

RESEARCH

Open Access



Modified PCR protocol to increase sensitivity for determination of bacterial community composition

Kayla M. Williamson¹, Brandie D. Wagner^{1,2}, Charles E. Robertson³, Mark J. Stevens², Marci K. Sontag⁴, Peter M. Mourani² and J. Kirk Harris^{2*} 

Abstract

Background: The objective of this project was to increase the sensitivity of sequence-based bacterial community determination without impacting community composition or interfering with cluster formation during sequencing. Two PCR protocols (standard and modified) were examined in airway samples where we observed a large range in bacterial load (3.1–6.2 log₁₀ 16S rRNA gene copies/reaction). Tracheal aspirate (TA) samples ($n = 99$) were collected from sixteen children requiring mechanical ventilation at a single center. DNA was extracted, and total bacterial load (TBL) was assessed using qPCR. Amplification of 16S rRNA was attempted with both protocols in all samples.

Results: PCR product was observed using both protocols in 52 samples and in 24 additional samples only with the modified protocol. TBL, diversity metrics, and prominent taxa were compared for samples in three groups based on success of the two protocols (successful with both, success with modified only, unsuccessful for both). TBL differed significantly across the three groups ($p < 0.001$). Specifically, the modified protocol allowed amplification from samples with intermediate TBL. Shannon diversity was similar between the two protocols, and Morisita-Horn beta diversity index showed high agreement between the two protocols within samples (median value 0.9997, range 0.9947 to 1). We show that both protocols identify similar communities, and the technical variability of both protocols was very low. The use of limited PCR cycles was a key feature to limit impact of background by exclusion of 24% of samples with no evidence of bacterial DNA present in the sample.

Conclusion: The modified amplification protocol represents a viable approach that increased sensitivity of bacterial community analysis, which is important for study of the human airway microbiome where bacterial load is highly variable.

Keywords: 16S rRNA gene, Microbiome, Amplicon, Mechanical ventilation

* Correspondence: Jonathan.Harris@cuanschutz.edu

²Section of Critical Care, Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, 13123 E. 16th Ave. Box B395, Aurora, CO 80045, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

High-throughput sequencing allows parallel examination of bacterial communities from large numbers of samples. This is accomplished by using indexed primers to tag individual samples with a short DNA index sequence that allows assignment of each sequence to the sample of origin. High-throughput sequencing platforms require additional extension of the indexed primers with oligonucleotides to facilitate sequencing. The additional primer oligonucleotides required for high-throughput sequencing impacts the efficiency of amplification. This is problematic for samples with low levels of bacterial DNA or in cases where large amounts of non-target DNA are present, which is common in clinical specimens. We utilized a qPCR-based approach to estimate the bacterial load [1, 2] rather than total DNA concentration to perform sequencing amplification. The bacterial load determined by qPCR is more informative than DNA concentration regarding the amount of bacteria present in clinical samples where the majority of DNA is non-bacterial.

To reduce the impact of background DNA on bacterial community composition limiting PCR cycle numbers is important [3]. We utilize a conservative approach that eliminates samples from analysis due to inability to amplify adequate PCR product in order to decrease the contribution of background DNA in sequencing projects. This is accomplished by validating the background level in each tube of PCR reagent by running positive and negative amplification controls prior to use with clinical samples. Amplification is limited to a cycle number below this tube-specific background (never exceeding 30 cycles), which limits the amount of background contribution to the sequencing experiments. We observed poor sensitivity with our standard amplification protocol in association with low bacterial load in human airway samples. To address this poor sensitivity, we designed a modified amplification protocol that incorporated primers specific to the target region without sequencing adapters.

The goal of the experimental modification was to increase the sensitivity of bacterial community analysis for samples containing low bacterial load with the expectation that bacterial community identification can be informative to underlying pathobiology of patients, even in low load settings. To test this idea, we utilized tracheal aspirate samples from children requiring invasive mechanical ventilation support that contained a mixture of high and low bacterial load DNA extractions. DNA amplification efficiency was tested using both protocols. To examine technical variation, we utilized samples with adequate load to run replicate libraries using both protocols.

Methods

Samples

Tracheal aspirate (TA) samples were collected from children requiring invasive mechanical ventilation at a single center under an IRB-approved protocol (COMIRB 14-1530). Written informed consent and HIPPA authorization were obtained from all subjects or their legal guardians if <18 years of age. The initial sample was collected within 24 h of initial intubation, and subsequent samples were collected daily for up to 14 days. Samples were collected during standard of care endotracheal tube clearance into a sterile mucous trap via inline suctioning equipment. Depth of suctioning was standardized by protocol. Per care guidelines up to 0.5 mL of non-bacteriostatic sterile saline could be used to clear tubing and facilitate collection. Specimens were transferred from the sterile mucous trap to 2 mL cryovials, within 5 min of collection, and stored at -80°C . Each sample was assigned an alphanumeric identifier that was utilized throughout the laboratory processing.

DNA extraction and determination of bacterial load

DNA was produced using the Qiagen EZ1 Advanced extraction platform. The frozen tracheal aspirate was thawed at 4°C , and a 200 μL aliquot was transferred into the tube supplied with the EZ1 DNA Tissue Kit (Qiagen). DNA extraction controls were generated using 200 μL of DEPC-treated water (same used for PCR) in parallel with samples. Extraction was performed using manufacturer's instructions, and DNA was eluted in 100 μL of Tris/EDTA (TE, supplied in the cartridge). Each DNA was diluted 1:40 in TE (to preserve DNA sample volume), and 4 μL of the dilution (dilution factor of 10) was used in triplicate reactions to determine total bacterial load based on the assay published by [2]. Briefly, this assay is based on TaqMAN qPCR approach and utilizes three conserved regions within the 16S rRNA gene (338/805 regions for primers, 515 region for probe). Copy number was determined using a cloned 16S rRNA gene (*Prevotella melaninogenica*) obtained from CF sputum [1, 4]. The coefficient of variation for the standards Ct value was $\sim 1\%$ across the range of copy number (10^3 to 10^8), and the assay efficiency was ≥ 0.86 for all plates, and the measured background was 192 ± 20 copies/reaction.

High-throughput DNA sequencing for microbiome analysis

16S amplicon library construction

Bacterial profiles were determined by broad-range amplification and sequence analysis of 16S rRNA genes [5–7]. Each DNA was amplified in triplicate along with an index-specific negative PCR control (single mastermix; termed “standard protocol” throughout). Each reaction

contained 1X HotmasterMix (5Prime), 150 nM each indexed 27F, and 338R primer in a 25 µL reaction volume. Cycling conditions were 94 °C 2 min followed by 30 cycles of 94 °C 20 s, 52 °C 20 s, 65 °C 60 s. After thermal cycling, the amplicons were assessed by agarose gel electrophoresis (pooled triplicates and negative control independently) for appropriately sized bands from the DNA template and no evidence of amplification from the negative controls. If any amplification was evident in the negative control, that specimen was repeated. In addition to our standard protocol, we utilized a second modified protocol to attempt to improve sensitivity for low bacterial load samples. The additional PCR was performed exactly as in the standard protocol, but with 15nM 27F/338R primers (10% spike) that did not contain required extension for sequencing (index, linker, Illumina adapters; termed “modified protocol” throughout). All DNA extracts utilized in this study were attempted using both protocols. *PCR products were normalized using agarose gel densitometry and pooled, gel purified, and concentrated using a DNA Clean and Concentrator Kit (Zymo, Irvine, CA). Pooled amplicons were quantified using Qubit Fluorometer 2.0 (Invitrogen, Carlsbad, CA). The pool was diluted to 4 nM and denatured with 0.2 N NaOH at room temperature. The denatured DNA was diluted to 20pM and spiked with 10% of the Illumina PhiX control DNA prior to loading the sequencer. Illumina paired-end sequencing was performed on the MiSeq platform using a 500 cycle version 2 reagent kit.*

Analysis of Illumina paired-end reads

Illumina Miseq paired-end reads were aligned to human genome reference genome hg19 with bowtie2 and matching sequences discarded [8]. As previously described, paired-end sequences were sorted by sample via indexes in the paired reads with a python script [6]. Sorted paired-end sequence data were deposited in the NCBI Short Read Archive under accession number SRP133576 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA436139>). The sorted paired reads were assembled using Phrap [9, 10]. Pairs that did not assemble were discarded. Assembled sequence ends were trimmed over a moving window of 5 nucleotides until average quality was met or exceeded 20. Trimmed sequences with more than 1 ambiguity or shorter than 250 nt were discarded. Potential chimeras identified with Uchime (usearch6.0.203_i86linux32) [11] using the Schloss Silva reference sequences [12] were removed from subsequent analyses. Assembled sequences were aligned and classified with SINA (1.3.0-r23838) [13] using the 418,497 bacterial sequences in Silva 115NR [14] as reference configured to yield the Silva taxonomy. Sequences with identical taxonomic assignments were grouped to produce operational taxonomic units (OTUs).

This process generated 28,453,582 sequences for 168 libraries (average sequence length 312 nt; average sample size 169,367 sequences/sample; minimum 17,660; maximum 441,328). The median Good’s coverage score was ≥ 99.86% at the rarefaction point of 17,660. The software package Explicet (v2.10.5, www.explicet.org) [15] was used for visual inspection of the sequence data and analysis (rarefied alpha diversity values).

Statistical analyses

Bacterial load was compared across the three groups, those who amplified with neither protocol, those who amplified only in the modified protocol, and those who amplified with both protocols, using a Kruskal-Wallis test. Samples where sequencing data was obtained using both the standard and the modified protocol were compared by evaluating differences in alpha diversity using paired *t*-tests. Beta diversity values between paired samples were evaluated using a one-sample *t*-test. Relative abundances for taxa were compared between protocols using generalized linear models with a negative binomial distribution, a log link, and generalized estimating equations with nested random effects (samples nested within subjects).

Results

Subject and samples

Sixteen subjects recruited within 24 h of intubation for mechanical ventilation support were included in this study. Each subject had tracheal aspirate samples collected daily at the first standard of care endotracheal tube clearance. DNA extractions were performed for 99 TA samples: 52 samples sequenced from both protocols, 24 samples sequenced from the modified protocol only, and 23 samples failed to amplify with either protocol. In addition, 10 samples were amplified in triplicate using each protocol to assess technical variability. Basic demographics for subjects included in this analysis are provided in Table 1.

Comparison of bacterial load

The median bacterial load in the 99 samples was 3.76 (range 3.13–6.22) log₁₀ rRNA copies per reaction. Seven total individuals had insufficient load to obtain sequencing (failed amplification, red dots), six individuals had at least one inadequate sample, and individual 15 failed

Table 1 Study participant demographic information

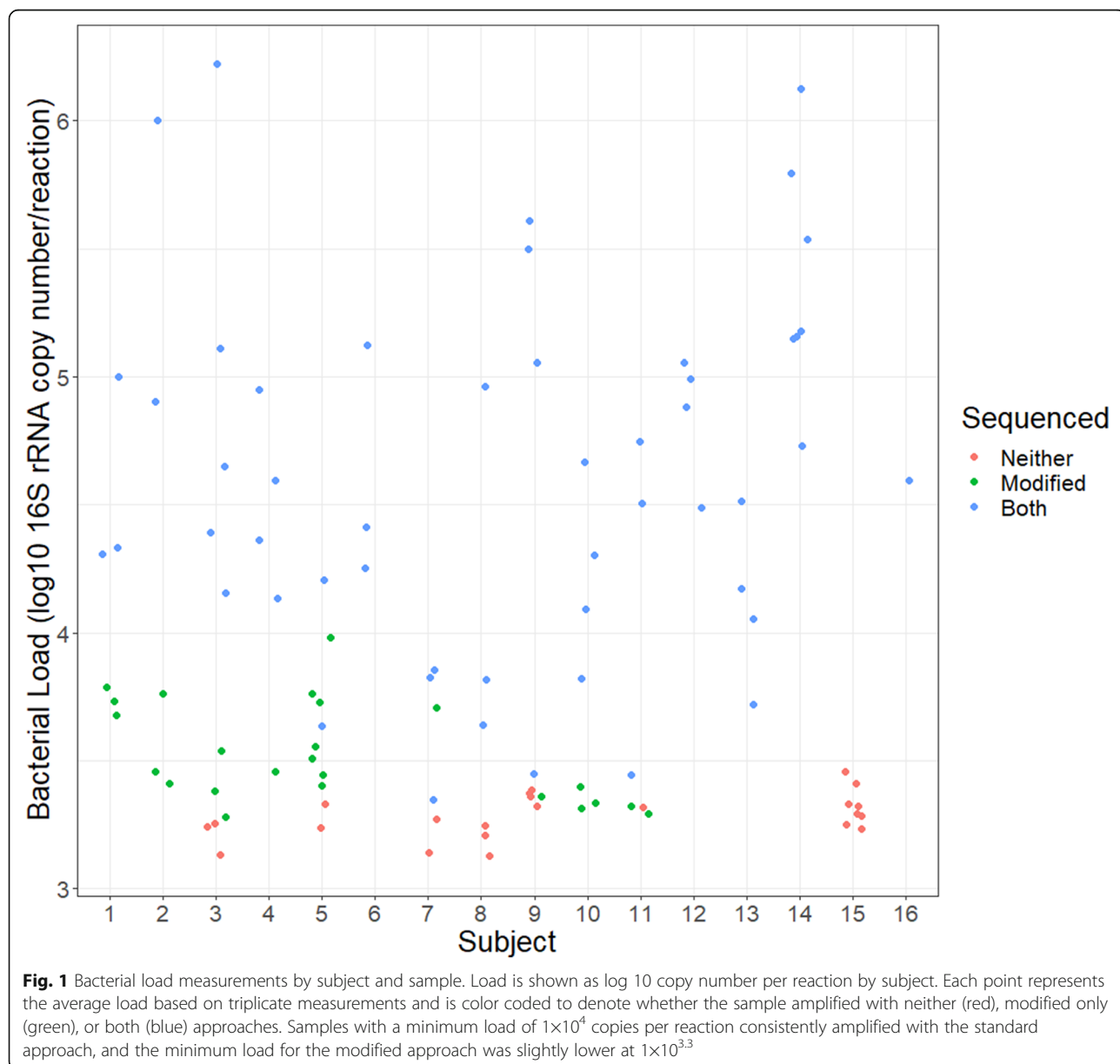
	Overall
<i>n</i>	16
Sex = male (%)	6 (37.5)
Age mean (SD)	3.98 (5.78)
On antibiotics during study period	16 (100)

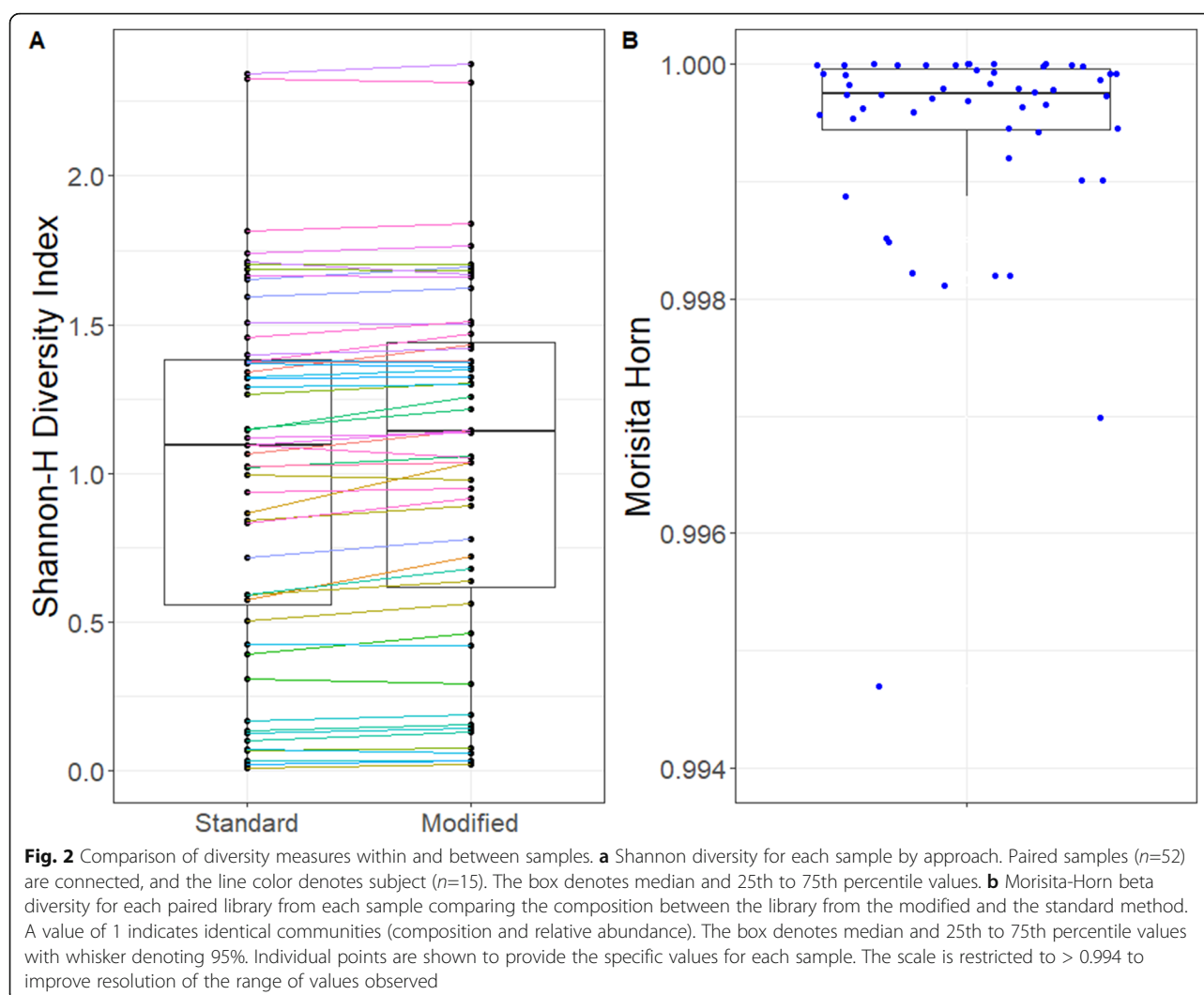
to amplify from any samples ($n=23$, Fig. 1 red dots). The remaining samples are shown based on whether both protocols (Fig. 1 blue dots) or the modified protocol only (Fig. 1 green dots) resulted in amplification. The transition to requiring the modified protocol for amplification was not consistent between subjects, but approximately 1×10^4 copies per reaction was required for the standard protocol to work consistently ($\sim 5 \times$ background). Not surprisingly, bacterial load was significantly higher in samples that successfully amplified with both protocols compared to those that only amplified with the modified protocol (median both 4.55 (IQR [4.15, 5.05]) \log_{10} rRNA copies per reaction, median modified 3.46 (IQR [3.38, 3.71]) \log_{10} rRNA copies per

reaction; p -value < 0.01). Samples that amplified with only the modified protocol had a significantly higher load than those samples that did not amplify with either protocol (median neither 3.29 (IQR [3.24, 3.33]) \log_{10} rRNA copies per reaction; p -value < 0.01).

Comparison of modified and standard protocols for paired samples

Both alpha and beta diversity measures indicate that the community composition is similar between pairs in the 52 samples that amplified with both the standard and modified protocol (Fig. 2). The relative abundances for the prominent taxa are higher when using the standard protocol (more points below the line at high RA, Fig. 3).





Only one taxon was identified as statistically different using the model-based approach (Table 2). *Enterobacteriaceae* had an estimated 1.9% relative abundance in the standard protocol and an estimated 2.4% relative abundance in the modified protocol. Samples from 10 subjects had a relative abundance of at least 1% for *Enterobacteriaceae*; further, only half of the samples were found to have *Enterobacteriaceae* detected (prevalence = 52%).

Evaluation of samples obtained only from the modified protocol

There were 24 samples from 9 subjects that only amplified using the modified protocol. Each of these 9 subjects had samples collected from other time points that did amplify using the standard protocol. Community comparisons from samples collected within subjects had consistent composition over time (Fig. 4).

Examination of background with the modified protocol

PCR and DNA extraction controls (16 each, 32 total libraries sequenced across two amplicon pools) were evaluated using the modified method to assess the modification on background. PCR controls utilized DEPC-treated water, and the DNA extraction controls were performed using the same DEPC-treated water extracted with the Qiagen EZ1 platform (methods). There was no apparent amplification in any of these controls, and they were included in the pool with the maximum volume used for samples. The raw sequence counts obtained from control libraries was low median 3146 sequences (IQR 1732–4612). Few taxa ($n=10$; 5 exclusive to DNA extraction, 2 exclusive to PCR) were present in all 32 replicate controls, which suggest the majority of taxa observed were derived from sources other than the reagents. The number of taxa observed in control libraries increased

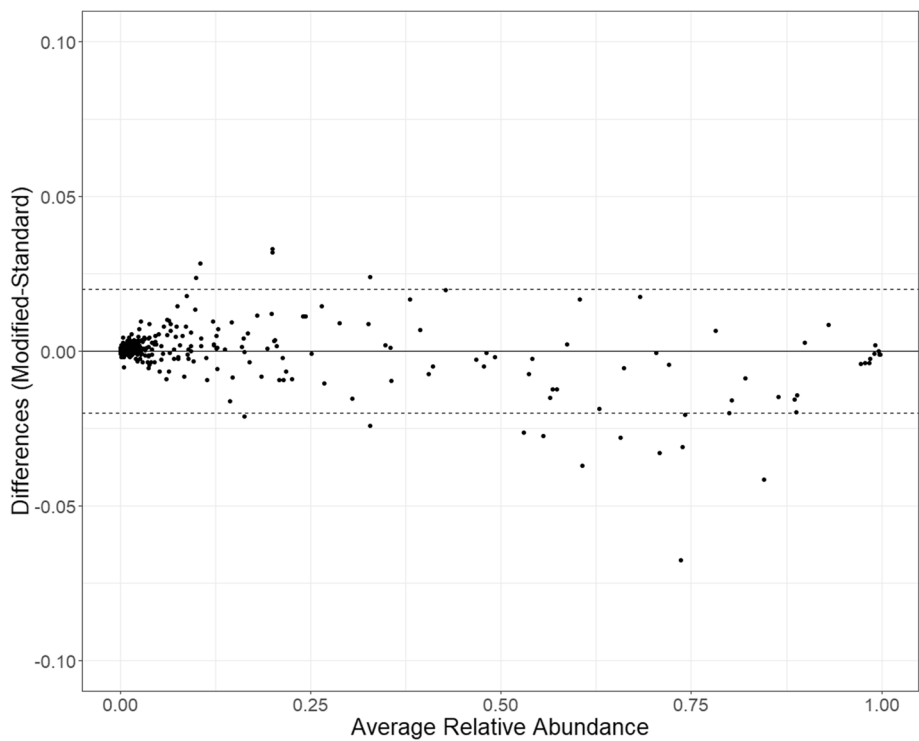


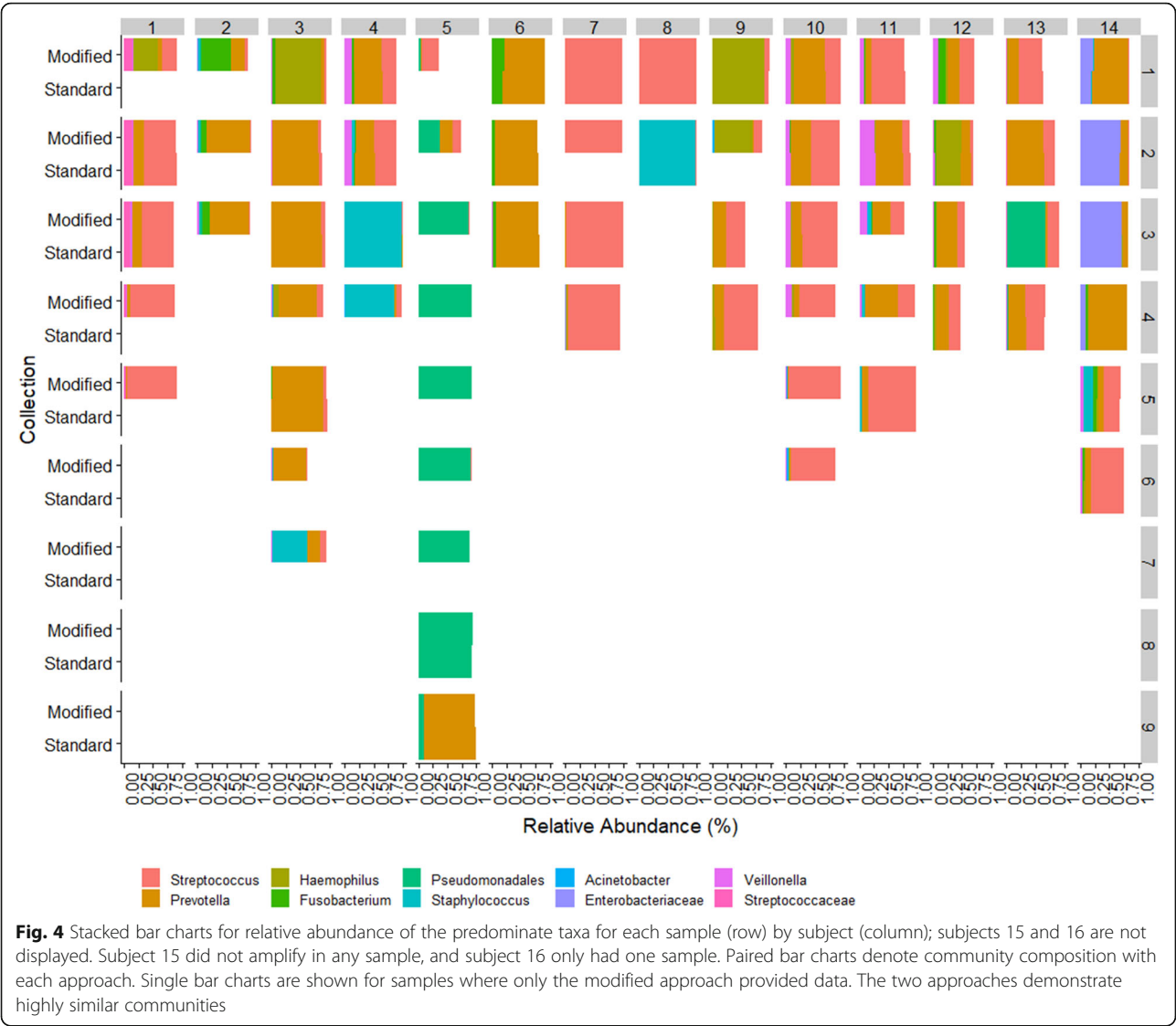
Fig. 3 Bland-Altman plot comparing the difference in relative abundance (modified–standard RA) versus RA for each taxon from paired samples. Taxa above zero indicate higher RA in the modified approach. The black solid line indicates no difference between the modified and standard approach. The black dotted lines represent a $\pm 2\%$ change

slightly ($n=27$; 11 exclusive to DNA extraction, 7 exclusive to PCR) if prevalence in the 32 control libraries was reduced to 75% (taxa observed in ≥ 24 libraries). There was a single prominent taxon in the DNA extraction controls (*Alcaligenaceae*), but otherwise, the relative abundance of taxa in the control libraries was generally variable (Fig. 5).

Reproducibility of protocols
Ten samples with adequate load were selected to amplify with both protocols using 3 independent PCRs (four additional libraries for each DNA). Figure 6 shows the relative abundance at which taxa were consistently detected in all three replicates for both protocols. For the standard and modified protocols, the relative abundance

Table 2 Parameter estimates for comparing relative abundances for specific taxa using generalized linear models with a negative binomial distribution, a log link, and generalized estimating equations with nested random effects

Parameter	Estimated relative abundance standard	Estimated relative abundance modified	Pr >
<i>Streptococcus</i>	0.261	0.266	0.853
<i>Prevotella</i>	0.234	0.219	0.603
<i>Haemophilus</i>	0.159	0.133	0.0858
<i>Fusobacterium</i>	0.079	0.066	0.6494
<i>Pseudomonadales</i>	0.025	0.071	0.0557
<i>Staphylococcus</i>	0.042	0.050	0.5286
<i>Actinobacteria</i>	0.031	0.025	0.7982
<i>Enterobacteriaceae</i>	0.024	0.019	0.0088
<i>Veillonella</i>	0.019	0.019	0.9239
<i>Streptococcaceae</i>	0.019	0.016	0.3846



at which taxa were repeatedly detected in all replicates was 0.01% and 0.027%, respectively (Figure S1).

Discussion

The modified protocol provided sequence data from 31.5% more samples compared to the standard protocol (52 vs. 76 of 99). For samples that produced sequencing results using both protocols, the bacterial communities were similar by alpha and beta diversity measures, but some variation was observed in the RA of more abundant taxa ($\leq 7\%$ RA). While *Enterobacteriaceae* RA was significantly different between protocols, this taxon was only observed in one individual making it challenging to generalize. Further, the absolute change in estimated RA was small (0.5%) and would not alter biological inferences, which is the primary factor for most analyses. Both protocols had similar reproducibility and the lower limits of RA where all replicates amplified were much lower than typical cutoffs used when evaluating specific taxa.

There is some controversy regarding the impact of the additional primer sequence on the bacterial community composition observed depending on the experimental design. Berry et al. [16] proposed an approach, termed 2-step, where standard primers were used in the first round of PCR to amplify sample DNAs, and then, a second round of PCR was performed to add the sequence platform-required oligonucleotides using the amplicons from the first round as template. They observed statistical differences in alpha diversity (Bray-Curtis, unweighted UniFrac). However, the prominent taxa identified did not change suggesting the alterations would not greatly impact biological conclusions. A limitation of this study was the absence of replication to determine how much influence biological and technical replicates introduce into the analysis. Our protocol modification emulated the

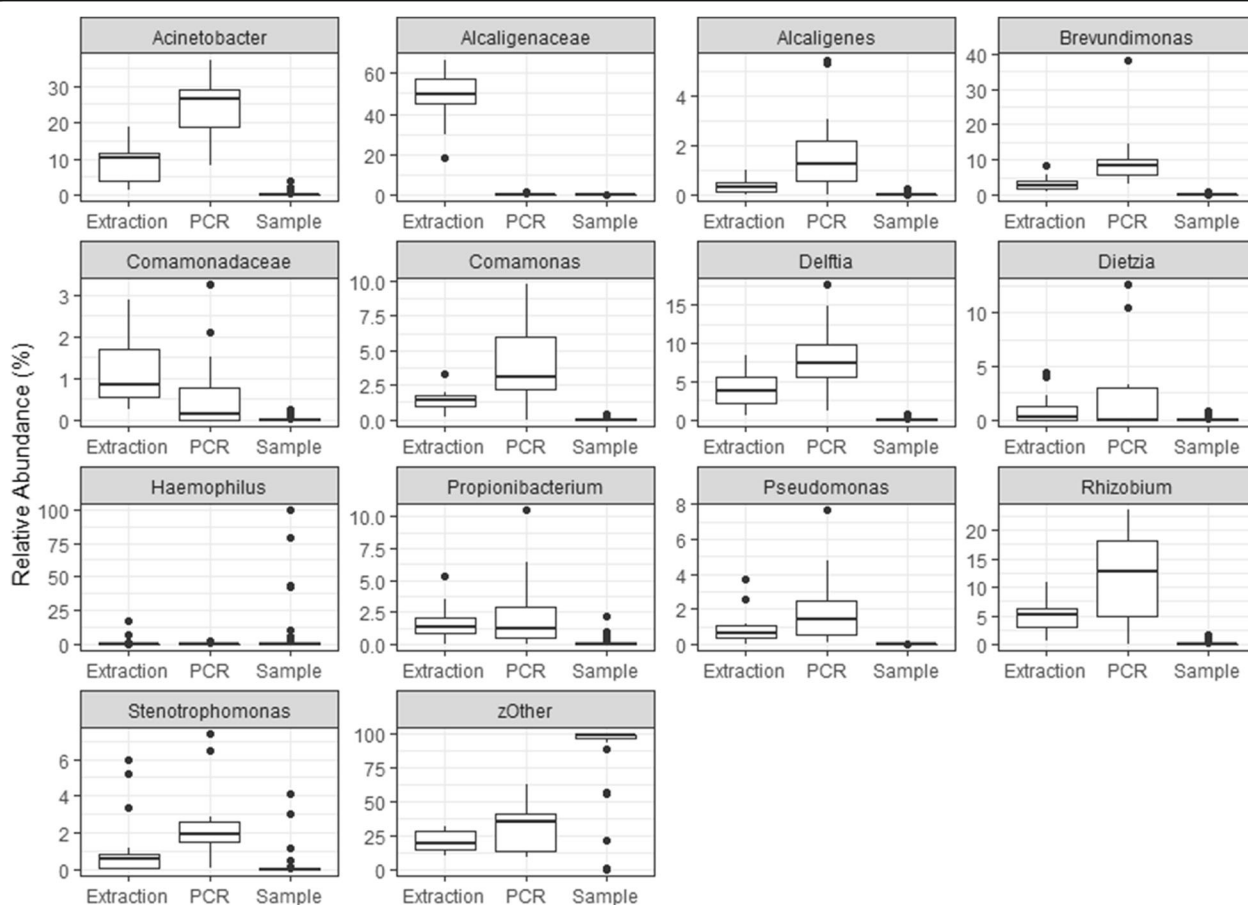


Fig. 5 Relative abundance of prevalent taxa within control samples. Boxplots show PCR and extraction controls for each taxon identified along with the distribution of RA from clinical samples. There is limited overlap between controls and clinical samples when comparing RA of these taxa

approach by Berry, but without running two separate PCR reactions. The experimental modification was to spike in 16S rRNA gene-specific primers at 10% of the standard primer concentration used for the longer sequencer-specific indexing primers. The rationale was the short primers would initiate PCR more efficiently in early cycles and increase the concentration of the 16S rRNA target sequences for the longer modified primers sufficiently to obtain adequate amplification for reliable sequencing.

The impact from background is a concern when attempting to amplify low load samples [3]. We demonstrated that the modification did not introduce significant background into the analysis, which is largely due to our very conservative approach to PCR amplification (mainly limited cycle number). We eliminated 23% of the samples prior to sequencing due to multiple independent pieces of evidence that inadequate bacteria were present (low TBL; no amplification apparent from either PCR protocols) for reliable data collection (essentially QNS, quantity not sufficient). This is in contrast to published approaches that suggest obtaining reliable sequence data from all samples is feasible. We agree that it

is possible to obtain sequence data in all cases if adequate protocol deviation is tolerated (particularly excessive PCR cycles) but question the utility given the difficulty in attributing the source of sequences derived from this type of experiment. The high level of variation observed in our background experiments suggests that reagents were a minor source of background at lower PCR cycle numbers, and the most likely source of significant portions of the background sequence is due to the samples amplified in parallel [3]. The DNA template added to adjacent wells represent the largest potential exposure, and that is why we utilize a negative PCR control for all samples (in contrast to plate level controls). This control does not guarantee that transfer between wells does not occur but does allow us to assess the magnitude of potential cross-contamination and take appropriate measures to mitigate in cases with clear contamination.

We observed a non-significant increase in alpha diversity, which may raise concerns about background. The increased efficiency of the PCR should yield an increase in richness particularly due to better sampling of rare



taxa; improved efficiency simply provides access further out the rank abundance curve independent of the source of DNA. Our replicate data demonstrates excellent agreement within technical replicates to RA levels much lower than that typically used of specific taxa (~0.005% RA versus >1% RA for specific taxa). Pragmatically, interpretation of low RA observations is challenging independent of the source (sample or background). These data demonstrate very good agreement for typically used metrics and provided data for 24 additional samples (51% of samples that failed to amplify initially) without major impacts to data integrity.

These data address several critical questions. First, inclusion of non-adapter-linked primers in the primary PCR could inhibit the ability to obtain sequence data. Amplicons without sequencing adapters are likely to co-purify with the target amplicons during processing steps

required for sequencing. We did appear to lose some cluster density with the modified protocol, but with limited sequencing runs, it is difficult to examine this critically. Based on the minimum library size (>17,000 sequences) and Good's coverage >99.86%, this was not an issue. Second, the anticipated mode of action for the modified protocol was to increase efficiency of early rounds of PCR, which should improve sensitivity. We had observed a shift of approximately three cycles in Ct when using the different primers (with and without sequencing adapters) in simple qPCR experiments (data not shown), which suggested about a 10-fold loss of efficiency for the primers with sequencing adaptors. These observations were the basis for the experiment described here, which does support the expected mode of action for the experimental intervention. Third, it was not known if increased efficiency in early rounds of PCR

would impact the community composition observed in the experiment. Variation in early amplification is known to impact amplicon composition and is the reason we run multiple independent reactions and pool them to mitigate any potential impact from this known issue [17]. The high beta diversity values observed indicate very good agreement between the two approaches. Further, the technical replicates demonstrated highly reproducible results from both PCR protocols.

Conclusion

We show that both protocols identify similar communities with typical airway composition, and the technical variability of both protocols was very low. The modified protocol was successful in amplifying more samples compared to the standard protocol especially for those samples with intermediate bacterial load. This is likely due to the improved efficiency of PCR using non-fusion primers. Therefore, our modified amplification protocol provides a viable alternative for increasing the sensitivity of bacterial community analysis in specimens with intermediate bacterial load without impacting data integrity.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-020-00958-y>.

Additional file 1: Figure S1. Lower limit of detection from triplicates.

The y axis is the number of triplicates in which a taxon was detected and the x-axis is the average RA across the triplicates for each taxon and sample. Points are jittered along the y-axis to better display the spread of points. Black dots correspond to the modified approach and blue dots to the standard approach. A) displays the lower range of RA values and B) displays the full range.

Authors' contributions

Conception and design: JH. Drafting of manuscript: KW, BW, JH. Review and revision of manuscript: KW, BW, CR, MJS, MKS, PM, JH. Data collection: CR, MJS, JH. Analysis and data interpretation: KW, BW, PM, JH. The authors read and approved the final manuscript.

Funding

Supported in part by the following cooperative agreements from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), and Heart Lung Blood Institute (NHLBI), National Institutes of Health (NIH): UG1HD083171 and (Dr. Mourani), 1R01HL124103 (Drs. Mourani and Sontag) UG1HD049983 (Dr. Carcillo), UG01HD049934 (Drs. Reeder, Locandro, and Dean), UG1HD083170 (Dr. Hall), UG1HD050096 (Dr. Meert), UG1HD63108 (Dr. Zuppa), UG1HD083116 (Dr. Sapru), UG1HD083166 (Dr. McQuillen), and UG1HD049981 (Dr. Pollack). The study sponsors were not involved in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the report for publication.

Availability of data and materials

Sorted paired-end sequence data were deposited in the NCBI Short Read Archive under accession number SRP133576 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA436139>).

Declarations

Ethics approval and consent to participate

This study follows an IRB-approved protocol (COMIRB 14-1530). Written informed consent and HIPPA authorization were obtained from all subjects or their legal guardians if <18 years of age.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado School of Medicine, 13001 17th Place, Mail Stop B119, Aurora, CO 80045, USA. ²Section of Critical Care, Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, 13123 E. 16th Ave. Box B395, Aurora, CO 80045, USA. ³Division of Infectious Diseases, School of Medicine, University of Colorado, 12700 East 19th Avenue, Mail Stop B168, Aurora, CO 80045, USA. ⁴Department of Epidemiology, Colorado School of Public Health, University of Colorado School of Medicine, 13001 17th Place, Mail Stop B119, Aurora, CO 80045, USA.

Received: 6 July 2020 Accepted: 6 December 2020

Published online: 13 April 2021

References

- Zemanick ET, Wagner BD, Sagel SD, Stevens MJ, Accurso FJ, Harris JK. Reliability of quantitative real-time PCR for bacterial detection in cystic fibrosis airway specimens. *PLoS One*. 2010;5(11):e15101. <https://doi.org/10.1371/journal.pone.0015101>.
- Nadkarni MA, Martin FE, Jacques NA, Hunter N. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology*. 2002;148(1):257–66. <https://doi.org/10.1099/00221287-148-1-257>.
- Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12(1):87. <https://doi.org/10.1186/s12915-014-0087-z>.
- Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, et al. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. *Proc Natl Acad Sci U S A*. 2007;104(51):20529–33. <https://doi.org/10.1073/pnas.0709804104>.
- Hara N, Alkanani AK, Ir D, Robertson CE, Wagner BD, Frank DN, et al. Prevention of virus-induced type 1 diabetes with antibiotic therapy. *J Immunol*. 2012;189(8):3805–14. <https://doi.org/10.4049/jimmunol.1201257>.
- Markle JGM, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolfe-Kampczyk U, et al. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science*. 2013;339(6123):1084–8. <https://doi.org/10.1126/science.1233521>.
- Zemanick ET, Wagner BD, Robertson CE, Ahrens RC, Chmiel JF, Clancy JP, et al. Airway microbiota across age and disease spectrum in cystic fibrosis. *Eur Respir J*. 2017;50(5):1700832. <https://doi.org/10.1183/13993003.00832-2017>.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9. <https://doi.org/10.1038/nmeth.1923>.
- Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error Probabilities. *Genome Res*. 1998;8(3):186–94. <https://doi.org/10.1101/gr.8.3.186>.
- Ewing B, Hillier L, Wendl M, Research PG-G. Undefined. Transcriptome analysis of *Schistosoma mansoni* larval development using serial analysis of gene expression (SAGE); 1998.
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. 2011;27(16):2194–200. <https://doi.org/10.1093/bioinformatics/btr381>.
- Schloss PD, Westcott SL. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol*. 2011;77(10):3219–26. <https://doi.org/10.1128/AEM.02810-10>.

13. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*. 2012;28(14):1823–9. <https://doi.org/10.1093/bioinformatics/bts252>.
14. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(Database issue):D590–6. <https://doi.org/10.1093/nar/gks1219>.
15. Robertson CE, Harris JK, Wagner BD, Granger D, Browne K, Tatem B, et al. Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics*. 2013;29(23):3100–1. <https://doi.org/10.1093/bioinformatics/btt526>.
16. Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl Environ Microbiol*. 2011;77(21):7846–9. <https://doi.org/10.1128/AEM.05220-11>.
17. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol*. 2005;71(12):8966–9. <https://doi.org/10.1128/AEM.71.12.8966-8969.2005>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

