Applied and Environmental Microbiology

Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification

David Berry, Karim Ben Mahfoudh, Michael Wagner and Alexander Loy

Appl. Environ. Microbiol. **2011, 77(21):7846. DOI: 10.1128/AEM.05220-11.**
Published Ahead of Print 2 September 2011.

Applied and Environmental Microbiology, Nov. 2011, p. 7846–7849 Vol. 77, No. 21
0099-2240/11/$12.00 doi:10.1128/AEM.05220-11

Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification▽†

David Berry, Karim Ben Mahfoudh, Michael Wagner, and Alexander Loy*

Department of Microbial Ecology, Vienna Ecology Center, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria

Received 6 May 2011/Accepted 24 August 2011

"Barcode-tagged" PCR primers used for multiplex amplicon sequencing generate a thus-far-overlooked amplification bias that produces variable terminal restriction fragment length polymorphism (T-RFLP) and pyrosequencing data from the same environmental DNA template. We propose a simple two-step PCR approach that increases reproducibility and consistently recovers higher genetic diversity in pyrosequencing libraries.

Recent advances in DNA sequencing technologies have created opportunities for sequencing at an unprecedented depth and breadth (12) and multiplex sequencing has emerged as a popular strategy for parallel sequencing of many different samples (14). In multiplex sequencing, a unique sample-specific identifier, or "barcode" sequence, is added to the DNA that is to be sequenced. After sequencing, reads are sorted into sample libraries via detection of the appropriate barcode. Multiplexing in amplicon sequencing, which is widely performed for diversity surveys of 16S rRNA or functional genes, can be performed either by ligating barcodes and sequencing adapters to amplicons created with "conventional" PCR primers (primers that consist only of the template-specific sequence) (13), or more simply by using long oligonucleotides that, in addition to conventional PCR primers, already include 5′ tags with barcodes and sequencing adapters, thereby eliminating the ligation step (2, 8). The latter approach is referred to here as "barcoded primer" PCR (bcPCR).

The implicit assumption behind the bcPCR approach is that the adapter and barcode nucleotide sequence adjacent to the template-specific PCR primer does not interact with the template strand in such a way as to promote template sequence-dependent selective amplification. Pyrosequencing-based genetic diversity studies are known to be affected by a number of factors, including template sequence (1), amplicon size and target region, choice of primers (4), pyrosequencing errors (5, 10, 16), and OTU clustering procedure (9), and it was recently demonstrated that this widely used approach suffers from a relatively low technical reproducibility (20). In order to test specifically whether bcPCR affects surveys of genetic diversity, we designed barcoded primers comprised of the Titanium FLX sequencing adapters, randomly selected 8-nucleotide barcode sequences from a published and widely cited list (7) (Table S1), and primers targeting a fragment of the 16S rRNA gene of most bacteria and spanning regions V6 to V9, which is sufficient for accurate microbial community characterization (11) and captures genetic diversity similarly to full-length 16S rRNA (17) (for details, see the supplemental material). We amplified DNA isolated from the mouse gut lumen and analyzed the resulting 16S rRNA gene amplicons using terminal restriction fragment length polymorphism (T-RFLP) and 454 pyrosequencing.

Each primer variant out of 11 randomly selected barcoded primers was tested in triplicate using T-RFLP (for details, see the supplemental material). T-RFLP was also conducted for three replicate DNA extractions (using the same extraction protocol) from the same homogenized sample in order to compare barcode-induced variation to a known source of technical variation. T-RFLP profiles were significantly less reproducible for primers that had different barcodes than for replicates of the same barcoded primer ($P < 0.0001$) (Fig. 1A). The resulting average pairwise distance of profiles obtained with primers carrying different barcodes was even greater than that observed for amplification of multiple DNA extractions with a single barcoded primer ($P < 0.0001$), indicating that the variability associated with amplification using primers with different barcodes is greater than that observed with replicate DNA extractions (Fig. 1A). The detection of T-RF peaks, which is a presence/absence measurement, was not significantly different for the 1-step and 2-step bcPCR on average ($P = 0.31$), but the overall variation was higher among amplicons produced with different barcoded primers (F test, $P = 0.029$), which indicates that the barcode sequence did affect the detection of some peaks (Fig. 1B).

In order to reduce the variability associated with different bar-coded primers, we reasoned that the presence of the overhanging pyrosequencing adapter and barcode region should be minimized during amplification. We therefore implemented a 2-step PCR procedure in which conventional PCR primers amplify the template to the desired yield in the first step, and a dilution of the amplicons from this first step then serves as a template in a successive low-cycle-number amplification using the appropriate barcoded primers (see Fig. S1 in the supplemental material). This 2-step protocol is similar to "reconditioning PCR" and therefore may be ex-

* Corresponding author. Mailing address: Department of Microbial Ecology, Vienna Ecology Center, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria. Phone: 43 4277 54207. Fax: 43 4277 54389. E-mail: loy@microbial-ecology.net.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 2 September 2011.

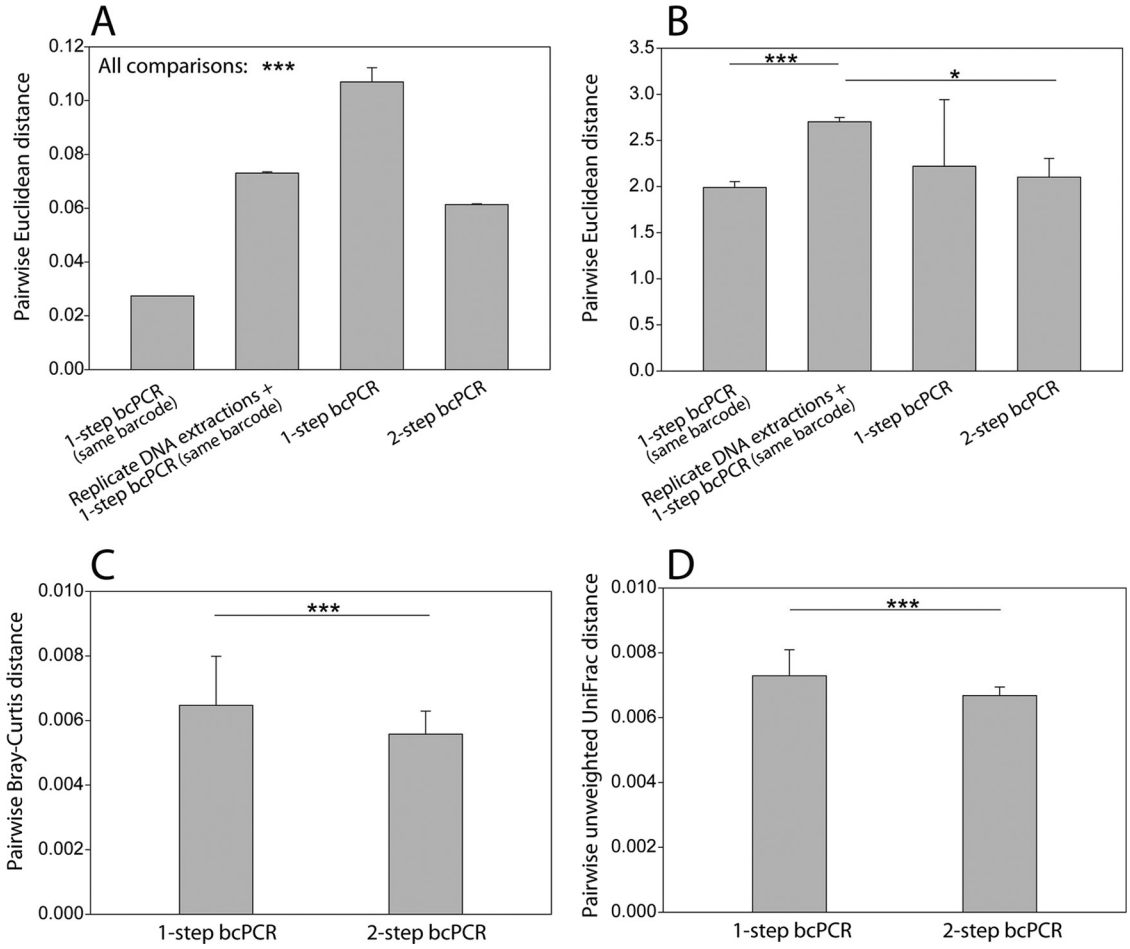


FIG. 1. Barcoded pyrosequencing primers affect reproducibility of community profiles obtained via T-RFLP (A and B) or 454 sequencing (C and D). All T-RFLP experiments were performed in triplicate. (A and B) Average pairwise Euclidean distances of T-RFLP profiles are shown, as measured by T-RF relative abundances (A) and T-RF presence/absence (B). From left to right, the bars show comparisons made using T-RFLP replicates obtained from application of a single barcoded primer for bcPCR using DNA from a single extraction, T-RFLP replicates obtained from application of a single barcoded primer for bcPCR using DNA extracted separately for each replicate from the same homogenized gut sample, and T-RFLP profiles obtained after amplification of DNA from a single extraction using a mixture of 11 randomly chosen barcoded PCR primers with either 1-step or 2-step bcPCR. (C and D) Average pairwise community similarities from 454 sequencing libraries prepared from the same DNA extraction (also DNA from the mouse gut lumen, but different extraction than that used for panels A and B) using 16 barcoded primers with either 1-step or 2-step bcPCR are compared. Bray-Curtis (C) and unweighted UniFrac (D) distances are shown. Error bars indicate standard deviations, and asterisks indicate statistical significance at P values of <0.05 (*) and <0.001 (***).

pected to have the additional benefit of reducing heteroduplex formation in mixed-template reactions (19), although in the present study we did not observe a significant effect of PCR procedure on the percentage of 454 pyrosequencing reads (see below) detected as chimeras by Chimera Slayer (6) ($9.0 \pm 2.3\%$, $n = 22$, $P = 0.25$). This protocol, which we refer to as "2-step bcPCR" to distinguish it from standard "1-step" bcPCR, produces barcoded amplicons that can be directly used for pyrosequencing.

To test this approach, we performed 20 cycles of amplification with conventional PCR primers and then used 1 μl of the PCR product of the first reaction (1:50 dilution) as the template for a 5-cycle amplification with barcoded primers. Compared to the 1-step bcPCR, the 2-step bcPCR protocol indeed significantly improved the reproducibility of T-RFLP profiles obtained after use of the same 11 different barcoded primers with the same DNA extract (t test, $P < 0.0001$). The T-RFLP profiles were also more similar to each other than profiles obtained from 1-step bcPCR amplification with a single barcoded primer were to each other using DNA from replicate extractions of the same homogenized sample ($P < 0.0001$) (Fig. 1A). However, the profiles from the 2-step bcPCR were still slightly less reproducible than those obtained with 1-step bcPCR using a single barcoded primer. The reason for this remaining minor bias introduced by barcoded primers even in the 2-step bcPCR is unknown, but it is unlikely to be connected with interactions between the barcode and the template. This is because the first step of amplification produces amplicons removed from their genomic context, and therefore in the second step of amplification, a template with neighboring sequence regions should no longer be present at relevant concentrations. Amplification using barcoded primers in both steps of the 2-step protocol confirmed that the presence of the barcoded primer was responsible for the reduced reproducibil-

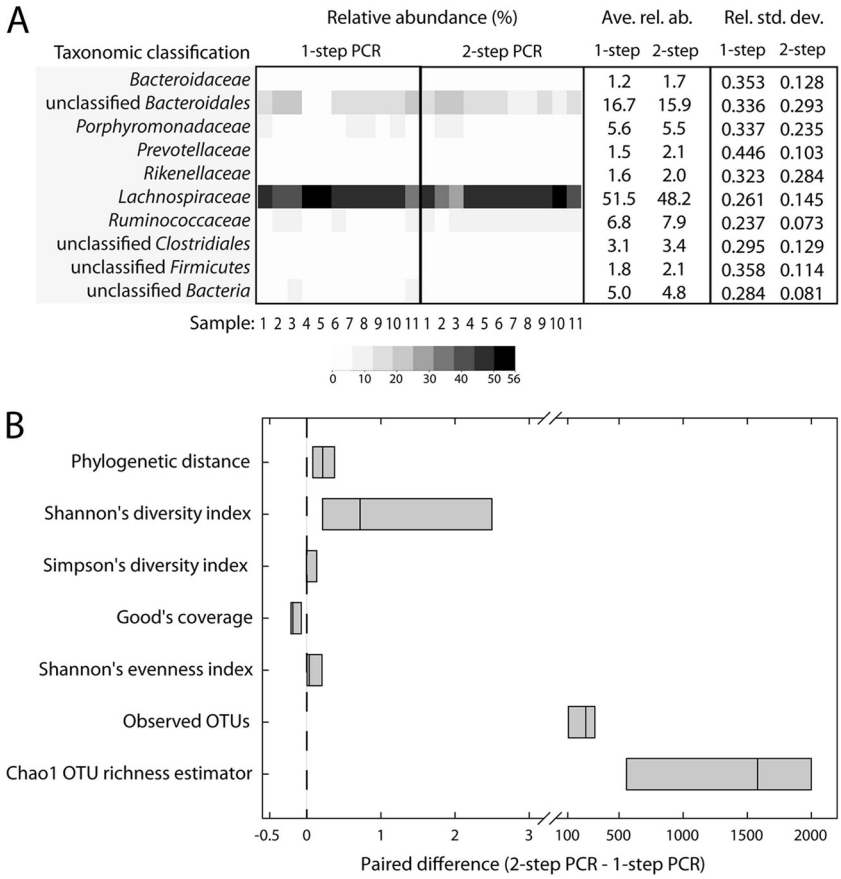


FIG. 2. The bcPCR method affects alpha diversity and reproducibility of taxonomic classification. (A) Comparison of 11 randomly selected barcoded primers (6 used for both bcPCR methods). Relative abundance for each taxon (family level) present on average at ≥1% is plotted on a heatmap for each barcode used. Average relative abundance and relative standard deviation are listed for each method and taxon. (B) A box plot of the paired difference for several alpha diversity metrics for operational taxonomic units (OTUs) is shown for 1-step and 2-step bcPCR using an identical set of 6 randomly selected barcoded primers (for details about metrics, see the supplemental material). The dashed line indicates a difference of zero.

ity of the 1-step bcPCR T-RFLP profiles (see Fig. S2 in the supplemental material), rather than the lack of a reconditioning step (19).

To test whether these results could be reproduced with sequencing data, pyrosequencing with 16 different barcodes using either 1-step or 2-step bcPCR was performed. Of the 16 barcodes, 6 were tested with both methods in order to make paired comparisons and the other 10 were tested with one of the PCR methods (5 for each method). The pyrosequencing data confirmed the T-RFLP result that 2-step bcPCR improves reproducibility, as measured by community similarity assessments with Bray-Curtis distance as well as unweighted UniFrac distance (Fig. 1C and D).

A widely used approach for 16S rRNA gene surveys is to classify sequences as belonging to specific taxa based on reference databases and compare their relative abundances (3, 18). The average relative representations of abundant taxa (>1% on average, classified at the family level) in the 1-step and 2-step bcPCR pyrosequence data sets were similar (Fig. 2A). However, we found that 2-step bcPCR reduced the relative standard deviation of relative abundance data for abundant families (Fig. 2A). Comparison of 6 barcoded primers evaluated using both methods revealed that 1-step bcPCR yielded reduced species richness, evenness, and phylogenetic distance (UniFrac tree branch length) (Fig. 2B), indicating that 2-step bcPCR recovers some sequence diversity missed by 1-step bcPCR. The extra diversity recovered by 2-step bcPCR shared similarity with high-quality 16S rRNA sequences in the SILVA database (SSU r106 Ref) (15) (mean sequence similarity, 91%) (see Fig. S3 in the supplemental material) and included two reads with 100% similarity to sequences in the database that had been recovered from rat feces, which indicates that the extra diversity is real and not a methodological artifact.

We explored whether any of the variation observed with different barcodes could be explained by known or predictable characteristics of the different barcoded oligonucleotides, but community structure was not determined by *in silico* folding stability, homodimer or heterodimer formation potential, or the identity of the nucleotide base on the 3′ end of the barcode (the base proximal to the template-specific PCR primer sequence) (perMANOVA, $P > 0.05$), and GC content was identical for all barcodes. This leads us to conclude that the bcPCR bias cannot be predicted by *in silico* secondary structure eval-

uation of the primer but is likely driven by selective or stochastic amplification caused by currently unknown template and barcoded primer interactions. While the present study evaluated the effect of varying the barcode region, the sequencing adapters would also be expected to contribute to selective amplification in bcPCR. This raises a possible concern for comparability of studies across different sequencing platforms as well as sequencing chemistries that use different adapters on the same platform.

The T-RFLP and pyrosequencing data clearly demonstrate that barcoded primers introduce biases in PCR that translate into less reproducible data sets. We have devised and evaluated a modified 2-step amplification procedure that improves this issue and outperforms the standard protocol. This modification can be easily incorporated into existing protocols and should be a valuable contribution to the production of high-quality multiplex amplicon libraries for high-throughput sequencing.

We thank Sebastian Lücker for design of bacterial primers, Holger Daims for helpful discussions, Christian Baranyi for technical assistance, and the Norwegian High-Throughput Sequencing Centre for pyrosequencing.

This work was financially supported by the Austrian Science Fund (P20185-B17 to A.L.) and the Austrian Federal Ministry of Science and Research (GEN-AU III InflammoBiota to D.B., M.W., and A.L.).

REFERENCES

1. **Amend, A. S., K. A. Seifert, and T. D. Bruns.** 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Mol. Ecol. **19:**5555–5565.
2. **Binladen, J., et al.** 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One **2:**e197.
3. **Caporaso, J. G., et al.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods **7:**335–336.
4. **Engelbrektson, A., et al.** 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J. **4:**642–647.
5. **Gomez-Alvarez, V., T. K. Teal, and T. M. Schmidt.** 2009. Systematic artifacts in metagenomes from complex microbial communities. ISME J. **3:**1314–1317.
6. **Haas, B. J., et al.** 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. **21:**494–504.
7. **Hamady, M., J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight.** 2008. Error-correcting Barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods **5:**235–237.
8. **Hoffmann, C., et al.** 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res. **35:**e91.
9. **Huse, S. M., D. M. Welch, H. G. Morrison, and M. L. Sogin.** 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. **12:**1889–1898.
10. **Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118–123.
11. **Liu, Z., C. Lozupone, M. Hamady, F. D. Bushman, and R. Knight.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. **35:**e120.
12. **Margulies, M., et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
13. **Meyer, M., U. Stenzel, and M. Hofreiter.** 2008. Parallel tagged sequencing on the 454 platform. Nat. Protoc. **3:**267–278.
14. **Parameswaran, P., et al.** 2007. A pyrosequencing-tailored nucleotide Barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res. **35:**e130.
15. **Pruesse, E., et al.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. **35:**7188.
16. **Quince, C., et al.** 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods **6:**639–641.
17. **Schloss, P. D.** 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput. Biol. **6:**e1000844.
18. **Schloss, P. D., et al.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75:**7537–7541.
19. **Thompson, J. R., L. A. Marcelino, and M. F. Polz.** 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR." Nucleic Acids Res. **30:**2083–2088.
20. **Zhou, J., et al.** 2011. Reproducibility and quantitation of amplicon sequencing-based detection. ISME J. **5:**1303–1313.

AEM
Journals.ASM.org

ERRATUM

Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification

David Berry, Karim Ben Mahfoudh, Michael Wagner, and Alexander Loy

Department of Microbial Ecology, Vienna Ecology Center, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria

Volume 77, no. 21, p. 7846–7849, 2011. Supplemental Materials and Methods: "5′-CCACTCAAAKGAATWGACGG-3′" should read "5′-ACTCAAAKGAATWGACGG-3′." Revised supplemental material is posted at http://aem.asm.org/content/suppl/2011/10/18/77.21.7846.DC1.

doi:10.1128/AEM.07448-11

SUPPLEMENTARY MATERIAL FOR

Barcoded primers used in multiplex amplicon pyrosequencing bias amplification

David Berry, Karim Ben Mahfoudh, Michael Wagner, and Alexander Loy

Samples and DNA extraction
Gut lumen contents of healthy C57BL/6N mice from a previous study were used (Berry et al. unpublished data). DNA was extracted using a standard phenol-chloroform bead-beating procedure .

PCR and sequencing
PCR primers targeting the 16S rRNA gene of most bacteria (909F: 5'-ACTCAAAKGAATWGACGG-3', 1492R: 5'-NTACCTTGTTACGACT-3') were used for amplification. For bcPCR, long oligonucleotides were used consisting of the gene-specific PCR primer sequences tagged with the sequencing adapters for GS FLX Titanium chemistry, which were designed as recommended by the manufacturer (454 Life Sciences) to include (5'-3'): Titanium adapter, 8 base barcode (only on the reverse primer), a linker sequence ("CC" for forward primer, "TA" for reverse primer), and the gene-specific PCR primer. The reverse primer included an 8 bp barcode identifier (Table S1) that was randomly selected from a subset of a published list of barcode sequences with identical GC content . *In silico* predictions of barcoded primer folding stability, homo- or heterodimer formation potential were made with UnaFold .

All PCRs were conducted in 50 µl volume using 1.25 U Taq DNA polymerase (Fermentas), 1 mM primers, 0.16 mM dNTP mix, 2 mM $MgCl_2$, and 0.1 µg bovine serum albumin. For one-step bcPCR, the template was amplified using barcoded PCR primers for 20 cycles. For two-step bcPCR, the template was amplified using non-barcoded PCR primers for 20 cycles, followed by a 1:50 dilution of the PCR product and 5 additional cycles of amplification with barcoded PCR primers. All PCR reactions were performed in triplicate and then pooled. The thermal program consisted of an initial 95 °C denaturation step for 4 min, a cycling program of 95 °C for 30 s, 52 °C for 30 s, and 72 °C for 60 s, and a final elongation step at 72 °C for 7 min. For sequencing, PCR amplicons were purified using Agencourt AMPure beads (Beckman Coulter Genomics) and quantified with a fluorescent stain-based kit (Quant-iT PicoGreen, Invitrogen). Amplicons were sequenced from the reverse side. Sequencing was performed on a GS FLX instrument using Titanium chemistry (454 Life Sciences) at the Norwegian High-Throughput Sequencing Centre (NSC).

Sequencing reads were filtered using LUCY (27.5 average and end PHRED score, minimum length =200 bp) , yielding an average of 6,600 reads per library. Operational taxonomic units (OTUs) were created by clustering reads at 97% identity with UCLUST and representative sequences were aligned with mothur using default settings . Chimera detection was performed using Chimera Slayer (mothur settings: minsnp=50, iters=1000, minbs=99) . Alpha and beta diversity metrics were calculated with QIIME using re-sampling and comparing samples at less

than the size of the smallest library (1360 reads) , and taxonomic assignments were made using the RDP naive Bayesian classifier . In addition to standard alpha diversity metrics, the phylogenetic distance, which is the total branch length of a phylogenetic tree of representative sequences for all OTUs in a sample , was calculated. Phylogenetic trees were produced using FastTree . Multivariate analysis of variance was performed with the vegan package in R .

Sequence reads are available as FASTA format together with associated metadata in a compressed file (.zip) included for download in the supplementary material.

Terminal-Restriction Fragment Length Polymorphism (T-RFLP)

PCR was conducted exactly as described above except that the forward primer was FAM-labeled (6-carboxyfluorescein). PCR products were purified according to manufacturer's instructions (Qiagen QIAquick PCR Purification Kit), 100 ng were digested with 5 U AluI restriction enzyme (Fermentas) for 3 h at 37 °C in a 20 µl reaction volume, and the enzyme was inactivated by heating to 65 °C for 20 min. Digested DNA was desalted (Sephadex GS50 SS, Sigma-Aldrich) and 4 µl was mixed with 10 µl formamide (Hi-Di formamide, Applied Biosystems) and 0.25 µl of MapMarker 1000 (Rox-labeled, Bio-Ventures, Inc.). The sample was denatured (3 min, 96°C), immediately cooled on ice, and run on an ABI 3130 XL Genetic Analyzer (Applied Biosystems) using the ABI program Fragment Analysis 50POP7. T-RF peaks were called using default parameters in the Peak Scanner software (version 1.0, Applied Biosystems) and peaks were processed and binned using the RawGeno package in R . Pair-wise Euclidean distances of peak relative abundance and presence/absence was calculated in R.

Table S1. Barcode sequences used in this study. Amplicons were sequenced from the reverse primer, and therefore barcoded primers were composed of (5'-3'): Titanium B adapter (CTATGCGCCTTGCCAGCCCGCTCAG), the 8 base barcode, a linker sequence (TA), and the reverse PCR primer (NTACCTTGTTACGACT). The forward primer was used with sequencing adapter and forward PCR primer (909F), but with no barcode (5'-CTATGCGCCTTGCCAGCCCGCTCAGCCACTCAAAKGAATWGACGG-3').

Barcode ID	Barcode (5' -3')	Method used
26	AACGTTGC	T-RFLP, Sequencing (1- and 2-step bcPCR)
86	ACAGCTGT	T-RFLP
149	ACGTCTAG	Sequencing (1- step bcPCR)
150	ACGTCTTC	Sequencing (1- step bcPCR)
155	ACGTTCCT	Sequencing (1- step bcPCR)
168	ACTCGAGT	Sequencing (1- step bcPCR)
198	AGACCTGT	Sequencing (1- step bcPCR)
332	ATCCATGG	T-RFLP
336	ATCCGGTA	T-RFLP
407	CAAGCAAG	Sequencing (2- step bcPCR)
409	CAAGCTAC	Sequencing (2- step bcPCR)
410	CAAGCTTG	Sequencing (2- step bcPCR)
418	CAAGTGGA	Sequencing (2- step bcPCR)
647	CTACAGCA	Sequencing (2- step bcPCR)
684	CTCACTGT	Sequencing (1- and 2-step bcPCR)
701	CTCTGACT	T-RFLP, Sequencing (1- and 2-step bcPCR)
709	CTGAACAC	Sequencing (1- and 2-step bcPCR)
755	CTTCTGCA	Sequencing (1- and 2-step bcPCR)
756	CTTCTGGT	Sequencing (1- and 2-step bcPCR)
887	GATGAGCA	T-RFLP
1023	GGTTCCAA	T-RFLP
1189	TACGTTCC	T-RFLP
1323	TCTCACTG	T-RFLP
1385	TGAGTGAG	T-RFLP
1464	TGTCTCTG	T-RFLP

Figure S1. Work flow and differences between 1-step (left column) and 2-step bcPCR (right column). Using the same amount of template DNA for the initial PCR and assuming 100% duplication of gene copy numbers after each cycle, barcoded amplicons obtained by 1-step and 2-step bcPCR have almost the same copy number (same order of magnitude) of the target gene. Note that 20 cycles and triplicate PCRs per template were performed in this study.

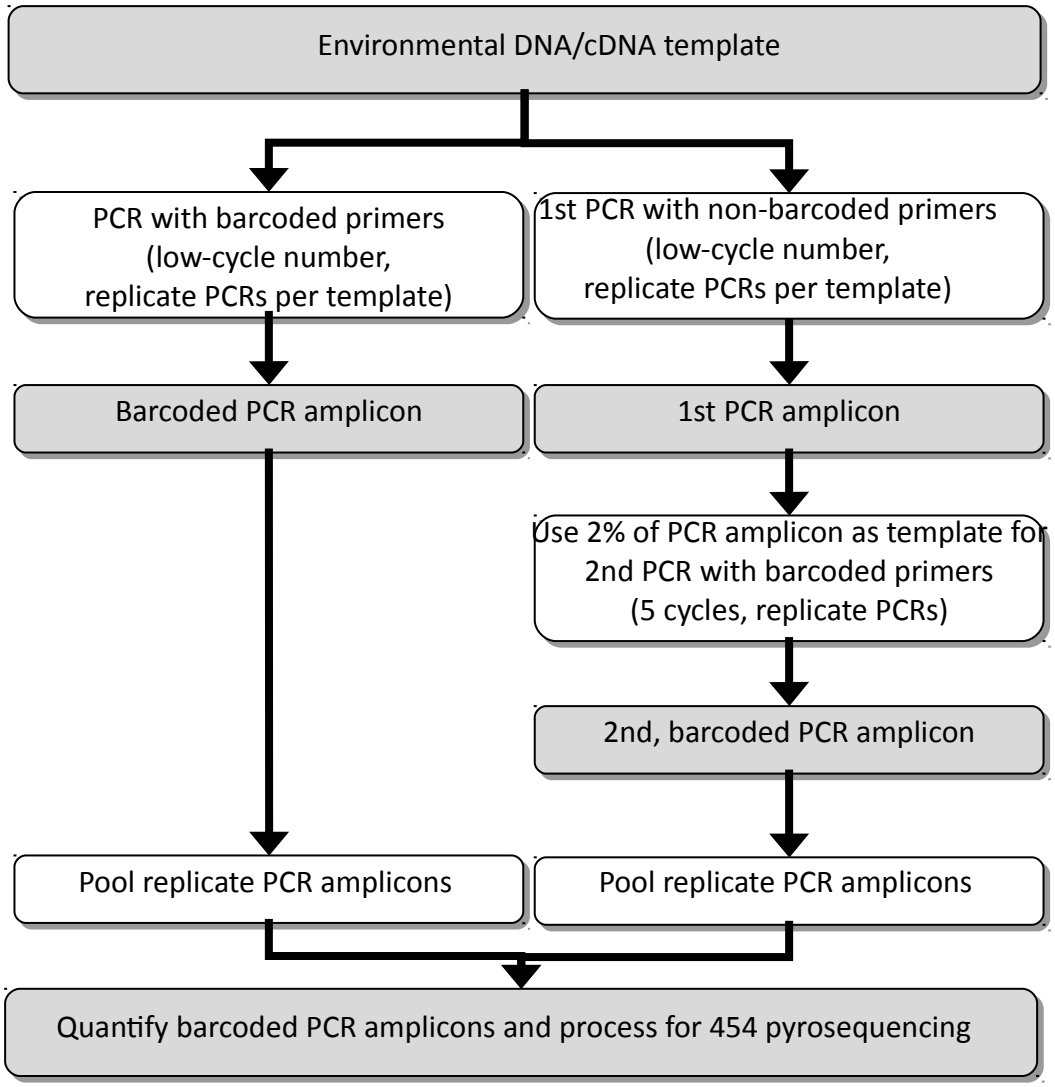


Figure S2. Performing a 2-step "reconditioning PCR" with barcoded pyrosequencing primers does not increase reproducibility of community profiles obtained via T-RFLP. Average pairwise Euclidean distance of T-RFLP profiles is shown, as measured by T-RF relative abundances. Comparisons were made using: T-RFLP profiles obtained after amplification of DNA from a single extraction using either 1-step or 2-step bcPCR, or using the 2-step protocol, but using barcoded primers in the first and second step (reconditioning bcPCR). All T-RFLP experiments were performed using 11 randomly chosen barcoded PCR primers. Error bars indicate standard deviation.

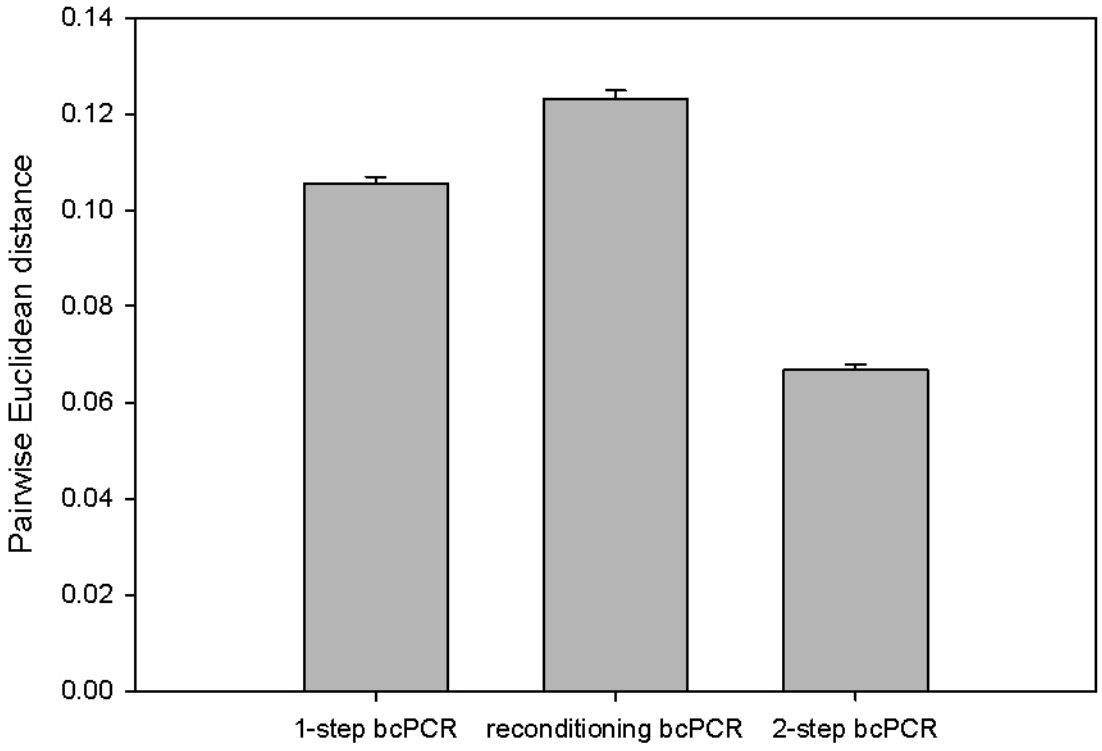
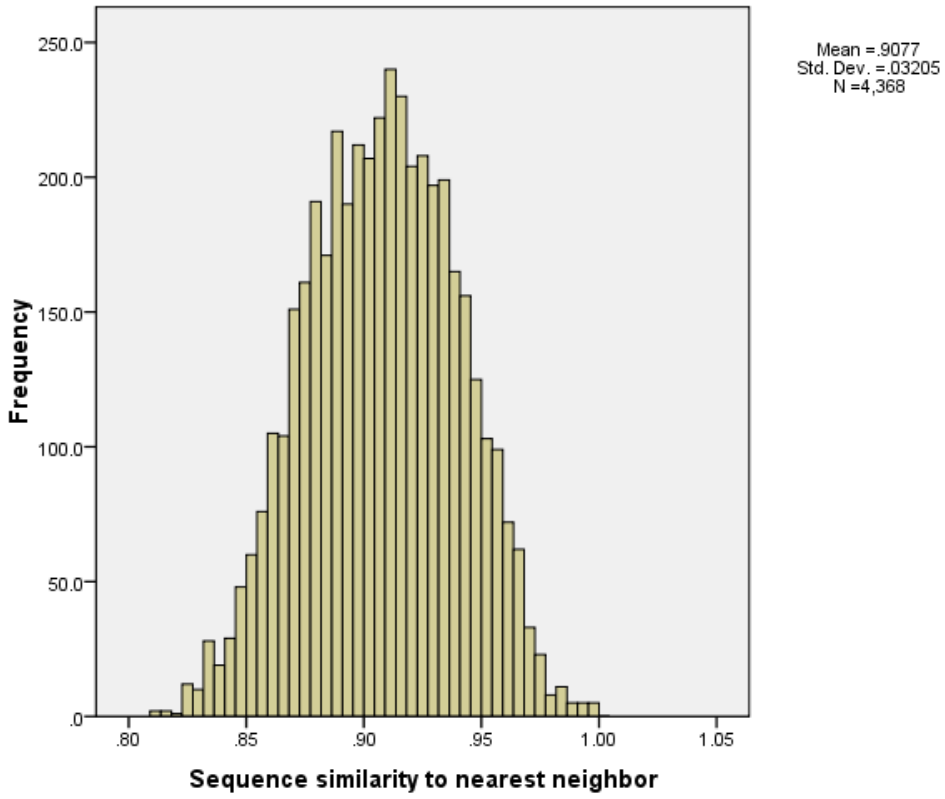


Figure S3. Histogram of sequence similarity of representative sequences of OTUs recovered in 2-step, but not 1-step, bcPCR sequencing libraries to high quality 16S rRNA sequences in the SILVA database (SSU r106 Ref) . Reads were aligned and nearest neighbors were identified using the SINA web aligner (http://www.arb-silva.de/aligner/).



References

Arrigo N, Tuszynski J, Ehrich D, Gerdes T, Alvarez N (2009). Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. *BMC Bioinformatics* **10:** 1-14.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335-336.

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

Griffiths RI, Whiteley AS, O'Donnell AG, Bailey MJ (2000). Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl Environ Microbiol* **66:** 5488-5491.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G *et al* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21:** 494-504.

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.

Lozupone C, Knight R (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71:** 8228-8235.

Markham NR, Zuker M (2008). UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Bio* **453:** 3-31.

Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL *et al* (2010). Vegan: community ecology package, R package version 1.17-4 edn.

Price MN, Dehal PS, Arkin AP (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26:** 1641-1650.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35:** 7188.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al* (2009). Introducing mothur: open-Source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537-7541.

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