Practical innovations for high-throughput amplicon sequencing

Derek S Lundberg[1,2,9], Scott Yourstone[1,3,9], Piotr Mieczkowski[4–6], Corbin D Jones[1–3,5,6] & Jeffery L Dangl[1,2,6–8]

We describe improvements for sequencing 16S ribosomal RNA (rRNA) amplicons, a cornerstone technique in metagenomics. Through unique tagging of template molecules before PCR, amplicon sequences can be mapped to their original templates to correct amplification bias and sequencing error with software we provide. PCR clamps block amplification of contaminating sequences from a eukaryotic host, thereby substantially enriching microbial sequences without introducing bias.

Microbes profoundly affect biological processes across Earth's ecological niches and are frequently identified through culture-independent methods using DNA purified directly from environmental samples[1]. Common PCR-based approaches target highly conserved rRNA genes, such as those encoding the 16S/18S and 28S subunits or the internal transcribed spacer (ITS) between them. These ubiquitous genes have diverged enough that polymorphisms across their 'hypervariable regions' (**Supplementary Fig. 1**) allow taxonomic classification. Amplicon sequencing is an important and widely used tool for inferring the presence of taxonomic groups in microbial communities, but poor estimates result from sequencing errors and biases introduced during amplification. Inefficiencies also result from the amplification of nontarget DNA. Here we describe methods that make rRNA amplicon sequencing more accurate and cost-effective.

Accurate base-calling on Illumina platforms requires sequence diversity at each nucleotide position[2]. Because amplicon libraries often lack diversity at specific positions owing to sequence conservation, it is common to spike sequencing runs with sheared genomic DNA from the virus phiX174. We created sequence diversity in 16S amplicons using a mix of primers that have frameshifting nucleotides (**Supplementary Figs. 2** and **3**). Despite recent upgrades to Illumina's base-calling procedure, this strategy remains useful for maximizing data yield as it devotes the entire sequencing effort to the amplicon of interest (**Supplementary Figs. 4** and **5**).

PCR and sequencing introduce sequence errors and sampling bias[3]. We adapted and validated a modified protocol that uniquely tags each template molecule with random nucleotides before PCR[4–7] (**Supplementary Figs. 2** and **6a**,**b**). Provided that there are enough random nucleotides, amplicons sharing the same tag are overwhelmingly likely to have originated from the same template molecule (the 'birthday paradox'[8]; **Supplementary Fig. 7**). Thus, by generating consensus sequences from each group of sequences sharing a molecule tag (MT), we can correct errors and infer the amplicon's probable template sequence (**Supplementary Fig. 6f**–**h**).

We verified that consensus sequences (ConSeqs) correct errors by amplifying a clonal plasmid-borne 16S template (**Fig. 1** and **Supplementary Table 1**). A dilution series ensured a variety of coverage depths for each MT (**Fig. 1a**). We found that a sample of 15,000 ConSeqs had fivefold lower mean error than a sample of 15,000 untreated (nonconsensus) 16S sequences (**Fig. 1b** and Online Methods).

We observed unexpectedly high numbers of singletons in the MT depth distributions for samples prepared from diluted templates, suggesting that some singletons arise from MT mutations in lower-quality reads. Consistent with this, the error rate among 15,000 singletons was more than twice that for untreated sequences. We also observed a lower error rate among the 4,777 available 'perfect ConSeqs' constructed from three or more reads with identical sequence sharing an MT, compared with all ConSeqs. Interestingly, this rate was not 0 because either all sequences in these perfect ConSeqs carried the same error, the template plasmid had some level of polymorphism that was accurately captured or a combination of these.

Operational taxonomic unit (OTU) clustering is a common approach both to corral noisy 16S sequence data into groups approximating microbial species and to reduce computational complexity[3]. Using data from the clonal 16S template, we clustered either 30,000 untreated sequences or 30,000 ConSeqs into OTUs using both 97% and 99% identity thresholds. ConSeqs clustered at 97% formed two OTUs, with the second OTU containing only six sequences (**Fig. 1c**). Untreated sequences at 97%, on the other hand, produced 66 OTUs, two of which were sufficient to capture 95% of the data. ConSeqs clustered at 99% formed 42 OTUs, and the first two OTUs contained 95% of the data, whereas untreated sequences produced 683 OTUs and required 66 OTUs to capture

---

[1]Department of Biology, University of North Carolina, Chapel Hill, North Carolina, USA. [2]Curriculum in Genetics and Molecular Biology, University of North Carolina, Chapel Hill, North Carolina, USA. [3]Program in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, North Carolina, USA. [4]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [5]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, USA. [6]Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, North Carolina, USA. [7]Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, North Carolina, USA. [8]Howard Hughes Medical Institute, University of North Carolina, Chapel Hill, North Carolina, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to J.L.D. (dangl@email.unc.edu).

RECEIVED 23 APRIL; ACCEPTED 31 JULY; PUBLISHED ONLINE 1 SEPTEMBER 2013; DOI:10.1038/NMETH.2634

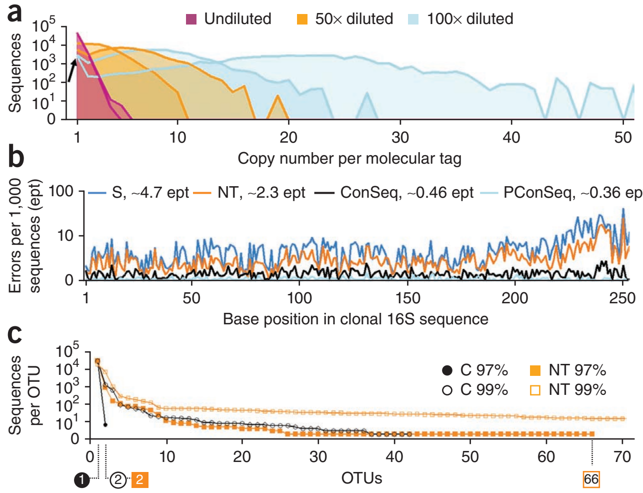

Figure 1 | Molecular tagging reduces sequence error for a clonal template. (**a**) Diluting template increases the coverage within each MT. Shown are two replicates each (overlaid in the same color) of undiluted, 50× diluted and 100× diluted clonal 16S template. All six samples were rarefied to 40,000 sequences, and the number of sequences collapsed into each MT was graphed as a density distribution for each sample. We noted more singleton MTs than expected by a unimodal Poisson distribution for the diluted samples (arrow). (**b**) Per-base error rates per 1,000 sequences were measured in pooled data from the 50× and 100× diluted template samples. We compared no-MT sequences (NT); ConSeqs from two or more sequences with identical MTs (ConSeq); perfect ConSeqs, for which all sequences in the alignments of three or more sequences were identical (PConSeq); and singleton MTs (S). Mean error per thousand (ept) for each MT treatment is shown in the color key. (**c**) 30,000 ConSeqs (C) or untreated sequences (NT) were clustered into OTUs at both 97% and 99% identity thresholds. Rank-abundance curves demonstrate the number of sequences per OTU. The position of the colored boxes and circles below the *x* axis, and the numbers in each, show the number of ranked OTUs necessary to represent 95% of the sequences for each condition.

95% of the data. Thus, ConSeqs were more homogenous than untreated sequences and tolerated stricter OTU definitions, a result suggesting that ConSeqs can be used to provide a more accurate picture of true microbial alpha diversity[3].

We applied our approach to samples amplified from pooled bulk wild Mason Farm soil DNA ('soil') and pooled root endophyte compartment DNA grown in that soil[9] ('root EC'; Online Methods). All 16S reads were processed into untreated sequences, ConSeqs and singletons as above, as well as a mix of 'ConSeqs plus adjusted singletons' (CASs), in which the singletons were downsampled in proportion to the ConSeqs collapse ratio (the ratio of the number of all ConSeqs to the number of all constituent sequences used to compute them). CASs thus retain the majority of singletons from template-overloaded samples, in which singletons contain the majority of high-quality reads; but they retain fewer singletons from dilute samples, in which the singletons are enriched for lower-quality outcasts. We generated OTUs at 97% and 99% identity thresholds and used rarefaction curves to observe the microbial richness (**Fig. 2a**). Within both root EC and the more complex soil communities, ConSeqs and CASs performed similarly and gave estimates of microbial richness lower than those of untreated sequences. This effect was particularly apparent at 99% clustering, but it was also evident at 97%, again demonstrating that ConSeqs correct overestimates of microbial alpha diversity[3].

MT treatments enhanced the technical reproducibility of independently amplified samples. Our data set comprised 12 pairs of root EC replicates and 12 pairs of soil replicates (**Supplementary Table 2a**,**d** and Online Methods). The OTU abundances of all samples were regressed against those of their replicates, and the coefficient of determination R^2 was graphed (**Fig. 2b**). Low-abundance OTUs were the least correlated[9–11]; as these were removed, R^2 increased quickly. Even before low-abundance OTUs were dropped, ConSeqs and CASs were more reproducible than untreated sequences and singletons, and their R^2 plateaued more quickly. Singletons formed many more small OTUs than did other classes, especially at 99% clustering. Thus, relatively more of the irreproducible singleton data were discarded at lower OTU abundance thresholds than for other MT classes, which explains the more rapid increase in technical reproducibility for singletons than for untreated sequences.

We compared our method directly to that of the Earth Microbiome Project (EMP), which uses primers without MTs[12]. Using both methods, we prepared libraries of the same sample composition, including independent soil samples from two sites, root EC samples from individual plants grown in one of the soils and the clonal 16S template used above (**Supplementary Table 2b–d**

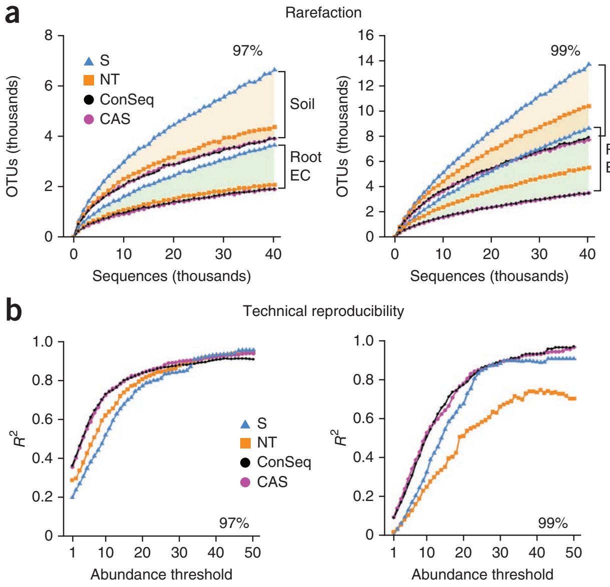

Figure 2 | Molecular tagging lowers estimates of alpha diversity and improves technical reproducibility. (**a**) 16S sequences with no MTs (NT), ConSeqs from two or more sequences with identical MTs (ConSeq), singleton MTs (S) and a combination of ConSeqs and a downsampled fraction of the residual singletons (CAS) were rarefied before OTU formation and clustered independently into OTUs at 97% (left) and 99% (right) identity. Bacterial reads from root EC or soil samples were pooled, producing a soil pool and a root EC pool per MT treatment at each identity threshold. These pools were rarefied at intervals of 1,000 sequences, and the number of OTUs observed at each depth were plotted. Beige shading connects soil samples; green shading connects EC root samples. (**b**) Progressive drop-out analysis displaying the coefficient of determination (R^2) of 24 intra-run technical replicates as OTUs with low read numbers are discarded. OTU tables are the same as in **a**, with the exception that plastid and mitochondrial OTUs were not removed.

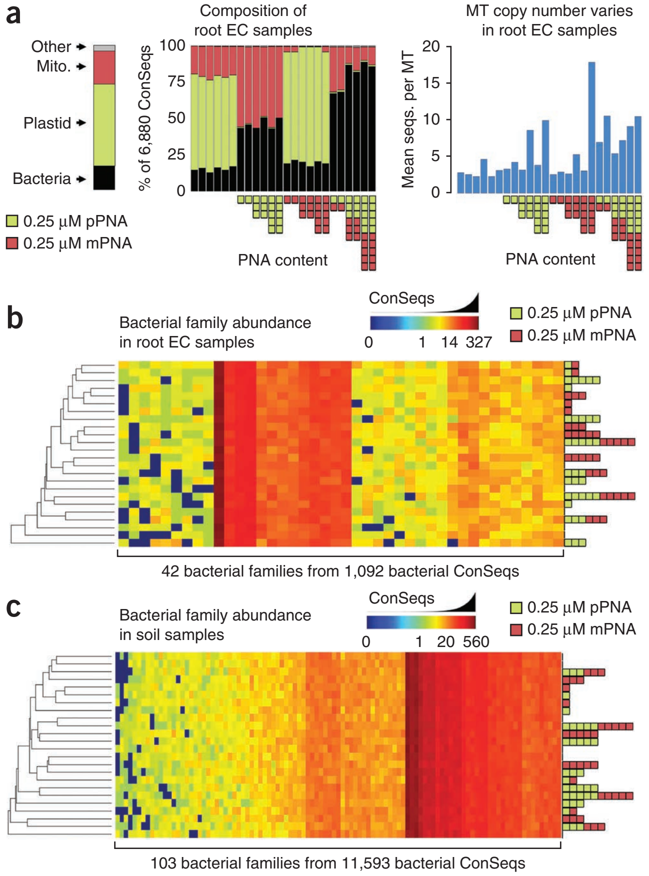

Figure 3 | PNA specifically blocks amplification of contaminant sequences. (**a**) The stacked bar chart legend (left) schematizes the relative abundance of ConSeqs classified as bacteria, plastid, mitochondria (Mito.) and other. PNA was titrated into PCR reactions of root EC DNA. Each green or red block below the histogram represents 0.25 μM of pPNA or mPNA in the final reaction, respectively. The sequence copy number per MT, and thus the mean number of sequences (seqs.) in each alignment used to compute the ConSeqs (blue bars, right), is determined by the sequencing depth and the amplifiable template concentration. (**b**) Root EC samples (rows) to which varying titrations of PNA had been applied (colored blocks) were clustered on the basis of the abundance of bacterial families (columns; family IDs not shown). The relative abundance of each bacterial family is displayed as a heat map. (**c**) Clustering and abundance as in **b** but with soil samples. Note that there is no clustering by PNA treatment in **b** and **c**.

and Online Methods). Major beta diversity conclusions from both methods were the same; the sample types grouped similarly after we performed principal-coordinates analysis based on weighted UniFrac distances (**Supplementary Fig. 8**). Also, the same clades formed on the basis of hierarchical clustering by Bray-Curtis dissimilarity. However, there were fewer OTUs using our method, which is consistent with our initial data (**Figs. 1a** and **2a**). Evidence that the extra OTUs are noise comes from the clonal 16S template, which formed one OTU with our method, as opposed to several with the EMP method.

Next we tackled a problem encountered when investigating microbial communities associated with a eukaryotic host, wherein 16S sequences originating from the host's genome, plastid or mitochondria can account for >80% of the sequences obtained[9,10,13]. Although modification of the bases in the 'universal' amplicon primers can mitigate amplification of the contamination, this can also lead to bias[14]. We instead developed peptide nucleic acid (PNA) PCR clamps[15]: synthetic oligomers that bind tightly and specifically to a unique signature in the contaminant sequence and physically block its amplification[13,16–18] (**Supplementary Fig. 9** and Online Methods). We designed PNAs to suppress plant host plastid and mitochondrial 16S contamination (**Supplementary Fig. 10**) and tested them using 24 samples amplified from pooled root EC DNA samples, in which ~85% of 16S sequences post-PCR were either plastid or mitochondria (**Fig. 3a**). Combining both PNAs in the same reaction blocked both types of contaminant and yielded approximately eightfold more bacterial 16S rRNA sequence as a fraction of total sequences.

Owing to an effective PNA-dependent template reduction, the mean number of sequences sharing an MT that were aligned and used to calculate each ConSeq was ~2.5-fold larger in the 12 samples containing anti-plastid PNA (pPNA; $P = 0.026$, permutation test of the means) (**Fig. 3a**). Neither the presence of pPNA or anti-mitochondrial PNA (mPNA) nor the related increase in the number of sequences per alignment affected clustering of root EC samples by bacterial families or OTUs (**Fig. 3b** and **Supplementary Fig. 11a**). There was also not a significant effect on the relative abundance of individual bacterial families or OTUs when the 12 samples amplified with each PNA were compared to the 12 samples amplified without it ($Q > 0.05$ for all permutation tests on the means with false discovery rate (FDR) correction; **Supplementary Figs. 11a** and **12a** and Online Methods). Using the same PNA concentrations for PCR of extremely diverse bulk soil[9], we observed that PNAs had no effect on clustering of samples by bacterial families or OTUs (**Fig. 3c** and **Supplementary Fig. 11b**) or the abundances of families or OTUs ($Q > 0.05$ for all permutation tests on the means with FDR correction; **Supplementary Figs. 11b** and **12b**), with one exception that was likely a false positive (**Supplementary Fig. 12b–d**).

Both the pPNA and mPNA sequences are conserved among higher plants and should function well for most plant microbiome projects (**Supplementary Fig. 13**). Many studies have demonstrated the potential of PNAs for a variety of research questions using low-resolution molecular methods[13,15,17–20], but a proof-of-concept study using deep sequencing has been lacking. A recent study showed the effectiveness of PNAs designed to block plastid and mitochondrial sequences for plant microbiome analysis using T-RFLP[13]. However, the authors considered only primer annealing-blocking regions that overlapped with conserved 16S primers, which limited the number of candidate PNAs and likely their target specificity (**Supplementary Fig. 14**).

Sequence features in the molecule tagging–frameshifting (MT-FS) primers can be used as additional barcodes. For example, nonintersecting sets of frameshifting primers on two samples sharing the same PCR barcode—or better, conventional barcoding bases in the MT-FS primers—allowed samples to be distinguished with >99.9% accuracy (**Supplementary Fig. 15b**). Each MT-FS barcode, or even unrelated template-tagging primers such as ITS region primers, can be used with the universal PCR barcodes, thereby enhancing the cost-effectiveness of our approach (**Supplementary Figs. 16** and **17**).

We also provide our validated MTToolbox: user-friendly software to merge overlapping paired-end reads, recognize and trim primer sequences, and process molecular tags into ConSeqs. MTToolbox is compatible with data produced by the related Safe-SeqS[6] and LEA-Seq[7] techniques. Downloads and source code can be accessed through SourceForge (http://sourceforge.net/projects/mttoolbox/), and user manuals and documentation can be found at https://sites.google.com/site/moleculetagtoolbox/).

In summary, our methods provided higher sequencing accuracy and technical reproducibility while increasing flexibility and savings. In the case of a MiSeq run of 96 root EC samples in which the PNAs were applicable, the combination of frameshifts, combinatorial barcoding and PNA yielded substantial cost reductions and provided greater flexibility to investigate new amplicons. These techniques can be adopted à la carte for a particular amplicon project and sequencing platform. The benefits of frameshifting and template tagging were independently described in a metagenomics context during the revision of this work[7], attesting to the need for improved amplicon sequencing methods.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence Read Archive: ERP003492.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Tremblay and S.G. Tringe of the Department of Energy Joint Genome Institute for early discussions regarding their independent invention and adoption of frameshifting primers. We thank S.H. Paredes, C. Jabara, S. Biswas and H. Kelkar for essential discussions and S.L. Lebeis, N.W. Breakfield, B.J. Campbell and C.W. Schadt for comments on the manuscript. This work was supported by US National Science Foundation Microbial Systems Biology grant IOS-0958245 to J.L.D. D.S.L. was supported by US National Institutes of Health (NIH) Training Grant T32 GM07092-34. S.Y. was supported by NIH Training Grant T32 GM067553-06. J.L.D. acknowledges the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation for funding (in part via grant GBMF3030 to J.L.D.).

AUTHOR CONTRIBUTIONS

D.S.L., P.M., C.D.J. and J.L.D. conceived wet-bench methods. S.Y. designed and wrote the informatics pipeline. D.S.L. and P.M. designed and performed experiments. D.S.L. and S.Y. analyzed data. D.S.L. wrote the manuscript with help from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Lozupone, C.A. & Knight, R. *Proc. Natl. Acad. Sci. USA* **104**, 11436–11440 (2007).
2. Krueger, F., Andrews, S.R. & Osborne, C.S. *PLoS ONE* **6**, e16607 (2011).
3. Patin, N.V., Kunin, V., Lidström, U. & Ashby, M. *Microb. Ecol.* **65**, 709–719 (2013).
4. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A. & Swanstrom, R. *Proc. Natl. Acad. Sci. USA* **108**, 20166–20171 (2012).
5. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2011).
6. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
7. Faith, J.J. *et al. Science* **341**, 1237439 (2013).
8. Sheward, D.J., Murrell, B. & Williamson, C. *Proc. Natl. Acad. Sci. USA* **109**, E1330 (2012).
9. Lundberg, D.S. *et al. Nature* **488**, 86–90 (2012).
10. Bulgarelli, D. *et al. Nature* **488**, 91–95 (2012).
11. Benson, A.K. *et al. Proc. Natl. Acad. Sci. USA* **107**, 18933–18938 (2010).
12. Caporaso, J.G. *et al. ISME J.* **6**, 1621–1624 (2012).
13. Sakai, M. & Ikenaga, M. *J. Microbiol. Methods* **92**, 281–288 (2013).
14. Sim, K. *et al. PLoS ONE* **7**, e32543 (2012).
15. von Wintzingerode, F., Landt, O., Ehrlich, A. & Göbel, U.B. *Appl. Environ. Microbiol.* **66**, 549–557 (2000).
16. Tanaka, T. *et al. Int. J. Cancer* **126**, 651–655 (2010).
17. Troedsson, C. *et al. Appl. Environ. Microbiol.* **74**, 4346–4353 (2008).
18. Ray, A. & Nordén, B. *FASEB J.* **14**, 1041–1060 (2000).
19. Chow, S. *et al. Mar. Biotechnol. (NY)* **13**, 305–313 (2011).
20. Terahara, T. *et al. PLoS ONE* **6**, e25715 (2011).

ONLINE METHODS

Additional detailed methods specific to figures. Additional details related to the creation of figures and supplementary figures is provided as a separate document (**Supplementary Note**).

Cloned 16S template. We amplified a 16S gene from a *Mycobacterium* sp. using primers 27F and 1492R and 25 PCR cycles, cloned the PCR product into pENTR/D-TOPO (Invitrogen) and selected a single transformed *Escherichia coli* colony. Plasmid DNA was prepped from a 3 mL culture using standard alkaline lysis, purified by silica column, quantified using a NanoDrop 1000 (Thermo Scientific) and sequenced using an ABI3130 genetic analyzer using 515F and 806R variable region 4 (V4) primers. The forward and reverse reads were overlapped and merged using Sequencher (http://genecodes.com/). Primer sequences were recognized and removed, thereby generating a high-quality sequence (**Supplementary Note**).

Root EC, soil and leaf DNA extraction and quantification. Mason Farm root endophyte compartment DNA (root EC), Mason Farm bulk soil DNA (soil), and Clayton bulk soil DNA (Clayton soil) were collected and extracted as previously described in Lundberg *et al.*[9]. All *Arabidopsis* DNA was made from the *Arabidopsis thaliana* Col-0 reference accession. *A. thaliana* and *Oryza sativa* leaf DNA were prepared in the same manner as root EC DNA, except that a similar quantity of whole leaves was prepped fresh, without sonication, bleaching or any other treatment to remove epiphytes. DNA templates were quantified using PicoGreen fluorescent dye (Invitrogen) and a fluorescence plate reader exciting at 475 nm and reading at 530 nm. Leaf DNA could not be reliably quantified, as it showed fluorescence at the limits of detection, and was therefore added without dilution in the template-tagging reactions (described below). For the individual samples used in the comparison of our method (Run C) to the Earth Microbiome Project (EMP) method (Run D), approximately 50 ng/μL was used for each sample.

Peptide nucleic acid (PNA) design. To identify candidate PNA oligo sequences, we fragmented *in silico* the full length *A. thaliana* plastid and mitochondrial 16S sequences into short *k*-mers for *k* of length 9, 10, 11, 12 and 13, and we queried for exact matches against the 4 February 2011 version of the Greengenes 16S training set comprising 35,430 unique, high-quality full-length bacterial sequences (**Supplementary Fig. 10**). *A. thaliana*–specific *k*-mers falling between the 515F and 806R 16S rRNA primers (V4 region) were considered candidates and were lengthened as necessary to increase the predicted melting temperatures and were screened for design characteristics[15,20].

A successful elongation arrest PNA clamp is generally between 13 bp and 17 bp and has an annealing temperature above that of the PCR primer whose extension it blocks and a melting temperature above that used for the extension cycle[20]. We designed 17-mer sequences to block the plastid and mitochondria, each with a predicted melting temperature around 80 °C (**Supplementary Table 1f**). Melting temperature, problematic hairpins, GC content and other design considerations were calculated using the Life Technologies PNA designer (http://www6.appliedbiosystems.com/support/pnadesigner.cfm).

The anti-mitochondrial PNA (mPNA) 5′-GGCAAGTGTTCTTCGGA-3′ and the anti-plastid PNA (pPNA) 5′-GGCTCAACCCTGGACAG-3′ (**Supplementary Table 1f**) were ordered from PNA Bio. Lyophilized PNA was resuspended in sterile water to a stock concentration of 100 μM. For PNA concentrations that were repeatedly tested, working stocks of 5 μM, 15 μM, 25 μM and 40 μM were prepared in water. All stocks were stored at −20 °C and heated to 65 °C before use to resolubilize any precipitate.

Primer design. All primers longer than 45 bases were Ultramers from Integrated DNA Technologies, purified by standard desalting. Shorter primers, such as the sequencing primers, were ordered from Eurofins MWG Operon and purified by the QuickLC method. Forward and reverse molecule tagging–frameshifting (MT-FS or Bc-MT-FS for nonbarcoded and barcoded, respectively) V4 16S primers and universal barcoding PCR primers are diagrammed and listed in **Supplementary Figure 2** and **Supplementary Table 1a**,**b**. Forward and reverse molecule-tagging ITS2 primers are diagrammed and listed in **Supplementary Table 1a**,**b**. Primers used for comparison to the EMP method are in **Supplementary Table 3**.

MT-FS primers and their barcoded versions, Bc-MT-FS primers, were designed with the frameshift and barcoding bases occurring within the molecular tag regions to break up the stretch of random bases and minimize unpredictable features related to annealing and secondary structure. We used 2-bp linkers to buffer the template-annealing 515F and 806R portions of the MT-FS primers from the rest of the primer. Ideal linkers have low homology to known microbial sequences, creating a short stretch of mispairing. Our linker sequences for the V4 16S region differ from those used in the EMP method[12] but are equally valid choices on the basis of the lack of matches to the Greengenes database (**Supplementary Fig. 17**).

The molecule-tagging ITS2 primers are similar but are of an earlier design that uses nine random bases for the forward primer and four random bases for the reverse primer. No frameshifting variants of the ITS2 primers were used.

The 9-bp barcodes we used for the universal barcoding PCR primers were adapted from the 12-bp Golay barcodes used by Caporaso and colleagues[12]. Of the 2,168 published Golay barcodes, we chose a subset of 96 that had a balanced mix of all bases at each position. We then extracted just the first 9 bases of these 12-bp barcodes; in our set of 96 barcodes of 9 bp, three or more SNPs would be needed to transform any one barcode into another. We chose to trim the Golay barcodes from 12 to 9 in order to shorten the primers; deeper barcoding can be accomplished by adding mini-barcodes in the MT-FS primers, such as the 3-bp barcodes we chose (**Supplementary Fig. 2b**), and combining each mini-barcode used during template tagging with the full suite of 96 universal barcodes in PCR.

Template tagging with molecular tagging–frameshifting primers. Template DNA was tagged with the MT-FS primers in two reactions: one for the reverse MT-FS primers and a subsequent reaction for the forward MT-FS or Bc-MT-FS primers, as described below. The purpose of using the tagging primers in two separate reactions, one for each primer, was to reduce the possibility of formation of difficult-to-remove heterodimers between the long MT-FS primers. The shorter reverse MT-FS primers were used to tag the template first because removal of shorter primers during PCR cleanup is more efficient. Although the use of separate tagging reactions discourages heterodimers, it is not strictly necessary; and in practice both forward- and reverse-tagging primers can be used in a single two-cycle template-tagging reaction with good results (not shown).

For reverse V4 16S tagging in Run B, the primary MiSeq run we analyzed, we prepared two working stocks of reverse MT-FS V4 16S primer in water, where each working stock contained an equimolar mix of three of our six primers such that the concentration of the mixed stock was 0.5 µM. These working stocks we designate "V4R_2-4-6" (806R_f2, 806R_f4, and 806R_f6) and "V4R_1-3-5" (806R_f1, 806R_f3, and 806R_f5). For Run C, which we used to compare our method directly to the EMP method, we used a mix of all six reverse MT-FS primers ("V4R_mix1-6") such that the concentration of the mixed stock was again 0.5 µM.

We used the KAPA 2G Robust HS PCR Kit with dNTPs (KK5518, Kapa Biosystems) in a 25 µL including 5 µL Kapa Enhancer, 5 µL Kapa Buffer A, 2 µL of 0.5 µM reverse-tagging primer mix ("V4R_1-3-5" or "V4R_2-4-6" for Run B or "V4R_mix1-6" for Run C), 0.5 µL Kapa dNTPs, 0.25 µL Kapa Robust Taq and 12.5 µL DNA template with water.

To minimize pipetting variation of small volumes, we used master mixes to prepare reagents whenever possible. Samples were incubated in a thermocycler using a program of denaturing at 95 °C for 1 min, reverse–MT-FS primer annealing at 50 °C for 2 min, and extension at 72 °C for 2 min, followed by a cooldown to 4 °C. The newly synthesized reverse-tagged strands, as well as the original DNA template molecules to which they were annealed, were cleaned to remove primers and PCR reagents with Agencourt AMPure XP beads (Beckman Coulter) using the manufacturer's protocol with the exception of an altered bead-to-DNA ratio: we used 15 µL of beads to clean the 25 µL of tagged template because this ratio (0.6:1) allowed size selection that more effectively eliminated the long tagging primers (data not shown). The DNA was eluted in 11 µL of water.

The cleaned, reverse-tagged DNA was next tagged with forward primers. For Run B, we made two forward MT-FS working stocks of three frameshift variants each (**Supplementary Figs. 2b** and **15a**), which we designate "V4F_2-4-6" (515F_f2, 515F_f4, and 515F_f6) and "V4F_1-3-5" (515F_f1, 515F_f3, and 515F_f5). For Run C, we made two forward Bc-MT-FS working stocks of six frameshift variants each, where each Bc-MT-FS mix differed by its 3-bp barcode (**Supplementary Figs. 2b** and **15c**). We designate these "V4F_TGA_mix1-6" and "V4F_ACT_mix1-6."

For samples to which PNA was applied (**Supplementary Table 2a–d**), PNA was included in reactions in only the forward-tagging step, as the PNA blocks the extension of the forward-tagging primers. The 25-µL forward-tagging reaction included 5 µL Kapa Enhancer, 5 µL Kapa Buffer A, 2 µL of 0.5 µM forward-tagging primer mix ("V4F_1-3-5" or "V4F_2-4-6" for Run B or "V4F_TGA_mix1-6" or "V4F_ACT_mix1-6" for Run C), 0.5 µL Kapa dNTPs, 0.25 µL Kapa Robust Taq, 2.5 µL PNA working stock (containing pPNA, mPNA, both mPNA and pPNA, or water) and 10 µL reverse-tagged DNA from above.

Samples were incubated in a thermocycler using a program of denaturing at 95 °C for 1 min, PNA annealing at 78 °C for 10 s, forward tagging–primer annealing at 50 °C for 2 min and extension at 72 °C for 2 min, followed by a cooldown to 4 °C. The DNA, now tagged with both forward- and reverse-tagging primers, was cleaned with Agencourt beads using 17.5 µL of beads to clean the 25 µL of tagged template. A marginally more conservative bead-to-DNA ratio of 0.7:1 was used to clean the dual-tagged template as compared to single-tagged template because the overall length of dual-tagged template (<500 bp) is shorter than that of single-tagged template (>1 kbp). The dual-tagged DNA was eluted in 16 µL of water.

ITS tagging was similar to that for V4 16S, except that there was only one reverse primer in the 0.5 µM reverse working stock and only one forward primer in the 0.5 µM forward working stock (**Supplementary Table 1a**,**b**).

PCR using tagged templates (our method). We performed PCR in a 50-µL reaction mix, in which the reverse primer differed for each individually barcoded sample (**Supplementary Table 2a–c**). The mix included 25 µL Kapa HiFi HotStart ReadyMix (KK2602, Kapa Biosystems), 2.5 µL PCR_F forward primer (from 5 µM working stock), 2.5 µL PCR_R_bc reverse primer (from 5 µM working stock), 5 µL mixed PNA working stock or water, and 15 µL DNA from the forward template–tagging step.

The PCR program was denaturation at 95 °C for 45 s followed by 34 cycles of denaturation at 95 °C for 15 s, PNA annealing at 78 °C for 10 s, primer annealing at 60 °C for 30 s, extension at 72 °C for 30 s and then a cooldown to 4 °C. All samples were cleaned with Agencourt beads using 35 µL of beads to clean the 50-µL PCR (0.7:1). DNA was eluted in 50 µL water.

PCR using untagged templates (EMP method). We used the primers and protocol available at http://www.earthmicrobiome.org/, with some exceptions to improve direct comparability with our method. The exact EMP primers used are listed in **Supplementary Table 3**. The first exception to the published protocol is that we used 2× Kapa HiFi Ready Mix for the PCR, which is the same polymerase we used for the PCR in our method. The second exception is that we altered the thermocycling conditions to be more similar to ours (with the exception of the primer annealing temperature) and to include a PNA annealing step. The altered EMP thermocycling conditions were denaturing at 95 °C for 45 s followed by 35 cycles of denaturation at 95 °C for 15 s, PNA annealing at 78 °C for 10s, primer annealing at 50 °C for 30 s, extension at 72 °C for 30 s and ending with a cooldown to 4 °C. All samples were cleaned with Agencourt beads using 35 µL of beads to clean the 50-µL PCR (0.7:1). DNA was eluted in 50 µL of water.

Quantification of PCR products and library mixing. From all cleaned PCR reactions, 1 µL was quantified in 96-well plate format using PicoGreen fluorescent dye (Invitrogen) and a fluorescence plate reader exciting at 475 nm and reading at 530 nm. The PCR reactions were mixed at equimolar ratios to make a pooled library for each run. For analysis purposes, in a setup run (Run A) and our primary run (Run B), we included from each run all potentially sequenceable material from low-yield and negative-control samples: low-quality material enriched for primer dimers and other abnormal amplicons that decrease the overall quality of the run. In Run C and Run D, samples with DNA below the detection limit were not used.

doi:10.1038/nmeth.2634

The mixed libraries were purified once more using Agencourt beads at a 0.7:1 bead-to-library ratio and were eluted in half the original volume to concentrate the final libraries. Each final library was quantified in triplicate using PicoGreen, and the values were averaged to reach a library quantification.

Library denaturation, dilution and sequencing. The final library was diluted to 4 nM, assuming an average amplicon length, including adaptors, of 448 bp. To denature the DNA, we mixed 5 µL of the 4 nM library with 5 µL of 0.2 N fresh NaOH and incubated 5 min at room temperature. 990 µL of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, 275 µL of the 20 pM library was mixed with 725 µL of chilled HT1 buffer to make a 5.5 pM sequenceable library, which was kept on ice until use. We noticed that 5.5 pM gave us a cluster density of between 700 K/mm^2 and 900 K/mm^2, which gave the best balance of quantity (which improves with higher cluster density) and quality (which improves with lower cluster density). The Illumina recommended range is 500 K/mm^2–1,200 K/mm^2. A 500-cycle v2 MiSeq reagent cartridge was thawed for 1 h in a water bath, inverted ten times to mix the thawed reagents, and stored at 4 °C a short time until use.

For sequencing in Run A, Run B and Run C using our method, the custom Illumina Nextera P1 primer ("Read1_seq"; **Supplementary Table 1e**), was used as the forward sequencing primer for read 1 and was prepared by mixing 3 µL of 100 µM stock into 597 µL HT1 buffer to make a 0.5 µM solution. A MiSeq v2 flow cell was rinsed with water and ethanol and polished dry with lens paper. The 5.5 pM library was loaded into the "Load Sample" well, and the custom Nextera primer solution was loaded into port 18 of the reagent cartridge. The "Settings" section of the sample sheet was modified to include "C1" as the "CustomRead1PrimerMix" and "5′-AGATCGGAAGAGCACACGTC-3′" as the adaptor. Read 2 was sequenced with the TruSeq read 2 sequencing primer already present in the reagent cartridge ("Read2_seq"; **Supplementary Table 1e**), and the barcode read was sequenced with the TruSeq Index Read Sequencing Primer ("Barcode_seq"; **Supplementary Table 1e**). The sample sheet along with sample names and the corresponding reverse complement of each nine-nucleotide barcode sequence was uploaded onto the MiSeq instrument before each run. The machine does not use the final base of the barcode read for annotation, and so each sample was associated with an 8-bp read sequence. The sample sheets used for Run B and Run C are available in **Supplementary Table 4**.

For Run A and Run B, we applied a feature in Real-Time Analysis (RTA v1.17.22) that allowed the machine to use a hardcoded matrix and phasing calculations. This modification improved the performance of low diversity libraries. In order to do this we altered the MiSeqConfiguration.xml file (this modification required assistance from an Illumina field application specialist). For Run C, we upgraded our machine to the new version of Real-Time Analysis (RTA v1.17.28) and used the default feature of the upgrade without additional hardcoded matrix or phasing modifications.

For sequencing in Run D (EMP method), all custom sequencing primers were prepared by mixing 3 µL of 100 µM primer stock into 597 µL HT1 buffer to make a 0.5 µM solution. The custom primer "EMP_Read1_seq" (**Supplementary Table 3**) was used as the forward sequencing primer for read 1 and was loaded into port 18 of the reagent cartridge. "EMP_Read2_seq" was used as the forward sequencing primer for read 2 and was loaded into port 20. "EMP_barcode_seq" was used to sequence the sample barcode and was loaded into port 19. The Settings section of the sample sheet was modified to include "C1" as the "CustomRead1PrimerMix," "C2" as the "CustomIndexPrimerMix," and "C3" as the "CustomRead2PrimerMix." The sample sheet along with sample names and the corresponding reverse complement of each 12-nucleotide barcode sequence was uploaded onto the MiSeq instrument before the run. The machine does not use the final base of the barcode read for annotation, and so each sample was associated with an 11-bp read sequence. The sample sheet used for Run D is available in **Supplementary Table 4**. As with Run C, Run D was completed using Real-Time Analysis (RTA v1.17.28) without additional software modifications.

Demultiplexing. Standard preprocessing and demultiplexing of PCR barcodes were performed with Consensus Assessment of Sequence and Variation (CASAVA) software (Illumina, v.1.8.2), allowing for 0 mismatches to the sample barcodes.

Raw sequence processing (our method). Paired-end overlapping and merging, as well as recognition of pattern-matching sequences and MT processing, were performed using MTToolbox, a freely available software package hosted by SourceForge (https://sourceforge.net/projects/mttoolbox/). Documentation and user manuals can be accessed via the MTToolbox web page (https://sites.google.com/site/moleculetagtoolbox/).

Paired ends were overlapped with FLASH[21] using parameters "-m 30 -M 250 -x 0.25 -p 33 -r 250 -f 310 -s 20" for all V4 16S samples and "-m 20 -M 250 -x 0.25 -p 33 -r 250 -f 400 -s 20" for all ITS samples. In the overlapping region, the bases with the highest quality score were chosen for the merged reads, with bases from Read1 preferred in the case of ties (**Supplementary Fig. 6e**).

In Run B, merged sequences in each sample were then matched to expected patterns for either V4 16S amplicons or ITS amplicons using the regular expressions "all_HQ_V4_sequences_RunB" or "all_HQ_ITS2_sequences" (**Supplementary Table 1g**). Because the merged V4 amplicons in Run C contained barcodes on the template-tagging Bc-MT-FS primers (**Supplementary Fig. 2b**), the slightly modified regular expression "all_HQ_V4_sequences_RunC" was used (**Supplementary Table 1g**). These expressions select sequences without ambiguous bases or errors in priming sequences.

From the pattern-matching sequences, the sequence fragment 5′ to the forward linker and the fragment 3′ to the reverse linker were extracted and concatenated to form that sequence's molecular tag (MT), and the sequence occurring between the forward and reverse template–specific primers was extracted for analysis (**Supplementary Fig. 6f**). We did not analyze sequences corresponding to the primers because we observed high sequence variability at the wobble bases, even when amplifying a clonal template, which indicated that the wobble base observed in the sequence is a poor indicator of the primed sequence (data not shown).

Each unique MT observed in a sample was considered a unique MT category (**Supplementary Fig. 6g**). Sequences sharing the same MT were classified as belonging to the same category, and for each category containing two or more sequences, a multiple sequence alignment was built using command line ClustalW[22] with parameters "-output=gde -outorder=input -case=upper -query -quicktree" (**Supplementary Fig. 6h**). A consensus sequence was calculated from the multiple sequence alignment by choosing the most common base at each position. For MT categories containing only two sequences (and for all other ties), the base with the highest average quality score was chosen; and if a tie could still not be resolved, an IUPAC base was used to indicate the tie in the consensus sequence. For each sample, a FASTA file of consensus sequences was built, with each consensus sequence given a composite name including the sample of origin, or "P_number_ID" followed by the MT of that consensus. For example: >P0_GGCTGACTTTAC-GGCAGTCAAT [Sequence].

MT categories in each sample that contained only one sequence (category depth = 1) could not be represented by a consensus, and the sequences in these categories, or 'singletons', were kept in a separate FASTA file with each sequence given a composite name including the sample the sequences came from, the sequence number within the corresponding sample, the MT sequence and the original read ID. For example: >P0_20176 GAGTAGGAATA-TCTAT UNC20:76:000000000-A315U:1:1101:14750:1667 1:N:0:GGCGCTTA [Sequence].

Raw sequence processing (EMP method). Paired ends were overlapped with FLASH[21] using parameters "-m 30 -M 250 -x 0.25 -p 33 -r 250 -f 310 -s 20."

EMP sequencing primers provide data between the highly conserved areas bound by the 515F and 806R primers; thus, regular expressions for these primers cannot be used to identify pattern-matching sequences. Therefore, we define high-quality sequences in the context of EMP data as sequence that successfully overlaps and merges and does not have ambiguous bases.

Operational taxonomic unit (OTU) formation. OTUs were built using OTUpipe, a collection of USearch (http://www.drive5.com/) commands encapsulated in a bash script that clusters sequences on the basis of their nucleotide identity and that removes chimeras that can form during PCR. First, FASTA files from samples to be clustered were concatenated into one file. OTUpipe was then run with nondefault parameters ABSKEW = 3 and MINSIZE = 1. For 99% OTU clustering, the following nondefault parameters were used: PCTID_ERR = 99, PCTID_OTU = 99, PCTID_BIN = 99. We did not make OTUs at higher than 99% because a single bacterial genome can harbor several copies of the 16S gene that differ on average by 0.55% (ref. 23), meaning that at identity thresholds higher than 99%, a single bacterium would form several OTUs even if error was eliminated.

OTU table construction. OTUs were built into OTU tables, and their taxonomy was assigned using functions in QIIME 1.5.0 (ref. 24). The OTUpipe output file "readmap.uc" was transformed into a QIIME cluster file by running the QIIME script "readmap-2qiime.py," generating the text file "qiime_otu_clusters.txt." This file was passed to the QIIME script "make_otu_table.py" to make a Biological Observation Matrix (BIOM) OTU table. Finally, the BIOM table was converted to a classic format OTU table using the QIIME script "convert_biom.py."

Assigning taxonomy to OTUs. Taxonomy was assigned to bacterial OTUs using the RDP classifier trained on the most recent (4 February 2011) Greengenes 97% identity taxonomy representatives and was accomplished by running the QIIME 1.5.0 script "assign_taxonomy.py" on OTU representative sequences using "greengenes_tax_rdp_train.txt" as the ID to taxonomy mapping file, "gg_97_otus_4feb2011.fasta" as the reference sequences and the parameter "-c 0.5."

Helpful instructions for running the QIIME scripts can be found by searching for the script name on the QIIME website (http://www.qiime.org/).

Owing to a focus on bacterial taxa, RDP trained on Greengenes did a poor job of recognizing plastid and mitochondrial sequences in our data. Rather than editing the training set, we further recognized plant contaminant OTUs by using BLAST to compare the representative sequences to a custom database containing the *Arabidopsis* 18S rRNA sequence as well as plastid and mitochondria 16S rRNA sequences from *Arabidopsis* and other plants (**Supplementary Table 5**). We used BLAST with an *E* value of 0.00001 and a percent identity of 94.

Predicting pPNA and mPNA utility across diverse plant families. The pPNA and mPNA sequences were tested for exact matches to representative chloroplast and mitochondrial 16S sequences from diverse plant species found in NCBI GenBank (**Supplementary Fig. 13** and **Supplementary Table 6**).

Subsampling. Normalization of FASTA files and all other subsampling was performed using the sample() function in the "base" library of R (http://www.r-project.org/). Rarefaction of OTU tables was performed using the function rrarefy() the "vegan" library of R, which also makes use of the sample() function.

Permutation tests. All permutation tests involved 24 samples and asked whether the mean value of 12 samples in "condition low" was lower than the mean value of 12 samples in "condition high." For each permutation test, the values from the 24 samples were randomly assigned into two groups of 12 using the sample() function in the base library of R, and the difference in the means of these groups was taken. This was repeated 10,000 times per test to form the probability distribution for each test. The *P* value was the fraction of 10,000 permutations in which the observed difference in the means would be as large due to chance.

A nonparametric test on the means was chosen in preference to a parametric *t*-test because of relatively low group size of 12 samples, which prevents accurate estimation of the underlying probability distributions and is not sufficiently large to make the assumption of normality under the Central Limit Theorem.

Correction for multiple testing. Permutation tests were used to test whether the relative abundances of bacterial families and bacterial OTUs were lower in PNA samples than in control samples, for all families and OTUs above the threshold (see figure-specific methods in the **Supplementary Note**). The green and red histograms of uncorrected *P* values display the results of these permutation tests for pPNA and mPNA (**Supplementary Figs. 11** and **12**). The *P* values within each histogram were corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) method as implemented by the p.adjust() function in the "stats" library of R, and the number of tests that

doi:10.1038/nmeth.2634

were included in each application of the FDR method is shown beneath each *P* value histogram.

Chi-squared tests. The green and red histograms of uncorrected *P* values display the results of permutation tests for pPNA and mPNA, respectively (**Supplementary Figs. 11** and **12a**,**b**). For root EC families and OTUs, and for soil families, there were ~100 or fewer tests, and ten bins were used for the *P* value histogram (**Supplementary Figs. 11a** and **12a**,**b**). For soil OTUs, there were 1,010 tests for each PNA, and 20 bins were used for higher resolution of the distribution histogram (**Supplementary Fig. 11b**). Each histogram was compared to the null flat distribution (equal number of *P* values in each bin of the histogram) using a Chi-squared test with 9 degrees of freedom for histograms with 10 bins or 19 degrees of freedom for histograms with 20 bins. Chi-squared tests were performed using the function chisq.test() in the stats library of R.

21. Magoč, T. & Salzberg, S.L. *Bioinformatics* **27**, 2957–2963 (2011).
22. Larkin, M.A. *et al. Bioinformatics* **23**, 2947–2948 (2007).
23. Pei, A.Y. *et al. Appl. Environ. Microbiol.* **76**, 3886–3897 (2010).
24. Caporaso, J.G. *et al. Nat. Methods* **7**, 335–336 (2010).

doi:10.1038/nmeth.2634