Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling

Sandra Reitmeier
Technische Universitat Munchen
Thomas CA Hitch
Rheinisch-Westfalische Technische Hochschule Aachen Medizinische Fakultat
Nikolaos Fikas
Hellenic Center for Marine Research
Bela Hausmann
Medizinische Universitat Wien
Amanda E Ramer-Tait
University of Nebraska-Lincoln
Klaus Neuhaus
Technische Universitat Munchen
David Berry
Universitat Wien
Dirk Haller
Technische Universitat Munchen
Ilias Lagkouvardos
Hellenic Center for Marine Research
Thomas Clavel (tclavel@ukaachen.de)
RWTH University Hospital https://orcid.org/0000-0002-7229-5595

Research

Keywords: Microbiomes; high-throughput amplicon sequencing; 16S rRNA gene; spurious taxa; singletons; standardization; effective richness

Posted Date: January 17th, 2020

DOI: https://doi.org/10.21203/rs.2.21240/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at ISME Communications on June 29th, 2021. See the published version at https://doi.org/10.1038/s43705-021-00033-z.

Abstract

Background: 16S rRNA gene amplicon sequencing is a very popular approach for studying microbiomes. However, varying standards exist for sample and data processing and some basic concepts such as the occurrence of spurious sequences have not been investigated in a comprehensive manner, which was done in the present study. Methods: Using defined communities of bacteria in vitro and in vivo , we searched for sequences not matching the expected species (i.e. , spurious taxa) and determine a threshold of occurrence relevant for adequate data analysis. The origin of spurious taxa was then investigated via large-scale amplicon queries. We also assessed the impact of varying sequence filtering stringency on diversity readouts in human fecal and peat soil communities. Results: 16S rRNA gene amplicon data processing based on Operational Taxonomic Units (OTUs) clustering and singleton removal, a commonly used approach that discards any taxa represented by only one sequence across all samples, delivered approx. 50% (mock communities) to 80% (gnotobiotic mice) spurious taxa on average. This spurious fraction of taxa was lower based on amplicon sequence variants (ASVs) analysis but varied depending on the gene region targeted and the barcoding system used. A relative abundance of 0.25% was identified as a threshold below which the analysis of spurious taxa can be prevented to a large extent. Most spurious taxa (approx. 70%) detected in simplified communities occurred in samples multiplexed in the same sequencing run and were present in only one of ten runs. Use of the 0.25% relative abundance threshold decreased the coefficient of variations calculated on richness in the same six human fecal samples across seven sequencing runs by 38% compared with singleton filtering. The output of beta -diversity analyses of human fecal communities was markedly affected by both the filtering strategy and the type of phylogenetic distances used for comparing samples. Importantly, major findings were confirmed by using data generated in a second sequencing facility. Conclusions: Handling of artifact sequences during bioinformatic processing of 16S rRNA gene amplicon data requires careful attention to avoid the generation of misleading findings. A threshold of relative abundance of 0.25% is more appropriate than singleton removal, although study-specific analysis strategies are mandatory. We propose the concept of effective richness, which will help comparing results across studies.

Introduction

Since the late 2000s, high-throughput sequencing of 16S rRNA gene amplicons has become the most popular method for rapid analysis of the diversity and composition of complex microbial communities [1]. Despite its popularity and usefulness, the method is prone to technical artifacts at various levels of the workflow, from sample processing to data analysis. For the latter, one common approach that has been used for decades [2] and is included in many freely available processing pipelines [3, 4] consists of building clusters of sequences representing single microbial entities, also known as operational taxonomic units (OTUs), at a defined level of sequence identity determined by the user (usually >97% used as proxy for species-level diversity) [5]. Other strategies, such as exact/amplicon sequence variant (ESV/ASV) analysis [6] are available, but do not refute the relevance of OTU-based approaches, as both can be applied in a synergistic manner and generate complementary readouts. Nonetheless, diversity measures derived from OTU-based datasets are strongly influenced by the choice of settings to filter both original sequences and built OTUs based on their occurrence. As a consequence, there has been for instance a lot of confusion in the field as to how many bacterial species can be detected in the human intestine by sequencing, with values ranging from a few hundred to several thousand [7, 8]. Reference studies based on low-error amplicon analysis protocols or shotgun metagenomics suggested the detection of 150 to 200 species in one individual sample, albeit based on sample size <200 [9, 10]. Despite the widespread use of 16S rRNA gene amplicon sequencing approaches, it is still unclear which thresholds of occurrence are most suitable to help excluding falsely detected taxa, hereon referred to as spurious taxa. One very widespread strategy is to remove so called singletons, defined as OTUs represented by only one sequence across all samples analyzed. However, OTU tables created by such an approach have a high likelihood of still containing spurious OTUs, the number of which inflates with increasing sample size and sequencing depth.

In the present study, we aimed at assessing filtering thresholds usable to exclude spurious taxa from high-throughput 16S rRNA amplicon datasets. We used mixtures of known bacteria both *in vitro* (mock communities) and from gnotobiotic mice to determine a consensus threshold and further studied the impact of various filtering approaches on final readouts from both literature datasets and in-house sequence data generated in two different sequencing facilities. We also investigated the occurrence of spurious taxa in various ecosystems to identify their potential origin. Of note, our purpose is neither to set rules that should be strictly followed by all nor to refute data published in the literature. We simply intend to draw attention to an analysis parameter that despite its trivial aspect is very important but often neglected, leading to false interpretations that rapidly spread throughout the scientific community and beyond. Filtering strategies must always depend on the specific aims of a given study and the type of samples analyzed.

Methods

Datasets and samples

A schematic view of the study is provided in **Fig. 1a**. For the determination of filtering cutoffs, two types of reference communities were used: *in vitro* mixtures of known bacteria (mock communities) and *in vivo* communities from gnotobiotic mice, *i.e.*, ex-germfree mice colonized with defined sets of known bacteria. Seven different mock communities from published studies with raw sequencing data available and two additional in-house generated datasets were used (**Table 1**). This was complemented by the analysis of amplicon datasets generated from fecal samples of gnotobiotic mice colonized with four different mixtures of bacteria (**Table 2**). The 16S rRNA gene sequences and taxonomies of microbes included in all these reference communities are provided in **Additional File 1** and **2** (**Table S1** and **S2)**. To further analyze the impact of different filtering strategies on data processing outcomes, two comprehensive studies with open access to their sequence data [11, 12] and data from six human fecal samples generated in multiple sequencing runs at the Microbiome Core Facility of the ZIEL Institute for Food and Health (Freising, Germany) were used (**Fig. 1b**). In order to validate findings, amplicon datasets from one Mock community and a peat soil sample were generated at the Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna (JMF) (Austria), including multiple replicates and sequencing runs (**Fig. 1c**).

Sample processing for sequencing at ZIEL

DNA extraction and library preparation of mock communities and samples generated in the present study were performed as described previously [13]. Briefly, DNA was purified on columns (Macherey-Nagel) after mechanical lysis (bead-beating) and the V3-V4 region of 16S rRNA genes was amplified in a two-step approach (15+10 cycles) [14] using primers 341F and 785R [15] following a combinatorial dual (CD) indexing strategy. Libraries were purified using magnetic beads (Beckman-Coulter), pooled in equimolar amounts and then sequenced in paired-end mode (2 x 275 nt) using the v3 chemistry on an Illumina MiSeq following the manufacturer's instructions. The platform was semi-automated (Biomek4000 pipetting robot, Beckman Coulter) to increase reproducibility. Moreover, the workflow systematically included two negative controls (a DNA-extraction control, *i.e.*, sample-free DNA-stabilization solution, and a PCR blank, *i.e.*, PCR-grade water as template) for each 46 samples sequenced.

Sample processing for sequencing at JMF

DNA extraction and library preparation were performed as described previously [16]. Briefly, Mock communities were ordered as extracted DNA standards (Zymo Research, cat. no. D6311), whilst a peat soil sample was extracted using a phenol-chloroform extraction method after mechanical lysis (bead-beating) [17]. The V3-V4 or V4 regions of 16S rRNA genes were amplified (30 cycles) using primers 341F and 785R [15] or 515F and 806R [18], respectively, modified with a linker sequence [16] and barcoded (8 cycles) in a combinatorial dual (CD) or unique dual (UD) setup. The barcoded samples were purified and normalized over a SequalPrep™ Normalization Plate Kit (Invitrogen) using a Biomek® NXP Span-8 pipetting robot (Beckman Coulter), pooled and concentrated on columns (Anlaytik Jena). Sequencing libraries were prepared with the Illumina TruSeq Nano Kit and sequenced in paired-end mode (2 x 300 nt; v3 chemistry) on an Illumina MiSeq following the manufacturer's instructions. The workflow systematically included four negative controls (PCR blanks, *i.e.*, PCR-grade water as template) for each 90 samples sequenced.

Data analysis

Raw amplicon data were analyzed using IMNGS (www.imngs.org) [4], a platform that integrates a UPARSE-based, *de novo* OTU-picking strategy [19]. A sequence identity threshold of 97% was used for clustering sequences. Additional parameters were: barcode mismatch tolerated, 1; no. of nucleotides trimmed at each the 5'- and 3'-end, 5; trim quality score, 3; max. expected errors, 3; min. read length, 0; max. read length, 600. Data were first processed without any filtering of OTUs. These primary outputs were further processed using the desired filtering cutoffs, *i.e.* (i) by removing singletons only (*i.e.* OTUs represented by only one sequence across all samples), which is a commonly used strategy [20], or (ii) by removing those OTUs that did not occur at least at a defined relative abundance in at least one sample (*e.g.* 0.5% was a threshold that we had proposed previously below which the variation of OTU-specific relative abundances between replicate samples increases exponentially [21]). Phylogenetic trees of resulting representative OTU sequences were constructed

in FastTree [22]. Whenever appropriate, closed-reference picking was performed in QIIME v1.9.1 using default settings [3]. Processed data were further analyzed using Rhea for the generation of diversity and composition readouts [21]. The identity of OTUs (*i.e.* their match to the reference sequences of species included in the defined communities) was assessed using BLAST [23], considering ≥97% sequence identify, ≥90% coverage, and an e-value < 0.00001 as positive hits. The taxonomy of spurious OTUs was assigned using SILVA [24]. Besides the OTU-based approach, the DADA2 pipeline version 1.12.1. was used on data from mock communities and gnotobiotes to generate amplicon sequence variants (ASV) with the recommended settings for paired-end sequences (adjusted options: maxEE, 3.3; truncQ, 3; maxN, 0; truncLeft, 10; truncRight, 20) [6]. Samples processed at JMF were analyzed using DADA2 version 1.14.0 following a previously described workflow [25] with pooling for each run (adjusted options: truncLen, 150 for V4; truncLen, 230 for V3/4; maxEE, adjusted for each run).

Large-scale amplicon sequencing studies

All spurious OTUs from the mock communities across ten sequencing runs were collapsed at 97 % sequence identity using UCLUST [5] to remove redundancy. Samples in IMNGS (build 1905) [4] with unambiguous origin were grouped into five categories (human, mouse, soil, freshwater, and marine). All pre-calculated OTUs in the selected IMNGS samples were searched against the spurious OTUs from each run in parallel and assigned to their best match with identity >97% over 90% of the query length. Results were merged into an occurrence map of all spurious OTUs in each IMNGS sample tested. Due to different primers being used across studies, there is no guarantee of overlap between the sequences of spurious OTUs and those from IMNGS samples. Hence, IMNGS samples with no hit to any of the spurious OTUs were not considered as it was unclear if spurious OTUs were indeed absent from these samples or regions were simply not matching. The prevalence of each spurious OTU in all sample categories was calculated as the percentage of samples in the given category that were positive at a threshold >0.25% relative abundance. When spurious OTUs occurred in different sample categories, a Z-test was used to determine whether sequences could be considered as exclusive to one of these sample categories ($p < 0.05$).

Statistics

All statistical tests were performed in R, version 3.4.0. P-values < 0.05 were considered as significant (after adjustment for multiple testing whenever appropriate using the Benjamini-Hochberg method). For microbial community analyses, detailed descriptions of statistical tests applied are provided in the Rhea support information and in the corresponding scripts (https://lagkouvardos.github.io/Rhea). Sequence counts were normalized according to the minimum sum count across the given OTU table prior to calculation of *alpha*-diversity parameters. *Beta*-diversity analyses were based on the calculation of unweighted and generalized UniFrac distances [26, 27].

Results

Filtering threshold for handling spurious sequences

We first used bacterial communities of known composition (simplified communities) to assess the occurrence of spurious taxa and to determine at which relative abundances they start to appear. To propose a cutoff potentially applicable to different 16S rRNA gene amplicon studies, we included reference data obtained with different variable regions and sequencing pipelines and originating from both *in vitro* an *in vivo* communities varying in number and type of species (max. 58) (**Table 1** and **Table 2**).

Without any filtering, sequence clustering generated an average of 508 ± 355 OTU (min. 52; max. 1,081) per mock community (10 to 58 target species in theory) and 105 ± 50 OTU (min, 55; max. 215) per gnotobiotic community (4 to 12 target species in theory). Up to 87% of these OTUs were spurious (*i.e.* they did not match the expected classification of species contained in the corresponding artificial community) (**Fig. 2a**). On average, the proportion of spurious OTU in both mock communities and samples from gnotobiotic mice was slightly lower after removing singletons, without reaching significance (50.8 *vs.* 64.3%, $p = 0.227$; 57.5% *vs.* 65.7%, $p = 0.70$). Interestingly, the proportion of spurious OTUs was higher in gnotobiotic mice independent of filtering ($p < 0.001$), suggesting that the matrix in which members of the defined communities are (here fecal material) influences the outcome. Besides the goal of removing spurious taxa, it is of course important to include as many true molecular species as possible into the analysis. Even without any cutoff,

not all target species could be detected: the percentage of positive hits was 94.9 and 92.3% for mock communities and gnotobiotic mice, respectively (**Fig. 2b**).

To determine a filtering threshold that allowed exclusion of most spurious taxa, we recorded the relative abundance of the first spurious OTU occurring in each of the reference community datasets (**Fig. 2c**). Median values of approx. 0.12% relative abundance were observed (**Fig. 2d**). Besides one outlier in the Mock communities (0.44% relative abundance), all values were below 0.25%, a threshold that was selected for all further analyses. The distribution of spurious OTUs and positive hits obtained after applying this new cutoff on all reference communities is depicted in **Fig. 2e**. Whilst the number of spurious taxa decreased drastically (4.0 *vs.* 50.8% for mock communities and 1.0 *vs.* 57.0% for gnotobiotes; $p \leq 0.01$), the number of positive hits was not affected significantly (87.2 *vs.* 93.7% for mock communities and 82.4 *vs.* 88.7% for gnotobiotes; $p > 0.50$) by the 0.25%-cutoff *vs.* singletons removal, respectively (**Fig. 2e**). Note that the diversity of reference communities in the gnotobiotic mice was relatively low (4-12 members; **Table 2**), resulting in a proportionally marked drop in the percentage of positive hit (8-25%) even if only one true member is excluded after filtering due to a low relative abundance (which is an expectable event considering a classical exponentially decreasing distribution of species occurrence in gut environments).

Next, we used another bioinformatic pipeline for ASV analysis to confirm the results aforementioned. Processing of the very same simplified communities generated a total number of 42 ± 25 ASVs (min. 16; max. 98) for mock communities (10 to 58 target species in theory) and 14 ± 8 ASVs (min. 4; max. 25) for gnotobiotes (4 to 12 target species in theory). Altogether, a marked decrease in spurious taxa was observed compared with OTU clustering, with an average of 8.6% ± 11.8 and 4.4% ± 6.4 spurious sequences after singletons removal for Mock and gnotobiotic communities, respectively (**Fig. 2f** compared with **Fig. 2a**). Of note, the DADA2 pipeline used for the ASV approach does not infer sequence variants only supported by a single read (singletons), due to a lack of confidence in their existence relative to sequencing errors. Hence, data corresponding to no filtering with the OTU-based approach were not generated. On average, the first spurious ASV occurred at a relative abundance of 0.10% ± 0.32. Applying the cutoff of 0.25% relative abundance completely removed spurious sequences (except for three outlying samples), albeit with a slight drop in positive hits for both Mock and gnotobiotic communities (**Fig. 2f**).

Ecology of spurious OTUs

We then investigated the diversity and origin of spurious taxa, *i.e.* those not matching any sequences of the reference species in the defined communities. Therefore, their taxonomic information and occurrence in >100,000 IMNGS-derived amplicon datasets [4] were combined and depicted in a dendrogram (**Fig. 3a**). Approximately half of the 678 non-redundant spurious OTUs belonged to the phylum Firmicutes followed by Bacteroidetes and Proteobacteria. Most of these OTUs were characterized by highest prevalence in human- and mouse-derived datasets, with values reaching up to 40 % in the thousands of tested samples (**Fig. 3a**). The majority (>20%) of spurious OTUs detected in human and mouse samples were exclusive to this respective category of habitats (**Fig. 3b**). This distribution implies that the type of samples multiplexed with target samples within a given sequencing run (in our case mouse and human gut samples) greatly influence the occurrence of spurious OTUs in target samples. Interestingly, >600 of the 678 spurious OTUs occurred in less than five of the ten sequencing runs tested, with approximately 450 of them occurring in only one run (**Fig. 3c**). This indicates that the majority of spurious taxa are sporadic cross-contaminations rather than generalist artifacts across sequencing runs. Whilst most of the spurious taxa were characterized by relative abundances between 0.25 and 2% in the IMNGS-amplicon datasets tested, they represented very dominant populations in a few samples (**Fig. 3d**).

Loose taxa filtering inflates alpha-diversity and increases heterogeneity

Spurious OTUs, as looked at in the present study, are *per se* low abundant: the cumulative relative abundance of spurious OTUs in the reference communities used above was approx. 1% on average. Hence, whilst spurious OTUs are not expected to substantially influence composition data, they can have a major influence on diversity (*e.g.* richness and evenness for *alpha*-diversity and between-sample distances for *beta*-diversity), as presented in the next sections. To assess the impact of filtering thresholds on analysis outcomes, we used recently published amplicon data from two comprehensive studies that included a substantial number of samples analyzed by Illumina sequencing of 16S rRNA gene amplicons and for which raw datasets could be retrieved from public repositories. The study by Flores *et al.* [12] (hereon referred to as Study-1) focused on dynamics of human body microbiomes over time, collecting samples weekly

from 85 college-age adults over a 3-month period; we focused only on gut samples in the present work. The second study published by Halfvarson *et al.* [11] (hereon referred to as Study-2) focused on shifts in the human fecal microbiota over time in patients with inflammatory bowel diseases *vs.* controls, including 683 fecal samples from 137 individuals. We emphasize again that the purpose of the present study is not to confirm or refute data from the literature, but rather to draw attention to an analysis parameter that can have a profound impact on results. In all following analyses, outcomes after the most commonly used approach of singletons filtering following *de novo* OTU clustering was compared with the 0.25%-cutoff introduced above (*i.e.* keeping only those OTUs occurring at a minimum relative abundance of 0.25% in at least one sample).

In both Study-1 and Study-2, filtering OTUs using the 0.25%-cutoff led to a significant decrease in richness by approx. two-fold, resulting in an average number of about 200 observed species per sample (**Fig. 4a**). More interestingly, when looking at individual-specific variations in richness by plotting inter-quartile ranges (IQR) across the different time points analyzed in the studies, the 0.25%-cutoff was associated with a significantly lower heterogeneity in richness (Study-1: IQR = 28.0 ± 17.8 *vs.* 70.6 ± 34.1, $p < 0.001$; Study-2: IQR = 17.0 ± 3.2 *vs.* 49.0 ± 10.4, $p = 2.5 \times 10^{-13}$) (**Fig. 4a**). Another helpful readout of *alpha*-diversity is the Shannon effective count, which takes into account the evenness of species distribution and can be, simply speaking, considered as a proxy for the number of most dominant species [21, 28]. Altogether, the trend observed for richness (less heterogeneity after 0.25%-filtering) was similar when considering Shannon effective counts (data not shown). However, lower effective counts after stringent filtering (0.25%) were not significantly different for Study-2, showing that Shannon effective counts can be useful to alleviate the impact of low abundant species.

In addition to these two literature studies, which focused on the analysis of distinct samples (different individuals with several time points in both studies), we also analyzed triplicates of six fecal samples from healthy human adults sequenced several times in-house. This dataset that consisted of the very same samples sequenced in seven different runs was ideal to test reproducibility depending on filtering thresholds. Across all runs, the coefficient of variations (CVs) calculated on richness values between the triplicates of each sample within a run were on average <5%, and lowest when applying the 0.25%-cutoff (**Fig. 4b**). In contrast, CVs of the richness within samples across sequencing runs increased to 20% on average with a peak at 40% when applying singletons filtering, which dropped to approx. 10% (average) and 30% (maximum) when applying the 0.25%-cutoff (Wilcoxon-Mann-Whitney test, $p = 0.004$) (**Fig. 4b**). This clearly indicates that 16S rRNA gene amplicon sequencing, at least as performed in our study, generates richness values that vary markedly between sequencing runs for the same sample, especially when following a loose OTU filtering strategy.

Between-sample comparisons are impacted by filtering strategies

We next assessed how filtering influenced *beta*-diversity analyses. To obtain reference data to which filtering strategies after *de novo* OTU clustering could be compared, a closed-reference protocol was also used, as in the published studies selected [11, 12].

In Study-1, the median unweighted distance across all individuals was approximately 0.5 after using reference-picking, including a broad range of within-host temporal variations (some individuals being characterized by more stable profiles than others) (**Fig. 4c**; left panel), as observed in the original study [12]. As expected, the strongest effect of filtering strategies was observed when using unweighted UniFrac distances: singletons removal was characterized by a higher overtime variation in profiles (median value of approx. 0.6 *vs.* 0.3 for the 0.25%-cutoff) (**Fig. 4c**; middle panel). Interestingly, whilst using generalized UniFrac distances leveled the difference between filtering approaches, it widened the range of individual-specific overtime variability around the median, potentially enhancing the discriminatory power between 'stable' and 'variable' individuals (**Fig. 4c**; right panel).

In Study-2, one of the main findings in the original work was that volatility (*i.e.* variations overtime within individuals) was highest in patients suffering from Crohn's disease with an ileal phenotype that underwent ileocecal resection (ICD-r) [11]. We confirmed this finding by using reference-based picking and unweighted distances, as done in the published manuscript (**Fig. 4d**; left panel). However, whenever applying *de novo* clustering, this difference could only be observed when using the 0.25%-cutoff in combination with unweighted distances (**Fig. 4d**; middle panel). The absence of a significant differences when using unweighted distances after singletons removal may indicate the confounding effect of spurious OTUs (**Fig. 4d**; middle panel). The absence of difference when applying generalized distances in general (**Fig. 4d**; right panel) suggests that individual-specific overtime variations in this study are bound to the presence/absence of taxa rather than to changes in composition.

Validation studies

In order to confirm utility of the 0.25%-cutoff inferred from data generated at the Core Facility Microbiome of the ZIEL Institute for Food and Health (TU Munich, Germany), additional samples were processed and analyzed independently at the Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna (JMF).

First, processing of a log-distributed version of the ZymoBIOMICS Microbial Community Standard (Zymo Research GmbH) containing eight bacterial strains confirmed the advantage of applying the 0.25%-filtering approach. Twenty-five replicates of the same DNA sample were sequenced on five sequencing runs (1-8 replicates per run) using either the V4 region combined with combinatorial dual (CD) barcoding or the V3/4 regions with unique dual (UD) barcoding (two and three runs, respectively). V4-CD *vs.* V3/4-UD yielded 31 ± 16/8 ± 2 ASVs (min: 13/5, max: 57/10), respectively. Spurious ASVs (*i.e.*, all sequences with a Hamming distance to the reference > 1) were greatly reduced using a 0.25% filtering step from 73 ± 8 to 2 ± 2 and from 13 ± 15 to 0 in V4-CD *vs.* V3/4-UD, respectively (**Fig. 5a**). This occurred at a loss of 15% true taxa in the case of V4-CD whilst no change was observed with V3/4-UD (**Fig. 5b**). As the highest relative abundance reached by any spurious ASV was 0.28% and the true taxa detected corresponded to dominant members of the standard community, the cumulative relative abundance of true taxa was high (>98%) in all cases (**Fig. 5b**).

Second, peat soil DNA [17] was analyzed to confirm suitability of our filtering approach for non-gut samples. One identical DNA sample was sequenced on three different runs (3-5 replicates per run) using primers 515F/806R (V4 region) and CD barcoding. The ASV table was rarefied to the minimum sum count (9,104) and analyzed without or with filtering (*i.e.*, only ASVs observed at a relative abundance >0.25% in at least one replicate were kept). Richness was calculated using ampvis2 [29]. Applying the 0.25%-cutoff decreased the number of observed ASVs from 408 ± 71 to 139 ± 5 and, more importantly, the IQR from 101 to 7 (**Fig. 5b**). Unweighted UniFrac distances within and between runs as calculated using ampvis2 were also compared before and after filtering. Sequences were aligned using MAFFT [30] and phylogeny was inferred using FastTree. Whilst the community makeup in the soil sample varied substantially between sequencing runs without additional filtering, the 0.25%-cutoff reduced this variation to the level observed within runs without filtering (**Fig. 5c**). Replicates within a run were very similar after applying the 0.25%-cutoff. Altogether, these data demonstrate that stringent filtering delivers more stable values obtained for the very same sample sequenced in replicates across several sequencing runs.

Discussion

The goal of our work was to investigate the occurrence of spurious taxa in high-throughput 16S rRNA gene amplicon datasets. The findings clearly stress the need for cautious handling of low abundant sequences.

The advent of high-throughput sequencing led researchers to conclude that as many as tens of thousands of species inhabit the human gastrointestinal tract [31] and that organs known to be sterile host microbiomes [32] before this concept was recently ruled out [33-35]. We clearly show that processed sequence datasets based on the very common approach of removing singletons still contain a high proportion of spurious taxa that inflates *alpha*-diversity values. Of course, enhancing the filtering stringency by increasing relative abundance thresholds comes with the risk of losing true diversity. Hence, analysis strategies should always be adapted to the main goal of a study and no universal threshold can be proposed. It is beyond the scope of the present work to dissect the contribution of each wet lab and *in silico* step to the introduction of spurious taxa into datasets. Nonetheless, we observed that many spurious taxa most likely originate from samples multiplexed with the defined communities in one sequencing run, despite the implementation of multiple negative controls and an automated sample processing workflow. One cause for spurious sequences, termed index-hopping, was previously identified to account for 0.47% of reads, with samples with the fewest reads being affected the most [36]. As defined in the present study by using defined communities as references, spurious taxa do not necessarily represent true artifacts (*i.e.* sequences not corresponding to real microbes). Remnant DNA in laboratory materials and reagents [37] or in the feed used for laboratory animals (including germfree models) do originate from existing microbes and may give rise to amplification products that can confound results (especially when the number of PCR cycles are ≥30, as often used).

There is an obvious need for standardizing sequencing-based microbiome studies [38, 39]. Whilst several research groups looked at the impact of sequence quality filtering and sequencing depth [40-44], the influence of low abundant, potentially spurious taxa on readouts have been studied in less detail. Variations between replicate samples after Illumina-based amplicon sequencing can be quite high even after singletons removal [45] and the present study stresses the importance of benchmarking platforms using reference communities. Filtering strategies for removal of spurious taxa primarily affect diversity readouts, especially richness, implying that variations in richness that have very often been associated with disturbed microbial ecosystems under disease conditions should be interpreted more carefully [46, 47]. Richness estimates are strongly dependent on parameters set during bioinformatic analysis. Due to the influence of

sequencing depth on measured richness, it is usually normalized for comparison across samples, typically by using the minimum depth across all samples in the study. However, this does not help comparisons between studies, for which a standardized normalization would help. Technically, all OTUs must be counted for the estimation of richness, but the existence of spurious OTUs in sequencing data requires the implementation of appropriate cutoffs. Legacy has favoured the use of singletons removal before estimating richness. However, a singleton from a sample with 100 reads should obviously not be weighted the same as a singleton from another sample with 100,000 reads. That is why proportional filtering thresholds have been applied, albeit with marked variations between studies with little to no justification. We showed here that the majority of spurious OTUs can be effectively removed by applying a 0.25% relative abundance cutoff. Although by no mean universal, we recommend its usage over singleton removal prior to alpha- and beta-diversity analyses. Such filtering is simple to implement and already available in IMNGS (www.imngs.org).

Although study-specific filtering is effective in reducing the number of spurious OTUs and their impact on diversity measures, its outcome depends on the number of samples included in the study (as any OTU occurring at a relative abundance above the selected threshold in at least one sample is kept) and the depth of sequencing per sample. Due to this, *alpha*-diversity measures, such as richness, are especially sensitive to the normalization and filtering applied, making it difficult to compare richness across studies. A sample-specific measurement of *alpha*-diversity that takes into account the effect of sequencing depth and spurious taxa would be very useful for comparative analysis between studies. Therefore, we propose the notion of 'Effective Microbial Richness' (EMR), defined as the number of taxa with a relative abundance greater than a set cut-off (per default 0.25%) in each microbial profile considered. EMR is unaffected by sequencing depth or normalization steps and thus allows inter-studies comparisons. It is equivalent to the count of taxa after normalization to 1,000 reads and removal of those occurring below 2.5 counts. EMR together with other established *alpha*-diversity measures such as Shannon effective counts are implemented in Rhea (https://lagkouvardos.github.io/Rhea).

Conclusions

Thresholds for filtering low abundant taxa can markedly influence the outcome of microbiota analysis by 16S rRNA gene amplicon sequencing, especially diversity readouts. We strongly recommend applying filtering strategies that go beyond singletons removal. 'Effective Microbial Richness' will help the comparison of *alpha*-diversity across studies.

Declarations

Acknowledgments

We are grateful to Caroline Ziegler and Angela Sachsenhauser from the Microbiome Core Facility of the ZIEL Institute for Food and Health for outstanding technical support with amplicon sequencing. We also thank Jasmin Schwarz, Gudrun Kohl, and Petra Pjevac from the JMF for sample processing and discussions.

Funding

TC was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): grant no. CL481/2-1 within Priority Program SPP1656 and Project-ID 403224013 – SFB 1382 ‚Gut-liver axis'. DH received funding from the DFG, Project no. 395357507 – SFB 1371 ‚Microbiome signatures'. Both TC and DH coordinated the DINAMIC project funded by the JPI-HDHL Microbiomics program. IL received funding from the Hellenic Foundation for Research and Innovation (HFRI). The JMF is funded by the Medical University of Vienna and the University of Vienna (Vienna, Austria).

Availability of data and materials

The 16S rRNA gene amplicon datasets generated in the present study are available in the European Nucleotide Archive (www.ebi.ac.uk/ena) under study accession number PRJEB34431 (data from the Microbiome Core Facility of ZIEL) and SRA accession numbers SRR10688001-37 (data from the JMF).

Authors' contributions

IL and TC developed the study concept and design. SR, TCAH, NF, and BH carried out the experiments and data analyses. ART, KN, DB, DH, IL, and TC provided guidance and access to materials and resources. DH and TC secured main funding. SR, TCAH, BH, and IL drafted text sections. TC coordinated the project and wrote the manuscript. All authors critically reviewed the manuscript and approved its final version.

Ethics approval and consent to participate

Sequencing datasets from mice either originated from a published study with an already approved protocol (Gnoto1 data) [48] or were derived from newly generated samples obtained during experiments approved by the Government of Upper Bavaria (Gnoto3 and 4; approval no. 55.2-1-54-2532-156-13/138-14) or by the Institutional Animal Care and Use Committee at the University of Nebraska-Lincoln (Gnoto2; Protocols 1215 and 1301).

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 2009; 19:1141-1152.
2. Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD, Dore J. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. Appl Environ Microbiol 1999; 65:4799-4807.
3. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI *et al*. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010; 7:335-336.
4. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, Clavel T. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci Rep 2016; 6:33721.
5. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010; 26:2460-2461.
6. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016; 13:581-583.
7. Avershina E, Rudi K. Confusion about the species richness of human gut microbiota. Benef Microbes 2015; 6:657-659.
8. Clavel T, Lagkouvardos I, Hiergeist A. Microbiome sequencing: challenges and opportunities for molecular medicine. Expert Rev Mol Diagn 2016; 16:795-805.
9. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL *et al*. The long-term stability of the human gut microbiota. Science 2013; 341:1237439.
10. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010; 464:59-65.
11. Halfvarson J, Brislawn CJ, Lamendella R, Vazquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A *et al*. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2017; 2:17004.
12. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vazquez-Baeza Y, Gonzalez A, Knight R *et al*. Temporal variability is a personalized feature of the human microbiome. Genome Biol 2014; 15:531.
13. Sircana A, Framarin L, Leone N, Berrutti M, Castellino F, Parente R, De Michieli F, Paschetta E, Musso G. Altered Gut Microbiota in Type 2 Diabetes: Just a Coincidence? Curr Diab Rep 2018; 18:98.
14. Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl Environ Microbiol 2011; 77:7846-7849.
15. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 2013; 41:e1.
16. Herbold CW, Pelikan C, Kuzyk O, Hausmann B, Angel R, Berry D, Loy A. A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. Front Microbiol 2015; 6:731.
17. Hausmann B, Knorr KH, Schreck K, Tringe SG, Glavina Del Rio T, Loy A, Pester M. Consortia of low-abundance bacteria drive sulfate reduction-dependent degradation of fermentation products in peat soil microcosms. ISME J 2016; 10:2365-2375.
18. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 2016; 18:1403-1414.

19. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013; 10:996-998.
20. Auer L, Mariadassou M, O'Donohue M, Klopp C, Hernandez-Raquet G. Analysis of large 16S rRNA Illumina data sets: Impact of singleton read filtering on microbial community description. Mol Ecol Resour 2017; 17:e122-e132.
21. Lagkouvardos I, Fischer S, Kumar N, Clavel T. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ 2017; 5:e2836.
22. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009; 26:1641-1650.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-410.
24. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glockner FO. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 2014; 42:D643-648.
25. Callahan B, Sankaran K, Fukuyama J, McMurdie P. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. F1000Research 2016; 5:1492.
26. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 2012; 28:2106-2113.
27. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005; 71:8228-8235.
28. Jost L. Partitioning diversity into independent alpha and beta components. Ecology 2007; 88:2427-2439.
29. Andersen K, Kirkegaard R, Karst S, Albertsen M. ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. bioRxiv 2019; doi.org/10.1101/299537.
30. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013; 30:772-780.
31. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A 2007; 104:13780-13785.
32. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med 2014; 6:237ra265.
33. de Goffau MC, Lager S, Sovio U, Gaccioli F, Cook E, Peacock SJ, Parkhill J, Charnock-Jones DS, Smith GCS. Human placenta has no microbiome but can contain potential pathogens. Nature 2019; 572:329-334.
34. Hornef M, Penders J. Does a prenatal bacterial microbiota exist? Mucosal Immunol 2017; 10:598-601.
35. Perez-Munoz ME, Arrieta MC, Ramer-Tait AE, Walter J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. Microbiome 2017; 5:48.
36. van der Valk T, Vezzi F, Ormestad M, Dalen L, Guschanski K. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol Ecol Resour 2019.
37. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014; 12:87.
38. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung FE *et al*. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 2017; 35:1069-1076.
39. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D *et al*. Best practices for analysing microbiomes. Nat Rev Microbiol 2018; 16:410-422.
40. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 2013; 10:57-59.
41. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M *et al*. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 2012; 6:1621-1624.
42. Lundin D, Severin I, Logue JB, Ostman O, Andersson AF, Lindstrom ES. Which sequencing depth is sufficient to describe patterns in bacterial alpha- and beta-diversity? Environ Microbiol Rep 2012; 4:367-372.
43. Ni J, Li X, He Z, Xu M. A novel method to determine the minimum number of sequences required for reliable microbial community analysis. J Microbiol Methods 2017; 139:196-201.

44. Xiao F, Yu Y, Li J, Juneau P, Yan Q. Necessary Sequencing Depth and Clustering Method to Obtain Relatively Stable Diversity Patterns in Studying Fish Gut Microbiota. Curr Microbiol 2018; 75:1240-1246.
45. Wen C, Wu L, Qin Y, Van Nostrand JD, Ning D, Sun B, Xue K, Liu F, Deng Y, Liang Y *et al*. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. PLoS One 2017; 12:e0176716.
46. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P *et al*. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006; 55:205-211.
47. Sze MA, Schloss PD. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. MBio 2016; 7.
48. Lengfelder I, Sava IG, Hansen JJ, Kleigrewe K, Herzog J, Neuhaus K, Hofmann T, Sartor RB, Haller D. Complex Bacterial Consortia Reprogram the Colitogenic Activity of Enterococcus faecalis in a Gnotobiotic Mouse Model of Chronic, Immune-Mediated Colitis. Front Immunol 2019; 10:1420.
49. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res 2017; 45:D408-D414.
50. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 2015; 43:e37.
51. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 2013; 79:5112-5120.
52. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res 2017; 45:e23.

Tables

Table 1 Mock communities used in the present study

Name	Seq. facility[b]	Gene region	Replicates	No. species	No. raw reads	No. sequences after processing	Total no. taxa[d] (no filtering)	1st spurious taxon[d] (% rel. abundance)	Reference
Mock-1	See ref.	V4	1	27	153,841	140,397	432	0.007	[49]
Mock-2	See ref.	V4	1	58	593,868	578,569	761	0.439	[50]
Mock-3	See ref.	V4	1	21	1,012,097	453,215	1,081	0.031	[51]
Mock-4	See ref.	V4	1	14	169,516	159,352	417	0.020	[52]
Mock-5	See ref.	V4	1	21	613,091	108,414	802	0.439	[51]
Mock-6	See ref.	V4	1	21	602,819	231,685	732	0.160	[51]
Mock-7	See ref.	V4	1	20	306,773	42,746	95	0.008	[40]
Mock-TUM	1	V3/4	7	13	25,640 ± 8,516	19,882 ± 8,163	77 ± 15	0.130 ± 0.138	This study
ZymoBIOMICS (cat. #D6300)[a]	1	V3/4	7	8[c]	67,465 ± 31,752	52,079 ± 24,759	177 ± 33	0.059 ± 0.026	This study
ZymoBIOMICS (cat. #D6311)[a]	2	V3/4 and V4	25	8[c]	16,093 ± 9,319	14,383 ± 9,083	17 ± 15	0.106 ± 0.085	This study

In case of replicates, data are shown as mean ± sd
The sequences and taxonomies of all species included in the respective Mocks are provided in the Supplementary online information

[a] Whilst D6300 corresponds to an evenly distributed mixture of the microbes, D6311 a log-distributed mixture of DNA from the same microbes.
[b] For studies from the literature, please refer to the corresponding listed reference. For in-house generated data in this study: 1, ZIEL Core Facility Microbiome, TU Munich, Freising, Germany; 2, Joint Microbiome Facility of the Medical

University of Vienna and the University of Vienna (JMF), Austria.

[c] Bacterial species only. The Mock community includes also two yeast species not considered in the present study (10 microbial species in total).

[d] All values refer to Operational Taxonomic Units (OTUs) clustered at 97% sequence identity, with exception of the last Mock community analyzed in sequencing facility 2, for which values refer to Amplicon Sequence Variants (ASVs).

Table 2 Gnotobiotic mouse communities used in the present study

	Gene region	Replicates	No. species	No. raw reads	No. sequences after processing	Total no. OTUs (no filtering)	1st spurious OUT (% rel. abundance)	Reference
GNOTO1	V3/V4	6	7	28,706 ± 4,904	25,869 ± 4,494	66 ± 4	0.101 ± 0.014	This study
GNOTO2	V3/V4	9	12	30,261 ± 11,434	21,148 ± 8,313	172 ± 24	0.009 ± 0.004	This study
GNOTO3	V3/V4	6	6	30,444 ± 42,325	27,632 ± 3,563	85 ± 10	0.116 ± 0.016	This study
GNOTO4	V3/V4	7	4	25,217 ± 6,514	47,505 ± 6,106	68 ± 10	0.249 ± 0.041	This study

In case of replicates, data are shown as mean ± sd

The sequences and taxonomies of species included in the respective Mocks are provided in the Supplementary online information

Figures

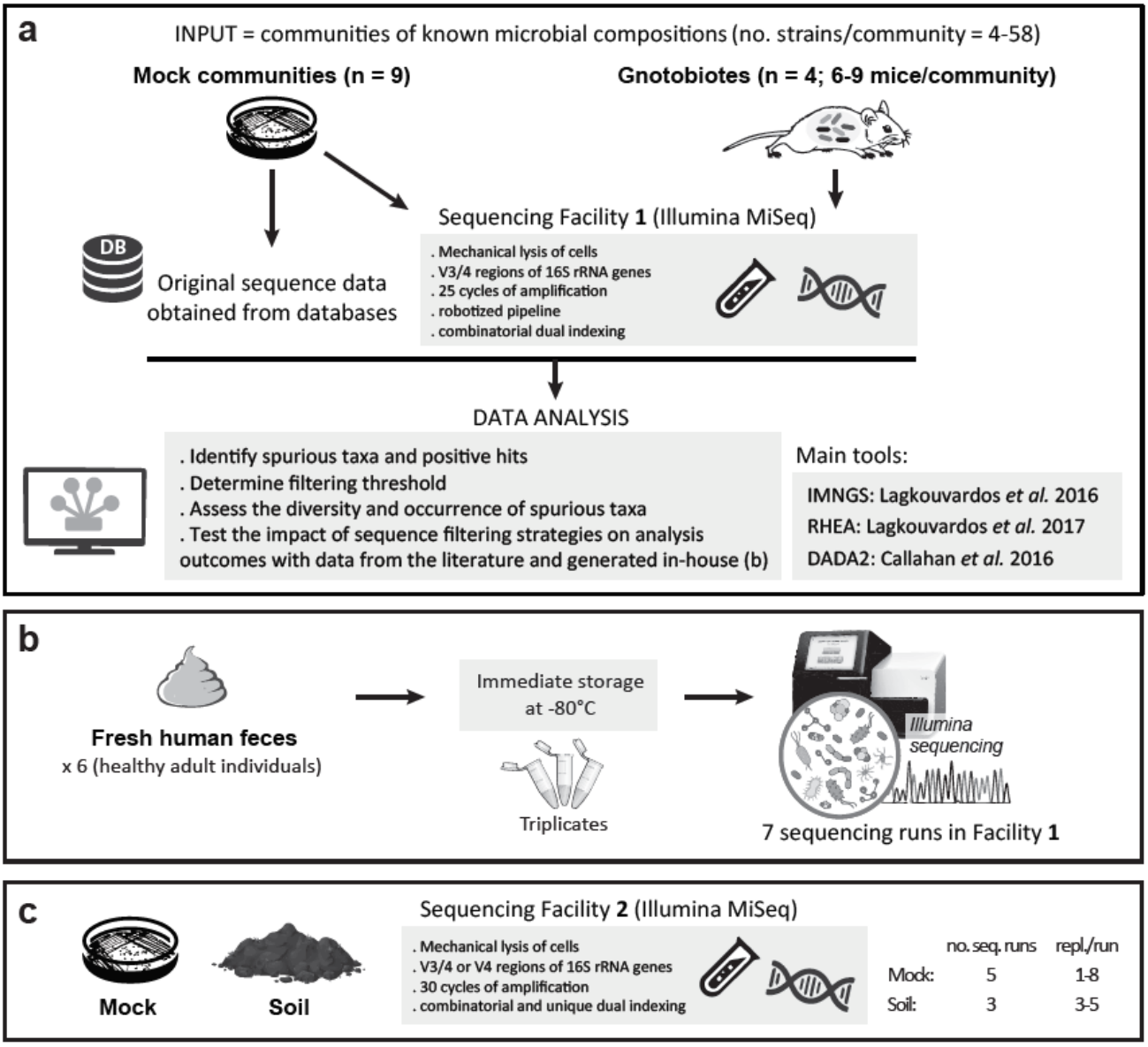

Figure 1

chematic overview of the work. a The use of reference communities of microbes in vitro and in vivo using data from the literature or generated in-house and analyzed using different bioinformatic pipelines allowed precise analysis of the occurrence of spurious taxa. b Several human fecal samples stored under different conditions and processed in triplicates in different sequencing runs allowed assessing the reproducibility of microbiota profiles generated by high-throughput 16S rRNA gene amplicon sequencing following different filtering thresholds to remove spurious taxa. c Sequencing of a Mock community and a soil sample, including several replicates and sequencing runs, followed by data analysis in a different facility were performed to validate findings. All technical details are given in the Methods section.

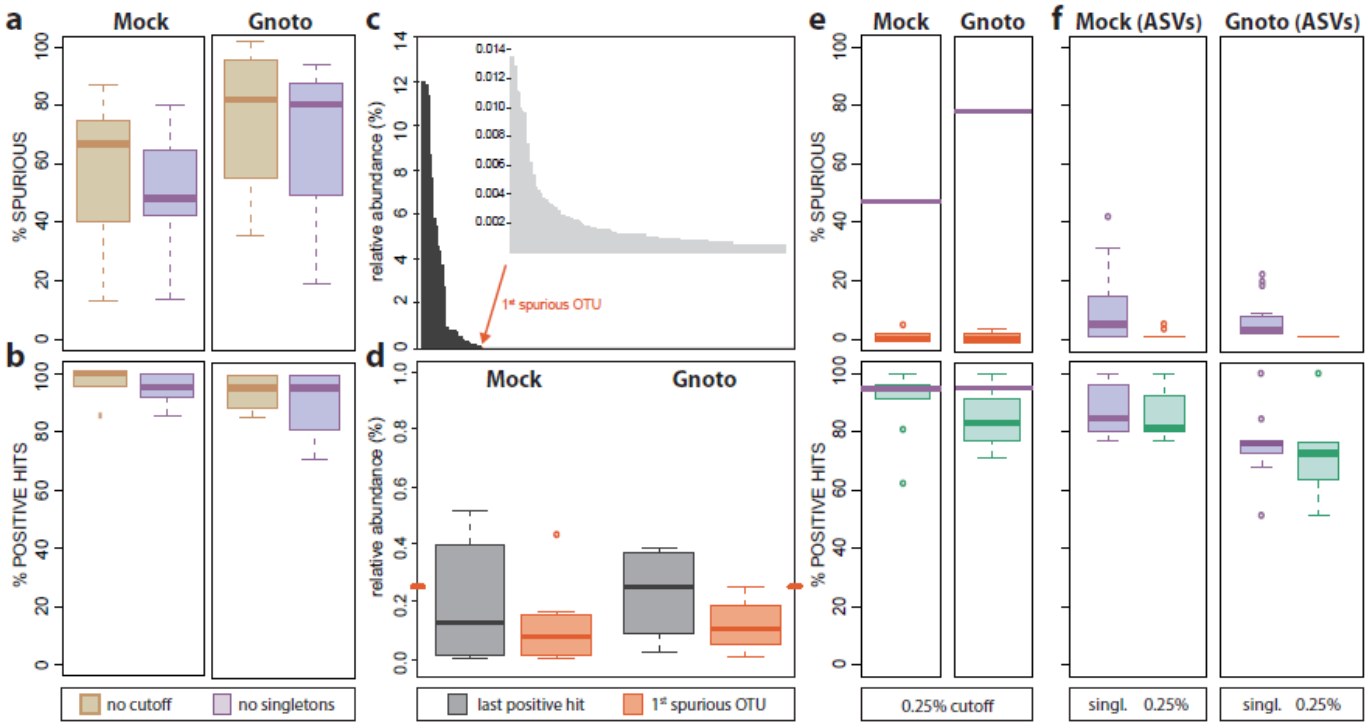

Figure 2

Determination of filtering thresholds using artificial communities of known composition in vitro (mock; n = 9 different types; 21 replicates in total) and in mice (gnotobiotes; n = 4 different communities; 28 mice in total). a Comparison of various standard filtering cutoffs (see explanations in the text) regarding the percentage of spurious OTUs (i.e. those molecular species not matching sequences of the known species contained in the artificial communities). b Corresponding percentages of positive hits retained by the different filtering strategies, positive hits being defined as the reference sequences found in the respective amplicon datasets. c Example of the relative abundance distribution of total OTUs detected without filtering in the gut of a gnotobiotic mouse [48]. The arrow indicates the position of the first spurious OTUs, all following OTUs being considered as having a high risk of being spurious (light grey bars). d Determination of a representative filtering threshold following the strategy shown in panel c using all available mock communities and samples from gnotobiotes. The orange lines on the y-axis indicate the consensus threshold of 0.25% relative abundance. e Percentage of spurious OTUs and positive hits for the determined filtering cutoff of 0.25% in mock and mouse communities. f Percentage of spurious OTUs and positive hits in the same reference communities using different filtering thresholds and the DADA2 pipeline for analysis based on Amplicon Sequence Variants (ASVs) [6].

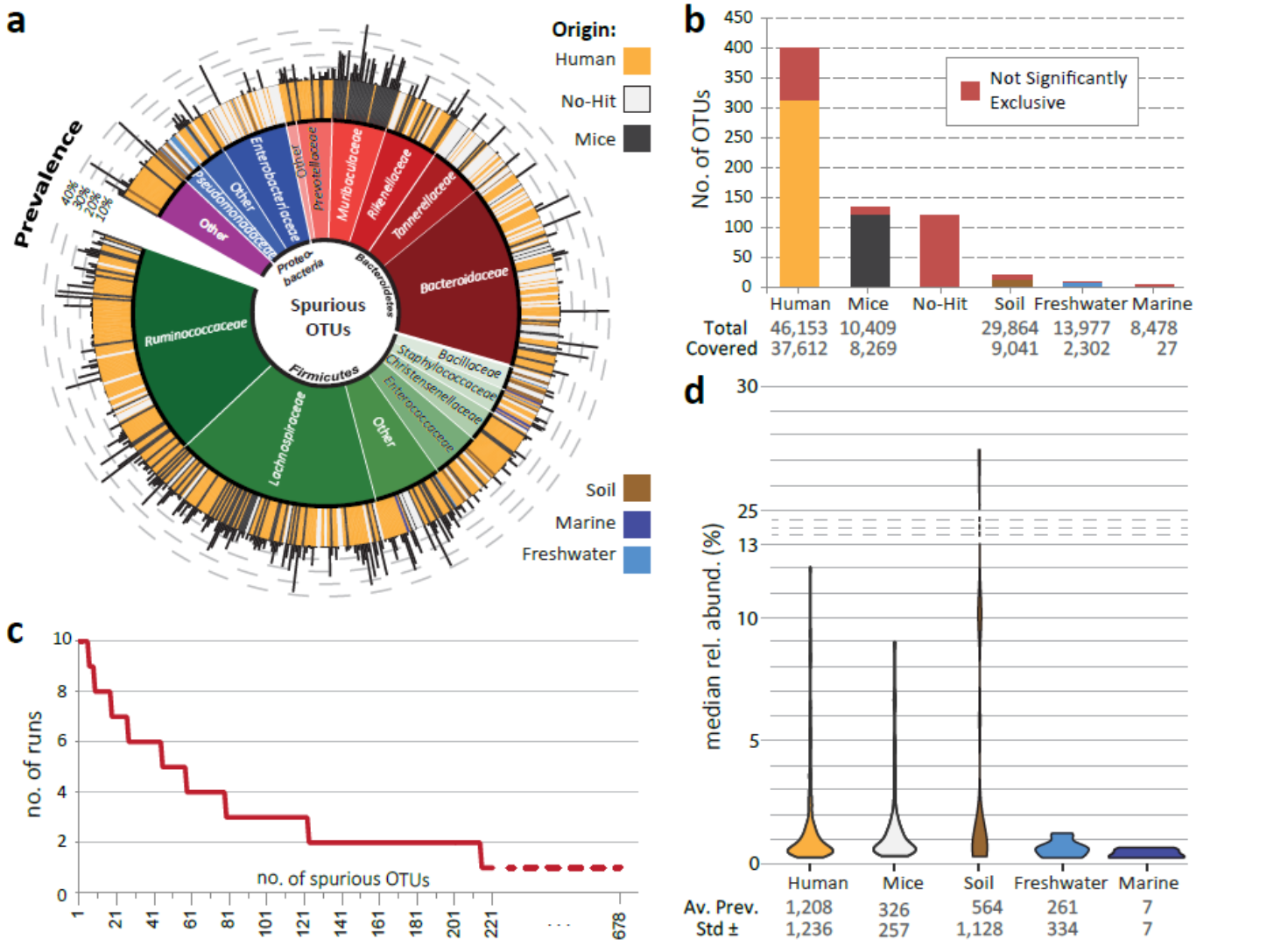

Figure 3

Origin and occurrence of spurious OTUs. a Taxonomic profile and ecological distribution. Inner ring: SILVA-based classification of all non-redundant spurious OTUs at the phylum and family level. Outer coloured ring: sample type characterized by the highest prevalence for the given OTU. Outer bars: corresponding highest prevalence values. Only samples with relative abundances >0.25% for any given OTU were counted as positive for prevalence calculation. The total numbers of samples considered were: human, 46,153; soil, 29,864; freshwater, 13,977; mouse, 10,409; marine, 8,478. b Distribution of the spurious OTUs across sample types. The exclusivity of each OTU for any given sample type was assessed using a Z-test: those OTUs with none-significant specificity of sample type appear in red ($p<0.05$). The total number of IMNGS samples considered for each sample type with at least one of any spurious OTUs matching sequences above 0.25% relative abundance was labelled as “Total” (equal numbers in panel a). The number of samples in each type covered by at least one spurious OTU with highest prevalence in this sample type was labelled as “Covered” (i.e. the remaining samples in that category contained also at least one spurious OTU, which was however characterized by highest prevalence in another sample type). c Redundancy of the spurious OTUs across 10 sequencing runs. d Violin plots of the distribution of median relative abundances of all OTUs within the given sample types as shown in panel b. The average prevalence of the spurious OTUs in each sample category is shown as mean ± sd below the x-axis.

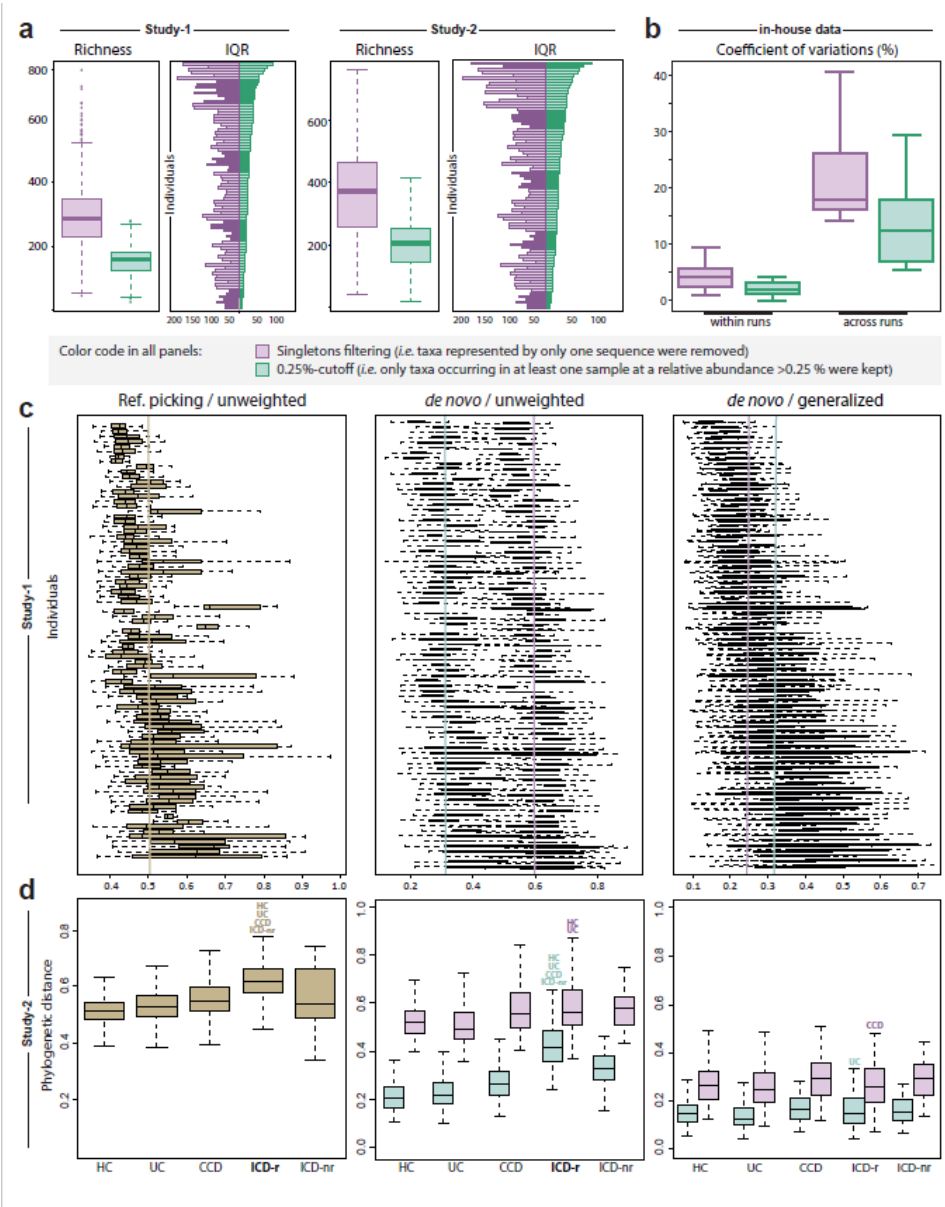

Figure 4

Impact of sequence filtering methods on study outcomes: literature data. Throughout the entire figure, the color code is as follows: (i) singleton removal, pale violet; (ii) 0.25%-cutoff filtering (i.e. keeping only those OTUs occurring in at least one samples at a relative abundance >0.25%), light green; (iii) closed-reference picking, light brown. a The box plots show richness distribution across all individual samples and time points. The bar plots show inter-quartile ranges (IQR = Q3-Q1) of individual samples (rows) as a proxy for richness variation across the various time points of a given sample. IQRs were ranked by decreasing values after applying the 0.25%-cutoff. b Coefficient of variations calculated on richness values obtained from six fecal samples each sequenced in triplicates in seven different sequencing runs. Box plots on the left indicate variations across triplicates within any given run for both filtering methods. Box plots on the right indicate variations between the same samples across runs. c Overtime variations in microbiota profiles for each individual from Study-1 [12] based on reference OTU-picking and unweighted UniFrac distances (left; as in the published study), de novo OTU-picking and unweighted UniFrac distances (middle) or generalized UniFrac distances (right). Bars indicate median distances across all individuals. Individuals were ordered by increasing average distance using the 0.25%-cutoff and generalized UniFrac (right panel). d Differences in the phylogenetic makeup of fecal microbiota as in panel c for Study-2 [11].

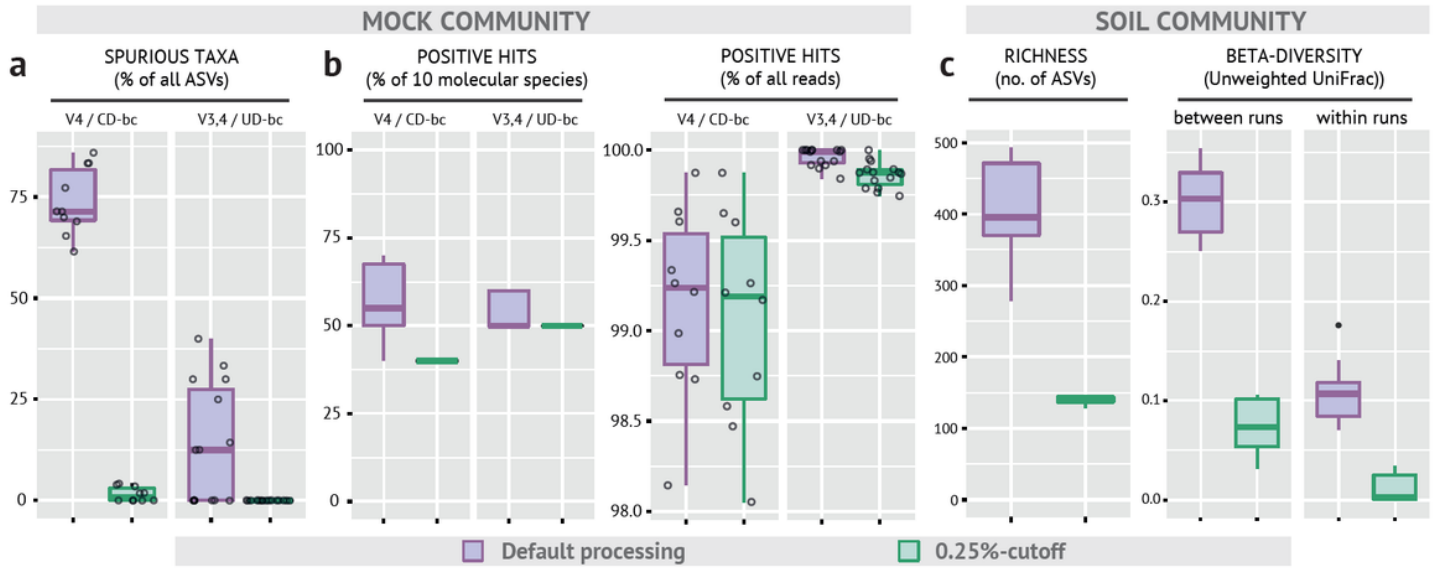

Figure 5

Validations studies in a second, independent sequencing facility. a Fraction of spurious taxa with (green) or without (violet) applying the 0.25%-cutoff displayed according to the targeted 16S rRNA gene regions and barcoding strategy used. b Corresponding fraction of positive hits, i.e., amplicons matching the reference strains contained in the Mock community. c Average and distribution of richness and distance values between replicates of the same soil sample processed in multiple sequencing runs. Abbreviations: ASV, amplicon sequence variant; bc, barcoding; CD, combinatorial dual; no., number; UD, unique dual.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1SequencesTaxonomiesMockcommunitiesR1.xlsx
- TableS2SequencesTaxonomiesGnotobiotesR1.xlsx