frontiers in Microbiology

REVIEW
published: 04 September 2017
doi: 10.3389/fmicb.2017.01561

Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing

Luisa W. Hugerth [1,2]* *and Anders F. Andersson* [2]*

[1] Department of Molecular, Tumour and Cell Biology, Centre for Translational Microbiome Research, Karolinska Institutet, Solna, Sweden, [2] Division of Gene Technology, Science for Life Laboratory, School of Biotechnology, KTH Royal Institute of Technology, Solna, Sweden

OPEN ACCESS

Edited by:
Steve Lindemann,
Purdue University, United States

Reviewed by:
Julien Tremblay,
National Research Council Canada (NRC-CNRC), Canada
Joseph Henry Vineis,
Marine Biological Laboratory, United States

***Correspondence:**
Luisa W. Hugerth
luisa.hugerth@scilifelab.se
Anders F. Andersson
anders.andersson@scilifelab.se

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: 16 April 2017
Accepted: 02 August 2017
Published: 04 September 2017

Citation:
Hugerth LW and Andersson AF (2017) Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Front. Microbiol. 8:1561. doi: 10.3389/fmicb.2017.01561

Microbial ecology as a scientific field is fundamentally driven by technological advance. The past decade's revolution in DNA sequencing cost and throughput has made it possible for most research groups to map microbial community composition in environments of interest. However, the computational and statistical methodology required to analyse this kind of data is often not part of the biologist training. In this review, we give a historical perspective on the use of sequencing data in microbial ecology and restate the current need for this method; but also highlight the major caveats with standard practices for handling these data, from sample collection and library preparation to statistical analysis. Further, we outline the main new analytical tools that have been developed in the past few years to bypass these caveats, as well as highlight the major requirements of common statistical practices and the extent to which they are applicable to microbial data. Besides delving into the meaning of select alpha- and beta-diversity measures, we give special consideration to techniques for finding the main drivers of community dissimilarity and for interaction network construction. While every project design has specific needs, this review should serve as a starting point for considering what options are available.

Keywords: bioinformatics, biostatistics, amplicon sequencing, microbiome, NGS, 16S rRNA, microbial ecology

A BRIEF HISTORY OF METHODOLOGIES FOR DETERMINING MICROBIAL COMMUNITY COMPOSITION

While humans have been selectively breeding bacteria and fungi for food fermentation for several centuries, the first observations of microbial organisms were made in the 1670's by Antony van Leeuwenhoek, who first observed microbes (that he called "animalcules") in saliva, and the first purposeful and successful isolation of bacteria for scientific purposes was attained by Robert Koch and Julius Petri in the 1870's. Both direct observation and culturing remain invaluable techniques to this day, albeit both have limitations.

Culturing is the gold standard for microbial characterization, as it provides large amounts of cells from a clonal population, and allows any number of functional tests on bacterial biochemistry, physiology and genetics to be performed. It was however evident even to Koch that different bacteria grow best in different settings, and by the early 1900's it was accepted that the vast majority of bacteria could not be cultivated with standard techniques, a phenomenon later dubbed "the great plate count anomaly" (Staley and Konopka, 1985). Therefore, most of what is known today about bacterial physiology stems from a very small subset of easily culturable bacteria of medical or veterinary importance which grow well in the presence of high nutrient loads (Lagier et al., 2015).

Reasons for refraction to culturing are many. Firstly, in the absence of knowledge of the specific growth requirements of an organism, trial-and-error is not a feasible way to determine them (Stewart, 2012), especially considering that many organisms have rather narrow windows of growth (Lagier et al., 2015). These microbes might survive in the environment in boom-and-bust cycles (Iluz et al., 2009; Gilbert et al., 2012) or grow at a pace so slow as to be nearly indistinguishable in the lab (Zengler et al., 2002; Lagier et al., 2015).

Laboratory settings can also generate toxic conditions, such as oxidative stress (Morris et al., 2011; Tanaka et al., 2014). In addition to this, organisms might fail to grow due to missing pathways (Nye et al., 1999), or be dependent on siderophores produced by other members of their community (D'Onofrio et al., 2010).

Even today, despite high-throughput dilution-to-extinction culturing techniques (Aakra et al., 1999; Rappé et al., 2002; Aoi et al., 2009; Liu et al., 2009), culture chambers that mimic natural environments (Zengler et al., 2002; Nichols et al., 2010; Sizova et al., 2012) and co-culturing approaches (Kaeberlein et al., 2002; Tanaka et al., 2004; Morris et al., 2008), isolating and culturing bacteria is a complex and time-consuming endeavor.

An alternative to culturing is to perform microscopy directly on environmental samples. High-resolution microscopy techniques such as electron microscopy, confocal microscopy and photoswitchable fluorophores allow a number of specific biological questions to be addressed directly from images of live or fixated bacteria (reviewed in Coltharp and Xiao, 2012). However, regardless of technology, with observation alone it can be extremely hard to achieve reasonable functional or taxonomic resolution for the diversity of microbes typically found in an environmental sample. It takes years of training as a taxonomist to excel in the visual identification of microbes, even ones with as much morphological diversity as protists; and even then there are strong observer effects (reviewed in Moreira and López-García, 2002; Silva, 2008).

To move beyond the difficulties of culturing and the limitations of microscopy, microbial ecologists have moved increasingly toward molecular fingerprinting. Starting in 1977, Woese and colleagues established the suitability of the small subunit (SSU) of the ribosomal RNA (rRNA) gene for inferring phylogenetic relationships between prokaryotic organisms, a property later verified to also apply to eukaryotes (Woese and Fox, 1977; Woese et al., 1985; Woese, 1987). Norman Pace and colleagues soon started applying the same technique to natural communities (Pace et al., 1985; Stahl et al., 1985). Together with the ribosomal internal transcribed spacer (ITS), this is still the most commonly used gene for community phylogenetic composition analysis (community fingerprinting). The advantages of using SSU rRNA for community fingerprinting are many: (i) This gene is found in all cellular life forms. (ii) It is a highly conserved gene, serving to a large degree as a reliable molecular chronometer. (iii) It is seldom transferred horizontally. (iv) It possesses both conserved and variable regions, so that the conserved regions can be targeted by polymerase chain reaction (PCR) primers and the variable ones be used as identifying markers. A handful of other genes, such as the large subunit (LSU) rRNA share these properties, but the length of ∼1,500 bp of the bacterial SSU rRNA made it amenable to early molecular techniques, and the impressive body of knowledge that has since accumulated on the basis of this gene makes a switch to other markers very impractical, except in certain sub-fields such as mycology, where ITS and LSU are widely used.

The 1990s saw the first high-throughput environmental fingerprinting approaches, sometimes referred to as microbiomics. It is the decade of techniques such as denaturing gradient gel electrophoresis (DGGE) (Muyzer et al., 1993), terminal restriction fragment length polymorphism (T-RFLP) (Liu et al., 1997) and automated ribosomal intergenic space analysis (ARISA) (Fisher and Triplett, 1999), all of which are based on the characteristic travel distance of PCR amplified DNA fragments (amplicons) in an electrophoretic device. These banding patterns can be used directly to compare broad changes in taxonomic composition of samples in different conditions. Even though, in each of these techniques, different organisms might give rise to identical bands, each band is treated as an operational taxonomic unit (OTU). To assign a tentative taxonomy to the OTU, high abundance bands can be selected for sequencing.

At around the same time, microarrays emerged as an alternative to fingerprinting. A downside of microarrays is that identification is restricted to sequences previously known and printed onto the array (Ehrenreich, 2006). While this limits its applications as a general environmental survey tool, microarrays can still be valuable tools in focused clinical, industrial or environmental monitoring settings (Humbert et al., 2010; Ricke et al., 2013; Zumla et al., 2014). Since microarrays can cover various regions of the genome, they can be used for distinguishing between closely related species or strains (Lehner et al., 2005; Singh and Mohapatra, 2008; Narihiro and Sekiguchi, 2011).

The rise of high-throughput DNA sequencing was a game changer for microbial ecology. In 2006, the first study was published using 454 pyrosequencing for assessing microbial communities, a survey of the microbial diversity in a marine water community (Sogin et al., 2006). This study, while sequencing relatively shallowly (6,505–22,994 sequences/sample), already presented two of the main characteristics of sequencing-based microbiomics that came to be seen as standards in the field: rarefaction curves very far from reaching saturation, which indicated a much larger microbial diversity than previously suspected; and a highly uneven community, with 3–4 orders of magnitude of difference in abundance between the least and most abundant tags. These previously unknown low abundance organisms were dubbed in the paper the "rare biosphere", a term still in use and whose biological relevance is much discussed (Lynch and Neufeld, 2015). Later, the introduction of sample-specific barcode sequences (Andersson et al., 2008; Hamady et al., 2008) allowed sequencing many samples in the same run and opened up the door to large-scale comparative microbiome studies.

SAMPLE COLLECTION, STORAGE, DNA EXTRACTION, LIBRARY PREPARATION AND SEQUENCING

While high-throughput amplicon-sequencing has proven powerful and accurate for microbial community analysis, random and systematic errors can be introduced in several of the steps along the analysis chain. Sampling itself can introduce biases, which has to be considered when collecting or analysing any sort of ecological data. Solid samples such as soil can have extreme short-distance heterogeneity (Certini et al., 2004). The amount of material used for extraction, and the definition of the sample (e.g., whether they're homogenized in bulk or kept separately) has to be suited to the research question at hand. As for aquatic samples, long term studies must contend with the issue of the flowing and mixing of water masses. A stationary sampling, fixed to geographical coordinates, faces the issue that changes observed in the microbial community can be a true change within a community or a replacement of one community by another as the water flows. As an alternative to the stationary eulerian sampling, it is possible to follow a water mass using a buoy and collect samples around it, a strategy termed Lagrangian sampling. This approach, however, is only effective for a few weeks, after which the water mass is mixed beyond the point where it can be considered coherent with the initial sample. The temporal dimension is crucial regardless of the sampling strategy, since the frequency of sampling should be (but often isn't) commensurate with the rate of the biological processes of interest.

Sample storage is also extremely important, to prevent bacterial overgrowth as well as taxonomically biased DNA damage or degradation. Bacteria should be inactivated as soon as possible without causing significant damage to their DNA (Choo et al., 2015; Song et al., 2016). While most laboratories correctly choose to keep their samples in the freezer, it is also necessary to consider the damage done by repeatedly freezing and thawing cells (Moré et al., 1994; Harju et al., 2004) and DNA (Thomson et al., 2010; Todorova et al., 2012). While short amplicons can usually still be amplified even from fragmented DNA, long amplicons, metagenomic libraries, and cDNA libraries have stricter requirements. Researchers should assess their study design and laboratory capacity and consider the possibility of storing samples in an appropriate preservation medium (Roberts, 2016).

The next source of bias and artifacts is the DNA extraction method. Extraction relates to sampling, since different methods require and tolerate different amounts of starting material. The physico-chemical characteristics of the environment and of the biological material in it will in turn interact with the extraction method, producing a more or less efficient disruption of cell walls and membranes and removal of contaminants. A failure to appropriately disrupt certain types of cell wall will cause those organisms to be underestimated in the community profile. A failure to remove contaminants such as other biomolecules and organic acids will inhibit the DNA amplification step, leading to amplification biases and eventually even sample loss (Weiss et al., 2014; Gorzelak et al., 2015; Reck et al., 2015; Walker et al., 2015). Finally, for samples of low microbial density, such as patient blood samples, minute amounts of DNA or cellular contamination in any reagent or piece of equipment used in extraction will generate spurious reads (Salter et al., 2014). **Figure 1** illustrates biases at each step of the sampling procedure.

It must also be noted that RNA can also be extracted and analyzed as complementary DNA (cDNA). DNA is a more stable molecule, so community signatures are less likely to experience radical change at the DNA level as a result of sample collection (Lim et al., 2014; Reck et al., 2015). On the other hand, different organisms, specially eukaryotes, can have an enormous range of copies of the rRNA gene in their genomes, which hinders a simple correlation between gene copies and cell numbers (Gong et al., 2013). The number of rRNA copies per cell, however, is largely independent of the number of gene copies in prokaryotic cells and is instead correlated to cell activity (Jones and Lennon, 2010). Activity, as measured by the ratio of 16S rRNA:16S rRNA gene copies, has in turn been shown to often be highest in low-abundance populations within a community (Campbell et al., 2011; Zhang et al., 2014). These differences can result in very different community profiles for cDNA and DNA analyses, for example in deep water layers, which at the DNA level are more affected by sinking dead and otherwise inactive cells (Zhang et al., 2014; Cram et al., 2015). When comparing cDNA and DNA profiles from spatially heterogenous biomes (such as soil), it may be important to do both types of analysis on the same actual samples (Roume et al., 2013).

After nucleic acid extraction (and cDNA synthesis when applicable), the region of interest must be amplified and prepared for sequencing. This is almost always achieved through PCR, a method which is sturdy and cost-effective, but may introduce large biases to the sample (Schirmer et al., 2015). PCR depends on primers, short DNA molecules (usually 15–30 bp) of defined sequence that bind to the ends of the DNA target region on the template strands and allow a DNA polymerase to synthesize a new DNA strand, complementary to the template and downstream of the primer's 3′ end. By flanking the region of interest with two primers, its copy number is doubled at every polymerization cycle, hence the term "polymerase chain reaction". This means that a DNA template which does not present complementarity to the primers will not be amplified and its corresponding organism will be a false negative in the microbiome profile. By applying a mixture of primers with base-level variations (degenerate primers), the percentage of taxa being targeted by the primers can be increased. On the other hand, this increases the odds of amplifying other DNA regions, creating artificial diversity. Furthermore, although mismatches in the primer sequence are the main cause of amplification bias, preferential binding of sequences containing C/G rather than A/T at a degenerate position is also a strong factor

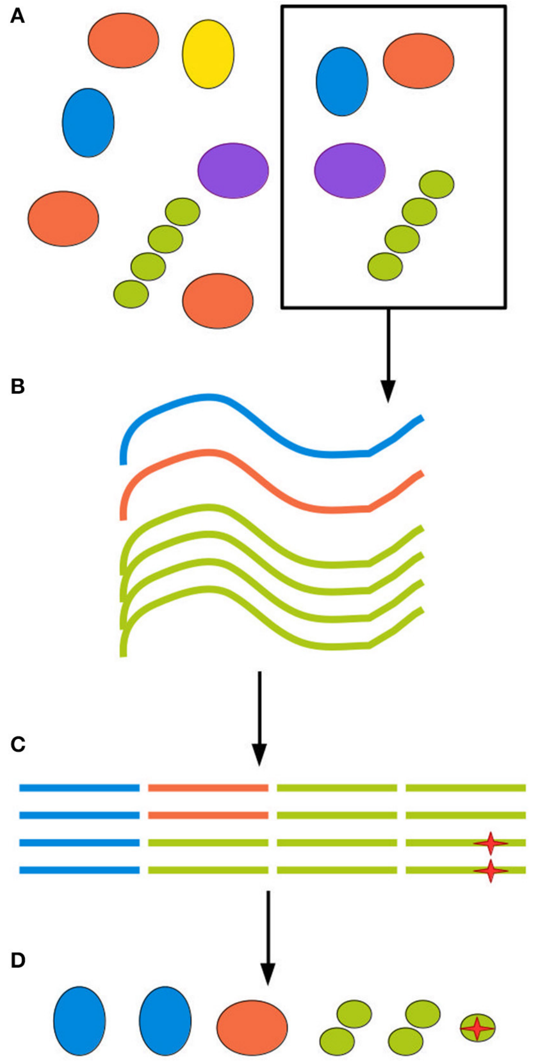

FIGURE 1 | All experimental steps can contribute to the total bias. **(A)** When retrieving an environmental sample, not all clades present in the environment will be present in the sample, and those who are present might be present in different proportions than in the environment as a whole. **(B)** DNA extraction methods inevitably have different efficiency rates for different clades, further distorting the sample. **(C)** Finally, amplification and sequencing both introduce biases and point mutations, represented here by a red star, as well as possible chimerism. **(D)** The final inferred community can be quite different from the ground truth.

(Lanzén et al., 2011). Therefore, the exact sequence of the PCR primer should be considered in terms of the community at hand and the acceptability of different biases in the resulting amplicon pool. The choice of primer is also very sensitive since the same community amplified with different high-quality primer pairs will still give a different profile, since different lineages may evolve at different rates in each variable region of their marker gene. Thus, studies focusing on different subregions aren't directly comparable (Nossa et al., 2010; Soergel et al., 2012; Yang et al., 2016) and some studies even choose to analyse different variable regions in parallel (Smith et al., 2012).

Several papers have been published that systematically assess the ability of primer sequences to amplify 16S (Baker et al., 2003; Wang and Qian, 2009; Youssef et al., 2009; Gantner et al., 2010; Nossa et al., 2010; Kumar et al., 2011; Soergel et al., 2012; Klindworth et al., 2013), 18S (Amaral-Zettler et al., 2009; Stoeck et al., 2010; Hugerth et al., 2014a), fungal ITS (Martin and Rygiewicz, 2005; Manter and Vivanco, 2007; Toju et al., 2012; Op De Beeck et al., 2014) and many other genes. PrimerProspector (Walters et al., 2011) and DegePrime (Hugerth et al., 2014b) are examples of computer programs that can aid in the design of broad taxonomic primers. DegePrime finds the primers with maximal taxonomic coverage while controlling the degeneracy to a user-specified level.

Another major issue with PCR is the formation of chimeras, that is, DNA molecules containing partial sequences from two or more biological sequences. This arises due to sequence similarity between amplicons and the nature of the chain reaction. Partially amplified molecules might break and anneal unspecifically to templates from other lineages, serving as primers for new, chimeric sequences. Factors that have been linked to increased frequency of chimeras include both laboratory settings, such as fast thermocycling during PCR (Stevens et al., 2013), and intrinsic characteristics of the sample, such as richness and diversity (Fonseca et al., 2012).

As more and more research groups started using short-read high-throughput gene tag sequencing, first with 454 pyrosequencing and later with Illumina and Ion Torrent technologies, it also became increasingly clear that these methods, while less biased than some of their predecessors, do produce a considerable number of artifacts, which can be very hard to detect and separate from true biological signal. As an example of this, by using the same filtering strategy as the seminal work of Sogin et al. (2006), Quince et al. (2011) observed c. three times more 97%-clustered OTU than were present in the mock community used. Three main kinds of errors can arise from PCR amplification and sequencing: substitutions (a base is read in place of another), insertions (a base is read more times than were actually present) and deletions (a base is skipped). Each sequencing platform has its characteristic error profiles and assorted suite of tools for handling them, which have been described elsewhere (Quince et al., 2009; Gilles et al., 2011; Carneiro et al., 2012; Bragg et al., 2013; Laver et al., 2015; Schirmer et al., 2015). In addition to this, reads from one sample might be assigned to a different one, due to contamination or sample-switching (exchange of index during library preparation or sequencing). This issue has been estimated to affect up to 2% of reads in certain datasets, and is hard to control for (Edgar, 2016a).

In addition to sampling and library preparation, the choice of sequencing platform has to be suitable for the environment and research questions of interest. Longer reads can increase the accuracy of phylogenetic placement (Okubo et al., 2012; Quick et al., 2015), while a larger number of reads might be needed to reduce the effect of random noise and increases the sensitivity of the approach. In addition to the short read technologies which are the focus of this work, long read approaches such as PacBio (Schloss et al., 2016) and Oxford Nanopore MinION are increasingly in use (Benítez-Páez et al., 2016; Lindberg et al., 2016; Hu et al., 2017).

OTU CLUSTERING AND TAXONOMIC ANNOTATION

The initial sequencing data processing step is filtering based on read quality scores and the presence of the expected primer and adapter sequences, and the removal of these non-biological sequences. Commonly used tools for this task are Fastx (Gordon and Hannon, 2009), TrimGalore (Krueger, 2017) and Cutadapt (Martin, 2012). Low quality bases, adapters, primer dimers, reads that are too short and obvious contaminants (e.g., human DNA) need to be removed. Then, for paired-end reads, it is common to merge them at their overlapping regions. Since read quality also falls toward the end of the read for most short-read technologies (Salipante et al., 2014; Schirmer et al., 2015), this is also a way to increase the confidence in the bases in this region (Salipante et al., 2014). Good stand-alone tools for this are Usearch (Edgar, 2013) and FLASH (Magoč and Salzberg, 2011), and it can also be achieved using MOTHUR or Qiime (Schloss et al., 2009; Caporaso et al., 2010). Single-end short reads should be trimmed to the same length for comparability, and reads shorter than the cutoff, discarded (Edgar, 2013). This approach can also be used for paired-end reads covering a region too long for merging; in this case, they can be trimmed to a fixed length and concatenated (Hugerth et al., 2014a). FastQC (Andrews, 2009) or MultiQC (Ewels et al., 2016) can be used to assess the quality of the data before and after quality filtering, including whether errors are randomly distributed or clustered at certain bases or regions of the flow cell.

The next step is usually to "pick OTU". In the case of sequencing, OTU are most often defined by clustering sequences according to similarity. This step is meant to eliminate erroneous sequences formed by PCR and sequencing errors, since these should deviate from a true sequence by only a few bases. This way, a sequence diversity is reduced to true biological diversity. Since small variations in sequence are observed among strains of a single species, and even among different operons of a single strain, it is assumed that tags differing by only a small percentage of their bases represent taxonomically equivalent cells. This is not always true, however, as in the well documented case of Escherichia *spp. a*nd S*higella spp.,* which despite having clearly distinct natural histories harbor the exact same sequence along the full length of their 16S rRNA gene (Zuo et al., 2013).

OTU picking procedures can be divided into closed reference, open reference and *de novo*. Closed reference means mapping reads to a database and assigning them to the best possible match. Reads that do not match with a sufficiently high score are discarded. On an open reference approach, those sequences that fail to match to the reference are submitted to a *de novo a*pproach.

De novo approaches, in their turn, can broadly be divided into hierarchical clustering (based on single, average or complete linkage) (Schloss and Handelsman, 2005) and heuristic strategies. In single-linkage, a sequence is placed in a cluster if it has a similarity above a threshold to at least one other sequence in the cluster. This procedure tends to form very large clusters with a lot of heterogeneity and is rarely used, except occasionally for very rapidly evolving genes such as the fungal ITS region (Lindahl et al., 2013). Complete linkage, on the other hand, requires that a sequence in a cluster has similarity above the threshold to all others. This method produces therefore much more and smaller OTU than the other, and tends to overestimate measures of community richness, especially for data with many errors. For average-linkage, finally, the average similarity between a sequence and all others in the same cluster has to be above the threshold.

Since it is computationally very demanding to run an all-against-all comparison on datasets of millions of reads, as is done in hierarchical clustering, heuristic approaches were developed, the most relevant of which being the Usearch/Uparse suite (Edgar, 2010). It approximates complete- or average-linkage approaches by only comparing each sequence to a "centroid" sequence within each cluster. By selecting a distance cutoff between this centroid sequence and the candidate sequences, an average-linkage clustering is approximated. A recent benchmarking found that average-linkage clustering produced the most meaningful and stable OTU, followed by the distance-based greedy clustering implemented in Usearch (Westcott and Schloss, 2015). A consequence of this is that the order in which sequences are handled affects the final result. Therefore, sequences are generally sorted by decreasing abundance before clustering, since abundant sequences are less likely to be artifacts. From the description of these methods, it is clear that the distribution of distances between sequences in clusters will differ depending on the approach used, although the same nominal similarity cutoff is applied, a fact that is often glossed over when discussing microbiomics.

Hierarchical clustering was first made widely available to the ecology community through the software DOTUR (Schloss and Handelsman, 2005), but can now be found in many implementations, most prominently its successor MOTHUR (Schloss et al., 2009) and the CD-HIT package (Fu et al., 2012). As for heuristic strategies, popular software packages for OTU picking are Usearch (Edgar, 2013), Qiime (Caporaso et al., 2010), which runs Uclust in the background, and Vsearch (Westcott and Schloss, 2015), an open-source alternative to Usearch. However, these and other approaches suffer from OTU instability, that is, the fact that the same sequence might be assigned to different OTU depending on the community context (He et al., 2015; Schmidt et al., 2015). Most OTU-picking pipelines include a chimera removal step, either by comparing sequences to a known database, or d*e novo* by flagging low abundance sequences that could be formed by a combination of two sequences of higher abundance within the dataset (Schloss et al., 2009; Caporaso et al., 2010; Edgar, 2010).

Very often, clusters are selected at 97% similarity (Gevers et al., 2005). At lower similarity levels, sequences are unlikely to be derived from the same species, and isolates are unlikely to display 70% DNA-DNA hybridization (a previously common heuristic for determining bacterial species assignment; Stackebrandt and Goebel, 1994; Gevers et al., 2005). However, 97% similarity over the full length of the ∼1,500 bp gene doesn't translate directly to 97% similarity over any given region of the gene (Schloss, 2010). Further, as discussed above, a 3% distance doesn't mean exactly the same thing across all packages. Finally, the 97% similarity cutoff is to a large degree arbitrary, since different taxa might have much less of a distance between their tags and still represent ecologically distinct clades (Fox et al., 1992; Gevers et al., 2005). In the field of eukaryotic microbiomics, higher degrees of similarity are often used (Not et al., 2009; Stoeck et al., 2010). This stems from an understanding of the different way taxonomy is applied to eukaryotes as compared to prokaryotes, ie clades with more morphological variety tend to be assigned a more fine-grained classification (Ciccarelli et al., 2006). Some degree of clustering could be deemed necessary when sequencing approaches had much lower throughput, and most reads were likely to be singletons or doubletons (OTU detected by only one or two reads, respectively), which would make statistical comparison between samples very difficult. This, however, is no longer the case. With increased sequence data quality, a 99% cut-off is increasingly common for bacteria as well. The appropriateness of any method is ultimately dependent on the research question being addressed, since OTU clustered at 97% similarity through different approaches have both been shown to recapitulate natural history well, when assessed from a global perspective, and to harbor extreme heterogeneity, when studied at a narrower scale (Koeppel and Wu, 2013; Schmidt et al., 2014). This issue is far from being resolved, as the very concept of species is the subject of much controversy between and within different branches of microbiology (Gevers et al., 2005). **Figure 2** illustrates the effect of different clustering approaches on raw sequencing data.

In an attempt to advance the methodological aspect of OTU picking, several approaches have been recently published which attempt to produce biologically meaningful OTU independently of a predefined level of similarity. Each of them has a different strategy to separate the noise introduced by PCR and sequencing from true biological diversity.

DADA2 (Divisive Amplicon Denoising Algorithm 2) initially divides amplicons by considering their abundance distribution (since common reads are more likely to be true sequences) and sequence distance from other reads (since errors are expected to occur at most a few times per read). Then, it uses the clusters generated and the quality scores of bases, produced by the sequencing platform, to calculate a substitution error model conditioned on quality scores for the sequencing run at hand. Finally, it uses this error model to "correct" reads, that is, assign low frequency reads to higher frequency reads from which they could with high probability be derived by substitution (Callahan et al., 2016).

Another strategy, Cluster-Free Filtering (CFF), uses a simpler but conceptually similar approach to figure out which are the true biological sequences in the dataset. Rather than correcting the other sequences, it removes them from the analysis. In addition, the CFF pipeline uses patterns of covariation across samples to infer whether correlated sequence types are most likely derived from different subpopulations or from different operons (or remaining sequencing errors) of the same subpopulation (Tikhonov et al., 2015). The idea in this case is that while different operon copies or erroneous sequences of the same subpopulation will have similar dynamics irrespective of condition, different subpopulations will react differently to different stimuli and their correlation will change between conditions.

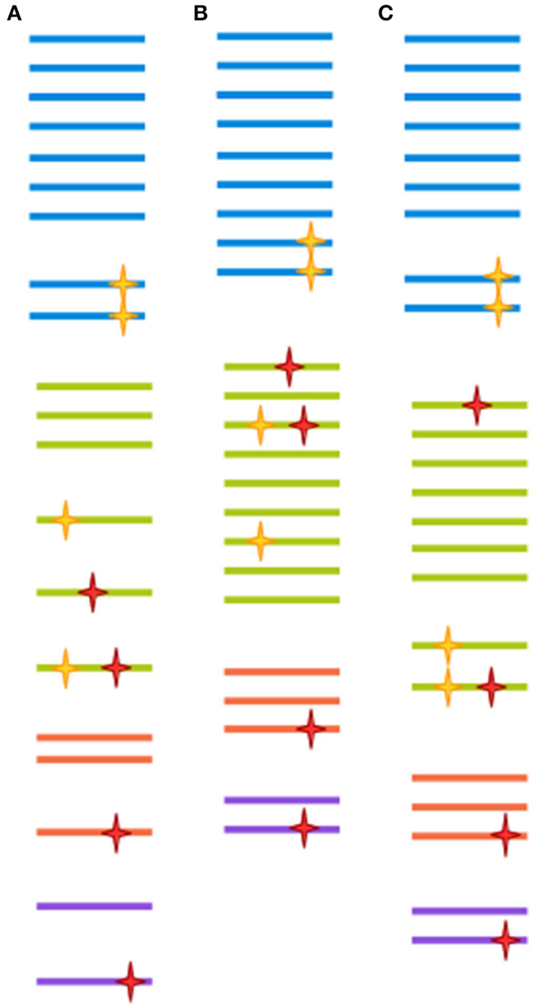

FIGURE 2 | Each clustering approach distorts the OTU composition differently. A hypothetical sample is clustered in three different ways. Each color represents a different clade. Yellow stars represent biological point mutations, and red stars are errors created during amplification and sequencing. The spacing between sequences indicates the inferred clusters. **(A)** Clustering at 100%-identity treats sequences bearing single mismatches as separate OTUs. **(B)** Clustering at 97% identity will remove the effect of amplification/sequencing errors, but will also cluster together sequence variants that represent different clades. **(C)** Modern techniques such as DADA2, cluster-free filtering and minimum entropy decomposition can preserve true diversity while eliminating most spurious mutations. Notice, however, that if mutations are sufficiently abundant (for instance, if they are generated at an early PCR cycle), they might still give rise to a spurious OTU.

Unoise is another denoising algorithm optimized for speed (Edgar, 2016b). It starts by eliminating low-abundance OTU (by default, <4 counts) and then fits an error model in which the larger the distance (substitutions and deletions) between a low-abundance unique sequence and a higher abundance one is, the larger their difference in relative abundance should be, exponentially. Due to this large dependence in abundances, Unoise is best used on full datasets, not individual samples. However, it can become prohibitively slow in large enough datasets, so that a pre-filtering step removing unique sequences with an abundance below, e.g., 10^{-6} can decrease the processing time from days to hours, depending on the evenness of the data. Pre-filtered reads can later be mapped back to the approved centroid sequences to prevent loss of quantitative information. Since DADA2 and Unoise have very different approaches to denoising, they can complement each other and have indeed been shown to produce best results when used in parallel, so that only sequences considered correct by both approaches are accepted as true (Edgar, 2016b).

Minimum-Entropy Decomposition is a denoising method based on the Shannon entropy of each position in an alignment of all sequences. Positions that are peaks in entropy trigger the algorithm to split sequences into smaller clusters of lower entropy. The procedure continues until no cluster exists with a significant entropy peak and a minimum number of sequence reads. Empirically, this approach has been shown to reveal community dynamics that would have been obfuscated by 97% OTU clustering (Eren et al., 2014).

After the OTUs in a study are determined, it is crucial to assign a taxonomic classification to them. This allows the OTUs' dynamics across samples to be interpreted in light of what is known about these taxa from previous studies and, more broadly, allows comparison across microbiomics studies. Unfortunately, there is also no consensus in the microbiology community about how to assign taxonomy to OTU tags. Certain workflows, such as QIIME (Caporaso et al., 2010) and MOTHUR (Schloss et al., 2009), include the classification step. Other softwares are dedicated exclusively to it. For instance, the Ribosomal Database Project (RDP) Classifier uses a naïve bayesian approach to classify sequences based on exact matches of 8-letter words and performs bootstrapping to give probability estimates of the correctness of the assignment (Wang et al., 2007). Another popular approach to sequence classification is the Silva Incremental Aligner (SINA) (Pruesse et al., 2012). SINA uses an initial k-mer based search similar to that of the RDP classifier, but then uses the subset of the reference sequences matched best by the k-mer search to construct a tree representing all unique selected sequences and calculates an exact alignment between the query sequence and the reference candidates. Finally, the sequence taxonomy is assigned as the least common ancestor of the top-scoring alignments.

It is also worth noticing that a classification is only as good as the underlying database. The RDP is maintained by the Center for Molecular ecology at the University of Michigan and it is updated periodically. It includes over three million bacterial and archaeal 16S rRNA genes as well as several thousand fungal 28S and two separate sets of fungal ITS sequences. Another frequently updated database is SILVA, maintained by the Microbial Genomics and Bioinformatics Group at MPI Bremen. It currently contains over five million SSU rRNA gene sequences (16S and 18S) and more than 700,000 LSU sequences (23S and 28S). It also presents subsets of these data including only full length sequences and non-redundant full length sequences. Both the SILVA and the RDP team rely heavily on the work of the Bergey's Manual for taxonomy (Goodfellow et al., 2012) and collaborate with the Bergey Trust. In addition to these, the Greengenes database (DeSantis et al., 2006) combines the NCBI taxonomy with the cyanoDB (Komárek and Hauer, 2013) for a high quality tree of life based on 16S rRNA genes, and it is the default choice for various tools, including the Qiime package and function predictors such as PICRUSt and Tax4Fun (Langille et al., 2013; Aßhauer et al., 2015); however, it hasn't had a new release since 2013, missing the bacterial lineages identified since then. Stand-alone versions of the RDP and SINA classifiers allow the construction of manually curated, personalized databases appropriate to the environment of interest, so the user is by no means limited to the one tool, one database paradigm. A good taxonomic database can also be used to remove spurious sequences, so chloroplast, mitochondria, host rRNA etc., should be included as sanity checks.

Standard approaches generally perform much more poorly for eukaryotes than prokaryotes, due both to more incomplete databases and to a more elaborate taxonomy. Therefore, databases and placement strategies for eukaryotic microbes are still being developed (Lanzén et al., 2012; Guillou et al., 2013; Hu et al., 2015). For well-studied environments of limited diversity, placing OTU directly over a phylogenetic tree is a good strategy for assigning last common ancestor taxonomy to OTU of interest, but this approach is computationally demanding and doesn't scale well for large datasets with high taxonomic diversity (Matsen et al., 2010).

As an alternative to defining and taxonomically classifying OTU *de novo* they can be mapped to reference sequences that have been clustered beforehand (i.e., closed- or open-reference clustering). Examples of such datasets are available in the Greengenes (DeSantis et al., 2006) and Silva (Pruesse et al., 2007) databases. Support for this type of analysis is provided within Qiime, or it can be performed by any blast-like mapping tool, such as Usearch. This approach works well for microbiomes that are well represented in rRNA databases, and makes it easy to add more samples to a comparative study without needing to redo the OTU generation from scratch. For more unexplored environments many reads may lack a close relative and a *de novo* OTU approach is preferable.

Any combination of methods and algorithms chosen to profile community microbiomes have their own intrinsic and unavoidable biases. Even the taxonomy levels considered in each database differ (Balvočiūtė and Huson, 2017), which is a major source of variability between pipelines (Siegwald et al., 2017). Which method produces the results closest to the underlying community is difficult to assess and depends on the specific community under study, but being aware of the biases produced by each method is crucial both for method selection and for data interpretation and comparison across studies.

UNEQUAL SAMPLE SIZES AND DATA NORMALIZATION

Multisample microbiomics data is generally summarized as a table of read counts per OTU per sample. These tables are often very sparse, especially for communities with a long tail of OTU belonging to the rare biosphere. The interpretation of counts of 0 is not straightforward, since they may represent the true absence of an OTU or its presence under the detection limit. Moreover, due to e.g., differences in yields between sequencing runs, and unequal representation of samples in pooled sequencing libraries, detection limits will vary between samples.

Due to this lack of clarity on the method's limit of detection, arbitrary approaches are often adopted. Both MOTHUR and Qiime (Schloss et al., 2009; Caporaso et al., 2010) include built-in functions to discard OTU with less than a given number of independent observations, or proportion of reads, or present in fewer than a given number of samples. However, where to set these cutoffs is far from obvious, as it depends on the environment of interest, the sampling scheme and the specific research question. Lundberg et al. (2012) provide in their Supplementary Material an interesting example on how to define these thresholds for a given study. It is also important to consider that eliminating rare OTU or including artifacts may have very large effects on alpha-diversity estimates (discussed below) (McCoy and Matsen, 2013).

This unequal sampling depth makes it often necessary to conduct some kind of normalization. One of the most popular approaches is to divide the counts by the total count of the sample. Doing this breaks the independence of observations, since an increase in the relative abundance of one OTU induces a perceived reduction in all others. A similar approach is using not the total count of reads for normalization, but a fixed percentile of them (Bullard et al., 2010), which should be less sensitive to events such as blooms.

As an alternative to calculating relative abundances, some authors perform random down-sampling of every sample to the smallest sample size of the cohort. This procedure is a recommended approach for comparing alpha-diversity between samples (Lundin et al., 2012), but downsampling also entails data waste and loss of statistical power for downstream analysis (McMurdie and Holmes, 2014). McMurdie and Holmes also demonstrate that normalizing each sample to 1 doesn't control for overdispersion and, in breaking data independence, increases the rate of false positives. Instead, these authors propose using available packages for mRNA and marker gene sequencing, such as DESeq, edgeR, and metagenomeSeq (Robinson et al., 2010; Paulson et al., 2013; Love et al., 2014). These packages model the dispersion of the data on appropriate models, thereby minimizing both the rate of false negatives and of false positives. Ideally, actual counts can be obtained by multiplying relative abundances by total cell counts, measured by e.g., flow cytometry or microscopy. But even if accurate counts of rRNA gene fragments could be obtained, the variable number of copies of these genes per genome across the tree of life means that this still wouldn't correspond to exact cell counts. Some tools, such as the RDP classifier, can take this into account, at least for well-known lineages. Any choice of normalization will affect how data is interpreted downstream (**Figure 3**; R code used for generating **Figures 3–6**, **8** are provided in Supplementary File 1).

ESTIMATING DIVERSITY WITHIN A SAMPLE (ALPHA-DIVERSITY)

The term "alpha-diversity" was first defined by Robert Whittaker in 1960 as "The richness in species of a particular stand or community, or a given stratum or group of organisms in a stand" (Whittaker, 1960). In microbial ecology, alpha-diversity is generally understood as the diversity within a single sample or set of replicates. The most naïve way to measure this is observed richness, that is, simply counting how many different OTU are in a sample. However, it is typically impossible to identify every single taxon in a microbial sample, which requires the use of techniques that take into account the incompleteness of the inventory. These can be borrowed from the field of macrobial ecology (Hughes et al., 2001).

One way to estimate the true richness of a sample is to take into account the tail of the species (or OTU) abundance distribution, more specifically the number of singletons (species observed once) and doubletons (species observed twice). This is done by the Chao1 estimator, defined as:

$$S_{est} = S_{obs} + \frac{f_1^2}{2f_2}$$

Where S_{est} is the estimated species richness, S_{obs} is the observed species richness, f_1 is the number of singletons and f_2 is the number of doubletons. Related to Chao1 is ACE (abundance-based coverage estimator), which considers the ratio not only of singletons and doubletons, but of all OTU observed up to an arbitrary count, most usually 10:

$$\begin{aligned}
S_{ace} &= f_{abund} + \frac{f_{rare}}{C_{ace}} + \gamma_{ace}^2 \frac{f_1}{C_{ace}} \\
C_{ace} &= 1 - \frac{f_1}{n_{rare}} \\
\gamma_{ace}^2 &= \max\left[0, \frac{f_{rare}\sum_{i=1}^{10} i\,(i-1)\,f_i}{C_{ace} n_{rare}\,(n_{rare}-1)} - 1\right]
\end{aligned}$$

Where f_{abund} is the number of OTU above the abundance threshold, f_{rare} is the number of OTU at or below the threshold, f_i is the number of OTU observed i times, n_{rare} is the total number of individuals in rare OTU, C_{ace} is a sample coverage estimator and γ_{ace}^2 is the estimated coefficient of variation for rare OTU.

In addition to the number of species in a sample, an important measure of diversity is how even their distribution is. Intuitively, a sample where 10 different OTU each compose 10% of the cells is more diverse than one where one OTU takes up 91% of the sample and the others, 1% each. The Simpson index is a way to quantify this (Simpson, 1949), and it corresponds to the odds that two individual microbes sampled at random will belong to the same OTU.

$$\lambda = \Sigma p_i^2$$

Where λ is the simpson index and p_i is the relative abundance of each OTU i.

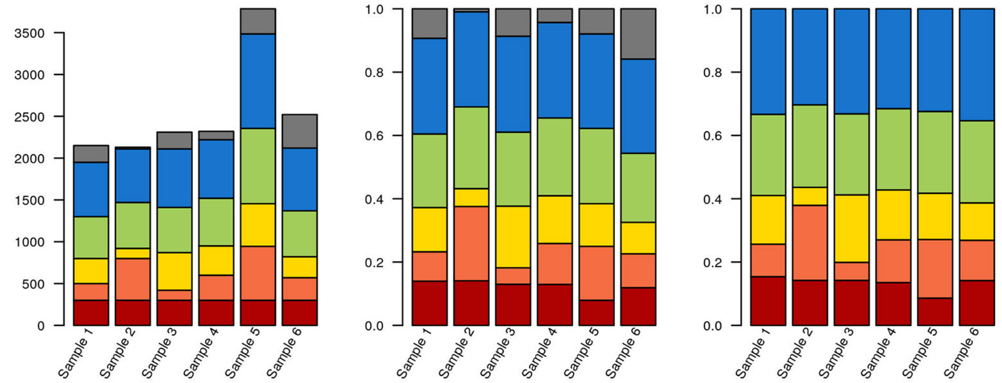

FIGURE 3 | Data selection and normalization affects data representation. The same hypothetical data from a sequencing experiment is depicted in all three panels. Each of the brighter colors represents a clade, and gray corresponds to unclassified sequences. In the leftmost panel, raw data is depicted. The red clade has identical quantities in each sample (300 reads). In the middle panel, the data has been normalized to unity. The blue clade now has the same proportion of reads in each sample (30%). Finally, in the third panel, where only classified sequences are depicted, the green clade has the same proportion of reads in each sample (25%). Also notice that, due to the stacked nature of the bar plots, it isn't necessarily obvious that the green or blue blocks are identical in their respective panels.

Another index that measures combined species richness and evenness is Shannon's diversity index, or Shannon entropy. Although originally intended for calculating entropy (uncertainty of information content) of strings of text (Shannon, 1948), Shannon entropy can easily be interpreted in ecology as the uncertainty involved in predicting the species of an individual sampled at random. Mathematically, it is defined as

$$H' = -\Sigma p_i \ln\left(p_i\right)$$

From the Shannon index, it is also possible to derive a measure of evenness, Pielou's evenness index, by dividing the observed value of the Shannon index by the highest possible value (that is, that which would be observed if all OTU were present in equal abundance; Pielou, 1966). Mathematically:

$$\begin{aligned} J' &= \frac{H'}{H'_{\max}} \\ H'_{\max} &= -\Sigma \frac{1}{S} \ln\left(\frac{1}{S}\right) = \ln\left(S\right) \end{aligned}$$

Where $H'_{\max}$ is the highest possible Shannon index for a sample with S number of OTU.

All of the metrics discussed above give equal weight to each OTU. This would give the same diversity values to a community composed of 10 species from a single genus as it would one composed of 10 different phyla. The phylogenetic diversity of a community can be considered by taking into account the sum of the branches of the phylogenetic tree that includes all OTU in the sample (Faith, 1992), which can also be weighted by the relative abundance of each clade in the sample (Cadotte et al., 2010). The R package Picante can be used to calculate Faith's phylogenetic diversity (Kembel et al., 2010). Building on Faith's original work, other authors have extended measures such as Simpson and Shannon into phylogenetically weighted equivalents (Warwick and Clarke, 1995; Allen et al., 2009). These were later generalized and shown to outperform standard measures at separating healthy from disease-associated human microbiome communities (McCoy and Matsen, 2013).

ESTIMATING COMMUNITY DISSIMILARITIES (BETA-DIVERSITY)

"Beta-diversity," as coined by Whittaker (1960), is "The extent of change of community composition, or degree of community differentiation, in relation to a complex gradient of environment, or a pattern of environments." In other words, beta-diversity is the degree to which two samples are different. This is a rather different issue than within-sample richness and evenness, and can be measured in many different ways.

The choice of beta-diversity metric can have important consequences to subsequent analyses, such as clustering and ordination. This is partially due to the interplay between distance metrics and normalization techniques, which can widen or reduce the apparent distance between samples (**Figure 3**).

A true distance metric is one that is always positive, in which the distance between a point and itself is 0, the distance between A and B is identical to the distance between B and A and the sum of the distance between A and B and between B and C is no greater than the distance between A and C. This last assumption is the one that often fails for other dissimilarity measures. The appropriate metric for a study might depend on the size of the effect of interest and on the depth of sampling.

The most widely known true distance metric is the euclidean:

$$d\left(S_1, S_2\right) = \sqrt{\sum \left(S_{1i} - S_{2i}\right)^2}$$

Where S_1 and S_2 are two samples and S_{1i}, S_{2i} are the abundance of OTU i in samples S_1 and S_2, respectively.

However, the euclidean distance requires very large effect sizes for statistical significance (Kuczynski et al., 2010) and doesn't perform well in datasets with many zeroes. A more appropriate metric is thus Jensen-Shannon's, a symmetric version of the Kullback-Leibler divergence. In Kullback-Leibler, the distance between S_1, S_2 is:

$$KL(S_1, S_2) = \sum S_{1i} \times ln\frac{S_{1i}}{S_{2i}}$$

Thus, Kullback-Leibler is not applicable for 0-rich datasets. However, since Jensen-Shannon's compares samples S_1 and S_2 to their average, the problem of 0's disappears:

$$JS\left(S_1, S_2\right) = \frac{1}{2} \times KL\left(S_1, \frac{S_1 + S_2}{2}\right) + \frac{1}{2} \times KL\left(S_2, \frac{S_1 + S_2}{2}\right)$$

This formulation also automatically satisfies the other requirements for a distance metric.

In microbial ecology it is also common to use correlation coefficients, such as Pearson's product moment (**Figure 4A**):

$$r\left(S_1, S_2\right) = \frac{\sum\left(S_{1i} - \bar{S_1}\right)\left(S_{2i} - \bar{S_2}\right)}{\sqrt{\sum\left(S_{1i} - \bar{S_1}\right)^2}\sqrt{\sum\left(S_{2i} - \bar{S_2}\right)^2}},$$

To minimize the influence of noise, other researchers prefer Spearman's rank correlation, which is identical to Pearson's except that instead of the measured values, their ranks are used (**Figure 4B**). Finally, Bray-Curtis dissimilarity, while not very sensitive, is appropriate for 0-inflated datasets (**Figure 4C**):

$$BC\left(S_1, S_2\right) = \frac{\sum\left|S_{1i} - S_{2i}\right|}{\sum\left(S_{1i} + S_{2i}\right)}$$

An alternative to OTU-based distances is to use phylogenetic distances. While these approaches also require several non-trivial choices, such as the underlying phylogenetic tree and the placement of OTU in it, evolutionary distances are often more biologically meaningful, not least because phylogenetic relatedness is often associated to trait conservation (Martiny et al., 2015). As is the case for OTU-based metrics, using a quantitative or qualitative approach to community comparison can lead to very different results (Lozupone et al., 2007). This can be ameliorated through an appropriate weighting procedure, such as generalized Unifrac (Chen et al., 2012). Recent work by Schmidt and colleagues does a thorough review of commonly used distance metrics and proposes new taxonomic and phylogenetic distances based on co-occurrence networks (Schmidt et al., 2016).

Different approaches to community dissimilarity, such as OTU-based vs. phylogenetic, may highlight different aspects of the community and its functioning. It can therefore be useful to combine these different analyses to gain deeper insight into the system under study.

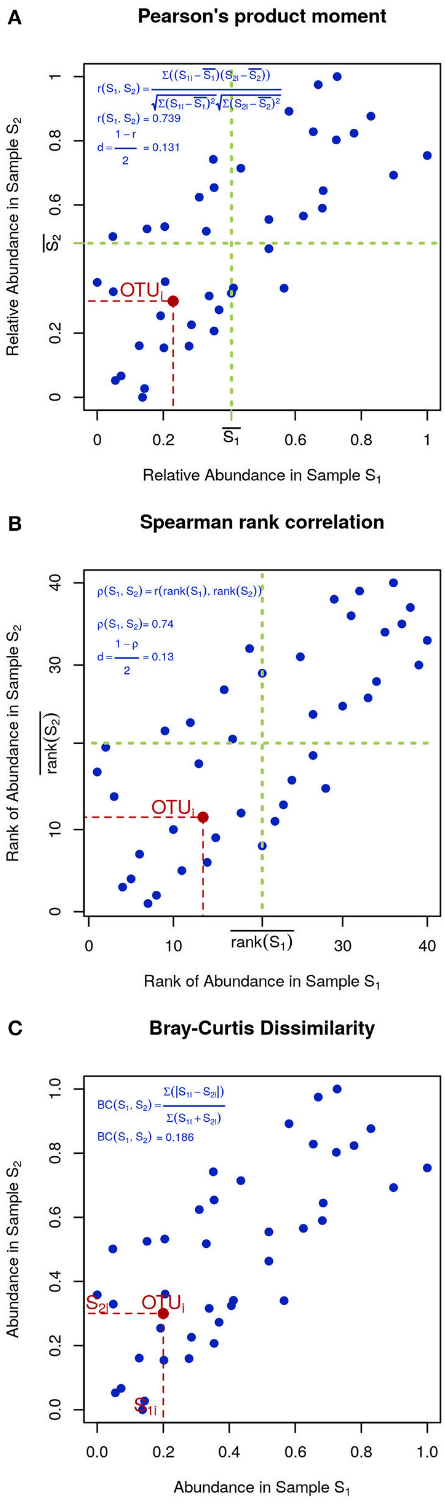

FIGURE 4 | Visual intuition to selected community dissimilarity metrics. In each panel, the same set of OTU (blue dots) is represented in a scatter plot from two highly correlated samples. Pearson's and Spearman's correlations can be intuitively thought as the degree to which the scatter deviates from a straight diagonal line, except Pearson is based on the numeric values of distances **(A)** and Spearman on their ranks **(B)**. Bray-Curtis dissimilarity is displayed in **(C)**.

VISUALIZING HIGH-DIMENSIONAL DATA

To be useful, a graphical representation of data must be more readily interpretable than the raw data. This is usually achieved by decreasing the level of detail in the data. A boxplot, for instance, contains much less information than a scatter-plot of the same data, but is often more easily apprehendable. In this specific case of unidimensional data with many data points, a balance between information-richness and interpretability can be achieved through the use of violin plots (**Figure 5**). In other cases, information is added to a plot, for instance through the use of color indicating data density or outlines highlighting groups of interest. Care needs to be taken then to not induce false conclusions. While the reader may be aware of which aspects of the graph aren't strictly informative, it is difficult to not be led, at least subconsciously, by these elements.

A simple visual inspection of the data is often the first step of any analysis. In microbiomics, this translates into plotting OTU abundance per sample. Since the number of OTU is generally incompatible with thorough plotting, they are often aggregated at higher taxonomic levels. It is usually recommendable to make these plots at various taxonomic levels, since community composition could, for instance, be stable at the phylum level but highly dynamic at the family level or, alternatively, highly stable for all but a few families which drive large phylum-level differences. A useful tool to avoid these constraints is Krona, which makes hierarchical interactive pie-charts representing several taxonomic levels at the same time (Ondov et al., 2011). However, when analysing a large number of samples, pie charts can make comparison across samples unintuitive, since there is no structured spatial organization of the data.

Stacked bar plots and line plots (stacked or not) are good alternatives for representing large numbers of samples. When comparing temporal series or data that is physically structured, line plots have the advantage of preserving the (temporal or geographical) distance between samples. On the other hand, the very existence of a line connecting points suggests that data changes smoothly across that interval, which may in fact be far from true. Barplots, on the other hand, preserve the discrete nature of data collection, but generally only display sample labels on the x-axis, making them less informative. Finally, authors must decide how to present the data in their bar plots, whether to normalize each sample to 1, or normalize the presented portion of the data (for instance, only OTU which have taxonomy at least at the domain level) to 1, for example (**Figure 3**). Each of these choices will highlight a different aspect of the data and must be clearly stated.

Since the human mind cannot process images in more than three dimensions, one of the main challenges in dealing with the vast number of OTU and/or samples in a microbiomics study is condensing the information into two- or three-dimensional spaces. A very good overview of techniques to achieve this was written by Paliy and Shankar (2016). One of the oldest and most common methods to achieve this is principal component analysis, or PCA (Ringnér, 2008). In it, variables are treated as axes in a euclidean multidimensional space and the first principal component is by definition placed on the direction representing the largest variation of the data. The second component is placed in the direction orthogonal to this that explains the largest amount of the remaining variation, and so on. The first few components often explain a large amount of the variation, allowing a visual inspection of the distance between samples in two- or three-dimensional space. Furthermore, the percentage of the variation explained by each axis indicates whether there are dominant drivers present or not. However, the euclidean distance is seldom appropriate for microbial data, since pairs of samples with many common counts of 0 are given a short distance. Bray-Curtis dissimilarity, in contrast, only considers OTU present in at least one of the samples being compared.

To avoid the constraint of the euclidean distance, principal coordinate analysis (PCoA) can be used together with any dissimilarity matrix, such as Bray-Curtis or UniFrac, which are

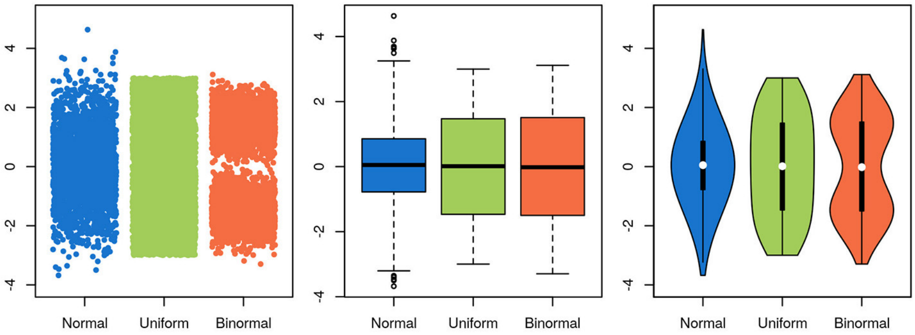

FIGURE 5 | Choosing a visualization technique is a balance between accuracy and clarity. The same three random samples from three underlying distributions are depicted in all panels. A normal distribution is depicted in blue, a uniform in green and a bimodal in orange. In the scatterplot in the leftmost panel, every single datapoint is depicted. While this shows the data accurately, it is harder to estimate the underlying distribution by eye, and adding more points will further obscure the distribution, as the overlap between them increases. Further, the width of the x-scatter, while depicted, is not informative. In the middle panel, the same data is depicted as boxplots. The problem of data overlap is solved by depicting the data in terms of quantiles, but now the binormal and the uniform distribution look very similar. Finally, in the rightmost panel, a violin plot is depicted with the corresponding box-plot on top (boxplot in black, median in white). While not representing the data as fine grained as in the scatter-plot, this depictions represents the data both thoroughly and accurately.

more appropriate for microbiome data. Another conceptually similar strategy is correspondence analysis (CA), where rather than maximizing the percentage of variance explained by each axis, the correspondence between rows and columns in the matrix is optimized. In the case of UniFrac, there are also specific methods developed for it, such as edge PCA, which directly selects high variability lineages as axes on the PCA, allowing a direct phylogenetic interpretation of the resulting plot (Matsen and Evans, 2013).

To assist in the graphical interpretation of a PCA or PCoA, it can be useful to plot each of the measured variables against the two main components. An extension of this that facilitates comparing separate PCA is the circle plot, which focuses only on the variables measured and how they relate to each other (independent, positive or negative correlation). PCA clusters can also become difficult to visualize if there are too many data points. In this case, an alternative to plotting each point is to use a density cloud, where the number of points per area unit is shown with the use of color. Finally, it can be useful to highlight clusters with a visible elliptic contour covering, for instance, 1–2 standard deviations from the cluster's center (**Figure 6**). This is useful to highlight *a priori* expected clusters and how they compare to the actual results, or in cases where clusters are poorly visible in just a few dimensions. Nevertheless, when interpreting plots containing contours it is important to discern which clusters are clearly visible and which are merely highlighted. It is also possible to calculate how many axes are shared from two or more PCA analyses, using the technique of common principal component analysis (CPCA).

Unlike these techniques, in multidimensional scaling (MDS), the number of dimensions to which the dataset should be reduced is chosen a priori and the algorithm finds the distribution of objects in the lower-dimension space that best corresponds to their distances in the full dimension, while also calculating a stress function representing the amount of the distortion between the true distances and the distances in the reduced space. If using

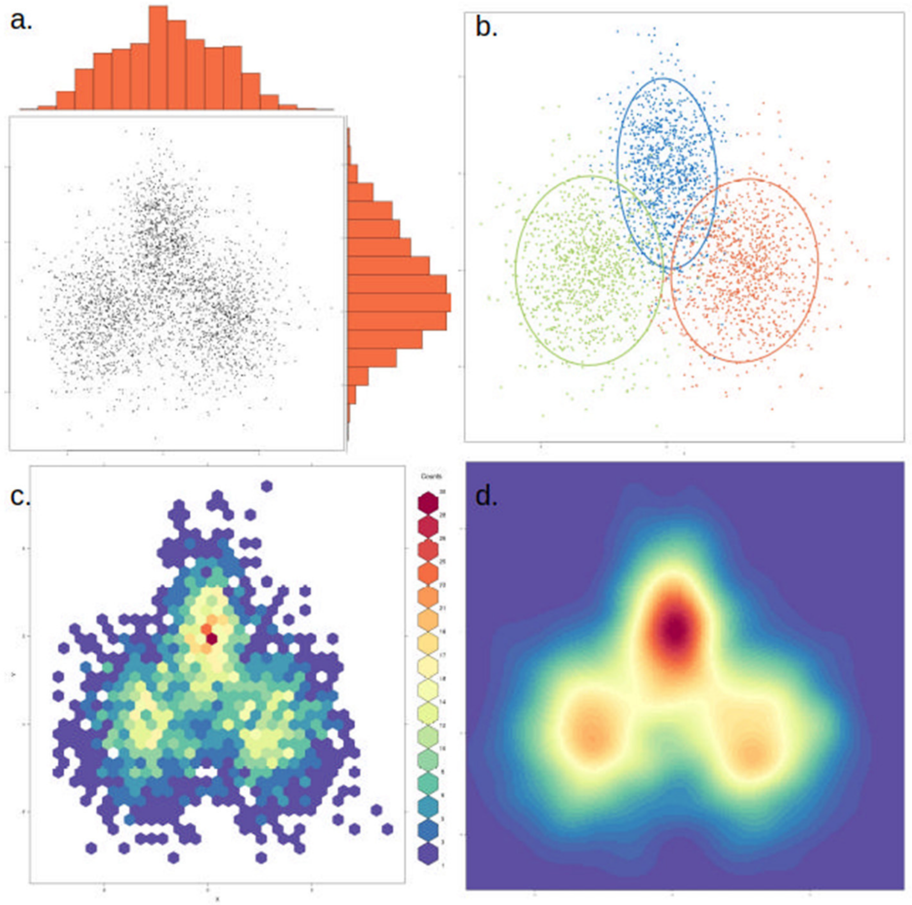

FIGURE 6 | The same data can be represented in different ways to highlight or dampen properties. The same two-dimensional distribution, composed of three overlapping normal distributions with different means and variances, is depicted in each of the four panels. **(a)** Simple scatter plot of the data, with histograms showing the density distribution in each dimension. **(b)** Data points are colored by cluster (representing the underlying distributions) and an ellipse marks the 85th percentile of each distribution. In this case clusters were defined *a priori*, but this could be achieved through various clustering methods. **(c)** The data is binned into hexagonal bins and the color of each hexagon corresponds to the density of points in the bin, as shown in the figure legend. **(d)** Similar to **(c)**, but data is smoothened and interpolated. While clusters are more visible, this approach might also artificially strengthen non-significant clusters.

euclidean distances, i.e., a "classical MDS," the result is identical to a PCA. However, other true distance metrics can be used in metric MDS, and non-metric MDS is an extension of this technique using the ranks of distances rather than their values. Clarke (1993) includes many interesting practical considerations in the interpretation of MDS plots.

Once relevant clusters are determined, either through exploratory techniques or through the use of previous knowledge, linear discriminant analysis (LDA or its multisample counterpart MDA) can be used to define which linear combination of quantitative descriptive variables best separates these clusters. If using clusters found by a dimension-reduction approach, however, it is crucial that the LDA is performed on independent data. The model generated by the LDA can later be used to partition new data into one of the known clusters. LDA will fail if any of the clusters has too few data points, if their variance is not independent from their mean and in the presence of categorical variables or significantly non-linear interactions between variables.

It is often not clear what the main driver of the community over the gradient is, or even how many overlapping gradients there are. These are the cases where exploratory methods are most needed, but also where the biases of each method can most affect the biological interpretation of results. For instance, the horseshoe effect, where sparse matrices driven by a single dominant gradient assume an arch-like pattern when submitted to PCA or CA, may mask other, more subtle gradients, and detrending techniques used to eliminate this effect often erase true patterns (Kuczynski et al., 2010). In datasets with many overlapping gradients, an NMDS will often produce a clearer overview of the data distribution than methods that don't limit the number of dimensions (Paliy and Shankar, 2016). It is therefore recommendable to try a variety of different approaches and retain not merely those which explain the largest proportion of the variation in the dataset, but also those that propose underlying biological mechanisms amenable to further investigation.

Another popular method for visualizing data clusters, specially when there are more variables than samples, is through the use of heatmaps, often associated to a dendrogram. The main problem with this practice is the use of the red-green color scale, which is inaccessible to up to 8% of the male population (Simunovic, 2010). This is however easily by-passed through the use of a yellow-blue color scale. Another problem is the associated use of dendrograms. Since the human mind gives much more emphasis to distance than to associated lines, people tend to perceive data rows or columns that are side-by-side as more similar than those with a smaller branch length, but further apart. This is unavoidable in the use of trees, and something that has to be kept in mind when inspecting one.

HYPOTHESIS TESTING

In addition to visually inspecting the data, it is important to have statistical tests to assess the plausibility of proposed hypotheses. Typical questions that a researcher might ask from these data are: does the microbial community cluster according to predefined sample groups, eg patients and healthy controls? Does the distribution of the community reflect the underlying contextual parameters, eg physicochemical environmental data? What component of the environmental or patient data corresponds to the largest shift in the microbial community? Conversely, what components of the microbial community correspond to the largest shift in health or environmental markers?

For instance, it might be important to assess whether a *priori* groupings of samples, such as different environments or treatment groups, correspond indeed to statistically different microbial communities. Most researchers are familiar with one- and two-way analysis of variance (ANOVA/MANOVA). However, due to the non-normality of most microbial data, non-parametric versions of these tests are needed. Kruskal-Wallis' *H*-test, also known as "ANOVA on ranks" is suitable when there are only two sample groups. For multiple comparisons, non-parametric MANOVA is often termed PERMANOVA, since permutations are used to assess significance. ANOSIM is a similar test, which assesses whether ranks of distances of objects within *a priori* defined classes are smaller than between those classes. Closely related to these tests, linear discriminant analysis (LDA) tests whether groups of samples are significantly different on multiple axes and then attempts to find one axis that optimally discriminates the groups. However, LDA requires that the groups' variance is independent from their mean, which is often not the case in microbiomics data.

There are also several tests available that assess how similar two matrices are. This is the mathematical equivalent to visually assessing the likeness of two PCAs (e.g., one based on OTU and one based on metadata) to say whether they likely reflect related phenomena. For instance, canonical correlation analysis (CCorA) tries to find the linear combinations of variables in two datasets that provide the maximum correlation between them. The non-parametric extension of CCorA is called BIOENV and is deemed more suitable for ecological data (Clarke, 1993; Clarke and Ainsworth, 1993). In Procrustes analysis, the same set of objects (e.g., samples) placed on different spaces (e.g., biological domains or metabolites) are moved, rotated and scaled to minimize the sum of distances between pairs of corresponding objects. A conceptually similar test is Mantel's, which calculates the correlation between two distance matrices and assesses significance by permutation.

When it is clear which are the explanatory variables and which are the response variables, methods can be constrained accordingly. Redundancy analysis (RDA) extracts and summarizes the extent of variation in a response dataset which can be accounted by an explanatory dataset. Likewise, canonical correspondence analysis (CCA) maximizes the correspondence between rows and columns in a table, constrained to the explanatory variables.

If one variable overwhelms the effect of all others, as can be the case in intervention studies in which all treated samples are clustered together and apart from the non-treated, a principal responses curve (PRC) can be used (van den Brink et al., 2009). This approach is also useful if an overlap of many potentially interacting gradients makes the visual interpretation of an RDA or CCA plot impossible. Other approaches with the same goal, such as partial least-squared regression and canonical inertia analysis are thoroughly discussed by Le Cao et al. (2009).

None of the strategies discussed here can distinguish correlation from causation, except perhaps in intervention studies. More importantly, clusters and gradients produced along artificial axes do not necessarily correspond to any underlying biological effect. From a mathematical perspective, variables of different types (e.g., metabolomics vs. microbiomics) will often have different variance-to-mean characteristics, which requires appropriate data transformation (Paliy and Shankar, 2016). New methods for testing hypotheses based on high-throughput data are still being developed, and understanding their strengths as well as their assumptions is a crucial and challenging issue for microbial ecologists. Detailed descriptions, assumptions, limitations and test cases of many popular statistical methods for ecological research can be found in the GUSTAME server (Buttigieg and Ramette, 2014), and in the review by Paliy and Shankar (2016).

ASSESSING THE ROLES OF SPECIFIC OTUS IN THE COMMUNITY

In many cases, it isn't enough to determine how contextual data interact with the microbiome at the community level. In addition, it may be important to determine which organisms contribute most to the community differences. Similarity percentages breakdown (SIMPER) measures the contribution of individual OTUs to Bray-Curtis dissimilarities between sample groups (Clarke, 1993). In other cases, even if the total community isn't significantly different, a subset of OTU might still display significant changes in abundance, such as pathogens or taxa with unusual metabolic capabilities. This is important in the development of diagnostic tools, environmental surveillance strategies and generally for generating testable hypotheses.

In some cases, identifying differentially abundant OTU can be straightforward. For instance, after detecting conditionally rare taxa, Shade et al. (2014) directly calculated which fraction of the distance between communities could be attributed to these specific OTUs. Often, however, there is no clear *a priori* choice of OTU to analyse, and specific methods have to be applied. From a mathematical perspective, there are three main challenges to identifying differentially abundant OTU: the variance of each OTU is not independent from its measured value (heteroskedasticity), most OTUs are below detection limit in most samples (0-inflation, or sparsity) and, due to normalization procedures, the observed value for each OTU in a sample depends on the others (non-independence). Additionally, different statistical tests perform quite differently in cases close to the detection limit, with e.g., the *t*-test failing when the count on either sample under analysis is 0, but Fisher's test performing as expected (Bullard et al., 2010).

Metastats, released in 2009, deals with sparsity by separately considering sparsely sampled OTU using Fisher's exact test (White et al., 2009). Instead of assuming data normality, a nonparametric *t*-test is used, with multiple testing correction performed by calculating the false discovery rate.

Many tools initially developed for RNA-sequencing data can also be used for microbiome studies (Jonsson et al., 2016). Released in 2010, edgeR explicitly models the underlying distribution of each feature (e.g., gene or OTU) as a negative binomial distribution, using an empirical Bayes procedure and conditioning each OTU's variance on their abundance (Robinson et al., 2010). Several other tools were released since then that model the distribution of each OTU using similar procedures. One of the most popular is DESeq2 (Love et al., 2014). In it, information is shared across OTU and it is assumed that OTU with similar abundance will have similar dispersions. This assumption is however over-ruled when the observed variance is more than two-fold different from the mean variance. DESeq2 also considers that noise is greater when counts are low, and is more aggressive in its variation shrinkage approach for low-abundance OTU. Significance of differentially abundant OTU is assessed via a Wald test and multiple testing correction is performed via Benjamini-Hochberg, but the false negative rate is minimized by previous removal of low abundance OTU (whose likelihood of being significantly differentially abundant is, in any case, low). DESEeq2 also includes a tool for making the variance of each OTU independent from its mean (regularized log normalization), a formal requirement for many of the machine learning and ordination methods discussed in this review.

As an alternative to methods such as edgeR or DESeq2 that depend on the negative binomial distribution, SAMSeq (Li and Tibshirani, 2011) was developed as a non-parametric method. SAMSeq conducts Mann–Whitney test on multiple resampling of the data to account for different sequencing depths. It has been reported to give fewer false positives than the negative-binomial tests but has low sensitivity in case of small sample sizes.

A Bioconductor package explicitly aiming at modeling OTU count data was released in 2013 as metagenomeSeq (Paulson et al., 2013). This package introduces two novelties. Firstly, instead of normalizing counts by the total sum of each sample, a percentile cut-off is used. This percentile is chosen automatically by selecting the highest percentile after which there is a large instability between expected values and observed values, suggestive of PCR biases. In addition to this, since microbiomics data is generally much more sparse than RNA sequencing data, a different distribution was chosen to model the data, namely a zero-inflated Gaussian. However, posterior work showed that the zero-inflated Gaussian has a higher rate of false positives than negative-binomial based approaches, and recommended either edgeR or DESeq2 as best practices (McMurdie and Holmes, 2014). These packages and others can be easily used in R in combination with other microbiomics tools through the wrapper package PhyloSeq, which also includes extensive documentation of its features (McMurdie and Holmes, 2013).

The linear discriminant analysis effect size method (LEfSe; Segata et al., 2011) takes a different approach by combining standard statistical tests with the usage of previous biological knowledge in its search for markers. After a first round of feature selection through Kruskal-Wallis' sum-rank test, which identifies OTU differentially abundant between conditions, LEfSe uses pairwise Wilcoxon's tests to discard OTU whose differential abundance isn't consistent across sub-conditions, a step intended to remove spurious correlations. Since these two tests are non-parametric, the possible non-normality of the data is not an issue. Finally, it uses linear discriminant analysis to estimate the effect size of each differentially abundant OTU, an important step in biomarker discovery, as even a highly statistically significant marker is unlikely to be driving environmental or host phenotypic changes if its effect size is too small. The particular setup of LEfSe emphasizes the need to address a range of relevant conditions within the characteristic under study, and is therefore an interesting example of a computational method driving study design.

COMMUNITY DYNAMICS AND NETWORK RECONSTRUCTION

While assessing the relationship between microbial communities and environmental parameters or treatments will always be of fundamental importance, there is mounting evidence that, within a given ecosystem, interactions between taxa play a more important role in driving community dynamics than environmental forcing (Gilbert et al., 2012; Lima-Mendez et al., 2015).

The most basic approach to hypothesizing interactions between microbial populations is through pairwise relationships, either as presence/absence ("checkerboard patterns") or through quantitative measures (**Figure 7**). The latter generally relies on measures of correlation such as Spearman's and Pearson's, while the hypergeometric distribution is appropriate for binary data (Chaffron et al., 2010; Freilich et al., 2010). WGCNA (Langfelder and Horvath, 2008) is a useful R package to create, analyse, compare and visualize correlation networks.

These simple correlation approaches are hampered by limited sampling depth and the ensuing compositionality of the data, which induces spurious correlations. SparCC is a tool built with this caveat in focus, and by-passes it by including the calculation of effective sample-size in its interaction estimation (Friedman and Alm, 2012). SparCC assumes that every OTU is present in every sample, but that they are often below detection limit. Therefore, OTU expected to be very rare and seldom present should not be included in its estimates. However, others have shown that increased rarefaction of data (leading to increase in proportion of 0 count OTU) greatly increases the rate of false positives for all methods tested (Weiss et al., 2016).

Regardless of the procedure adopted, the underlying hypothesis is that, if there is an interaction between two species, and given similar environments with similar resources, these two species will co-occur more likely than expected by chance if their interaction is beneficial (mutualism or commensalism) and co-occur less likely than expected by chance if their interactions is prejudicial (competition or amensalism). However, two of the most important types of interactions in natural systems, predation and parasitism, are beneficial to one of the parts (the predator or parasite) and prejudicial to the other (the prey or host), complicating the ecological interpretation of co-occurrence patterns. Furthermore, given the intricacies of microbial metabolism, it is seldom clear if a species is excluded from a niche due to negative interactions with other organisms or due to environmental constraints. Nevertheless, mapping pairwise correlations can be a useful first step in developing an interaction hypothesis.

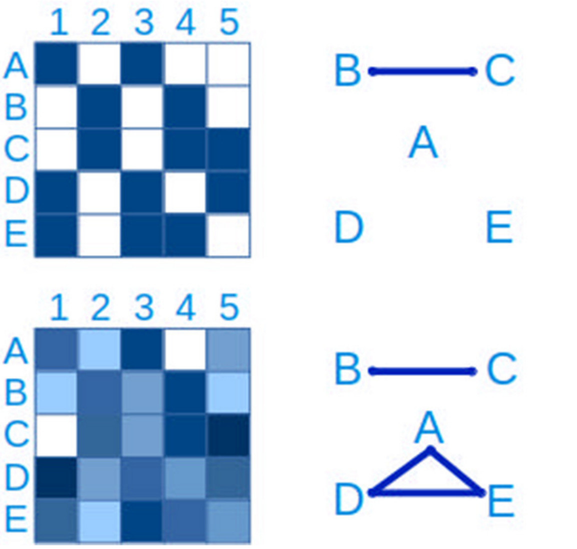

FIGURE 7 | Connections between taxa can be inferred from checkerboard patterns. The same data is represented on the left side. In the upper panel, a presence-absence table of clades A–E in samples 1-5 is represented, where blue is presence and white is absence. From this data it is possible to infer a connection between clades B and C. In the lower left panel, the presence-absence table has been replaced by a quantitative table, with a lower detection limit, more typical of microbial ecology studies. In this case, in addition to the connection between clades B and C, a three-way connection between clades A,D and E is inferred.

Other available approaches do not depend on monotonic correlations. Maximal Information Coefficient (MIC) (Reshef et al., 2011) is a non-parametric approach designed to detect associations and to give similar scores to associations with similar noise levels, regardless of their shape (linear, exponential, periodic etc.). Intuitively, this is achieved by plotting the abundance of OTUs against each other, pairwise, and over each plot defining a grid which splits the sections of the graph that contain data from those which do not. Mutual information—a measure of the predictability of two variables in relation to each other—is then calculated for each section of the grid. The MIC algorithm penalizes overly complex relations by decreasing the score according to the number of partitions in the grid.

Since, in the typical case, thousands of correlations and anticorrelations will be tested, the significance of any association has to be tested and subjected to multiple testing correction. This is often done by randomizing the interaction network and calculating the distribution of scores. It is however still not clear what the correct randomization procedure is for this type of data (Faust and Raes, 2012). One alternative to reduce the rate of false positive inferences is to combine different approaches and keep only links supported by multiple sources of evidence.

A tool for doing this was introduced by Faust et al. (2012) and is alternatively called CoNet, Reboot or CCRePe, depending on its implementation. However, in further work, the same authors used CoNet as only one of the elements in a more elaborate ensemble approach with superior results (Weiss et al., 2016).

The pairwise interactions inferred by the techniques described above can be used to build networks where each OTU or measured environmental parameter is a node and interactions between them are links. In addition to being a rich representation of interactions between particular nodes, properties of the network itself can contain information about the system. For instance, microbial networks are generally modular, scale-free and have short average path length (Faust and Raes, 2012). How these mathematical properties translate into biological properties is still open to debate. It is not clear, for instance, whether a node with a high degree (i.e., linked to many nodes in the network) represents a keystone clade whose demise would severely perturb the entire system, or whether the levels of redundancy and plasticity in biological systems are enough to functionally replace these hubs without much propagation of perturbation. In the case of bacteria in particular, not only does the community present a certain level of plasticity, but single populations and even individual cells can dramatically alter their life strategy in response to disturbances, decoupling to a large extent a community's taxonomic composition from its functional profile (Shade et al., 2012; Comte et al., 2013). Network properties also interact with community characteristics such as richness and evenness, and often have opposite effects in the resulting resistance and resilience of the community to perturbation, so that broad natural laws of community stability might be impossible to obtain (Shade et al., 2012).

A powerful approach to gain insight into the internal mechanisms of a natural microbial community is sampling a time-series with appropriate intervals and length, and using techniques such as Local Similarity Analysis (Ruan et al., 2006; Steele et al., 2011) or auto- and cross-correlation (Fuhrman et al., 2006; Gilbert et al., 2012; David et al., 2014). If a system has an intrinsic periodicity, such as annual cycles, a few full cycles should be included in the study to separate recurring patterns, random fluctuations and system drift (time decay; Gilbert et al., 2012; Kara et al., 2013). Also important is to consider that different processes might take place at different rates, corresponding to one or more sampling intervals or, conversely, that associations that are significant in the short term can be irrelevant at longer time-spans (Steele et al., 2011; Needham et al., 2013).

Strong seasonal recurrence has been reported in several sites, with the rate of interannual decay declining with the length of the time-series (Gilbert et al., 2012; Cram et al., 2015). Due to seasonality, the time-frame which is relevant for most free-living microbial assemblages are those which are one-year apart, i.e., in the same season. Stochastic factors mean that there is significant loss of signal from one year to the next. However, environmental and biological constraints maintain community variation within certain boundaries (**Figure 8**). In addition to recurrent and linear (time-decay) patterns, dramatic but rare events can occasionally also be observed in long time-series (Gilbert et al., 2012; Vergin et al., 2013; Lindh et al., 2015).

Local Similarity Analysis is a strategy optimized for time-series data and non-linear interactions (Ruan et al., 2006), and is available as the stand-alone package eLSA (Xia et al., 2013). It is robust to data sparsity and was evaluated by Weiss et al. (2016) to be the overall best approach for time-series data. They note, however, that the frequency of sampling plays an important part in the algorithm's accuracy. While eLSA can fill in missing data by interpolation, it is not clear to what extent this can mask or induce spurious correlations.

Weiss et al. (2016) further noted that none of the strategies currently available is sufficiently robust. Even the choice of sequencing technology was observed to significantly affect the output of network inference algorithms. Experimental validation is therefore crucial, but can be very hard to conduct, especially if the interactions do not involve physical contact of the cells. Carefully considering the sampling procedure and adapting it to the needs of the network inference algorithm is therefore recommended. Furthermore, it is generally still not possible to accurately predict interactions involving one or more OTU which is present in only a small fraction of the samples, so computational time and statistical power can be saved by

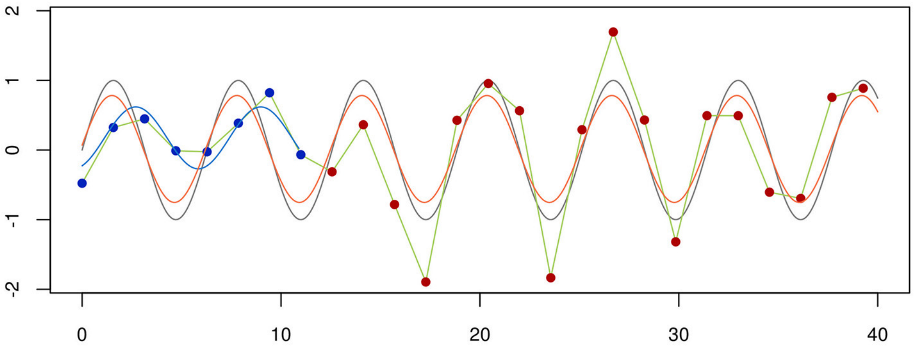

FIGURE 8 | Temporal decay is dampened in longer time-series. The gray line represents an oscillator, e.g., a microbial community subjected to strong seasonal variation. Each point is one sample, representing the intrinsic oscillation of the community plus stochastic deviations from it. Simply connecting the points (green line) doesn't give any mechanistic insight. A first mechanistic hypothesis can be generated from a few cycles (blue dots, blue line), but it is significantly worse than a more complete temporal series (all circles, orange line).

removing rare OTU (present in less than 30–60% of samples; Weiss et al., 2016) before network inference.

A very different approach to generating predictive models for OTUs based on environmental data and/or other OTUs is through artificial neural networks (ANN) or Random Forests. Briefly, an ANN is a layered series of computing units, analogous to neurons in a real neural circuit. Raw data is fed to an initial layer and is then relayed non-linearly through each layer of the ANN. At each layer, each computing unit receives data from each unit of the previous layer, and performs a weighting procedure to its input and then another non-linear operation. At the output, a classification or numeric prediction is made. Despite very promising results (Larsen et al., 2012), this approach has not been widely adopted by the field. A more thorough discussion of neural networks and applications to computational biology can be found in Angermueller et al. (2016).

Random Forests, on the other hand, are machine learning strategies based on decision trees. A decision tree starts with a table of pre-classified data. Based on that, it determines decision criteria for classifying new data. In addition to that, a decision tree can be used to fit, e.g., linear models to each partition, if the data of interest is quantitative. Random Forests are an extension of decision trees where several random subsets of the total data are given as input to different trees (thereby creating a forest of decision trees). This increases the robustness of the prediction and allows the estimation of classification accuracy based on the training data. For datasets with a large number of parameters, improved predictions can sometimes be achieved with a pre-selection criterion (Lima-Mendez et al., 2015).

SUMMARY AND PERSPECTIVES

Life on earth was exclusively microbial for most of its history, and is still predominantly so. Microbiologists have been striving to catalog, understand and manage this wealth of life for almost 250 years, and yet been severely limited by technical development. Historically, while general ecology has been based on direct observation combined with mathematical modeling, breakthroughs in microbial ecology have been coupled to technological advance. With recent advances in technologies such as microfluidics and high-throughput DNA sequencing, as well as the steady growth of computational methods and processing capacity, the pace of advance in microbial ecology has been greatly increased. In addition to these approaches, microbiologists can now use metagenomics, metatranscriptomics, metaproteomics, metabolomics, single-cell genome sequencing, genome binning, flow cytometry, cell sorting, high-throughput image analysis and nanoSIMS (nanoscale mass spectrometry), together providing a wide array of complementary techniques for assessing microbial phylogeny and activity in bulk as well as at the single-cell level.

While the work of mapping and modeling microbial life on earth will remain an open field of basic scientific inquiry, it is important to also consider the potential medical and technological applications of these studies. From alternative fuel sources to environmental decontamination, antibiotic resistance to prevention and treatment of immunological and metabolic disorders, many of the biggest challenges of our times may soon find their answers in the myriad of strategies microorganisms adapt to survive, compete, cooperate and thrive on earth. It is therefore crucial that the full potential, as well as the caveats and biases, of established and nascent microbiology approaches are understood.

AUTHOR CONTRIBUTIONS

LH and AA participated in the study design and the elaboration of the manuscript.

FUNDING

This work was supported by BONUS BLUEPRINT project, supported by BONUS (Art 185), funded jointly by the eU and the Swedish Research Council FORMAS. It is also funded by the Swedish Research Council VR (grant 2011-5689).

ACKNOWLEDGMENTS

The authors greatly appreciate Anders Lanzén's careful proofreading of the manuscript. We thank Fredrik Boulund and Mattias Wickberg for guidance in improving the quality of our figures. Finally, we are grateful to Rafael Sola De Paula De Angelo Calsaverini for vital conversations on distance metrics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01561/full#supplementary-material

Supplementary File 1 | R code used in figures. **Figures 3–6**, **8** were created in R (v.3.3.1; R Development Core Team, 2016) using packages RGP (v.0.4-1; Flasch et al., 2014), Vegan (v.2.4-3; Oksanen et al., 2017) vioplot (v0.2; Adler, 2005), car (v.2.1-4; Fox and Weisberg, 2011), hexbin (v.1.27.1; Carr et al., 2016), RColorBrewer (v.1.1-2; Neuwirth, 2014), gplots (v.3.0.1; Warnes et al., 2016), and MASS (v7.3-47; Venables and Ripley, 2002). The code used is available in the Supplementary File.

REFERENCES

Aakra, A., Utåker, J. B., Nes, I. F., and Bakken, L. R. (1999). An evaluated improvement of the extinction dilution method for isolation of ammonia-oxidizing bacteria. *J. Microbiol. Methods* 39, 23–31. doi: 10.1016/S0167-7012(99)00094-9

Adler, D. (2005). *vioplot: Violin plot.* Available online at: https://cran.r-project.org/web/packages/vioplot/index.html

Allen, B., Kon, M., and Bar-Yam, Y. (2009). A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *Am. Nat.* 174, 236–243. doi: 10.1086/600101

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, W., and Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* 4:e6372. doi: 10.1371/annotation/50c43133-0df5-4b8b-8975-8cc37d4f2f26

Andersson, A. F., Lindberg, M., Jakbosson, H., Bäckhed, F., Nyrén, P., and Engstrand, L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* 3:e2836. doi: 10.1371/journal.pone.0002836

Andrews, S. (2009). *FastQC. A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12:878. doi: 10.15252/msb.20156651

Aoi, Y., Kinoshita, T., Hata, T., Ohta, H., Obokata, H., and Tsuneda, S. (2009). Hollow-fiber membrane chamber as a device for in situ environmental cultivation. *Appl. Environ. Microbiol.* 75, 3826–3833. doi: 10.1128/AEM.02542-08

Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 31, 2882–2884. doi: 10.1093/bioinformatics/btv287

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

Balvočiūtė, M., and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT – how do these taxonomies compare? *BMC Genomics* 18:114. doi: 10.1186/s12864-017-3501-4

Benítez-Páez, A., Portune, K. J., and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION^{TM} portable nanopore sequencer. *Gigascience* 5:4. doi: 10.1186/s13742-016-0111-z

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., and Tyson, G. W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* 9:e1003031. doi: 10.1371/journal.pcbi.1003031

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550. doi: 10.1111/1574-6941.12437

Cadotte, M. W., Jonathan Davies, T., Regetz, J., Kembel, S. W., Cleland, E., and Oakley, T. H. (2010). Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol. Lett.* 13, 96–105. doi: 10.1111/j.1461-0248.2009.01405.x

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

Campbell, B. J., Yu, L., Heidelberg, J. F., and Kirchman, D. L. (2011). Activity of abundant and rare bacteria in a coastal ocean. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12776–12781. doi: 10.1073/pnas.1101405108

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., and DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* 13:375. doi: 10.1186/1471-2164-13-375

Carr, D., Nicholas Lewin-Koh, P., Maechler, M., and Deepayan Sarkar, C. C. L. F. W. (2016). *hexbin: Hexagonal Binning Routines* . Available online at: https://CRAN.Rproject.org/package=hexbin

Certini, G., Campbell, C. D., and Edwards, A. C. (2004). Rock fragments in soil support a different microbial community from the fine earth. *Soil Biol. Biochem.* 36, 1119–1128. doi: 10.1016/j.soilbio.2004.02.022

Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* 20, 947–959. doi: 10.1101/gr.104521.109

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28, 2106–2113. doi: 10.1093/bioinformatics/bts342

Choo, J. M., Leong, L. E. X., and Rogers, G. B. (2015). Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* 5:16350. doi: 10.1038/srep16350

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287. doi: 10.1126/science.1123061

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* 18, 117–143. doi: 10.1111/j.1442-9993.1993.tb00438.x

Clarke, K. R., and Ainsworth, M. (1993). A method of linking multivariate community structure to environmental variables. *Mar. Ecol. Prog. Ser.* 92, 205–219. doi: 10.3354/meps092205

Coltharp, C., and Xiao, J. (2012). Superresolution microscopy for microbiology. *Cell. Microbiol.* 14, 1808–1818. doi: 10.1111/cmi.12024

Comte, J., Fauteux, L., and Del Giorgio, P. A. (2013). Links between metabolic plasticity and functional redundancy in freshwater bacterioplankton communities. *Front. Microbiol.* 4:112. doi: 10.3389/fmicb.2013.00112

Cram, J. A., Chow, C.-E. T., Sachdeva, R., Needham, D. M., Parada, A. E., Steele, J. A., et al. (2015). Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J.* 9, 563–580. doi: 10.1038/ismej.2014.153

D'Onofrio, A., Crawford, J. M., Stewart, E. J., Witt, K., Gavrish, E., Epstein, S., et al. (2010). Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chem. Biol.* 17, 254–264. doi: 10.1016/j.chembiol.2010.02.010

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15:R89. doi: 10.1186/gb-2014-15-7-r89

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C. (2016a). UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. *bioRxiv*, 088666. doi: 10.1101/088666

Edgar, R. C. (2016b). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, 081257. doi: 10.1101/081257

Ehrenreich, A. (2006). DNA microarray technology for the microbiologist: an overview. *Appl. Microbiol. Biotechnol.* 73, 255–273. doi: 10.1007/s00253-006-0584-2

Eren, M. A., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi: 10.1016/0006-3207(92)91201-3

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8:e1002606. doi: 10.1371/journal.pcbi.1002606

Fisher, M. M., and Triplett, E. W. (1999). Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl. Environ. Microbiol.* 65, 4630–4636.

Flasch, O., Mersmann, O., Bartz-Beielstein, T., Stork, J., and Zaefferer, M. (2014). *rgp: R Genetic Programming Framework*. Available online at: https://CRAN.Rproject.org/package=rgp

Fonseca, V. G., Nichols, B., Lallias, D., Quince, C., Carvalho, G. R., Power, D. M., et al. (2012). Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Res.* 40:e66. doi: 10.1093/nar/gks002

Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.* 42, 166–170. doi: 10.1099/00207713-42-1-166

Fox, J., and Weisberg, S. (2011). *An R Companion to Applied Regression.* Available online at: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Freilich, S., Kreimer, A., Meilijson, I., Gophna, U., Sharan, R., and Ruppin, E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.* 38, 3857–3868. doi: 10.1093/nar/gkq118

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

Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V., and Naeem, S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13104–13109. doi: 10.1073/pnas.0602399103

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gantner, S., Andersson, A. F., Alonso-Sáez, L., and Bertilsson, S. (2010). Novel primers for 16S rRNA-based archaeal community analysis in environmental samples. *J. Microbiol. Methods* 84, 12–18. doi: 10.1016/j.mimet.2010.10.001

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., et al. (2005). Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733–739. doi: 10.1038/nrmicro1236

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., et al. (2012). Defining seasonal marine microbial community dynamics. *ISME J.* 6, 298–308. doi: 10.1038/ismej.2011.107

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.-F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12:245. doi: 10.1186/1471-2164-12-245

Gong, J., Dong, J., Liu, X., and Massana, R. (2013). Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of Oligotrich and Peritrich Ciliates. *Protist* 164, 369–379. doi: 10.1016/j.protis.2012.11.006

Goodfellow, M., Kampfer, P., Busse, H.-J., Trujillo, M. E., Suzuki, K.-I., Ludwig, W., et al. (2012). *Bergey's Manual of Systematic Bacteriology*. New York, NY:Springer.

Gordon, A., and Hannon, G. (2009). *FASTX-toolkit*. Available online at: http://hannonlab.cshl.edu/fastx_toolkit/

Gorzelak, M. A., Gill, S. K., Tasnim, N., Ahmadi-Vand, Z., Jay, M., and Gibson, D. L. (2015). Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS ONE* 10:e0134802. doi: 10.1371/journal.pone.0134802

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* 41, D597–D604. doi: 10.1093/nar/gks1160

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* 5, 235–237. doi: 10.1038/nmeth.1184

Harju, S., Fedosyuk, H., and Peterson, K. R. (2004). Rapid isolation of yeast genomic DNA: Bust n' Grab. *BMC Biotechnol.* 4:8. doi: 10.1186/1472-6750-4-8

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., et al. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 3:20. doi: 10.1186/s40168-015-0081-x

Hu, Y. O. O., Karlson, B., Charvet, S., and Andersson, A. F. (2015). Diversity of Pico- to Mesoplankton Along the 2000 km Salinity Gradient of the Baltic Sea. *Front. Microbiol.* 7:679. doi: 10.1101/035485

Hu, Y., Ndegwa, N., Alneberg, J., Johansson, S., Logue, J., Huss, M., et al. (2017). *Stationary and Portable Sequencing-Based Approaches for Tracing Wastewater Contamination in Urban Stormwater Systems.* Available online at: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1085975&dswid=6256 (Accessed on: June 14, 2017).

Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., et al. (2014a). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS ONE* 9:e95567. doi: 10.1371/journal.pone.0095567

Hugerth, L. W., Wefer, H. A., Lundin, S., Jakobsson, H. E., Lindberg, M., Rodin, S., et al. (2014b). DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Appl. Environ. Microbiol.* 80, 5116–5123. doi: 10.1128/AEM.01403-14

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* 67, 4399–4406. doi: 10.1128/AEM.67.10.4399-4406.2001

Humbert, J. F., Quiblier, C., and Gugger, M. (2010). Molecular approaches for monitoring potentially toxic marine and freshwater phytoplankton species. *Anal. Bioanal. Chem.* 397, 1723–1732. doi: 10.1007/s00216-010-3642-7

Iluz, D., Dishon, G., Capuzzo, E., Meeder, E., Astoreca, R., Montecino, V., et al. (2009). Short-term variability in primary productivity during a wind-driven diatom bloom in the Gulf of eilat (Aqaba). *Aquat. Microb. Ecol.* 56, 205–215. doi: 10.3354/ame01321

Jones, S. E., and Lennon, J. T. (2010). Dormancy contributes to the maintenance of microbial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5881–5886. doi: 10.1073/pnas.0912765107

Jonsson, V., Osterlund, T., Nerman, O., and Kristiansson, E. (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* 17:78. doi: 10.1186/s12864-016-2386-y

Kaeberlein, T., Lewis, K., and Epstein, S. S. (2002). Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. *Science* 296, 1127–1129. doi: 10.1126/science.1070633

Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L., and McMahon, K. D. (2013). A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *ISME J.* 7, 680–684. doi: 10.1038/ismej.2012.118

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., et al. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463–1464. doi: 10.1093/bioinformatics/btq166

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

Koeppel, A. F., and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res.* 41, 5175–5188. doi: 10.1093/nar/gkt241

Komárek, J., and Hauer, T. (2013). *CyanoDB.cz - On-Line Database of Cyanobacterial Genera*. Available online at: http://www.cyanodb.cz/

Krueger, F. (2017). T*rim Galore!* Available online at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* 7, 813–819. doi: 10.1038/nmeth.1499

Kumar, P. S., Brooker, M. R., Dowd, S. E., and Carmelento, T. (2011). Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS ONE* 6:e20956. doi: 10.1371/journal.pone.0020956

Lagier, J.-C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M., and Raoult, D. (2015). Current and past strategies for bacterial culture in clinical microbiology. *Clin. Microbiol. Rev.* 28, 208–236. doi: 10.1128/CMR.00110-14

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. doi: 10.1186/1471-2105-9-559

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821.

Lanzén, A., Jørgensen, S. L., Bengtsson, M. M., Jonassen, I., Ovreås, L., and Urich, T. (2011). Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. *FEMS Microbiol. Ecol.* 77, 577–589. doi: 10.1111/j.1574-6941.2011.01138.x

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., et al. (2012). CREST–classification resources for environmental sequence tags. *PLoS ONE* 7:e49334. doi: 10.1371/journal.pone.0049334

Larsen, P. E., Field, D., and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* 9, 621–625. doi: 10.1038/nmeth.1975

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., et al. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3, 1–8. doi: 10.1016/j.bdq.2015.02.001

Le Cao, K.-A., Martin, P. G. P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformat.* 10:34. doi: 10.1186/1471-2105-10-34

Lehner, A., Loy, A., Behr, T., Gaenge, H., Ludwig, W., Wagner, M., et al. (2005). Oligonucleotide microarray for identification of enterococcus species. *FEMS Microbiol. Lett.* 246, 133–142. doi: 10.1016/j.femsle.2005.04.002

Li, J., and Tibshirani, R. (2011). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* 22, 519–536. doi: 10.1177/0962280211428386

Lim, Y. W., Haynes, M., Furlan, M., Robertson, C. E., Harris, J. K., and Rohwer, F. (2014). Purifying the impure: sequencing metagenomes and metatranscriptomes from complex animal-associated samples. *J. Vis. Exp.* e52117. doi: 10.3791/52117

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015). Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* 348:1262073. doi: 10.1126/science.1262073

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., et al. (2013). Fungal community analysis by high-throughput sequencing of amplified markers–a user's guide. *New Phytol.* 199, 288–299. doi: 10.1111/nph.12243

Lindberg, M. R., Schmedes, S. E., Hewitt, F. C., Haas, J. L., Ternus, K. L., Kadavy, D. R., et al. (2016). A Comparison and Integration of MiSeq and MinION Platforms for Sequencing Single Source and Mixed Mitochondrial Genomes. *PLoS ONE* 11:e0167600. doi: 10.1371/journal.pone.0167600

Lindh, M. V., Sjöstedt, J., Andersson, A. F., Baltar, F., Hugerth, L. W., Lundin, D., et al. (2015). Disentangling seasonal bacterioplankton population dynamics by high frequency sampling. *Environ. Microbiol.* 17, 2459–2476. doi: 10.1111/1462-2920.12720

Liu, W. T., Marsh, T. L., Cheng, H., and Forney, L. J. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol.* 63, 4516–4522.

Liu, W., Kim, H. J., Lucchetta, E. M., Du, W., and Ismagilov, R. F. (2009). Isolation, incubation, and parallel functional testing and identification by FISH of rare microbial single-copy cells from multi-species mixtures using the combination of chemistrode and stochastic confinement. *Lab Chip* 9, 2153–2162. doi: 10.1039/b904958d

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi: 10.1128/AEM.01996-06

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488, 86–90. doi: 10.1038/nature11237

Lundin, D., Severin, I., Logue, J. B., Ostman, O., Andersson, A. F., and Lindström, E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity? *Environ. Microbiol. Rep.* 4, 367–372. doi: 10.1111/j.1758-2229.2012.00345.x

Lynch, M. D. J., and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* 13, 217–229. doi: 10.1038/nrmicro3400

Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Manter, D. K., and Vivanco, J. M. (2007). Use of the ITS primers, ITS1F and ITS4, to characterize fungal abundance and diversity in mixed-template samples by qPCR and length heterogeneity analysis. *J. Microbiol. Methods* 71, 7–14. doi: 10.1016/j.mimet.2007.06.016

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

Martin, M. (2012). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Bioinformat. Action* 17, 10–12. doi: 10.14806/ej.17.1.200

Martiny, J. B. H., Jones, S. E., Lennon, J. T., and Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:aac9323. doi: 10.1126/science.aac9323

Matsen, F. A. IV., and Evans, S. N. (2013). Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE* 8:e56859. doi: 10.1371/annotation/40cb3123-845a-43e7-b4c0-9fb00b6e2212

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformat.* 11:538. doi: 10.1186/1471-2105-11-538

McCoy, C. O., and Matsen, F. A. IV. (2013). Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ* 1:e157. doi: 10.7717/peerj.157

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

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

Moré, M. I., Herrick, J. B., Silva, M. C., Ghiorse, W. C., and Madsen, E. L. (1994). Quantitative cell lysis of indigenous microorganisms and rapid extraction of microbial DNA from sediment. A*ppl. Environ. Microbiol.* 60, 1572–1580.

Moreira, D., and López-García, P. (2002). The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* 10, 31–38. doi: 10.1016/S0966-842X(01)02257-0

Morris, J. J., Johnson, Z. I., Szul, M. J., Keller, M., and Zinser, E. R. (2011). Dependence of the cyanobacterium Prochlorococcus on hydrogen peroxide scavenging microbes for growth at the ocean's surface. *PLoS ONE* 6:e16805. doi: 10.1371/journal.pone.0016805

Morris, J. J., Kirkegaard, R., Szul, M. J., Johnson, Z. I., and Zinser, E. R. (2008). Facilitation of robust growth of prochlorococcus colonies and dilute liquid cultures by "Helper" Heterotrophic Bacteria. *Appl. Environ. Microbiol.* 74, 4530–4534. doi: 10.1128/AEM.02479-07

Muyzer, G., de Waal, E. C., and Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. A*ppl. Environ. Microbiol.* 59, 695–700.

Narihiro, T., and Sekiguchi, Y. (2011). Oligonucleotide primers, probes and molecular methods for the environmental monitoring of methanogenic archaea. *Microb. Biotechnol.* 4, 585–602. doi: 10.1111/j.1751-7915.2010.00239.x

Needham, D. M., Chow, C.-E. T., Cram, J. A., Sachdeva, R., Parada, A., and Fuhrman, J. A. (2013). Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J.* 7, 1274–1285. doi: 10.1038/ismej.2013.19

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. Available online at: https://CRAN.Rproject.org/package=RColorBrewer

Nichols, D., Cahoon, N., Trakhtenberg, E. M., Pham, L., Mehta, A., Belanger, A., et al. (2010). Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. *Appl. Environ. Microbiol.* 76, 2445–2450. doi: 10.1128/AEM.01754-09

Nossa, C. W., Oberdorf, W. E., Yang, L., Aas, J. A., Paster, B. J., Desantis, T. Z., et al. (2010). Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J. Gastroenterol.* 16, 4135–4144. doi: 10.3748/wjg.v16.i33.4135

Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009). New insights into the diversity of marine picoeukaryotes. *PLoS ONE* 4:e7143. doi: 10.1371/journal.pone.0007143

Nye, K. J., Fallon, D., Gee, B., Messer, S., Warren, R. E., and Andrews, N. (1999). A comparison of blood agar supplemented with NAD with plain blood agar and chocolated blood agar in the isolation of *Streptococcus pneumoniae* and *Haemophilus influenzae* from sputum. Bacterial Methods evaluation Group. *J. Med. Microbiol.* 48, 1111–1114. doi: 10.1099/00222615-48-12-1111

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2017). *vegan: Community Ecology Package*. Available online at: https://CRAN.Rproject.org/package=vegan

Okubo, T., Ikeda, S., Yamashita, A., Terasawa, K., and Minamisawa, K. (2012). Pyrosequence read length of 16S rRNA gene affects phylogenetic assignment of plant-associated bacteria. *Microb. Environ.* 27, 204–208. doi: 10.1264/jsme2.ME11258

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformat.* 12:385. doi: 10.1186/1471-2105-12-385

Op De Beeck, M., Lievens, B., Busschaert, P., Declerck, S., Vangronsveld, J., and Colpaert, J. V. (2014). Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. *PLoS ONE* 9:e97629. doi: 10.1371/journal.pone.0097629

Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1985). Analyzing natural microbial populations by rRNA sequences. *ASM News* 51, 4–12.

Paliy, O., and Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Mol. Ecol.* 25, 1032–1057. doi: 10.1111/mec.13536

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *J. Theor. Biol.* 13, 131–144. doi: 10.1016/0022-5193(66)90013-0

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi: 10.1093/bioinformatics/bts252

Pruesse, E., Quast, C., Knittel, K., Fuchs, B., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* 16:114. doi: 10.1186/s13059-015-0677-2

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., et al. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6, 639–641. doi: 10.1038/nmeth.1361

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12:38. doi: 10.1186/1471-2105-12-38

Rappé, M. S., Connon, S. A., Vergin, K. L., and Giovannoni, S. J. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418, 630–633. doi: 10.1038/nature00917

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing .

Reck, M., Tomasch, J., Deng, Z., Jarek, M., Husemann, P., and Wagner-Döbler, I. (2015). Stool metatranscriptomics: a technical guideline for mRNA stabilisation and isolation. *BMC Genomics* 16:804. doi: 10.1186/s12864-015-1694-y

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., et al. (2011). Detecting novel associations in large data sets. *Science* 334, 1518–1524. doi: 10.1126/science.1205438

Ricke, S. C., Khatiwara, A., and Kwon, Y. M. (2013). Application of microarray analysis of foodborne Salmonella in poultry production: a review. *Poult. Sci.* 92, 2243–2250. doi: 10.3382/ps.2012-02740

Ringnér, M. (2008). What is principal component analysis? *Nat. Biotechnol.* 26, 303–304. doi: 10.1038/nbt0308-303

Roberts, J. P. (2016). *Nucleic Acid extraction—Keeping It Stable and Intact.* Available online at: http://www.biocompare.com/Editorial-Articles/187876-Nucleic-Acid-Extraction-Keeping-It-Stable-and-Intact/ (Accessed on: June 14, 2017).

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Roume, H., Muller, E. E. L., Cordes, T., Renaut, J., Hiller, K., and Wilmes, P. (2013). A biomolecular isolation framework for eco-systems biology. *ISME J.* 7, 110–121. doi: 10.1038/ismej.2012.72

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., and Sun, F. (2006). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 22, 2532–2538. doi: 10.1093/bioinformatics/btl417

Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., et al. (2014). Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.* 80, 7583–7591. doi: 10.1128/AEM.02206-14

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43:e37. doi: 10.1093/nar/gku1341

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.* 6:e1000844. doi: 10.1371/journal.pcbi.1000844

Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71, 1501–1506. doi: 10.1128/AEM.71.3.1501-1506.2005

Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L., and Highlander, S. K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* 4:e1869. doi: 10.7717/peerj.1869

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hatmann, M., Hollister, E. B., et al. (2009). Introducing Mothur: Open-source, platform-independent community- supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Schmidt, T. S. B., Matias Rodrigues, J. F., and von Mering, C. (2014). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput. Biol.* 10:e1003594. doi: 10.1371/journal.pcbi.1003594

Schmidt, T. S. B., Matias Rodrigues, J. F., and von Mering, C. (2015). Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ. Microbiol.* 17, 1689–1706. doi: 10.1111/1462-2920.12610

Schmidt, T. S. B., Rodrigues, J. F. M., and von Mering, C. (2016). A Family of Interaction-Adjusted Indices of Community Similarity. *bioRxiv*, 040097. doi: 10.1101/040097

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., et al. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio* 5:e01371-14. doi: 10.1128/mBio.01371-14

Shade, A., Peter, H., Allison, S. D., Baho, D., Berga, M., Buergmann, H., et al. (2012). Fundamentals of microbial community resistance and resilience. *Front. Microbiol.* 3:417. doi: 10.3389/fmicb.2012.00417

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Sys. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., and Caboche, S. (2017). Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. *PLoS ONE* 12:e0169563. doi: 10.1371/journal.pone.0169563

Silva, P. C. (2008). Historical review of attempts to decrease subjectivity in species identification, with particular regard to algae. *Protist* 159, 153–161. doi: 10.1016/j.protis.2007.10.001

Simpson, E. H. (1949). Measurement of Diversity. *Nature* 163:688. doi: 10.1038/163688a0

Simunovic, M. P. (2010). Colour vision deficiency. *Eye* 24, 747–755. doi: 10.1038/eye.2009.251

Singh, D. V., and Mohapatra, H. (2008). Application of DNA-based methods in typing Vibrio cholerae strains. *Future Microbiol.* 3, 87–96. doi: 10.2217/17460913.3.1.87

Sizova, M. V., Hohmann, T., Hazen, A., Paster, B. J., Halem, S. R., Murphy, C. M., et al. (2012). New approaches for isolation of previously uncultivated oral bacteria. *Appl. Environ. Microbiol.* 78, 194–203. doi: 10.1128/AEM.06813-11

Smith, B. C., McAndrew, T., Chen, Z., Harari, A., Barris, D. M., Viswanathan, S., et al. (2012). The cervical microbiome over 7 years and a comparison of methodologies for its characterization. *PLoS ONE* 7:e40425. doi: 10.1371/journal.pone.0040425

Soergel, D. A., Dey, N., Knight, R., and Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440–1444. doi: 10.1038/ismej.2011.208

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Song, S. J., Amir, A., Metcalf, J. L., Amato, K. R., Xu, Z. Z., Humphrey, G., et al. (2016). Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. *mSystems* 1:e00021-16. doi: 10.1128/mSystems.00021-16

Stahl, D. A., Lane, D. J., Olsen, G. J., and Pace, N. R. (1985). Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl. Environ. Microbiol.* 49, 1379–1384.

Staley, J. T., and Konopka A. (1985). Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. A*nnu. Rev. Microbiol.* 39, 321–346. doi: 10.1146/annurev.mi.39.100185.001541

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44, 846–849.

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., et al. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.* 5, 1414–1425. doi: 10.1038/ismej.2011.24

Stevens, J. L., Jackson, R. L., and Olson, J. B. (2013). Slowing PCR ramp speed reduces chimera formation from environmental samples. *J. Microbiol. Methods* 93, 203–205. doi: 10.1016/j.mimet.2013.03.013

Stewart, E. J. (2012). Growing unculturable bacteria. *J. Bacteriol.* 194, 4151–4160. doi: 10.1128/JB.00345-12

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D., Breiner H. W., et al. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* 19, 21–31. doi: 10.1111/j.1365-294X.2009.04480.x

Tanaka, T., Kawasaki, K., Daimon, S., Kitagawa, W., Yamamoto, K., Tamaki, H., et al. (2014). A hidden pitfall in the preparation of agar media undermines microorganism cultivability. *Appl. Environ. Microbiol.* 80, 7659–7666. doi: 10.1128/AEM.02741-14

Tanaka, Y., Hanada, S., Manome, A., Tsuchida, T., Kurane, R., Nakamura, K., et al. (2004). Catellibacterium nectariphilum gen. nov., sp. nov., which requires a diffusible compound from a strain related to the genus Sphingomonas for vigorous growth. *Int. J. Syst. Evol. Microbiol.* 54, 955–959. doi: 10.1099/ijs.0.02750-0

Thomson, L. K., Fleming, S. D., Barone, K., Zieschang, J.-A., and Clark, A. M. (2010). The effect of repeated freezing and thawing on human sperm DNA fragmentation. *Fertil. Steril.* 93, 1147–1156. doi: 10.1016/j.fertnstert.2008.11.023

Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* 9, 68–80. doi: 10.1038/ismej.2014.117

Todorova, T., Pesheva, M., Stamenova, R., Dimitrov, M., and Venkov, P. (2012). Mutagenic effect of freezing on nuclear DNA of *Saccharomyces cerevisiae*. *Yeast* 29, 191–199. doi: 10.1002/yea.2901

Toju, H., Tanabe, A. S., Yamamoto, S., and Sato, H. (2012). High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. *PLoS ONE* 7:e40863. doi: 10.1371/journal.pone.0040863

van den Brink, P. J., den Besten, P. J., bij de Vaate, A., and ter Braak, C. J. (2009). Principal response curves technique for the analysis of multivariate biomonitoring time series. *Environ. Monit. Assess.* 152, 271–281. doi: 10.1007/s10661-008-0314-6

Vergin, K. L., Done, B., Carlson, C. A., and Giovannoni, S. J. (2013). Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. *Aquat. Microb. Ecol.* 71, 1–13. doi: 10.3354/ame01661

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Available at: http://www.stats.ox.ac.uk/pub/MASS4

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., and Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* 3:440. doi: 10.1186/s40168-015-0087-4

Walters, W. A., Caporaso, J. G., Lauber, C. L., Berg-Lyons, D., Fierer, N., and Knight, R. (2011). PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* 27, 1159–1161. doi: 10.1093/bioinformatics/btr087

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wang, Y., and Qian, P. Y. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE* 4:e7401. doi: 10.1371/journal.pone.0007401

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., et al. (2016). *gplots: Various R Programming Tools for Plotting Data.* Available online at: https://CRAN.R-project.org/package=gplots

Warwick, R. M., and Clarke, K. R. (1995). New "biodiversity" measures reveal a decrease in taxonomic distinctness with increasing stress. *Mar. Ecol. Prog. Ser.* 129, 301–305. doi: 10.3354/meps129301

Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J., and Knight, R. (2014). Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol.* 15:1704. doi: 10.1186/s13059-014-0564-2

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235

Westcott, S. L., and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487. doi: 10.7717/peerj.1487

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr.* 30, 279–338. doi: 10.2307/1943563

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51, 221–271.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5088–5090. doi: 10.1073/pnas.74.11.5088

Woese, C. R., Stackebrandt, E., Macke, T. J., and Fox, G. E. (1985). A phylogenetic definition of the major eubacterial taxa. *Syst. Appl. Microbiol.* 6, 143–151. doi: 10.1016/S0723-2020(85)80047-3

Xia, L. C., Ai, D., Cram, J., Fuhrman, J. A., and Sun, F. (2013). Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* 29, 230–237. doi: 10.1093/bioinformatics/bts668

Yang, B., Wang, Y., and Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 17:135. doi: 10.1186/s12859-016-0992-y

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A., and Elshahed, M. S. (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl. Environ. Microbiol.* 75, 5227–5236. doi: 10.1128/AEM.00592-09

Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E. J., Short, J. M., et al. (2002). Cultivating the uncultured. *Proc. Natl. Acad. Sci. U.S. A.* 99, 15681–15686. doi: 10.1073/pnas.252630999

Zhang, Y., Zhao, Z., Dai, M., Jiao, N., and Herndl, G. J. (2014). Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. *Mol. Ecol.* 23, 2260–2274. doi: 10.1111/mec.12739

Zumla, A., Al-Tawfiq, J. A., Enne, V. I., Kidd, M., Drosten, C., Breuer, J., et al. (2014). Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections–needs, advances, and future prospects. *Lancet Infect. Dis.* 14, 1123–1135. doi: 10.1016/S1473-3099(14)70827-8

Zuo, G., Xu, Z., and Hao, B. (2013). Shigella strains are not clones of *Escherichia coli* but sister species in the genus Escherichia. *Genomics Proteomics Bioinformatics* 11, 61–65. doi: 10.1016/j.gpb.2012.11.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.