Biol Fertil Soils (2017) 53:485–489
DOI 10.1007/s00374-017-1205-1

EDITORIAL

Analysis of soil microbial communities based on amplicon sequencing of marker genes

Anne Schöler[1] · **Samuel Jacquiod**[2] · **Gisle Vestergaard**[1] · **Stefanie Schulz**[1] · **Michael Schloter**[1,3]

Received: 10 January 2017 / Revised: 21 March 2017 / Accepted: 21 April 2017 / Published online: 4 May 2017

Abstract The use of cultivation independent methods has revolutionized soil biology in the last decades. Most popular approaches are based on directly extracted DNA from soil and subsequent analysis of PCR-amplified marker genes by next-generation sequencing. While these high-throughput methods offer novel possibilities over cultivation-based approaches, several key points need to be considered to minimize potential biases during library preparation and downstream bioinformatic analysis. This opinion paper highlights crucial steps that should be considered for accurate analysis and data interpretation.

Keywords Next-generation sequencing · Amplicon sequencing · 16S rRNA · Sequencing depth · Bioinformatics

Introduction

Most advances achieved in microbial ecology over the last decades are based on methodological progresses enabling a fine description of complex microbial community composition and function in their particular environment.

✉ Michael Schloter
schloter@helmholtz-muenchen.de

[1] Research Unit for Comparative Microbiome Analysis, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

[2] INRA Dijon, UMR1347 Agroécologie, Dijon, France

[3] Chair for Soil Science, Technische Universität München, Munich, Germany

Besides direct sequencing of the total extracted DNA (metagenome) or RNA via cDNA (metatranscriptome), targeted approaches are often used, which include amplification of single genes before sequencing. Due to its broad presence in all organisms, and its adequate level of conservation, the ribosomal operon is often used as "golden standard" for diversity inventories (Amann et al. 1995). Amplicon-based approaches targeting variable regions of specific markers (e.g., 16S, ITS, or 18S) are widely used to describe bacterial, archaeal, fungal (Lindahl et al. 2013), and micro-eukaryote community composition (Lentendu et al. 2014). This approach was transposed for functional studies, e.g., targeting enzyme-coding genes catalyzing C, N, and P cycles, for example, β-glucosidases (Pathan et al. 2015), protease genes (Baraniya et al. 2016), or alkaline phosphatases (Bergkemper et al. 2016).

In this opinion paper, we highlight major pitfalls in the generation, processing, and interpretation of amplicon sequencing data, keeping in mind key ecological concepts. For other issues including soil sampling strategies, DNA extraction, and metadata collection, we refer to Vestergaard et al. (2017).

Library preparation

When performing amplicon sequencing, several key points deserve consideration: Standard PCR conditions should be applied with minimized cycle numbers if possible (Schmidt and Rothhämel 2012) and DNA templates should be quantified, as 10–20 ng of template DNA is sufficient for amplification of ribosomal marker genes. Higher amounts might be required if the targeted genes are rare. Potential inhibition of the PCR may be assessed via preliminary tests using serial dilutions of template DNA. Moreover, to minimize PCR-introduced biases, it is recommended to perform technically

replicated PCR reactions for each sample, which are subsequently pooled before sequencing.

Including negative controls is essential, as recent studies highlighted microbial DNA contaminations originating from DNA extraction kits that can highly bias the obtained sequencing output (Lusk 2014; Salter et al. 2014). Contamination issues become most problematic when working with low DNA amounts.

Primer design

For the assessment of prokaryote diversity, well-established 16S rRNA gene primer pairs are available (Table 1). The use of de facto standard primers is recommended as it increases inter-study reproducibility and comparability (Caporaso et al. 2012). For the assessment of fungal diversity, the Internal Transcribed Spacer (ITS) regions of the ribosomal DNA have been established in the last years as standard, due to the high variability compared to the 18S rRNA gene (Martin and Rygiewicz 2005). For the analysis of micro-eukaryotes, multi-barcoding approaches with several primer sets targeting different groups have been successfully applied (Lentendu et al. 2014).

Conversely, as a large fraction of soil microbial genomes is still uncharacterized, designing universal primers for functional genes with current database coverage is challenging (Fredriksson et al. 2013; Tremblay et al. 2015), often giving partial views on the real diversity only (Wei et al. 2015). Primer degeneration is offering more options for binding sites and diversity recovery, while non-PCR-based alternatives also do exist, e.g., DNA-DNA hybridization or screen of metagenomes (Jacquiod et al. 2014). Designing primers to bind to a region of the gene which codes for the catalytic domain of an enzyme, which is usually conserved between microbes, can increase the detection of a greater diversity of microbes carrying the gene of interest. If the catalytic domain of the enzyme is however highly conserved, compared to other regions of the enzyme, the phylogenetic resolution is greatly decreased. Despite the current lack of coverage in databases, evaluating designed primers using in silico PCR can give important insights about the phylogenetic resolution and specificity.

Amplicon length is crucial, as longer sequences will considerably increase annotation accuracy and phylogenetic resolution (Singer et al. 2016). If libraries will be sequenced using Illumina® Miseq (2 × 300 bp) or HiSeq (2 × 250 bp) paired-end sequencing technology, amplicon sizes up to 550 bp are possible.

Finally, it is important to keep in mind that DNA-based sequencing approaches do not provide information about the expression levels of the amplified gene. Thus, any speculations related to the "functionality of a gene" must be avoided or mitigated. The analysis of mRNA or 16S rRNA via cDNA,

Table 1 Primer systems frequently used for the partial amplification of the 16S rRNA gene from bacteria and archaea in studies based on Illumina MiSeq or HiSeq sequencing

Forward primer		Reverse primer		Variable region	Reference
Name	Sequence 5′-3′	Name	Sequence 5′-3′		
515f Modified	GTGYCAGCMGCCGCGGTAA	926r	CCGYCAATTYMTTTRAGTTT	4–5	(Walters et al. 2016)
S-D-Bact-0008-a-S-16 (27F)	AGAGTTTGATCMTGGC	S-D-Bact-0343-a-A-15 (357R)	CTGCTGCCTYCCGTA	1–2	(Klindworth et al. 2013)
S-D-Bact-0341-b-S-17 (341F)	CCTACGGGNGGCWGCAG	S-D-Bact-0785-a-A-2 (B805R)	GACTACHVGGGTATCTAATCC	3–4	(Klindworth et al. 2013)

as well as meta-proteomics are efficient alternatives to address such questions.

Bioinformatic processing

The typical amplicon bioinformatic pipeline consists of several steps including: quality filtering, clustering, and comparison to a reference database. Key quality control steps include adapter trimming, quality and length filtering, chimeric sequence and contaminant removal (e.g., PhiX and host DNA), and are discussed in detail by Vestergaard et al. (2017). Bokulich et al. (2013) provide guidelines for Illumina® 16s rRNA amplicon sequencing analysis, showing that filtering for high-quality read length and using abundance cut-offs significantly improve diversity estimates.

Quality-filtered sequences are typically clustered into operational taxonomic units (OTUs). It is suggested to remove OTUs that only have one read in the entire data set (singletons) as they might reflect sequencing errors (Zhou et al. 2011). The clustering into OTUs can be done on any level of similarity. For a meaningful interpretation of amplicon sequencing data, an adequate coverage of the microbial community is required. Here, rarefaction curves, describing the increase of observed OTUs as a function of sequenced reads (Fig. 1) are highly informative. In general, decent coverage is achieved between 10,000 and 100,000 reads per soil sample, depending on the complexity of the microbiome, the targeted gene, and the desired resolution. Whereas for the 16S rRNA gene, typically similarity levels of 97–99% (Poretsky et al. 2014) are used to analyze bacteria or archaea on the level of "species," for micro-eukaryotes thresholds from 80 to 95% are used, when 18S rRNA genes or ITS regions are analyzed (Lentendu et al. 2014; Wang et al. 2014). Enzyme-coding genes sequences are often translated into amino acid sequences and compared to custom databases using tools like DIAMOND or blastP. Positive hits are then clustered into OTUs using similarity thresholds of 80% or even lower (Kielak et al. 2013). However, a clear taxonomic assignment is often not possible, due to the unknown frequency of horizontal gene transfer events of the genes in question. Here, the database choice for annotation may strongly influence the outcome due to limiting annotation possibilities (Jacquiod et al. 2014). Alternatively, translated sequences can be scanned for the presence of protein families using hidden Markov models (Bergkemper et al. 2016). An interesting approach for amplicons obtained from enzyme-coding genes is the possibility to estimate evolutionary rates of a particular gene compared to the 16S rRNA gene resulting in parameters for OTU clustering and taxonomic assignment (Kim and Liesack 2014).

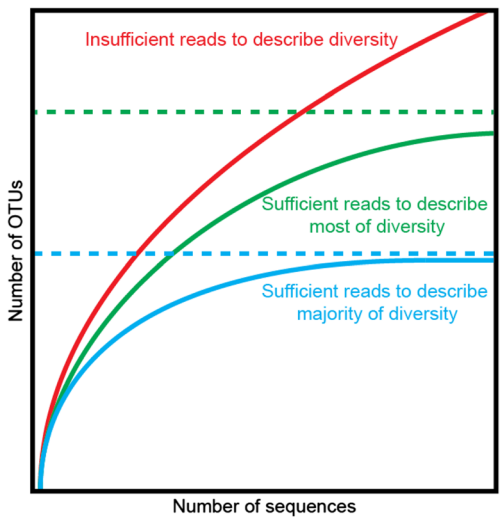

Fig. 1 Rarefaction curves obtained after sequencing of PCR amplicons and their consequences for data analysis

Another important issue is the semi-quantitative aspect of amplicon sequencing data. Indeed, for 16S rRNA-based data, it has been shown that due to fluctuation of ribosomal operon copy numbers between microbial species, an absolute quantification of a given OTU is not feasible (Jacquiod et al. 2016). The same issue may hold true for enzyme-coding genes, but here, baseline data is still missing. The PICRUST tool, originally designed to predict metagenomic profiles from amplicon data, implements a normalization procedure correcting count data based on sequenced genome information available in databases for achieving better estimations of absolute abundances (Langille et al. 2013). Nevertheless, analysis between different treatments provides relative semi-quantitative information useful and relevant for comparative studies when performed appropriately (Nunes et al. 2016).

Guidelines for biostatistic analysis and data interpretation

Data analysis is typically based on a contingency table and an associated distance matrix (UNIFRAC or other metrics). One major issue is dealing with a varying number of reads between samples. These differences may introduce a large bias when estimating diversity indices (e.g., richness, Shannon, Chao-1), especially when an asymptotic trend of the curves (Fig. 1) is not fulfilled. QIIME and other pipelines account for this issue by even random resampling (Weiss et al. 2015), while others argue that more advanced models should be used (McMurdie and Holmes 2014).

To extract significantly responding taxa across experimental designs, unappropriated Gaussian-based models like ANOVA are often used. These models are based on the normality assumption and independency of each variable, which is often not the case for microbial communities which are mostly non-normally distributed, have many correlated variables and a very strong mean to variance relationship (e.g., many zeros and rare observations). Thus, Gaussian-based models result in biased conclusions due to attribution of a

higher significance to taxa with a high variance, which persist even after data transformation (e.g., log10 or center-scaling). Instead, the correlations between variables can be assessed using permutations (e.g., Monte-Carlo simulations) while the mean to variance relationship can be predicted using generalized linear models and the non-normality of data modeled by logistic regressions (e.g., negative binomial distribution), followed by appropriated statistics like likelihood ratio tests.

All these approaches are available in the R packages mvabund (Wang et al. 2012), edgeR (Robinson et al. 2010), and phyloseq (McMurdie and Holmes 2013). Basic principles on these procedures are summarized in a YouTube tutorial by David Warton about data normalization in ecology using mvabund (https://www.youtube.com/watch?v=KnPkH6d89l4).

Network visualization is becoming more frequent to predict the complex interactions within microbial communities (Karimi et al. 2017). It is important to keep in mind that networks show data co-occurrences and co-exclusions, which is no evidence of biological interactions. Moreover, adequate appropriated correlation indices (e.g., Spearman, Pearson, SparCC) and permutation validation are required to avoid biases that might arise from the count data itself, like random correlations and compositional biases (Berry and Widder 2014; Friedman and Alm 2012). Besides coefficient choice and statistical validity, stringent high cut-offs can be applied on the correlation strengths (>|0.6|) to avoid integration of weak observations and false positives, at the cost of excluding many false negative interactions. Metadata addition (e.g., soil pH; carbon, and nitrogen content) to network visualizations can be an important factor to further assess biological linkage relevance. Here, patterns can be found, which reinforce the idea of specific niche observations, and also the detection of important topographic features such as keystone species, outliers/satellites, modules of highly co-occurring taxa, and even hubs of connected modules that might positively or negatively correlate (Jacquiod et al. 2016).

As the field of microbial ecology is growing, accumulation of knowledge and large amounts of data enable the testing of important ecological theories that were previously out of reach (Prosser et al. 2007). For instance, multivariate data generated from environmental microbial communities can be used to test macro-ecology concepts like Functional Response Groups (FRGs, group of organisms sharing similar response to environmental changes) and Functional Effect Groups (FEGs or guilds, groups of organisms contributing to a similar ecosystem function) (Lehsten et al. 2009). Data on 16S rRNA genes or transcripts can be used to identify FRG and allow the identification of bacterial taxa with a similar response pattern to an environmental clue (Nunes et al. 2016). In addition, functional genes and/or transcripts can be used to determine FEG contributing to a specific ecosystem function (e.g., chitin or cellulose degradation). Working with groups defined from multivariate data is very powerful for understanding the basis of microbial behavior such as trophic lifestyles (e.g., copiotrophs or oligotrophs), growing habits (e.g., fast growers or slow growers), or to reveal environmental bioindicators.

Outlook

Improvements and developments of novel sequencing technologies add more detailed data, allowing new questions to be asked. Third-generation sequencing platforms can provide even greater detail to amplicon-based approaches. For instance, sequencing the entire 16S rRNA gene using PacBio® technology resulted in a higher phylogenetic resolution compared to shorter amplicons which are mostly generated from Illumina sequencing (Singer et al. 2016). However, new sequencing approaches often need new protocols and new bioinformatic tools. Here the construction of synthetic or "mock" communities is a valuable tool to validate and optimize sequencing workflows and subsequent data analyses (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012).

Finally, the information that can be gained from sequence-based approaches heavily relies on the quality and completeness of reference databases. Therefore, sequencing approaches should be complemented with classical isolation and cultivation-based approaches as well as functional validation using enzyme assays to further improve the quality of databases for a better characterization of the dark microbial matter.

Acknowledgements Gisle Vestergaard is supported by a Humboldt Research Fellowship for postdoctoral researchers.

References

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Baraniya D, Puglisi E, Ceccherini MT, Pietramellara G, Giagnoni L, Arenella M, Nannipieri P, Renella G (2016) Protease encoding microbial communities and protease activity of the rhizosphere and bulk soils of two maize lines with different N uptake efficiency. Soil Biol and Biochem 96:176–179 DOI: 10.1016/j.soilbio.2016.02.001

Bergkemper F, Kublik S, Lang F, Kruger J, Vestergaard G, Schloter M, Schulz S (2016) Novel oligonucleotide primers reveal a high diversity of microbes which drive phosphorous turnover in soil. J Microbiol Methods 125:91–97. doi:10.1016/j.mimet.2016.04.011

Berry D, Widder S (2014) Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol 5 doi:ARTN 219 10.3389/fmicb.2014.00219

Bokulich NA et al (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57–59. doi:10.1038/nmeth.2276

Caporaso JG et al (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6: 1621–1624. doi:10.1038/ismej.2012.8

Fredriksson NJ, Hermansson M, Wilen BM (2013) The choice of PCR primers has great impact on assessments of bacterial community diversity and dynamics in a wastewater treatment plant. PLoS One 8:e76431. doi:10.1371/journal.pone.0076431

Friedman J, Alm EJ (2012) Inferring correlation networks from genomic survey data. PLoS Comput Biol 8:e1002687. doi:10.1371/journal.pcbi.1002687

Jacquiod S et al (2014) Characterization of new bacterial catabolic genes and mobile genetic elements by high throughput genetic screening of a soil metagenomic library. J Biotechnol 190:18–29. doi:10.1016/j.jbiotec.2014.03.036

Jacquiod S, Stenbaek J, Santos SS, Winding A, Sorensen SJ, Prieme A (2016) Metagenomes provide valuable comparative information on soil microeukaryotes. Res Microbiol 167:436–450. doi:10.1016/j.resmic.2016.03.003

Jumpstart Consortium Human Microbiome Project Data Generation Working G (2012) Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One 7:e39315. doi:10.1371/journal.pone.0039315

Karimi B, Maron PA, Chemidlin-Prevost Boure N, Bernard N, Gilbert D, Ranjard L (2017) Microbial diversity and ecological networks as indicators of environmental quality. Environ Chem Lett. doi:10.1007/s10311-017-0614-6

Kielak AM, Cretoiu MS, Semenov AV, Sorensen SJ, van Elsas JD (2013) Bacterial chitinolytic communities respond to chitin and pH alteration in soil. Appl Environ Microbiol 79:263–272. doi:10.1128/AEM.02546-12

Kim Y, Liesack W (2014) DAFGA: diversity analysis of functional gene amplicons. Bioinformatics 30:2820–2821. doi:10.1093/bioinformatics/btu394

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41:e1. doi:10.1093/nar/gks808

Langille MG et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821. doi:10.1038/nbt.2676

Lehsten V, Harmand P, Kleyer M (2009) Fourth-corner generation of plant functional response groups. Environ Ecol Stat 16:561–584. doi:10.1007/s10651-008-0098-4

Lentendu G, Wubet T, Chatzinotas A, Wilhelm C, Buscot F, Schlegel M (2014) Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach. Mol Ecol 23:3341–3355. doi:10.1111/mec.12819

Lindahl BD et al (2013) Fungal community analysis by high-throughput sequencing of amplified markers—a user's guide. New Phytol 199: 288–299. doi:10.1111/nph.12243

Lusk RW (2014) Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. PLoS One 9: e110808. doi:10.1371/journal.pone.0110808

Martin KJ, Rygiewicz PT (2005) Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. BMC Microbiol 5:28. doi:10.1186/1471-2180-5-28

McMurdie PJ, Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. doi:10.1371/journal.pone.0061217

McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 10: e1003531. doi:10.1371/journal.pcbi.1003531

Nunes I et al. (2016) Coping with copper: legacy effect of copper on potential activity of soil bacteria following a century of exposure. FEMS Microbiol Ecol 92 doi:10.1093/femsec/fiw175

Pathan SI et al (2015) Maize lines with different nitrogen use efficiency select bacterial communities with different beta-glucosidase-encoding genes and glucosidase activity in the rhizosphere. Biol Fert Soils 51:995–1004

Poretsky R, Rodriguez RL, Luo C, Tsementzi D, Konstantinidis KT (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 9:e93827. doi:10.1371/journal.pone.0093827

Prosser JI et al (2007) Essay—the role of ecological theory in microbial ecology. Nat Rev Microbiol 5:384–392. doi:10.1038/nrmicro1643

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. doi:10.1093/bioinformatics/btp616

Salter SJ et al (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87. doi:10.1186/s12915-014-0087-z

Schmidt H, Rothhämel S (2012) Polymerase-Kettenreaktion. Gentechnische Methoden:135–171

Singer E et al (2016) High-resolution phylogenetic microbial community profiling. ISME J 10:2020–2032. doi:10.1038/ismej.2015.249

Tremblay J et al (2015) Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771. doi:10.3389/fmicb.2015.00771

Vestergaard G, Schulz S, Schöler A, Schloter M (2017) Making big data smart—how to use metagenomics to understand soil quality. Biol Fert Soils. doi:10.1007/s00374-017-1191-3

Walters W et al. (2016) Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys mSystems 1 doi:10.1128/mSystems.00009-15

Wang Y, Naumann U, Wright ST, Warton DI (2012) mvabund—an R package for model-based analysis of multivariate abundance data. Methods Ecol Evol 3:471–474. doi:10.1111/j.2041-210X.2012.00190.x

Wang Y, Tian RM, Gao ZM, Bougouffa S, Qian PY (2014) Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. PLoS One 9:e90053. doi:10.1371/journal.pone.0090053

Wei W et al (2015) Higher diversity and abundance of denitrifying microorganisms in environments than considered previously. ISME J 9:1954–1965. doi:10.1038/ismej.2015.9

Weiss SJ et al (2015) Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. Peer J Preprint. doi:10.7287/peerj.preprints.1157v1

Zhou J et al (2011) Reproducibility and quantitation of amplicon sequencing-based detection. ISME J 5:1303–1313. doi:10.1038/ismej.2011.11