JA. (2015). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time-series and global field samples. *Environ Microbiol*; e-pub ahead of print 14 August 2015; doi:10.1111/1462-2920.13023.

Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M *et al.* (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* **35**: e130–e130.

Pimenov NV, Kallistova AY, Rusanov II, Yusupov SK, Montonen L, Jurgens G *et al.* (2010). Methane formation and oxidation in the meromictic oligotrophic Lake Gek-Gel (Azerbaijan). *Microbiology* **79**: 247–252.

Pruesse E, Peplies J, Glockner FO. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. (2011). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130–D135.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR *et al.* (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers. *BMC Genom* **13**: 1–1.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P *et al.* (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Racine JS. (2012). RStudio: A Platform-Independent IDE for R and Sweave. *J Appl Econ* **27**: 167–172.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schmid M, Twachtmann U, Klein M, Strous M, Juretschko S, Jetten M *et al.* (2000). Molecular evidence for genus level diversity of bacteria capable of catalyzing anaerobic ammonium oxidation. *Syst Appl Microbiol* **23**: 93–106.

Soergel DAW, Dey N, Knight R, Brenner SE. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J* **6**: 1440–1444.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stackebrandt E, Goodfellow M (1991). Nucleic acid techniques in bacterial systematics. In: Stackebrandt E, Goodfellow M (eds), *16S rRNA Sequencing Primers*. Wiley: New York, NY, USA, pp 132–136.

Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M. (2014). Development of a prokaryotic universal primer for simultaneous analysis of bacteria and archaea using next-generation sequencing. *PLoS One* **9**: e105592.

Theodore MG, Wardle LP. (2012). *Wastewater Chemical/Biological Treatment Method for Openwater Discharge Earth Renaissance Technologies*. Publication number US8192626 B2; application number US 12/927, 168.

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T *et al.* (2015). Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* **6**: 771.

Tringe SG, Hugenholtz P. (2008). A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**: 442–446.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. (2007). The Human Microbiome Project. *Nature* **449**: 804–810.

Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. (2011). PrimerProspector: *de novo* design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**: 1159–1161.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Wright JJ, Lee S, Zaikova E, Walsh DA, Hallam SJ. (2009). DNA extraction from 0.22 μM Sterivex filters and cesium chloride density gradient centrifugation. *J Vis Exp*; e-pub ahead of print 18 September 2009; doi:10.3791/1352.

Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer K-H *et al.* (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**: 635–645.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M *et al.* (2012). Human gut microbiome viewed across age and geography. *Nature* **486**: 222–227.

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**: 5227–5236.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)