A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing

A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing

Wenke Smets[a], Jonathan W. Leff[b,c], Mark A. Bradford[d], Rebecca L. McCulley[e], Sarah Lebeer[a], Noah Fierer[b,c*]

[a]Department of Bioscience Engineering, University of Antwerp, Antwerp, Belgium

[b]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA

[c]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA

[d]School of Forestry and Environmental Studies, Yale University, New Haven, CT, USA

[e]Department of Plant and Soil Sciences, University of Kentucky, Lexington, KY, USA

*Corresponding author: Noah Fierer

Address: University of Colorado, 216 UCB, CIRES, Boulder, CO 80309-0216, USA.

Phone number: 303-492-5615.

Email address: Noah.fierer@colorado.edu.

Keywords: microbial community analyses, 16S rRNA gene sequencing, bacterial abundances, soil bacteria

Abstract

Many recent studies rely on 16S rRNA-based sequencing approaches to analyze bacterial or archaeal communities found in soil and other environmental samples. While this approach is valuable for determining the relative abundances of different microbial taxa found in a given sample, it does not provide information on how the abundances of targeted microbes differ across samples. Here we demonstrate how the simple addition of an internal standard at the DNA extraction step allows for the quantitative comparison of how the total abundance of bacterial 16S rRNA genes varies across samples. The reliability of this method was assessed in two ways. First, we spiked a dilution series of two different soils with internal standards to ascertain whether we could accurately quantify differences in cell abundances. We tested two different internal standards, adding DNA from *Aliivibrio fischeri* or *Thermus thermophilus*, bacterial taxa unlikely to be found in soil. The total abundances of 16S rRNA genes in soil were calculated from the number of 16S rRNA genes of the internal standard recovered in the sequence data. Both standards allowed us to accurately quantify total gene abundances in soil as there was a strong positive correlation between total 16S rRNA gene estimations and the different starting amounts of soil extracted. We then tested whether we could use this approach to quantify differences in microbial abundances across a wide range of soil types; comparing estimated 16S rRNA gene abundances measured using this approach to microbial biomass determined with more standard methods: phospholipid fatty acid (PLFA) analysis and substrate induced respiration (SIR) analysis. The gene abundances estimated with the internal standard sequencing approach were significantly correlated with the independent biomass measurements, and were in fact better correlated to SIR and PLFA estimates than either of these two biomass measurements were correlated with one

another. Together, these results demonstrate that adding a DNA internal standard to soil or other environmental samples prior to DNA extraction is an effective method for comparing bacterial 16S rRNA gene abundances across samples. Given the ease of adding DNA internal standards to soil samples prior to high-throughput marker gene sequencing, 16S rRNA gene abundances and bacterial community composition can now be determined simultaneously and routinely.

1. Introduction

High-throughput sequencing has revolutionized the field of soil microbial ecology. In particular, 16S rRNA gene sequencing has provided unprecedented insight into the diversity of soil microbial communities, and it is now commonly used to characterize bacterial and archaeal communities in soil or related environments. However, one important limitation of this approach is that it only provides estimates of the proportional abundances of taxa – it provides no information on how the total amounts of microbial DNA may vary across samples. This can cause problems in the interpretation of results. For example, consider two samples that both have 20% of 16S rRNA sequences assigned to the phylum *Acidobacteria*. Such information can be used to assess how the abundance of *Acidobacteria* compares to other taxa, but this does not tell us anything about the total numbers of bacteria in these two samples and one sample could have far more *Acidobacteria* than another if it had a higher total number of bacterial cells. Likewise, two samples could appear to have very different proportional abundances of 16S rRNA reads assigned to a bacterial taxon of interest (e.g., a bacterial pathogen), but these two samples could actually have the same number of cells belonging to that taxon. Changes in bacterial community composition can also be difficult to interpret when relying on proportional abundances since it is often

difficult to differentiate between one taxon increasing in abundance and another decreasing in abundance. Thus, the proportional nature of marker gene datasets makes it impossible to compare how the total amounts of microbes (or their marker genes) vary across environmental samples that could have very different total cell numbers.

While there are a myriad of non-nucleic acid techniques for estimating total microbial biomass in soil or other environments (including phospholipid fatty acid analysis, substrate induced respiration, chloroform fumigation, and direct counting) – these methods are difficult to relate directly to the relative abundances as determined by 16S rRNA gene sequencing. This is due to the fact that these methods do not provide direct information on the amounts of 16S rRNA gene copies in a given sample, but instead represent other metrics of cell abundances (e.g., amounts of phospholipid fatty acids, amounts of chloroform-extractable microbial biomass carbon, or numbers of visible cells). Perhaps more importantly, these approaches for estimating soil microbial biomass can require significant added effort, they may require analyzing fresh (unfrozen) samples, and they often do not discriminate between prokaryotes and other organisms that can be abundant in soil (like fungi, protists, plants, or soil fauna). Likewise, total DNA yield is not typically considered a useful estimate of bacterial biomass as much of the DNA in soil comes from soil organisms that are not bacteria (Leckie et al., 2004) and DNA extraction efficiencies can vary dramatically across soil types (Frostegård et al., 1999; Cruaud et al., 2014). While quantitative PCR is frequently used to determine the number of 16S rRNA genes in a given sample (e.g. Fierer et al. 2005), this requires an additional step in the analyses, and the quantitative PCR analyses are also subject to biases due to differences in DNA extraction efficiencies and varying DNA amplification efficiencies across samples (Martin-Laurent et al., 2001; Fierer et al., 2005).

Clearly it is useful to have a method in place to directly relate 16S rRNA gene sequence data to estimates of the total amount of 16S rRNA genes found in a given sample to improve our understanding of the changes in taxon abundances underlying spatial or temporal differences in community composition. Here, we describe a method that couples sequencing-based analyses of the 16S rRNA gene for the assessment of microbial diversity with the determination of the variability in 16S rRNA gene abundances across samples, thus allowing the comparison of the actual abundances of different microbial taxa across samples, not just their proportional abundances. The approach involves adding DNA from a bacterium unlikely to be found in soil to the soil sample at the DNA extraction step. We then use the proportional representation of 16S rRNA reads from this organism to calculate how the total abundances of 16S rRNA genes vary across samples (Fig. 1). By adding this simple step to pipelines for 16S rRNA gene analyses, one can simultaneously compare how the diversity and abundances of soil bacteria vary across samples.

2. Materials and methods

2.1 General Description of the Approach

Our general approach is outlined in Fig. 1. For convenience, we define cell abundances in this manuscript as 16S rRNA gene abundances, although we recognize that the conversion from 16S rRNA gene numbers to bacterial cell abundances should be done with caution as the number of 16S rRNA gene copies per cell can vary from one to fifteen (Lee et al., 2009). Prior to DNA extraction, we add an internal standard to permit estimation of 16S rRNA gene abundances across soil samples. This internal standard is DNA from a bacterial strain unlikely to be found in soil and the same amount of internal standard DNA is added to each sample prior to DNA extraction. After DNA extraction, a portion of the 16S

rRNA gene is PCR amplified following a standard protocol described previously (Caporaso et al., 2012) with sequencing of the amplicon DNA pool conducted on the Illumina platform. From the DNA sequence data, we used the percentage of 16S rRNA reads of the internal standard versus the 16S rRNA reads from the native soil bacterial community to estimate the total abundance of 16S rRNA genes in a given sample.

2.2 Internal standard

The internal standards were aliquots of DNA extracted from bacterial strains that do not typically occur in soil samples and are not closely related to any common soil bacterial taxa. We selected two such taxa, *Aliivibrio fischeri* and *Thermus thermophilus*, as these taxa are not considered typical soil inhabitants, but instead are commonly associated with marine animals (Visick and McFall-Ngai, 2000) and geothermal environments (Oshima and Imahori, 1974), respectively. We checked whether these species were indeed uncommon in soil by looking for 16S rRNA sequences from these taxa in previously published datasets: 52 soils collected from across the globe and 52 soils from Central Park in New York City (Ramirez et al., 2014) plus a set of grassland soils collected from 25 sites from across the globe (Prober et al., 2015). Both bacterial strains were purchased from DSMZ (Braunschweig, Germany). *A. fischeri* (DSM 507) was grown at room temperature on tryptic soy agar (BD, Franklin Lakes, NJ, USA) complemented with 3% of NaCl (Fisher Scientific, Pittsburgh, PA, USA) (Tavares et al., 2010). *T. thermophilus* (DSM 46338) was grown at 72°C on tryptic soy agar complemented with Bacto Agar (BD, Franklin Lakes, NJ, USA), 4 g L^{-1} of yeast extract (Sigma, St. Louis, MO, USA), and 3% of NaCl; the pH of this medium was adjusted to 7.5 (Wilquet et al., 2004). DNA from colonies was extracted using the PowerSoil® DNA Isolation Kit (MoBio, Carlsbad, CA, USA) following the manufacturer's instructions. DNA concentrations of these

internal standards were determined using the PicoGreen dsDNA assay (Invitrogen, Carlsbad, CA, USA).

2.3 Molecular Analyses

Before DNA extraction, genomic DNA (gDNA) of *A. fischeri* or *T. thermophilus* was added to the mixture of soil and the bead solution at the first step of the PowerSoil® DNA Isolation Kit (MoBio, Carlsbad, CA, USA). The amount of internal standard added to each of the soil samples was calculated using the average DNA extraction yield of the 116 distinct soil samples (described below), with DNA concentrations quantified via a PicoGreen dsDNA assay. We added gDNA from the internal standards at an amount equivalent to either 0.1% or 1.0% of this average DNA yield (28 µg DNA per gram soil across the subset of soils tested). Thus, the specific amount of internal standard DNA added to each soil was held constant at either 28 or 280 ng DNA per gram soil (for the 0.1 and 1.0% levels, respectively). These internal standard amounts were selected as we wanted to have enough 16S rRNA reads from the internal standard so we could estimate the relative abundance of these reads with confidence but not so many that we would be prevented from assessing the composition of the native soil bacterial communities.

After DNA extraction, a portion of the 16S rRNA gene was amplified using the 515f/806r primer set with the primers containing the appropriate Illumina adapters and 12-bp barcodes to permit multiplexed sequencing on the Illumina platform (Caporaso et al., 2012). This primer set is designed to amplify the V4–V5 region of the 16S rRNA gene from both Archaea and Bacteria and is commonly used for the taxonomic analyses of bacterial and archaeal communities found in environmental samples (Caporaso et al., 2012). PCR conditions followed those described previously (Caporaso et al., 2011). Each sample was

amplified in triplicate and the amplicons from each sample were pooled prior to normalization using the SequalPrep™ Normalization Plate Kit, 96 Well (Invitrogen). Amplicons were then pooled together and sequenced on the Illumina MiSeq platform at the University of Colorado Next Generation Sequencing Facility, generating 2 x 150 bp paired-end reads. The sequences were demultiplexed using a custom Python script ('prep_fastq_for_uparse_paired.py', available at: https://github.com/leffj/helper-code-for-uparse). The UPARSE pipeline (Edgar, 2013) was used to merge the demultiplexed sequences, conduct quality filtering, and cluster sequences into operational taxonomic units (OTUs). To merge the paired-end sequences, we set a minimum overlap of 20 bp with a maximum of one mismatch and merged reads had to to be more than 200 bp in length. A maximum per sequence expected error frequency value of 0.5 was used to quality-filter sequences. Clustering was done at the ≥97% similarity level, and the merged raw reads were mapped to the clustered *de novo* database at 97% similarity. Taxonomy was assigned using the RDP classifier (Wang et al., 2007) trained on the Greengenes database (version 13_08). Chloroplast and mitochondrial OTUs were removed and samples with < 2,500 reads were excluded from downstream analyses.

To determine the number of 16S rRNA genes found in each extracted soil sample (*x*) from the sequence data, we used the following equation:

$$\frac{R_i}{R_s} = \frac{\frac{w_i}{g_i} * c_i}{\mathrm{x}} \quad \Rightarrow \quad \mathrm{x} = \frac{R_s * (\frac{w_i}{g_i} * c_i)}{R_i}$$

where R_i is the number of reads assigned to the bacterial taxon used as the internal standard (either *A. fischeri* or *T. thermophilus*), R_s is the number of reads assigned to other taxa that were found in the soil, w_i is the weight of internal standard gDNA added to the samples, g_i is

the weight of the genome of the internal standard, c_i is the 16S copy number of the internal standard, and *x* is the (initially unknown) number of 16S rRNA genes per sample. To calculate *x*, we assumed a 16S rRNA gene copy number of eight and a weight of $4.49*10^{-15}$ g per *A. fischeri* genome, and a 16S rRNA gene copy number of two and a weight of $2.16*10^{-15}$ g per *T. thermophilus* genome (Dolezel et al., 2003; Hallin and Ussery, 2004). The DNA of the internal standard was added before DNA extraction so that if DNA extraction yields varied between samples, such variation would not interfere with this calculation as long as we assume that the recovery of DNA from the internal standard and the recovery of DNA from the native bacterial cells are similar. Even when this condition is not met, the comparison of bacterial 16S rRNA gene abundances between samples should be unaffected by higher (or lower) extraction efficiencies of the internal standard than of the soil bacteria as long as the bias in extraction efficiencies is held reasonably constant across all of the samples tested. That is, as long as the efficiency of recovering the DNA from the internal standard is similar across all of the samples examined, we should still be able to quantify changes in total 16S rRNA gene abundances across a collection of soil samples.

2.4 Testing the approach with a soil dilution series

To test whether the 16S rRNA gene abundance estimates were correlated with the amounts of starting material analyzed , we added standards to different amounts of the same soil, using a 5-fold dilution series. As it is difficult to directly weigh out and quantitatively extract DNA from increasingly small amounts of soil, we effectively extracted DNA from soil samples ranging in size from 0.03 g to 0.50 g by serially diluting two individual soils in buffer and then extracting DNA from each of the dilutions. We generated serial dilutions from two different soil types. One grassland soil was collected at a depth of

approximately 20 cm from Table Mountain near Boulder, CO (40.13 °N, 105.24 °W) on January 28, 2015. Another soil was collected from the main campus of the University of Colorado at Boulder (40.01 °N, 105.27 °W), at a depth of approximately 5 cm, on February 27, 2015. Samples were stored at -20°C until DNA extraction.

We made a dilution series of each soil type by suspending the soil in DNA-protecting buffer (the 'Bead Solution' from the MoBio DNA isolation kit), homogenizing the slurry by vortexing, and serially diluting the slurries. Negative controls consisting only of the bead solution were also included. We added gDNA from either *A. fischeri* or *T. thermophilus* to each individual extraction tube. The amounts corresponded to either 0.1% or 1% of total DNA that we expected to extract from the soil sample, adding the same amount of gDNA from *A. fischeri* or *T. thermophilus* to all soils included in each dilution series. We added four different amounts and types of internal standard DNA (0.1% *A. fischeri*, 0.1% *T. thermophilus*, 1% *A. fischeri* and 1% *T. thermophilus*), with all samples processed in triplicate yielding 60 samples in total. We compared the weights of the extracted soil samples from the dilution series to the estimated 16S rRNA gene numbers using linear regression analyses as implemented in the R software package (R Development Core Team, 2012).

2.5 Testing the approach across a wide range of soil types

We assessed the efficacy of the internal standard method by comparing estimates of 16S rRNA abundances obtained with this method to microbial biomass estimates obtained using PLFA and SIR on a wide range of soils. In addition, we used these data to determine whether adding gDNA from the internal standards altered assessments of bacterial and archaeal community composition across the soils. We used 116 soil samples collected from 10 locations across the United States (Fig. S1; forested and non-forested samples from each

location). These soils were chosen as they represent a broad range in edaphic properties and microbial biomass levels (Table S1 and Table S2). Individual samples from each location were exposed to glucose and temperature manipulation experiments (Table S1) to test our methodology in a range of soil sample types incubated under different conditions.

Microbial biomass was measured in these samples using two different non-nucleic acid methods that are commonly used by soil ecologists. PLFA analyses were used to determine bacterial biomass per gram of dry soil, as described previously (Crowther et al., 2014), by summing the abundance of 36 individual fatty acid methyl esters identified by MIDI software (Microbial Identification Inc., Newark, Delaware) as being produced by either gram-negative or gram-positive bacteria and actinomycetes. In addition, we used the SIR method that estimates microbial biomass from measured rates of microbial respiration in soils amended with a labile carbon substrate following Fierer et al. (2003).

To assess potential effects of internal standard additions on the determination of bacterial community composition in the samples, DNA from each soil sample was extracted twice: once without the DNA from the internal standard added and once with 7 ng of internal standard DNA added (which corresponds to approximately 0.1% of the average DNA yield of a sample). For these tests, we used gDNA from *A. fischeri* as the internal standard. The estimated concentration of 16S rRNA genes was determined using the internal standard method described above. Spearman rank correlation coefficients between PLFA or SIR estimates of microbial biomass and our estimates of 16S rRNA gene abundances were determined using R software package (R Development Core Team, 2012). We used Spearman rank correlations as the data were not normally distributed and did not meet the assumption of homogeneity in variance.

To assess whether the addition of an internal standard affects the determination of community composition, the internal standard reads (those reads classified as *A. fischeri*) were removed from the original data, and the samples were rarefied by sub-sampling a set number of 16S rRNA sequences per sample to control for differences in library size (Goodrich et al., 2014). The composition of the communities in the spiked versus the non-spiked replicates of each soil sample were compared using the Mantel test (10000 permutations) with differences in community composition assessed using pair-wise Bray-Curtis distances (using the vegan package in R (vegan.r-forge.r-project.org/)). In addition, the relative abundances of the main phyla and sub-phyla were compared between the soils with and without internal standard. The corresponding linear models were evaluated and the median and range of the relative abundances of each of the dominant bacterial groups were calculated for the group of samples receiving internal standard and for the group of samples that did not receive the internal standard DNA.

3. Results and Discussion

3.1 Selection of internal standards to add to soil

We confirmed that our sources of internal standard DNA, from *A. fischeri* and *T. thermophilus*, do not commonly occur in soil by looking for these taxa (or closely related taxa) in previously published datasets that span a wide range of soil types. No sequences >97% similar to 16S rRNA gene sequences from *A. fischeri* and *T. thermophilus* were found in any of these collections of soil samples. In addition, when we screened the 116 soil samples used to test this method (Table S1), 16S rRNA sequences >97% similar to *A. fischeri* or *T. thermophilus* were very rare, averaging < 0.0028% and < 0.0003% of all 16S rRNA sequences obtained from these samples, respectively. Hence, we assume that the abundance of 16S

rRNA gene sequences from the addition of these internal standards should be largely unaffected by the presence of these taxa native to the soil samples.

3.2 Dilution series

The dilution series were set up to assess whether the internal standard read abundance accurately reflects changes in the amount of soil extracted (from 0.03 g to 0.5 g of soil). To verify whether the method works in different conditions, two types of internal standards were added to two different soils. The linear regressions (Fig. 2) show that the amount of soil extracted and the estimated 16S rRNA gene abundances, determined using our internal standard approach, were well correlated (R^2 = 0.67-0.96, $P < 0.001$). These correlations were found for both internal standard strains (either *A. fischeri* or *T. thermophilus*) and regardless of whether we added 0.1% or 1% of internal standard DNA to each sample (Fig. 2).

These results indicate that the abundances of internal standard reads allowed us to estimate expected differences in soil rRNA gene abundances, a proxy for microbial cell biomass. The estimates of total 16S rRNA gene abundances did, however, vary depending on the bacterial strain used for the internal standard (Fig. 2). These different estimates are likely due to differences in DNA extraction efficiencies and/or amplification efficiencies between the two internal standard strains. Hence, the different results of *A. fischeri* versus *T. thermophilus* indicate that all samples in a given study should be analyzed using the same strain as an internal standard to allow comparison of 16S rRNA gene abundances between samples.

3.3 Cross-sample analyses

Correlations between the two independent measures of microbial biomass, PLFA and SIR, and the 16S rRNA gene abundance estimates obtained using the internal standard/sequencing approach were compared for the 110 distinct soil samples (Fig. 3). Across all of the soils analyzed, there was a high degree of variability in estimated abundances, regardless of the metric used (Fig. 3). For example, PLFA estimates of bacterial abundances varied by 30-fold across all samples and estimates of 16S rRNA gene abundances varying by 210-fold across the samples (Fig. 3). This level of variability was expected given that the soils were collected from a wide variety of ecosystem types and represent a broad range in geographical locations, incubation conditions, and edaphic characteristics (Tables S1 and S2). We found significant correlations between 16S rRNA gene concentrations determined using the internal standard method and biomass estimates determined using SIR ($r = 0.57$; $P < 0.001$) and PLFA analysis ($r = 0.71$; $P < 0.001$; Fig. 3).

PLFA quantifies cell membrane constituents and is a commonly used method for determining microbial biomass in soils and it allows for the discrimination of biomass derived from bacteria versus other soil organisms (Frostegård et al., 2011). The SIR method measures catabolic activity (CO_2 production shortly after addition of a labile substrate) and does not discriminate between bacteria and other soil organisms, like fungi, that could account for a large proportion of catabolic activity. In addition, SIR does not detect bacteria that are alive, but not capable of rapidly catabolizing the added substrate. The PLFA values can therefore be considered a better reflection of total bacterial biomass than SIR (Bailey et al., 2002) and indeed our PLFA estimates of bacterial biomass showed a stronger correlation with the estimated 16S rRNA gene abundances. Interestingly, the estimates of 16S rRNA gene abundances were better correlated with both PLFA and SIR data than these non-molecular biomass estimates were correlated with one another. It is nearly impossible to

know true microbial abundances in soil as every one of these techniques has important limitations (including techniques not tested here, such as direct counting and chloroform fumigation-extraction (Bradford et al., 2009)), with previous work showing that the correlation in biomass estimates obtained using non-molecular methods is often surprisingly low (Bailey et al., 2002). Thus, these results highlight that using the internal standard to estimate cell abundances is at least as effective as other commonly-used methods to determine microbial abundances in soil. However, we note that the approach described here has the added advantage of simultaneously providing information on the diversity and composition of soil bacterial communities, with minimal added effort.

We tested whether the addition of an internal standard biases the determination of microbial community composition. If adding the genomic DNA for the internal standard strongly alters the apparent composition of the microbial communities, this approach may not be reliable for the simultaneous determination of both community composition and 16S rRNA gene abundances. After removing the *A. fischeri* reads from the spiked samples, all samples were rarefied to a sequencing depth of 9,067 16S rRNA reads per sample. We compared the pairwise dissimilarities between the 110 remaining samples that received the internal standard and the same samples that were processed in an identical manner but did not receive any internal standard. There was a very strong relationship between these dissimilarities ($r = 0.98$; $P < 0.001$) highlighting that our assessments of the overall composition of the microbial communities were generally unaffected by the addition of the internal standard DNA (Fig. 4). This was also evident when we compared the relative abundances of major bacterial phyla in soils processed with and without the internal standard. In general, the relative abundances of individual phyla were strongly correlated across the samples (Table 1) even though there is likely some inherent variability between

the replicate DNA extractions in our estimates of taxon abundances. Again, these results highlight that the addition of the *A. fischeri* gDNA as an internal standard did not alter our ability to assess the overall structure of the soil microbial communities.

3.4 Caveats

There are important caveats to consider when using this internal standard approach for determining total abundances of 16S rRNA genes in soil microbial communities. First, we are still only estimating the abundance of 16S rRNA genes in a given sample, and since bacteria can vary with respect to the number of 16S rRNA gene copies per cell, it remains difficult to infer the total number of cells per taxon or the total number of bacterial cells in a given sample. Furthermore, it is important to determine *a priori* how much internal standard DNA to add to each sample as our ability to detect 16S rRNA gene abundances is only as good as our ability to detect the internal standard in our sequence data. If the internal standard is too rare (i.e. represented by low number of sequences per sample), we would not be able to accurately estimate concentrations of 16S rRNA genes in a given sample. Likewise, if the internal standard was too abundant in the resulting libraries, we would not be able to assess microbial community composition due to an insufficient number of reads from the soil bacteria and archaea. The ideal amount of internal standard DNA to add depends on multiple factors, such as taxonomic richness, sequencing depth, and research questions. Thus, it is important to determine the most appropriate amount of internal standard DNA to add by first estimating the quantity of DNA in the samples. This can be done relatively quickly by extracting DNA from a subset of new samples and measuring soil DNA concentrations. Another important caveat is that the internal standard technique does not provide an accurate estimate of true absolute abundances, as both internal standards – i.e.

the DNA of *A. fischeri* and DNA of *T. thermophilus* – result in different absolute abundances for the same samples (Fig. 2). If the objective is to determine absolute abundances of a certain taxon in a sample, quantitative PCR with taxon-specific primers is the advised method for now, even though these abundance estimates can also be influenced by differences in amplification efficiencies, choice of primers, DNA extraction efficiencies, and other factors (Fierer et al., 2005; Feinstein et al., 2009; Smith and Osborn, 2009; Tremblay et al., 2015). Nevertheless, our work suggests it is possible to compare 16S rRNA gene abundances between samples by adding an internal standard prior to DNA extraction and amplicon sequencing as long as all samples being compared are analyzed using the same methods (including the same sampling strategy, internal standard, and bioinformatics pipeline).

4. Conclusions

Whereas until now, 16S rRNA gene sequencing only gave information on the relative abundances of taxa within a sample, with the internal standard it is now possible to estimate how the total amounts of 16S rRNA genes may vary across samples. The described method is likely to be broadly useful when conducting analyses of microbial studies in soil, and potentially other environments, where the ability to compare the total amounts of bacteria and archaea between samples is critical. Further testing is needed to determine the usefulness of our approach for other applications, such as when analyzing fungal communities via sequencing the internal transcribed spacer region (Schoch et al., 2012) or when analyzing microbial communities found in other environments (including the atmosphere, skin, or in the built environment).

Acknowledgments

We thank Jessica Henley for her assistance with sample processing and Tess Brewer for her assistance with lab work and graphic design.

This work was supported by grants to N.F., R.L. M., and M.A.B. from the U.S. National Science Foundation (DEB-1021222, DEB-1021098, DEB-0953331, and DEB-1021112) with funding to W.S. and S.L. provided from the University of Antwerp.

References

Bailey, V.L., Peacock, A., Smith, J.L., Bolton, H., 2002. Relationships between soil microbial biomass determined by chloroform fumigation–extraction, substrate-induced respiration, and phospholipid fatty acid analysis. Soil Biology & Biochemistry 34, 1385-1389.

Bradford, M.A., Wallenstein, M.D., Allison, S.D., Treseder, K.K., Frey, S.D., Watts, B.W., Davies, C.A., Maddox, T.R., Melillo, J.M., Mohan, J.E., 2009. Decreased mass specific respiration under experimental warming is robust to the microbial biomass method employed. Ecology Letters 12, E15-E18.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME Journal 6, 1621-1624.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences, USA 108, 4516-4522.

Crowther, T.W., Maynard, D.S., Leff, J.W., Oldfield, E.E., McCulley, R.L., Fierer, N., Bradford, M.A., 2014. Predicting the responsiveness of soil biodiversity to deforestation: a cross-biome study. Global Change Biology 20, 2983-2994.

Cruaud, P., Vigneron, A., Lucchetti-Miganeh, C., Ciron, P.E., Godfroy, A., Cambon-Bonavita, M.-A., 2014. Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems. Applied and Environmental Microbiology 80, 4626-4639.

Dolezel, J., Bartos, J., Voglmayr, H., Greilhuber, J., 2003. Nuclear DNA content and genome size of trout and human. Cytometry Part A, 51, 127-128.

Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods 10, 996-998.

Feinstein, L.M., Sul, W.J., Blackwood, C.B. 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. Applied & Environmental Microbiology 75: 5428-5433.

Fierer, N., Jackson, J.A., Vilgalys, R., Jackson, R.B., 2005. Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. Applied and Environmental Microbiology 71, 4117-4120.

Fierer, N., Schimel, J.P., Holden, P.A., 2003. Variations in microbial community composition through two soil depth profiles. Soil Biology & Biochemistry 35, 167-176.

Frostegård, Å., Courtois, S., Ramisse, V., Clerc, S., Bernillon, D., Le Gall, F., Jeannin, P., Nesme, X., Simonet, P., 1999. Quantification of bias related to the extraction of DNA directly from soils. Applied and Environmental Microbiology 65, 5409-5420.

Frostegård, Å., Tunlid, A., Bååth, E., 2011. Use and misuse of PLFA measurements in soils. Soil Biology & Biochemistry 43, 1621-1625.

Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R., Ley, R.E., 2014. Conducting a microbiome study. Cell 158, 250-262.

Hallin, P.F., Ussery, D.W., 2004. CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. Bioinformatics 20, 3682-3686.

Leckie, S.E., Prescott, C.E., Grayston, S.J., Neufeld, J.D., Mohn, W.W., 2004. Comparison of chloroform fumigation-extraction, phospholipid fatty acid, and DNA methods to determine microbial biomass in forest humus. Soil Biology & Biochemistry 36, 529-532.

Lee, Z.M., Bussema, C., Schmidt, T.M., 2009. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Research 37, D489-D493.

Martin-Laurent, F., Philippot, L., Hallet, S., Chaussod, R., Germon, J., Soulas, G., Catroux, G., 2001. DNA extraction from soils: old bias for new microbial diversity analysis methods. Applied and Environmental Microbiology 67, 2354-2359.

Oshima, T., Imahori, K., 1974. Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. International Journal of Systematic Bacteriology 24, 102-112.

Prober, S.M., Leff, J.W., Bates, S.T., Borer, E.T., Firn, J., Harpole, W.S., Lind, E.M., Seabloom, E.W., Adler, P.B., Bakker, J.D., 2015. Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. Ecology Letters 18, 85-95.

R Development Core Team, 2012. R: A language and environment for statistical computing.

Ramirez, K.S., Leff, J.W., Barberán, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., 2014. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. Proceedings of the Royal Society of London B: Biological Sciences 281, 20141988.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., Crous, P.W., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences, USA 109, 6241-6246.

Smith, C.J., Osborn, A.M., 2009. Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. FEMS Microbiology Ecology 67: 6-20.

Tavares, A., Carvalho, C., Faustino, M.A., Neves, M.G., Tomé, J.P., Tomé, A.C., Cavaleiro, J.A., Cunha, Â., Gomes, N., Alves, E., 2010. Antimicrobial photodynamic therapy: study of bacterial recovery viability and potential development of resistance after treatment. Marine Drugs 8, 91-105.

Tremblay J., Singh K., Fern A. Kirton, E.S., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J.F., Tringe, S.G., 2015. Primer and platform effects on 16S rRNA tag sequencing. Frontiers in Microbiology. 6:771.

Visick, K.L., McFall-Ngai, M.J., 2000. An exclusive contract: specificity in the *Vibrio fischeri*-*Euprymna scolopes* partnership. Journal of Bacteriology 182, 1779-1787.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73, 5261-5267.

Wilquet, V., Van de Casteele, M., Gigot, D., Legrain, C., Glansdorff, N., 2004. Dihydropteridine reductase as an alternative to dihydrofolate reductase for synthesis of tetrahydrofolate in *Thermus thermophilus*. Journal of Bacteriology 186, 351-355.

Table 1. Comparison of relative taxon abundances between samples processed with and without the internal standard (IS) added. Results were summarized for all bacterial phyla (and proteobacterial sub-groups) with mean relative abundances > 2.5%. The second column shows the correlation coefficients between relative abundances of the taxon in samples that received internal standards and those that did not. All relationships were significant (P < 0.001). Also shown are the median and range in relative abundances of the taxa in all of the samples with and without the internal standard added.

Taxon	R^2	Median without IS (%)	Median with IS (%)	Range without IS (%)	Range with IS (%)
Acidobacteria	0.838	15	18	3-37	3-37
Actinobacteria	0.646	7	5	1-22	1-16
Alphaproteobacteria	0.854	15	14	4-59	4-67
Bacteroidetes	0.725	7	7	0-22	0-17
Betaproteobacteria	0.896	6	6	1-37	1-41
Chloroflexi	0.937	4	3	0-41	0-36
Deltaproteobacteria	0.879	3	3	0-10	0-13
Firmicutes	0.914	1	1	0-49	0-32
Gammaproteobacteria	0.947	6	6	1-40	1-36
Verrucomicrobia	0.902	8	9	0-24	0-31
Other	0.898	14	13	4-46	4-41

Figure 1: Overview of the workflow. Two hypothetical soil samples, one with high and another with low total amounts of bacteria are processed with an internal standard. This method results in both an overview of community composition and an estimate of cell abundances per sample.

Figure 2. Relationships between the quantity of soil material extracted and the estimates of 16S rRNA gene abundances, based on the number of internal standard reads. Results from two dilution series of the 'Table Mountain' soil are shown in panel A: one with 0.1% *T. thermophilus* and the other with 0.1% *A. fischeri*. Results from two other dilution series of the 'campus' soil are depicted in panel B: one with 1% *T. thermophilus* and the other with 1% *A. fischeri*. The results for 0 g of soil were determined using the blanks consisting only of the dilution solution and internal standard.

Figure 3. Plots and corresponding Spearman rank correlation coefficients (r_S) of SIR, PLFA and millions of 16S rRNA genes. All relationships were significant ($P < 0.001$). Sample 71 was identified as an outlier using a Bonferroni test ($P < 0.001$) and it is marked in the plots.

Figure 4. The relationship in pairwise Bray-Curtis dissimilarities between the samples with and the samples without internal standard DNA added. The line indicates where dissimilarities are exactly the same for both with-internal-standard and without-internal-standard samples.

Figure S1: Map displaying the locations of the sampling sites. These locations (characteristics in Table S2) represent a wide range of different soils (Crowther et al. 2014).