Title
Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates

Author
Kunin, Victor

Publication Date
2009-09-01

Peer reviewed

Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates

Victor Kunin[1], Anna Engelbrektson[1], Howard Ochman[2], and Philip Hugenholtz[1¶]

[1] Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, 94598 CA, USA, [2] Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, USA.

[¶]Corresponding author: fax 925-296-5720 • email: phugenholtz@lbl.gov

Email addresses:

Victor Kunin: vkunin@lbl.gov

Anna Engelbrektson: AEngelbrektson@lbl.gov

Howard Ochman: hochman@email.arizona.edu

Philip Hugenholtz: PHugenholtz@lbl.gov

SUMMARY

Massively parallel pyrosequencing of the small subunit (16S) ribosomal RNA gene has revealed that the extent of rare microbial populations in several environments, the "rare biosphere", is orders of magnitude higher than previously thought. One important caveat with this method is that sequencing error could artificially inflate diversity estimates. Although the per-base error of 16S rDNA amplicon pyrosequencing has been shown to be as good as or lower than Sanger sequencing, no direct assessments of pyrosequencing errors on diversity estimates have been reported. Using only *Escherichia coli* MG1655 as a reference template, we find that 16S rDNA diversity is grossly overestimated unless relatively stringent read quality filtering and low clustering thresholds are applied. In particular, the common practice of removing reads with unresolved bases and anomalous read lengths is insufficient to ensure accurate estimates of microbial diversity. Furthermore, common and reproducible homopolymer length errors can result in relatively abundant spurious phylotypes further confounding data interpretation. We suggest that stringent quality-based trimming of 16S pyrotags and clustering thresholds no greater than 97% identity should be used to avoid overestimates of the rare biosphere.

INTRODUCTION

Pyrosequencing (Margulies et al., 2005) is one of the leading technologies supplanting Sanger sequencing for comparative genomics and metagenomics. One emerging application is the pyrosequencing of 16S rRNA genes ("16S pyrotags") to profile the phylogenetic diversity within microbial communities. The large number of reads produced in a single pyrosequencing run provides unprecedented sampling depth, leading to the conclusion that the rare biosphere, *i.e*. the tail of the species abundance distribution, is substantially larger and more diverse than previously appreciated (Sogin et al., 2006).

One caveat, however, is that the intrinsic error rate of pyrosequencing could lead to overestimates of the number of rare phylotypes. Unlike genome sequencing projects in which sequencing errors can be corrected by assembly and sequencing depth, each read in a pyrotag analysis is interpreted as a unique identifier of a community member and therefore errors will potentially inflate diversity estimates. Sogin and coworkers, appreciating this risk, invested considerable effort to determine the error rates of first generation GS20 pyrosequencing using a mixture of 43 reference templates (Huse et al., 2007). They concluded that quality filtering based on the removal of reads with one or more unresolved bases (N's), errors in the barcode or primer sequence, and/or atypically short or long reads is sufficient to ensure per-base error rates lower than conventional Sanger sequencing while retaining >90% of the reads. Ideally, the number of operational taxonomic units (OTUs) from their analysis should have been 43, however they did not report OTU estimates of their synthetic community based on pre- or post-filtered pyrosequencing reads. Here we assess the effect of error rates in second generation FLX pyrosequencing on diversity estimates using pyrotags PCR-amplified from two regions of the 16S rRNA amplicons of a well-characterized laboratory isolate of *E. coli*.

RESULTS

Approximately 300-bp regions from the 5' and 3' ends of the 16S rRNA genes of *E. coli* MG1655 were PCR-amplified using adaptor-modified standard primer sets (A-27F/B-342R and B-1114F/A-1392R) and pyrosequenced from the 27-forward or 1392-reverse primers, producing a total of 9,781 reads. Of these, 4,254 and 4,244 (87% of the total reads) could be unambiguously assigned to the 5'-forward and 3'-reverse regions of the 16S rRNA molecule, respectively, based on the presence of error-free barcode and primer sequences.

Read quality filtering. Reads were quality filtered by applying either the current practice of removing reads with unresolved bases and/or anomalous read length, or quality score-based end-trimming at different stringencies (3% to 0.1% per base error probabilities). After quality filtering and trimming to a uniform length of 244 bp to enable comparisons across samples and regions, the resulting reads were compared to the 16S rRNA sequences from the *E. coli* MG1655 genome to determine error rates. The extent of improvement and data loss after applying such quality filtering and length trimming is presented in Figure 1. The 5'-forward region had, on average, 15% more reads with one or more errors than did the 3'-reverse region at each quality-filtering treatment (Table 1). This difference is due to the higher number of homopolymers in the 5'-forward region relative to the 3' region (62 *vs*. 50), because homopolymer miscounts are the major source of errors in pyrosequence data (Margulies et al., 2005; Huse et al., 2007).

The lower quality of data from the 5'-forward region resulted in ~15% fewer usable reads than from the 3'-reverse region. The now standard practice of removing reads with undetermined bases (*i.e*., N's) resulted in only a marginal improvement (~1%) in errorless reads. In contrast, we found that trimming based on quality scores had a more pronounced effect on error rate when relatively stringent per-base error probabilities were applied (≤0.2% producing >4% improvement in errorless reads; Table 1). The number of usable reads decreased sharply when the most stringent (0.1%) error probability was applied, indicating that the benefits of increasing the stringency of quality filtering stringency were not offset by data loss beyond 0.2% error probability for this dataset.

Clustering evaluation. Reads were aligned and clustered at various identity thresholds ranging from 100% (unique sequences) down to 90% (sequences that differ by 10% are clustered into a single OTU) (Table 1, Fig. 1). Assuming no sequencing errors, the theoretical number of clusters (OTUs) should correspond to the actual number of 16S phylotypes in the sample; and in the case of *E. coli* MG1655, the number of unique OTUs should be five in the 5'-forward region and one in 3'-reverse region (Table 1). Remarkably, unfiltered reads overestimate this diversity by two orders of magnitude, producing 643 and 385 unique OTUs from the 5'-forward and 3'-reverse regions, respectively (Table 1, Fig. 1). Moreover, we note that increases in the size of the dataset will increase the observed number of OTUs (Fig. S1).

In ranking the abundance of OTUs in our samples, the majority of reads possess the exact sequence of the corresponding region in an *E. coli* 16S rRNA gene; however, rank-abundance distributions for both regions were flanked by a long tail of OTUs containing one or more insertion and/or substitution errors relative to the *E. coli* reference sequences, and in the case of the 5'-forward region, two putative chimeric OTUs formed between different *E. coli* 16S operons (Fig. 2). A remarkable feature of the 5'-forward region distribution is that between the abundant error-free OTUs and the rare erroneous OTUs and singletons, there were several moderately abundant clusters, together constituting ~6% of the reads. These OTUs contain the same re-occurring homopolymer error; 6 instead of 5 guanines spanning *E. coli* positions 200 to 204 (Fig. 2).

The primary effect of clustering at different levels of sequence identity was to recruit erroneous OTUs and singletons into larger clusters, thereby decreasing exponentially the number of OTUs as identity thresholds were relaxed (Fig. 1). But even at the most relaxed threshold, there were two 5'-forward and one 3'-reverse OTUs that did not match *E. coli*. The closest matches (>98% identity) to these OTUs were members of the *Saprospirales* (Bacteroidetes), *Bradyrhizobiales (*Alphaproteobacteria*)* and *Peptostreptococcaceae* (Firmicutes). All other sequences clearly originated from *E. coli* and represent the overwhelming majority (99.97%) of the sequence data.

DISCUSSION

Despite a rigorous analysis of error rates in 16S rRNA pyrosequences of known templates (Huse et al., 2007), there have been no reports of the effect of pyrosequencing errors on diversity estimates (number of inferred phylotypes), and therefore, no way to gauge the accuracy of diversity reported in individual studies or to compare the observed variation of communities across studies. To resolve this issue, we chose to examine a single bacterial strain both to remove the complication of inter-species chimera formation (Huber et al., 2004) and to focus solely on the effect of pyrosequencing error on diversity estimates. Even with a fairly modest number of second generation 454 FLX reads from two regions of the 16S rRNA genes of *Escherichia coli* MG1655 (~4250 reads per region), we find that sequencing errors inflate estimates of the actual diversity by two orders of magnitude when considering unique reads (Fig. 1).

This overestimation is consistent with a high percentage of reads with one or more errors; ~15% and ~30% of reads for the 3'-reverse (V8) and 5'-forward (V1&2) regions, respectively (Table 1) also detected in prior analysis of the V6 region in which 18% of reads had ≥1 error (Huse et al., 2007). A large proportion of these artefacts is attributable to miscounted homopolymeric runs that occur in otherwise high quality regions of the read, and are therefore not removed by end-trimming based on quality scores (see below) or by culling reads with unresolved bases or anomalous lengths. Moreover, some of these errors are highly reproducible and produce phantom OTUs with large numbers of reads (Fig. 2), indicating that not only will false phylotypes be detected, but that, in some cases, spurious phylotypes will be relatively abundant (≥1%) at least in the case of 100% OTUs.

In practice, 100% sequence identity is rarely used as a threshold for defining OTUs, but rather, reads are usually grouped at some lower level of sequence identity (often 97% sequence identity (Stackebrandt and Goebel, 1994), which clusters sequences differing by as much as 3% into a single OTU). This has the effect of absorbing much of the observed sequencing errors. We tested a range of clustering thresholds, and as expected, clustering greatly reduces the overestimation of diversity (Fig. 1). However, we find that the current practice of removing reads with undetermined bases and/or anomalous read lengths is not adequate to ensure accurate diversity estimates at a 97% clustering threshold (Fig. 1). This occurs despite the comparable or lower per base error rates observed for 454 pyrosequencing when compared to conventional Sanger sequencing (Huse et al., 2007).

Recent improvements in error estimation of pyrosequence data (Brockman et al., 2008) allow the use of trimming programs, such as LUCY (Chou and Holmes, 2001), that are based on the per-nucleotide quality score. Only when the LUCY end-trimming stringency was increased to ≤0.2% per base error probability (equivalent to a *phred* quality score of ≥27), combined with clustering at ≤97% identity, did the number of OTUs approach the expected number of *E. coli* MG1655 rRNA operons. The slightly overestimated number of OTUs at these settings were, in fact, not sequencing artefacts, but most likely due to experimental contamination introduced during the PCR amplification, as seen previously with no-template PCR controls (Tanner et al., 1998). These contaminants represent only 0.03% of the reads obtained in the present study and suggest that all PCR-based surveys that use broad-specificity primers will likely suffer from similar low-level background contamination, a point worth bearing in mind when interpreting rare biosphere data.

Based on our analyses, we propose the use of quality trimming to 0.2% error probability and a clustering threshold of 97% identity when applying 454 pyrosequencing to community profiling. These parameters should substantially reduce artefactual inflation of diversity estimates due to pyrosequencing errors. Raising the trimming stringency from 0.2% to 0.1% error probability results in a sharp decrease in usable reads with little additional improvement in error reduction (Table 1). We note, however, that error rates are sequence specific (Fig. 1a *vs*. 1b) and that the spurious inflation of OTU numbers will increase with the size of the dataset (Fig. S1). Therefore, the proposed parameters may be insufficient to prevent overestimates of diversity using very large pyrotag datasets from regions of the 16S rRNA gene with a high fraction of homopolymers. Overall, we anticipate that the use of high stringency quality-based trimming and clustering thresholds ≤97% will be the simplest, least computationally intensive means to ensure that 16S pyrotag analyses provide accurate, high sensitivity phylogenetic profiling of microbial communities.

MATERIALS AND METHODS

DNA Extraction. *Escherichia coli* MG1655 was grown overnight at 37°C in 10 ml of LB and harvested by centrifugation at 10,000xg for 5 min. Cells were treated with proteinase K (20 mg/ml) and lysozyme (5 mg/ml), and DNA was isolated using a standard phenol-chloroform extraction, followed by ethanol precipitation.

PCR Amplicon Library Construction and Sequencing. One 5' and one 3' region of the 16S rRNA gene were targeted using the broad-specificity oligonucleotide primer pairs 27F/342R and 1114F/1392R (Stackebrandt and Goodfellow, 1991). Primer sequence (small caps) were modified by addition of the Roche 454 A or B adaptor sequences (lower case) and a five nucleotide identifying barcode (bolded uppercase) to distinguish different amplicons in the same sequencing reaction, as follows: A-27F, 5'-gcc tcc ctc gcg cca tca g**AC GTC** AGA GTT TGA TCM TGG CTC AG-3', B-342R, 5'-gcc ttg cca gcc cgc tca gCT GCT GCS YCC CGT AG-3', A-1392R, 5'-gcc tcc ctc gcg cca tca g**TG CTG** ACG GGC GGT GTG TRC-3' and B-1114F, 5'-gcc ttg cca gcc cgc tca gGC AAC GAG CGC AAC CC-3'. 20 µL PCR reactions were performed in triplicate for each primer pair, using 0.5 units Taq (GE Healthcare), 2 µL of supplied 10X buffer, 0.4 µL of 10 mM dNTP mix (MBI Fermentas), 0.6 µL of 10 mg/mL BSA (New England Biolabs), 0.2 µL of each 10 µM primer, and 10 ng of *E. coli* genomic DNA per reaction. Thermocycling proceeded as follows: 95ºC for 3 mins followed by 30 cycles of 95ºC for 30 sec, 55ºC for 45 sec, and 72ºC for 90 sec and final extension at 72ºC for 10 min. Upon completion, the three reactions for each primer pair were pooled, and amplicons were purified with the Qiagen MinElute PCR cleanup kit and quantified on a Qubit fluorometer (Invitrogen).

Barcoded amplicons were mixed in equal proportions prior to emulsion PCR in preparation for GS FLX pyrosequencing.

Informatic Analysis. Pyrosequencing flowgrams were converted to sequence reads using the standard software provided by 454 Life Sciences. Reads were either used directly (which served as the unfiltered control) or quality filtered in one of two ways: (*i*) reads with any unresolved nucleotides (N's) were removed from the dataset, or (*ii*) reads were end trimmed based on quality scores over a range of accuracy thresholds (0.1 to 3% per base error probabilities) using LUCY (Chou and Holmes, 2001). This resulted in eight quality filtered datasets (Table 1).

To compare sequences across samples, all reads in each of the datasets were truncated from their 3'end to 244 bp, and reads less than 244 bp were discarded. In the same step, barcodes and primer sequences were trimmed from the 5' end, and any read with a sequence error in its barcode and/or primer was removed. This resulted in 5'-forward reads spanning positions 28 to 246 (*E. coli* numbering), which encompasses variable regions 1 and 2, and the 3' reads spanning positions 1168 to 1391, which encompasses variable region 8 of the 16S rRNA molecule. From the remaining uniform length sequences, all redundant sequences were removed yielding a dereplicated dataset containing only unique phylotypes (termed the 100% OTUs in subsequent steps).

Unique truncated reads were aligned using a modified Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) and clustered along a range of identity thresholds (90, 95, 97, 98 and 99%) using MCL, executed with default parameters (Van Dongen, 2000). Sequencing errors in each of the unique reads were determined by BLAST alignment (Altschul et al., 1997) to the known 16S rRNA gene sequences of *E. coli* MG1655, assuming that any mismatches derived from the most similar of the seven *E. coli* operons. For the 5'-forward region considered, there are five unique 16S rRNA sequences in *E. coli*, and for the 3'-reverse region considered, all *E. coli* 16S sequences are identical. A subset of reads was manually inspected in ARB (Ludwig et al., 2004) to confirm the specific type and location of the BLAST-determined errors and to identify putative chimeras.

ACKNOWLEDGMENTS

We thank Suzan Yilmaz for extracting *E. coli* DNA and Alex Copeland for discussions on quality-score based filtering. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. VK was supported in part by NSF grant OPP0632359.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L. et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* **18**: 763-770.

Chou, H.H., and Holmes, M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* **17**: 1093-1104.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**: 2317-2319.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363-1371.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Needleman, S.B., and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R. et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* **103**: 12115-12120.

Stackebrandt, E., and Goodfellow, M. (1991) *Nucleic acid techniques in bacterial systematics*. Chichester ; New York: Wiley.

Stackebrandt, E., and Goebel, B.M. (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bacteriol* **44**: 846-849.

Tanner, M.A., Goebel, B.M., Dojka, M.A., and Pace, N.R. (1998) Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl Environ Microbiol* **64**: 3110-3113.

Van Dongen, S. (2000) Graph Clustering by Flow Simulation. In: University of Utrecht.

Figures and Legends

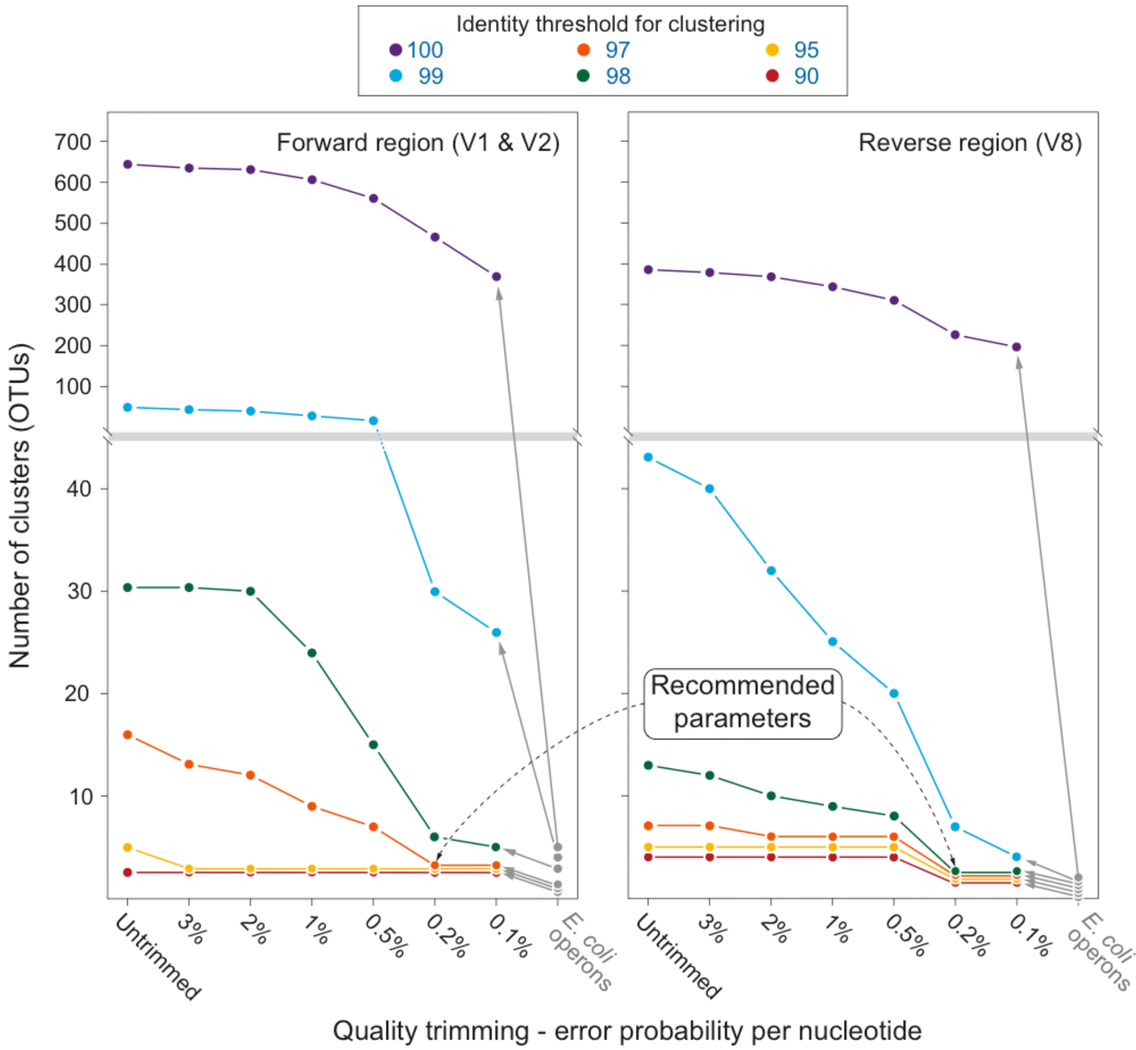
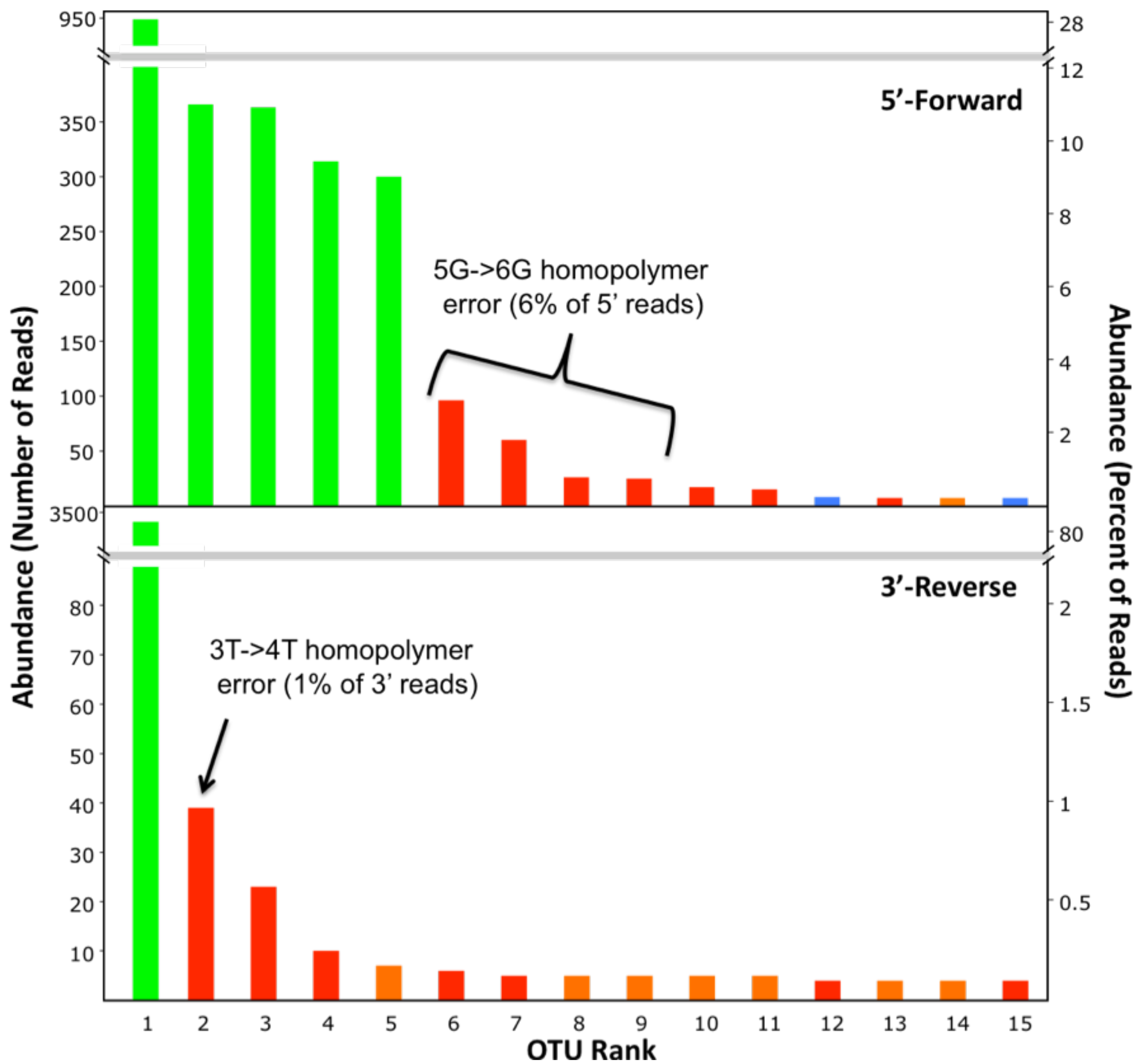


Fig. 1. Graphical representation of effect of quality filtering and clustering on diversity estimates of an *E. coli* “community” using pyrotags from a 5’-forward (1A) and 3’-reverse (1B) region of the 16S rRNA molecule.



285 **Fig. 2.** Rank abundance distribution and error types of the top 15 unique phylotypes (100% OTUs) from unfiltered 5'-forward and 3'-reverse 16S pyrotags. Colors denote errorless (green) reads, chimeras (blue) and reads with homopolymer length (red) or substitution (orange) errors.

290

Table 1. Effect of quality filtering and clustering on diversity estimates (OTU number), error rate and data loss of pyrotags amplified from two regions of *E. coli* MG1655 16S rRNA genes. Diversity estimates should be considered relative to the theoretical number of OTUs from *E. coli*.

Read filtering		Number of OTUs at % identity thresholds						% errorless	% reads
		100	99	98	97	95	90	reads	used
5' forward (V1&2)									
theoretical number		5	4	3	1	1	1		
no quality filtering		643	95	31	16	5	3	68.7	77.9
reads with Ns removed		600	85	29	14	4	3	69.8	76.7
quality score-based filtering (% per base error probability)	3	638	92	31	13	3	3	68.9	77.7
	2	632	90	30	14	3	3	69.0	77.6
	1	609	79	24	9	3	3	69.1	77.3
	0.5	562	66	15	7	3	3	70.7	75.3
	0.2	469	30	6	**3**	3	3	73.2	70.8
	0.1	372	26	5	3	3	3	77.8	57.8
3' reverse (V8)									
theoretical number		1	1	1	1	1	1		
no quality filtering		385	43	13	7	5	4	84.6	94.4
reads with Ns removed		361	40	12	6	4	3	85.3	93.6
quality score-based filtering (% per base error probability)	3	378	40	12	7	5	4	84.8	94.2
	2	368	32	10	6	5	4	85.1	93.8
	1	342	25	9	6	5	4	85.3	93.3
	0.5	310	20	8	6	5	4	87.5	89.5
	0.2	236	7	2	**2**	2	2	89.6	82.1
	0.1	196	4	2	2	2	2	90.7	70.6

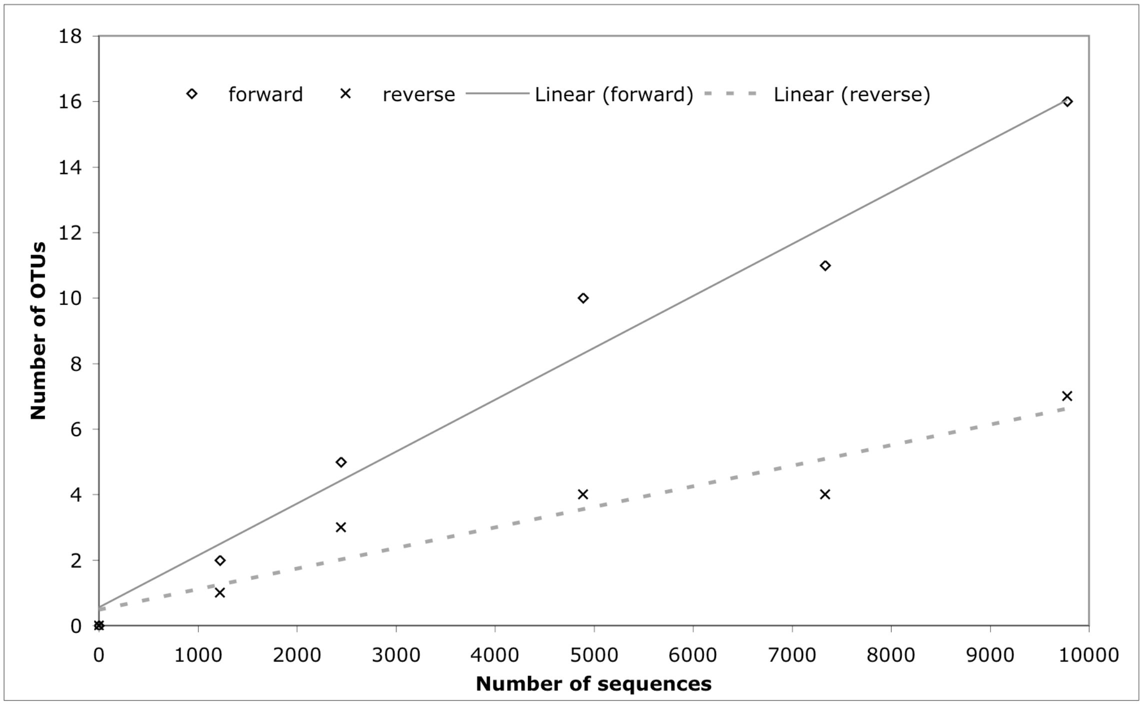


Fig. S1. Effect of pyrotag sample size on OTU number estimates from the 5'-forward and 3'-reverse regions of *E. coli* 16S rRNA genes.