

## ORIGINAL ARTICLE

# Illumina-based analysis of microbial community diversity

Patrick H Degnan and Howard Ochman

Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

**Microbes commonly exist in milieus of varying complexity and diversity. Although cultivation-based techniques have been unable to accurately capture the true diversity within microbial communities, these deficiencies have been overcome by applying molecular approaches that target the universally conserved 16S ribosomal RNA gene. The recent application of 454 pyrosequencing to simultaneously sequence thousands of 16S rDNA sequences (pyrotags) has revolutionized the characterization of complex microbial communities. To date, studies based on 454 pyrotags have dominated the field, but sequencing platforms that generate many more sequence reads at much lower costs have been developed. Here, we use the Illumina sequencing platform to design a strategy for 16S amplicon analysis (iTags), and assess its generality, practicality and potential complications. We fabricated and sequenced paired-end libraries of amplified hyper-variable 16S rDNA fragments from sets of samples that varied in their contents, ranging from a single bacterium to highly complex communities. We adopted an approach that allowed us to evaluate several potential sources of errors, including sequencing artifacts, amplification biases, non-corresponding paired-end reads and mistakes in taxonomic classification. By considering each source of error, we delineate ways to make biologically relevant and robust conclusions from the millions of sequencing reads that can be readily generated by this technology.**

*The ISME Journal* (2012) 6, 183–194; doi:10.1038/ismej.2011.74; published online 16 June 2011

**Subject Category:** microbial ecology and functional diversity of natural habitats

**Keywords:** iTags; pyrotags; 16S ribosomal RNA

## Introduction

Microbes in nature typically exist in complex communities, which can be highly variable in both the composition and the abundance of their constituent species. Initial surveys of microbial community diversity applied cultivation-based methods, but these were often biased and of limited value owing to the inability to cultivate many (more likely, the majority of) naturally occurring species. Such methods were supplanted by molecular approaches that did not rely on cultivation but were based rather on the PCR amplification, cloning and Sanger sequencing of universally conserved molecules, usually the 16S ribosomal RNA gene. Current sequence databases contain over a million full-length 16S rRNA sequences spanning a broad phylogenetic spectrum than can serve as a benchmark for assessing the bacterial taxa (also referred to as ‘phylotypes’ or ‘ribotypes’) present in environments worldwide (Cole *et al.*, 2009).

Although the traditional 16S-cloning-and-sequencing approach has the potential to provide an in-depth view of the richness and evenness of bacterial species within a community, its application is somewhat laborious and costly, with the result that most applications have assayed on the order of only 100 sequences per sample, which may not be sufficient to fully characterize all but the simplest communities. This has led to the development of alternative techniques for assessing rRNA variation (for example, ARISA, DGGE, tRFLPs) (Muyzer *et al.*, 1993; Liu *et al.*, 1997; Fisher and Triplett, 1999); however, the most revolutionary innovation has been the application of 454 pyrosequencing to survey hundreds of thousands of 16S rRNA gene amplicons (termed ‘pyrotags’) in a single sequencing run (Sogin *et al.*, 2006). Since its initial introduction, pyrotag analysis has been widely implemented (for example, Dethlefsen *et al.*, 2008; Bowers *et al.*, 2009; Cardenas *et al.*, 2010; McLellan *et al.*, 2010) and further optimized in order to obtain thousands of sequences from multiple samples in parallel (Parameswaran *et al.*, 2007; Hamady *et al.*, 2008; Costello *et al.*, 2009; Engelbrekton *et al.*, 2010).

Advancements in sequencing technologies, by offering greater numbers of sequencing reads at much lower costs, might eventually supplant pyrotag analysis for the characterization of microbial

Correspondence: PH Degnan, Department of Ecology and Evolutionary Biology, Yale University West Campus, PO Box 27388, West Haven, CT 06516-7388, USA.

E-mail: patrick.degnan@yale.edu

Received 14 March 2011; revised 2 May 2011; accepted 2 May 2011; published online 16 June 2011

communities. Recent interest has focused on the applicability of other sequencing methodologies, most notably Solexa/Illumina (Lazarevic *et al.*, 2009; Claesson *et al.*, 2010; Gloor *et al.*, 2010; Caporaso *et al.*, 2011; Zhou *et al.*, 2011), which is currently less than 1/100 the cost per read than 454 pyrosequencing. Although realizing much shorter read lengths, the Illumina technology can be tailored to yield sequences of increased lengths, as can be obtained by merging the paired-end reads generated from the same amplicon (Gloor *et al.*, 2010; Rodrigue *et al.*, 2010; Zhou *et al.*, 2011). By integrating sample-identifying barcodes into the amplification primers, the Illumina platform, like 454 pyrosequencing, is amenable to a high level of multiplexing, which further increases its utility for examining large and complex sets of samples (Gloor *et al.*, 2010).

Here, we use the Illumina sequencing platform for 16S rRNA amplicon analysis (called 'iTags') and assess its generality, practicality and potential complications. We first fabricate and sequence a paired-end library of hyper-variable 16S rDNA fragments amplified from samples that varied in contents from a single bacterium, an artificial community and a highly complex natural community. The inclusion of identifying barcodes and multiple primer pairs, combined with the selected amplicon lengths, allowed us to evaluate the level of sequencing error, amplification biases and association between paired-end reads, among other variables. Despite recovering over 30 million reads, our imposition of stringent quality filters resulted in up to an 85% reduction in the number of potentially informative reads. Nevertheless, tens to hundreds of thousands of informative reads were returned for each of the multiplexed samples. These results uncovered both the major advantages and obstacles in the application of these methods, and provide insights into the ways that such impediments can be ameliorated prior to the application of these procedures.

## Materials and methods

### *Designing amplification primers*

During the development of this project, the Illumina GAIIx sequencer could produce reads of up to 100 nucleotides (nt) in length from both ends of single DNA fragments or amplicons (that is, paired-end reads). We exploited this technology to assemble reads spanning a >100-nt region of the 16S rDNA. To accomplish this, we identified candidate universal primer pairs spanning the 16S rRNA variable loops (Neefs *et al.*, 1991) that would maximize sequence information and be short enough to allow adequate read overlap.

The phylogenetic distribution of each primer sequence was analyzed using the RDP Probe Match (Cole *et al.*, 2009). The goal was to obtain broadly

distributed primers that amplify regions of approximately 100–150 bp in length such that the paired-end reads from an amplicon would show some overlap. Of primer sets meeting these criteria, two flanking the variable loop-6 (V6) were chosen (XXX denotes sample-specific barcodes): 967F 5'-XXX CAACGCGAAGAACCTTACC-3' and 1046R 5'-XXX CGACAGCCATGCANCACCT-3' (Short 1 V6 (S<sub>1</sub>V6); 98 bp in *Escherichia coli*) (Sogin *et al.*, 2006), and 917F 5'-XXX GAATTGACGGGGRCCCGC-3' and 1061R 5'-XXX CACGRCACGAGCTGACGAC-3' (Long V6 (LV6); 163 bp in *E. coli*) (Keijser *et al.*, 2008). We further modified the S<sub>1</sub>V6 primer pair to increase both the extent of overlap between paired-reads and the number of potentially informative sites as follows: 970F 5'-XXX CGCGAAGAACCT TACC-3' and 1050R 5'-XXX ACGACAGCCATG CANC-3' (Short 2 V6 (S<sub>2</sub>V6); 96 bp in *E. coli*).

Sample-identifying barcode sequences were included at the 5' ends on both the forward and the reverse primers. The barcodes were designed (1) to be short, 3 or 4 bp, to minimize loss of sequence information; (2) to have two or more differences from one another (such that a single sequencing error could not convert one into another); and (3) to have the 3' nucleotide of the barcode be a low-frequency match of nucleotide upstream from the 5' end of the priming site. The barcodes for the primer pairs were as follows: S<sub>1</sub>V6–ATG, CAG, TCG, ACT, CGT, GCA, GAT, and TGA; S<sub>2</sub>V6–AGC, ATT, GTC, TCC, CCAT, CGGC, GCGT and TTAT; and LV6–GTC, AAC, CCA, TAA, AGGA, GCGC, TTGC and ATAC.

### *Selecting sequencing templates*

We tested 16S primers on sets of samples representing a range of microbial diversity. First, we used two well-defined samples containing either DNA extracted from a single sequenced strain of *E. coli* K-12 (MG1655) (Blattner *et al.*, 1997) or from a mixture of genomic DNAs of known concentrations from 19 cultivated strains of bacteria (Table 3). Next, we analyzed a sample of unknown microbial diversity: a fecal sample from a laboratory mouse. The QIAamp DNA stool mini kit (Qiagen, Valencia, CA, USA) was used following the manufacturer's protocol to extract total DNA from the feces of an individual lab-reared mouse, *Mus domesticus* strain WSB.

### *Preparing samples for 16S amplification*

PCR amplifications were performed in triplicate 30- $\mu$ l reactions using 30–50 ng of template DNA and one of two DNA polymerases. Samples amplified with the barcoded S<sub>1</sub>V6 primer pairs were performed using 1.2 U of *Taq* DNA polymerase (5 PRIME) and final concentrations of 1X polymerization buffer, 1 mM dNTPs and 0.4  $\mu$ M of each primer. The reaction mixtures were subjected to an initial denaturation cycle of 95 °C for 2 min, followed by 25 cycles at 95 °C

for 30 s, 55 °C for 45 s and 72 °C for 1 min, and a final extension step at 72 °C for 10 min. The samples amplified with the barcoded S<sub>2</sub>V6 or LV6 primer pairs used the high-fidelity Advantage 2 polymerase mix and the Advantage 2 SA PCR buffer (Clontech, Mountain View, CA, USA) following the manufacturer's instructions. Amplifications with the S<sub>2</sub>V6 and LV6 primer pairs were performed in a laminar-flow hood using screw cap tubes to minimize contamination. The thermal cycling conditions were identical to those detailed above except that the annealing temperature was 58 °C.

Triplicate reactions for each sample were pooled, visualized on 1% agarose gels, purified with the MinElute PCR purification kit (Qiagen) and eluted in 11 µl of EB buffer. The concentrations of the amplification products were quantified on a Bio-analyzer using the DNA 1000 LabChip kit (Agilent Technologies, Santa Clara, CA, USA). The barcoded products of the S<sub>1</sub>V6 and S<sub>2</sub>V6 primer sets were combined in equimolar amounts, re-purified with the MinElute kit, eluted in 30 µl of EB buffer and submitted for 2 × 75-nt paired-end sequencing at the Yale Center for Genomic Analysis. Similarly, the LV6 products were mixed in equimolar amounts, but these longer amplicons were subjected to 2 × 100-nt paired-end Illumina sequencing.

#### Processing and assembly of sequencing reads

Initial base-calling using Bustard (v 1.7) and quality filtering of paired-end reads were performed by the Yale Center for Genomic Analysis. Upon receipt, these reads were evaluated for quality (mean quality scores; presence of sites denoted as 'N'), and identifiable primers and barcode sequences. Read pairs with perfectly matching primer and barcodes were extracted and binned by barcode. The resulting Illumina *fastq* files, containing quality values equivalent to Phred scores, were converted to Sanger *fastq* files using Maq ('maq\_ill2sanger', v 0.7.1) (Li *et al.*, 2008). Individual read pairs were converted to *fasta* and *qual* files, and then assembled using Phrap (v 1.080812) using the following parameters: '-vector\_bound 0 -trim\_start 0 -forcelevel 3 -preassemble -bandwidth 10 -repeat\_stringency 0.98 -ace -minmatch 5 -maxmatch 20 -minscore 10'.

To investigate the effect of quality scores on estimates of diversity, the assembled reads for each barcoded sample were further filtered by consensus quality scores. Barcodes were pruned from sequences and the data sets were aligned in the RDP pyrosequence aligner (Cole *et al.*, 2009), which implements the Infernal algorithm (Nawrocki and Eddy, 2007). The resulting alignments were edited to remove primer sequences and clustered to generate operational taxonomic units (OTUs) using the RDP clustering algorithm. The taxonomic assignments of OTUs were derived from the RDP Classifier using a 50% bootstrap cutoff, as recommended for reads less than 100 bp. Custom PERL scripts were

written to generate a pipeline to analyze, organize and format these data. The PERL scripts and raw sequence data presented and analyzed here are available at [www.yale.edu/ochman/data](http://www.yale.edu/ochman/data).

## Results

### High read recovery, low read usability

Over 30 million paired-reads were collected from each of the two multiplexed Illumina sequencing runs. Initial filtering removed 85% of the reads from *run1* and 54% from *run2* owing to errors stemming from several sources (Table 1), which is comparable to other studies that report filtering 40–70% of their initial reads (Gloor *et al.*, 2010; Caporaso *et al.*, 2011). We secured a total of 4.5 and 16.0 million paired-end reads for the two runs in spite of the high number of reads that were discarded because they did not meet our thresholds for accuracy; this, in turn yielded tens of thousands to millions of raw read-pairs per multiplexed sample (*run1*  $\bar{x} = 380\,253 \pm 176\,319$ ; *run2*  $\bar{x} = 3\,936\,966 \pm 2\,299\,439$ ) (Table 2). Because the runs were founded with equimolar concentrations of amplicons, some of the variation in the number of reads returned per sample is most likely attributable to amplification biases that occurred during the synthesis of the Illumina sequencing library. Future experiments might avoid this bias by implementing an amplification-free library approach (Kozarewa *et al.*, 2009).

We used a strict set of quality filters to minimize the impact of erroneous reads on measures of microbial diversity (Kunin *et al.*, 2010). Errors in the barcode/primer sequence of either of the paired-end reads necessitates the removal of both reads in the pair; and in the majority of read-pairs that were culled, either only one of the reads perfectly matched the primer sequence or the paired reads lacked identical barcodes. The primer and barcodes are not informative for sequence classification and together constitute over 20% of the sequencing read, but the inclusion of both sequences provide a useful filter for removing erroneous reads and estimating sequencing error rates.

**Table 1** Summary statistics of Illumina paired-end runs

	run1 2 × 75	run2 2 × 100
	Number (%)	Number (%)
Total reads	30 024 020	35 372 083
Reads with N's	85 581 (0.3)	1 776 183 (5.0)
Only one primer	8 595 494 (28.6)	5 825 314 (16.5)
No primers	146 399 (0.5)	451 498 (1.3)
Mismatched barcodes	16 654 891 (55.5)	8 734 300 (24.7)
Low read quality	58 421 (0.2)	2 582 009 (7.3)
Incorrect barcode	2267 (0.01)	16 474 (0.05)
Binned, barcoded	4 480 967 (14.9)	15 986 305 (45.2)

**Table 2** Binning, assembly and quality filtering of iTags for the analyzed samples

Sample	Run	Barcode and primer	Initially binned		Assembled		Quality filtered				
			No.	% of run <sup>a</sup>	No.	% of binned	Threshold	No.	% of assembled	No. of 100% OTUs	
<i>E. coli</i> K-12	1	ATG-S <sub>1</sub> V6	487 169	2%	486 958	100%	q00	486 939	100%	8322	
							q20	464 022	95%	6189	
							q30	405 413	83%	5252	
	1	AGC-S <sub>2</sub> V6	317 746	1%	317 746	100%	q00	312 277	98%	5072	
							q20	293 527	92%	3849	
							q30	225 505	71%	3081	
	1	CCAT-S <sub>2</sub> V6	300 261	1%	300 128	100%	q00	299 660	100%	4090	
							q20	278 748	93%	2983	
							q30	229 091	76%	2467	
	2	GTC-LV6	2 311 017	7%	2 262 281	98%	q00	2 035 748	90%	159 224	
							q20	684 897	30%	4075	
							q30	72 955	3%	775	
19 Strains	1	CAG-S <sub>1</sub> V6	721 809	2%	721 367	100%	q00	721 066	100%	16 598	
							q20	675 333	94%	12 501	
							q30	556 184	77%	10 120	
	1	ATT-S <sub>2</sub> V6	226 600	1%	226 510	100%	q00	226 505	100%	6177	
							q20	210 797	93%	4930	
							q30	162 125	72%	3869	
	2	AAC-LV6	5 562 915	17%	5 472 113	98%	q00	5 422 964	99%	516 427	
							q20	2 378 803	43%	27 347	
							q30	407 771	7%	6658	
	<i>M. domesticus</i> WSB	1	GCA-S <sub>1</sub> V6	384 480	1%	384 096	100%	q00	384 040	100%	15 623
								q20	354 604	92%	12 272
								q30	277 670	72%	9549
1		TCC-S <sub>2</sub> V6	223 708	1%	223 610	100%	q00	223 488	100%	7082	
							q20	210 455	94%	5748	
							q30	163 335	73%	4398	

Abbreviation: OTU, operational taxonomic unit.

<sup>a</sup>Percentage based on the fraction of total paired-end reads not containing an 'N' *run1* = 29 938 439; *run2* = 33 595 900.

Error frequencies were estimated for the barcode/primer regions of all reads with high sequence quality ( $\geq q30$ ) for the first 25 nt and with single-nucleotide differences from the expected sequence. This included reads initially filtered out for not containing a perfect primer or barcode sequence (see above). Across the 14 barcode/primer combinations analyzed from *run1*, the error rates per nucleotide varied over three-fold ( $1.6 \times 10^{-4}$  to  $5.6 \times 10^{-4}$ ), and rates varied only two-fold ( $4.1 \times 10^{-4}$  to  $8.4 \times 10^{-4}$ ) among the four barcode/primer combinations from *run2* (Supplementary Table 1). For two reasons, we ascribe these variants to errors generated during sequencing rather than to artifacts that arose during PCR amplification: (1) Oligonucleotide synthesis is typically very accurate, such that errors in primer regions are more likely generated during the sequencing step; (2) we performed independent experiments using two different *Taq* polymerases each having unique error rates, but obtained approximately the same rate (analysis of covariance;  $F_{1,15} = 1.07$ ,  $P = 0.317$ ) and spectrum of mutations with each. The fact that error rates were independent of the source of *Taq* polymerase further suggests that errors are introduced during DNA sequencing.

There were large numbers of paired-end reads with mismatched barcodes, but owing to our use of

barcodes that require at least two mutations to convert one barcode into another, not many of these mismatched read-pairs were attributable to sequencing errors. In fact, many read-pairs contained recognizable barcodes that differed between the corresponding reads, which could result from (1) over-clustering during the sequencing run; (2) the production of chimeric amplicons during the amplification step of the Illumina library or (3) primer contamination, the likelihood of each of which is discussed below.

*Over-clustering* can result from low nucleotide diversity in the first four sequenced nucleotides and high cluster density on the flow cell. Although the first four bases of our barcodes and primers did not contain equal frequencies of each of the four nucleotides, over-clustering is an unlikely cause of the mismatched barcodes detected. An analysis of the run showed that read signal intensities and the percentage of clusters passing filter were both above normal expectations (intensity > 200 for first cycle, > 80% clusters passing filter). Alternatively, chimeric amplicons are a common by-product of 16S rDNA amplifications (Qiu *et al.*, 2001); however, our protocols used a low number of amplification cycles (10–12) and the high-fidelity *Taq* polymerase for library construction, both of which reduce the frequency of chimeric amplicons. This leaves primer

**Table 3** Characterization of the bacterial strains used in the '19 Strains' community

No.	Bacterial species	Strain	Family	rRNA operons	Genome G+C <sup>a</sup>	rRNA G+C	100% OTU-V6	Accession no. <sup>b</sup>	% DNA
1	<i>Bacillus subtilis</i>	168	Bacillaceae	10	43	54.9	1 <sup>c</sup>	<b>NC_000964</b>	3.4
2	<i>Bacillus cohnii</i>	RSH	Bacillaceae	NA	33.5–35.0	54.3	1	X76437	0.7
3	<i>Bacillus clarkii</i>	JP395	Bacillaceae	NA	42.4–43.0	56	1	X76444	0.2
4	<i>Bacillus gibsonii</i>	PD138	Bacillaceae	NA	40.6–41.7	55	1	X76446	0.6
5	<i>Staphylococcus intermedius</i>	H11/68	Staphylococcaceae	NA	33.5	52	1	D83369	0.9
6	<i>Staphylococcus vitulinus</i>	DD756	Staphylococcaceae	NA	34	51.1	2	AB009946	0.7
7	<i>Staphylococcus arlettae</i>	BP47	Staphylococcaceae	NA	31–32.6	51.4	2	AB009933	1.2
8	<i>Streptococcus suis</i>	735	Streptococcaceae	4	41	52.8	1	AB002525, <b>NC_009442*</b>	4.8
9	<i>Streptococcus cristatus</i>	CR311	Streptococcaceae	NA	42.6–43.2	52.8	1	EU156757	0.7
10	<i>Agrobacterium vitis</i>	CG47	Rhizobiaceae	4	57.57	54.7	1	<b>NC_011989*</b>	0.5
11	<i>Agrobacterium radiobacter</i>	K84	Rhizobiaceae	3	60.59	55.1	1	<b>NC_011985</b>	0.5
12	<i>Agrobacterium tumefaciens</i>	C58	Rhizobiaceae	4	59.59	54.7	1	<b>NC_003062</b>	0.7
13	<i>Escherichia fergusonii</i>	2022	Enterobacteriaceae	7	49	54.5	1	<b>NC_011740*</b>	63.1
14	<i>E. coli</i>	K-12, MG-1655	Enterobacteriaceae	7	50	54.4	2	<b>NC_000913</b>	0.4
15	<i>Salmonella enterica</i> subsp. enterica sv. Typhimurium	14028S	Enterobacteriaceae	7	52	54.4	3	<b>CP001363</b>	0.2
16	<i>S. enterica</i> subsp. enterica sv. Paratyphi C	SARB49	Enterobacteriaceae	7	52	54.5	2	<b>NC_012125</b>	1.9
17	<i>S. enterica</i> subsp. enterica sv. Agona	SARB1	Enterobacteriaceae	7	52	54.7	2	<b>NC_011149*</b>	11.5
18	<i>S. enterica</i> subsp. enterica sv. Typhi	SARB63	Enterobacteriaceae	7	52	54.6	1	<b>NC_003198*</b>	6.4
19	<i>Xylella fastidiosa</i>	2679 PLM G83	Xanthomonadaceae	2	52	54	1	AF203388, <b>NC_002488*</b>	1.6

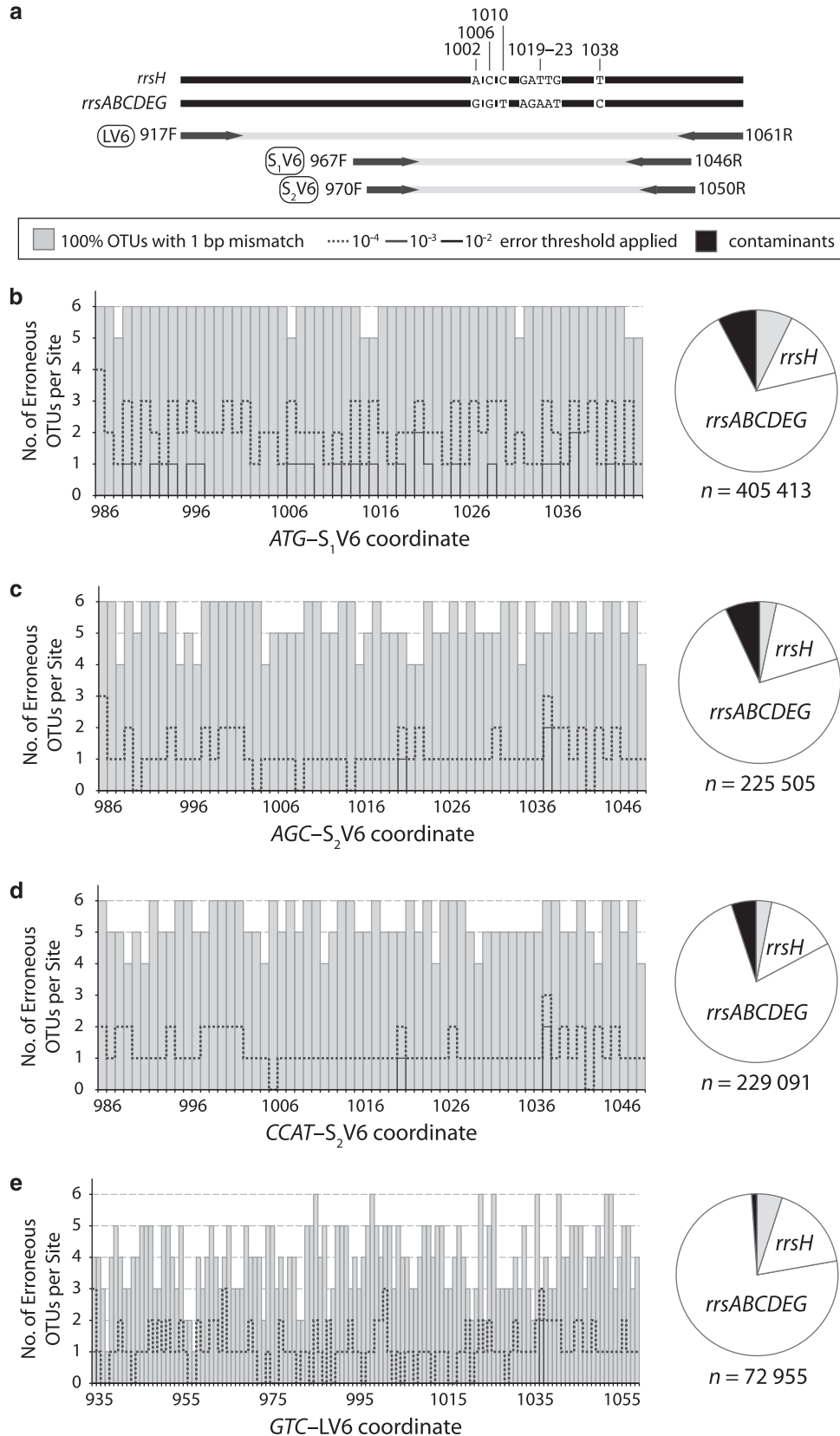
<sup>a</sup>The genome G+C contents for strains without genome sequence are based on the species description citation. The values separated by a comma indicate the G+C contents of chromosome-I and II, respectively.

<sup>b</sup>The numbers in bold indicate completely sequenced genomes and the asterisks (\*) indicate sequenced strains that are similar to the one used.

<sup>c</sup>The *B. subtilis* genome contains one divergent 16S rRNA gene that was not found as an amplicon. Data not available are indicated by 'NA'.

contamination during the initial PCR amplifications as the principal contributor of mismatched barcodes. Although we implemented precautionary

measures intended to minimize contamination (using screw cap tubes; removing primers and templates individually; setting up the reactions in



a laminar flow hood), contamination occurred in these samples. With current interest in multiplexing scores (or even hundreds) of samples, we let these results serve to forewarn investigators that highly stringent procedures are necessary to prevent cross-contamination.

#### Recapitulation of *E. coli* OTUs

We first evaluated the efficacy and accuracy of the iTag approach by sequencing a single template, *E. coli* K-12 (MG1655). The *E. coli* genome contains seven rRNA operons, six of which have identical alleles at V6 and the seventh encoding a V6 allele with nine nucleotide differences (Figure 1a). At both 100% and 97% identity thresholds only two OTUs are expected.

The V6 loop of *E. coli* was amplified using three primer pairs ( $S_1V6$ ,  $S_2V6$  and LV6), and we performed a technical replicate of  $S_2V6$  (Table 2). Nearly all reads assembled with well-supported overlaps of their distal 3' regions ( $S_1V6$  and  $S_2V6$ ,  $\approx 50$  bp; LV6,  $\approx 30$  bp). Alignment and clustering of the assembled iTag sequences showed that each of the *E. coli* samples contained much more than the two predicted phylotypes ( $n = 4090\text{--}159\,224$ ; Table 2).

Previous work has shown that quality trimming of pyrotags is essential for accurate prediction of community composition (Kunin *et al.*, 2010); therefore, we filtered the assembled reads according to consensus quality scores estimated by Phrap. We compared two quality filters, requiring that all sites had error probabilities of either  $\leq 10^{-2}$  or  $\leq 10^{-3}$  (corresponding to Phred scores of  $\geq q20$  or  $\geq q30$ ). These measures reduced the number of unique phylotypes (aka 100% OTUs), but failed to remove all of the spurious OTUs (Table 2).

Among these confounding OTUs, we found that 3–7% of high-quality reads ( $\geq q30$ ) differed by only 1 nt from the expected *E. coli* alleles (Figures 1b–e). Variants occurred at every nucleotide position regardless of the primer pair or the source of *Taq* polymerase. Given the numbers of variants, it is not likely that each represents a new mutation that occurred during growth in culture. Therefore, these phylotypes are attributable to errors that occurred during sequencing.

The very large number of sequence variants made them impossible to cluster into the two expected OTUs, even when applying very relaxed thresholds

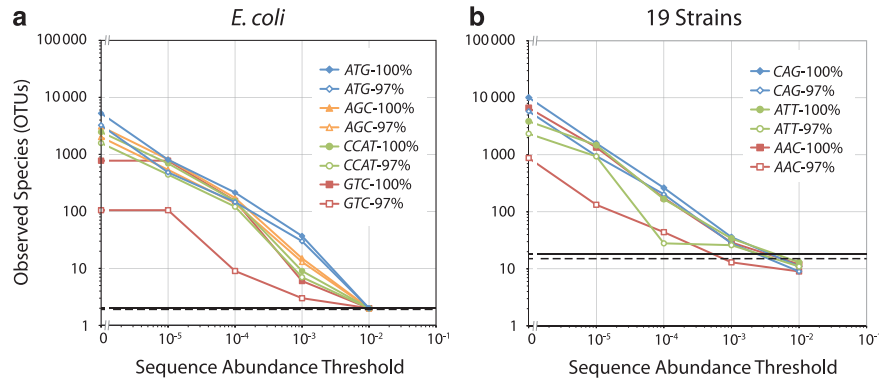
(for example, 90% OTUs). Most of the remaining OTUs (Figure 1) represent phylotypes that were detected in other samples and are best explained by primer contamination (discussed above). In addition, some of the contaminants in the LV6 samples were because of nonspecific amplification of other chromosomal loci; for example, we recovered a 123-bp fragment of uridylyltransferase (*glnD*) from *E. coli* and a 78-bp fragment of an ATP transport permease from *Streptococcus* sp.

To rid samples of artifactual phylotypes generated by sequencing errors and PCR contamination, we used a sequence abundance threshold in a manner similar to that of Caporaso *et al.* (2011). We based this threshold on the representation of a unique phylotype (100% OTU) given the number of q30 reads for that barcode. Through stepwise increments in the abundance threshold, we followed the degree to which spurious reads were eliminated (Figures 1 and 2). Eventually, by applying a 1.0% abundance threshold (that is, removing all OTUs occurring at frequencies less than 1.0%) we were able to recapitulate the two phylotypes expected to occur in the *E. coli* genome.

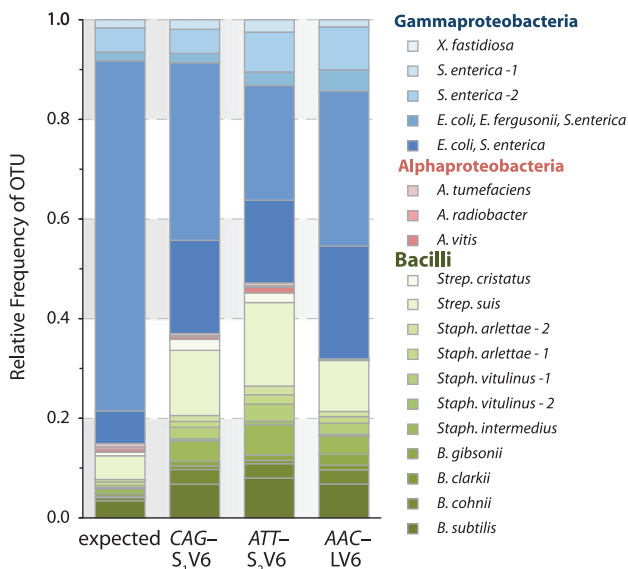
#### Diversity in a defined community

Having addressed the sequence variation detected within a single species by the iTag approach, we analyzed the data generated for a synthetic community of 19 strains of bacteria (Table 3). This sample was amplified using each of the three primer sets, which should yield 19 100% OTUs, and 16 ( $S_1V6$ ,  $S_2V6$ ) or 15 (LV6) 97% OTUs. A similar pattern showing an excess of erroneous and contaminating high-quality OTUs emerged even after stringent quality filtering (Table 3). Although the application of a 1% abundance threshold removed artifactual reads, it also removed several of the expected OTUs that occurred below this threshold (Figure 2). The initial template DNA concentrations of 12 of the 19 strains represented less than 1% of the total DNA mass, so this result was anticipated (Table 3); however, we also found that the actual read abundances for the 19 100% OTUs departed from the expected values (Figure 3). The likely cause of the deviation between the relative amount of input DNA and the relative number of reads could be primer specificity. For example, the *Streptococcus cristatus* OTU was similarly abundant in the CAG– $S_1V6$  (1.7% = 2754/162 125) and ATT– $S_2V6$

**Figure 1** iTag design and diversity in *E. coli*. (a) Three primer combinations were tested mapping to  $\approx 100$ - or  $\approx 160$ -bp regions spanning the V6 loop of bacterial 16S rRNA. The schematic representation illustrates the two alleles (black bars differing at nine positions) in the 16S rRNA genes of *E. coli* MG1655 (*rrsA-H*) that are differentiated by the resulting amplicons (LV6,  $S_1V6$  and  $S_2V6$ , with forward and reverse primer positions designated). In panels (b–e), the location and count of high-quality, 100% OTUs with single sequencing errors relative to the expected *E. coli* reference alleles are plotted for each of four *E. coli* iTag samples. The dashed and solid lines indicate the numbers of mutations after application of increasingly stringent sequence abundance thresholds ( $10^{-4}$ ,  $10^{-3}$ ,  $10^{-2}$ ) used to remove erroneous OTUs. The pie charts indicate the relative abundance of iTags corresponding to expected *E. coli* OTUs (white), tags with 1-bp errors (gray) and putative contaminants (black) without the application of an abundance threshold.



**Figure 2** The thresholds required to attain the actual rRNA diversity in a sample. Observed species estimates (OTUs) at 100% (filled) and 97% (open) levels are plotted at a series of sequence abundance thresholds for (a) *E. coli* and (b) 19 Strain samples. The solid and dashed lines represent the expected numbers of OTUs at 100% and 97% identity thresholds, respectively. An identical pattern emerges when the numbers of estimated species (Chao1) are used (data not shown).



**Figure 3** Frequencies of OTUs from defined communities. The stacked bar (left) indicates the expected frequencies of each of the 19 species based on input DNA, followed by the frequencies obtained in each of three iTag analyses. The abbreviations are as follows: *X.*, *Xylella*; *S.*, *Salmonella*; *E.*, *Escherichia*; *A.*, *Agrobacterium*; *Strep.*, *Streptococcus*; *Staph.*, *Staphylococcus*; *B.*, *Bacillus*.

(1.8% = 10 207/556 184) data sets, but nearly absent from the AAC-LV6 data set (0.0022% = 9/407 785).

Short 16S tags can be difficult to assign taxonomically and alone provide limited phylogenetic resolution (Liu *et al.*, 2007). We found that the RDP Classifier accurately identified all of the 19 expected OTUs to taxonomic Class, but its success rate progressively decreased for lower taxonomic divisions, with only about 60–90% assigned correctly. Alternative BLAST-based classification approaches can increase accuracy provided related full-length 16S rRNA sequences are available (Huse *et al.*, 2008). Of course, neither of these

approaches are able to distinguish among species that have identical V6 loop sequences (for example, *E. coli*, *Escherichia fergusonii* and *Salmonella enterica*; Figure 3).

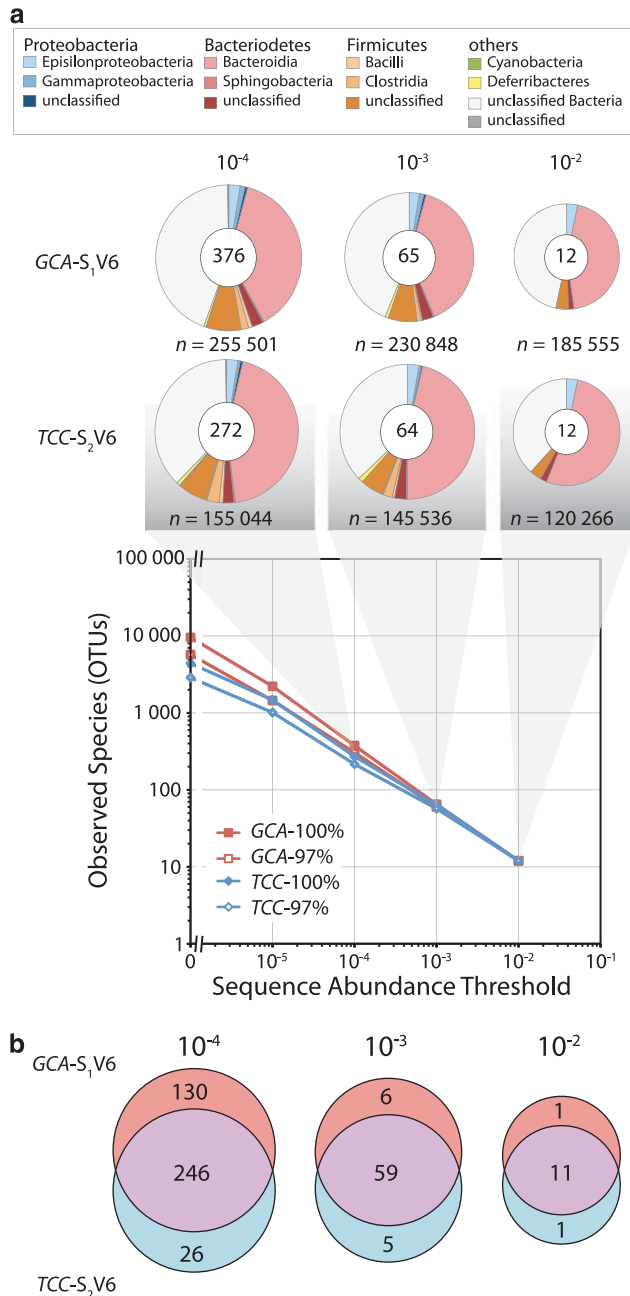
#### Microbial diversity in a mammalian host

By way of analyzing a complex microbial community whose composition and constituents are uncharacterized, we assessed the V6 sequence diversity within a fecal sample from the house mouse *M. domesticus* strain WSB/Eij. This sample was amplified with *GCA-S<sub>1</sub>V6* and *TCC-S<sub>2</sub>V6*, generating, respectively, 384 480 and 223 708 binned reads. As before, even with rigorous quality filtering, a large number of low-frequency 100% OTUs remained (Table 2). We applied similar abundance thresholds ( $10^{-5}$  to  $10^{-2}$ ), which resulted in a dramatic reduction in phylotypes (Figure 4). The utilization of the same threshold as that applied to the *E. coli* samples (that is,  $10^{-2}$ ) leaves only 12 OTUs, which represent 67% (*GCA-S<sub>1</sub>V6*) and 74% (*TCC-S<sub>2</sub>V6*) of the high-quality reads. Thus, it appears that application of a 1.0% threshold is too stringent for actual biological samples and probably removes many tags that are likely to represent the actual microbial constituents of these communities. The two amplified samples show an extensive overlap in 100% OTUs (Sørensen distance, 0.010; Jaccard distance, 0.020), which results in a highly similar taxonomic distribution regardless of the abundance threshold applied (Figure 4).

## Discussion

Massively parallel sequencing platforms supply an efficient and sensitive means of characterizing microbial communities, but the sequencing depths offer by these methods are accompanied by artifacts that can affect the interpretation of community





**Figure 4** iTag diversity and taxonomic representation in the colon of house mouse. **(a)** The pie charts indicate the relative abundances and taxonomic affiliation of the resultant OTUs based on the RDP Classifier labeled according to the key. The plot shows the numbers of OTUs after application of different abundance and clustering thresholds on the 100% OTUs derived for the *M. domesticus* WSB samples GCA-S<sub>1</sub>V6 and TCC-S<sub>2</sub>V6. **(b)** The numbers of phylotypes common to the 100% OTUs predicted at three increasingly stringent sequence abundance thresholds (10<sup>-4</sup>, 10<sup>-3</sup>, 10<sup>-2</sup>).

diversity. We have identified and quantified several sources of errors inherent to 16S amplicon sequencing (iTags) on the Illumina platform (contamination, amplification and sequencing errors), thereby enhancing the utility of these data for assessing

community diversity. Most recent efforts to interrogate microbial communities have relied upon 16S rRNA gene sequences generated by 454 pyrosequencing (for example, Dethlefsen *et al.*, 2008; Bowers *et al.*, 2009; Costello *et al.*, 2009; Cardenas *et al.*, 2010; McLellan *et al.*, 2010); however, Illumina sequencing technologies currently produce much larger numbers of much shorter reads for a fraction of the cost and will likely supplant the use of 454 pyrotags.

Our experimental iTag design used barcoded primers flanking the V6 segment of 16S rRNA, a region commonly amplified in pyrotag experiments (Sogin *et al.*, 2006; Keijsers *et al.*, 2008; Galand *et al.*, 2009). Individual samples were amplified, mixed and then used as templates to construct and sequence two standard Illumina paired-end libraries. Although Illumina sequence quality decays along the length of reads, our motivation for assembling paired-end reads was to increase the quality and confidence of the overlapping region (Gloor *et al.*, 2010; Rodrigue *et al.*, 2010; Zhou *et al.*, 2011). Therefore the amplified V6 region was restricted to 100–160 bp to ensure an adequate overlap of the forward and reverse paired-end reads. We subsequently filtered out low-quality tags, leaving only assembled reads that contained barcode and primer sequences at each end, which bounded a phylogenetically informative amplified region. This design allowed the simultaneous estimation of sequencing error rates and microbial diversity.

Over half of > 30 million paired-end reads from each of Illumina runs were subsequently discarded owing to either (1) sequencing errors in one or both of the primer regions or (2) paired-end reads that contained mismatched barcodes (Table 1). Other Illumina-based 16S rRNA studies have encountered similarly high sequencing error rates, making such extreme read filtering necessary, but these studies did not report a high incidence of mismatched barcodes (Gloor *et al.*, 2010; Caporaso *et al.*, 2011). After ruling out over-clustering and sequence chimeras, primer contamination during the initial sample amplifications was viewed as the most likely cause of the mismatched barcodes (see above). For example, > 60% of the contaminants detected in the *E. coli* libraries were present in one or more of the other samples, and frequently these contaminants represented the more abundant phylotypes in the other samples (data not shown). The sequencing depth afforded by the Illumina platform greatly increases the likelihood of detecting contaminants; however, we stress the utility of barcoding and sequencing both ends of multiplexed amplicons in order to detect this source of errors. In the absence of paired barcodes, removing reads that lack a significant match to a reference database can be implemented. Although this latter method is of limited utility when examining novel or poorly characterized communities.

### Analysis of contrived samples

To establish baseline levels of sensitivity of the iTag approach, we performed several controls, including the amplicons generated from a single bacterial genome and from a defined mixture of DNA from 19 strains. In both cases, we found an excess of high-quality, 100% phylotypes regardless of the primer pair, *Taq* polymerase or sequencing run. However, sequencing reads bearing a single mutation relative to the two expected *E. coli* phylotypes, or to the 19 phylotypes from the mixture of strains, represented 2.2–8.2% of the reads. Based on these reads we estimate a sequencing error rate of  $1.3 \times 10^{-3}$  to  $4.1 \times 10^{-4}$  per nucleotide (Supplementary Table 2), which is slightly greater than the rate estimated from errors confined to the primer regions (*run1*  $\bar{x} = 3.8 \times 10^{-4}$ ; *run2*  $\bar{x} = 6.4 \times 10^{-4}$ ). This difference is in line with expectations as, in the Illumina platform, error rates are known to increase with distance from the priming site (Rodrigue *et al.*, 2010; Zhou *et al.*, 2011). These errors impact taxonomic identification using the RDP Classifier, particularly reducing the ability to assign Order, Family and Genus identities to the shorter S<sub>1</sub>V6 and S<sub>2</sub>V6 tags.

To account for the large number of spurious reads in these samples, we applied strict abundance thresholds to our samples. We came the closest to recapitulating our expected diversity from these samples when we applied an abundance threshold at which all phylotypes at frequencies <1% were removed. This cutoff is based on the total number of high-quality reads for a given sample and more-or-less corresponds to the 0.01% threshold applied by Caporaso *et al.* (2011), whose threshold is based on the total number of read-pairs recovered for their entire experiment (10 000 out of 87 507 177 reads). Application of such stringent thresholds is not particularly satisfying, but will remain a necessary measure until sequencing error rates are improved.

### Measurements of mouse intestinal flora

Mammalian guts house a diverse array of resident microbes and are numerically dominated by the Firmicutes and Bacteroidetes (Ley *et al.*, 2008). A previous pyrosequencing effort identified 1000 bacterial phylotypes (97% OTUs) in healthy *Mus musculus* individuals using the same S<sub>1</sub>V6 primers (Antonopoulos *et al.*, 2009); however, our sample from a lab-reared, wild-derived inbred strain of *M. domesticus* WSB yielded only 12 phylotypes (97% OTUs) after application of the equivalent quality and abundance filters derived from the contrived samples ( $q_{30}$  and  $10^{-2}$ ). Reducing the stringency of the abundance threshold increased the number of recovered iTag phylotypes, but did not alter the taxonomic distribution of the iTags (Figure 4). (Lowering the abundance threshold from  $10^{-2}$  to  $10^{-4}$  yielded about 20% more iTags containing only a single-nucleotide difference relative to the 12 high abundance OTUs.) The source of the

difference in the number of phylotypes detected in the *Mus* fecal samples is unknown, and direct comparisons between iTags and pyrotags can only be assessed from the corresponding analyses of 16S diversity within *E. coli* (Kunin *et al.*, 2010; this study). Therefore, we suggest the amplification and sequencing of a control sample (for example, *E. coli*) during each multiplex experiment to empirically determine appropriate abundance thresholds and to account for differences in sequencing error rates between runs.

### Alternative iTag implementations

Interest in leveraging the Illumina sequencing platform to examine 16S rRNA gene diversity has led to the recent development of several strategies and observations about its implementation (Claesson *et al.*, 2010; Gloor *et al.*, 2010; Caporaso *et al.*, 2011; Zhou *et al.*, 2011). Until recently, the Illumina sequence read length was limited to  $\approx 100$  nt. As a result, paired-end sequencing of single, short variable loops (V6) has been assembled (Gloor *et al.*, 2010; Zhou *et al.*, 2011; this study), and longer variable region(s) (255–465 bp) were artificially merged or analyzed separately (Claesson *et al.*, 2010; Caporaso *et al.*, 2011). Assembly of overlapping, paired-end reads results in phylotypes of considerably better quality because it identifies and rectifies problems that arise from the deterioration of read quality as a function of read length (Claesson *et al.*, 2010; Gloor *et al.*, 2010; Zhou *et al.*, 2011). But because PCR and sequencing errors persist, application of abundance thresholds are still necessary to avoid inflated estimates of species richness (Gloor *et al.*, 2010; Zhou *et al.*, 2011). Moreover, conservative estimates of error rates using contrived DNA samples vary six-fold, from  $2.4 \times 10^{-3}$  to  $4.1 \times 10^{-4}$  per nucleotide (Zhou *et al.*, 2011; this study).

The iTag approach holds enormous promise as means to investigate microbial diversity, but in its current applications there are three major obstacles: (1) contamination, (2) the utility and classification of short read lengths, and (3) sequencing error rates. Contamination can likely be mitigated by changes in laboratory procedures, including use of liquid-handling robots and sterile work environments. Although not considered here, contamination derived from reagents or consumables is also possible and would require alternative strategies for its detection and elimination. Despite the prevalence of studies using the V6 region and evidence that short reads perform adequately for community analyses (Liu *et al.*, 2007), short V6 tags appear to systematically overestimate species richness (Youssef *et al.*, 2009). Current Illumina read lengths ( $\approx 150$  nt) will allow the recovery and assembly of larger (V4) or more (V6+V7) 16S variable regions that better reflect the microbial diversity obtained when analyzing the entire 16S rRNA molecule. Additionally,

longer read lengths will facilitate the use of additional barcodes that will allow multiplexing hundreds of samples in a single run. Although the observed error rates are of little consequence to most genomic studies, as errors are resolved through sequencing depth, amplicon tagging approaches treat each unique sequence read as a novel OTU. Until there are improvements to the Illumina sequencing chemistry and analysis programs, erroneous reads will need to be filtered through the application of abundance thresholds (Gloor *et al.*, 2010; Caporaso *et al.*, 2011). The cost and capacity of the Illumina sequencing make it the preeminent platform for assessing microbial community diversity, and we have called attention to the errors that need to be resolved in its application and in the interpretation of results.

## Acknowledgements

We thank Michael Nachman, Nancy Moran and Zakee Sabree for contributing the template material used for library construction, and Zakee Sabree for coining the term 'iTag'. Funding for this project was provided by the National Institutes of Health to HO.

## References

- Antonopoulos DA, Huse SM, Morrison HG, Schmidt TM, Sogin ML, Young VB. (2009). Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infect Immun* **77**: 2367–2375.
- Blattner FR, Plunkett III G, Bloch CA, Perna NT, Burland V, Riley M *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453–1462.
- Bowers RM, Lauber CL, Wiedinmyer C, Hamady M, Hallar AG, Fall R *et al.* (2009). Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. *Appl Environ Microbiol* **75**: 5121–5130.
- Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108**(Suppl 1): 4516–4522.
- Cardenas E, Wu WM, Leigh MB, Carley J, Carroll S, Gentry T *et al.* (2010). A combined massively parallel sequencing—indicator species approach revealed significant association between sulfate-reducing bacteria and uranium-reducing microbial communities. *Appl Environ Microbiol* **76**: 6778–6786.
- 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.
- Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.
- Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.
- Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**: e280.
- Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H *et al.* (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**: 642–647.
- Fisher MM, Triplett EW. (1999). Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* **65**: 4630–4636.
- Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* **106**: 22427–22432.
- Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R *et al.* (2010). Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS One* **5**: e15406.
- 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.
- Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4**: e1000255.
- Keijsers BJ, Zaura E, Huse SM, van der Vossen JM, Schuren FH, Montijn RC *et al.* (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* **87**: 1016–1020.
- Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**: 291–295.
- 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**: 118–123.
- Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M *et al.* (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**: 266–271.
- Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS *et al.* (2008). Evolution of mammals and their gut microbes. *Science* **320**: 1647–1651.
- Li H, Ruan J, Durbin R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851–1858.
- Liu W, Marsh T, Cheng H, Forney L. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* **63**: 4516–4522.
- 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.
- McLellan SL, Huse SM, Mueller-Spitz SR, Andreishcheva EN, Sogin ML. (2010). Diversity and population

- structure of sewage-derived microorganisms in wastewater treatment plant influent. *Environ Microbiol* **12**: 378–392.
- 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.
- Nawrocki EP, Eddy SR. (2007). Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol* **3**: e56.
- Neefs JM, Van de Peer Y, De Rijk P, Goris A, De Wachter R. (1991). Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res* **19**(Suppl): 1987–2015.
- 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.
- Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM *et al.* (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* **67**: 880–887.
- Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ *et al.* (2010). Unlocking short read sequencing for metagenomics. *PLoS One* **5**: e11840.
- 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.
- Youssef N, Sheik CS, Krumholz LR, Najjar 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.
- Zhou H-W, Li D-F, Tam NF-Y, Jiang XT, Zhang H, Sheng H-F *et al.* (2011). BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J* **5**: 741–749.

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