PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls

Fabian Bonk[†], Denny Popp[†,*], Hauke Harms, Florian Centler

Department of Environmental Microbiology, UFZ - Helmholtz Centre for Environmental Research, Permoserstraße 15, 04318 Leipzig, Germany

[†] Authors contributed equally

* Corresponding author: Denny Popp, Permoserstraße 15, 04318 Leipzig, Germany. Tel: +49 341 235 482247; e-mail address: denny.popp@ufz.de

Abstract

The quantification of relative and absolute taxa-specific abundances in complex microbial communities is crucial for understanding and modeling natural and engineered ecosystems. Many errors inherent to this quantification are, though well-known, still insufficiently addressed and can potentially lead to a completely different interpretation of experimental results. This review provides a critical assessment of next generation sequencing (NGS) of amplicons and quantitative real-time PCR for the quantification of relative and absolute taxa-specific genome abundances. Starting from DNA extraction, the following error sources were considered: DNA extraction efficiency, PCR-associated bias, variance of strain-specific 16S rRNA operon copy number per genome, and analysis of quantitative real-time PCR and NGS data. Tools and methods for estimating and minimizing these errors are presented and demonstrated on published data. In conclusion, amplicon sequencing and qPCR of 16S rRNA genes are valuable tools to determine relative and absolute taxa-specific genome abundances, but results can deviate by several orders of magnitudes from the true values if the reviewed error sources are ignored. Many of these errors can be minimized in a cost-efficient manner and large errors can be easily identified by plausibility checks as shown in this review. Finally, the accurate conversion of genome abundances to cell numbers and microbial biomasses was pointed out as an important future research topic for the integration of PCR-based abundances into mathematical models.

Keywords: absolute abundance quantification; biomass quantification; 16S rRNA gene amplicon sequencing; quantitative real-time PCR (qPCR); 16S rRNA gene copy number variation; ploidy

1. Introduction

Microbial communities are the hidden champions in many natural and engineered ecosystems, driving global elemental cycles, waste removal in waste water treatment plants, and methane production in biogas plants to name a few examples. To understand and model these systems, the accurate quantification of taxa-specific abundances is of crucial importance. Abundances can hereby refer to gene abundances, genome abundances, cell numbers and biomasses. In microbial ecology studies, abundances are often based on 16S rRNA gene copy numbers while mechanistic mathematical models rather consider microbial abundances as biomass measured as dry weight.

Various culture-independent molecular biology based techniques have been used to study microbial communities such as metagenomics (Eloe-Fadrosh et al., 2016; Shakya et al., 2013), metaproteomics (Kleiner et al., 2017), flow cytometry (Lambrecht et al., 2017), fluorescence in-situ hybridization (FISH) (Nettmann et al., 2010), and amplicon sequencing (Klassen et al., 2017). For amplicon sequencing, 16S rRNA genes are commonly used to identify community composition, but also other genes have been targeted, for example the *mcrA* gene to focus on methanogenic archaea (Steinberg and Regan, 2008). 16S rRNA gene amplicon sequencing informs on relative taxa-specific gene abundances, which can be converted to relative taxa-specific genome abundances using strain-specific 16S rRNA operon copy number information. Relative gene and genome abundances are commonly used to analyze the relationship between microbial community composition and environmental parameters.

However, relative abundances are of limited use if total abundances are unknown (Props et al., 2017) (see Additional file 1, chapter A.1). Absolute taxa-specific genome abundances can be determined by quantitative real-time PCR (qPCR) for individual taxa (Yu et al., 2005) and for all taxa of a community at once by combining 16S rRNA gene amplicon sequencing with qPCR (Dannemiller et al., 2014). Both amplicon sequencing and qPCR can lead to substantial

errors of which many have been widely discussed in the literature. However, their impact is still often neglected and not systematically corrected for when reporting on relative and absolute genome abundances.

To guide practitioners, we critically evaluate potential errors of 16S rRNA gene amplicon sequencing and qPCR in quantitative terms in this review. Moreover, tools and guidelines to estimate and minimize these errors are provided. Errors discussed in this review are associated with DNA extraction, PCR, and analysis of NGS and qPCR data. Furthermore, methods are presented to estimate absolute abundances to check the plausibility of experimental results. Finally, we addressed the errors associated with the conversion of absolute taxa-specific genome abundances to taxa-specific cell numbers and biomasses to be used in mathematical modelling. To illustrate the errors above, we mainly refer to examples from anaerobic digestion which is driven by a complex microbial community. However, the problems and solutions discussed in this study directly extend to any complex prokaryotic community.

2. Errors associated with both amplicon sequencing and qPCR

2.1. Avoidance, quantification, and correction of DNA extraction efficiency associated errors

Extracting DNA of a diverse community from a complex matrix is a challenging task. Lysis conditions have to be harsh enough to break up all types of cells but should not cause damage to the DNA. The strain-specific DNA extraction efficiency, i.e. the recovered amount of genomic DNA divided by the total amount of genomic DNA present in the sample for a certain strain, depends on the sample matrix, species' morphology, and the extraction method.

Inter-strain differences in DNA extraction efficiency, for example due to different cell walls and membranes, lead to biased relative and absolute genome abundances. This error can be minimized by spiking a microbial mock community of known and representative composition to a sample and testing various DNA extraction methods (Willner et al., 2012). The mock community should contain a variety of morphologies representative of the sample and *E. coli* (if used to determine overall extraction efficiency, see below). The extraction method leading to the least biased relative taxa-specific abundances of the mock community members should be chosen.

After minimizing taxa-specific extraction biases, the overall extraction efficiency of prokaryotic DNA needs to be determined. If DNA loss during extraction is neglected, absolute genome abundances in the sample will be underestimated. Proposed standards for efficiency estimation commonly rely on spiking a known number of *E. coli* cells to the sample. A target gene in the genomic DNA or on a plasmid of *E. coli* is then quantified. The overall extraction efficiency is calculated by the number of detected target genes divided by the number of spiked target genes. If *E. coli* is already a member of the community to be analyzed, its originally present target genes in the sample needs to be subtracted before calculating the overall extraction efficiency. Using *E. coli* spikes to cattle manure, overall extraction efficiencies were between 38% and 99.97% for different extraction methods (Lebuhn et al., 2016). The extraction efficiency of 38% would result in an error on the absolute genome abundances of 263%.

Previous to DNA extraction, an additional step for removing extracellular DNA from dead cells can be performed using for example propidium monoazide (Emerson et al., 2017). While leading to more accurate results, the application of propidium monoazide to turbid samples remains a challenge in practice (Kirkegaard et al., 2017).

In conclusion, inter-species variations in DNA extraction efficiency can lead to substantially inaccurate relative genome abundances. This error cannot be corrected a posteriori. A mock community can be used to choose the best DNA extraction method for minimizing this error. Thereafter, the overall DNA extraction efficiency needs to be estimated by spiking a known amount of a standard, typically *E. coli*, to the sample prior to DNA extraction.

2.2. Avoiding PCR-associated biases

After extraction, the genomic DNA is used as a template for PCR amplification, either for amplicon sequencing of 16S rRNA genes or in qPCR. The PCR amplification step can be a major source of bias. These errors are associated with the PCR template properties (template concentration, GC content), primer choice (primer coverage and mismatch), polymerase choice, and the PCR protocol (annealing temperature and PCR cycle number).

PCR template properties

Diluting the DNA template concentration can positively affect the PCR efficiency if inhibitors are present in the sample. However, diluting the DNA template concentration for PCR for amplicon sequencing can exclude rare taxa from the detection and therefore results in a lower observed richness (Wu et al., 2010). The potential error of using too low DNA template concentrations can neither be estimated nor corrected for rare taxa and should therefore be avoided.

A high GC content of template DNA of a certain taxon can lead to an under-representation of this taxon (Pinto and Raskin, 2012) because high GC contents cause a less efficient initial denaturation during PCR leading to a lower amplification efficiency (Laursen et al., 2017). The resulting error of the genomic GC content on the relative gene abundances is hard to estimate as it is intertwined with other error sources and can be up to one order of magnitude.

Increasing the time for initial denaturation during PCR can reduce this error but cannot suppress it completely (Pinto and Raskin, 2012).

Primer coverage

Targeting complex microbial communities requires primer pairs covering all bacterial and archaeal 16S rRNA gene sequences present in the sample. As indicated by several studies, there are no universal primer pairs covering both domains equally well (Baker et al., 2003; Bru et al., 2008; Takahashi et al., 2014). One suggested solution is the use of separate primer pairs for each domain (Fischer et al., 2016). Furthermore, the use of several primer pairs for bacteria is recommended because primer pairs covering all bacterial phyla are not available (Klindworth et al., 2013) and might even not be feasible. An analysis of 500 bacterial 16S rRNA gene sequences revealed that apart from a very limited region (position 788 to 798) the absolute base conservation is restricted to four or less consecutive bases (Baker et al., 2003). Hence, no primer having a suitable length can be designed to cover all bacterial and even less so all prokaryotic 16S rRNA genes sequences.

For designing highly universal primer pairs, *in silico* coverages are calculated using tools like Silva TestPrime, see for example Klindworth et al. (2013). However, comparisons of predicted and measured coverages show high disagreements for some taxa (Claesson et al., 2010; Fischer et al., 2016; Thijs et al., 2017). Hence, *in silico* coverages cannot reliably predict *in situ* coverage and therefore cannot be applied to correct insufficiently covered taxa *a posteriori.*

The coverage of primers is determined by sequence similarity of primer and target sequence. A single primer target mismatch can substantially decrease the PCR amplification efficiency (up to 1000-fold), leading to a significant underrepresentation of the specific taxon (Bru et al., 2008).

Many sequencing techniques have limited read lengths. That is why usually only a limited number of variable regions of 16S rRNA genes are targeted, for example the V3/V4 regions (Albertsen et al., 2015) or the V1-V3 regions (Cai et al., 2013). In a 16S rRNA gene database study, targeting partial regions provided lower phylogenetic resolution than full-length sequences (Kim et al., 2011). In contrast, another study reported similar community compositions when targeting the V1 to V4 regions compared to full-length sequences (Kraková et al., 2016). As a consequence of these conflicting results, there are no indisputable recommendations on which variable regions to address to characterize complex microbial communities. A recent method targeting the SSU rRNA molecules instead of their rRNA genes resulted in a million of full-length rRNA gene sequences free of the primer bias (Karst et al., 2018). However, this method is laborious and complex and hence, not yet applicable as standard method.

In conclusion, it is advisable (i) to use primer pairs which are specific for either bacteria or archaea and target each domain separately and (ii) to use two primer pairs for each domain targeting different variable regions.

Choice of polymerase

DNA polymerases used for elongation are characterized by their fidelity and proofreading activity. Fidelity expressed as substitution error rate is in the range of 1 falsely incorporated base per 2 kb or lower (Potapov and Ong, 2017). This means that this error can be neglected considering the typical amplicon length of 300 to 500 bp. In theory, this error could be corrected *a posteriori* as the substitution error profiles are known (Shagin et al., 2017). However, such bioinformatics tools are not available yet. The proofreading activity of polymerase can lead to primer editing (Gohl et al., 2016). If this happens early during PCR, taxa which show a primer target mismatch will still be amplified efficiently, avoiding the

potential 1000-fold bias due to the mismatch as described above. Hence, polymerases with a high fidelity and proofreading activity are desirable.

PCR protocol

Studies on the influence of PCR annealing temperature on relative abundances in microbial communities show conflicting results. No significant effect of lowering the annealing temperature was found for complex communities like chicken caecal samples (Sergeant et al., 2012) while an increased number of OTUs was observed in activated sludge (Albertsen et al., 2015). Its effect apparently depends on the sequence similarity between primer and template DNA. In case of a perfect match, lowering the annealing temperature had no effect on the PCR product ratios for a mock combination of two bacteria, while an exponential deviation from the initial 1:1 ratio of two bacterial DNA templates was reported when there was one mismatch (Sipos et al., 2007). In conclusion, high annealing temperatures should be avoided as mismatches can lead to biased community composition results.

The effect of varying PCR cycle numbers on PCR products is also ambiguous. For activated sludge, no effect was found (Albertsen et al., 2015). This contrasts another study reporting increasing apparent richness when more PCR cycles were performed (Ahn et al., 2012). However, this might be due to the formation of PCR artifacts like chimeric sequences which is supposed to happen when PCR components become limiting. Chimeric sequences are generated when the elongation is not completed within one PCR cycle and the DNA fragment serves as primer for the next cycle binding to template DNA of another taxon present in the sample. Chimeric sequences can be a substantial part of the PCR products as libraries with up to 45% chimeras were found, which can be misinterpreted as additional taxa (Ashelford et al., 2006). The chimera formation cannot be quantified in complex communities and therefore, the associated error cannot be corrected. In conclusion, it is beneficial to reduce chimera

formation to a minimum by restricting the number of PCR cycles and to filter at best remaining chimeras by appropriate sequence analysis as discussed below.

Another type of PCR artifacts are heteroduplex molecules which arise from hybridization of heterologous sequences and which are less frequent than chimeras (Qiu et al., 2001). They will not influence the relative abundance obtained by common NGS methods like Illumina sequencing and pyrosequencing as DNA is denatured before sequencing. This might be different for other methods using dsDNA for sequencing like SMRT PacBio sequencing or Oxford Nanopore sequencing. Formation of heteroduplexes can be reduced by reducing the number of PCR cycles and avoiding high DNA template concentrations as mentioned above (Thompson, 2002).

2.3. Quantification and correction of errors associated with variations of strain-specific 16S rRNA operon copy numbers per genome

An innate bias of using 16S rRNA genes for relative and absolute taxa-specific genome abundance quantification are variations of strain-specific 16S rRNA operon copy numbers per genome (Kembel et al., 2012), with reported ranges of 1-17 for bacteria and 1-4 for archaea (Stoddard et al., 2015). An unaccounted higher 16S rRNA operon copy number per genome of a specific taxon will appear as a higher relative and absolute genome abundance of this taxon. Hence, a correction for strain-specific 16S rRNA operon copy numbers per genome is necessary, for example by using the rrnDB database (Stoddard et al., 2015).

The correction for strain-specific 16S rRNA operon copy numbers per genome is incomplete as data is not available for all strains and as operational taxonomic units (OTUs) cannot be taxonomically assigned on strain level. If the strain-specific number is missing, either median or mean copy number for the respective higher taxonomic rank can be used. However, this leads to inaccuracies.

Here, we used three example data sets A (Klassen et al., 2017), B (Maus et al., 2017) and C (Müller et al., 2016) to illustrate the effect of correcting for strain-specific 16S rRNA gene operon copy numbers per genome on relative abundances (see Additional File 1, Chapter A.2, for details). In our examples, the copy number correction changed the relative genome abundances compared to the relative 16S rRNA gene abundances by 22% on average though not uniformly for all taxa (Figure 2a). In a standard analysis of data set A without 16S rRNA gene operon copy number correction, the phylum *Bacteroidetes* appeared to be most dominant followed by *Firmicutes* and *Chlorobi*. In contrast, after 16S rRNA operon copy number correction *Chlorobi* was the most dominant phylum, followed by *Bacteroidetes* and *Firmicutes*. This was due to the low average 16S rRNA operon copy number per genome for *Chlorobi* (2.0) compared to the other dominant phyla *Bacteroidetes* (3.7) and *Firmicutes* (7.0). Hence, negligence of the strain-specific 16S rRNA operon copy numbers per genome results in biased relative abundances and wrong dominance rankings.

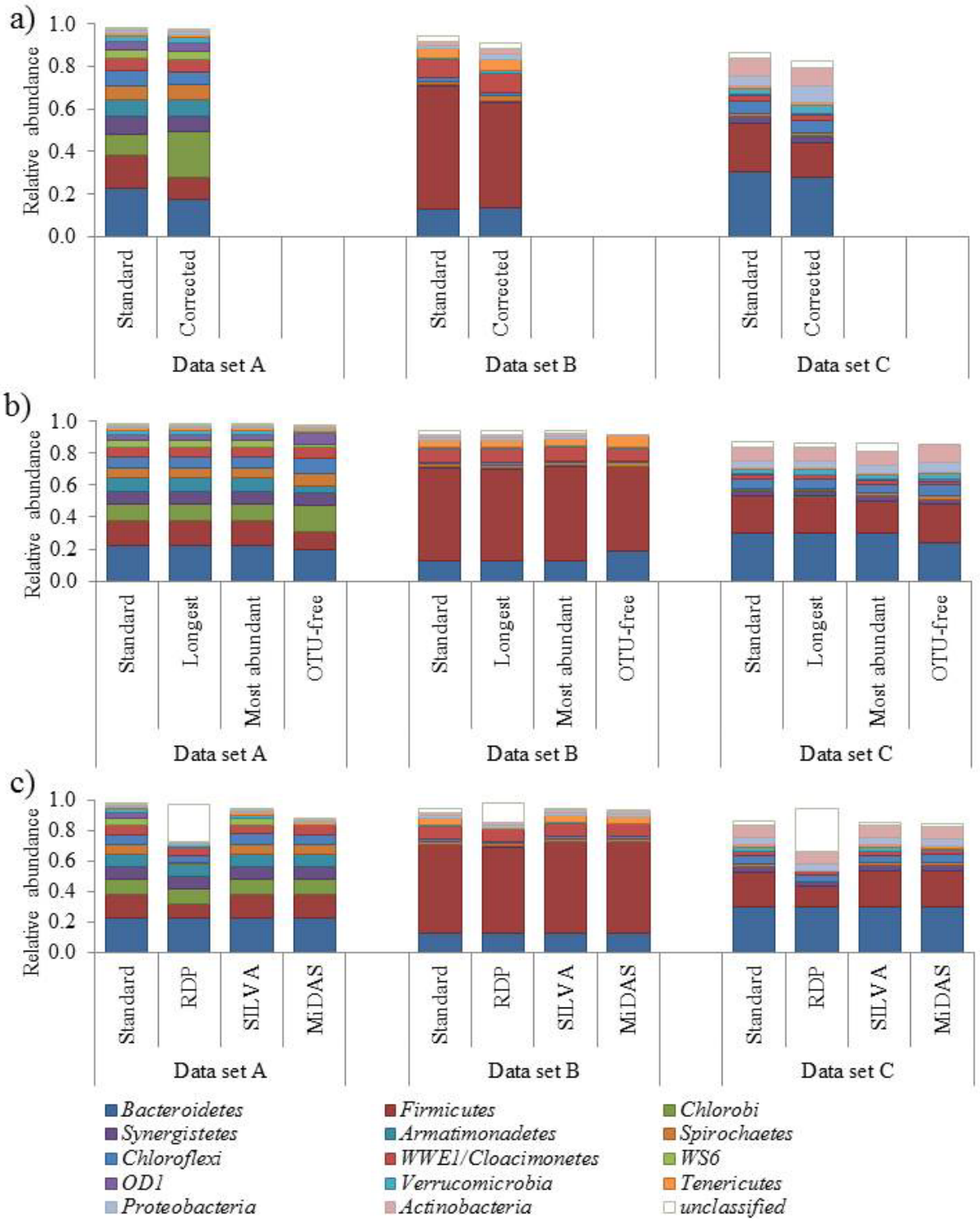

Figure 2: Relative abundance of the 15 most abundant bacterial phyla for three example data sets A (Klassen et al., 2017), B (Maus et al., 2017) and C (Müller et al., 2016). a) Comparison of relative 16S rRNA gene abundance (the standard approach) with relative genome abundance after correction for strain-specific 16S rRNA operon copy number per genome

(corrected). b) Comparison of the relative 16S rRNA gene abundances after use of different strategies for selecting the representative sequence for each OTU. The standard approach uses the seed sequence which was used during the clustering process. Alternatively, the longest or the most abundant sequence per OTU was selected as the representative. Furthermore, an OTU-free approach was applied as implemented in the DADA2 pipeline. c) Comparison of relative 16S rRNA gene abundances obtained with different databases used for taxonomic assignment.

3. Errors solely associated with amplicon sequencing

3.1. Influence of sequencing technology on relative gene abundances

NGS platforms like 454 pyrosequencing, Illumina, IonTorrent, and PacBio employ different principles for sequencing DNA (Goodwin et al., 2016). This in turn results in substantial differences in length and quality of reads as well as sequencing depth, which, in turn, influences the inferred community composition. As the sequence length is limited due to the sequencing technology, only subsets of the variable regions of the 16S rRNA genes are targeted. This is strongly linked to the choice of primers as discussed above. The read quality heavily depends on the sequencing platform. For example, the frequently used MiSeq Illumina sequencing platform is associated with a sequencing error rate of less than 1% (Schirmer et al., 2016). The sequencing depths of NGS platforms depend on their read numbers which usually range from thousands to millions of sequences. More reads generate higher apparent richness as more rare taxa are detected (Claesson et al., 2010). Nonetheless, there is a trade-off between the efficiency of rare taxa detection and artefactual taxa removal (erroneous sequences) during bioinformatic analyses, see (Zhan and MacIsaac, 2015) for further discussion. Apart from the rare taxa, the same taxa were detected by different

sequencing platforms when the same primer pairs were applied (D'Amore et al., 2016; Tremblay et al., 2015).

In conclusion, a sufficient sequencing depth is necessary to target rare taxa and sequencing error is considered only as a minor source of error. However due a limited read length, the sequencing platform restricts the primer choice which has a larger impact than the NGS platform itself (Hiergeist et al., 2016).

3.2. Avoidance of errors associated with NGS data analysis

Next to biases from wet-lab procedures, errors can be introduced during data processing. For the analysis of amplicon sequencing data, several pipelines are available. Different pipelines lead to considerably different relative 16S rRNA gene abundancies (Golob et al., 2017; Plummer and Twin, 2015; Werner et al., 2012). Here, we focused on general biases which can be introduced during analysis of amplicon sequencing data and which are not restricted to a specific analysis pipeline.

Filtering low quality and chimeric sequences

Quality control of raw sequences, i.e. filtering of low quality and chimeric sequences is common to all pipelines. As chimeric sequences can make up a substantial portion of all reads, their removal is of importance. Sequences can be filtered for chimeras using a reference database (e.g., the Gold reference collection) containing 16S rRNA gene sequences of cultivated bacteria and archaea. However, not all chimeric sequences are detected by this approach (Ahn et al., 2012). Furthermore, this reference chimera detection will perform poorly on microbial communities if they contain yet uncultivated organisms (Schloss et al., 2011). To increase chimera detection efficiency, *de novo* detection methods can be applied, optionally in combination with reference-based detection (Schloss et al., 2011). However, as

still not all of the chimeras can be detected (Haas et al., 2011), it is important to reduce chimera formation by adjusting the PCR conditions appropriately as described above.

In conclusion, chimera formation needs to be reduced by optimizing PCR conditions in the first place. A substantial part, but not all of the remaining chimeric sequences can be removed during data analysis. For microbial communities with a lot of undescribed microorganisms, *de novo* detection methods optionally combined with reference-based detection should be applied.

OTU clustering

Filtered sequences are clustered into OTUs according to sequence similarity. For clustering, a 16S rRNA gene sequence similarity threshold of 97% is commonly used to differentiate between species, though it represents an arbitrary threshold rather than being based on a commonly accepted species definition (Callahan et al., 2017). Hence, it may assign different species to the same OTU or one species to several OTUs due to sequence variation in multiple 16S rRNA operons.

After clustering, one representative sequence per OTU is selected and taxonomically assigned. The effect of the choice of representative sequences for each OTU is illustrated in Figure 2b. As standard in QIIME analysis, the centroid sequence which has been used for defining the OTU is taken as the representative. Alternatively, the longest and the most abundant sequence can be chosen. For the example data sets A (Klassen et al., 2017), B (Maus et al., 2017) and C (Müller et al., 2016), the selection strategy has only a minor impact on the obtained community composition. Relative 16S rRNA gene abundances change by 0.3% on average except for data set C if the most abundant sequences per OTU are selected. Here, the average change of 5% is caused by a higher fraction of unclassified OTUs.

In order to avoid arbitrary threshold setting for OTU clustering resulting in misassignments, recently developed analysis pipelines like DADA2 avoid OTU clustering altogether and instead account for sequence variation down to one nucleotide sequence differences (Callahan et al., 2016). Community compositions of the example data sets derived from the OTU clustering approach and exact sequence inference differed by 45% on average comparing the relative 16S rRNA genes abundances, see Figure 2b. A similar difference was reported for gut microbiome samples (Allali et al., 2017).

In conclusion, based on our example data sets, OTU clustering-free approaches should be used to avoid setting arbitrary sequence similarity thresholds. If OTU clustering is desired, selecting the most abundant sequence should be avoided. Instead, the centroid sequence of each OTU should be taken as the representative sequence.

Influence of 16S rRNA databases on taxonomic assignment

After selecting representative or inferring exact sequences, these are classified against a taxonomic database. The choice of database has a strong influence on the observed community composition for our example data sets A (Klassen et al., 2017), B (Maus et al., 2017) and C (Müller et al., 2016), see Figure 2c. Relative 16S rRNA gene abundances changed on average by 9% when using SILVA or the SILVA-based MiDAS taxonomies instead of the Greengenes taxonomy. Previous reports also showed that different databases for taxonomic assignment give different community compositions (Werner et al., 2012). The latest Greengenes version is from 2013 [25] and is problematic for the example data set, because it lacks 16S rRNA sequences of microorganisms which were described more recently, for example syntrophic acetate oxidizing bacteria. SILVA, RDP, and MiDAS are more up to date. Using the RDP database results in a substantially higher number of unclassified OTUs for the example data sets, and therefore does not seem to be recommendable. The MiDAS database is built on the SILVA database and was amended for

taxa present in activated sludge, anaerobic digesters, and influent wastewater (McIlroy et al., 2015, 2017), making it the database of choice for the example data set and anaerobic digester samples in general. Similar to MiDAS, other dedicated databases exist, for example for human intestinal (Ritari et al., 2015), human oral (Chen et al., 2010), and bee intestinal (Newton and Roeselers, 2012) microbial communities.

4. Errors solely associated with qPCR data analysis

For qPCR, two reporter systems are commonly used, hybridization probes (also called *Taq*Man^{TM} probes) and intercalating dyes such as SYBR Green (Smith and Osborn, 2009). Intercalating dyes bind non-specifically to all amplicons. Therefore, post-PCR melting curve analyses need to confirm that only target genes were quantified and not non-specific PCR products such as primer-dimers (Smith and Osborn, 2009). Hybridization probes are designed to bind to a conserved site on the target gene, ensuring that only the target gene is quantified (Smith and Osborn, 2009). Such a conserved site, however, might not exist for a target gene present in various members of a mixed community with potentially unknown members. Here, a probe might bind unequally to the 16S rRNA genes of all members and lead to biased results.

For absolute gene quantification with qPCR, an external standard is necessary. From the fluorescence of the sample compared to that of the standard, the 16S rRNA gene copy number in the sample is determined. A qPCR standard contains a known number of 16S rRNA genes, either from a single member or a mixture of members of the microbial community of interest. The 16S rRNA genes can be in the form of a purified PCR product or a plasmid insert. Plasmid standards can be linearized which gave more accurate results when quantifying microalgae (Hou et al., 2010). In that study, the non-linearized standard led to an overestimation of the target gene copy number by 777%. However, a study targeting two bacterial and two archaeal species did not find a systematic overestimation by using circular

plasmids, and similar absolute gene numbers were obtained using linearized, supercoiled or nicked plasmids or a purified PCR product (Oldham and Duncan, 2012).

Ideally, a new standard representing the community composition is produced from each sample individually. However this is resource intensive and therefore, it is common to use a standard from a single species which neglects the possibility that the qPCR efficiency of the standard could be atypical for the mixture of target genes of the community. However, for example a 10% lower amplification efficiency of the sample can give rise to 275% overestimated absolute gene copy numbers (Pérez et al., 2013). Lower amplification efficiencies can also be caused by inhibitors in the sample matrix. The amplification efficiencies of all individual samples and standards can be determined with for example the LinRegPCRProgram as described in the literature (Brankatschk et al., 2012). If a difference in amplification efficiency is detected, it is necessary to correct it for example by one-point-calibration (Brankatschk et al., 2012) or by repeating the analysis with diluted DNA template concentrations. A convenient spreadsheet to apply the one-point-calibration method is provided in Additional file 2. This spreadsheet can also be used to estimate the error of the differences in efficiencies and other errors discussed above. The problem of amplification efficiency differences between standard and sample in qPCR can be avoided by using digital PCR (dPCR) because it does not require a standard (Kim et al., 2015).

The absolute genome abundance of a single taxon can be obtained by qPCR with a specific primer pair. However, in a mixed culture other taxa than the targeted one can be additionally amplified leading to inaccurate quantification results. The use of hybridization probes can reduce this problem. In addition, amplicon sequencing using the same primer pair can be used to identify and correct the influence of unspecific amplification on the absolute genome abundance of a single taxon.

Publication of qPCR data should follow the “minimum information for publication of quantitative real-time PCR experiments” (MIQE) (Bustin et al., 2009). In particular, the publication of the qPCR raw data is desirable to enable post-publication error estimates and corrections for amplification efficiency differences.

In conclusion, there are several error sources in qPCR that can sum up to several orders of magnitude. Differences in amplification efficiency between standard and sample are often neglected but can be corrected by one-point calibration. Either linearized plasmids, circular (supercoiled or nicked) plasmids or PCR products can be used as a standard for prokaryotes. A decision tree helping to avoid the above mentioned errors is provided in Figure 3.

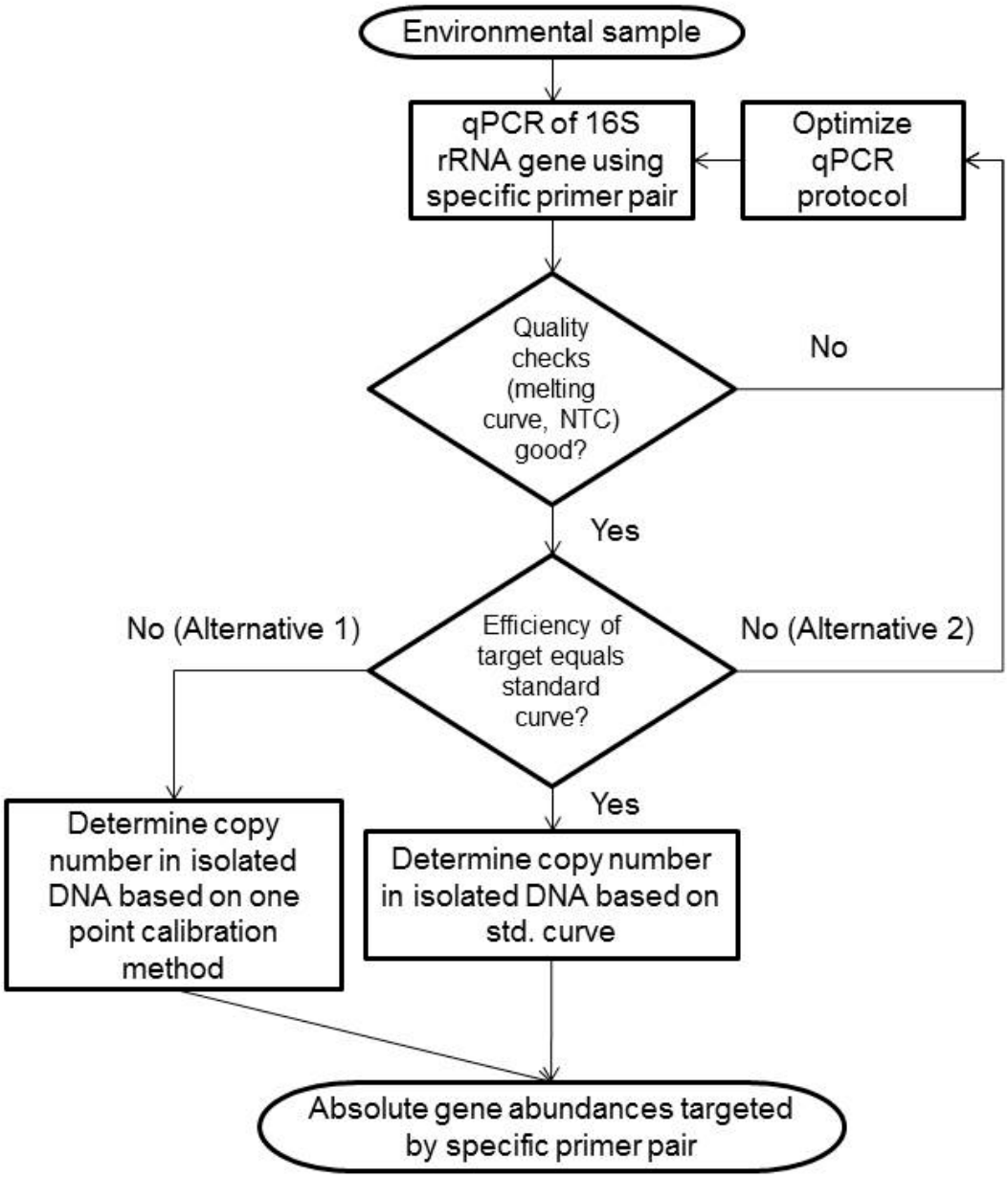

Figure 3: Guideline for minimizing absolute 16S rRNA gene abundance quantification errors in qPCR

5. Error identification with plausibility checks based on environmental parameters

The quantification of absolute taxa-specific genome abundances can yield errors of up to several orders of magnitude. Plausibility checks are thus highly desirable to validate results and to avoid conclusions based on erroneous data. Several methods are available to estimate

absolute biomass concentrations based on process parameters and environmental conditions. Mechanistic mathematical models, such as the Anaerobic Digestion Model No.1 (Batstone et al., 2002) can be used to predict microbial biomasses. However, such models require extensive information as input. For systems with scarce information, black box approaches are more suitable, in particular methods including thermodynamic considerations (Heijnen, 2013; Kleerebezem and Van Loosdrecht, 2010).

An example for a simple plausibility check is presented in Figure 4. Absolute archaeal 16S rRNA gene copy numbers measured in anaerobic digesters (Lee et al., 2011; Nettmann et al., 2010) were compared with minimum gene copy numbers estimated by a black box approach. These estimates were based on the assumption that the catabolism of any microbial cell is limited by a maximum rate for electron transport (Heijnen, 2002). Given this assumption, the minimum cell number required to produce the amount of methane measured in the digesters was calculated (see Additional File 1, section A.4, for details). For the conversion of cell numbers to gene copies, each cell was conservatively assumed to contain one archaeal 16S rRNA gene copy.

The archaeal 16S rRNA gene copy numbers measured by Nettmann et al. (2010) lie well above the estimated minimum copy numbers. However, the copy numbers measured by Lee et al. (2010) are orders of magnitude below the estimated minimum. This indicates that their results likely underestimate absolute quantities in the sampled digesters. DNA extraction efficiencies were not considered in that study, which might be a reason for the implausibly low copy numbers. This example illustrates how even simple plausibility checks can be used to identify implausible quantification results. Such plausibility checks are not restricted to methanogenic environments. Kleerebezem and Van Loosdrecht (2010) for example provided biomass yield estimates for 61 organic compounds with either oxygen, nitrate, sulfate and carbon dioxide as electron acceptors.

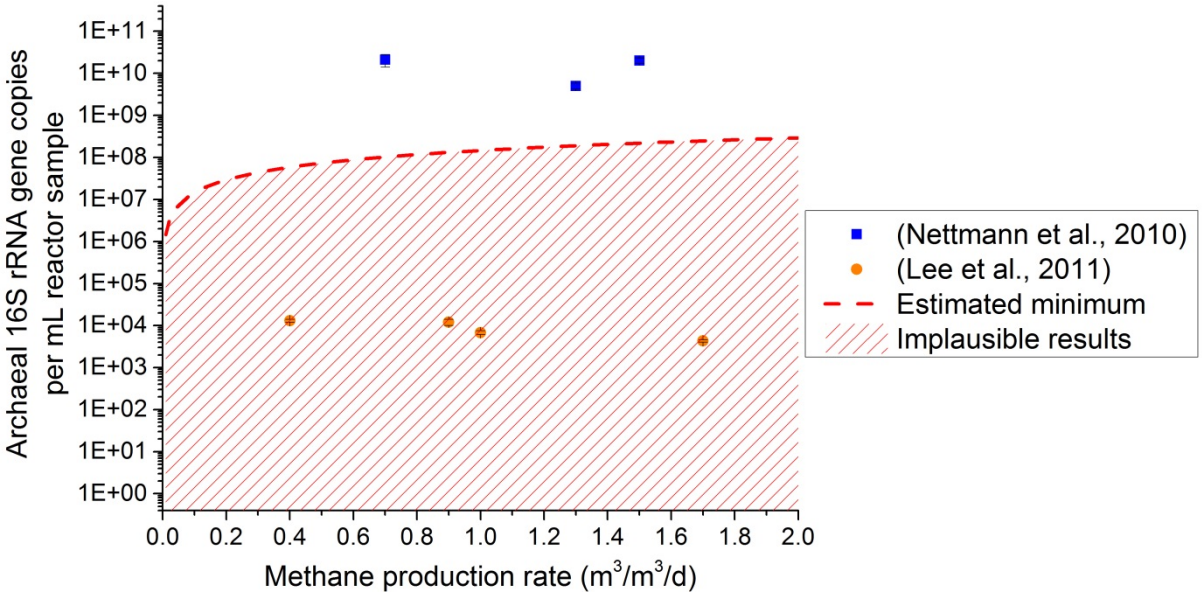

Figure 4: Example for checking the plausibility of absolute 16S rRNA gene quantification. Comparison of experimentally derived archaeal 16S rRNA gene copies in anaerobic digesters (Lee et al., 2011; Nettmann et al., 2010) with estimated minimum gene copy number based on the maximum electron transfer rate (Heijnen, 2002) and the methane production rates measured in the digesters (normalized to the digesters working volume).

6. Errors associated with converting genome abundances to cell numbers and biomasses

Ideally, taxa-specific genome abundances could be converted to more tangible cell numbers. However, this requires taxa-specific information on the ploidy, i.e. the number of genome copies per cell. Prokaryotes have historically been considered as monoploid (1 genome copy per cell) (Pecoraro et al., 2011). However, several recent studies have found oligoploid (<10 genome copies per cell) and polyploid (>10 genome copies per cell) archaea and bacteria and it appears that monoploid prokaryotes are rather the exception than the rule (Soppa, 2014).

A correction for ploidy is difficult. The ploidy can differ even within one genus, for example within *Desulfovibrio* and within *Neisseria* (Pecoraro et al., 2011). Furthermore, the ploidy of a species does not only vary between one and two during the cell cycle but can differ greatly between different growth phases, for example from 3-15 genome copies per cell in the exponential phase to 2-4 genome copies in the stationary phase for *Methanocaldococcus jannaschii* (Pecoraro et al., 2011). Taxa-specific ploidy has been determined experimentally for pure cultures by combining qPCR with cell counting (Pecoraro et al., 2011). For complex communities, qPCR could be combined with fluorescence-activated cell sorting in the future.

Deriving taxa-specific cell numbers from complex communities was recently suggested by combining absolute cell numbers derived from flow cytometry with taxa-specific 16S rRNA genome abundances (Props et al., 2017). However, this method requires the same ploidy of all taxa which is unlikely given the high variance of ploidy found in prokaryotes (Pecoraro et al., 2011).

As mentioned above, mechanistic models often consider microorganisms not as cells but as biomass, but it is difficult to determine experimentally the taxa-specific biomasses of species-rich complex microbial communities. Nevertheless, the distinction between cell number and biomass must not be ignored, because the prokaryotic cell masses can vary over several orders of magnitude (Loferer-Krößbacher et al., 1998). Taxa-specific biomasses can be inferred from cell volumes determined by FISH combined with digital image analysis (Daims, 2009). In addition to FISH, metaproteomics has recently been suggested as a method for taxa-specific biomass quantification (Kleiner et al., 2017).

In conclusion, taxa-specific average genome copies and biomasses per cell can vary substantially between taxa and neglecting this fact can lead to errors of up to 10,000%. Consequently, the accurate conversion between genome abundances and cell numbers as well as biomasses remains a challenge.

7. Conclusions

Relative and absolute taxa-specific genome abundances are important parameters for studying microbial community dynamics, but their quantification with PCR based approaches has a number of potential errors that can reach several orders of magnitudes. These errors and suggested measures to avoid or reduce them are summarized in Table 1. Many errors can already be reduced by proper data analysis. Others require additional experimental effort, such as spiking of known microorganisms to estimate DNA extraction efficiencies. Using different primer pairs for bacteria and archaea is essential for accurate analyses but adds substantial experimental effort. The accurate conversion of taxa-specific genome abundances to cell numbers and biomasses is important for their use in mathematical models but remains a challenge. Flow cytometry, FISH and metaproteomics might bring valuable culture-independent contributions in the future to solve this problem.

494

Additional files

- **Additional file 1: Further information on the studied errors and concepts, .**pdf, Details on the example NGS data sets and the data analysis. Details on the minimum gene number estimate. Illustration of the difference in using relative and absolute cell numbers for interpreting the influence of distinct taxa on the process performance. Additional information on correcting strain-specific 16S rRNA operon copy numbers.
- **Additional file 2: Tool for qPCR data analysis und error estimation, .**xlsx, Excel sheet to simplify the analysis of qPCR data for absolute quantification and error estimation.

Funding

This work was funded by the German Federal Ministry of Education and Research (e:Bio project “McBiogas”, FKZ 031A317).

Conflict of interest

The authors declare they have no conflict of interest.

References

Ahn, J.-H., Kim, B.-Y., Song, J., and Weon, H.-Y. (2012). Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J. Microbiol.* 50, 1071–1074. doi:10.1007/s12275-012-2642-z.

Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H., and Nielsen, P. H. (2015). Back to basics – the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS One* 10, e0132783. doi:10.1371/journal.pone.0132783.

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., et al. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol.* 17, 194. doi:10.1186/s12866-017-1101-8.

Ashelford, K. E., Chuzhanova, N. a., Fry, J. C., Jones, A. J., and Weightman, A. J. (2006). New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl. Environ. Microbiol.* 72, 5734–5741. doi:10.1128/AEM.00556-06.

Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541–555. doi:10.1016/j.mimet.2003.08.009.

Batstone, D. J., Keller, J., Angelidaki, I., Kalyuzhnyi, S. V., Pavlostathis, S. G., Rozzi, A., et al. (2002). *Anaerobic Digestion Model No.1. (ADM1)*. London: IWA Publishing.

Brankatschk, R., Bodenhausen, N., Zeyer, J., and Bürgmann, H. (2012). Simple absolute quantification method correcting for quantitative PCR efficiency variations for microbial community samples. *Appl. Environ. Microbiol.* 78, 4481–4489. doi:10.1128/AEM.07878-11.

Bru, D., Martin-Laurent, F., and Philippot, L. (2008). Quantification of the detrimental effect of a single primer-template mismatch by real-Time PCR using the 16S rRNA gene as an example. *Appl. Environ. Microbiol.* 74, 1660–1663. doi:10.1128/AEM.02403-07.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* 55, 611–622. doi:10.1373/clinchem.2008.112797.

Cai, L., Ye, L., Tong, A. H. Y., Lok, S., and Zhang, T. (2013). Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. *PLoS One* 8, e53649. doi:10.1371/journal.pone.0053649.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi:10.1038/ismej.2017.119.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010). The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* 2010, baq013-baq013. doi:10.1093/database/baq013.

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., et al. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res.* 38, e200–e200. doi:10.1093/nar/gkq873.

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., et al. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* 17, 55. doi:10.1186/s12864-015-2194-9.

Daims, H. (2009). Use of fluorescence in situ hybridization and the daime image analysis program for the cultivation-independent quantification of microorganisms in environmental and medical samples. *Cold Spring Harb. Protoc.* 4, 1–7. doi:10.1101/pdb.prot5253.

Dannemiller, K. C., Lang-Yona, N., Yamamoto, N., Rudich, Y., and Peccia, J. (2014). Combining real-time PCR and next-generation DNA sequencing to provide quantitative comparisons of fungal aerosol populations. *Atmos. Environ.* 84, 113–121. doi:10.1016/j.atmosenv.2013.11.036.

Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., and Kyrpides, N. C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.* 1, 15032. doi:10.1038/nmicrobiol.2015.32.

Emerson, J. B., Adams, R. I., Román, C. M. B., Brooks, B., Coil, D. a., Dahlhausen, K., et al. (2017). Schrödinger's microbes: tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome* 5, 86. doi:10.1186/s40168-017-0285-3.

Fischer, M. A., Güllert, S., Neulinger, S. C., Streit, W. R., and Schmitz, R. A. (2016). Evaluation of 16S rRNA gene primer pairs for monitoring microbial community structures showed high reproducibility within and low comparability between datasets generated with multiple archaeal and bacterial primer pairs. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01297.

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* 34, 942–949. doi:10.1038/nbt.3601.

Golob, J. L., Margolis, E., Hoffman, N. G., and Fredricks, D. N. (2017). Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics* 18, 283. doi:10.1186/s12859-017-1690-0.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi:10.1101/gr.112730.110.

Heijnen, J. J. (2002). "Bioenergetics of Microbial GrowthMicrobial Growth," in *Encyclopedia of Bioprocess Technology*, eds. M. C. Flickinger and S. W. Drew (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 267–291. doi:10.1002/0471250589.ebt026.

Heijnen, J. J. (2013). "A thermodynamic approach to predict black box model parameters for microbial growth," in *Biothermodynamics* (EPFL Press), 443–472.

Hiergeist, A., Reischl, U., and Gessner, A. (2016). Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center

variability. *Int. J. Med. Microbiol.* 306, 334–342. doi:10.1016/j.ijmm.2016.03.005.

Hou, Y., Zhang, H., Miranda, L., and Lin, S. (2010). Serious overestimation in quantitative PCR by circular (supercoiled) plasmid standard: microalgal pcna as the model gene. *PLoS One* 5, e9545. doi:10.1371/journal.pone.0009545.

Kembel, S. W., Wu, M., Eisen, J. A., and Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* 8, e1002743. doi:10.1371/journal.pcbi.1002743.

Kim, M., Morrison, M., and Yu, Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J. Microbiol. Methods* 84, 81–87. doi:10.1016/j.mimet.2010.10.020.

Kim, T. G., Jeong, S.-Y., and Cho, K.-S. (2015). Development of droplet digital PCR assays for methanogenic taxa and examination of methanogen communities in full-scale anaerobic digesters. *Appl. Microbiol. Biotechnol.* 99, 445–458. doi:10.1007/s00253-014-6007-x.

Kirkegaard, R. H., McIlroy, S. J., Kristensen, J. M., Nierychlo, M., Karst, S. M., Dueholm, M. S., et al. (2017). The impact of immigration on microbial community composition in full-scale anaerobic digesters. *Sci. Rep.* 7, 9343. doi:10.1038/s41598-017-09303-0.

Klassen, V., Blifernez-Klassen, O., Wibberg, D., Winkler, A., Kalinowski, J., Posten, C., et al. (2017). Highly efficient methane generation from untreated microalgae biomass. *Biotechnol. Biofuels* 10, 186. doi:10.1186/s13068-017-0871-4.

Kleerebezem, R., and Van Loosdrecht, M. C. M. (2010). A generalized method for thermodynamic state analysis of environmental systems. *Crit. Rev. Environ. Sci. Technol.* 40, 1–54. doi:10.1080/10643380802000974.

Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., et al. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* 8, 1558. doi:10.1038/s41467-017-01544-x.

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

Kraková, L., Šoltys, K., Budiš, J., Grivalský, T., Ďuriš, F., Pangallo, D., et al. (2016). Investigation of bacterial and archaeal communities: novel protocols using modern sequencing by Illumina MiSeq and traditional DGGE-cloning. *Extremophiles* 20, 795–808. doi:10.1007/s00792-016-0855-5.

Lambrecht, J., Cichocki, N., Hübschmann, T., Koch, C., Harms, H., and Müller, S. (2017). Flow cytometric quantification, sorting and sequencing of methanogenic archaea based on F420 autofluorescence. *Microb. Cell Fact.* 16, 1–15. doi:10.1186/s12934-017-0793-7.

Laursen, M. F., Dalgaard, M. D., and Bahl, M. I. (2017). Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. *Front. Microbiol.* 8, 1–8. doi:10.3389/fmicb.2017.01934.

Lebuhn, M., Derenkó, J., Rademacher, A., Helbig, S., Munk, B., Pechtl, A., et al. (2016). DNA and RNA extraction and quantitative real-time PCR-based assays for biogas biocenoses in an interlaboratory comparison. *Bioengineering* 3, 7. doi:10.3390/bioengineering3010007.

Lee, I. S., Parameswaran, P., and Rittmann, B. E. (2011). Effects of solids retention time on methanogenesis in anaerobic digestion of thickened mixed sludge. *Bioresour. Technol.*

102, 10266–10272. doi:10.1016/j.biortech.2011.08.079.

Loferer-Krößbacher, M., Klima, J., and Psenner, R. (1998). Determination of bacterial cell dry mass by transmission electron microscopy and densitometric image analysis. *Appl. Environ. Microbiol.* 64, 688–694.

.

Maus, I., Kim, Y. S., Wibberg, D., Stolze, Y., Off, S., Antonczyk, S., et al. (2017). Biphasic study to characterize agricultural biogas plants by high-throughput 16S rRNA gene amplicon sequencing and microscopic analysis. *J. Microbiol. Biotechnol.* 27, 321–334. doi:10.4014/jmb.1605.05083.

McIlroy, S. J., Kirkegaard, R. H., McIlroy, B., Nierychlo, M., Kristensen, J. M., Karst, S. M., et al. (2017). MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. *Database* 2017. doi:10.1093/database/bax016.

McIlroy, S. J., Saunders, A. M., Albertsen, M., Nierychlo, M., McIlroy, B., Hansen, A. A., et al. (2015). MiDAS: the field guide to the microbes of activated sludge. *Database* 2015, bav062. doi:10.1093/database/bav062.

Müller, B., Sun, L., Westerholm, M., and Schnürer, A. (2016). Bacterial community composition and fhs profiles of low- and high-ammonia biogas digesters reveal novel syntrophic acetate-oxidising bacteria. *Biotechnol. Biofuels* 9, 48. doi:10.1186/s13068-016-0454-9.

Nettmann, E., Bergmann, I., Pramschufer, S., Mundt, K., Plogsties, V., Herrmann, C., et al. (2010). Polyphasic analyses of methanogenic archaeal communities in agricultural biogas plants. *Appl. Environ. Microbiol.* 76, 2540–2548. doi:10.1128/AEM.01423-09.

Newton, I. L., and Roeselers, G. (2012). The effect of training set on the classification of honey bee gut microbiota using the Naïve Bayesian Classifier. *BMC Microbiol.* 12, 221. doi:10.1186/1471-2180-12-221.

Oldham, A. L., and Duncan, K. E. (2012). Similar gene estimates from circular and linear standards in quantitative PCR analyses using the prokaryotic 16S rRNA gene as a model. *PLoS One* 7, e51931. doi:10.1371/journal.pone.0051931.

Pecoraro, V., Zerulla, K., Lange, C., and Soppa, J. (2011). Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PLoS One* 6. doi:10.1371/journal.pone.0016392.

Pérez, L. M., Fittipaldi, M., Adrados, B., Morató, J., and Codony, F. (2013). Error estimation in environmental DNA targets quantification due to PCR efficiencies differences between real samples and standards. *Folia Microbiol. (Praha).* 58, 657–662. doi:10.1007/s12223-013-0255-5.

Pinto, A. J., and Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* 7, e43093. doi:10.1371/journal.pone.0043093.

Plummer, E., and Twin, J. (2015). A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *J. Proteomics Bioinform.* 8. doi:10.4172/jpb.1000381.

Potapov, V., and Ong, J. L. (2017). Examining sources of error in PCR by single-molecule sequencing. *PLoS One* 12, 1–19. doi:10.1371/journal.pone.0181128.

Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman,

W., et al. (2017). Absolute quantification of microbial taxon abundances. *ISME J.* 11, 584–587. doi:10.1038/ismej.2016.117.

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, a. V., Tiedje, J. M., et al. (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol.* 67, 880–887. doi:10.1128/AEM.67.2.880-887.2001.

Ritari, J., Salojärvi, J., Lahti, L., and de Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* 16, 1056. doi:10.1186/s12864-015-2265-y.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17, 1–15. doi:10.1186/s12859-016-0976-y.

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6, e27310. doi:10.1371/journal.pone.0027310.

Sergeant, M. J., Constantinidou, C., Cogan, T., Penn, C. W. C. W., and Pallen, M. J. M. J. (2012). High-throughput sequencing of 16S rRNA gene amplicons: effects of extraction procedure, primer length and annealing temperature. *PLoS One* 7, e38094. doi:10.1371/journal.pone.0038094.

Shagin, D. a., Shagina, I. a., Zaretsky, A. R., Barsova, E. V., Kelmanson, I. V., Lukyanov, S., et al. (2017). A high-throughput assay for quantitative measurement of PCR errors. *Sci. Rep.* 7, 2718. doi:10.1038/s41598-017-02727-8.

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* 15, 1882–1899. doi:10.1111/1462-2920.12086.

Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* 60, 341–350. doi:10.1111/j.1574-6941.2007.00283.x.

Smith, C. J., and Osborn, a. M. (2009). Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol. Ecol.* 67, 6–20. doi:10.1111/j.1574-6941.2008.00629.x.

Soppa, J. (2014). Polyploidy in archaea and bacteria: About desiccation resistance, giant cell size, long-term survival, enforcement by a eukaryotic host and additional aspects. *J. Mol. Microbiol. Biotechnol.* 24, 409–419. doi:10.1159/000368855.

Steinberg, L. M., and Regan, J. M. (2008). Phylogenetic comparison of the methanogenic communities from an acidic, oligotrophic fen and an anaerobic digester treating municipal wastewater sludge. *Appl. Environ. Microbiol.* 74, 6663–6671. doi:10.1128/AEM.00553-08.

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K., and Schmidt, T. M. (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* 43, D593–D598. doi:10.1093/nar/gku1201.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., and Nishijima, M. (2014). Development of a prokaryotic universal primer for simultaneous analysis of bacteria and archaea using

738 next-generation sequencing. *PLoS One* 9, e105592. doi:10.1371/journal.pone.0105592.

739 Thijs, S., Op De Beeck, M., Beckers, B., Truyens, S., Stevens, V., Van Hamme, J. D., et al.
740 (2017). Comparative evaluation of four bacteria-specific primer pairs for 16S rRNA gene
741 surveys. *Front. Microbiol.* 8. doi:10.3389/fmicb.2017.00494.

742 Thompson, J. R. (2002). Heteroduplexes in mixed-template amplifications: formation,
743 consequence and elimination by "reconditioning PCR." *Nucleic Acids Res.* 30, 2083–
744 2088. doi:10.1093/nar/30.9.2083.

745 Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., et al. (2015). Primer and
746 platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* 6.
747 doi:10.3389/fmicb.2015.00771.

748 Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., et
749 al. (2012). Impact of training sets on classification of high-throughput bacterial 16s
750 rRNA gene surveys. *ISME J.* 6, 94–103. doi:10.1038/ismej.2011.82.

751 Willner, D., Daly, J., Whiley, D., Grimwood, K., Wainwright, C. E., and Hugenholtz, P.
752 (2012). Comparison of DNA extraction methods for microbial community profiling with
753 an application to pediatric bronchoalveolar lavage samples. *PLoS One* 7, e34605.
754 doi:10.1371/journal.pone.0034605.

755 Wu, J.-Y., Jiang, X.-T., Jiang, Y.-X., Lu, S.-Y., Zou, F., and Zhou, H.-W. (2010). Effects of
756 polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity
757 analysis using the deep sequencing method. *BMC Microbiol.* 10, 255. doi:10.1186/1471-
758 2180-10-255.

759 Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005). Group-specific primer and probe sets to
760 detect methanogenic communities using quantitative real-time polymerase chain
761 reaction. *Biotechnol. Bioeng.* 89, 670–679. doi:10.1002/bit.20347.

762 Zhan, A., and MacIsaac, H. J. (2015). Rare biosphere exploration using high-throughput
763 sequencing: research progress and perspectives. *Conserv. Genet.* 16, 513–522.
764 doi:10.1007/s10592-014-0678-9.

765

766 Table 1: Estimation, avoidance and correction of errors in PCR-based taxa-specific absolute genome abundance quantification

Error type and source	Potential error size	Estimation of error	Avoidance of error a priori	Correction of error a posteriori	Additional effort for error avoidance/correction
NGS and qPCR associated errors					
Strain-specific DNA extraction bias	N/A	Spike mock community	Optimize extraction protocol	Impractical	High
Extraction efficiency	~1000 %	Spike standard (*E. coli* cells)	Optimize extraction protocol	Quantifying spiked standard	Medium
DNA Template concentration PCR for amplicon sequencing	Low (only rare taxa)	Not possible (?)	Avoid too low concentrations	Impractical	Low
Primer coverage	~100,000 %	Impractical	(1) Separate primer pairs for bacteria and archaea; (2) Additional primer pair for each domain covering different variable region	Impractical	High for (1) Very high for (1) + (2)
PCR	~1000%	Impractical	Reduce cycles to reduce PCR artifacts, use high fidelity polymerase	Identification of chimeras and their removal	low
NGS platform	low	Impractical	Impractical	Impractical	N/A
NGS data analysis					
OTU clustering	~10 %	N/A	Use OTU free clustering approach		Low (only computational)
Taxonomic database choice	~10%	N/A	Use dedicated database		Low (only computational)
Strain-specific 16S rRNA operon copy number per genome	~100%	N/A	N/A	Use amplicon sequencing data and rrnDB database	Low (only computational)
qPCR data analysis					
PCR efficiency standard vs sample	~100-1000%	Use amplification curves	Optimize PCR protocol + standard. Use dPCR instead.	Use one-point calibration method	Low (only computational)

767