Experimental factors affecting PCR-based estimates of microbial species richness and evenness

Anna Engelbrektson[1], Victor Kunin[1], Kelly C. Wrighton[2], Natasha Zvenigorodsky[1], Feng Chen[1], Howard Ochman[3], & Philip Hugenholtz[1]

[1]Department of Energy Joint Genome Institute, Walnut Creek, CA

[2]Department of Plant and Microbial Biology, University of California, Berkeley, CA

[3]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ

Subject Category: Microbial populations and community ecology

Abstract

Pyrosequencing of 16S rRNA gene amplicons for microbial community profiling can, for equivalent costs, yield greater than two orders of magnitude more sensitivity than traditional PCR-cloning and Sanger sequencing. With this increased sensitivity and the ability to analyze multiple samples in parallel, it has become possible to evaluate several technical aspects of PCR-based community structure profiling methods. We tested the effect of amplicon length and primer pair on estimates of species richness (number of species) and evenness (relative abundance of species) by assessing the potentially tractable microbial community residing in the termite hindgut. Two regions of the 16S rRNA gene were sequenced from one of two common priming sites, spanning the V1-V2 or V8 regions, using amplicons ranging in length from 352 to 1443 bp. Our results demonstrate that both amplicon length and primer pair markedly influence estimates of richness and evenness. However, estimates of species evenness are consistent among different primer pairs targeting the same region. These results highlight the importance of experimental methodology when comparing diversity estimates across communities.

Keywords: 16S rRNA/ Diversity estimates/ Microbial community/ Pyrosequencing/ Termite

Introduction

Next-generation sequencing technologies have generated renewed interest in culture-independent 16S rRNA-gene-based community profiling (Tringe and Hugenholtz, 2008). 16S amplicon pyrosequencing permits much deeper sampling of microbial communities by providing orders of magnitude more sequence information than the more traditional Sanger sequencing of PCR-clone libraries. Moreover, barcoded 16S amplicons from multiple samples can be analyzed in parallel and provide greater sensitivity than PCR clone libraries (Parameswaran *et al.*, 2007; Sogin *et al.*, 2006). This makes it possible to examine the effects of several variables on community composition estimates, such as biases due to DNA extraction or PCR conditions.

Here we investigate two technical issues: the effects of primer choice and amplicon length on assessments of bacterial species richness and evenness. The termite P3 hindgut lumen community of *Nasutitermes corniger* was chosen to assess these factors in that this community has been characterized extensively by PCR clone library analysis (~1700 near full length sequenced clones) and has potentially tractable diversity (Warnecke *et al.*, 2007). As part of this investigation we explored two hypotheses. First, we anticipated that shorter amplicons produce higher richness estimates. Given that amplicons can compete with primers for binding sites in the PCR reaction, shorter amplicons may accumulate and inhibit their own production in earlier cycles allowing rarer templates to amplify in later cycles, thereby increasing the apparent richness. Additionally, we hypothesize that primer choice will have a marked effect on species evenness due to variable priming specificities and annealing kinetics (Suzuki and Giovannoni, 1996).

Material and Methods

DNA extraction

To obtain termite hindgut community DNA, the gut tracts of 25 *Nasutitermes corniger* worker specimens were extracted from the exoskeleton using sterile forceps. A hemi-transverse incision of the P3 hindgut compartment was made with a needle, and 2 µL of 100 mM PBS were mixed with luminal contents squeezed out of the P3 compartment. The samples were pooled, maintained on ice, and DNA was isolated using aluminum ammonium sulfate added to cetyl trimethylammonium bromide (CTAB) nucleic acid extraction protocol followed by a polyethylene glycol (PEG) precipitation (Wrighton *et al.*, 2008).

PCR and 454 sequencing

Eleven amplicons ranging from 352 to 1443 nucleotides in length (Fig. 1) were produced using combinations of the following broadly conserved 16S primers (note 454 adaptor sequences and barcodes are not shown here): 27F (5'-agagtttgatcMtggctcag-3'), 357F (5'-ctcctacgggaggcagcag-3'), 530F (5'-gtgccagcMgccgcgg-3'), 803F (5'-attagataccctggtagtc-3'), 926F (5'-aaactY aaaKgaattgacgg-3'), 1114F (5'-gcaacgagcgcaaccc-3'), 342R (5'-ctgctgcSYcccgtag-3'), 519R (5'-gWattaccgcggcKgctg-3'), 787R (5'-ctaccagggtatctaat-3'), 907R (5'-ccgtcaattcMtttRagttt-3), 1100R (5'-gggttgcgctcgttg-3'), and 1392R (5'-acgggcggtgtgtRc-3').

To multiplex amplicons for inclusion on a single sequencing run, the common primer in each reaction (27F or 1392R) was barcoded on the 5' end with five unique bases between the 454 A-adaptor sequence (5'-gcctccctcgcgccatcag-3') and the conserved 16S rRNA primer sequence. Rules for barcoding were established in order to reduce the likelihood of ambiguities due to potential homopolymeric errors; (i) barcodes cannot start with the same nucleotide as the 454 adaptor ends, (ii) barcodes cannot end with the same nucleotide as the first nucleotide in the

16S primer, (iii) there can be no more than two successive occurrences of the same nucleotide within the barcode, and (iv) each barcode must differ from other barcodes by at least two bases. The other primer in each pair was not barcoded but did incorporate the 454 B-adaptor (5'-gccttgccagcccgctca-3') at its 5' end.

For each primer pair, PCR was performed in triplicate and pooled to minimize random PCR bias. Each 20 μL reaction consisted of 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 template DNA. Each reaction proceeded under the following conditions: 95ºC for 3 min; 25 cycles of 95ºC for 30 sec, 55ºC for 45 sec, and 72ºC for 90 sec; followed by a final extension at 72ºC for 10 min. Amplification products were purified on Qiagen MinElute PCR columns following the manufacturer's instructions and quantified with a Qubit fluorometer (Invitrogen). To obtain a similar number of reads from each sample, amplicons were mixed in equal concentrations. Emulsion PCR and sequencing were performed using a GS FLX emPCR amplicon kit according to the manufacturer's protocols.

Informatic Analysis

Pyrosequencing flowgrams were converted to sequence reads using the standard software provided by 454 Life Sciences. Reads were processed using the computational pipeline described in (Kunin *et al.*, 2009). Briefly, the reads were end-trimmed with LUCY (Li and Chou, 2004) using an accuracy threshold of 0.5% per base error probability, and then barcode and primer sequences were removed from the 5' end of the read. Reads lacking exact matches to a barcode and primer were discarded. All remaining reads were uniformly truncated to 220 nucleotides based on the length histogram of the quality-trimmed reads (not shown). Reads shorter than 220 nt were excluded from further analyses. Identical 220-nt reads were removed and unique

sequences compared by *blastn* using a word length of 25. The *blast* output was filtered to remove all pairwise matches with similarities <97% across the entire read length and clustered using the Markov Cluster algorithm using default parameters (van Dongen, 2000). 97% OTUs were classified taxonomically by *blastn* comparisons against the *greengenes* database (DeSantis *et al.*, 2006) using a word length of 25. Pass rates for each step of the processing pipeline were recorded.

To assess richness, rarefaction curves and bootstrapping were performed using an in-house script that plots randomly sampled clustered reads as a function of the number of 97% OTUs. To assess evenness, rank abundance curves were prepared using 97% OTUs with ≥0.5% relative abundance, averaged for each region. Simpson's measure of evenness ($E_{1/D;}$ $D=\sum(n/N)^2$, where n=number of organisms for a given species, and N=number of organisms for all species) was calculated for each amplicon using the statistical program R (Team, 2008). This metric is insensitive to the taxa richness and ranges from 0-1, with 0 representing complete dominance and 1 representing an evenly structured community. We statistically compared differences in evenness estimates between the primer regions using a two sample t-test (Minitab Inc, State College, PA). For the F_573 amplicon with anomalous OTU evenness (see Fig. S2), we estimated the presence of a mismatch in the 519R primer for each OTU based on the closest matched full-length *greengenes* sequence. Relative abundances of phyla were calculated using the *greengenes* classifications of the OTUs. ANOSIM in the statistical package Primer V (Plymouth Marine Laboratory, Plymouth UK), an analogue to the standard univariate 1-way ANOVA designed for ecological data, was performed on phylum-level pyrosequencing data to statistically assess assemblage differences between primer pairs (Clarke et al, 1993). For all

statistical hypothesis testing, either a two-sample t-test or ANOSIM, a significance level of 0.05 was used and probability (p-value) of observation is reported.

Data Submission

454 GS FLX flowgrams (sff files) were submitted to the Short Read Archive database at NCBI and have the accession SRA009438.

Results and discussion

To determine the effect of primer pair and amplicon length on 16S rRNA-based community composition estimates, we assessed operational taxonomic unit (OTU) richness (number of OTUs) and evenness (relative abundance of OTUs) for a range of amplicons obtained from a termite hindgut community using barcoded pyrosequencing (Parameswaran *et al.*, 2007; Sogin *et al.*, 2006). 16S rRNA genes were PCR-amplified using a combination of broad-specificity (domain or universal) primers with 454 FLX adaptor sequences. Eleven amplicons, ranging in size from 352 to 1443 bp, were prepared with primer sets that spanned either the V1 and V2 (sequenced forward from 27F) or the V8 regions (sequenced reverse from 1392R) of the 16S rRNA gene (Fig. 1). Technical replicates of four amplicons (F_394, F_573, R_352, and R_544; Fig. 1) were used to compare differences between amplicon datasets with the variation inherent in the method.

Amplicon pyrosequencing data processing and statistics

A total of 680,744 pyrosequence reads (termed "pyrotags") were produced from the 15 amplicons (Table S1). Quality-based trimming resulted in the loss of 18.6% of the data, and a further 17.6% were lost due to short sequence length or no identifiable barcode. In total, 36.2% of the reads were removed after these quality-filtering steps (Table S1); however, this loss was

not uniform since longer amplicons contributed disproportionately to those eliminated (Fig. S1). Amplicons greater than one kilobase in length had failure rates >90% and were not included in subsequent analyses. It is worth noting, however, that amplicons as long as 963 bp had pass rates greater than 50% (Table S1) despite manufacturers' recommendations to limit amplicons to less than 500 bp.

Sequences that passed quality filtering were trimmed to a uniform length of 220 bases to facilitate comparative analyses. Trimmed reads were grouped into clusters with a 97% identity threshold producing a total of 2,269 OTUs of which 1,617 and 652 are from the forward (V1-V2) and reverse (V8) regions respectively. The applied quality-filtering and clustering parameters were previously demonstrated to minimize the effect of pyrosequencing errors on microbial diversity estimates (Kunin *et al.*, 2009).

Species richness

The first hypothesis was that, for a given number of reads, shorter 16S rRNA gene amplicons yield greater species richness than longer amplicons. To estimate species richness, rarefaction curves were generated by randomly sampling reads and plotting the number of novel 97% OTUs against the number of reads sampled (Fig. 2). Most noticeably, forward amplicons all produced markedly (~3 fold) higher OTU richness estimates than did reverse amplicons (Fig. 2). This is due to the higher sequence variability in the V1-V2 region than the reverse V8 region as has recently been noted by Youssef *et al.* (2009). Therefore, OTU richness estimates provided by 16S pyrotags can vary according to the particular region surveyed, and absolute richness estimates based on different portions of the 16S rRNA gene should not be compared directly.

Within-region comparisons revealed that the shortest amplicons produced higher richness estimates than longer amplicons (Fig. 2) although this trend does not appear to hold beyond 400

bp fragments (e.g. F_963 appears to indicate higher richness than F_839). The apparently much lower richness estimate of F_573 compared to the other forward amplicons is likely due to a mis-priming effect biasing against many phylotypes (see below). It also should be noted that technical replicates of some of the shorter amplicons resulted in different richness estimates, suggesting that rare populations were not reproducibly sampled despite the relative simplicity of the community studied. A similar relationship between amplicon length and estimated richness was observed by Huber *et al*. (2009) using 16S rRNA gene clone library data from two hydrothermal vent fluid samples. In this case, 100-bp amplicons produced significantly higher estimates of richness than 400-bp or 1000-bp products.

Species evenness

Our second hypothesis posited that primer choice affects relative abundance of OTUs (i.e., species evenness). Whereas many primers designed to amplify 16S rRNA genes are broadly conserved, no primer pair is truly universal due to base pairing exceptions present in one or more lineages targeted by broadly conserved primers (Hugenholtz and Goebel, 2001). However, the extent of this problem on PCR-based community profiling has not been systematically addressed.

Rank abundance curves of the dominant OTUs (>0.5% relative abundance in at least one primer set) were prepared by averaging OTU abundances across amplicons (excluding F_573, see below). This was performed separately for each region since the forward and reverse data are not directly comparable. The relative abundance of the dominant reverse OTUs was greater than the dominant forward OTUs (Fig. 3) due to higher sequence conservation in the V8 region (Youssef *et al.*, 2009) resulting in larger clusters of reverse reads at the 97% identity threshold than forward reads. To corroborate our interpretation of the curves, we calculated Simpson's

inverse index of diversity ($E_{1/D}$) for each amplicon (Table S2). A two-sample t-test performed on $E_{1/D}$ values for each amplicon confirmed the 16S rRNA regions (V1-V2 vs V8) evaluated in this study resulted in statistically different estimates of evenness ($p<0.05$; $p=0.012$).

Remarkably, all of the primer pairs within either the forward or the reverse region (with the exception of F_573) produced very similar estimates of species evenness for the dominant OTUs, as evidenced by the low standard error for most OTUs (Fig. 3). The single outlier to this trend (*viz* amplicon F_573) produced markedly different OTU abundances (Fig. S2) attributable to a C:A mismatch between the 519R primer and 16S rRNA gene templates at *E. coli* position 534, three bases from the 5' end of the primer sequence. Such mismatches are customarily considered as having little or no impact on PCR because extension occurs from the 3' end (Bru *et al.*, 2008). However, the addition of the 18-bp 454 B-adaptor to the 5' end appears to have sufficiently destabilized the binding of this primer to mismatched templates thereby favoring the amplification of perfectly matched templates. This resulted in a consistent overrepresentation of perfectly matched (T:A) templates coupled with an underrepresentation of C:A mismatched templates (Fig. S2). Therefore, primer selection can significantly affect species evenness if base variations in templates are not accounted for by degeneracies in the primer sequence. However, when these variations are addressed, evenness of dominant OTUs is highly reproducible between different primer pairs targeting the same region.

To compare the phylogenetic diversity uncovered in pyrosequence data from each region to previous estimates of the community structure in the termite hindgut, we classified all OTUs by *blastn* against the *greengenes* database (DeSantis *et al.*, 2006) and then amalgamated the OTUs at phylum-level (Fig. 4). With the exception of F_573, estimates of the *Nasutitermes* hindgut community structure from the forward and reverse amplicons were not significantly

different ($p>0.05$; $p=0.10$ $R=0.556$) despite the difference in OTU granularity between regions (Fig. 3). The dominant phylum, the Spirochaetes, comprises 67 to 71% of the reads in each amplicon dataset followed by the Fibrobacteres (16 to 25%) and a handful of other phyla each representing >1% of reads, including Proteobacteria, Firmicutes, Bacteroidetes, Acidobacteria, and candidate phylum ZB3 (Fig. 4). These results are consistent with previous PCR-based (Warnecke *et al.*, 2007) and FISH-based (Hongoh *et al.*, 2006) profiles of *Nasutitermes* spp. hindgut communities. In contrast, spirochetes were significantly underrepresented and fibrobacters significantly overrepresented in the F_573 sample due to the aforementioned C:A mismatch in most spirochetes and T:A match in most fibrobacters.

Conclusion

This study tested the hypotheses that shorter pyrotag amplicons produce higher richness estimates and that primer choice affects species evenness. Our results show that the shortest amplicons tested (<400 bp) produce higher richness estimates than longer amplicons. However, regional variation in the 16S rRNA molecule has a much greater effect on apparent richness. Within a common region, primer choice had little effect on evenness of dominant OTUs (>0.5% abundance), provided that template mismatches are accommodated for by degeneracies in the primer. This surprising reproducibility may have been facilitated by the use of a common primer. However, pronounced differences in evenness were observed between the two regions of 16S rRNA tested due to differences in sequence conservation. Despite the observed inconsistencies in both richness and evenness estimates between variable regions, the inferred community structure at higher taxonomic ranks (phylum) was consistent between amplicons and regions and to previous estimates of community structure from the termite hindgut. We conclude that species (97% OTUs) evenness and richness should not be directly compared between different regions of

the 16S rRNA molecule. However species evenness estimated using different primer pairs targeting the same region may be reliably compared.

Acknowledgements

We thank Rudi Scheffrahn and Falk Warnecke for providing termites and Rebecca Daly for conversations and statistical analyses pertaining to microbial diversity. VK was supported in part by NSF grant OPP0632359 and KW by a Chang-Lin Tien Scholarship in Environmental Sciences and Biodiversity. The work was also supported by a grant from the Simon Family Fund and 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. Dedicated to the memory of my father, FWN Hugenholtz (1924-2009)- PH.

References

Bru D, Martin-Laurent F, Philippot L (2008). Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Appl Environ Microbiol* **74:** 1660-3.

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:** 5069-72.

Hongoh Y, Deevong P, Hattori S, Inoue T, Noda S, Noparatnaraporn N *et al* (2006). Phylogenetic diversity, localization, and cell morphologies of members of the candidate phylum TG3 and a subphylum in the phylum Fibrobacteres, recently discovered bacterial groups dominant in termite guts. *Appl Environ Microbiol* **72:** 6780-8.

Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB (2009). Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environ Microbiol* **11:** 1292-302.

Hugenholtz P, Goebel BM (2001). The polymerase chain reaction as a tool to investigate microbial diversity in environmental samples. In: Rochelle PA (ed). *Environmental Molecular Microbiology: Protocols and Applications*. Horizon Scientific Press.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2009). Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environmental Microbiology* **9999**.

Li S, Chou HH (2004). LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics* **20:** 2865-6.

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.

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 U S A* **103:** 12115-20.

Suzuki MT, Giovannoni SJ (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625-30.

Team RDC (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Tringe SG, Hugenholtz P (2008). A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11:** 442-6.

van Dongen SM. (2000). *Vol. PhD*. University of Utrecht: Utrecht.

Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT *et al* (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450:** 560-5.

Wrighton KC, Agbo P, Warnecke F, Weber KA, Brodie EL, DeSantis TZ *et al* (2008). A novel ecological role of the Firmicutes identified in thermophilic microbial fuel cells. *ISME J* **2:** 1146-56.

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:** 5227-36.

Figure Legends

Fig 1. Experimental design showing amplified regions of the 16S rRNA gene. Amplicon names to the left of the figure denote amplicon length including primers and the orientation of sequencing, forward (F) or reverse (R). Representations of amplicons show the region sequenced (blue) and forward (F) or reverse (R) primers (grey) use to produce the amplicon. Universal primers are presented in red type-face and domain-level primers are in black.

Fig 2. Rarefaction curves of the 97% OTUs for different length amplicons from forward (V1&V2) and reverse (V8) regions of the 16S rRNA molecule. Technical replicates are displayed as dashed lines, and colored hatching represent 95% confidence intervals.

Fig 3. Rank abundance curves of the top 97% OTUs in the forward V1-V2 (blue diamonds) amplicons and reverse V8 (red diamonds) amplicons. Standard errors are shown.

Fig 4. Relative abundance of bacterial phyla in the termite hindgut for each amplicon. Data from each technical replicate pair were averaged.

Fig S1. Pass rates of 454 reads for amplicons of different lengths at two points in the quality filtering process. Red diamonds represent the pass rate of amplicons after LUCY quality-trimming, and blue diamonds represent the pass rate for each amplicon length after barcode/primer and uniform length filtering.

Fig S2. Rank abundance curves of the top one hundred 97% OTUs in the forward region. Blue diamonds represent the average relative abundance of all forward (V1-V2) amplicons excluding F_573. Green and red diamonds represent the relative abundance of the F_573 OTUs, with green diamonds denoting sequences that likely match the primer sequence

(T:A) at *E. coli* position 534 (5’ GW**A**TTACCGCGGCKGCTG 3’) and red diamonds representing those that likely have a mismatch (C:A) at the same position.

Figure 1

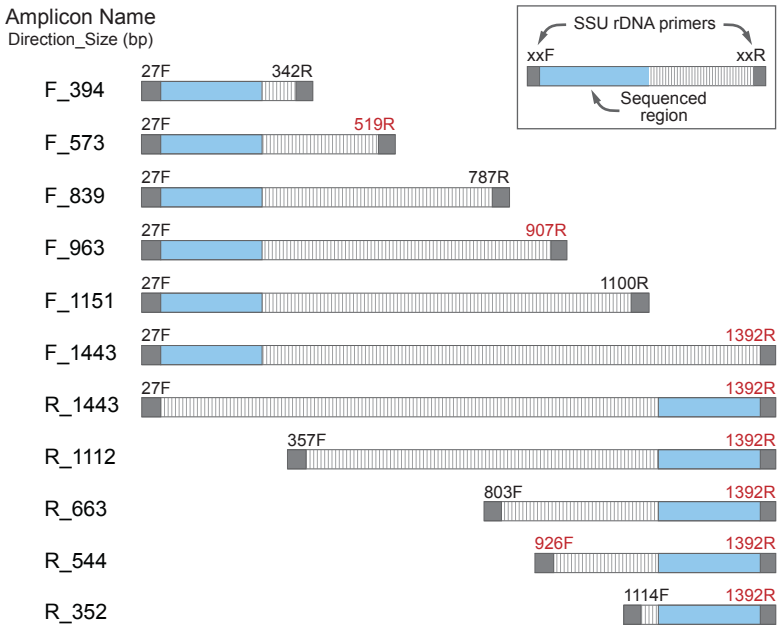

Amplicon Name
Direction_Size (bp)
SSU rDNA primers
xxF
xxR
Sequenced region
F_394
27F
342R
F_573
27F
519R
F_839
27F
787R
F_963
27F
907R
F_1151
27F
1100R
F_1443
27F
1392R
R_1443
27F
1392R
R_1112
357F
1392R
R_663
803F
1392R
R_544
926F
1392R
R_352
1114F
1392R

Figure 2

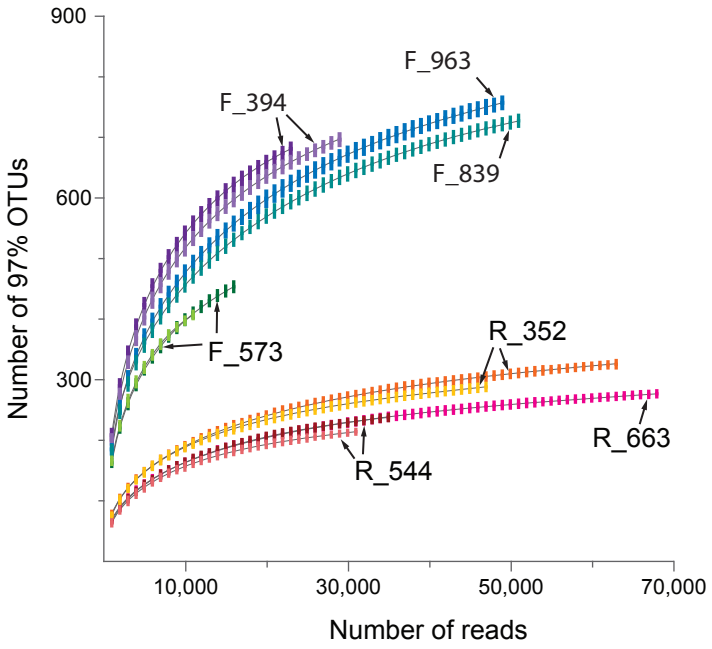

900
600
300
Number of 97% OTUs
F_963
F_394
F_839
F_573
R_352
R_663
R_544
10,000
30,000
50,000
70,000
Number of reads

Figure 3

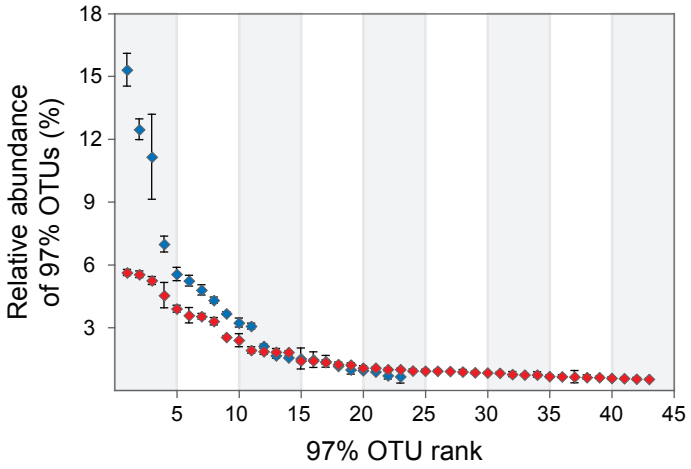
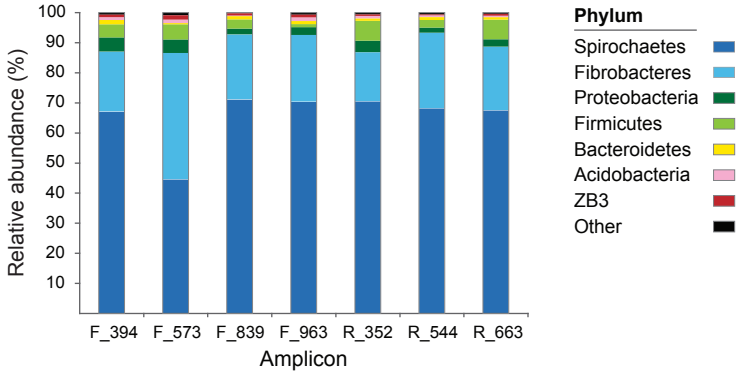

Relative abundance of 97% OTUs (%)
18
15
12
9
6
3
5
10
15
20
25
30
35
40
45
97% OTU rank

Figure 4


Relative abundance (%)
100
90
80
70
60
50
40
30
20
10
F_394
F_573
F_839
F_963
R_352
R_544
R_663
Amplicon
Phylum
Spirochaetes
Fibrobacteres
Proteobacteria
Firmicutes
Bacteroidetes
Acidobacteria
ZB3
Other