Pure

Scotland's Rural College

The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies

Pollock, J; Glendinning, L; Wisedchanwet, T; Watson, M

Published in:
Applied and Environmental Microbiology

DOI:
10.1128/AEM.02627-17

First published: 02/02/2018

Document Version
Peer reviewed version

Link to publication

Citation for pulished version (APA):
Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Applied and Environmental Microbiology*, *84*(7), [e02627-17]. https://doi.org/10.1128/AEM.02627-17

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 27. Aug. 2022

AEM Accepted Manuscript Posted Online 2 February 2018
Appl. Environ. Microbiol. doi:10.1128/AEM.02627-17
Copyright © 2018 Pollock et al.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies

Jolinda Pollock[a,b*#], Laura Glendinning[b*], Trong Wisedchanwet[b] & Mick Watson[b]

Animal and Veterinary Sciences, Scotland's Rural College (SRUC), Edinburgh, United Kingdom[a]; The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom[b]

Running Head: Consensus "best practice" for 16S studies

#Corresponding Author: Jolinda Pollock – jolinda.pollock@sruc.ac.uk

*JP and LG contributed equally to manuscript preparation

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

Accepted Manuscript Posted Online

Abstract

The development and continuous improvement of high-throughput sequencing platforms has stimulated interest in the study of complex microbial communities. Currently, the most popular sequencing approach to study microbial community composition and dynamics is targeted 16S rRNA gene metabarcoding. To prepare samples for sequencing, there are a variety of processing steps, each with the potential to introduce bias at the data analysis stage. In this short review, key information from the literature pertaining to each processing step is described and consequently, general recommendations for future 16S rRNA gene metabarcoding experiments are made.

Introduction

In recent years, the emergence of high-throughput sequencing platforms has revolutionised the study of complex microbial communities. Most commonly, marker genes (e.g. 16S rRNA and 18S rRNA genes) are amplified and sequenced, providing both qualitative and quantitative (i.e. relative abundance) data. However, the variety of methodologies which can be used to carry out marker gene analysis can be overwhelming. Each methodological stage, from sampling to data analysis, can introduce biases, and such biases can skew datasets by introducing changes in the relative abundances observed and can affect the perception of community diversity. This short review includes key information from current literature on sample collection, sample storage and processing, and sequencing and data analysis; specifically for the study of bacterial communities using 16S rRNA gene metabarcoding. By collating fundamental research from each of these areas, we aim to try to ensure that scientists entering this field are better informed to make decisions on experimental design for 16S rRNA gene sequencing studies.

Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology
Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Sample collection

Sampling method is obviously dependant on sample type and as such, the factors which may introduce bias will also vary between different types of microbiome studies. Clearly, study-specific concerns cannot be entirely covered in this review. However, the overarching factors which should be taken into account will be briefly covered in this section.

Firstly, it is important to consider the proposed sampling site. Bacterial community composition varies even within a specific environment, for example at different sites within the gastrointestinal tract (1), the respiratory tract (2) and at different soil depths (3, 4). Since the magnitude of inter-individual variation is very much dependant on sampling site (5), this can have implications for experimental design, specifically when considering the number of subjects and the number of samples to be taken.

Secondly, there are conflicting results in the literature with regards to the variation introduced by different sample collection methodologies. For example, there have been attempts to replace invasive sampling with less invasive methods; however, significant differences have been found in microbial populations when comparing swab and biopsy samples from human intestines (6), when comparing breath condensate and lung brushings (7) and when comparing rumen fluid samples obtained via oral stomach tubing and a fistula (8). However, other work contradicts these findings, with two studies showing no statistically significant differences when studying the rumen microbiota in cattle using a variety of sampling methods (9, 10). Additionally, no significant differences were evident in microbial composition when comparing sino-nasal swabs and biopsy samples (11) and rectal swabs and stool samples (12). This kind of conflict in the literature is not uncommon, which leads to a lack of consensus and standardisation.

A final consideration is whether samples should be homogenised, which appears to be most critical in studies on gut contents (8, 13) and on soil (14), since varying microbial

Accepted Manuscript Posted Online

compositions have been observed in different stool fractions and in soils with varying particle size.

Although the literature is generally conflicting with regards to sampling methodology, it is important to consider that comparing data obtained using different approaches should be avoided.

Sample storage

There is conflicting evidence on whether different storage conditions alone can have an impact on microbial community studies (15–18). It is often not practical to extract DNA from fresh samples, therefore samples are generally stored for varying durations prior to DNA extraction. Conventionally, it is assumed that rapid freezing to -80^{o}C is best practice (18, 19) but this is not feasible for all study designs, for example, at remote sites where low temperature storage is unavailable (20). Several studies have been carried out to assess the effects of storage conditions on study findings, which will be summarised in this section.

Fresh versus frozen samples

A couple of studies showed that freezing samples appeared to cause an increase in the Firmicutes to Bacteroidetes ratio in comparison with fresh samples (15, 19). Conversely, in a study by Fouhy *et al*, the only bacterial groups differentially expressed between fresh and snap frozen faecal samples were the *Faecalibacterium* and *Leuconostoc* genera, with no significant differences being evident at phylum or family levels (18). No significant effects on microbial composition or diversity were observed in faecal samples refrigerated for 24 hours (21) or 72 hours (20) prior to DNA extraction.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

The impact of storage duration has also been explored in various studies. Lauber *et al* stored soil, faeces and skin samples at various temperatures and found that storage duration had no significant impact on overall bacterial community structure or diversity (17). In samples which were stored at -80^{o}C for 2 years, a small number of changes in the microbial communities were observed with increased abundances of lactobacilli and bacilli, and a reduction in the total number of operational taxonomic units (OTUs) (for a definition of OTUs, please see section entitled "operational taxonomic unit picking methods").

When considering the data presented in the literature, generally processing fresh samples is the best approach but when this is not possible, samples should be frozen for an unequal amount of time and processed in one batch or frozen for an equal amount of time and processed in multiple batches. The decision on how to proceed will be dependent on the duration of the sample collection phase and on study design, but regardless of processing method, storage duration and DNA extraction batch should be recorded to enable this to be taken into account during analysis.

Use of cryoprotectant

McKain *et al* explored the effects of using a cryoprotectant (i.e. glycerol/phosphate buffered saline) to store ruminal digesta samples and found that freezing samples without cryoprotectant caused a significant loss in Bacteroidetes when measuring 16S rRNA gene copy number by quantitative PCR (15). The authors consequently suggest that simply storing samples without a cryoprotectant and carrying out DNA extraction at a later date would impact downstream results when considering archaeal and bacterial community composition. Choo *et al* explored the effects of using several common preservative buffers (i.e. RNAlater, OMNIgene.GUT and Tris-EDTA) relative to samples stored dry at -80^{o}C on faecal

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

microbiota composition (20). Samples stored in the OMNIgene.GUT buffer diverged the least from the samples stored dry at -80^{o}C and the results obtained from the samples stored in Tris-EDTA diverged the most, with associated changes in relative abundances of biologically important bacterial species such as *Escherichia-Shigella*, *Citrobacter* and *Enterobacter*. Additionally, RNAlater has previously been shown to be unsuitable for storage of samples subject to microbial community analysis, with samples stored in RNAlater being the least similar to fresh samples and samples immediately frozen at -80^{o}C (22, 23).

Consequently, when considering the use of a cryoprotectant for storage, it is important to ensure that all samples are stored in the same manner.

DNA extraction

During DNA extraction, it is important to consider that some microbial cells may be more resistant to lysis, such as bacterial endospores (24) and Gram-positive bacteria, which will have an impact on DNA extraction efficiency. The presence of inhibitors has also been found to directly impact DNA extraction efficiency (e.g. debris in environmental samples, organic matter in soil and faeces) and can affect the efficiency of PCR downstream (reviewed in detail by Schrader *et al* (25)). Common inhibitors include inorganic material (e.g. calcium ions), with the majority of inhibitors being organic matter such as humic acid, bile salts and polysaccharides. These issues will vary according to sample type, therefore, matrix-specific DNA extraction protocols should be optimised as part of a 16S rRNA gene metabarcoding experiment.

Besides phenol-chloroform DNA extraction methods, there are many commercial extraction kits available which incorporate mechanical and/or chemical/enzymatic lysis steps. Numerous authors have demonstrated that the abundances of specific bacterial groups vary

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

when comparing different DNA extraction methodologies (8, 26–31). Specifically, variations in DNA yield and quality are obtained which can lead to different results in downstream analyses (28).

One key DNA extraction step which can introduce bias is the presence or absence of a mechanical lysis step. The inclusion of a bead-beating step has been linked to a higher DNA yield (8, 29, 32), higher bacterial diversity (29, 32) and more efficient extraction of DNA from Gram-positive and spore-forming bacteria (29, 33, 34). Consequently, some authors suggest that samples subject to different DNA extraction methods are not comparable (8, 28, 35).

Ultimately, the best approach is to utilise a method which extracts the highest yield and quality of DNA as possible without biasing the method towards particular bacterial taxa. To achieve this, inclusion of a bead beating step and prior optimisation of the DNA extraction method to ensure optimal DNA yield and quality is recommended prior to carrying out 16S rRNA gene sequencing.

Sequencing strategy

Library preparation

Since the entire 16S rRNA gene cannot be sequenced using short-read second-generation sequencing platforms, a short region of the gene must be selected for PCR amplification and sequencing. There is currently no consensus on the most appropriate hypervariable region(s) and several studies have been carried out to determine the advantages and disadvantages of each. Importantly, the choice of hypervariable region(s) and the design of the "universal" PCR primers have an effect on phylogenetic resolution (36–40). Indeed, no primer set is truly universal, with some commonly used 16S rRNA gene primers proving ineffective at

Accepted Manuscript Posted Online

amplifying biologically relevant bacteria (34, 41). Fouhy *et al* explored the effects of primer choice (as well as DNA extraction and sequencing platform) on microbial composition data using a mock bacterial community and three primer sets (42), with differences in relative abundances and richness being observed.

Further biases can be introduced during PCR amplification due to the presence of PCR inhibitors (described in the DNA extraction section), with the number of PCR cycles and the use of a high-fidelity polymerase (43) also having an impact on results. The formation of chimeras occurs in later PCR cycles when the highest concentration of incompletely extended primers compete with the original primers. Consequently, the potential for chimera formation can be reduced by lowering the number of PCR cycles (44). Previous work found that bacterial richness increased as the PCR cycle number increased (45, 46), but that cycle number had no significant effect on community structure (46). A lower number of PCR artefacts were found when using a high-fidelity polymerase compared to a standard polymerase (43). The use of different polymerases has also been found to significantly affect PCR efficiencies for particular bacterial groups and the overall bacterial community structures (46). Finally, the quantity of input DNA into a PCR reaction has also been found to have a significant effect on observed bacterial community structure (31).

In summary, there is not a "gold standard" hypervariable region for 16S sequencing but it is important to consider that PCR reagents and PCR conditions should be optimised and kept consistent across a study.

Sequencing platforms

D'Amore *et al* have studied the choice of sequencing platform most recently (47) and we would refer the reader to that manuscript for a more in depth analysis. Illumina technology

Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology
Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

(primarily the MiSeq) has become the most common sequencing platform for 16S rRNA gene metabarcoding. This is because the MiSeq, in general, produces the most accurate, longest reads and has a much higher throughput than the other platforms, which enables more samples to be sequenced at higher depth or cheaper cost. Indeed, whilst D'Amore *et al* caution that the choice of sequencer depends on the question being asked, they note that the MiSeq is likely to be the platform of choice in most cases. The Roche 454 sequencer was, for a long time, the platform most used for 16S studies. The potential longer reads of this technology have some advantages; however, it is now no longer available as Roche retired the product in 2013. The 454 unfortunately suffered from an elevated error rate due to mis-calling of homopolymers. The Ion Torrent and Ion Proton platforms are often available at low capital cost, and produce data more quickly than the MiSeq. However, the lower throughput and higher error rates mean that many researchers prefer to select the MiSeq. Whilst Illumina offers the highest quality data, there are some reported problems with the platform. Illumina error rates are often thought to be around 0.01%, however Kozich *et al* showed the actual error rates can be as high as 10%, and recommend a complete overlap of 250 bp reads to correct for this (48). D'Amore *et al* similarly showed library-dependent error rates in either read 1 or read 2 (but not the overlap) in MiSeq data, albeit at a lower rate (2-3%) (47). An improvement has been suggested to this involving a heterogeneity spacer that improves sequence diversity in the library (49).

PacBio and Oxford Nanopore technologies are able to sequence the full length of the 16S gene, which is of course very powerful. However, again error rates are an issue, in the range of 5-15% for both technologies, which can cause subsequent errors in downstream analysis. Despite the high error rate of long-read single molecule sequencing systems (50–52), studies are beginning to appear to show their utility for 16S rRNA gene sequencing (53–56). For example, Schloss *et al* were able to reduce the observed error rate for the V1–V9 region from

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

0.69 to 0.027% for PacBio data, which is comparable to Illumina, 454 and Ion Torrent systems (54). One of the drawbacks of the PacBio technology is throughput, which means that the number of samples that can be run on the platform simultaneously and at reasonable cost is much lower than the MiSeq.

When planning a 16S sequencing study, three key considerations are the quality of sequence data, the cost of sequencing and the length of generated reads, as detailed already in this section. A final factor is the number of samples which can be analysed per sequencing run. When considering Illumina platforms specifically, it is possible to use multiplexing strategies by implementation of unique single-indexed (57) or dual-indexed (48) (or barcoded) primers for library preparation. If the number of samples per run is increased, this is associated with a lower coverage (or number of sequences generated) per sample. If the coverage per sample is too low, then the diversity of the microbial community being studied is likely to be under-represented, as rarer members of the community are less likely to be detected. Therefore, guidance on the number of samples to be included per run should be obtained from small pilot studies (and observation of the resultant rarefaction curves) or published literature. In larger studies, more than one sequencing run may be required and Caporaso *et al* showed that data were highly reproducible across sequencing lanes (57).

The appropriate sequencing platform should be selected based upon the aims of the experiment and the error rates associated with the available platforms. Another key consideration is sequencing coverage and its relation to the number of samples to be run. When studying core members of a microbial community, lowering the amount of coverage by increasing the number of samples in a sequencing run may be an effective way to decrease costs. However, if rarer members of a community are of interest lower sample numbers leading to increase coverage may be more appropriate.

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Mock bacterial communities

As part of 16S microbiome studies, it is useful to include a mock community control composed of pre-determined ratios of DNA from a mixture of bacterial species. This not only allows the quantification of sequencing error (58) but also allows bias introduced during the sampling and library preparation processes to be identified (42, 47, 59, 60). For example, a mock community containing bacterial taxonomies which are of specific interest to the research group can be used to calculate whether these taxonomies are likely to be over or under represented in samples. Similar to mock communities, spike-in standards can also be used to analyse bias and the reproducibility of methodologies (61). However, unlike mock communities, these standards are added directly to samples and therefore quality control can be performed on a per sample basis. However, there is a risk of crossover between the 16S rRNA gene sequences contained in the standards and those which may be found in samples. Consequently, care must be taken to select bacteria which are highly unlikely to occur in the samples of interest (62, 63) or which have been designed *in silico* and are dissimilar to sequences found in 16S databases (61).

There are a variety of sources which provide mock bacterial communities for use in research; however some researchers choose to create their own mock communities in-house which more accurately reflect bacteria of interest and scientific importance. Pre-prepared bacterial communities are available in two different formats – DNA mock communities and whole cell mock communities. The latter is useful for establishing the efficiency of the DNA extraction step, whereas the former will only assess the efficiency of PCR, clean up, sequencing and analysis steps. At the time of writing, mock communities are available from the American Type Culture Collection (ATCC) and Zymo Research.

When planning a 16S study, the inclusion of a mock community is strongly encouraged.

Analysis strategy

Comparing pipelines

The analysis of large and complex 16S rRNA gene sequencing data sets requires the use of bioinformatic tools. There are many pipelines available to process and analyse 16S rRNA gene sequencing data, including the commonly used QIIME (64), MG-RAST (65), UPARSE (66) (URL: https://www.drive5.com/usearch/manual/uparse_pipeline.html) and mothur (67). These packages contain sets of tools which facilitate the complete analysis of 16S rRNA gene data, from quality control to operational taxonomic unit (OTU) clustering. Where they differ is predominantly in their accessibility to those with limited computational knowledge and in the availability of documentation.

Nilakanta *et al* compared seven different packages (mothur, QIIME, WATERS, RDPipeline, VAMPS, Genboree, and SnoWMan) and concluded that while all of these packages provide effective pipelines for 16S rRNA gene analysis, the extensive documentation which accompanies mothur and QIIME provides them with an advantage over the other packages (68). Plummer and Twin analysed a single data set using QIIME, mothur and MG-RAST and found that there were few differences in the results when considering taxonomic classification and diversity (69). However, there were differences in the ease of use of each of these packages and the time required for analysis, with QIIME being the quickest analysis package (approximately 1 hour) and MG-RAST being the slowest (approximately 2 days, due to the need for manual quality control to remove multiple annotations of reads). The authors do state that although MG-RAST is the slowest analysis method, it is perhaps the most suitable package for users with no command line experience.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest
Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

Accepted Manuscript Posted Online

Ultimately, the choice of analysis package will be made on the basis of the user's level of experience in bioinformatics and on the available resources at the user's host institution.

Quality control, alignment and taxonomic assignment

It is essential to carry out quality filtering to remove DNA sequences which are of unexpected length, have long homopolymers, contain ambiguous bases or do not align to the correct 16S rRNA gene region. Critically, sequences should then be screened for chimeras, as the presence of chimeric sequences can affect the interpretation of the final dataset and could, for example, over-inflate perception of community diversity (70). A variety of tools have been developed to remove chimeric sequences such as UCHIME (66) and Chimera Slayer (70). By including a mock bacterial community in a sequencing run, since the true sequences in these are known, the number of chimeric sequences can be calculated (58).

Sequences should then be aligned to a reference alignment, or assigned to a suitable reference using a sequence classifier such as the RDP classifier which uses a naïve Bayesian approach based on 8-mers (71). Schloss showed that alignment quality can significantly impact diversity and can artificially inflate the number of bacterial OTUs, and advised against using alignments which do not take into account the secondary structure of the 16S gene (72). Of the three most commonly used alignments which are guided by secondary structure (i.e. greengenes (73), RDP (74) and SILVA (75)), the greengenes alignment was observed to be of poor quality, leading to significantly greater richness and diversity estimates.

Post-alignment, sequences and OTUs are assigned taxonomies based upon their similarity to training sets, most commonly constructed from the greengenes, RDP and SILVA databases. Errors within these databases, caused by sequencing/PCR errors (76) or by the incorrect

Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology
Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

labelling of sequences (77), may lead to the misidentification of sequences. Another issue when relying on databases for taxonomic assignment is their bias towards bacteria which are clinically relevant in humans, meaning that researchers investigating non-human hosts or environmental samples may struggle to assign taxonomy to their sequences. For example, in a study of the honey bee gut microbiota, disagreement was found between the three databases listed above upon carrying out taxonomic assignments (78). At genus level, the three databases concurred in their assignments for only 13% of sequences. The classification of sequences was improved by including bee-specific full length 16S rRNA gene sequences in the training set, highlighting the need to include more representative sequences from a greater number of habitats.

This has been highlighted by Werner *et al* who advised using the largest and most diverse database possible (79). This group also found that trimming the reference sequences to the primer region of interest improved classification depth. However, in a more extensively studied environment such as the human intestine, Ritari *et al* found that making a personalised reference database containing only bacterial species which were known to inhabit that niche led to an increase in lower taxonomic level assignments, probably due to less competition among sequences compared to large databases (80).

Operational taxonomic unit picking methods

Operational taxonomic units, or OTUs, are the common currency of 16S or marker gene studies of microbiomes. The term was originally coined by Sokal and Sneath (81), and in its more general usage refers simply to groups of organisms that are closely related. There are two major methods for defining OTUs – reference-based and *de novo*. In reference-based clustering, sequences from a community are clustered against a known reference database, and in *de novo* clustering, the sequences are clustered according to pairwise distance

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

measures. Reference-based OTUs are sometimes referred to as "*phylotypes*" (82). As with many areas of microbiome analysis, the evidence is mixed as to which of the two approaches is best. It has been found that *de novo* methods perform better when considering the quality of OTU assignments (83), with another study showing that de novo OTUs were unstable (84). However, Westcott and Schloss argued that OTUs can be stable yet still incorrect, and in particular showed that some reference-based techniques were sensitive to the order of sequences in the database. Sul *et al* found that reference-based techniques produced similar results to *de novo*, with the added benefit of low computational overheads and the ability to compare datasets from different variable regions (85). Indeed, perhaps the major difference between reference- and *de novo* based methods is that the latter has a significantly greater computational overhead, with the need to compare every sequence to every other sequence in its most naïve form.

Even within clustering tools, the choice of parameters has been shown to have a critical impact on the results. Whilst a threshold of 97% has become standard, Patin *et al* have shown that 16S rRNA gene sequences as similar as 99% can represent functionally distinct microorganisms, which means that functionally diverse species would be clustered at the 97% threshold (86). However, that may rely on accurate sequences, and if those don't exist, the 97% threshold can help avoid over-estimation of biodiversity (87). Susceptibility to differing parameters may also be pipeline-dependent (88). Given the controversy and potential biases of clustering sequences, some have suggested methods and models for using individual sequences to represent OTUs (i.e. remove the clustering step entirely) (89–92).

Correcting for gene copy number

Different bacterial species also have varying copy numbers of the 16S rRNA gene (93, 94) which can lead to misinterpretations when comparing the abundance of bacterial OTUs or

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

attempting to construct a "true" description of the microbial community within a sample (95). It is unusual in 16S rRNA gene studies to accurately know the copy numbers for all identified OTUs. Therefore, tools have been developed which seek to correct for copy number variation using sequence databases and phylogenetic information to give a more accurate picture of the relative abundances of these OTUs. These include Copyrighter (96), rrNDB (93), functions in the picante R package and pplacer (97) and part of the PICRUSt package (98).

As these techniques are reliant on databases the same problems are present as for taxonomic identification. Principally, lesser studied bacterial taxonomies are less likely to be represented. It is also important to note that when comparing OTUs between samples rather than within a sample (e.g. when comparing treatment effects), the impact of copy number variation is reduced as the under or over representation of OTUs would be consistent across samples as long as the same methodology had been used.

Contamination issues

Microbial DNA contamination arising from DNA extraction kits, PCR reagents and the lab environment may have a particularly large effect when studying low microbial biomass samples. Salter *et al* found that contamination in DNA extraction kits not only varied by manufacturer but by individual lot and that samples processed in separate laboratories contained different types of contaminating DNA (99). This lack of predictability led the authors to suggest that "negative" (or reagent-only) controls should be run alongside samples in all 16S rRNA gene metabarcoding studies. If reagent-only controls are not included, this can lead to the misinterpretation of results. When Salter *et al* analysed a dataset comparing nasopharyngeal microbiota samples from children at two time-points they found that while the time-points appeared to cluster separately, this effect was mainly due to bias caused by

Accepted Manuscript Posted Online

contamination from the extraction kits used. Randomisation of samples prior to processing may help avoid the introduction of this type of bias. Contamination could also lead to the false identification of microbial communities where they do not in fact exist (100) and could affect our understanding of which bacteria are relevant in clinical samples (101).

The amplification of background contaminants from PCR reagents could perhaps be avoided via the use of primer-extension PCR (102) but this would have no effect on contamination originating from other sources. Several methods have been suggested to remove contaminating DNA from reagents and the lab environment including: UV and γ radiation (103–107); DNA intercalation by 8-methoxypsoralen, ethidium monoazide and propidium monoazide (104, 106–108); enzymatic treatments (105–107, 109–111) silica-based membrane filtration (112); $CsCl_2$ density gradient centrifugation (111) and bleach/CoPA solution treatment (105). These methods have shown variable effects on contamination levels and PCR sensitivity and the inclusion of reagent-only controls alongside these decontamination measures is still recommended.

What should be done with sequencing data from reagent-only controls is still under debate. It is often not appropriate to simply remove all of the bacterial OTUs found in controls as these may overlap with OTUs which can genuinely be found in samples (108). Other methods have been suggested which take into account the abundance of OTUs to predict the likelihood of sequence reads having originated from contamination. These include an adaptation of the neutral community model (12) and combining qPCR data with OTU relative abundance data to compare the absolute abundance of contaminating OTUs in controls and samples (113). However, the field is rapidly reaching consensus that, due to contamination issues, not including reagent-only controls can negatively impact the quality control of sequence data.

When planning a 16S study, the inclusion of reagent-only controls (i.e. DNA extraction kit and PCR controls) is advised.

Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology
Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Conclusions

The study of complex microbial communities using high-throughput sequencing platforms has allowed better understanding of a variety of biological systems and the impact of various conditions (e.g. disease states) on the host microbiome. When considering the literature, it is clear that bias can be introduced into microbiota studies at all methodological stages, from sampling to bioinformatic analysis. While the variety of different 16S rRNA gene metabarcoding methodologies might seem overwhelming, the main factor to keep in mind when designing a microbiota study is consistency. It is paramount to use consistent methodology throughout a study to minimise potential biases which could lead to spurious results.

The volume of studies attempting to define best practice for various stages of the microbiome experimental process is large, and we cover only some of the literature in this review. Unfortunately, as can be seen, there is little consensus, and further studies are unlikely to find any. The reality is that many of the biases described in this review are context- and environment- specific, and whilst individual studies may be true within their context, their conclusions may not be transferable to other studies. Clearly, with biases possible at every step, a good experimental design is essential. Recording and publication of all experimental metadata is essential for understanding microbiome studies, and unfortunately many currently published studies lack these data.

Trying to find consensus in the literature is challenging, with many studies producing conflicting evidence about the effects of various steps in the experimental process. It is therefore essential that consistency is maintained within a study, and there must be an acceptance that comparison between studies may not be possible.

In summary, we recommend extracting DNA from fresh samples if possible; if not, samples should be stored in a consistent manner (i.e. at the same temperature, for the same duration and with or without cryprotectant) with appropriate metadata being recorded. The use of a mechanical lysis step is recommended to minimise potential biases due to some microbial cells being more resistant to lysis. The selection of appropriate primers should be made after careful consideration of the literature, but it is important to note that even universal primers will not amplify all bacteria in a given sample. Sequencing both mock bacterial communities and "negative"/reagent-only controls is important for determining background contamination and sequencing error rate, and should at least be included for each sequencing run and even better, for every batch of commercial reagents/kits. To reduce the chance of OTU inflation caused by sequencing errors, consider complete overlap of MiSeq reads, which translates as targeting a single hypervariable region. Finally, and to re-iterate – record every aspect of your experiment and report it in the methods section and remember that the critical consideration is consistency in methodology at each stage.

Acknowledgments

This project was supported by the Biotechnology and Biological Sciences Research Council (BBSRC: BB/N016742/1 (PI: Prof. Mick Watson), BB/N01720X/1 (PI: Prof. Mick Watson), BB/K501591/1 (PI: Prof. Jos Houdijk, SRUC), BB/J01446X/1 (PI: Dr. Gerry McLachlan, The Roslin Institute), including institute strategic programme and national capability awards to The Roslin Institute (BBSRC: BB/P013732/1, BB/J004235/1, BB/J004243/1). SRUC receives support from Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS).

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology
Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

References

1. Hong PY, Croix JA, Greenberg E, Gaskins HR, Mackie RI. 2011. Pyrosequencing-based analysis of the mucosal microbiota in healthy individuals reveals ubiquitous bacterial groups and micro-heterogeneity. PLoS One 6:e25042.

2. Glendinning L, Wright S, Pollock J, Tennant P, Collie D, McLachlan G. 2016. Variability of the sheep lung microbiota. Appl Environ Microbiol 82:3225–3238.

3. Mendes LW, Tsai SM. 2014. Variations of bacterial community structure and composition in mangrove sediment at different depths in Southeastern Brazil. Diversity 6:827–843.

4. Steven B, Gallegos-Graves LV, Belnap J, Kuske CR. 2013. Dryland soil microbial communities display spatial biogeographic patterns associated with soil depth and soil parent material. FEMS Microbiol Ecol 86:101–113.

5. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vázquez-Baeza Y, Gonzalez A, Knight R, Dunn RR, Fierer N. 2014. Temporal variability is a personalized feature of the human microbiome. Genome Biol 15:531.

6. Araújo-Pérez F, Mccoy AN, Okechukwu C, Carroll IM, Smith KM, Jeremiah K, Sandler RS, Asher GN, Keku TO. 2012. Differences in microbial signatures between rectal mucosal biopsies and rectal swabs. Gut Microbes 3:530–535.

7. Glendinning L, Wright S, Tennant P, Gill AC, Collie D, McLachlan G. 2017. Microbiota in exhaled breath condensate and the lung. Appl Environ Microbiol 83:AEM.00515–17.

8. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, Waghorn GC, Janssen PH. 2013. Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. PLoS One 8:e74787.

9. Paz HA, Anderson CL, Muller MJ, Kononoff PJ, Fernando SC. 2016. Rumen bacterial community composition in holstein and jersey cows is different under same dietary condition and is not affected by sampling method. Front Microbiol 7:1206.

10. Ramos-Morales E, Arco-Pérez A, Martín-García AI, Yáñez-Ruiz DR, Frutos P, Hervás G. 2014. Use of stomach tubing as an alternative to rumen cannulation to study ruminal fermentation and microbiota in sheep and goats. Anim Feed Sci Technol 198:57–66.

11. Bassiouni A, Cleland EJ, Psaltis AJ, Vreugde S, Wormald P-J. 2015. Sinonasal microbiome sampling: a comparison of techniques. PLoS One 10:e0123216.

12. Bassis CM, Moore NM, Lolans K, Seekatz AM, Weinstein RA, Young VB, Hayden MK. 2017. Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. BMC Microbiol 17.

13. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. 2015. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. PLoS One 10:1–14.

14. Portillo MC, Leff JW, Lauber CL, Fierer N. 2013. Cell size distributions of soil bacterial and archaeal taxa. Appl Environ Microbiol 79:7610–7617.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

15. McKain N, Genc B, Snelling TJ, Wallace RJ. 2013. Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. J Microbiol Methods 95:381–383.

16. Rubin BER, Gibbons SM, Kennedy S, Hampton-Marcell J, Owens S, Gilbert JA. 2013. Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS One 8:e70460.

17. Lauber CL, Zhou N, Gordon JI, Knight R. 2011. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett 307:80–86.

18. Fouhy F, Deane J, Rea MC, O'Sullivan Ó, Ross RP, O'Callaghan G, Plant BJ, Stanton C. 2015. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. PLoS One 10:e0119355.

19. Bahl MI, Bergström A, Licht TR. 2012. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. FEMS Microbiol Lett 329:193–7.

20. Choo JM, Leong LEX, Rogers GB. 2015. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5:1–10.

21. Tedjo DI, Jonkers DMAE, Savelkoul PH, Masclee AA, Van N. 2015. The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. PLoS One 10(5):e0126685.

22. Hale VL, Tan CL, Knight R, Amato KR. 2015. Effect of preservation method on spider monkey (*Ateles geoffroyi*) fecal microbiota over 8 weeks. J Microbiol Methods 113:16–26.

23. Dominianni C, Wu J, Hayes RB, Ahn J. 2014. Comparison of methods for fecal microbiome biospecimen collection. BMC Microbiol 14:1–6.

24. Kuske C, Banton K, Adorada D, Stark P, Hill K, Jackson P. 1998. Small-scale DNA sample preparation method for field PCR detection of microbial cells and spores in soil. Appl Environ Microbiol 64:2463–72.

25. Schrader C, Schielke A, Ellerbroek L, Johne R. 2012. PCR inhibitors - occurrence, properties and removal. J Appl Microbiol 113:1014–1026.

26. Desneux J, Pourcher AM. 2014. Comparison of DNA extraction kits and modification of DNA elution procedure for the quantitation of subdominant bacteria from piggery effluents with real-time PCR. Microbiologyopen 3:437–445.

27. Mackenzie BW, Waite DW, Taylor MW. 2015. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Front Microbiol 6:130.

28. Gerasimidis K, Bertz M, Quince C, Brunner K, Bruce A, Combet E, Calus S, Loman N, Ijaz UZ, Kennedy N, Walker A, Berry S, Salonen A, Nikkila J, Jalanka-Tuovinen J, Boer R, Peters R, Gierveld S, McOrist A, Jackson M, Bird A, Nechvatal J, Ram J, Basson M, D'Amore R, Ijaz U, Schirmer M, Quince C, Ijaz U, Loman N, Reichardt N, Barclay A, Weaver L, Godon J, Zumstein E, Dabert P, Gerasimidis K, Bertz M, Hanske L, Caporaso J, Lauber C, Walters W, Masella A, Bartram A, Truszkowski J, Wang Q, Garrity G, Tiedje J, Katoh K, Standley D,

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

Price M, Dehal P, Arkin A, McMurdie P, Holmes S, Paradis E, Claude J, Strimmer K, Schliep K, Cardoso P, Rigal F, Borges P. 2016. The effect of DNA extraction methodology on gut microbiota research applications. BMC Res Notes 9:365.

29. Guo F, Zhang T. 2013. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. Appl Microbiol Biotechnol 97:4607–4616.

30. Hart ML, Meyer A, Johnson PJ, Ericsson AC. 2015. Comparative evaluation of DNA extraction methods from feces of multiple host species for downstream next-generation sequencing. PLoS One 10:e0143334.

31. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM, Satsangi J, Flint HJ, Parkhill J, Lees CW, Hold GL. 2014. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One 9:1–9.

32. Maukonen J, Simoes C, Saarela M. 2012. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. FEMS Microbiol Ecol 79:697–708.

33. Salonen A, Nikkilä J, Jalanka-Tuovinen J, Immonen O, Rajilić-Stojanović M, Kekkonen RA, Palva A, de Vos WM. 2010. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods 81:127–134.

34. Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome 3:26.

35. Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, Licht TR. 2014. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome 2:19.

36. Yang B, Wang Y, Qian P-Y. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17:135.

37. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. 2015. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:1–15.

38. Cruaud P, Vigneron A, Lucchetti-Miganeh C, Ciron PE, Godfroy A, Cambon-Bonavita MA. 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. Appl Environ Microbiol 80:4626–4639.

39. Ghyselinck J, Pfeiffer S, Heylen K, Sessitsch A, De Vos P. 2013. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. PLoS One 8:e71360.

40. Chakravorty S, Helb D, Burday M, Connell N, Alland D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods 69:330–339.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

41. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso G, Walters WA, Knight R, Fierer N. 2012. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem 43:1450–1455.

42. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 2016. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol 16:123.

43. Gohl D, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould T, Clayton J, Johnson T, Hunter R, Knights D, Beckman K. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol 34:942–949.

44. Kanagawa T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96:317–323.

45. Ahn J, Kim B, Song J, Weon H. 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.

46. Wu J, Jiang X, Jiang Y, Lu S, Zou F, Zhou H. 2010. Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. BMC Microbiol 10:255.

47. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17:1–20.

48. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112–20.

49. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:6.

50. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, DeWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133–138.

51. Loman NJ, Watson M. 2015. Successful test launch for nanopore sequencing. Nat Methods 12:303–304.

52. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. 2015. Improved data analysis for the MinION nanopore sequencer. Nat Methods 12:351–356.

53. Fichot EB, Norman RS. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome 1:10.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
AEM
Applied and Environmental Microbiology
Applied and Environmental Microbiology

54. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ 4:e1869.

55. Benítez-Páez A, Portune KJ, Sanz Y. 2016. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. Gigascience 5:4.

56. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol 16:274.

57. Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert J a, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–4.

58. Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One 6:e27310.

59. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414.

60. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA, Vaginal Microbiome C. 2015. The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 15:66.

61. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. 2016. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res 45:e23.

62. Smets W, Leff JW, Bradford MA, McCulley RL, Lebeer S, Fierer N. 2016. A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. Soil Biol Biochem 96:145–151.

63. Stammler F, Glasner J, Hiergeist A, Holler E, Weber D, Oefner PJ, Gessner A, Spang R. 2016. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 4:28.

64. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich K, Gordon JI, Huttley GA, Kelley ST, Knights D, Jeremy E, Ley RE, Lozupone CA, Mcdonald D, Muegge BD, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA. 2011. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336.

65. Meyer F, Paarmann D, D’Souza M, Etal. 2008. The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes. BMC Bioinformatics 9:1–8.

66. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200.

67. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–41.

68. Nilakanta H, Drews KL, Firrell S, Foulkes MA, Jablonski KA. 2014. A review of software for analyzing molecular sequences. BMC Res Notes 7:830.

69. Plummer E, 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 8:283–291.

70. Haas BJ, Gevers D, Earl a M, Feldgarden M, Ward D V, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ, The Human Microbiome Consortium, Petrosino JF, Knight R, Birren BW. 2011. Chimeric 16S rRNA sequence formation and detection in sanger and 454-pyrosequenced PCR amplicons. Genome Res 21:494–504.

71. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267.

72. Schloss PD. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 6:e1000844.

73. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–72.

74. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y. 2013. Ribosomal database project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642.

75. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

76. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol 71:7724–7736.

77. Kozlov AM, Zhang JJ, Yilmaz P, Glockner FO, Stamatakis A. 2016. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res 44:5022–5033.

78. Newton ILG, Roeselers G. 2012. The effect of training set on the classification of honey bee gut microbiota using the Naive Bayesian Classifier. BMC Microbiol 12:221.

79. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. 2012. Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. ISME J 6:94–103.

80. Ritari J, Salojarvi J, Lahti L, de Vos WM. 2015. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 16:1056.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

81. Sokal RR, Sneath PHA. 1965. Principles of numerical taxonomy. J Mammology 46:111–112.

82. Schloss PD, Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77:3219–26.

83. Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 3:e1487.

84. He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, Zhou H-W. 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome 3:20.

85. Sul WJ, Cole JR, Jesus EDC, Wang Q, Farris RJ, Fish JA, Tiedje JM. 2011. Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. Proc Natl Acad Sci U S A 108:14637–42.

86. Patin NV., Kunin V, Lidström U, Ashby MN. 2013. Effects of OTU clustering and PCR artifacts on microbial diversity estimates. Microb Ecol 65:709–719.

87. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12:118–123.

88. Schmidt TSB, Matias Rodrigues JF, von Mering C. 2015. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. Environ Microbiol 17:1689–1706.

89. Tikhonov M, Leach RW, Wingreen NS. 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. ISME J 9:68–80.

90. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker gene data analysis. ISME J doi: 10.1038/ismej.2017.119.

91. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583.

92. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.

93. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. 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.

94. Větrovský T, Baldrian P. 2013. The variability of the 16S rRNA gene in bacterial genomes and Its consequences for bacterial community analyses. PLoS One 8:e57923.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online

95. Props R, Kerckhof F-M, Rubbens P, De Vrieze J, Hernandez Sanabria E, Waegeman W, Monsieurs P, Hammes F, Boon N. 2016. Absolute quantification of microbial taxon abundances. ISME J 11:584–587.

96. Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. 2014. CopyRighter: A rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2:11.

97. Kembel SW, Wu M, Eisen JA, Green JL. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comput Biol 8:e1002743.

98. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RL V, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821.

99. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87.

100. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, Leite R, Elovitz MA, Parry S, Bushman FD. 2016. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome 4:1–11.

101. Laurence M, Hatzis C, Brash DE. 2014. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One 9:e97876.

102. Chang SS, Hsu HL, Cheng JC, Tseng CP. 2011. An efficient strategy for broad-range detection of low abundance bacteria without DNA decontamination of PCR reagents. PLoS One 6:e20303.

103. Tamariz J, Voynarovska K, Prinz M, Caragine T. 2006. The application of ultraviolet irradiation to exogenous sources of DNA in plasticware and water for the amplification of low copy number DNA. J Forensic Sci 51:790–794.

104. Humphrey B, McLeod N, Turner C, Sutton JM, Dark PM, Warhurst G. 2015. Removal of contaminant DNA by combined UV-EMA treatment allows low copy number detection of clinically relevant bacteria using pan-bacterial real-time PCR. PLoS One 10:e0132954.

105. Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl E-M. 2010. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS One 5:e13042.

106. Klaschik S, Lehmann L, Raadts A, Hoeft A, Stuber F. 2002. Comparison of different decontamination methods for reagents to detect low concentrations of bacterial 16S DNA by real-time-PCR. Mol Biotechnol 22:231–242.

107. Corless CE, Guiver M, Borrow R, Edwards-Jones V, Kaczmarski EB, Fox AJ. 2000. Contamination and sensitivity issues with a real-time universal 16S rRNA PCR. J Clin Microbiol 38:1747–1752.

Applied and Environmental Microbiology

AEM

Applied and Environmental Microbiology

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest

Accepted Manuscript Posted Online
Applied and Environmental Microbiology
AEM
Applied and Environmental Microbiology

108. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. 2016. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog 8:24.

109. Czurda S, Smelik S, Preuner-Stix S, Nogueira F, Lion T. 2016. Occurrence of fungal DNA contamination in PCR reagents: Approaches to control and decontamination. J Clin Microbiol 54:148–152.

110. Mennerat A, Sheldon BC. 2014. How to deal with PCR contamination in molecular microbial ecology. Microb Ecol 68:834–841.

111. Rand KH, Houck H. 1990. Taq polymerase contains bacterial DNA of unknown origin. Mol Cell Probes 4:445–450.

112. Mohammadi T, Reesink HW, Vandenbroucke-Grauls C, Savelkoul PHM. 2005. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. J Microbiol Methods 61:285–288.

113. Lazarevic V, Gaïa N, Girard M, Schrenzel J. 2016. Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. BMC Microbiol 16:1–8.

Downloaded from http://aem.asm.org/ on February 12, 2018 by guest