Wu et al. BMC Microbiology (2015) 15:125
DOI 10.1186/s12866-015-0450-4

BMC Microbiology

METHODOLOGY ARTICLE Open Access

Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis

Liyou Wu[1†], Chongqing Wen[1,3†], Yujia Qin[1†], Huaqun Yin[1,4,5], Qichao Tu[1], Joy D. Van Nostrand[1], Tong Yuan[1], Menting Yuan[1], Ye Deng[1,7] and Jizhong Zhou[1,2,6*]

Abstract

Background: Although high-throughput sequencing, such as Illumina-based technologies (e.g. MiSeq), has revolutionized microbial ecology, adaptation of amplicon sequencing for environmental microbial community analysis is challenging due to the problem of low base diversity.

Results: A new phasing amplicon sequencing approach (PAS) was developed by shifting sequencing phases among different community samples from both directions via adding various numbers of bases (0–7) as spacers to both forward and reverse primers. Our results first indicated that the PAS method substantially ameliorated the problem of unbalanced base composition. Second, the PAS method substantially improved the sequence read base quality (an average of 10 % higher of bases above Q30). Third, the PAS method effectively increased raw sequence throughput (~15 % more raw reads). In addition, the PAS method significantly increased effective reads (9–47 %) and the effective read sequence length (16–96 more bases) after quality trim at Q30 with window 5. In addition, the PAS method reduced half of the sequencing errors (0.54–1.1 % less). Finally, two-step PCR amplification of the PAS method effectively ameliorated the amplification biases introduced by the long barcoded PCR primers.

Conclusion: The developed strategy is robust for 16S rRNA gene amplicon sequencing. In addition, a similar strategy could also be used for sequencing other genes important to ecosystem functional processes

Keywords: Next generation sequencing, Low diversity sample, Amplicon sequencing, Illumina Miseq, Microbial community, Phasing primer, Microbial ecology

Background

Microorganisms are the most diverse group of life known and can inhabit almost every imaginable environment on Earth [1]. Due to their vast diversity and as-yet uncultivated status, detecting, characterizing and quantifying zicroorganisms in natural settings are of grand challenges. Advances in high-throughput sequencing enabled microbiologists to address many research questions that were previously unanswerable. A major application of high-throughput sequencing in microbial ecology is sequencing amplified gene markers (e.g., 16S ribosomal RNA, nifH) [2, 3] to determine the phylogenetic/functional diversity of a microbial community [4–10]. Although various high-throughput or next generation sequencing technologies are available, the Illumina platform (e.g., HiSeq 2000, HiSeq 2500, MiSeq) has become an attractive option due to lower cost, rapid analysis, and higher accuracy [4–6, 9, 11–15]. It is anticipated that the MiSeq platform in particular will be a dominant sequencing technology for microbial ecology studies due to its great flexibility, fast-turnaround time, longer sequence reads and high accuracy [5, 16–20].

To decrease experimental cost, different community samples are sequenced together in a single Hiseq lane or Miseq run via the use of barcodes, which are added during PCR amplification [21]. However, low sequence

* Correspondence: jzhou@ou.edu
†Equal contributors
[1]Institute for Environmental Genomics, and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, USA
[2]State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China
Full list of author information is available at the end of the article

diversity or unbalanced base composition in template DNA sequences, which affects sequence output, quality, and error rate due to problems in cluster identification, focusing, phasing/pre-phasing and color matrix estimation, and high signal noise, are inherently problematic in amplicon sequencing with Illumina sequencing technologies [21]. Illumina addressed some of the issue caused by low diversity by improving the Real Time Analysis (RTA) software and providing a new reagents kit [14] although this is still a challenging issue and a certain amount of Phix as an additive is still needed to reduce the problem. Different length of barcodes (3–6 bases, three bases difference) [22] and short spacers (0–5 bases) [14] have been used to shift sequences in template DNA, but these shifts are inadequate, especially for the region with continuous homopolymers (Additional file 1: Figure S1 and Additional file 1: Figure S2). For example, there are five 'GGG', three 'GGGG', and one 'GGGGG' homopolymers within the 16S rRNA gene v4 region; and if the primer pair amoA-1 F, 5′-GGG GTTTCTACTGGTGGT and amoA-2R, 5′-CCCCTC KGSAAAGCCTTCTTC-3′ [23] are used for bacterial amoA gene amplicon sequencing, there will be 'GGGG' and 'CCCC' homopolymers at the beginning of the forward and reverse primer, respectively. In parallel to this study, recently, longer spacers (0–7 bases) were used in a dual-indexing primer design for reducing the number of barcoded primers in multiplex 16S rRNA gene amplicon sequencing and higher quality of sequence reads were reported [24, 25]. This design put spacers of 0–7 bases after indices of 12 bases in both forward and reverse primers, which are positioned after the Illumina HP10 or HP11 (Illumina, San Diego, CA, USA) sequencing primers.

Therefore, the sequencing for both forward and reverse reads starts at the indices of the forward and reverse primers, sacrificing a total 24 bases of the paired end reads, which will be essential for some long amplicon sequencing if assembly of the paired end reads is desired.

Here, we developed a new 16S rRNA gene-based amplicon sequencing strategy to ameliorate the problems associated with low diversity. In our phasing primer design, spacers of 0–7 bases are arranged in a complementary fashion in the forward and reverse primers so that the total length of the spacers is 7 bases in all paired end reads. With this spacer design, the total number of added bases between the forward and reverse primers is limited to 7 bases as to maximize the useful length of each amplicon sequence and to minimize any quality bias among sequence reads resulting from using different primer combinations. The single index of 12 bases is positioned between the Illumina adapter, which is used to hybridize the template DNA to the oligo on the Miseq flow cell, and the HP11 sequencing primer in the reverse primer. The index is sequenced separately so that it does not take spaces in the paired end sequence reads. In addition, a two-steps PCR amplification procedure is used to eliminate possible bias introduced by extra components in the long phasing primers (besides the bias introduced by target gene primers). A systematic comparison was made between Miseq runs of phasing and un-phasing methods in terms of throughput, sequence length, error rates and biases. Our results indicated that this strategy substantially increases sequence output, reads number and quality, and decreases sequencing errors, and hence can serve as a robust approach for reliably sequencing amplicons of large scale samples from various communities.

Methods

Samples, DNA extraction, and mock community DNA

Samples, including soils, ground waters, sea waters, bioreactor cultures, and saliva samples, used for PAS and non-PAS comparisons were collected from various locations and experiments. A neutral black soil planted with maize collected from Hailun, China in 2011 was used to compare one- and two-step PCR. Community DNA was extracted by freeze-grinding plus sodium dodecyl sulfate (SDS) lysis as described previously [26]. Crude DNA extracts were purified by electrophoresis on a 0.7 % low melting agarose gel, followed by phenol extraction [27]. DNA quality was assessed based on the absorbance ratios of 260/280 nm and 260/230 nm using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and the DNA concentration was quantified using a PicoGreen (Life Technologies, Grand Island, NY, USA) assay [28] with a FLUOstar Optima (BMG Labtech, Jena, Germany).

The mock community (Additional file 1: Table S1), which contained plasmids carrying near full length 16S rRNA gene sequences of 33 bacteria from different phyla or species at 10^9 copies/μl, was a gift from Dr. Lutgarde Raskin, University of Michigan [29].

PCR primers

The primers used for library preparation for the non-phasing sequencing runs were gifts from Dr. Rob Knight, University of Colorado, Department of Chemistry & Biochemistry, the design of which was described previously [5]. These primers contained the Illumina adapter, a pad and a linker of two bases and barcodes on the reverse primers [5].

For the two-step PCR amplification, primers [515F, 5′-GTGCCAGCMGCCGCGGTAA-3′ and 806R, 5′-GG ACTACHVGGGTWTCTAAT-3′ (Additional file 1: Figure S3A)] targeting the V4 region of both bacterial and archaeal 16S rDNA without added components

were used in the first step to avoid extra bias introduced by spacers and other added component.

The base diversity of sequences in sample libraries affects MiSeq amplicon sequencing in both data throughput and quality. The first 11 bases are particularly critical for cluster identification (first 7 bases) and color matrix estimation (first 11 bases). To increase the base diversity in sequences of sample libraries within V4 region, phasing primers were designed and used in the second step of the two-step PCR. Spacers of different length (0–7 bases) were added between the sequencing primer and the target gene primer in each of the 8 forward and reverse primer sets (Additional file 2: Table S2; Additional file 1: Figure S3E). To ensure that the total length of the amplified sequences do not vary with the primer set used, the forward and reverse primers were used in a complementary fashion so that all of the extended primer sets have exactly 7 extra bases as the spacer for sequencing phase shift. Barcodes were added to the reverse primer between the sequencing primer and the adaptor (Additional file 2: Table S2A, B; Additional file 1: Figure S3E-G). The reverse phasing primers contained (5′ to 3′) an Illumina adapter for reverse PCR (24 bases), unique barcodes (12 bases), the Illumina reverse read sequencing primer (35 bases), spacers (0–7 bases), and the target reverse primer 806R (20 bases). The forward phasing primers included (from 5′ to 3′) an Illumina adapter for forward PCR (25 bases), the Illumina forward read sequencing primer (33 bases), spacers (0–7 bases), and the target forward primer 515F (19 bases). These primers were then used in the second step PCR (Additional file 2: Table S2A, B; Additional file 1: Figure S3E-G).

PCR amplification and purification

Tagged PCR products were generated using primer pairs with unique barcodes through either one or two-step PCR with non-phasing or phasing primers. The addition of extra components (spacers, adaptors, barcodes, etc.) to primers may introduce additional PCR bias due to their varying affinities to the upstream sequences of the target region. To minimize the potential additional bias, a two-step PCR (Fig. 1) was used for library preparation of phasing sequencing runs. In this strategy, target-only primers were used in the first PCR reaction to amplify the target gene and that product was then used in the second PCR using primers containing all of the additional components.

In the one-step PCR, reactions were carried out in a 50 μl reaction: 5 μl 10 × PCR buffer II (including dNTPs), 0.5 U high fidelity AccuPrime™ Taq DNA polymerase (Life Technologies), 0.4 μM of both forward and reverse primers, 10 ng soil DNA or 1 μl mock community of 20x dilution (start solution contained 1×10^9 copies per μl). Samples were amplified using the following program: denaturation at 94 °C for 1 min, and 30 cycles of 94 °C for 20 s, 53 °C for 25 s, and 68 °C for 45 s, with a final extension at 68 °C for 10 min.

In the two-step PCR, the first round was carried out in a 50 μl reaction as described above using target-only forward and reverse primers. Reactions were performed in triplicate and the sample amplification program described above was used except that only 10 cycles were performed. To remove residual first step PCR primers, the genomic DNA templates, and those uncompleted short PCR products, the triplicate products from the first round PCR were combined, purified with an Agencourt® AMPure XP kit (Beckman Coulter, Beverly, MA, USA), eluted in 50 μl water, and aliquoted into three new PCR tubes (15 μL each). The second round PCR used a 25 μl reaction (2.5 μl 10 × PCR buffer II (including dNTPs), 0.25 U high fidelity AccuPrime™ Taq DNA polymerase (Life Technologies), 0.4 μM of both forward and reverse primers, 15 μl aliquot of the first-round purified PCR product). Phasing primers were used in this second round PCR with the barcode on the reverse primers. The amplifications were cycled 20 times following the above program. Positive PCR products were confirmed by agarose gel electrophoresis. PCR products from triplicate reactions were combined and quantified with PicoGreen.

PCR products from samples to be sequenced in the same MiSeq run (generally 3 × 96 = 288 samples) were pooled at equal *molality*. The pooled mixture was purified with a QIAquick Gel Extraction Kit (QIAGEN Sciences, Germantown, MD, USA) and re-quantified with PicoGreen. To keep the PCR product measurements consistent, PCR mixtures that had been previously sequenced were used as standards when a new PCR mixture was quantified. The concentration of the new PCR mixture was adjusted based on the current measurements and previous measurements of the standard PCR mixtures [adjusted new PCR mixture concentration = the measured concentration of the new PCR mixture × (the current measurement of the standard PCR mixture/ the previous measurement of the standard PCR mixture)].

Sequencing

Sample libraries for sequencing were prepared according to the MiSeq™ Reagent Kit Preparation Guide (Illumina, San Diego, CA, USA) as described previously [5]. Briefly, first, the combined sample library was diluted to 2 nM. Then, sample denaturation was performed by mixing 10 μl of the diluted library and 10 μl of 0.2 N fresh NaOH and incubated 5 min at room temperature. 980 μL of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, the 20pM library was further adjusted to reach the desired concentration for sequencing, for example, 625 μl of the

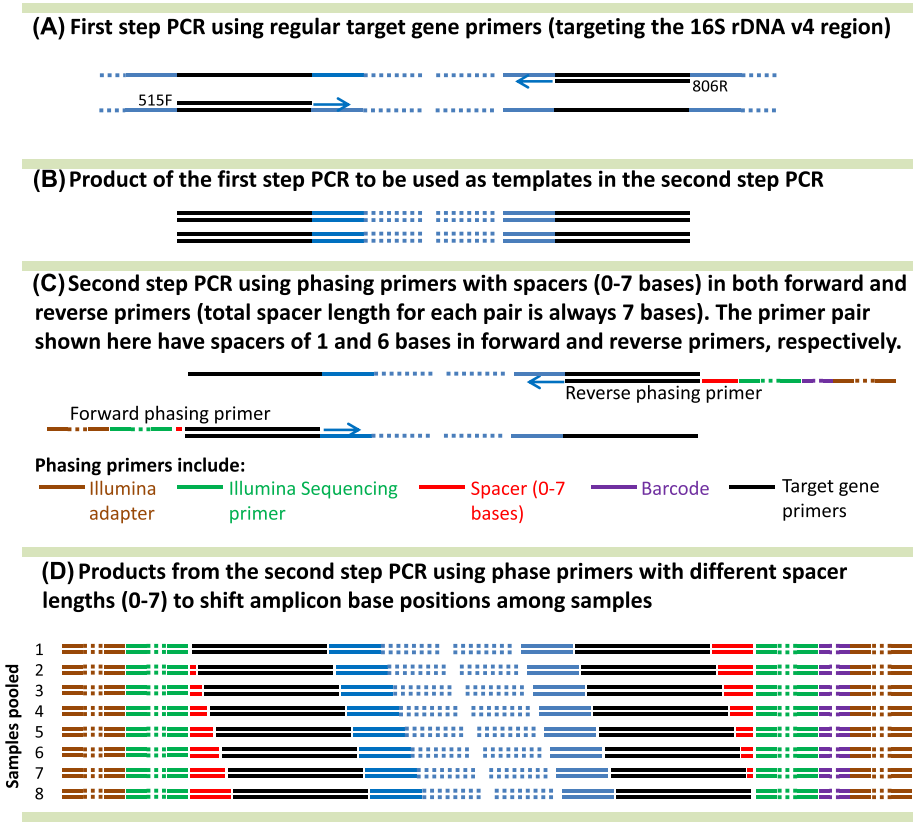

Fig. 1 Two-step PCR scheme used in the PAS method for 16S rRNA gene amplicon sequencing. (**a**) First-step PCR using regular target gene primers: 515F and 806R to target the v4 region only; (**b**) Products from the first-step PCR, which do not have the up and down stream sequences of the original template DNA, thus avoiding extra PCR biases resulting from binding of the added components of the long phasing primers to the template DNA; (**c**) Second-step PCR using complementary phasing primers with spacers of 0–7 bases (generating primer sets of 8 primers, each with a total of 7 bases between the two spacers) in both forward and reverse primers, which provides the sequence position frame shift among samples; (**d**) Products from the second-step PCR showing the sequence position frame shift among samples

20 pM library was mixed with 375 μl of chilled Illumina HT1 buffer to make a 12.5 pM library. The final concentration of the library used for sequencing was determined based on the targeted cluster density. Based on manufacture protocol, the range of cluster density of 500 K/mm^2–1,200 K/mm^2 is recommended. The library for sequencing was mixed with a proportion of a Phix library of the same concentration. For the sequencing runs using Illumina's MiSeq Control Software version 1.1.1 and Real Time Analysis (RTA) version earlier than v1.17.28, Phix DNA spikes were adjusted to 10–20 % for phasing runs and 30–50 % for non-phasing. The incorrect hardcode matrix and phasing estimations were avoided by altering the MiSeq Configuration.xml file to use hardcode matrix and phasing/pre-phasing rates from a normal PhiX DNA run (Additional file 1: Note S1). For the sequencing runs using MiSeq Control Software v2.2.0 with RTA v1.17.28 or later, PhiX DNA was adjusted to about 10–15 % for all runs.

A 500-cycle v1 or v2 MiSeq reagent cartridge (Illumina) was thawed for 1 h in a water bath, inverted ten times to mix the thawed reagents, and stored at 4 °C for a short time until use. For non-phasing primer runs, customized sequencing primers for forward, reverse, and index reads were added to the corresponding wells on the reagent cartridge prior to being loaded as described previously [5]. Sequencing was performed for 251, 12, and 251 cycles for forward, index, and reverse reads, respectively.

Sequencing runs were monitored in real time using the Illumina Sequencing Viewer for cluster density, percentage of clusters passing filter, phasing/pre-phasing ratios, % base, error rates, % reads with quality score ≥30, and other parameters. RTA software v1.17.28 or earlier versions uses the first 4 bases for initial identification of clusters, and the first 11 bases for cluster variation. (http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq_v2.2_software_release_notes.pdf). RTA v1.18.42 uses the first 7 bases for cluster identification and the first 11 cycles for color matrix estimation (http://supportres.illumina.com/documents/documentation/

system_documentation/miseq/miseq-updater-v2-3-software-release-notes.pdf).

Sequence data processing

Raw sequence data was processed using an in-house pipeline which was built on the Galaxy platform and incorporated various software tools. First, the quality of the raw sequence data was evaluated with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then, demultiplexing was performed to remove PhiX sequences, and to sort the sequences to the appropriate samples based on their barcodes, allowing for 1 to 2 mismatches. Quality trimming was done using Btrim [30] prior to merging the forward and reverse reads. Average lengths of all reads and the number of effective reads [with at least 80 % of the maximum theoretical length (200 bp for 2 × 250 bp kits)] were calculated for forward and reverse reads, respectively, after quality trimming. Paired end reads of sufficient length (minimum 20 base overlap between forward and reverse reads) were merged into full length sequences by FLASH v1.2.5 [31]. To test trimming strategies, different trimming window sizes (window 5 and 2) and cutoffs (Quality score 20 and 30) were used. These steps were followed in order to avoid issues of over estimating sequencing error rates based on how FLASH (prior to v1.2.8) assigned quality scores of mismatches within the overlap region. The current pipeline has been updated with FLASH v1.2.8. Sequences were removed if they were too short or contained ambiguous bases. Chimeric sequences were discarded based on prediction by Uchime (usearch v5.2.3) [32] using the reference database mode. OTUs were clustered using Uclust (usearch v5.2.32) [33] at a 97 % similarity level. Final OTUs were generated based on the clustering results, and taxonomic annotation of individual OTUs were achieved based on representative sequences using RDP's 16S Classifier 2.5 [34].

Statistical analyses

To determine the significance of differences among microbial communities, three different complementary non-parametric analyses for multivariate data were used: analysis of similarity (ANOSIM) [35], non-parametric multivariate ANOVA (Adonis) using distance matrices [36] and multiresponse permutation procedure (MRPP) [37]. We used both Jaccard and Bray-Curtis dissimilarity indices to calculate the distance matrix for ANOSIM, Adonis and MRPP analyses. Error rates of the sequences of the mock community were calculated based on sequence alignments of each of the 33 strains to their reference sequences. The significance of the differences in error rates, average read length between phasing runs and non-phasing runs were determined by a two tailed *T*-test.

Results and discussion

Overall strategy

To overcome the problem of low base diversity in 16S rRNA gene amplicon libraries, a novel amplicon sequencing strategy was developed using phasing primers to shift sequencing phases among community samples (Fig. 1; Additional file 1: Figure S3E; Additional file 2: Table S2). The major attributes of this method and other phasing methods are listed in Additional file 1: Table S3. The phasing primers designed in this work have nucleotide spacers between the sequencing and template amplification primers of both the forward and reverse primers (Fig. 1; Additional file 1: Figure S3; Additional file 2: Table S2). The combined length of the forward and reverse spacers is 7 bases, with individual spacers of 0–7 bases in both the forward and reverse primers, which provides a complete frame shift among the sequences of the template DNA. This length and strategy allow for the use of 8 spacer combinations (i.e., 0 bases in the forward, 7 in the reverse; 1 in the forward, 6 in the reverse; etc.) so that the nucleotide bases (A, G, T, C) at each sequencing position can be evenly distributed across the sequences in a sequencing pool (Additional file 1: Figure S3E; Additional file 2: Table S2). For each spacer pair, there is a single forward primer and twelve reverse primers, each having a unique barcode (Additional file 2: Table S2). Distribution of the eight sets of forward and reverse primers among samples in the sequencing libraries provides a sequence shift across sample sets and not just within a single sample. As a result, within a single sequencing run, the barcoded amplicon sequences from different communities are determined in different sequencing phases (Additional file 1: Figure S3I). We refer to this as phasing amplicon sequencing (PAS). In addition, in order to avoid extra PCR bias resulting from the presence of spacers and other components of the long Illumina sequencing primers, a two-step PCR strategy was used for generating the amplicon libraries (Fig. 1; Additional file 1: Figure S3A-G).

Base composition and fluorescence signal distribution

The V4 region of the 16S rRNA gene (Fig. 1; Additional file 1: Figure S3) is commonly targeted for sequencing with the primer set 515F and 806R, which has high sequence coverage for both bacteria and archaea [6, 34, 38] and produces an appropriately sized amplicon (253 bp by excluding primers) for Illumina sequencing. For optimal sequencing results, the base diversity across a set of amplicon sequences would have an even diversity at each position so that each base (A, T, C, G) would be present in 25 % of the sequences at any given positon; however, the base diversity in this region of the 16S rRNA gene is very low. Of the first 100 base positions, 63 and 79 % of positions in the forward and reverse

sequences, respectively, have one base with frequencies greater than 75 % (meaning that the same base is present in that position in 75 % of sequences found in public databases) (Additional file 1: Table S4), while 49 and 63 % of these positions, respectively, have one base with frequencies greater than 90 % (Fig. 2a, b, blue lines). To overcome this problem of unbalanced base distribution, several groups have attempted to shift the sequencing phases of amplicons by using staggered barcodes (3–6 bases) [39] or spacers of 1–5 bases [14]. However, these methods achieved not sufficient sequence position shifts based on the simulation of the base distribution after adding bases to the 5′ end of the primers (Additional file 1: Figure S1 and Additional file 1: Figure S2). Using a 1–5 base spacer, there would be only 6 primers available (i.e., 0 bases, 1 base, 2 bases, etc.), so the base distribution would still be unbalanced even in the first base position (Additional file 1: Figure S1B&E) since 6 is not a multiple of 4 (the number of bases available). For example, when using the 1–5 base spacer, amplicons of the amoA-bacterial gene amplified by the amoA F1 and amoA R2 primer pair [23] have 11 and 10 positions in the forward and reverse reads, respectively, with a single base having frequencies over 60 % (Additional file 1: Figure S1B&E). A similar problem exists with the 3–6 staggered base barcode design. With the same amoA primers mentioned above and the 3–6 staggered base barcode, there are 14 and 12 positions in the forward and reverse reads, respectively, with one base having frequencies over 50 % (Additional file 1:

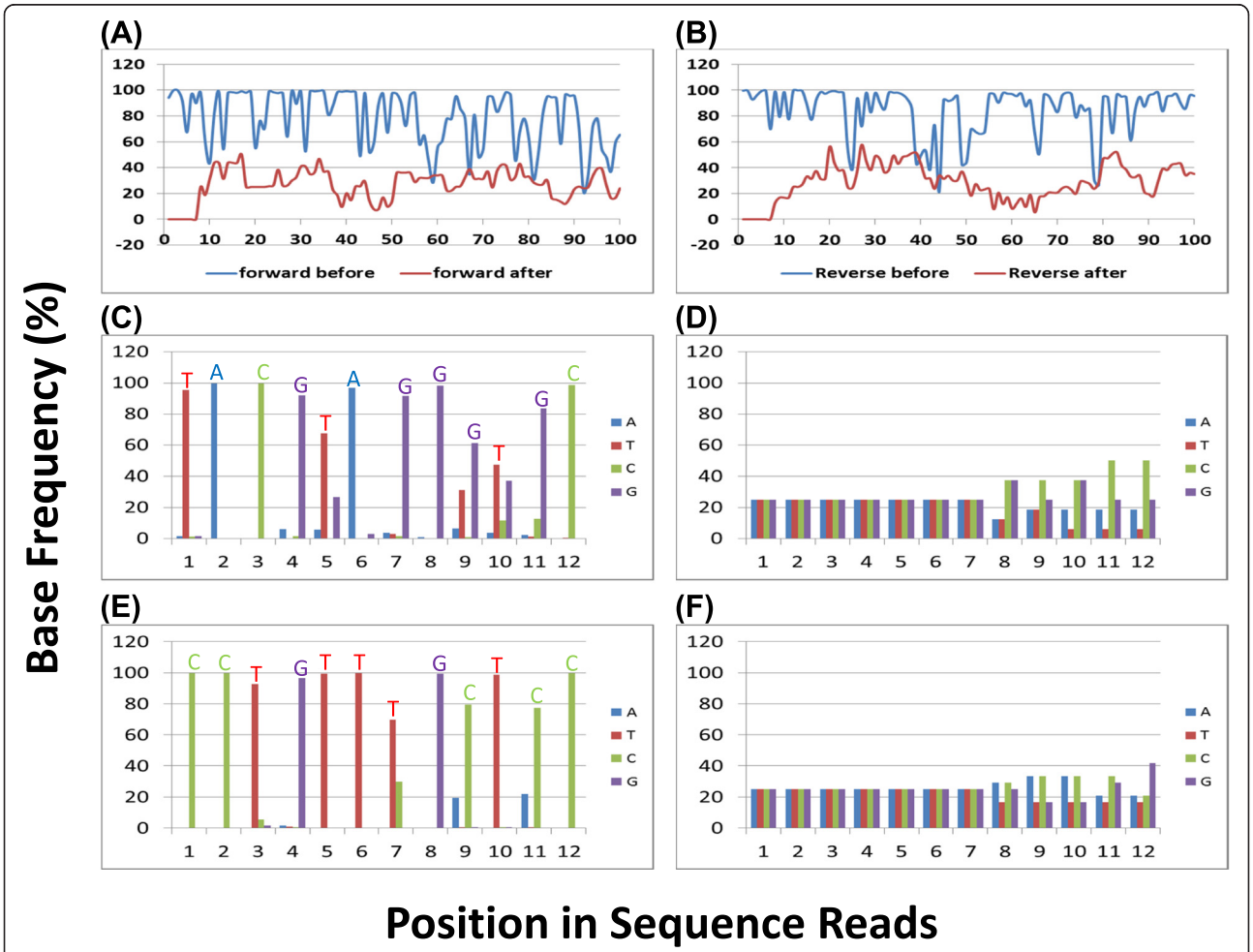

Fig. 2 Impact of phasing primers on base frequency distributions within the V4 region of the 16S rRNA gene. The V4 region is typically amplified with the primer set 515F and 806R, which generates an amplicon of 253 bp (excluding primers). Due to high sequence variation, reliable alignments are difficult to obtain for the entire region. Thus, the base frequencies for the first 100 bp from both directions were estimated based on all 16S gene sequences (96,489 for forward direction, 95,071 for reverse direction) from GreenGenes. Differences between the maximum and minimum base frequencies at each sequence position were estimated before and after primer shift for forward (**a**), and reverse sequences (**b**). Base frequencies of the first 12 positions of the forward sequences before (**c**) and after (**d**) primer shift, and the reverse sequences before (**e**) and after (**f**) primer shift

Figure S1C&F). In addition, there is one position in both the forward and reverse reads (if using the same barcode design as the forward reads) with a base frequency of 100 % due to the 4-base homopolymer at the 5′ end of both the forward (i.e., GGGG) and reverse (i.e., CCCC) primers (Additional file 1: Figure S1C&F). Similar theoretical base distributions were predicted for amplicons of the 16S rRNA gene amplified by the primer pair 515F and 806R (v4 region) [6] as well (Additional file 1: Figure S2). These findings suggested that a larger frame shift of at least 8 bases would be necessary to increase base diversity across the length of the entire amplicon. An additional concern is that using primers of varying length will results amplicon sequences of different length and quality bias among amplicon sequences due to their length differences. So, to address these issues, the PAS strategy developed here uses a complimentary spacer pair containing a variable number of bases (0–7 bp, but always equaling 7 bases between the two) inserted in both the forward and reverse primers between the sequencing and target amplification sections of the primer to shift sequencing phases among different community samples, increasing the base diversity at individual positions (Fig. 1; Additional file 1: Figure S3; Additional file 2: Table S2). After adding these spacers, the base composition in this region is more balanced and the difference in nucleotide frequency for most positions is < 30 % (Fig. 2a, b, red lines). This increased diversity is especially important for the first 11 bases (or sequencing cycles), as they are critical for cluster identification (first 7 bases) and final validation (first 11 bases). The PAS strategy substantially improved the base composition balance for both forward (Figs. 2c vs 2d) and reverse reads (Fig. 2e vs 2f).

It is expected that the improved balance of base composition would increase the evenness of fluorescent signals and the observed base diversity (Additional file 1: Figure S4, Additional file 1: Figure S5A). To confirm this, a PAS run and a non-PAS run both with moderate sequence cluster densities (~800 k/mm2) and similar amounts of spiked PhiX (~15 %) were compared. In the non-PAS run, fluorescent signals among the four bases were very uneven for both forward and reverse reads (Additional file 1: Figure S4I) while, in the PAS run, fluorescent signals were more even (Additional file 1: Figure S4II). The PAS method also increased the base diversity (Additional file 1: Figure S5A, right) compared to the non-PAS run (Additional file 1: Figure S5A, left). In the non-PAS run, the frequency of the four nucleotides for most positions (80 %) were >75 % (Additional file 1: Figure S5A), while they were <40 % in the PAS run (Additional file 1: Figure S5A). As a result, the sequence read base quality was substantially increased for the PAS run as indicated by the percentage of bases above Q30 at the end of the run (Additional file 1: Figure S5B), or distributed along the positions of both forward and reverse reads (Additional file 1: Figure S5C, and D).

Effective reads and read length

To determine whether PAS is consistently better than non-PAS in terms of sequence output, sequence quality and effective read lengths, the experimental data from different PAS (31; 5 were run before RTA was upgraded to RTA1.17.28; PhiX DNA spike was 20 %) and non-PAS (10; 3 were run before RTA was upgraded to RTA1.17.28; PhiX DNA spike was 50 %) runs were analyzed. These sequencing runs were used to determine the diversity of 8,731 microbial communities from diverse habitats such as soil, sediment, groundwater, bioreactors, waste water treatment plants, and human oral and gut. Although the percentage of sequence clusters passing the filter decreased with cluster density for both PAS and non-PAS runs, the decrease was much sharper for non-PAS runs. For example, when cluster densities reached 1000 K/mm^2, the percentage of sequence clusters passing the filter remained above 90 % for PAS runs, but dropped below 80 % for non-PAS runs (Fig. 3a). As expected, the number of sequence reads increased as the cluster density increased for both approaches (Fig. 3a), but more reads were obtained for PAS than non-PAS runs ($p < 0.0001$) (Fig. 3a). In addition, the average percentage of bases with > Q30 at the last cycle was significantly higher ($p < 0.001$) for PAS runs (forward, 93.5 %; reverse, 88.4 %) than for non-PAS runs (forward, 86.3 %; reverse, 78.5 %) (Fig. 3b). These results indicated that the PAS method provided high resolution for sequence cluster identification, and therefore, maximized the sequence read output, and significantly improved sequence read quality due to the balanced fluorescence signal intensity.

The PAS method was further evaluated by comparing the average read length after quality trim at Q30 and Q20 with the trimming window set at 5 or 2. The percentage of effective sequence reads, which refer to those sequences for which at least 80 % of all bases in the theoretical length have scores of > Q30 or > Q20 (e.g. 200 bp for 2 × 250 bp paired end reads), were also evaluated. The average read length for both forward and reverse sequences were significantly longer after quality trimming in PAS runs than in non-PAS runs. This was especially obvious at high cluster densities and at Q30 with the quality trimming window set at 5 (Fig. 3c; Additional file 1: Figure S6). More importantly, the percentage of effective reads were considerably higher for PAS runs than for non-PAS runs for both forward and reverse sequences and for combined full length sequences (253 bp) at all cluster densities compared, particularly at high sequence cluster densities and at Q30 for the reverse reads (Fig. 3d; Additional file 1: Figure S7).

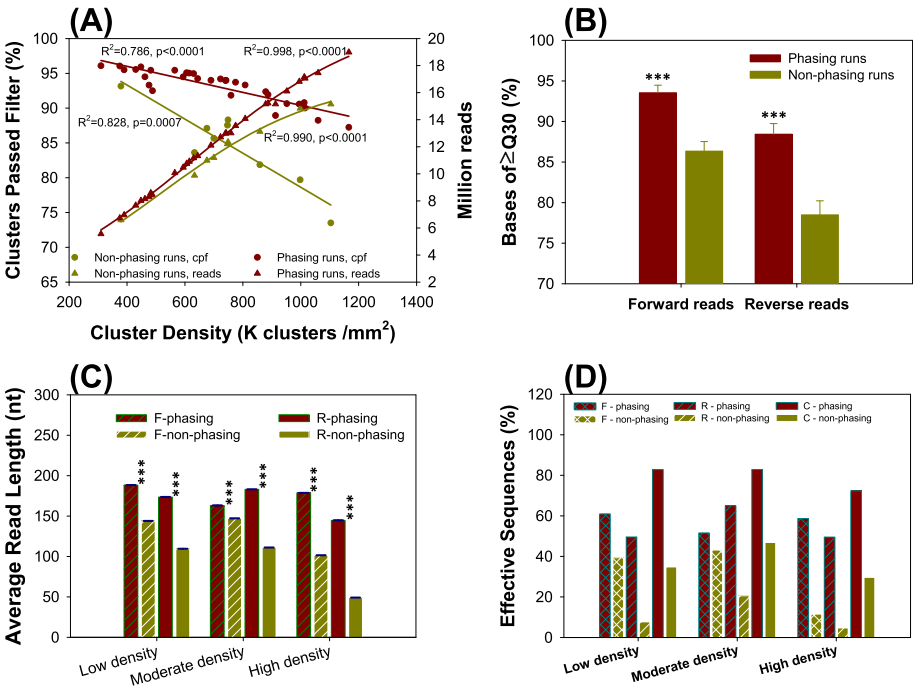

Fig. 3 Sequence output, read length and sequence quality comparisons between the PAS and non-PAS approaches. (**a**) Relationship between sequencing cluster density and percentage of clusters passing filter (● or ●), or read numbers (▲ or ▲). The dark red symbols are for phasing sequencing runs while the dark yellow symbols are for non-phasing sequencing runs. The relationship between cluster density and read number was fitted with a non-linear model: $y = \frac{a}{1+e^{-\left(\frac{x-x0}{b}\right)}}$. (**b**) Effect of PAS on read quality. The average percentage of bases above Q30 were estimated based on 31 PAS runs and 10 non-PAS runs. (**c**) Effect of PAS on average read length. *Error bars* indicate standard deviation of triplicate runs. Symbols may be larger than error bars; and (**d**) Impact of PAS on the percentage of effective sequences under different cluster density levels: low, ~400 k/mm^2; moderate, ~800 k/mm^2; and high, > 1000 k/mm^2. All runs were done with RTA version 1.17.28 and with a Phix DNA spike of 10 %. Here, effective sequences refers to sequences with 80 % of the theoretical length (200 bp for 2 × 250 bp kits) having quality scores ≥ 30. All raw sequences were trimmed if the average sequence quality score was < Q30 with a window size of 5. The Student *t* test was used to test significance: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. F, forward; R, reverse; C, combined

The difference between the number of effective combined sequences in the PAS and non-PAS methods was less than that between either the forward or reverse reads. This was most likely due to the relatively short amplicons generated from the 16S rDNA v4 region. Short reads are still a concern for amplicon sequencing with Illumina platforms even with the 2x300 bp paired end kit. If there is a relatively low base diversity, read length after quality trimming will be much shorter than expected, especially when the quality trimming is done under highly stringent conditions, e.g. Q30. For many functional genes, such as *nir*K, *nir*S, *amo*A, and *dsr*, it is difficult to find primers to generate amplicons of appropriate length, so relatively longer amplicons (over 500 bp) must be selected. The results here indicated that PAS method effectively improved sequence read quality and length, which are critical for sequencing longer amplicons, assembling paired-end reads and increasing overall sequencing accuracy.

Sequence error rates

To determine whether PAS affects sequencing error, a mock community containing full length, plasmid-borne 16S rDNA sequences from 33 different bacterial phyla or classes [29] (Additional file 1: Table S1) was sequenced using both PAS and non-PAS methods (both sequencing runs were performed after the Illumina RTA software was upgraded to version 1.17.28). Based on the results, the PAS method reduced sequencing errors. The average sequencing error rate of the raw sequence reads was significantly lower ($p < 0.0001$) for PAS than non-PAS runs (1.17 vs 1.71 % for forward sequences, 0.77 vs 1.87 % for reverse sequences) (Fig. 4a). Much higher error rates were observed for non-PAS runs both before the 100th cycle and in the last 97 cycles (Fig. 4a; Additional file 1: Figure S8). The higher raw sequence error rates for both forward and reverse reads in the non-PAS run was comparable to other reported error rates [25]. Also, although sequence quality trimming

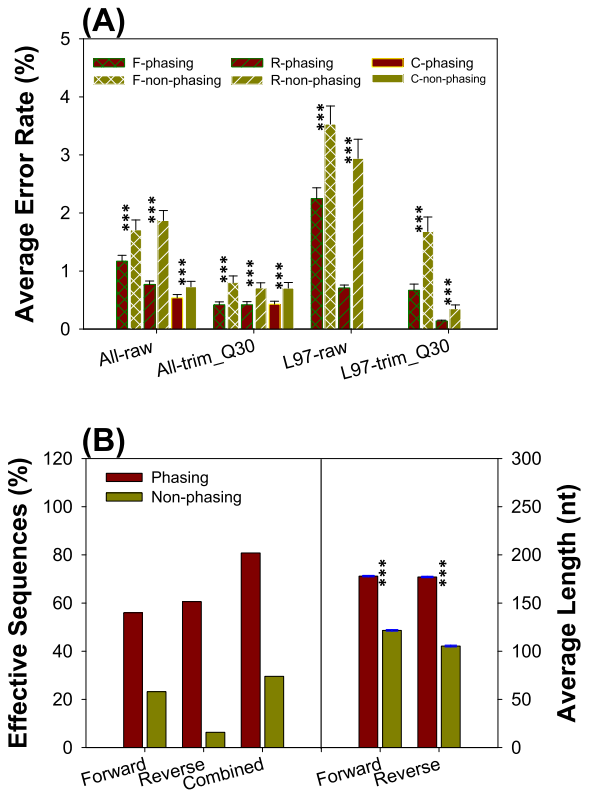

Fig. 4 Effect of PAS on sequence errors as determined with a mock community. (**a**) Average sequence error rates of entire sequence reads, the last 97 bases of the forward or reverse reads, and the combined sequence reads for both PAS and non-PAS methods; (**b**) The percentage of effective sequences for forward and reverse reads, and combined sequences after trimming for both PAS and non-PAS methods (*left panel*), and the average length of forward and reverse reads for both PAS and non-PAS methods (*right panel*). All raw sequences were trimmed if the average sequence quality score was < Q30 with a window size of 5. The Student *t* test was used to test significance: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. F, forward; R, reverse; C, combined

significantly reduced error rates for both approaches, error rates were still considerably higher for non-PAS than PAS runs (Fig. 4a; Additional file 1: Figure S8C-F and Additional file 1: Figure S9). In addition, due to higher sequencing errors and subsequently stricter quality trimming, the percentage of effective sequence reads and combined sequences (Fig. 4b, left panel; Additional file 1: Figure S10, left panel) and the average sequence length (Fig. 4b, right panel; Additional file 1: Figure S10, right panel) was substantially lower for non-PAS runs than PAS runs. These results indicate that the PAS method not only increased the number of effective sequence reads and read length but also reduced sequencing errors. Thus, using phasing primers is an effective and necessary strategy for reducing errors of amplicon sequencing, a major concern of users [11, 13], on the Miseq and other platforms. One way that the PAS method reduces sequence error rates could be the higher quality of the sequencing reads obtained using this method. Another reason could be that PAS has a relatively lower percentage of chimera formation during PCR amplification due to fewer amplification cycles at both amplification steps [40] and preliminary evidence indicates that fewer chimera are present with PAS (Wu et al., unpublished data).

Minimizing possible PCR bias

Since spacers and other components were added to the phasing primers before the target primer sequences (Additional file 1: Figure S3E; Additional file 2: Table S2), additional PCR amplification bias could be introduced [41]. We hypothesized that such biases could be minimized using a two-step PCR amplification strategy in which the target gene is amplified with standard primers (e.g., 515F, 806R) at a low cycle number (e.g. 10 cycles), followed by a second PCR amplification using the PCR products from the first step PCR and long barcoded primers with spacers (Additional file 1: Figure S3A-G). This strategy should reduce biases because the standard primers do not have added components, thus avoiding biases introduced by those components. And then in the second PCR reaction, the PCR products from the first PCR are used as targets and these products do not have up- or down-stream sequences, thus avoiding biases introduced by interaction between those regions and the added primer components. We noted that the method developed by Fadrosh et al. [24] used a single PCR step for amplicon library preparation. So, to test whether the added primer components result in additional

Table 1 Dissimilarity analysis of mock and soil community OTUs among phasing primer sets*[a]

Samples	Methods	Jaccard**			Bray-Curtis		
		MRPP	Anosim	Adonis	MRPP	Anosim	Adonis
Mock community	One-step PCR	1	1	n/a	0.001	0.001	0.001
	Two-steps PCR	1	1	n/a	0.379	0.448	0.374
Neutral black soil	One-step PCR	0.024	0.026	0.027	0.008	0.011	0.002
	Two-steps PCR	0.142	0.252	0.166	0.341	0.373	0.356

*Data in the table are *p* values

**For mock community, *p* value = 1 due to all OTUs were shared by all primer sets

[a]Both the mock community and the neutral black soil were amplified using the 8 phasing primer sets and 3 barcodes (as replicates)

PCR amplification bias and whether a two-step PCR amplification could reduce this type of bias, a mock community (Additional file 1: Table S1) and a soil sample were amplified using three barcoded primers from each of the eight sets of phasing primers with either a one-step or a two-step PCR strategy. Theoretically, if no additional amplification bias is present with the use of long phasing primers, then there will be no differences observed among the different primer sets given that the same template community DNA was used. The community composition and structure were significantly different ($p < 0.01$) among the different primer sets for both mock and soil communities with the one-step PCR amplification (Table 1). In contrast, no significant differences were observed among the different primer sets with the two-step PCR amplification (Table 1). These results indicated that the long primers with added components did introduce extra amplification biases with one-step PCR amplification while no apparent bias was introduced by the two-step PCR amplification. In addition, PCR amplification bias among technical replicates was also present with the one-step PCR when primers without spacers were used (data not shown). These results suggest that a two-step PCR approach could minimize amplification biases due to the introduction of additional components to PCR primers. The use of a two-step PCR approach is necessary if phasing primers or primers with added components are used for amplicon library preparation.

Conclusions

In summary, although the Illumina MiSeq and other high-throughput sequencing technologies are promising and powerful tools, adopting these technologies for analyzing microbial communities is challenging. A novel amplicon sequencing approach was developed by shifting sequencing phases among different community samples from both directions via adding a total of 7 bases to both forward and reverse primers as spacers. Our results indicate that this approach effectively increases raw sequence throughput, read quality and effective read sequence length, and reduces sequencing errors. Analysis of MiSeq sequencing runs showed that PAS provides a robust approach for reliably analyzing microbial communities of diverse composition from a variety of habitats. In addition, our results indicate that a two-step PCR amplification strategy effectively ameliorates PCR amplification biases introduced by the use of long barcoded PCR primers. The use of a single barcode makes it easy to utilize the complementary phasing primers among samples, but multiplex amplicon sequencing requires a large number of barcoded primers, increasing the upfront costs of this method. However, despite this initial outlay, the cost per sample for the PAS method is similar to other methods. After a careful comparison of the PAS method described in this paper and other phasing methods [14, 25, 39, 42], the PAS method has the following unique features: i) sufficient sequence position frame shift among samples to increase base diversity across the entire sequence; ii) minimum base sacrifice by sequencing barcodes in separate reads (index reads); iii) a complementary spacer design that adds a combined 7-base spacer to both the forward and reverse primers, minimizing the total number of bases added, maximizing the amplicon sequence length, and avoiding quality biases caused by differences in amplicon sequence lengths; iv) a two-step PCR strategy that eliminates the potential extra PCR bias caused by added PCR primer components, v) lower PCR cycles in both first and second step PCR to reduce chimeras. In addition, this study is the first time to systematically and thoroughly evaluate a phasing method for Miseq amplicon sequencing in terms of data output, sequence quality, error rate, and bias. While this strategy was developed and tested on the 16S rRNA gene, it has also been used successfully on ITS for fungi, 18S rRNA genes for protist, and other functional genes including bacterial and archaeal *amo*A *nif*H, *mcr*A, and *pmo*A (not shown here), indicating its applicability for sequencing many different genes.

Additional files

Additional file 1: Supporting Data is available with LabArchives at https://mynotebook.labarchives.com/share/wuliyou/MjAuOHw5MTgxMC8xNi9UcmVlTm9kZS8yNzMyMDY5NDU1fDUyLjg=, and with the doi:10.6070/H4KD1VW7. Table S1. Bacteria mock community. **Table S3.** Comparison of Technical Aspects between the PAS Method and Other Phasing Methods. **Table S4.** Theoretical base frequencies of 16S rDNA V4 region amplicons generated using target primers. **Table S5.** Technical Comparison of the PAS Method and Other Phasing Methods. **Figure S1.** Theoretical base distribution of the first 16 positions in both the forward and reverse reads of the amoA gene amplicon amplified by the primer pair amoA F1 (5′-GGGGTTTCTACTGGTGGT) and R2 (5′-CCCCTCKGSAAAGCCTTCTTC) with different frame shift lengths. **Figure S2.** Theoretical base distribution of the first 16 positions of both the forward and reverse reads of the 16S rRNA gene amplified by the primer pair of 515F (5′-GTGCCAGCMGCCGCGGTAA) and 806R (5′-GGACTACHVGGGTWTCTAAT) (v4 region) with frame shifts of different lengths. **Figure S3.** Overview of the phasing amplicon sequencing strategy. **Figure S4.** Thumbnail images showing fluorescence signal. **Figure S5.** Performance comparison between a non-phasing run (left) and a phasing run (right), both with moderate cluster densities (~800 k/mm2) and similar amounts of spiked PhiX (~15 %). **Figure S6.** Effects of quality trimming on sequence read length. **Figure S7.** Effects of quality trimming on effective read sequences. **Figure S8.** Sequence error rates at each sequence position for phasing (dark red) and non-phasing (dark yellow) runs. **Figure S9.** Sequence error rates from mock community sequencing at lower quality trimming standards. **Figure S10.** Impact of PAS on the number of effective sequences and the read length at lower quality trimming standards (Q20, window size of 5). **Note S1.** Altering the MiSeqConfiguration.xml file to hardcode 750 matrix and phasing.

Additional file 2: Table S2a. Forward PCR primers for the preparation of amplicon libraries for sequencing of the 16S rDNA V4 region (separate excel file). **Table S2b.** Reverse PCR primers for the preparation of amplicon libraries for sequencing of the 16S rDNA V4 region (separate eel file).

Abbreviations
PAS: Phasing amplicon sequencing approach; RTA: Real Time Analysis.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
LW and JZ conceived the Phasing Amplicon Sequencing methods and designed the experiments. CW, HY, YT, MY, and JDV performed experiments. YJ and YD designed and developed the sequence processing pipeline. LW, YJ, QT, and JDV analyzed data. JZ, LW, and JDV. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This method development work was supported by the OBER Biological Systems Research on the Role of Microbial Communities in Carbon Cycling Program (DE-SC0010715), and the U.S. National Science Foundation MacroSystems Biology program under the contract (NSF EF-1065844). We thank R. Knight (Department of Chemistry and Biochemistry, University of Colorado) for providing us barcoded non-phasing 16S rDNA PCR primers. We thank L. Raskin (Department of Civil and Environmental Engineering, University of Michigan University) for providing the 16S rDNA mock community.

Author details
[1]Institute for Environmental Genomics, and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, USA. [2]State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China. [3]Fisheries College, Guangdong Ocean University, Zhanjiang, Guangdong, China. [4]School of Minerals Processing and Bioengineering, Central South University, Changsha, Hunan, China. [5]Key Laboratory of Biometallurgy of the Ministry of Education, Changsha, Hunan, China. [6]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [7]CAS Key Laboratory of Environmnetal Biotechnology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China.

Received: 3 March 2015 Accepted: 18 May 2015
Published online: 19 June 2015

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci. 1998;95:6578–83.
2. Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R *et al.* (2010). Microbiome Profiling by Illumina Sequencing of Combinatorial Sequence-Tagged PCR Products. Plos One 2010;5(10):e15406.
3. Pereira e Silva MC, Schloter-Hai B, Schloter M, van Elsas JD, Salles JF (2013). Temporal Dynamics of Abundance and Composition of Nitrogen-Fixing Communities across Agricultural Soils. Plos One 2013;8(9):e74500.
4. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-End illumina reads. Appl Environ Microbiol. 2011;77:3846–52.
5. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. Isme J. 2012;6:1621–4.
6. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011;108:4516–22.
7. Degnan PH, Ochman H. Illumina-based analysis of microbial community diversity. Isme J. 2012;6:183–94.
8. Deng Y, He Z, Xu M, Qin Y, Van Nostrand JD, Wu L, et al. Elevated carbon dioxide alters the structure of soil microbial communities. Appl Environ Microbiol. 2012;78:2991–95.
9. Zhou H-W, Li D-F, Tam NF-Y, Jiang X-T, Zhang H, Sheng H-F, et al. BIPES, a cost-effective high-throughput method for assessing microbial diversity. Isme J. 2011;5:741–9.
10. Zhou J, Wu L, Deng Y, Zhi X, Jiang Y-H, Tu Q, et al. Reproducibility and quantitation of amplicon sequencing-based detection. Isme J. 2011;5:1303–13.
11. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human Gut microbiota. Science. 2013;341:44–U53.
12. Grossmann V, Roller A, Klein H-U, Weissmann S, Kern W, Haferlach C, et al. Robustness of amplicon deep sequencing underlines its utility in clinical applications. J Mol Diagn. 2013;15:473–84.
13. Hadd AG, Houghton J, Choudhary A, Sah S, Chen L, Marko AC, et al. Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens. J Mol Diagn. 2013;15:234–47.
14. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL. Practical innovations for high-throughput amplicon sequencing. Nat Methods. 2013;10(10):999–+.
15. Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat. 2013;34(7):1035–42.
16. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J (2014). Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. Plos One. 2014;9(4):e94249.
17. Ling AL, Robertson CE, Harris JK, Frank DN, Kotter CV, Stevens MJ, Pace NR, Hernandez MT. Carbon dioxide and hydrogen sulfide associations with regional bacterial diversity patterns in microbially induced concrete corrosion. Environmental Sci Technol. 2014;48(13):7357–64.
18. Liang B, Cheng H, Van Nostrand JD, Ma J, Yu H, Kong D, et al. Microbial community structure and function of Nitrobenzene reduction biocathode in response to carbon source switchover. Water Res. 2014;54:137–48.
19. Koskey AM, Fisher JC, Traudt MF, Newton RJ, McLellan SL. Analysis of the gull fecal microbial community reveals the dominance of catellicoccus marimammalium in relation to culturable enterococci. Appl Environ Microbiol. 2014;80(2):757–65.
20. Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc Natl Acad Sci U S A. 2014;111(22):8007–12.
21. Krueger F, Andrews SR, Osborne CS (2011). Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. Plos One. 2011;6(1):e16607.
22. Hummelen R, Fernandes AD, Macklaim JM, Dickson RJ, Changalucha J, Gloor GB *et al.* (2010). Deep Sequencing of the Vaginal Microbiota of Women with HIV. Plos One 5.
23. Rotthauwe JH, Witzel KP, Liesack W. The ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. Appl Environ Microbiol. 1997;63:4704–12.
24. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, et al. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome. 2014;2:1–7.
25. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq illumina sequencing platform. Appl Environ Microbiol. 2013;79:5112–20.
26. Zhou JZ, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. Appl Environ Microbiol. 1996;62:316–22.
27. Xie J-p, Wu L-y, van Nostrand JD, He Z-l, Lu Z-m, Yu H, et al. Improvements on environmental DNA extraction and purification procedures for matagenomic analysis. J Cent South Univ. 2012;19:3055–63.
28. Pinto AJ, Raskin L (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. Plos One 7.
29. Ahn SJ, Costa J, Emanuel JR. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. Nucleic Acids Res. 1996;24:2623–5.
30. Kong Y. Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics. 2011;98:152–3.
31. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27:2957–63.
32. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27:2194–200.
33. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
34. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73:5261–7.

35. Clarke KR. Nonparametric multivariate analyses of changes in community structure. Aust J Ecol. 1993;18:117–43.
36. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26:32–46.
37. Zimmerman GM, Goetz H, Mielke PW. Use of an improved statistical-method for group comparisons to study effects of prairie fire. Ecology. 1985;66:606–11.
38. Wang Y, Qian P-Y (2009). Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies. Plos One 4.
39. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, et al. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome. 2014;2.
40. Qiu XY, Wu LY, Huang HS, McDonel PE, Palumbo AV, Tiedje JM, et al. Evaluation of PCR-generated chimeras: Mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microbiol. 2001;67:880–7.
41. Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl Environ Microbiol. 2011;77:7846–9.
42. Hummelen R, Macklaim JM, Bisanz JE, Hammond J-A, McMillan A, Vongsa R *et al.* (2011). Vaginal Microbiome and Epithelial Gene Array in Post-Menopausal Women with Moderate to Severe Dryness. Plos One 6.