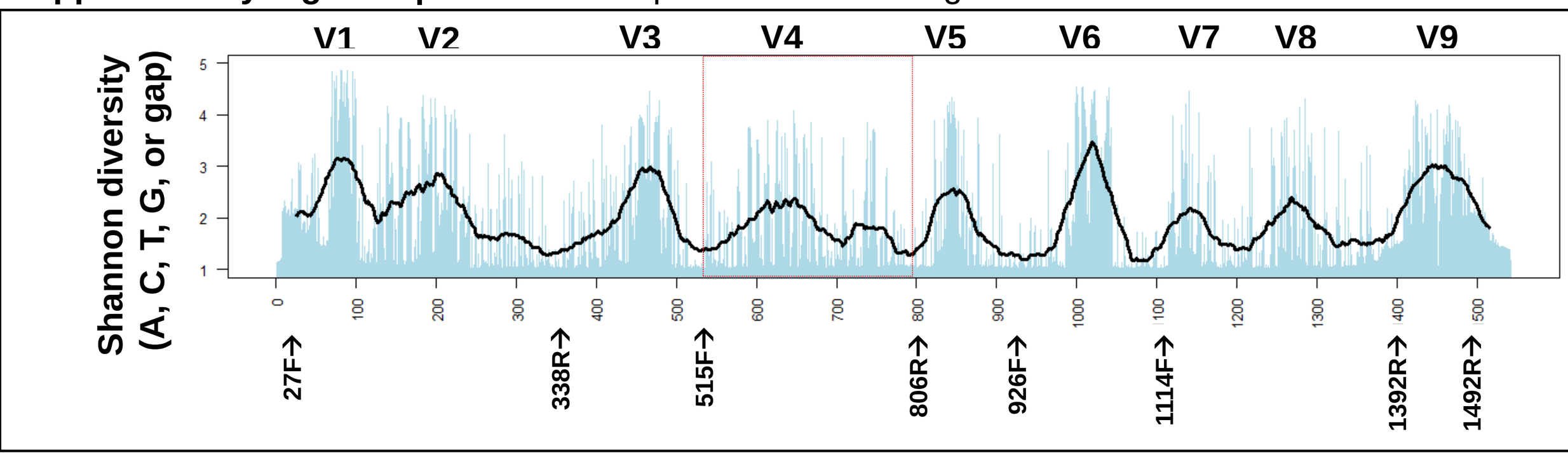
Supplementary Information for:

Practical innovations for high-throughput amplicon sequencing

Derek S. Lundberg*, Scott Yourstone*,
Piotr Mieczkowski, Corbin D. Jones, Jeffery L. Dangl

* contributed equally

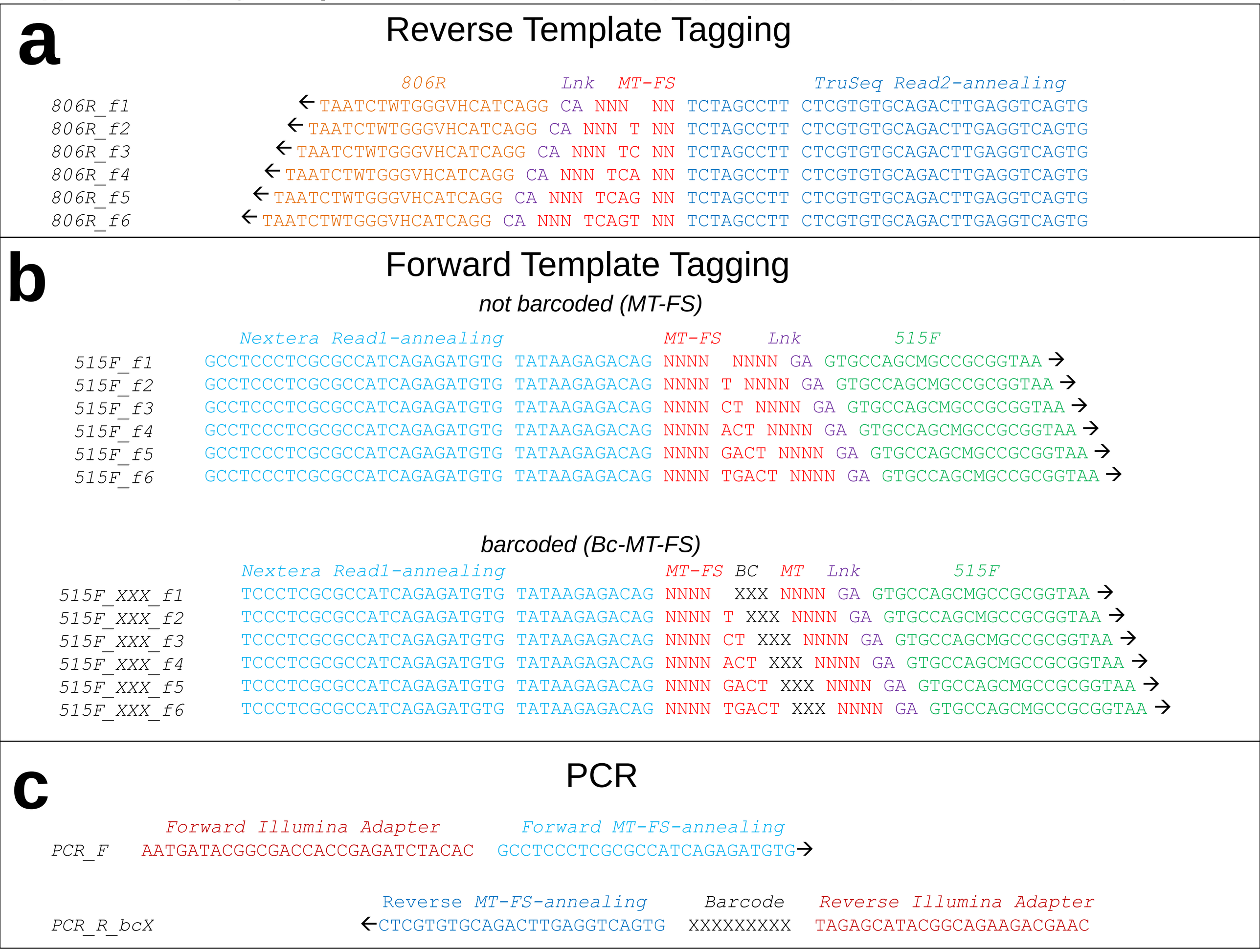
Supplementary Figure 1 | Reference map of the 16S rRNA gene.



Supplementary Figure 1 | Reference map of the 16S rRNA gene. Map shows variable regions V1-V9 (above chart) and the locations of common primers (based on conventional *E. coli* numbering, below chart). For each base present in *E. coli*, the Shannon Diversity of bases or gaps for that position is graphed in light blue histograms. The average Shannon Diversity based on a 50 bp sliding window is charted as a black line, displaying the classic 16S variable regions. The variable region V4 used in this study is boxed in red. Diversity was calculated by comparison to the Greengenes 97% representatives (most recent Feb. 4 2011 version) database of full length 16S sequences (Online Methods).

Nature Methods: doi:10.1038/nmeth.2634

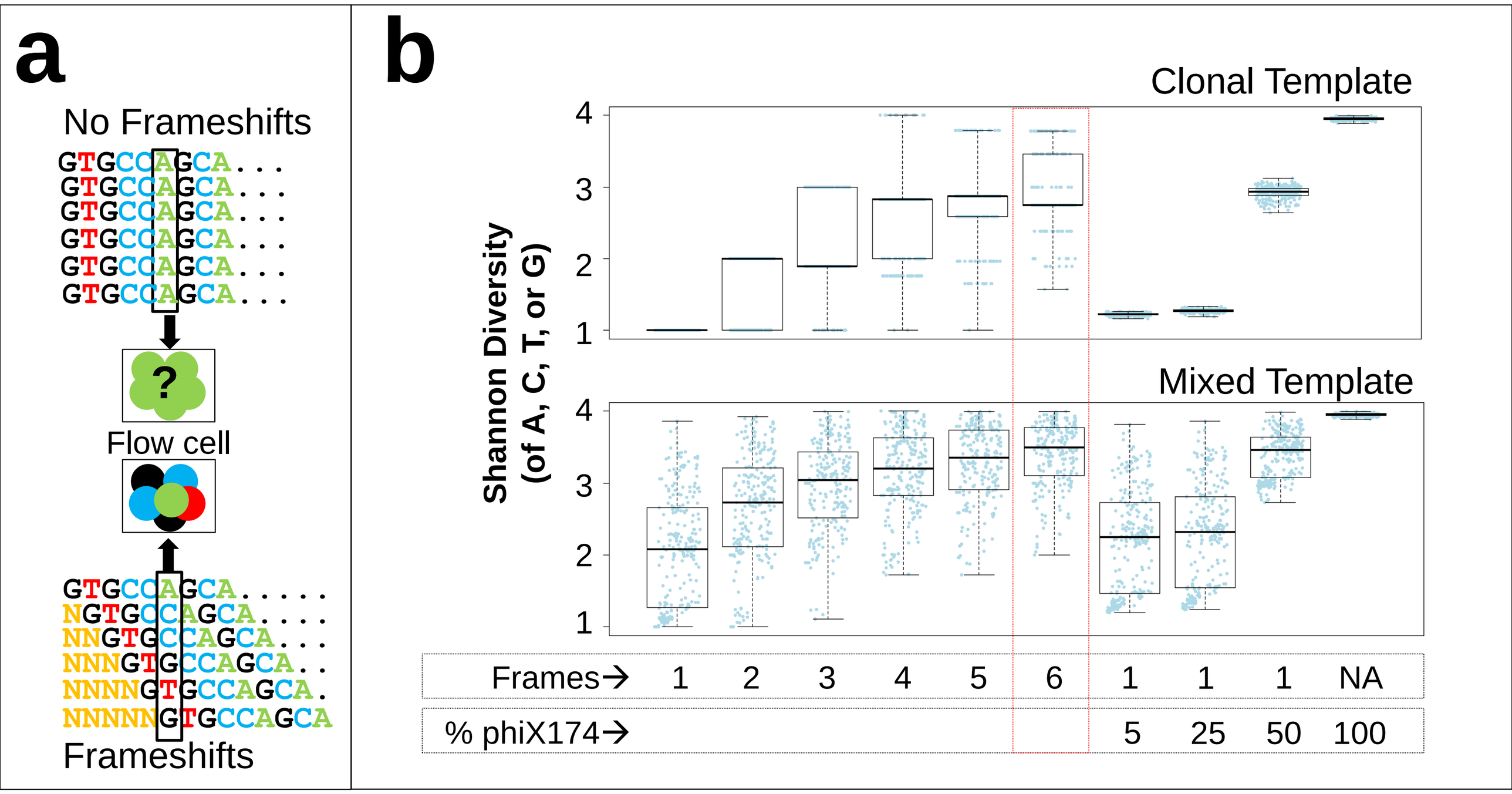
Supplementary Figure 2 | Schematic of Molecular Tagging - Frameshifting primers and PCR primers.



Supplementary Figure 2 | Schematic of molecular tagging - frameshifting template tagging primers. (**a**) MT-FS V4 16S reverse template-tagging primers. (**b**) Forward “MT-FS” V4 16S template-tagging primers (top), and forward barcoded “Bc-MT-FS” V4 16S template-tagging primers (bottom), where “XXX” is a three base pair barcode. MT-FS = Molecular tag and frameshifting bases. Lnk = Linker. “N” = MT random sequence (**c**) PCR primers. All primer sequences are available in **Supplementary Table 1a-c**.

Nature Methods: doi:10.1038/nmeth.2634

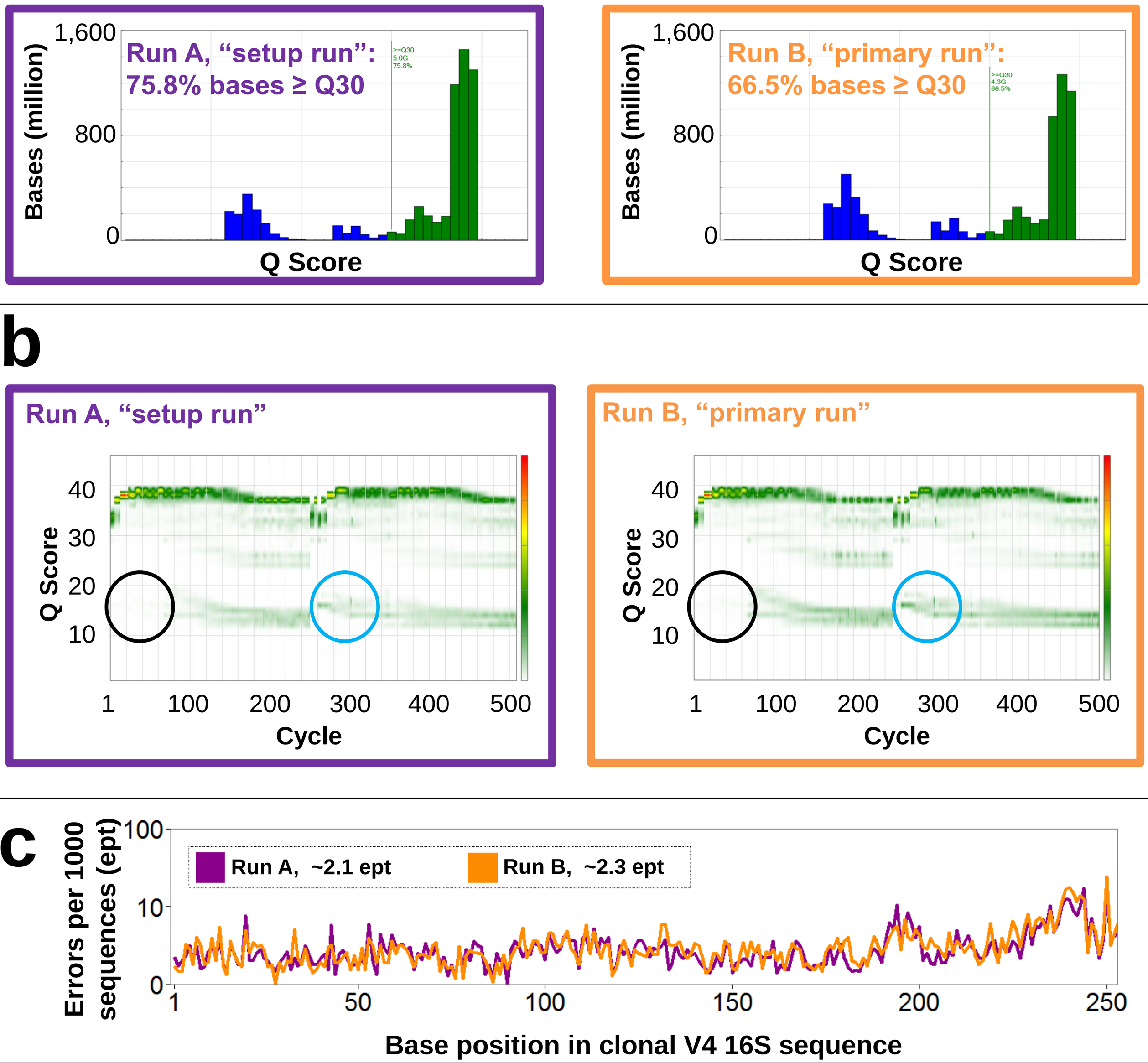
Supplementary Figure 3 | Frameshifting primers enhance library diversity.



Supplementary Figure 3 | Frameshifting primers enhance library diversity. (**a**) Schematic showing that frameshifts can impose diversity on a low-diversity library. (**b**) Diversity per sequenced base for simulated libraries made from a perfect clonal template (top) or a low-complexity template of 1000 real V4 bacterial 16S rRNA sequences (bottom). For each simulated library, subsets of 1000 sequences were randomly assigned to equally-sized groups to which six frameshifting treatments of 0-5 additional 5' bases were applied, creating between 1 and 6 frames ("Frames", below x-axis). Some libraries received simulated fragments of phiX174 genomic DNA in place of a fraction of the 1000 V4 16S sequences ("%phiX174", below x-axis). For each library, the Shannon diversity for each of the first 250 sequences was graphed (light blue dots), and the distribution summarized with a box-and-whiskers plot showing the extremes, upper and lower quartiles, and the median. Six frameshifts and no phiX174 were used in the remainder of this study (red box).

Nature Methods: doi:10.1038/nmeth.2634

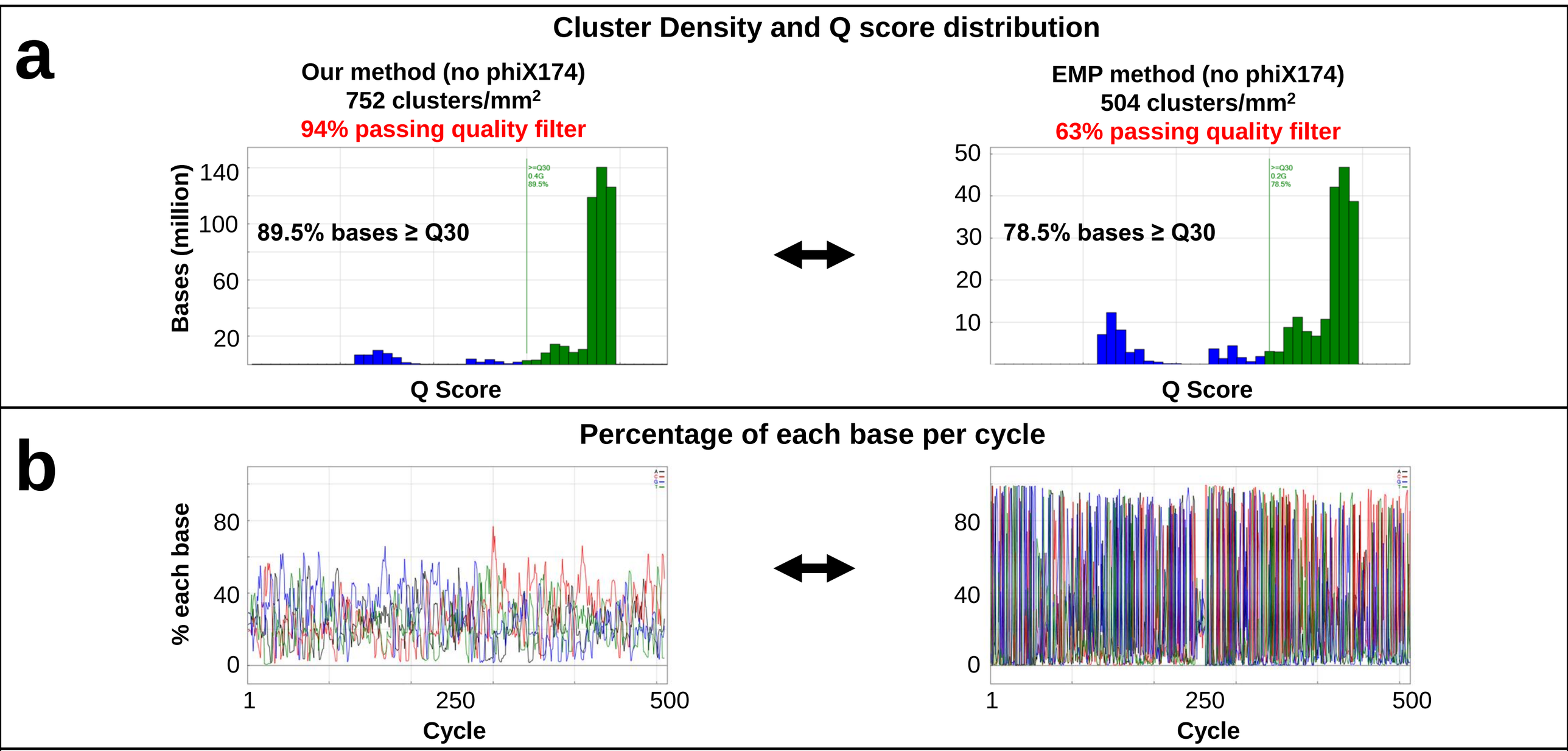
Supplementary Figure 4 | MiSeq run quality for Run A (setup run) and Run B (primary run).



Supplementary Figure 4 | MiSeq run quality for Run A (setup run) and Run B (primary run). Run A, a setup run, met Illumina quality specifications of sequencing pure phiX174 DNA and Run B, the primary run we analyze, came close. (**a**) Illumina MiSeq performance specifications for a 2 × 250 run of phiX174 is >75% of total bases above Q30 (not per cycle). A setup run without any phiX174 DNA, but containing a sample composition differing only in the initial concentration of several templates and library mixing (Online Methods), met the advertised specifications based on the machine’s statistics (top, purple box). The primary run we analyze (bottom; orange box), made up of a nearly identical composition of samples, was close (**Supplementary Table 2a**). This was despite deliberate includision in these runs of all potentially-sequenceable material from low-yield and negative control samples (**b**) Q Score heatmaps for setup run A (left; purple) and primary run B (right; orange). Both runs show sustained high quality, with diminishing quality towards the end of each run, and lower quality at the beginning of Read2 than of Read1 (circles). (**c**) Analysis of error rate across merged reads of a plasmid-borne clonal 16S rRNA template sample present in both runs reveals that the sequencing quality is similar in both runs. The mean error rate for pattern-matching (Online Methods) in each run is ~2.2 errors per thousand (ept) (color key), or Q27, with the error rate increasing towards the 3’ end of the read representing the non-overlapping portion of read 2, as expected.

Nature Methods: doi:10.1038/nmeth.2634

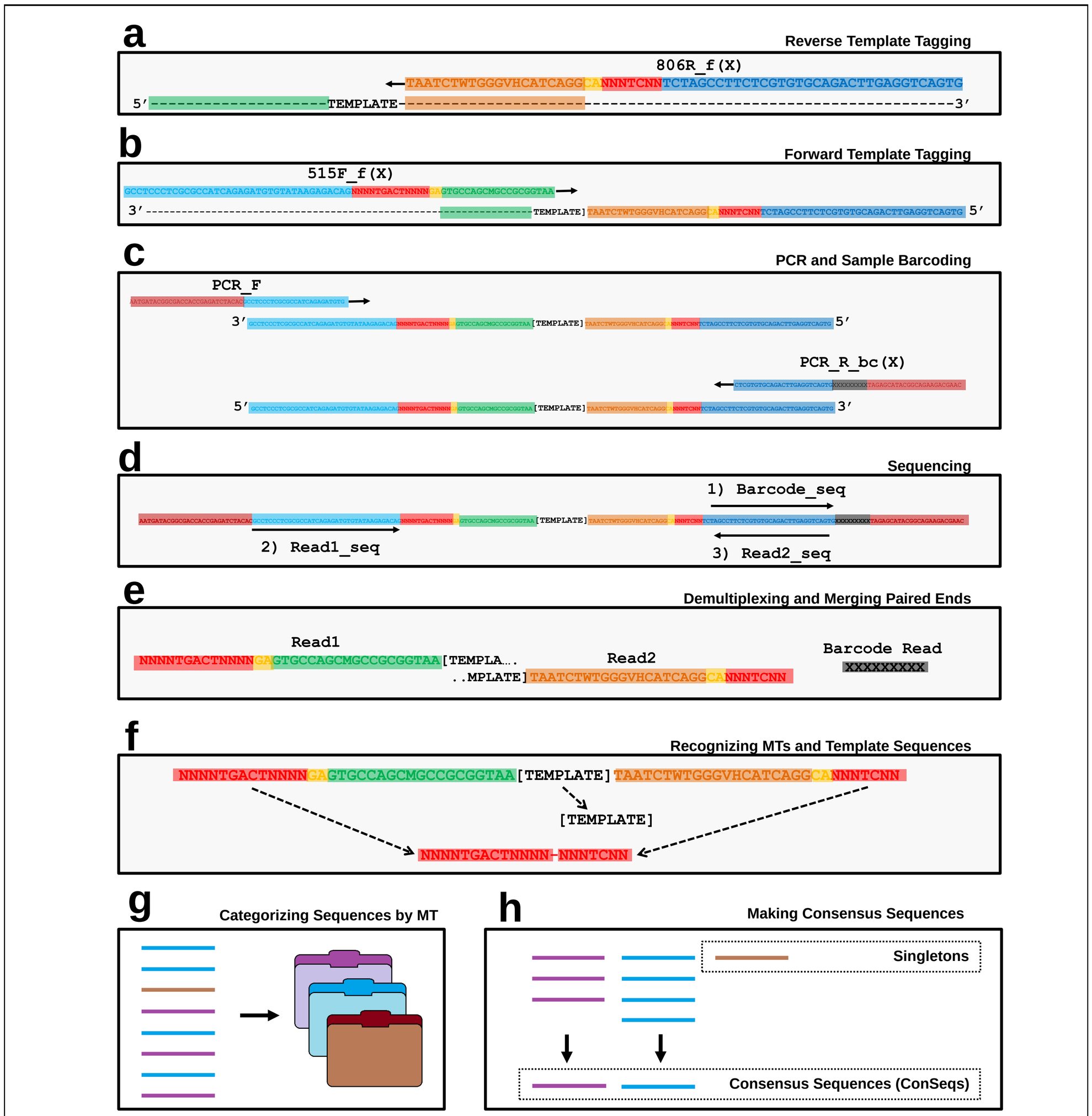
Supplementary Figure 5 | MiSeq run quality for Run C (our method) and Run D (Earth Microbiome Project method)



Supplementary Figure 5 | MiSeq run quality for Run C (our method) and Run D (Earth Microbiome Project method). The runs were consecutive, on a machine that had the Illumina May 2013 software upgrade to Real-Time Analysis v1.17.28. The recommended 5% phiX174 spike was not used for either run. Our method (left) and the EMP method (right) were each used in parallel to amplify 16S rRNA from the same set of samples (**Supplementary Table 2b,c**, Online Methods). Amplicons from each method were mixed to make two independent libraries. (**a**) The EMP library was loaded at a lower cluster density than the library prepared by our method – although this is expected to reduce crowding and improve cluster recognition, significantly fewer clusters passed the machine's quality filter. Of the high quality clusters, the percent of bases above Q30 was higher for the library prepared by our method. Both "nano" runs had more bases above Q30 than Run B used for the majority of analysis, likely the combined consequence of faster cycling due to the "nano" reagent kit, the software upgrade, and the fact that low-quality samples such as blanks were not mixed into the libraries, though they were in Run B. (**b**) As predicted from simulation (**Supplementary Fig. 2b**), observed base diversity is much higher for our method, resulting in no base approaching 100% representation in each cycle. In contrast, the EMP method results in much lower diversity. (**c**) The percentage of bases above Q30 on a per-cycle basis demonstrates a faster drop in quality for the EMP method for both read 1 (cycle 1-250) and read 2 (cycle 251-500). (**d**) Q score heatmaps demonstrating the full distribution of Q scores per cycle.

Nature Methods: doi:10.1038/nmeth.2634

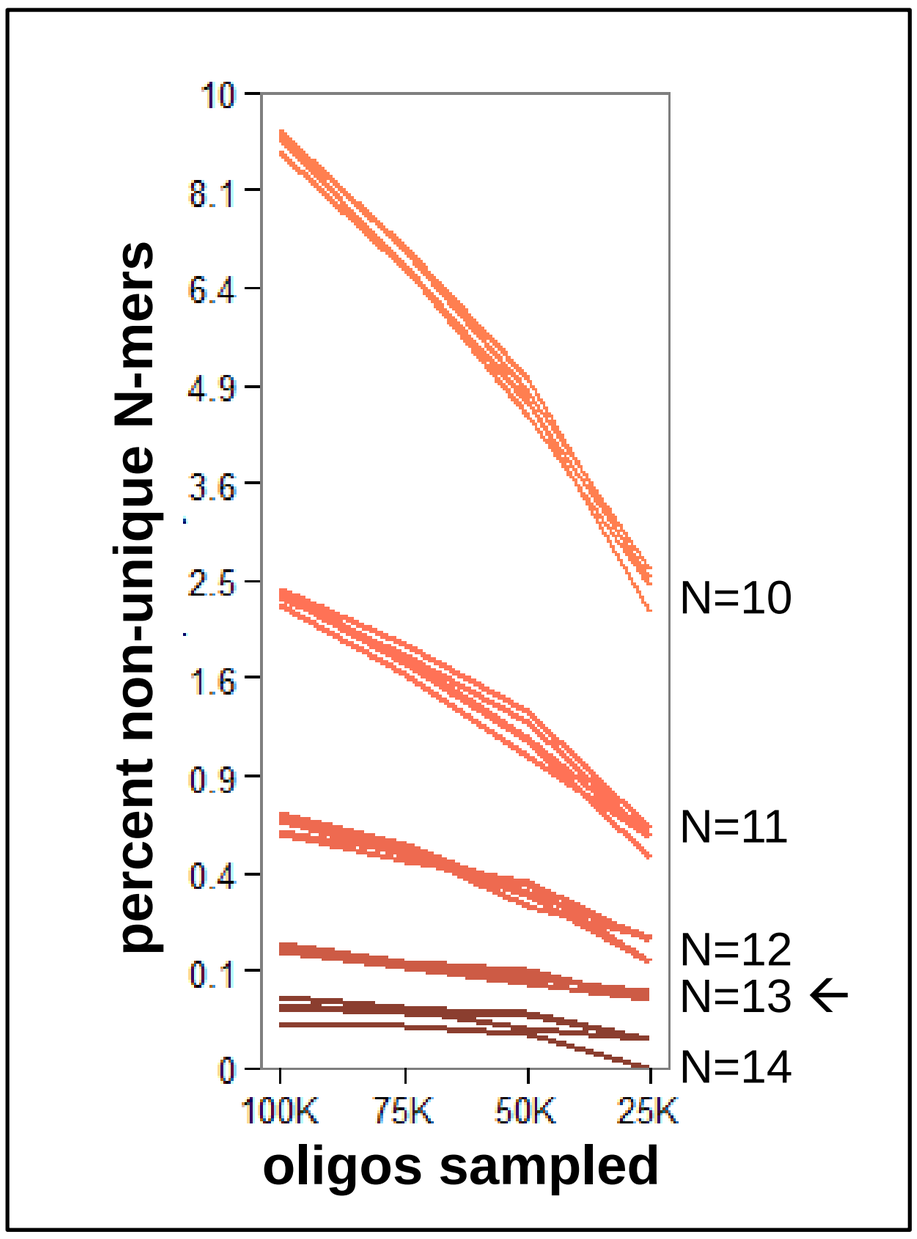
Supplementary Figure 6 | Template Tagging, PCR, sequencing, and Molecular Tag (MT) processing workflow.



Supplementary Figure 6 | Template Tagging, PCR, sequencing, and Molecular Tag (MT) processing workflow. Primer components colored as in **Supplementary Fig. 2**. (**a**) Template is tagged with reverse MT-FS primers using one extension cycle, and residual primer is removed. (**b**) The reverse-tagged template is tagged with forward MT-FS primers using one extension cycle, and residual primer is removed. (**c**) Dual-tagged template is amplified using universal primers that add sample barcodes. Residual primers are removed and samples are quantified and mixed to a final library. (**d**) Amplicons are sequenced in three reads. First, the 9 bp sample barcodes are read following priming with “Barcode_seq”. The 250 bp forward read is sequenced following priming with “Read1_seq”, and the 250 bp reverse read is sequenced following priming by “Read2_seq”. (**e**) All sequenced are de-multiplexed based on the “Barcode_seq” read which captures the sample barcode. For each sample, Read1 and Read2 are merged. (**f**) Regular expressions (**Supplementary Table 1g**) find all sequences in the set of merged sequences that match the expected patterns, and then extract the MT and template sequence from these pattern-matching sequences. (**g**) Sequences (colored lines) sharing the same molecular tag sequence (color) are grouped into the same MT category (each colored folder). (**h**) Sequences in the same MT category are aligned and a consensus sequence is built to represent that MT category. Singleton MT categories are kept in a separate file from consensus sequences.

Nature Methods: doi:10.1038/nmeth.2634

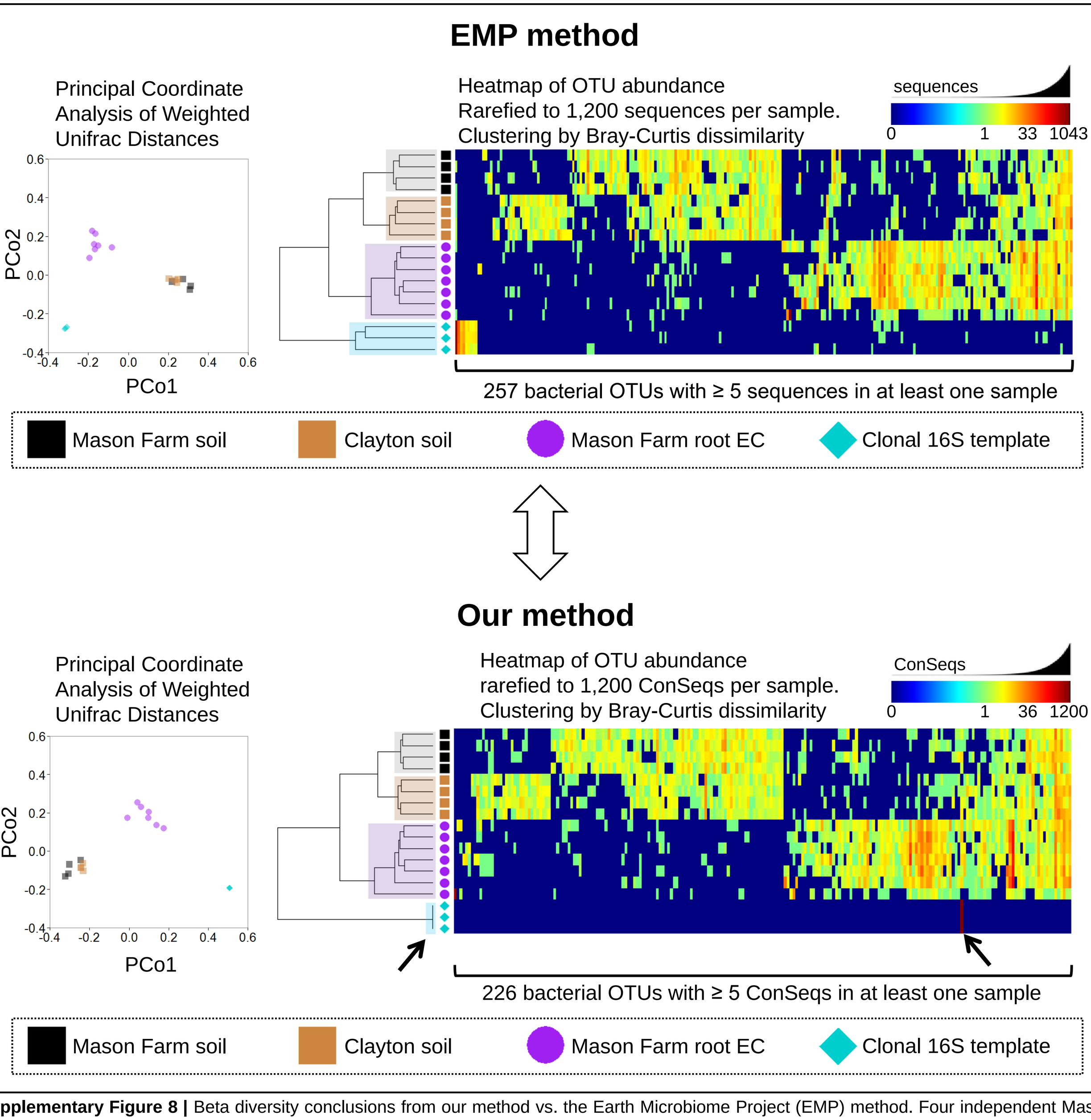
Supplementary Figure 7 | A MT of 13 random bases is sufficiently unique.



Supplementary Figure 7 | A MT of 13 random bases is sufficiently unique. Monte Carlo simulation at four sampling depths showing the percentage of non-unique oligonucleotide (A, C, T, or G) *N*-mers for *N*'s of 10, 11, 12, 13, and 14. The simulation was repeated 5 times (multiple lines within each hue). A randomer of *N* = 13 (second line from bottom) has about 140 non-unique oligos for every 100,000 sampled (~0.1%), which group into 70 duplicates. In the case of a template-overloaded sample sequenced to a depth of 100,000 reads or greater, these duplicate tags will lead to the unwanted classification of unrelated sequences as originating from the sample template. The consensus sequence made from the multiple sequence alignments will favor the overrepresented MT, often correcting the problem. Furthermore, each multiple sequence alignment can be assigned a quality score based on the average deviation of each sequence in the alignment from the consensus sequence for that alignment. Because multiple sequence alignments made from falsely-grouped independent templates will in general have worse alignment scores, these can be removed from the dataset by thresholding the worst alignments. Choice of random-mer length must be a balance between uniqueness on the one hand, versus costs in terms of sequence length and oligo chaos caused by longer lengths of *N*. It is more important to minimize non-unique *N*-mers than attempt to eliminate them; samples for which deep sequencing is needed can be multiplexed over several barcodes to increase depth, allowing unique molecular tagging without increasing random-mer length.

Nature Methods: doi:10.1038/nmeth.2634

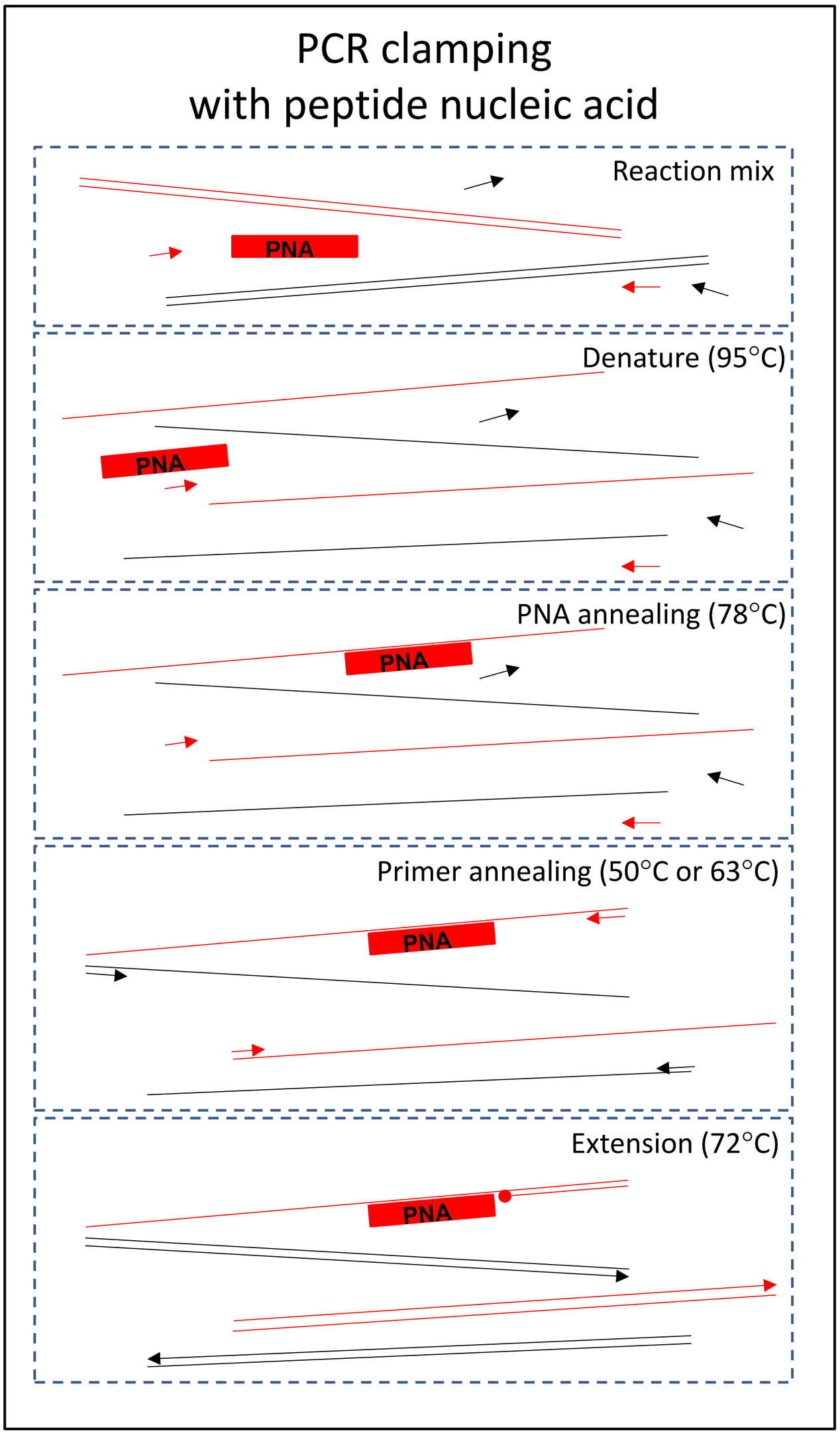
Supplementary Figure 8 | Beta diversity conclusions from our method vs. the Earth Microbiome Project (EMP) method.



Supplementary Figure 8 | Beta diversity conclusions from our method vs. the Earth Microbiome Project (EMP) method. Four independent Mason Farm soil samples (back squares), four indepdent Clayton soil samples (brown squares), seven Mason Form root endopyte compartment samples from separate plants (puple circles) and 3 technical PCR replicates of a cloned 16S template were each phylotyped using the EMP method (top) or our method (bottom) (**Supplementary Table 2b-d**, Online Methods). OTUs were formed at 97% identity and all samples were rarefed to 1,200 sequences or 1,200 ConSeqs. Principal coordinates analysis based on weighted unifrac distances (left) demonstrates that for both methods, the first two principal coordinates capture a similar separation of sample types. For heatmap visualization, the OTUs were thresholded such that only those OTUs contaning at least 5 sequences or ConSeqs in at least one sample are displayed. Heatmap rows and columns are ordered based on unsupervised clustering by Bray-Curtis dissimilarity. The heirarchical clustering results in the same separation of sample types as the Unifrac ordination for both methods, demonstrating that the same major beta-diversity conclusions can be reached with both methods. However, the ConSeqs from our method represent less noise, clearly evident from the single OTU formed for the clonal 16S template. In contrast, the EMP method produed produce several low-abundance OTUs from the clonal template, and 31 more OTUs overall using the same thresholding parameters (x-axis of heatmap, Online Methods).

Nature Methods: doi:10.1038/nmeth.2634

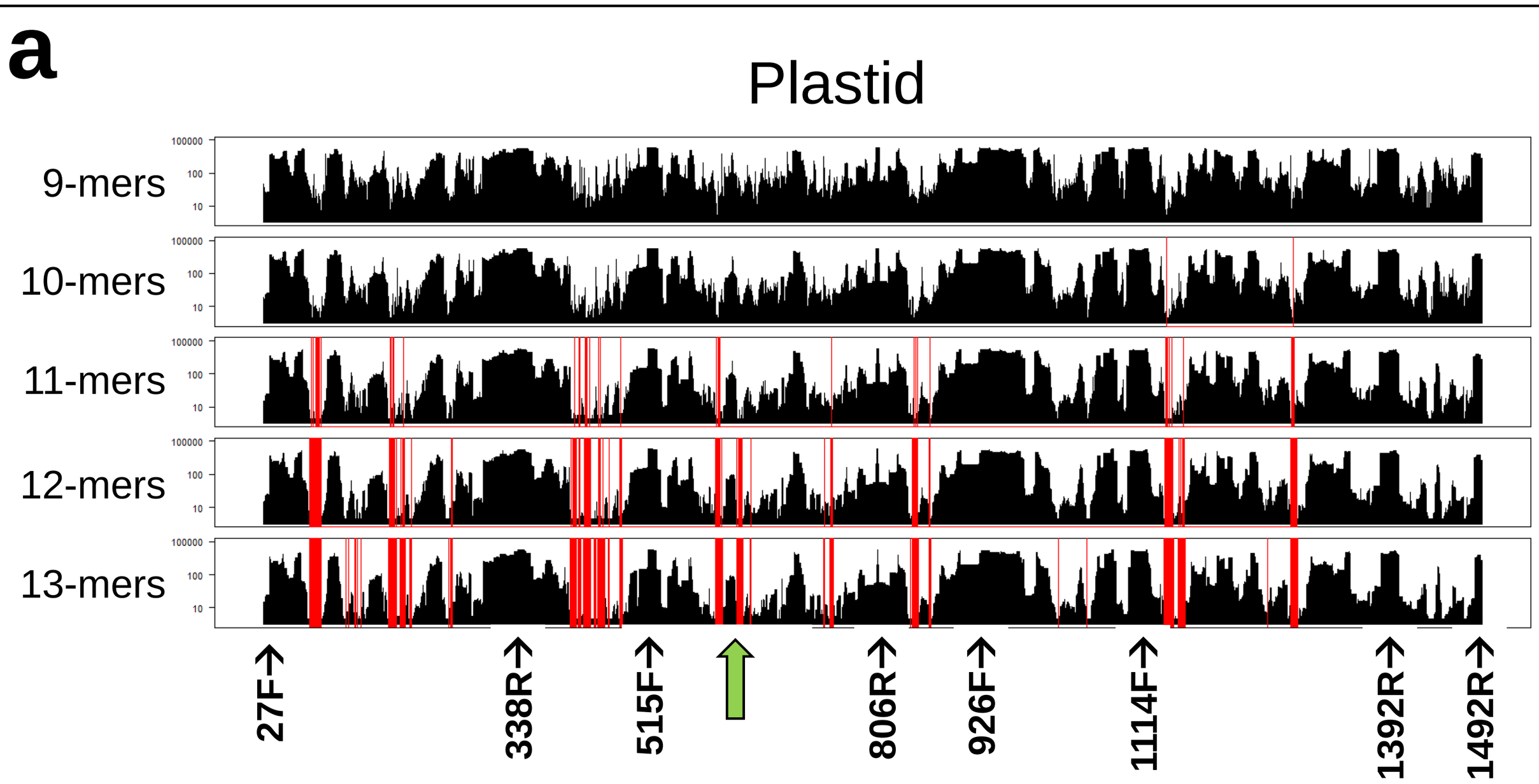
Supplementary Figure 9 | PNA schematic.

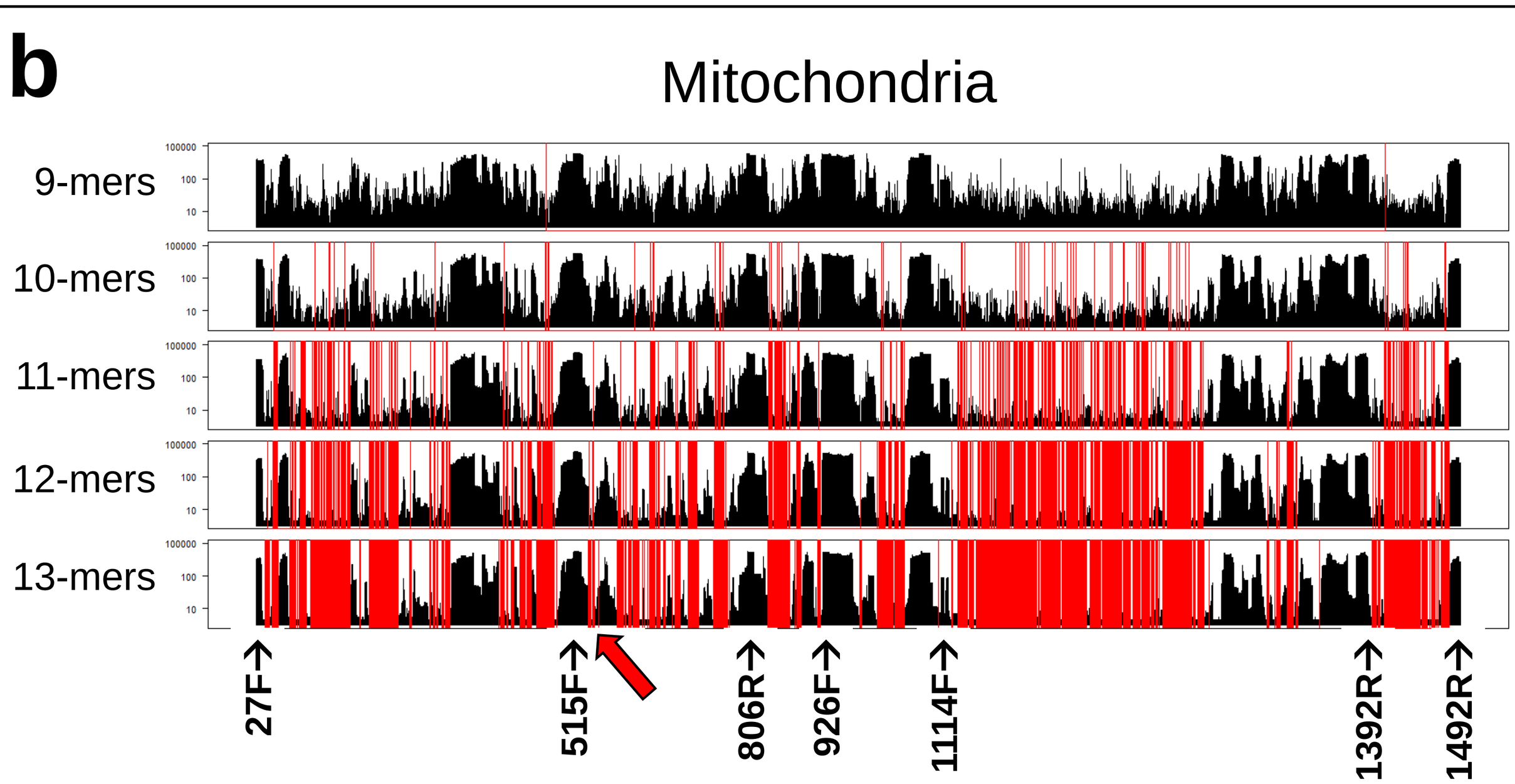


Supplementary Figure 9 | PNA schematic. PNA functions as an additive in the PCR reaction mix (top). After denaturation, PNA anneals specifically to templates via base pairing. As long as the PNA has a higher melting temperature than the primers, it anneals to template prior to the primers (middle). Depending on design, PNA either directly blocks primer annealing or blocks extension of the nascent strand.

Nature Methods: doi:10.1038/nmeth.2634

Supplmentary Figure 10 | Exhuastive search for PNA oligo candidates

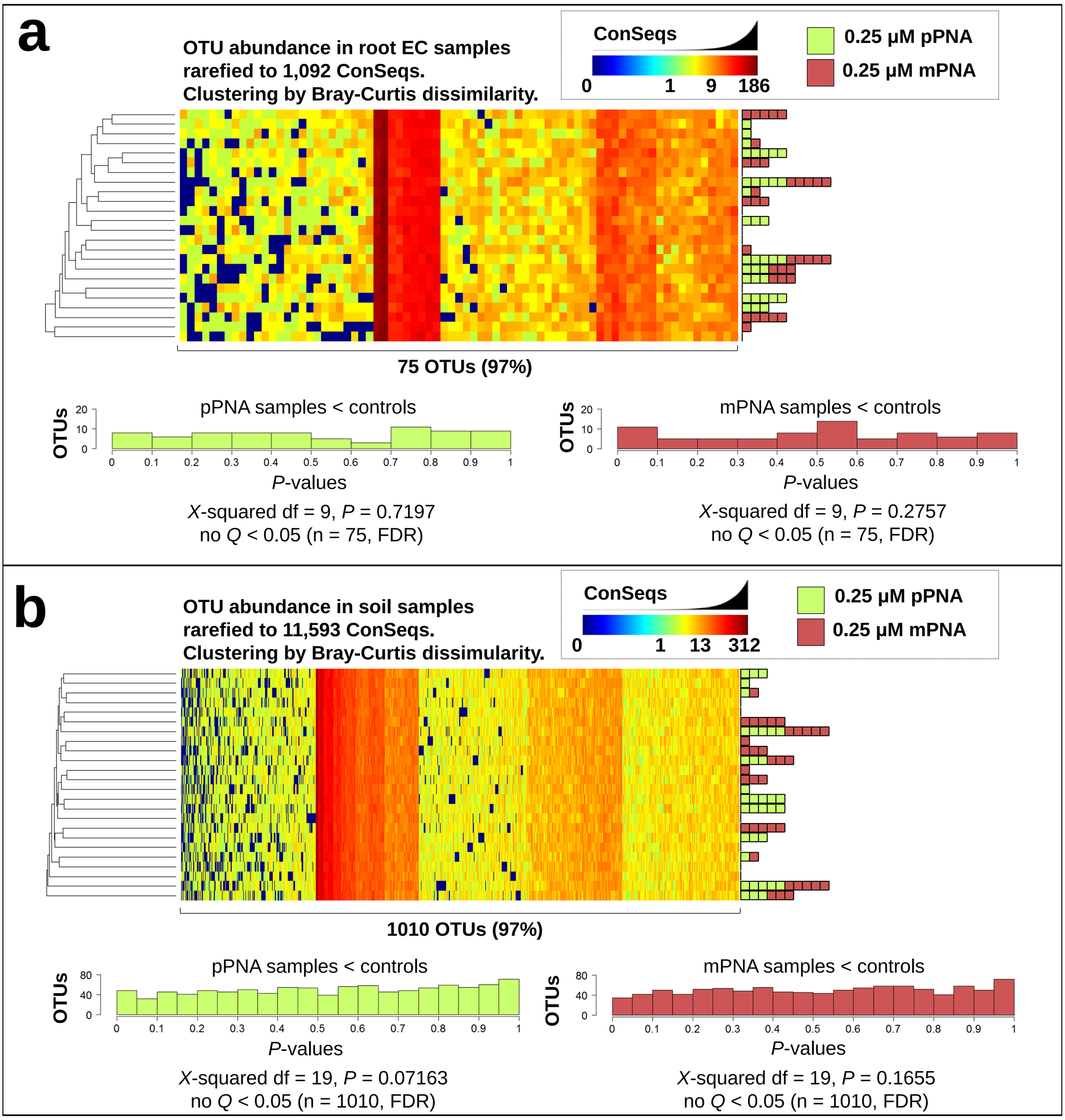




Supplmentary Figure 10 | Exhuastive search for PNA oligo candidates. (**a**) The full length chloroplast 16S rRNA sequence was split *in silico* into all possible 9-mers, 10-mers, 11-mers, 12-mers, and 13-mers. Each fragment was searched against the full length sequence for all sequences in the Greengenes 97% representatives microbial database, and the number of matches was graphed (black; log scale). Fragments of each length matching no sequences are marked with a red vertical line; these represent the best candidates for PCR clamping. The location of common 16S primers is shown beneath each histogram, and the location of the “pPNA” used in this study is shown with a green arrow. (**b**) As above, but for the mitochondrial 16S sequence. The location of the “mPNA” used in this study is shown with the red arrow.

Nature Methods: doi:10.1038/nmeth.2634

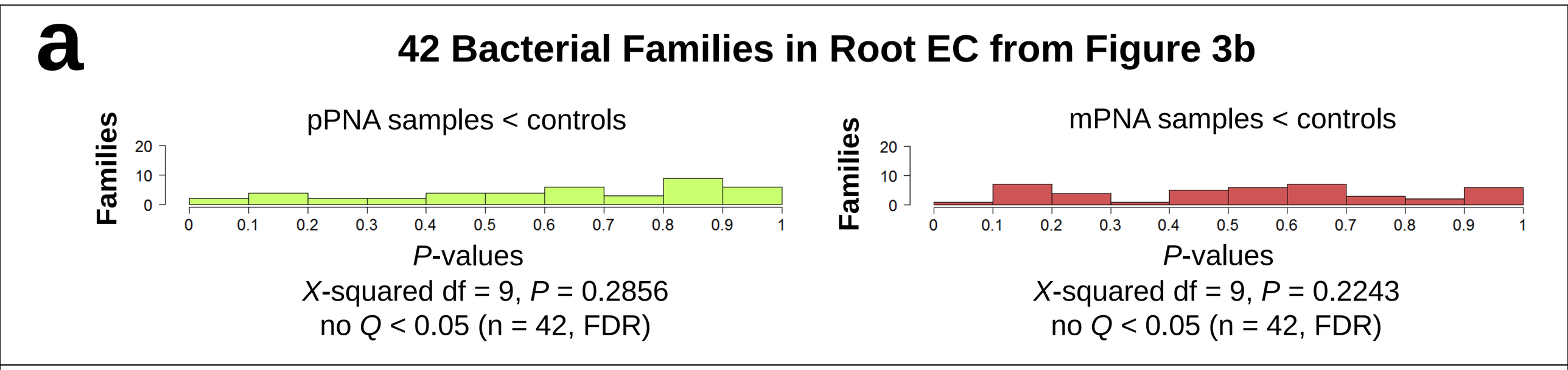
Supplementary Figure 11 | No bacterial OTU abundances are affected by pPNA or mPNA.



Supplementary Figure 11 | No bacterial OTU abundances are affected by pPNA or mPNA. (**a**) Root EC samples were clustered by the abundance of the 75 bacterial OTUs with ≥5 ConSeqs in at least one of the 24 samples. The heatmap shows the relative abundance of each OTU (columns) for each of the samples (rows) with the PNA doses shown (colored blocks). For each OTU, the 12 samples containing pPNA were tested for lower abundance than the 12 samples containing no PNA or only mPNA (left; green). Similarly, the 12 samples containing mPNA were tested for lower abundance than the 12 samples containing no PNA or only pPNA (right; red). *P*-values were obtained with a permutation test on the means using 10,000 permutations, and the *P*-value distribution was plotted across 10 bins (histograms). *P*-values were corrected for multiple testing with the FDR method; no OTUs were found significant. Each *P*-value distribution was shown not to deviate from the null flat distribution with a Chi-squared test (*P*-values for Chi-squared below histograms). (**b**) Same as in **a**, but for the 1,010 OTUs in soil samples with ≥5 ConSeqs in at least one of the 24 samples. Owing to the much greater number of OTUs the *P*-value distributions were plotted across 20 bins (histograms). The Chi-squared *P*-values, both for pPNA and mPNA comparisons, supported the null hypothesis of a flat distribution. *P*-values were corrected for multiple testing with the FDR method; limited OTUs in soil samples had significant *Q*-values (bold, red). Consistent with these statistics, there is no clustering (based on Bray-Curtis dissimilarity and group average linkage) by PNA treatment.

Nature Methods: doi:10.1038/nmeth.2634

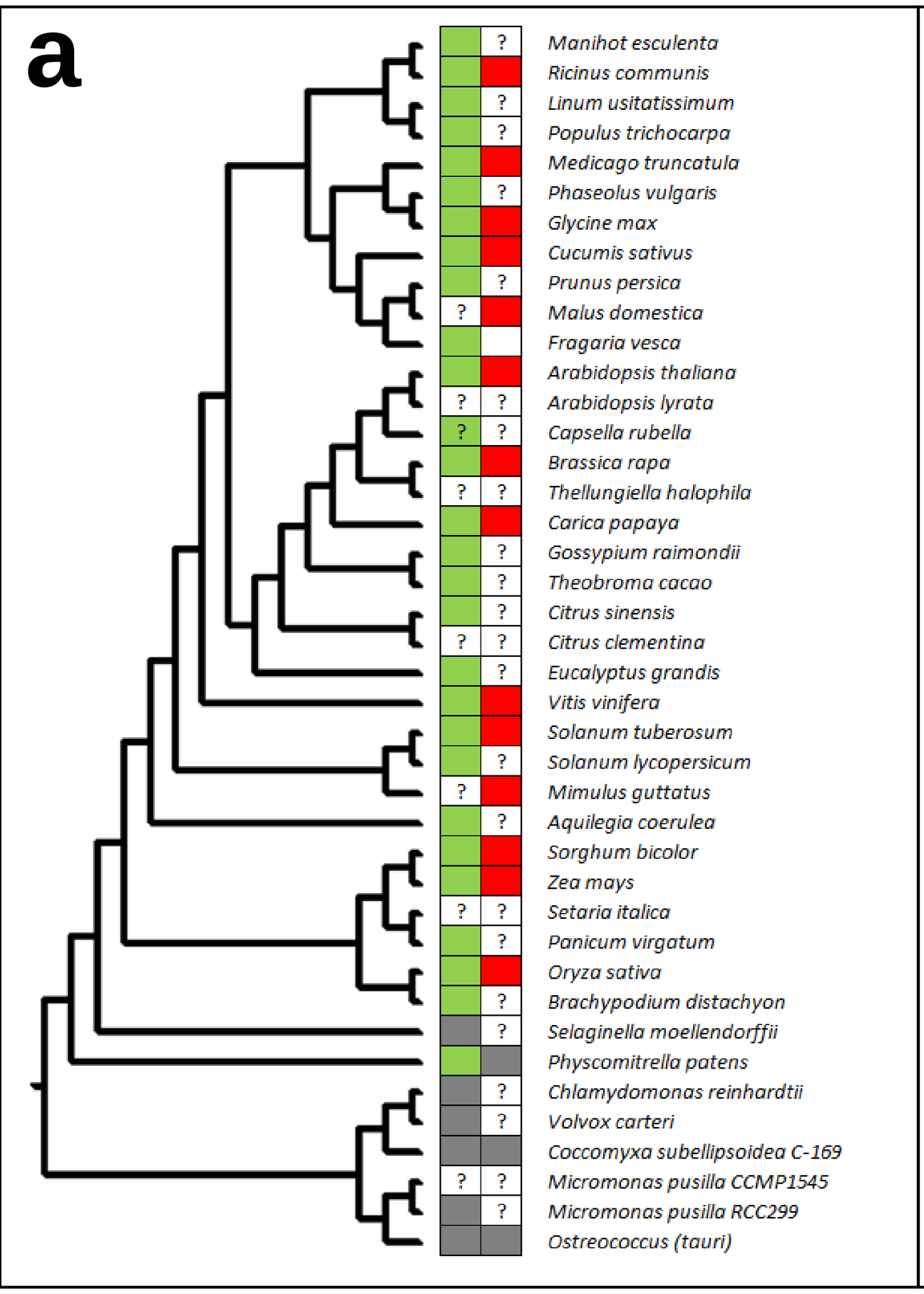
Supplementary Figure 12 | No bacterial family abundances are affected by pPNA or mPNA.

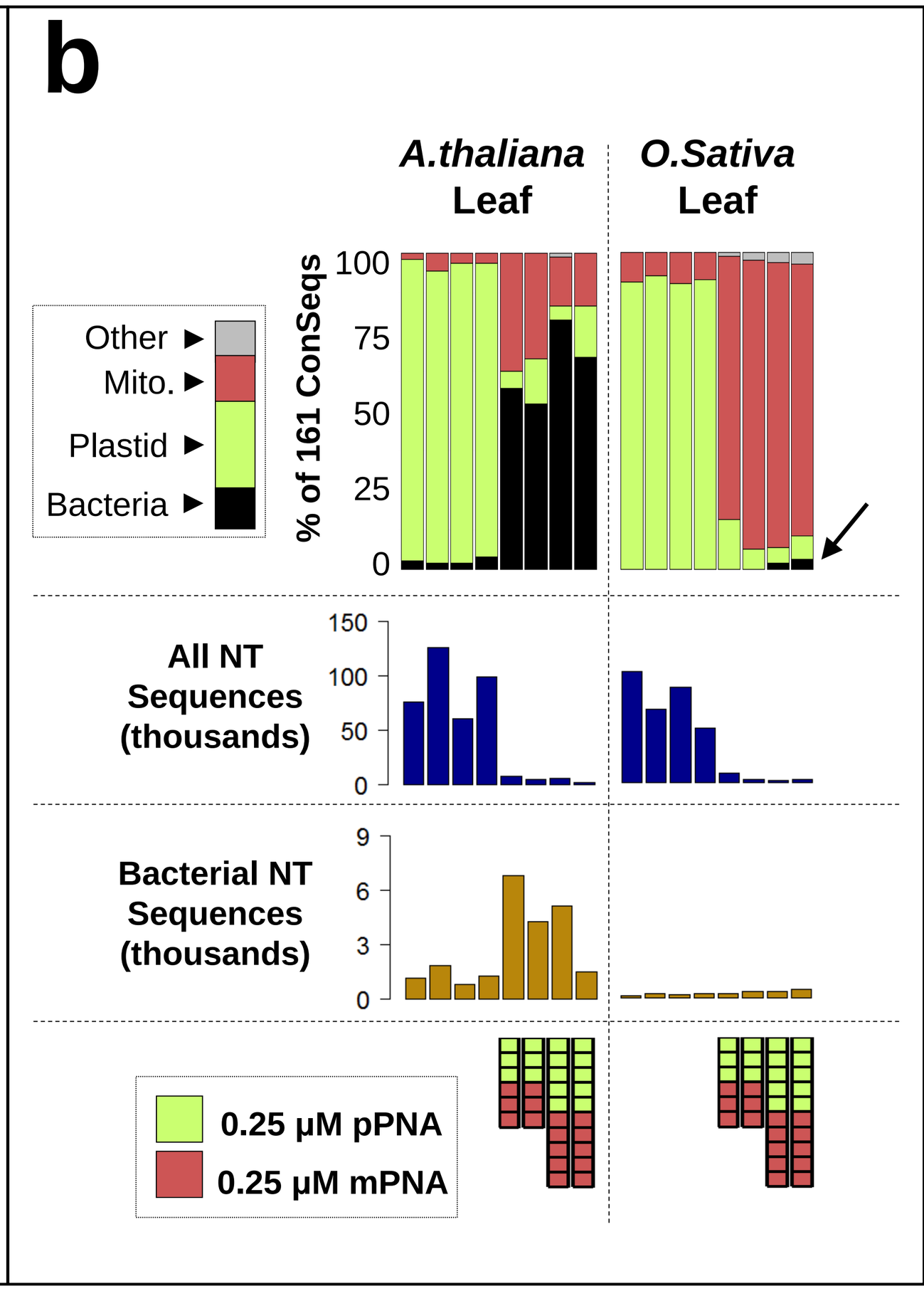


Supplementary Figure 12 | No bacterial family abundances are affected by pPNA or mPNA. (**a**) The abundance of each bacterial family with ≥5 ConSeqs in at least one of the 24 samples in different PNA treatments (**Fig. 3b**) was compared for root EC. For each bacterial family, the 12 samples containing pPNA were tested for lower abundance than the 12 samples containing no PNA or only mPNA (left; green). Similarly, the 12 samples containing mPNA were tested for lower abundance than the 12 samples containing no PNA or only pPNA (right; red). *P*-values were obtained with a permutation test on the means using 10,000 permutations, and the *P*-value distribution was plotted across 10 bins (histograms). The *P*-values were corrected for multiple testing with the FDR method; none of the resulting corrected *Q*-values were significant. Each *P*-value distribution was shown not to deviate from the null flat distribution with a Chi-squared test (*P*-values for Chi-squared below histograms). (**b**) Same as **a**, but analyzing bacterial families in soil (**Fig. 3c**). One *Q*-value for the mPNA test, corresponding to the family Bdellovibrionaceae, was significant. (**c**) The actual ConSeqs per soil sample are shown for family Bdellovibrionaceae in **b**, split by samples containing mPNA (left, hues of red) and control samples (right, white). Mean abundance in each sample group is shown with a horizontal black line. The mean abundance in the mPNA group is lower, but further addition of mPNA has no effect on the abundance (dark red, red, and light red, color legend). Rather, the sample in the mPNA group that is most similar to the mean of the control group has the highest mPNA dose (dark red), while the sample furthers from the mean of the control group has the lowest mPNA dose (light red), consistent with Bdellovibrionaceae being a false positive. (**d**) For each of 118 OTU representative sequences classified as family Bdellovibrionaceae in Run B, the 17 bp mPNA was aligned step-wise to every base position in the sequence, using forward, reverse, complemented, and reverse complemented orientations of the mPNA. The best identity score (maximum number of identical bases the OTU shared with the mPNA, x-axis) was recorded, and the number of Bdellovibrionaceae sequences at each identity score (y-axis) was plotted (red line). This was repeated for 118 independent OTU representative sequences from a mix of different families in the same phylogenetic class (Deltaproteobacteria, gray line), and a random sample of 118 sequences from bacteria from varied families (black line). The Bdellovibrionaceae do not show any higher identity to the mPNA than other bacterial groups, in contrast to what would be expected if their lower abundance were due to mPNA.

Nature Methods: doi:10.1038/nmeth.2634

Supplementary Figure 13 | Diverse plant species for which the PNAs in this study should block organelle 16S amplification.

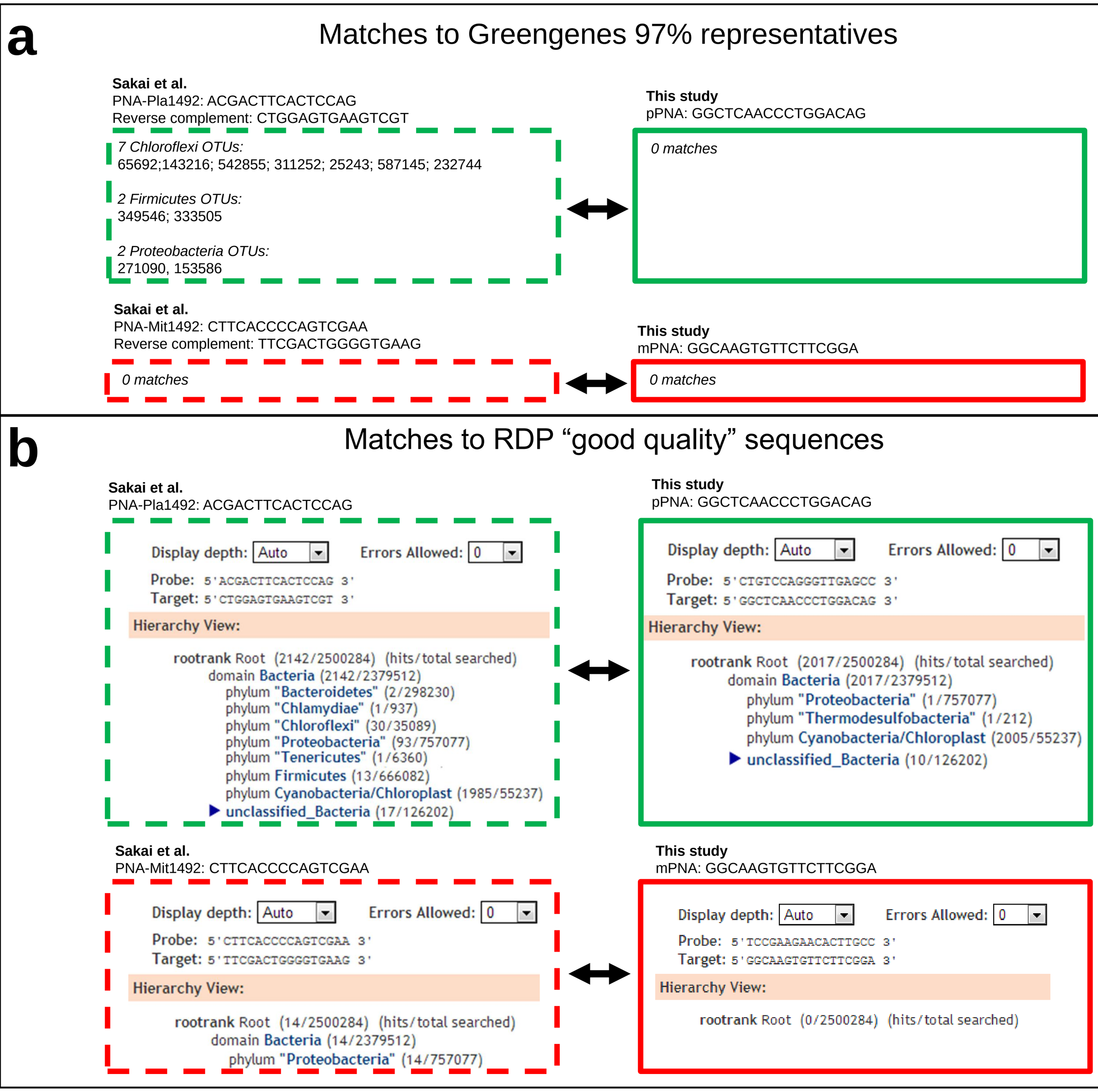




Supplementary Figure 13 | Diverse plant species for which the PNAs in this study should block organelle V4 16S amplification. (**a**) Diverse plant species for which the PNAs in this study should block organelle 16S amplification based on an exact sequence match. Phylogenetic tree and choice of plant taxa adapted from Phytozome v9.1 (http://www.phytozome.net/). Branch lengths are not meaningful. Plastid and mitochondrial organelle sequences for each plant in the phylogeny, or a relative in the same genus if the Phytozome species was not available, were collected from NCBI GenBank (**Supplementary Table 6**). The pPNA and mPNA sequences were queried against all collected plastid and mitochondrial sequences, respectively. Green squares represent exact matches of the pPNA to the plastid sequence; red squares represent exact matches of the mPNA to the mitochondrial sequence; grey squares represent a mismatch; white squares filled with "?" mean that the organelle sequence is not publicly available. (**b**) Leaf samples from *A.thaliana* (left) and *O.sativa* (right) (**Supplementary Table 2a,d**) were amplified with or without a mix of both pPNA and mPNA. Despite the extreme host contamination present in DNA from ground leaves (98.3% and 99;8% for *A.thaliana* and *O.sativa* respectively), addition of PNA increased the relative abundance of bacterial reads (top). Although the effect appears modest for *O.sativa*, the use of 1.25µM of both PNAs (arrow) represents a more than 20-fold increase in detectable templates. As with *A.thaliana* leaves, PNAs blocked the amplification of the majority of contaminant, and hence, template molecules of *O.sativa*, resulting in less sequenceable material (dark blue bars). However, more total bacterial sequences were nonetheless recovered (brown bars). These results are consistent with the PNAs functioning to block chloroplast and mitochondria, but not bacteria, in *O.sativa.*

Nature Methods: doi:10.1038/nmeth.2634

Supplementary Figure 14 | Predicted specificity of PNAs used in Sakai et al. vs. those used in this study.



Supplementary Figure 14 | Predicted specificity of PNAs used in Sakai et al. vs. those used in this study. (**a**) Summaries of searches to the Greengenes 97% database (most recent Feb. 4 2011 version) for anti-plastid PNAs (top) and anti-mitochondria PNAs (bottom), considering 29,556 non-chloroplast sequences. Sakai et al. PNAs[2] are shown in dotted boxes (left), while PNAs used in this study are shown in solid boxes (right). (**b**) Edited screenshots from RDP probe match (http://rdp.cme.msu.edu/probematch/search.jsp) showing all perfect matches to 2,500,284 “good quality” sequences for anti-plastid PNAs (top) and anti-mitochondria PNAs (bottom), in Sakai et al. (left) and this study (right). Phyla not matched are not displayed.

Nature Methods: doi:10.1038/nmeth.2634

Supplementary Figure 15 | Template tagging primer variants were evenly mixed and properly recovered.

a

All MT-FS V4 16S non-barcoded primers (as used in MiSeq Run B)

	Nextera Read1-annealing	*MT-FS*	*Lnk*	*515F*
515F_f1	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f2	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	T NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f3	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	CT NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f4	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	ACT NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f5	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	GACT NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f6	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	TGACT NNNN GA GTGCCAGCMGCCGCGGTAA→

	806R	*Lnk*	*MT-FS*	*TruSeq Read2-annealing*	
806R_f1	←TAATCTWTGGGVHCATCAGG	CA NNN	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f1
806R_f2	←TAATCTWTGGGVHCATCAGG	CA NNN T	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f2
806F_f3	←TAATCTWTGGGVHCATCAGG	CA NNN TC	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806F_f3
806R_f4	←TAATCTWTGGGVHCATCAGG	CA NNN TCA	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f4
806R_f5	←TAATCTWTGGGVHCATCAGG	CA NNN TCAG	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f5
806R_f6	←TAATCTWTGGGVHCATCAGG	CA NNN TCAGT	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f6

MT-FS group 1: V4F_1-3-5 with V4R_2-4-6

	Nextera Read1-annealing	*MT-FS*	*Lnk*	*515F*
515F_f1	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f3	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	CT NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f5	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	GACT NNNN GA GTGCCAGCMGCCGCGGTAA→

	806R	*Lnk*	*MT-FS*	*TruSeq Read2-annealing*	
806R_f2	←TAATCTWTGGGVHCATCAGG	CA NNN T	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f2
806R_f4	←TAATCTWTGGGVHCATCAGG	CA NNN TCA	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f4
806R_f6	←TAATCTWTGGGVHCATCAGG	CA NNN TCAGT	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f6

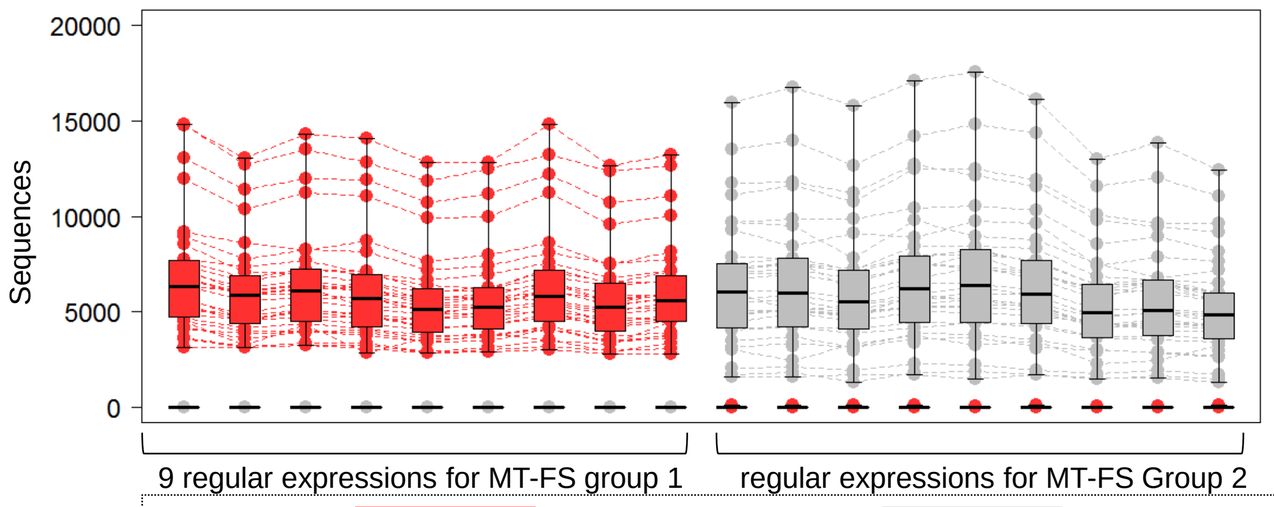
MT-FS group 2: V4F_2-4-6 with V4R_1-3-5

	Nextera Read1-annealing	*MT-FS*	*Lnk*	*515F*
515F_f2	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	T NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f4	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	ACT NNNN GA GTGCCAGCMGCCGCGGTAA→
515F_f6	GCCTCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	TGACT NNNN GA GTGCCAGCMGCCGCGGTAA→

	806R	*Lnk*	*MT-FS*	*TruSeq Read2-annealing*	
806R_f1	←TAATCTWTGGGVHCATCAGG	CA NNN	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f1
806F_f3	←TAATCTWTGGGVHCATCAGG	CA NNN TC	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806F_f3
806R_f5	←TAATCTWTGGGVHCATCAGG	CA NNN TCAG	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f5

b

Sequences from MiSeq Run B matching each potential combination of forward and reverse MT-FS V4 16S primers



From a sample tagged with MT-FS group 1 primers

From a sample tagged with MT-FS group 2 primers

99.86% **99.97%** Percent of regular expression-matching sequences coming from expected samples

c

Two barcoded mixes of Bc-MT-FS primers, each paired with the same reverse MF-FS mix (as used in MiSeq Run C)

Forward Bc-MT-FS mix with “TGA” barcode

	Nextera Read1-annealing	*MT-FS*	*BC*	*MT*	*Lnk*	*515F*
515F_TGA_f1	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	TGA	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_TGA_f2	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN T	TGA	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_TGA_f3	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN CT	TGA	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_TGA_f4	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN ACT	TGA	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_TGA_f5	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN GACT	TGA	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_TGA_f6	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN TGACT	TGA	NNNN GA	GTGCCAGCMGCCGCGGTAA→

Forward Bc-MT-FS mix with “ACT” barcode

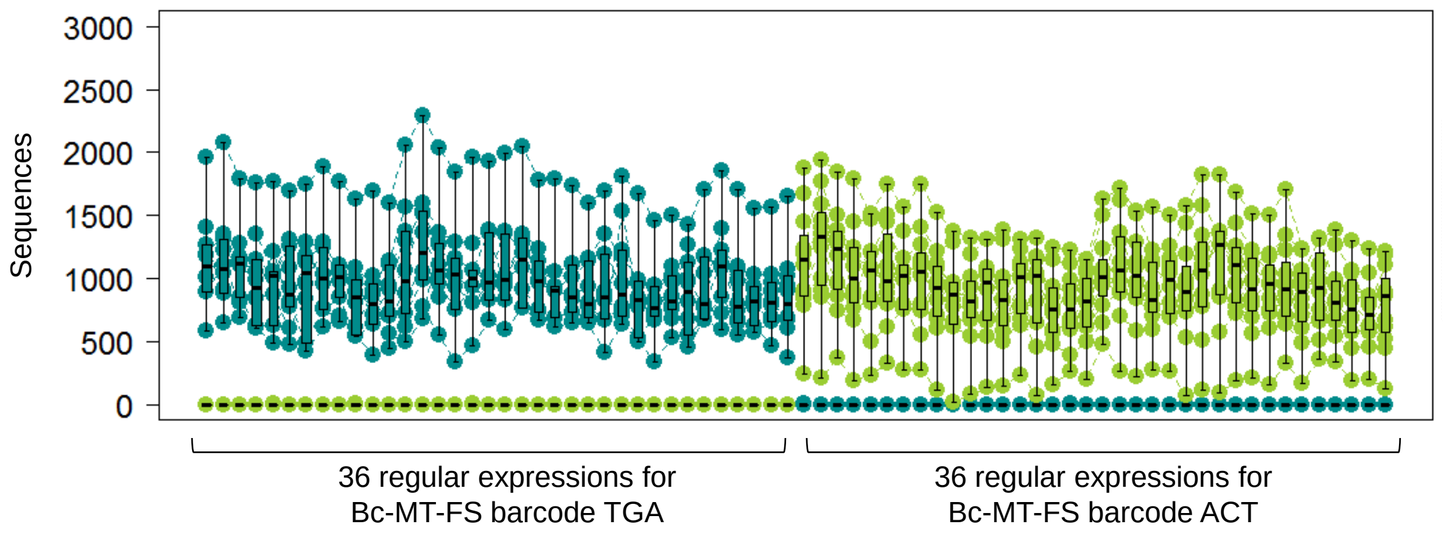
	Nextera Read1-annealing	*MT-FS*	*BC*	*MT*	*Lnk*	*515F*
515F_ACT_f1	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN	ACT	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_ACT_f2	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN T	ACT	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_ACT_f3	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN CT	ACT	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_ACT_f4	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN ACT	ACT	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_ACT_f5	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN GACT	ACT	NNNN GA	GTGCCAGCMGCCGCGGTAA→
515F_ACT_f6	TCCCTCGCGCCATCAGAGATGTG	TATAAGAGACAG	NNNN TGACT	ACT	NNNN GA	GTGCCAGCMGCCGCGGTAA→

Reverse MT-FS mix

	806R	*Lnk*	*MT-FS*	*TruSeq Read2-annealing*	
806R_f1	←TAATCTWTGGGVHCATCAGG	CA NNN	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f1
806R_f2	←TAATCTWTGGGVHCATCAGG	CA NNN T	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f2
806F_f3	←TAATCTWTGGGVHCATCAGG	CA NNN TC	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806F_f3
806R_f4	←TAATCTWTGGGVHCATCAGG	CA NNN TCA	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f4
806R_f5	←TAATCTWTGGGVHCATCAGG	CA NNN TCAG	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f5
806R_f6	←TAATCTWTGGGVHCATCAGG	CA NNN TCAGT	NN	TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG	806R_f6

d

Sequences from MiSeq Run C matching each potential combination of barcoded forward and reverse Bc-MT-FS V4 16S primers



From a sample tagged with “TGA” Bc-MT-FS mix

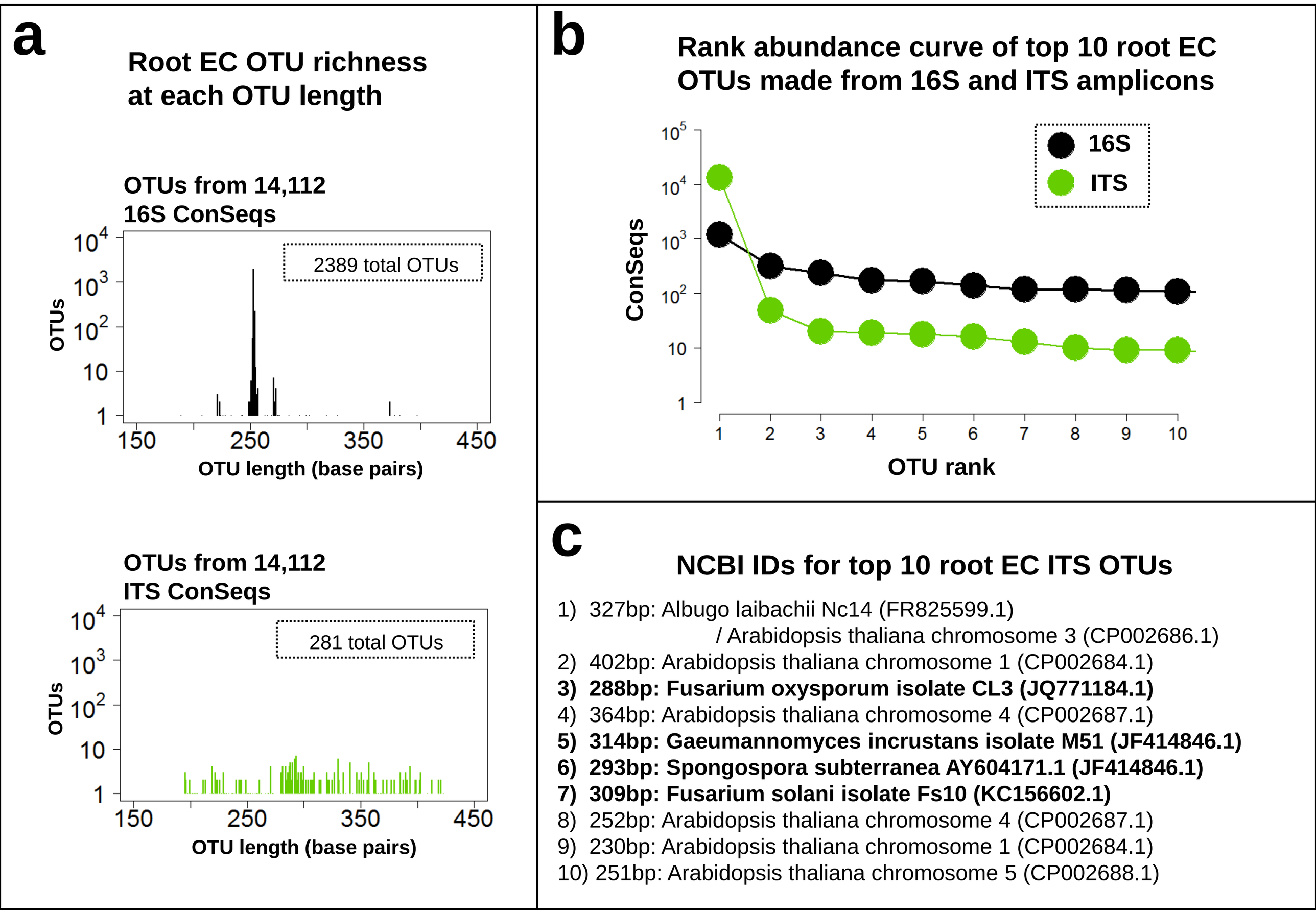
From a sample tagged with “ACT” Bc-MT-FS mix

99.97% **99.98%** Percent of regular expression-matching sequences coming from expected samples

Nature Methods: doi:10.1038/nmeth.2634

Supplementary Figure 15 | Template tagging primer variants were evenly mixed and properly recovered. (**a**) The full set of six forward and six reverse non-barcoded MT-FS V4 16S primers from **Fig. 1a,b** was subdivided into two groups called MT-FS group 1 (red) and MT-FS group 2 (gray). (**b**) Each independently-barcoded V4 16S sample in the run was tagged with a equimolar mix of primers from either MT-FS group 1 or MT-FS group 2 (**Supplementary Table 2a**, Online Methods); each MT-FS group can form 9 possible pairings of forward and reverse primers on the same molecule, for a total of 18 possible pairings. These 18 pairings can be recognized with regular expressions (x-axis, **Supplementary Table 1g**). For all samples containing 5000 or more pattern-matching sequences (**Supplementary Table 2g**), the number of sequences matching each regular expression is shown (y-axis) for MT-FS group 1 (red) and MT-FS group 2 (gray). Sequences from the same sample are connected with dotted lines; the flatness of the lines demonstrates even amplification and sequencing of each pairwise combination. As expected, sequences matching regular expressions for MT-FS group 1 came from samples originally amplified with primers from MT-FS group 1; the same is true for MT-FS group 2, demonstrating that groups of frameshifting primers are sufficient to distinguish samples and could serve as additional barcodes. (**c**) A mix of six forward Bc-MT-FS V4 16S primers with barcode “TGA” (dark blue) *or* a mix of six forward primers with barcode “ACT” (light green) was paired with the six reverse MT-FS V4 16S primers to make two barcoded groups. (**d**) Same as **b**, except that samples containing 1000 or more pattern-matching sequences (**Supplementary Table 2g**) were included, and regular expressions were used to match barcode groups rather than frameshift groups. For each pairing of six mixed forward with six mixed reverse primers, 36 pairings are possible. Although the frameshift groups in **b** performed similarly to barcode groups in **d** in terms of percent of reads correctly matched, barcodes are more robust because single base deletions are common primer synthesis errors.

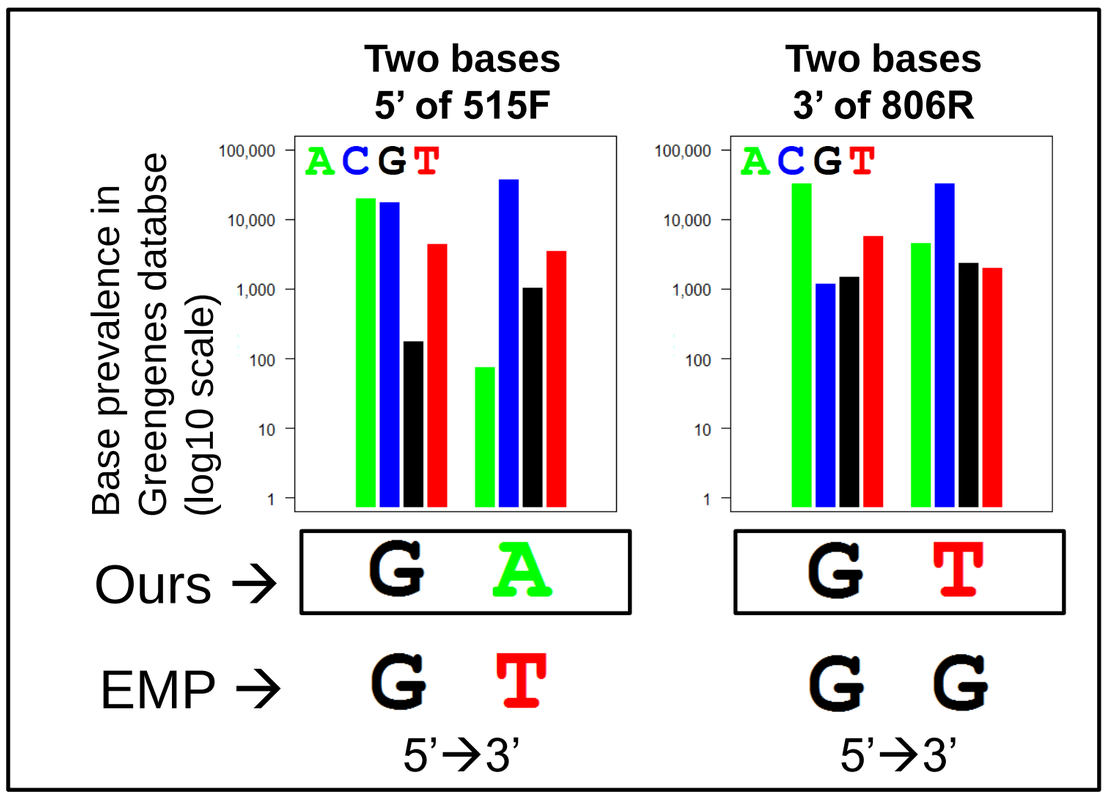
Supplementary Figure 16 | Universal PCR primers can be used to amplify and barcode other tagged templates.



Supplementary Figure 16 | Universal PCR primers can be used to amplify and barcode other tagged templates. (**a**) Root EC DNA was tagged with either V4 16S MT-FS primers or ITS2 MT primers (**Supplementary Table 1a,b**). Tagged template was amplified with universal PCR primers, sequenced, and MTs were used to form ConSeqs. For 16S (top, black) and ITS (bottom, green), the OTUs present among 14,112 ConSeqs were classified by their sequence length (x-axis), and the number of OTUs present at each length was plotted (y-axis). The total number of OTUs for each amplicon is inlaid in each plot. Although there were more V4 16S OTUs, the distribution of amplicon lengths is much narrower than for ITS. (**b**) The OTUs of 16S ConSeqs (black) and ITS ConSeqs (green) were ranked by their relative abundance and the number of sequences (log y-axis) is shown for the 10 most-abundant OTUs (x-axis). (**c**) The ITS OTUs shown in **b** were queried against the NCBI database using BLAST and the OTU length in base pairs and the best-scoring hit is shown. Several Arabidopsis OTUs demonstrate host contamination, but other eukaryotic and fungal OTUs are clearly present.

Nature Methods: doi:10.1038/nmeth.2634

Supplementary Figure 17 | Primer linkers.



Supplementary Figure 17 | Primer linkers. Our linkers differ from those used by the Earth Microbiome Project[1]. Ideal linkers should lack identity to the majority of microbial sequences in order to buffer the other elements of the template-tagging primer from the template. Our choices are equally or more divergent from sequences in the Greengenes database than are the EMP primers.

Supplementary Figure References

1. Caporaso, J.G. et al. ISME J 6, 1621-1624 (2012).
2. Sakai, M. & Ikenaga, M. Journal of Microbiological Methods 92, 281-288 (2013).

Supplementary Note:

Additional figure-specific analysis details

- Figure 1a: Copy number per MT for clonal 16S samples
- Figure 1b and Supplementary Figure 4c: Error rate for clonal 16S samples
- Figure 1c: OTU analysis of clonal 16S samples
- Figure 2a: Rarefaction curves of different MT treatments
- Figure 2b: Progressive drop-out analysis of technical reproducibility
- Figure 3a, left: Relative abundance of contaminant sequences
- Figure 3a, right: Mean number of sequences per multiple sequence alignment
- Figure 3b and 3c, and Supplementary Figures 8 and 11: Heatmaps
- Supplementary Figure 1: Variable regions in the 16S gene
- Supplementary Figure 3b: Library diversity simulation
- Supplementary Figure 4a-b, 5a-d: Q Score histograms, plots, and heatmaps, and base diversity per cycle
- Supplementary Figure 7: Monte Carlo simulation of MT uniqueness
- Supplementary Figure 8: Principal Coordinate Analysis of Weighted Unifrac distances
- Supplementary Figure 10: PNA design
- Supplementary Figures 11 and 12: Bacterial family and OTU relative abundance for different PNA treatments
- Supplementary Figure 13: Use of PNA on *A. thaliana* and *O. sativa* leaf DNA
- Supplementary Figure 15: Mixing and pattern-recognition of template-tagging primers
- Supplementary Figure 16: Internal Transcribed Spacer (ITS) amplicons
- Supplementary Figure 17: Primer linkers

Figure 1a: Copy number per MT for clonal 16S samples

Pattern-matching sequences from Run B of all the clonal 16S template samples, including both replicates of the no dilution, 50× dilution, and 100× dilution samples (**Supplementary Table 2a,d**) were rarefied to 40,000 sequences per sample. The sequences were then categorized by their MT (as in **Supplementary Fig. 6g**), but were not made into ConSeqs. A histogram was plotted of the number of sequences falling at each discrete MT category depth. The chart was produced with the geom_density() and geom_line() functions in the "ggplot2" library of R[1].

Figure 1b: Error rate for clonal 16S samples

Pattern-matching sequences of the clonal 16S template samples from only the 50× dilution and 100× dilution samples, as well as their replicates for a total of four samples (**Supplementary Table 2a,d**) were gathered from Run B. The MT and amplified template sequences were extracted (as in **Supplementary Fig. 6f**). The sequences were processed four ways to form four comparison groups:

1) *"NT", or no tag*, contained a mix of all pattern-matching sequences from all four samples, regardless of the MT.

2) *"ConSeqs", or ConSeqs of two or more sequences*, in which pattern-matching sequences in each sample were categorized by their MT and ConSeqs were constructed from the multiple sequence alignments. All ConSeqs made from MT categories containing 2 or more sequences were pulled from each sample and pooled.

3) *"S", or singletons*, in which pattern-matching sequences in each sample were categorized by their MT, and all the sequences with a unique single-copy molecular tag were pulled from each sample and pooled.

4) *"PConSeqs", or perfect ConSeqs made from three or more sequences*, in which pattern-matching sequences in each sample were categorized by their MT and ConSeqs were constructed from the multiple sequence alignments as described above. Only alignments of three or more sequences in which all constituent sequences were 100% identical were considered, and the ConSeqs (in this case PConSeqs) of these perfect alignments were pulled from each sample and pooled.

The four comparison groups were then each rarefied to 15,000 sequences, with the exception of the PConSeqs, which were a rarer class and used in full because only 4,777 were available in the run. The sequences were all aligned to a common set of pre-aligned templates using PyNAST, with default parameters as implemented in the QIIME script "align_seqs.py". The Sanger sequence of the clonal 16S template (sequence below), trimmed to the region between the 515F and 806R primers, was also aligned using PyNAST.

>Mycobacterium_16S_clone
TACGTAGGGTCCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCGCGTTGTTC
GTGAAAACTCACAGCTTAACTGTGGGCGTGCGGGCGATACGGGCAGACTTGAGTACTGCAGGGGAGACTG
GAATTCCTGGTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGG
CAGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGG

For each aligned comparison group, positions that were gaps in all aligned sequences *and* the Sanger reference sequence were not considered. In every case, gaps in the PyNAST comparison group alignments matched gaps in the Sanger alignment, indicating that the frequency of insertion errors in this sequence was extremely low. Next, for each base in the aligned Sanger sequence, the SNPs and gaps for all other sequences in the 15,000 sequences (or 4,777 for PConSeqs) of the comparison group were counted. This value was divided by 15 (or 4.777 for PConSeqs) to generate the errors per thousand (ept) at each base. The mean error rates per thousand were calculated by taking the mean of the per-base errors per thousand across each of 253 bases of the sequence. The chart was produced with the geom_line() function in the "ggplot2" library of R[1].

Figure 1c: OTU analysis of clonal 16S samples

Pattern-matching sequences from Run B of the clonal 16S template samples from only the 50× dilution and 100× dilution samples (**Supplementary Table 2a,d**) were unprocessed (NT) or processed into ConSeqs (ConSeqs). Each comparison group was rarefied to 30,000 sequences per sample and these sequences were clustered into OTUs at 97% or 99% identity. The OTUs were ordered by their relative abundance, and the number of sequences in each ranked OTUs is graphed for each comparison group. To determine the number of OTUs necessary to represent 95% of the data, the sequences in the OTUs were summed, starting with the most abundant, until 28,500 sequences (95% of 30,000) were accounted for. The chart was produced with the geom_line() and geom_point() functions in the "ggplot2" library of R[1].

Figure 2a: Rarefaction curves of different MT treatments

Pattern-matching sequences were gathered from Run B and the MT and amplified template sequences were extracted (as in **Supplementary Fig. 6f**). The sequences were processed four ways to form four comparison groups:

Comparison groups:
1) *"NT", or no tag*, as described for **Fig. 1b** and **Supplementary Fig. 4c**.

2) *"ConSeqs", or ConSeqs of two or more sequences*, as described for **Fig. 1b**.

3) *"S", or singletons*, as described for **Fig. 1b**.

4) "CAS", *or ConSeqs with adjusted singletons*. For each sample, sequences were categorized by their MT, and ConSeqs were constructed from the multiple sequence alignments. The *ConSeqs collapse ratio*, or the number of ConSeqs divided by the number of constituent sequences in the multiple sequence alignments, was calculated. Next, the singletons were quantified. The number of singletons was multiplied by the *ConSeqs collapse ratio*, rounded to the nearest integer, and then the singles were down-sampled to this integer. This adjustment thus keeps the ratio of singles to all other sequences constant, even as all other sequences are collapsed into their ConSeqs.

OTUs tables were formed from each comparison group, using 97% and 99% identity thresholds for clustering. Because the number of sequences in each comparison group varied substantially, with the NT group having many more sequences than the other groups, the FASTA files containing sequences from each comparison group were each normalized to 500,000 sequences. Each comparison group was then clustered independently at 97% or 99% sequence identity to produce 4 OTU tables. Plastid and mitochondrial OTUs were removed computationally, and bacterial reads for root EC and soil samples (**Supplementary Table 2a,d**) across all tables were pooled, producing a soil pool and a root EC pool per OTU table. These pools were rarefied at intervals of 1,000 sequences and the number of OTUs observed at each depth was plotted. The chart was produced with the geom_line() and geom_point() functions in the "ggplot2" library of R [1].

Figure 2b: Progressive drop-out analysis of technical reproducibility

The same four OTU tables representing the four comparison groups were used as in Figure 2a, with four exceptions. First, plastid and mitochondrial OTUs were *not* removed computationally. Second, in each OTU table, we considered the technical reproducibility of 12 pairs of root EC samples and 12 pairs of soil samples, for a total of 24 pairs, where each member of a pair was independently template-tagged, treated with water or PNA, and amplified (**Supplementary Table 2a,d**). These 24 pairs were chosen because these samples had good sequencing depth and reasonably diverse microbial composition. Third, each sample was rarefied to a common inter-table depth. Fourth, within each table, the more deeply-sequenced pair member for each of the 24 technical replicate pairs was rarefied to the number of sequences of the less-sequenced sample in the pair, such that the sequencing depth of the pair members was equal.

For each comparison group, the relative abundance of each OTU in one technical replicate pair member was log10-transformed to correct for heteroscedasticity and plotted against the log10-transformed relative abundance of that same OTUs in the other technical replicate pair member. This was repeated for all 24 pairs on the same set of axis, generating a densely-populated linearly-correlated scatterplot for each comparison group, similar to that previously published [2]. The R^2 coefficient of determination was then calculated for the scatterplot and graphed.

The low-abundance OTUs in an OTU table either represent rare but real sequences, or sequence errors, and are less-reproducible than larger OTUs[2]. We dropped OTUs from the scatterplot that did not meet the threshold abundance (x-axis) in at least one pair member in at

least one of the 24 pairs, and recalculated R^2 at each threshold, generating the upward-sloping curves. The chart was produced with the geom_line() and geom_point() functions in the "ggplot2" library of R [1].

Figure 3a, left: Relative abundance of contaminant sequences

Pattern-matching sequences in Run B were processed into ConSeqs which were clustered at 97% identity to form an OTU table. The twelve root EC samples with the PNA titrations and their technical replicates (**Supplementary Table 2a,d**) were extracted from this OTU table and rarefied to the smallest sample of the 24 (6,880 sequences). The relative abundance of bacterial sequences, plastid sequences, mitochondrial sequences, and other sequences were expressed as a percentage. The stacked bar chart was produced with the geom_bar() in the "ggplot2" library of R [1].

Figure 3a, right: Mean number of sequences per multiple sequence alignment

Pattern-matching sequences from root EC and soil samples, the same used in Figure 3a, left, (**Supplementary Table 2a,d**) were gathered from Run B and the MT and amplified template sequences were extracted (as in **Supplementary Fig. 6f**). The sequences were processed into ConSeqs, but just prior to formation of the ConSeqs from the multiple sequence alignments, the number of sequences in all multiple sequence alignments was counted. The average number of sequences per multiple sequence alignment per sample is graphed. The bar chart was produced with the geom_bar() in the "ggplot2" library of R [1].

Figure 3b and 3c, and Supplementary Figures 8 and 11: Heatmaps

Pattern-matching sequences from Run B were processed into ConSeqs, which were clustered at 97% identity to form an OTU table. All contaminant OTUs were removed, leaving only bacterial OTUs. For root EC heatmaps, 12 root EC samples and their technical replicates (**Supplementary Table 2a,d**) were extracted from this OTU table and rarefied to the smallest sample of the 24 (1,092 bacterial ConSeqs). For soil heatmaps, 12 soil samples and their technical replicates (**Supplementary Table 2a,d**) were extracted from this OTU table and rarefied to the smallest sample of the 24 (11,593 bacterial ConSeqs). For **Fig. 3b** and **3c**, the bacterial OTUs in each table were then reclassified at the family level, and OTUs from the same bacterial family were combined to convert the OTU table into a family-level table. Bacterial families that did not have an abundance of 5 ConSeqs in at least one of the 24 samples were removed to avoid visualizing rare families prone to sampling artifacts. For **Supplementary Fig. 11**, the bacterial OTUs were *not* reclassified at the family level, and OTUs that did not have an abundance of 5 ConSeqs in at least one of the 24 samples were removed. For better visualization in all heatmaps, abundances were transformed to log2 *per mille* $\log_2(1000x+1)$ prior to color assignment – this transformation is reflected in the color key. *The log2 transformation was for visualization only and transformed data was not used for statistical tests.* All heatmaps were made using the function heatmap.2() from the "gplots" library of R [3]. Hierarchical clustering of rows and columns in the heatmaps is based on Bray-Curtis dissimilarity and uses group-average linkage.

Supplementary Figure 1: Variable regions in the 16S gene

All sequences without unambiguous bases in the Greengenes training set of full length 16S sequences (29,846 sequences) were aligned to the pre-aligned Greengenes core set using PyNAST with default parameters as implemented in the QIIME script "align_seqs.py". The *E. coli* 16S sequence (PMID: CP002967.1) was also aligned. For each base position (non-gap position) in the *E. coli* alignment, the number of A, C, T, G, and gap characters in the corresponding position of all sequences in the Greengenes alignment was counted and the

Shannon diversity for this position was calculated and graphed (light blue vertical needles). The moving average of the Shannon diversity (thick waving black line in was calculated by taking the mean Shannon diversity at each *E. coli* base position considering a 50 bp sliding window stretching 25 bases 5' and 25 bases 3' of the base position considered – for this reason the black line is not graphed for the first 25 and the last 25 bases of the alignment. We note that this interpretation does not show the Shannon diversity at positions in the alignment for which the *E. coli* sequence shows a gap. Location of the hypervariable regions was based on mapping information in Chakravorty et al., 2007[4]. Degenerate regular expressions of common primers (**Supplementary Table 1g**) were used to map primer locations to the *E.coli* reference on the x-axis. The charts were produced with the plot() and points() functions in the "base" library of R [5].

Supplementary Figure 3b: Library diversity simulation

We simulated in *silico* a PCR template composed of 1,000 identical copies of a single V4 16S sequence, as well as a more realistic template composed of 1,000 real bacterial V4 16S sequences from a root EC sample. To mimic the effect of using frameshifting primers to PCR each template, subsets of the 1,000 sequences were randomly assigned to equally-sized groups to which six frameshifting treatments of 0-5 additional 5' bases were applied. To visualize the effect of mixing in phiX174 genomic DNA post-PCR, the phiX174 genome [NCBI GenBank ID: NC_001422] was randomly fragmented and the fragments were used to replace specific fractions of the 1,000 V4 16S sequences in the simulated PCRs. For each treatment of frameshifts and / or phiX174, the first 250 bp of each sequence was considered. Shannon diversity at each base position was calculated from the number of A, C, T, and G bases present at that position. The charts were produced with the boxplot() and points() functions in the "base" library of R [5].

Supplementary Figure 4a-b, 5a-d: Q Score histograms, plots, and heatmaps, and base diversity per cycle

Illumina Q scores are equivalent to 10 times the log10 of the reciprocal of the error rate. Q score histograms, plots, and heatmaps, and the graph of % base at each cycle, were generated from raw data on the MiSeq machine using Sequence Analysis Viewer version 1.8.11

Supplementary Figure 4c: Error rate for clonal 16S samples

Identical to Figure 1b, except that only reads from Run A and Run B processed by method 1, (NT), were used.

Supplementary Figure 7: Monte Carlo simulation of MT uniqueness

A custom R script was written to generate 100,000 oligonucleotide (A, C, T, or G) *N*-mers each for N's of 10, 11, 12, 13, and 14. For each *N-mer* length, the number of non-unique oligos in the set was divided by 100,000 to give the fraction of non-unique oligos, and then multiplied by 100 to give the percentage that is graphed. This process was also repeated for depths of 75,000, 50,000, and 25,000 *N*-mers. The entire simulation was then repeated 4 additional times, and all 5 replicates for each *N*-mer length were graphed. The chart was produced with the geom_line() function in the "ggplot2" library of R[1].

Supplementary Figure 8: Principal Coordinate Analysis of Weighted Unifrac distances

ConSeqs from our method in Run C, or high quality sequences from the EMP method in Run D, were clustered into OTUs with OTUpipe as described above using a 97% identity threshold, forming a separate OTU table for each run. Each sample in the OTU table from our method was rarefied to 1,200 ConSeqs, while each sample in the OTU table from the EMP method was

rarefied to 1,200 high quality sequences. For each run, the OTU representative sequences were aligned to a common set of pre-aligned templates using PyNAST, with default parameters as implemented in the QIIME script "align_seqs.py". The full alignments were then filtered and clustered into phylogenetic trees using the QIIME script "filter_alignment.py" followed by "make_phylogeny.py". The phylogenetic trees and the OTU tables were used in the QIIME script "beta_diversity.py" to return, for each OTU table, a pairwise matrix of weighted Unifrac distances between all samples. Principal Coordinate Analysis ordination was performed using the pcoa() function in the "ape" library of R[6], and the first two principal coordinates were plotted using the geom_point() function in the "ggplot2" library of R [1].

Supplementary Figure 10: PNA design

Method described under "Peptide Nucleic Acid (PNA) design". The black histogram of *k*-mer matches to the database was made using the plot() function in the "base" library of R [5], with the abline() function used to add the vertical red lines. Degenerate regular expressions of common primers (**Supplementary Table 1g**) were used to map primer locations to the plastid or mitochondrial sequence along the x-axis.

Supplementary Figures 11 and 12: Bacterial family and OTU relative abundance for different PNA treatments

In panel a (root EC) and b (soil) for both Supplementary Figures, the relative abundance of each bacterial family or OTU was compared between the 12 samples amplified using either pPNA (left) or mPNA (right) and the remaining 12 samples not containing pPNA or mPNA respectively. The test used was a permutation test on the means (Online Methods). Histograms of *P*-values were made with the hist() function in the "base" library of R [5]. The distribution of *P*-values was compared to the null flat distribution using a Chi-squared test (Online Methods). The dot plot in **Supplementary Fig. 12c** was made with the plot() function in the "base" library of R [5]. For **Supplementary Fig. 12d**, representative sequences for all 118 Bdellovibrionaceae OTUs in the dataset were compared to 118 representative sequences for independent OTUs in the class Deltaproteobacteria and 118 representative sequences from bacterial OTUs sampled at random. The mPNA, in forward, reverse, complemented, and reverse complemented orientations, was aligned stepwise to every base position in each OTU and the best match was recorded. All orientations of PNA were used to capture matches to the complementary strand and because there are some reports of PNA binding to DNA in the reverse orientation. The plot summarizing the best alignment scores for the 118 OTUs in each group was made using the geom_line() function in the "ggplot2" library of R [1].

Supplementary Figure 13: Use of PNA on *A. thaliana* and *O. sativa* leaf DNA

In **a**, the chloroplast and mitochondrial 16S sequences used to determine if pPNA and mPNA respectively were likely to function were taken from NCBI GenBank; the sequences and their GenBank ID are in **Supplementary Table 6.**

In **b**, pattern-matching sequences in Run B were processed into ConSeqs which were clustered at 97% identity to form an OTU table. Sixteen leaf samples from *A. thaliana* and *O. sativa* (**Supplementary Table 2a,d**) were extracted from this OTU table and rarefied to the smallest sample of the 16 (161 sequences). The relative abundance of bacterial sequences, plastid sequences, mitochondrial sequences, and other sequences were expressed as a percentage. The stacked bar chart was produced with the geom_bar() in the "ggplot2" library of R [1]. Next, pattern-matching sequences in Run B were processed into NT-sequences which were clustered at 97% identity to form an OTU table, and the 16 leaf samples were extracted. The number of all NT sequences in each sample, without normalization, is graphed in the dark blue bars. Contaminant OTUs were removed and the number of usable bacterial reads, without

normalization, is graphed in brown bars. The dark blue and brown bar plots were produced with the function barplot() in the "graphics" library of R [5].

Supplementary Figure 15b,d: Mixing and pattern-recognition of template-tagging primers

Regular expressions that recognize all 9 pairings of forward and reverse MT-FS group 1 primers or all 9 pairings of forward and reverse MT-FS group 2 primers (**Supplementary Table 1g**), or alternatively, regular expressions recognizing the template barcodes in the Bc-MT-FS primers (**Supplementary Table 1g**), were used to query sequence files from each sample (**Supplementary Table 2a,d**). The charts were produced with the boxplot() and points() function in the "base" library of R [5].

Supplementary Figure 16: Internal Transcribed Spacer (ITS) amplicons

Pattern-matching sequences from Run B were processed into ConSeqs, which were clustered at 97% identity to form an OTU table. Root EC samples amplified with V4 16S primers and root EC samples amplified with ITS2 primers were pooled and each pool was rarefied to a common value of 14,112 ConSeqs. The number of bases in each OTU was calculated (OTU length) for ITS and 16S OTUs, and then the number of OTUs at each OTU length was graphed in panel a using the plot() function in the "base" library of R [5], as was the rank-abundance curve in panel b.

Supplementary Figure 17: Primer linkers

Two bases 5-prime of the 515F primer and two bases 3' of the 806R primer were extracted from all sequences without unambiguous bases in the Greengenes 97% training set of full length 16S sequences (29,846 sequences) that matched expected patterns for V4 amplicons (**Supplementary Table 1g**), and the frequency of each base at all four positions was graphed. The figure represents the + strand, and so the reverse complement of the linkers in both our 806R primers and the Earth Microbiome Project [7] primers are displayed.

Supplementary Note References

1. Hadley Wickham, ggplot2: elegant graphics for data analysis (Springer New York, 2009).
2. Derek S. Lundberg, Sarah L. Lebeis, Sur Herrera Paredes et al., *Nature* **488** (7409), 86 (2012); Davide Bulgarelli, Matthias Rott, Klaus Schlaeppi et al., *Nature* **488** (7409), 91 (2012); Andrew K. Benson, Scott A. Kelly, Ryan Legge et al., *Proceedings of the National Academy of Sciences* **107** (44), 18933 (2010).
3. Gregory R. Warnes, gplots: Various R programming tools for plotting data (2011).
4. Soumitesh Chakravorty, Danica Helb, Michele Burday et al., *Journal of Microbiological Methods* **69** (2), 330 (2007).
5. R Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2012).
6. E. Paradis, J. Claude, and K. Strimmer, *Bioinformatics* **20**, 289 (2004).
7. J. Gregory Caporaso, Christian L. Lauber, William A. Walters et al., *ISME J* **6** (8), 1621 (2012).