FEMS Microbiology Ecology, 95, 2019, fiz045

doi: 10.1093/femsec/fiz045
Advance Access Publication Date: 9 April 2019
Minireview

MINIREVIEW

Issues and current standards of controls in microbiome research

Bastian V.H. Hornung[1,2,*], Romy D. Zwittink[1,2] and Ed J. Kuijper[1,2,3]

[1]Department of Medical Microbiology, Leiden University Medical Center, PO Box 9600, 2300RC, Leiden, The Netherlands, [2]Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center, PO Box 9600, 2300RC, Leiden, The Netherlands and [3]Netherlands Donor Feces Bank, Leiden University Medical Center, PO Box 9600, 2300RC, Leiden, The Netherlands

*Corresponding author: Center for Microbiome Analysis and Therapeutics, Leiden University Medical Center, PO Box 9600, 2300RC, Leiden, The Netherlands. Tel: +31(0)715261229; E-mail: bastian.hornung@gmx.de

One sentence summary: Current issues and standards with positive and negative controls in microbiome research are discussed.

Editor: Marcus Horn

ABSTRACT

Good scientific practice is important in all areas of science. In recent years this has gained more and more attention, especially considering the 'scientific reproducibility crisis'. While most researchers are aware of the issues with good scientific practice, not all of these issues are necessarily clear, and the details can be very complicated. For many years it has been accepted to perform and publish sequencing based microbiome studies without including proper controls. Although in recent years more scientists realize the necessity of implementing controls, this poses a problem due to the complexity of the field. Another concern is the inability to properly interpret the information gained from controls in microbiome studies. Here, we will discuss these issues and provide a comprehensive overview of problematic points regarding controls in microbiome research, and of the current standards in this area.

Keywords: positive control; negative control; microbiome; microbiota; metagenomics; best practices; contamination

INTRODUCTION

The microbiome field is a relatively new field of research. The hallmark publication by (Venter *et al.* 2004) is not even 15 years old. It was, and still is, exciting to sequence the microbial community of an environment to its full extent, and to learn about all its inhabitants and their possible functions. The subsequent rise in interest in this field was tremendous, leading to numerous publications in highly respected journals, e.g. (Tringe *et al.* 2005; Turnbaugh *et al.* 2009; van Nood *et al.* 2013).

The use of shotgun sequencing or targeted DNA amplicon sequencing to characterize the microbiome has its complications though. Most researchers are aware of the tremendous impact that DNA extraction has on the outcome of any microbiome study (Costea *et al.* 2017; Sinha *et al.* 2017a). It is also clear that these various DNA extraction methods might not produce very overlapping results (e.g. (Angelakis *et al.* 2016), among many others), which makes their interpretation difficult. There are various other concerns which are not specific to this field, but also apply here, like the crisis in reproducibility (Schloss 2018), or the public availability of data in a FAIR (Findable, Accessible, Interoperable and Reusable) way (Langille, Ravel and Fricke 2018).

One particular problem exists, which has been overlooked for a long time: the lack of controls in microbiome research. It is good practice to perform experiments with controls, to

Received: 7 December 2018; **Accepted:** 5 April 2019

ensure that all procedures were correctly performed, and that none of the steps have introduced false positive or false negative results. In microbiome research, however, controls have not been included in the majority of published studies. To clarify the extent of this, we reviewed all publications from the 2018 issues of Microbiome and the ISME journal (manually, as well as keyword searches for 'mock', 'blank' and 'control'). From the 265 publications utilizing high-throughput community sequencing of any type (16S, metagenomics, metatranscriptomics, 18S, ITS, virome, PacBio, Nanopore, others), only 30% (79) reported using any type of negative control, and only 10% (27) reported using a positive control. In some of these cases, however, it was unclear if the negative controls were actually sequenced (e.g. when samples were also used for qPCR), or the descriptions were insufficient to judge whether controls were adequate (e.g. 'appropriate controls were used at all steps', without further details or 'methods have previously been validated with positive controls'). During our research we also still noticed several high impact publications, which report results which are potentially indistinguishable from contaminations. Common to all these publications is the lack of controls, paired with the investigation of relatively low biomass microbiomes like mucosa (Zuo *et al.* 2019), amniotic fluid (Wang *et al.* 2018) (despite being heavily discussed earlier (Lauder *et al.* 2016)), gastric environment (Ferreira *et al.* 2018), pancreatic cancer (Pushalkar *et al.* 2018), respiratory tract (Nicola *et al.* 2017), human milk (Drago *et al.* 2017) (although the authors report a clean negative control) and sometimes pure in-silico studies might identify contaminations as biologically relevant (Tackmann *et al.* 2018). These examples come in addition to further publications mentioned in (de Goffau *et al.* 2018), which also did not use negative controls in their study setups. The substantiality of microbiome studies without controls might have been due to a combination of lack of knowledge, unavailability of positive controls and perceived costs associated with the inclusion of non-biological control samples. Indeed, there are numerous challenges with both types of controls in microbiome research, which will be summarized here, together with current developments in this field.

POSITIVE CONTROLS IN MICROBIOME RESEARCH

Selection of organisms

For a long time, positive controls were not used in microbiome research due to their unavailability. Positive controls are now commercially available in the form of defined synthetic communities, but their validity for specific research questions is uncertain, depending on the exact microbiome under investigation. For example, the controls from BEI resource, ⟨0:ext-link 3:href="https://www.beiresources.org/Catalog.aspx?f_instockflag=In±Stock%23~%23Discontinued%23~%23Temporarily±Out±of±Stock&q=mock%20community"0:ext-link-type="uri"⟩https://www.beiresources.org/Catalog.aspx?f_instockflag=In±Stock%23~%23Discontinued%23~%23Temporarily±Out±of±Stock&q=mock%20community⟨/0:ext-link⟩, and ATCC, ⟨0:ext-link 3:href="https://www.lgcstandards-atcc.org/en/Products/Cells_and_Microorganisms/By_Focus_Area/Microbiome_Research/Mock_Microbial_Communities.aspx"0:ext-link-type="uri"⟩https://www.lgcstandards-atcc.org/en/Products/Cells_and_Microorganisms/By_Focus_Area/Microbiome_Research/Mock_Microbial_Communities.aspx⟨/0:ext-link⟩, contain only bacteria, while the ZymoResearch control, ⟨0:ext-link 3:href="https://www.zymoresearch.eu/products/microbiomics"0:ext-link-type="uri"⟩https://www.zymoresearch.eu/products/microbiomics⟨/0:ext-link⟩, contains bacteria and fungi. Though the manufacturers took care to select relevant bacterial species, including pathogens and Gram-negative and -positive bacteria, it needs to be considered whether such a control is a valid representative for the investigated environment since archaea, viruses and other eukaryotes are not included. Therefore, the presence of archaea, viruses and other eukaryotes might be overlooked in the investigated samples (Bakker 2018). These microbes also pose their own challenges like variable amplicon length and within-species divergence (Palmer *et al.* 2018). New bacterial phyla are also still being discovered, sometimes even multiple phyla with various representatives at once (Karst *et al.* 2018), and we do not know anything about their physiology, including how resistant these organisms will be to the current DNA extraction methods and how they compare to well-studied organisms used in mock communities. This is not a problem of the controls, but of the microbiome field itself, where many of the microbes in these environments are unknown or uncultured. For a proper positive control, we would actually already need to know what microbes are present in a sample. All of this needs to be considered when selecting a positive control. When commercially available controls are unsuitable, custom designed positive controls might be needed, but standardized protocols to do so are currently lacking.

The interdependency between kits and controls

Another problem is the interdependency between the DNA extraction kit manufacturers and the positive control manufacturers. A kit or a method will be benchmarked using a positive control and this must be a part of the development process of the kit itself. At the end of development, the kit will be able to extract the correct proportions of DNA from the corresponding positive control. Despite the rigorous testing, it cannot be guaranteed that the kit will be able to extract correct DNA proportions from any type of community. Other factors within the community, like physical interactions between different cells or different kinds of metabolites (e.g. glycans) might interfere with the extraction (Angelakis *et al.* 2016), and these factors will vary between communities. This also applies to other potential positive controls, which might have different properties. Using one extraction kit on one mock community will therefore in some cases only indicate its performance on this mock community, and does not necessarily indicate its suitability for real biological samples or even other mock communities. The performance of different kits on varying mock communities has not been tested, and some kits might perform well on some mock communities, and rather poorly on others. The test of one kit with one mock community might therefore not in all cases be sufficient.

Amplification bias and related errors

If an approach is used which uses PCR amplification during library preparation (including all amplicon technologies), then it also needs to be considered that this step could introduce problems for accurately determining the composition of the microbiome. It has been shown that amplification biases exist, and that DNA fragments with a high or low GC content are not amplified in the same rate as fragments with an average GC content (Aird *et al.* 2011; Benjamini and Speed 2012). The possibly presence of this issue can be discovered with positive controls, since some organisms might be less efficiently

amplified. Both ZymoResearch and ATCC offer a pre-extracted DNA mix, which can be used to verify sequencing-related procedures, e.g. library preparation (Jones *et al.* 2015).This will help the researcher to distinguish amplification bias issues from DNA extraction issues (Costea *et al.* 2017; Sinha *et al.* 2017a), but can also give insight into issues related to using varying amounts of DNA for amplification (Bowers *et al.* 2015). Furthermore the various sequencing machines used in the field display different kinds of sequencing errors (Minoche, Dohm and Himmelbauer 2011), which can be specific to the sequenced DNA (Nakamura *et al.* 2011). Errors related to sequencing might furthermore appear randomly, without being reproducible (Yeh *et al.* 2018). This error might lead to the same faulty interpretation like amplification bias, which is that some bacteria might be absent in the sequenced samples, or less abundant than they are. Usage of a positive control will prevent this issue.

Influence of bioinformatics processing—clustering and filtering

Bioinformatics processing of the sequencing data also contributes to the issues of accurately determining community composition. There are no fully agreed upon standards for raw data processing, and the various parameters of 16s rRNA gene sequence analysis software, like QIIME (Caporaso *et al.* 2010), often need to be tweaked. Ideally, positive controls in microbiome research aid in correct data processing and the parameters can be optimized with working positive controls. A frequently considered parameter is the OTU similarity level for clustering, e.g. 97%, 98.5% or 100% (Patin *et al.* 2013). But any type of clustering based on a similarity of less than 100% might lump two sequences that differ by at least one nucleotide into a single OTU and produce inaccurate results (He *et al.* 2015). Not clustering sequences is not necessarily the solution to this problem, since there can be heterogeneity within the rRNA genes in the same microorganism, and this and other effects might inflate the number of OTUs detected (Nguyen *et al.* 2016; Nearing *et al.* 2018).

Influence of bioinformatics processing—taxonomic assignment and binning

Usage of a positive control will furthermore enable the researcher to pinpoint a limited amount of possible issues in the taxonomic assignment. As it is known, the public sequence databases contain various errors, including contaminations (Sheik *et al.* 2018), and sequences with incorrectly assigned taxonomy or incorrectly assigned names (Nilsson *et al.* 2006; Federhen 2015). If such an error falls into the taxonomic range of the positive control, the taxonomic assignment might come back faulty. Since the content of the control is known, the researcher might be able to find the sequences which obstruct the correct assignment, and remove these from their database. The general occurrence of this error is not likely, since many of the used pipelines for the assignment of taxonomy based on 16S will discard assignments which are not predominant. In case of metagenomics, this case could be less clear. The full genomic diversity of a species is often not sampled, with only a few specimen being sequenced, and one misassigned species might disturb the taxonomic assignment severely.

Another issue related to this topic is the potential depth of assignment. In some circumstances it is not possible to assign sequences to the correct genus or species, e.g. when genera are too closely related (e.g. *Escherichia* and *Shigella*). If this is the case for some species in the positive control, the researcher will get assignments on e.g. the respective family level, and will be made aware of this restriction and can interpret his data with this in mind. In the case of 16S amplicon data, most researchers will be aware that e.g. their assignment in the family of Enterobacteriaceae might be their specific organism of interest like *E. coli*, but for metagenomics, this problem can be more prevalent. A positive control can obviously aid here only in a limited way, but it makes a researcher potentially aware of this issue.

If binning (Sangwan, Xia and Gilbert 2016) of a genome will be attempted in a metagenomic sample, then it needs to be also considered that a different type of positive control might be more applicable. Having multiple more strongly related organisms (e.g. two strains of the same species) of different abundances in a positive control would be beneficial to in-depth confirm the result of the binning process, especially when more organism-specific parameters like GC content are one of the factors used during the binning process.

NEGATIVE CONTROLS IN MICROBIOME RESEARCH

Negative controls in microbiome research face also different problems. Some of these are the same as for the positive controls. E.g. it needs to be considered at which step which control is necessary, and which way to sample these is the most suitable. The details of the most important steps will be described below.

Sampling controls

The first step in which negative controls should be taken into account is at sampling. If a cohort of patients is sampled from a similar site (such as oral swabs) by a trained researcher then a chance exists that either the researcher, the used sampling equipment or the environment could contaminate the samples. Including an appropriate negative control is, however, not always easy. In this setting, a control swab could easily be taken, unpacked in the same surrounding and handled by the same researcher without taking any sample. But the question arises how a negative control would be taken if for example faecal material is considered, since no material to perform a negative DNA extraction would be available, and sampling the air or container with any other kind of instrument like a swab would not be a true negative control. Being consistent when taking samples (Vandeputte *et al.* 2017) is therefore key to minimize technical variation.

Contamination by the researcher

One of the potential sources of contamination, the researcher itself, also needs to be taken into account. With large numbers of samples, we cannot assume that potential contamination by the researcher will be evenly distributed over all samples during processing. If *Cutibacterium* (formerly *Propionibacterium*) *acnes* (a common skin commensal) unexpectedly appears in samples, but not in the negative controls, it cannot necessarily be concluded that it is not a contaminant from the sample processing, since the possibility exists that only some samples in a bigger cohort are contaminated. This could be resolved by culturing the suspicious organisms from the original sample, but specific bacterial species may be difficult to culture (Staley and Konopka

1985; Rinke *et al.* 2013; Boers *et al.* 2018) from samples containing many different bacterial species. Therefore, a negative culture would not give us definitive proof of absence. Furthermore it does not exclude that the contamination occurred in the original sample, and not during processing.

The 'kitome'

The most influential discovery regarding the necessity of including negative controls, is the recognition of a 'kitome' (Salter *et al.* 2014). It has become apparent that various DNA extraction kits contain their own unique microbiome, which may be indistinguishable from the real microbiome (e.g. often discussed (Bhatt *et al.* 2013) and mentioned in (Salter *et al.* 2014), as well as multiple examples in (de Goffau *et al.* 2018)). This means that even if the researcher in the laboratory has worked in a complete sterile environment with appropriate techniques, outcomes are affected by the DNA in the extraction kit. While this has been studied extensively, and specific organisms have been identified as common kit contaminants (Laurence, Hatzis and Brash 2014), this might cause problems when environments are studied where these organisms could be present (as mentioned in Jousselin *et al.* 2016), or where it is unknown if they could be present. While these extraction kits are not sold as 'sterile', it should also be noted that even if material is sold as 'sterile', it might still contain bacterial DNA, and therefore cannot be excluded as a source of contamination (van der Horst *et al.* 2013). Therefore extraction controls need to be performed, but index hopping (see further below) might interfere with this procedure.

Index hopping

Index hopping is another problem in microbiome research (Costello *et al.* 2018; MacConaill *et al.* 2018), as well as in other fields (Sinha *et al.* 2017b; Griffiths *et al.* 2018). It can occur 0%–10% of the sequenced data (Sinha *et al.* 2017a), depending on the used Illumina platform. If samples are multiplexed during the same run, the possibility exists that indexes from one sample will be incorrectly assigned to another sample. This is caused by non-ligated adapters from one sample (possible a low-biomass sample, where more adapters were in the sample than actual DNA), which will randomly ligate to free DNA from another sample on the same sequencing run. Negative controls might therefore contain data, which incorrectly originates from the other samples during the same sequencing run (although not always, since this can depend on the way the samples are loaded onto the sequencer). In practice, the negative controls might contain exactly the same profile as the sequenced samples. In these cases it is impossible to distinguish between true contamination and index hopping, making the controls (negative as well as positive) potentially useless. Also, proposed practices like the subtraction of shared OTUs between negative controls and samples (Edmonds and Williams 2017) would falsify the results in this situation, since the contamination is derived from the samples, and not from the researcher, DNA extraction kit, or the environment. This could be circumvented by sequencing negative controls on a separate run, but again, in practice this will increase the price of sequencing to an unacceptable high level, and is therefore highly unlikely to happen. If the sequencing is not performed in-house, this would also complicate the procedure even more, since it would be necessary to request different lanes for different samples. The final result would also no longer be a control of the whole process, since samples will not be sequenced at the same time and on the same machine anymore. While evaluating a sample, it therefore needs to be considered if data in a low biomass sample and negative controls could be derived from a high biomass sample or positive control, and results need to be interpreted with caution.

CURRENT STANDARDS AND DEVELOPMENTS

No easy solutions exist for many of the above mentioned problems. It can be advised to take negative controls during the sampling process and at all further steps, if feasible (for example Galan *et al.* 2016; Jousselin *et al.* 2016; Zhong *et al.* 2018). New methodologies, that reduce the amount of contamination during processing, might also be necessary, e.g. (Boers, Hays and Jansen 2017; Minich *et al.* 2018), but their implementation in the laboratory is not always easy.

For the data analysis of negative controls it is mainly advised to focus on the number of reads obtained. A clean negative control should have few reads, which excludes major contaminations from all possible sources. Kitome components or actual contaminations in these negative controls are in such cases unlikely to be abundant, and unlikely to have a significant impact on the analysis, even if present. The outcome of the negative controls are particularly relevant for samples with low microbial biomass (Biesbroek *et al.* 2012; Lusk 2014; Glassing *et al.* 2016; Karstens *et al.* 2018; Velasquez-Mejia, de la Cuesta-Zuluaga and Escobar 2018), since even low amounts of contamination could have an impact here. Especially in these cases the connection between the samples and the corresponding negative controls needs to be carefully evaluated. A possibility to resolve this contamination is to remove OTUs identified in the negative controls from the actual samples (e.g. Edmonds and Williams 2017). This is only applicable, when it can be ensured that these are actually contaminants, and not e.g. derived from another sample, as explained in the index hopping section. If a positive control in the same run has potentially acquired contaminations, then this would help in resolving this issue partially. Additional low abundance OTUs in the positive controls will give the researcher an idea which level of minor abundances can be reasonably filtered out. This again applies mainly for samples with a high biomass, while for low biomass samples more caution is needed.

Further developments to prevent index hopping by usage of multiple indexes at the same time (Costello *et al.* 2018; MacConaill *et al.* 2018) will hopefully make this interpretation easier in the future, since then it would be possible to also exclude contamination from high biomass samples on a run with low biomass samples. Currently this is not yet the case, which makes resolving the source of contamination often complicated.

The observation that the concentration of contaminants is inversely correlated to amplicon concentration has been reported multiple times (Salter *et al.* 2014; Jervis-Bardy *et al.* 2015; Lazarevic *et al.* 2016). While it has not yet been implemented, the microbiome research community might need to consider using a dilution series of single samples as a control, to make the identification of contaminants easier, although specific challenges also apply (Multinu *et al.* 2018). Standards need to be agreed upon, since having different dilution ratios might make the comparison between studies complicated. The dilution steps will most likely be dependent on the expected biomass, and a general protocol how to determine the appropriate steps will be necessary. Software solutions to address the outcome of such negative controls have already been developed (Davis *et al.* 2018),

and seem to be performing well (Karstens *et al.* 2018), so that researchers are directly able to deal with their control samples.

Another indicator of contamination in metagenomic samples can be derived from the insert size of the reads (Olm *et al.* 2017). Olm *et al.* identified a contaminant in their dataset by aberrant insert size distribution, indicating another origin of the DNA. The low heterogeneity within the contaminant data may also be an useful indicator, though it is not applicable in all cases, e.g. when strain transfer between microbiomes is a part of the research question (Smillie *et al.* 2018).

The idea of adding spike-in controls (addition of external microbial DNA to a sample) could be an interesting approach. It will not only allow tracking of samples, but also aid absolute quantification and quantification of cross-contamination/index hopping (Galan *et al.* 2016; Hardwick *et al.* 2018; Palmer *et al.* 2018; Tourlousse, Ohashi and Sekiguchi 2018), and should probably be implemented for low-biomass samples. Furthermore, it seems that some kind of quality control regarding misassigned sequences is even possible after the experiment has been conducted (Wright and Vetsigian 2016), but as the authors point out, this might need to be investigated on a per experiment basis, and therefore does not seem to be very practical in most circumstances. Another computational method to correct for misassigned sequences has been developed for single-cell data (Larsson *et al.* 2018), but its potential applicability to metagenomics data still needs to be shown.

For positive controls, the most diverse mock community available should be chosen, if it is applicable for the proposed research project. This should minimally prevent overfitting of a protocol to the used control, although it does not deal with the microbial dark matter. The possibility of creating one's own mock community can be considered, in case these samples will be processed more often and in case the available mock communities are not enough comparable to the investigated microbiome.

If expected organisms do not appear in the mock community, then a PCR for identifying species specific genes in the mock community needs to be performed, to ensure that this organism was actually present in the sequenced sample. The same applies for unexpected occurrences of contamination in the mock community.

In general, for good scientific practice, controls should be included at all steps. This should be done with the consideration that the microbiome field is developing rapidly and during the progress of a research project new methods might be published, which will help with proper interpretation. Even if proper interpretation is not possible, publishing the results from the positive and negative controls needs to be mandatory, so that the reader can interpret the findings with caution if necessary.

CONCLUSIONS

While the field of microbiome studies has become an established research area, and many best practices exist (Goodrich *et al.* 2014; Boers, Jansen and Hays 2016; Kim *et al.* 2017; Knight *et al.* 2018; Martin *et al.* 2018; Pollock *et al.* 2018), they have not gone in-depth regarding the usage of controls. We summarized the issues regarding good scientific practice in light of controls, to make scientists aware of these issues, so that they can be addressed in their own research. Many of the problems regarding the lack of controls have been recognized very recently, many even last year, and new strategies to prevent these are likely to be developed soon. While no standards exist for the processing and interpretation of controls, any further studies in this field need to include controls to prevent erroneous results and improve data interpretation. We therefore urge that future study designs must include all the necessary controls, as listed below, until the scientific community has agreed upon better standards:

- A negative control for sampling, to ensure that sampling equipment (tubes, swaps, etc) are not contaminated.
- A negative control for DNA extractions, to ensure identification of potential kit contaminations.
- A negative control for sequencing, to ensure that no large-scale cross-contamination between samples takes place.
- A positive control for DNA extraction, to ensure that the contained organisms can be sufficiently extracted with the used methods.
- A positive control for sequencing (a pre-extracted DNA mix), to ensure that the sequencing itself did not introduce any errors.

We furthermore would like to appeal to the journals in the microbiome field to introduce a screening procedure/checklist for these items, e.g. some journals have done to ensure proper replication within their publications.

ACKNOWLEDGEMENTS

We thank Debby Bogaert (University of Edinburgh), Mark Davids (Amsterdam UMC) and Prokopis Konstanti (Wageningen University & Research) for their thoughts and experiences regarding this topic. We acknowledge the ESCMID Study Group for Host And Microbiota Interaction (ESGHAMI) for stimulating discussion about this topic. We would also like to thank the reviewers for their useful comments.

We acknowledge the publication by (McDermott, Partridge and Bromberg 2018) for sparking the idea for the graphical abstract, and the webcomic xkcd, xkcd.com, for further inspiration. Furthermore we want to acknowledge the websites clker.com and publicdomainvectors.org for providing the used graphics.

We would furthermore like to point the readers to the publication by (Eisenhofer *et al.* 2019), which got published during the work on this manuscript, and which covers partially a similar subject.

FUNDING

BH is supported by an unrestricted grant from Vedanta Biosciences Inc.

Conflict of interest. None declared.

REFERENCES

Aird D, Ross MG, Chen WS *et al.* Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biol* 2011;**12**:R18.

Angelakis E, Bachar D, Henrissat B *et al.* Glycans affect DNA extraction and induce substantial differences in gut metagenomic studies. *Sci Rep* 2016;**6**:26276.

Bakker MG. A fungal mock community control for amplicon sequencing experiments. *Mol Ecol Resour* 2018;**18**:541–56.

Benjamini Y, Speed TP. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Res* 2012;**40**:e72.

Bhatt AS, Freeman SS, Herrera AF *et al*. Sequence-based discovery of bradyrhizobium enterica in cord colitis syndrome. *N Engl J Med* 2013;**369**:517–28.

Biesbroek G, Sanders EA, Roeselers G *et al*. Deep sequencing analyses of low density microbial communities: Working at the boundary of accurate microbiota detection. *PLoS One* 2012;**7**:e32942.

Boers SA, Hays JP, Jansen R. Novel micelle pcr-based method for accurate, sensitive and quantitative microbiota profiling. *Sci Rep* 2017;**7**:45536.

Boers SA, Hiltemann SD, Stubbs AP *et al*. Development and evaluation of a culture-free microbiota profiling platform (mycrobiota) for clinical diagnostics. *Eur J Clin Microbiol Infect Dis* 2018;**37**:1081–9.

Boers SA, Jansen R, Hays JP. Suddenly everyone is a microbiota specialist. *Clin Microbiol Infect* 2016;**22**:581–2.

Bowers RM, Clum A, Tice H *et al*. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 2015;**16**:856.

Caporaso JG, Kuczynski J, Stombaugh J *et al*. Qiime allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**:335–6.

Costea PI, Zeller G, Sunagawa S *et al*. Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol* 2017;**35**:1069–76.

Costello M, Fleharty M, Abreu J *et al*. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 2018;**19**:332.

Davis NM, Proctor D, Holmes SP *et al*. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data, *Microbiome* 2018;**6**:226.

de Goffau MC, Lager S, Salter SJ *et al*. Recognizing the reagent microbiome. *Nat Microbiol* 2018;**3**:851–3.

Drago L, Toscano M, De Grandi R *et al*. Microbiota network and mathematic microbe mutualism in colostrum and mature milk collected in two different geographic areas: Italy versus burundi. *ISME J* 2017;**11**:875–84.

Edmonds K, Williams L. The role of the negative control in microbiome analyses. *The FASEB Journal* 2017;**31**

Eisenhofer R, Minich JJ, Marotz C *et al*. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* 2019;**27**:105–17. S0966-842X(18)30253-130497919.

Federhen S. Type material in the ncbi taxonomy database. *Nucleic Acids Res* 2015;**43**:D1086–1098.

Ferreira RM, Pereira-Marques J, Pinto-Ribeiro I *et al*. Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. *Gut* 2018;**67**:226.

Galan M, Razzauti M, Bard E *et al*. 16s rrna amplicon sequencing for epidemiological surveys of bacteria in wildlife. *mSystems* 2016;**1**:e00032–16.

Glassing A, Dowd SE, Galandiuk S *et al*. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog* 2016;**8**:24.

Goodrich JK, Di Rienzi SC, Poole AC *et al*. Conducting a microbiome study. *Cell* 2014;**158**:250–62.

Griffiths JA, Richard AC, Bach K *et al*. Detection and removal of barcode swapping in single-cell rna-seq data. *Nat Commun* 2018;**9**:2667.

Hardwick SA, Chen WY, Wong T *et al*. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun* 2018;**9**:3096.

He Y, Caporaso JG, Jiang XT *et al*. Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. *Microbiome* 2015;**3**:20.

Jervis-Bardy J, Leong LE, Marri S *et al*. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of illumina miseq data. *Microbiome* 2015;**3**:19.

Jones MB, Highlander SK, Anderson EL *et al*. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A* 2015;**112**:14024–9.

Jousselin E, Clamens AL, Galan M *et al*. Assessment of a 16s rrna amplicon illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. *Mol Ecol Resour* 2016;**16**:628–40.

Karstens L, Asquith M, Davin S *et al*. Controlling for contaminants in low biomass 16s rrna gene sequencing experiments. *bioRxiv* 2018.

Karst SM, Dueholm MS, McIlroy SJ *et al*. Retrieval of a million high-quality, full-length microbial 16s and 18s rrna gene sequences without primer bias. *Nat Biotechnol* 2018;**36**:190–5.

Kim D, Hofstaedter CE, Zhao C *et al*. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 2017;**5**:52.

Knight R, Vrbanac A, Taylor BC *et al*. Best practices for analysing microbiomes. *Nat Rev Microbiol* 2018;**16**:410–22.

Langille MGI, Ravel J, Fricke WF. "Available upon request": Not good enough for microbiome data! *Microbiome* 2018;**6**:8.

Larsson AJM, Stanley G, Sinha R *et al*. Computational correction of index switching in multiplexed sequencing libraries. *Nat Methods* 2018;**15**:305–7.

Lauder AP, Roche AM, Sherrill-Mix S *et al*. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* 2016;**4**:29.

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

Lazarevic V, Gaïa N, Girard M *et al*. Decontamination of 16s rrna gene amplicon sequence datasets based on bacterial load assessment by qpcr. *BMC Microbiol* 2016;**16**:73.

Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One* 2014;**9**:e110808.

MacConaill LE, Burns RT, Nag A *et al*. Unique, dual-indexed sequencing adapters with umis effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* 2018;**19**:30.

Martin TC, Visconti A, Spector TD *et al*. Conducting metagenomic studies in microbiology and clinical research. *Appl Microbiol Biotechnol* 2018;**102**:8629–46.

McDermott JE, Partridge M, Bromberg Y. Ten simple rules for drawing scientific comics. *PLoS Comput Biol* 2018;**14**:e1005845.

Minich JJ, Zhu Q, Janssen S *et al*. Katharoseq enables high-throughput microbiome analysis from low-biomass samples. *mSystems* 2018;**3**:e00218–00217.

Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biol* 2011;**12**:R112.

Multinu F, Harrington SC, Chen J *et al*. Systematic bias introduced by genomic DNA template dilution in 16s rrna gene-targeted microbiota profiling in human stool homogenates. *mSphere* 2018;**3**:e00560–17.

Nakamura K, Oshima T, Morimoto T *et al*. Sequence-specific error profile of illumina sequencers. *Nucleic Acids Res* 2011;**39**:e90.

Nearing JT, Douglas GM, Comeau AM *et al*. Denoising the denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 2018;**6**:e5364.

Nguyen NP, Warnow T, Pop M *et al*. A perspective on 16s rrna operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* 2016;**2**:16004.

Nicola I, Cerutti F, Grego E *et al*. Characterization of the upper and lower respiratory tract microbiota in piedmontese calves. *Microbiome* 2017;**5**:152.

Nilsson RH, Ryberg M, Kristiansson E *et al*. Taxonomic reliability of DNA sequences in public sequence databases: A fungal perspective. *PLoS One* 2006;**1**:e59.

Olm MR, Butterfield CN, Copeland A *et al*. The source and evolutionary history of a microbial contaminant identified through soil metagenomic analysis. *MBio* 2017;**8**:e01969–16.

Palmer JM, Jusino MA, Banik MT *et al*. Non-biological synthetic spike-in controls and the amptk software pipeline improve mycobiome data. *PeerJ* 2018;**6**:e4925.

Patin NV, Kunin V, Lidstrom U *et al*. Effects of otu clustering and pcr artifacts on microbial diversity estimates. *Microb Ecol* 2013;**65**:709–19.

Pollock J, Glendinning L, Wisedchanwet T *et al*. The madness of microbiome: attempting to find consensus "best practice" for 16s microbiome studies. *Appl Environ Microbiol* 2018;**84**:e02627–17.

Pushalkar S, Hundeyin M, Daley D *et al*. The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discov* 2018;**8**:403–16.

Rinke C, Schwientek P, Sczyrba A *et al*. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**:431–7.

Salter SJ, Cox MJ, Turek EM *et al*. Reagent and laboratory contamination can critically impact sequence-based microbiome analysis. *BMC Biol* 2014;**12**:87.

Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 2016;**4**:8.

Schloss PD. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio* 2018;**9**:e00525-18.

Sheik CS, Reese BK, Twing KI *et al*. Identification and removal of contaminant sequences from ribosomal gene databases: Lessons from the census of deep life. *Front Microbiol* 2018;**9**:840.

Sinha R, Abu-Ali G, Vogtmann E *et al*. Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (mbqc) project consortium. *Nat Biotechnol* 2017a;**35**:1077–86.

Sinha R, Stanley G, Gulati GS *et al*. Index switching causes "spreading-of-signal" among multiplexed samples in illumina hiseq 4000 DNA sequencing. *bioRxiv* 2017b.

Smillie CS, Sauk J, Gevers D *et al*. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* 2018;**23**:229–40 e225.

Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* 1985;**39**:321–46.

Tackmann J, Arora N, Schmidt TSB *et al*. Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites. *Microbiome* 2018;**6**:192.

Tourlousse DM, Ohashi A, Sekiguchi Y. Sample tracking in microbiome community profiling assays using synthetic 16s rrna gene spike-in controls. *Sci Rep* 2018;**8**:9095.

Tringe SG, von Mering C, Kobayashi A *et al*. Comparative metagenomics of microbial communities. *Science* 2005;**308**:554–7.

Turnbaugh PJ, Hamady M, Yatsunenko T *et al*. A core gut microbiome in obese and lean twins. *Nature* 2009;**457**:480–4.

Vandeputte D, Tito RY, Vanleeuwen R *et al*. Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol Rev* 2017;**41**:S154–67.

van der Horst J, Buijs MJ, Laine ML *et al*. Sterile paper points as a bacterial DNA-contamination source in microbiome profiles of clinical samples. *J Dent* 2013;**41**:1297–301.

van Nood E, Vrieze A, Nieuwdorp M *et al*. Duodenal infusion of donor feces for recurrent *clostridium difficile*. *N Engl J Med* 2013;**368**:407–15.

Velasquez-Mejia EP, de la Cuesta-Zuluaga J, Escobar JS. Impact of DNA extraction, sample dilution, and reagent contamination on 16s rrna gene sequencing of human feces. *Appl Microbiol Biotechnol* 2018;**102**:403–11.

Venter JC, Remington K, Heidelberg JF *et al*. Environmental genome shotgun sequencing of the sargasso sea. *Science* 2004;**304**:66–74.

Wang J, Zheng J, Shi W *et al*. Dysbiosis of maternal and neonatal microbiota associated with gestational diabetes mellitus. *Gut* 2018;**67**:1614.

Wright ES, Vetsigian KH. Quality filtering of illumina index reads mitigates sample cross-talk. *BMC Genomics* 2016;**17**:876.

Yeh YC, Needham DM, Sieradzki ET *et al*. Taxon disappearance from microbiome analysis reinforces the value of mock communities as a standard in every sequencing run. *mSystems* 2018;**3**:e00023-18.

Zhong ZP, Solonenko NE, Gazitua MC *et al*. Clean low-biomass procedures and their application to ancient ice core microorganisms. *Front Microbiol* 2018;**9**:1094.

Zuo T, Lu X-J, Zhang Y *et al*. Gut mucosal virome alterations in ulcerative colitis. *Gut* 2019. gutjnl-2018-318131.