A primer and discussion on DNA-based microbiome data and related bioinformatics analyses

Gavin M. Douglas[1] and Morgan G. I. Langille[2,*]

[1] *Department of Microbiology and Immunology*
[2] *Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada*
[*] *Email for correspondence: morgan.langille@dal.ca*

Abstract

The past decade has seen an eruption of interest in profiling microbiomes through DNA sequencing. The resulting investigations have revealed myriad insights and attracted an influx of researchers to the research area. Many newcomers are in need of primers on the fundamentals of microbiome sequencing data types and the methods used to analyze them. Accordingly, here we aim to provide a detailed, but accessible, introduction to these topics. We first present the background on marker-gene and shotgun metagenomics sequencing and then discuss unique characteristics of microbiome data in general. We highlight several important caveats resulting from these characteristics that should be appreciated when analyzing these data. We then introduce the many-faceted concept of microbial functions and several controversies in this area. One controversy in particular is regarding whether metagenome prediction methods (i.e. based on marker gene sequences) are sufficiently accurate to ensure reliable biological inferences. We next highlight several underappreciated developments regarding the integration of taxonomic and functional data types. This is a highly pertinent topic because although these data types are inherently connected, they are often analyzed independently and primarily only linked anecdotally in the literature. We close by providing our perspective on this topic in addition to the issue of reproducibility in microbiome research, which are both crucial data analysis challenges facing microbiome researchers.

Background

Microbial communities encompass most of the genetic and species-level diversity on Earth. These communities are commonly characterized through DNA sequencing, which can be used to identify the presence and relative abundance of microbes in a community. These communities, including both the microbes, their constituent genes, and metabolites, are referred to as microbiomes. Due to technological improvements and the reduced cost of sequencing, the number of sequenced microbiomes has substantially grown in recent years. For instance, in 2017 the Earth Microbiome Project published a meta-analysis of 23,828 sequencing samples from all seven continents (Thompson et al. 2017). These data represented 109 environmental groupings and 21 major biomes, such as animal secretions, saline water, and soil. A key goal of microbial ecology research is to robustly analyze and correctly interpret these and other such microbial profiles.

But is DNA sequencing the best method for characterizing microbial communities? It is commonly observed that microbiome research would benefit from more emphasis on culturing, which enables individual microbes to be isolated and precisely studied in the lab. Traditionally, microbial communities were difficult to study by culturing alone because the vast majority of environmental microbes, particularly bacteria, could not be grown under standard culturing conditions (Staley and Konopka 1985). This issue remains unresolved even after gradual improvements to standard culturing conditions; a recent evaluation of six major environments identified only 34.9% of bacteria as culturable under standard conditions (Martiny 2019). However, modified culturing conditions can largely resolve this problem. By systematically applying 66 different conditions it was demonstrated that 95% of bacterial species in human stool samples could be grown in the lab (Lau et al. 2016). Therefore, it is no longer true for human stool samples, and likely other environments as well, that the majority of constituent bacteria cannot be cultured.

Despite these advances, a clear remaining advantage of DNA sequencing is that it enables microbial communities to be characterized in place, which theoretically enables the exact community relative abundances to be profiled. In practice, biases during sample collection and sequencing library preparation can perturb microbial relative abundances (Jones et al. 2015; Bukin et al. 2019; Watson et al. 2019). But nonetheless, DNA sequencing provides a more accurate view of the relative abundances of the community members than would be possible from culturing alone. For this reason, DNA sequencing remains the predominant method for characterizing microbial communities, although it is well-complemented by culturing (Lau et al. 2016).

DNA sequencing data is typically analyzed to identify specific associations between individual features (e.g. individual microbes) and sample groups of interest. Most commonly, researchers are interested in identifying associations between disease states and the relative abundance of features. A similar goal is often to investigate whether different measures of diversity in the studied dataset are associated with the sample groups. These measures of diversity are divided into alpha and beta diversity (Goodrich et al. 2014). Alpha diversity metrics refer to within-sample measures, such as richness (i.e. the number of taxa), and the Shannon diversity index (or entropy), which incorporates both the abundance and evenness of taxa within a sample (Jost 2006). In contrast, beta diversity refers to metrics that summarize variation between samples, which is most often performed by metrics that take the presence and abundance of features into account, such as the Bray-Curtis dissimilarity metric (Goodrich et al. 2014). Other microbiome-specific metrics have also been developed, such as the weighted UniFrac distance, which also takes the phylogenetic distance between taxa into account (Lozupone and Knight 2005). There is often more statistical power to detect overall differences based on alpha and beta diversity metrics than to detect associations with individual features, but diversity-level insights are also less actionable (Shade 2017).

There are many sub-categories of DNA sequencing approaches for characterizing microbial communities. One key distinction is between approaches that aim to characterize taxa (i.e. a group of organisms) and those that characterize genes and pathways, referred to as functions, that could be active in the community. These data types are referred to as taxonomic and functional microbiome data, respectively. Biologically this dichotomy is counter-intuitive; clearly genes are encoded in the genomes of taxa. So why does this distinction exist?

The reason is entirely related to methodological challenges. The most common and cost-effective sequencing approach focuses on sequencing marker genes. This method provides no direct information on the genomes of sequenced microbes, and instead is used to profile taxa. In contrast, shotgun metagenomics sequencing (MGS) provides information on all DNA present in a sample. MGS data can be used for analyzing both taxonomic and functional

profiles. However, it is difficult to integrate the two data types, largely due to the complexity of microbial communities and the fragmented nature of DNA sequencing: it is relatively straight-forward to identify genes in MGS data but challenging to determine from which genomes they originated.

Herein we introduce the key forms of these data types and highlight important caveats that should be considered when they are analyzed. We first cover the fundamentals of microbiome data analysis, starting with marker-gene sequencing, and then move to recently developed tools that could be leveraged to conduct joint analyses of taxonomic and functional data types. We conclude by highlighting two important challenges that must be addressed in microbiome data analysis.

Marker-gene sequencing

The earliest developed and most common form of microbiome sequencing is marker-gene sequencing, also known as amplicon sequencing. Under this approach specific genes are PCR-amplified and then sequenced. There are two key requirements for robust marker genes. First, they must be encoded by all (or at least most) taxa of interest. Second, the observed sequence divergence between orthologs should be approximately equal to the neutral mutation fixation rate multiplied by double the divergence time between orthologs (Woese 1987). Note that the divergence time should be doubled because mutations could accumulate in either lineage since the organisms diverged. Genes displaying this second requirement have been referred to as molecular chronometers. This term highlights the close link between these marker genes and the concept of the molecular clock (Zuckerkandl and Pauling 1965): given equal mutation rates and equal fixation rates for neutral mutations, the number of neutral substitutions between organisms is directly proportional to the evolutionary divergence between them.

However, there are many reasons why a gene might be an unreliable molecular chronometer (Janda and Abbott 2007). One reason is that if a gene varies in function across taxa then contrasting selection pressures could result in different non-synonymous substitution rates (Wheeler et al. 2016). For instance, as previously observed (Woese 1987), the cytochrome complex gene is a useful molecular chronometer in eukaryotes, but suffers from drawbacks. This gene was shown to be useful for building early phylogenetic trees that represented both long evolutionary distances across eukaryotes and short distances between human populations (Fitch and Margoliash 1967). However, within prokaryotes the cytochrome complex systematically varies in size, which is believed to be due to positive selection (Ambler et al. 1979). Because positive selection is likely driving divergence between orthologous cytochrome complexes, in at least some cases it would be an invalid molecular chronometer to study in prokaryotes. Similarly, if a gene is sufficiently divergent between organisms then it can be difficult to accurately align residues. Misalignments lead to inaccurate estimates of evolutionary divergence, which is particularly true if the gene accumulates insertions and deletions. Such highly divergent regions, particularly in areas under no selective constraint, have been referred to as "evolutionary stopwatches" (Woese 1987), because they are useful only at short evolutionary distances. Therefore, to select a robust marker gene one should adhere in some ways to the Goldilocks principle: some nucleotide conservation is needed, but not too much.

The 16 Svedberg (16S) ribosomal RNA (rRNA) gene fits well with this principle. This gene features highly conserved regions surrounding nine less conserved regions (referred to as variable regions). It is also encoded by all prokaryotes and represents 50 helical RNA regions encoded by approximately 1,500 base-pairs (Woese et al. 1980). This high number of independent functional domains is valuable in a marker gene (Woese 1987). This is because if there are non-random substitutions within a single domain, but substitutions in the majority of other domains are driven by random processes, there would likely be little effect on estimates of evolutionary divergence. This gene also encodes a highly conserved function across both prokaryotes and eukaryotes (where it is called the 18S rRNA gene). The 16S rRNA molecule is part of the 30S small subunit (SSU) of the ribosome, which helps initiate protein synthesis by binding the Shine-Dalgarno sequence in messenger RNA (mRNA) to align the ribosome with the encoded start codon. Many changes in the highly conserved region of the 16S rRNA gene affect its binding affinity to the ribosome and mRNA. The strong negative selection acting against such substitutions makes these regions valuable for detecting rare substitutions between distant relatives, anchoring alignments of 16S rRNA genes, and for primer design (Wang et al. 2013).

Since the 16S rRNA gene was identified as a useful molecular chronometer, it has been the prime marker gene used to develop phylogenetic models of the tree of life. Most famously, an alignment of 16S (and 18S) rRNA gene sequences from across life lead to distinguishing archaea, bacteria, and eukaryotes into distinct domains (Woese and Fox 1977). In these

early days, research focused on analyzing the rRNA sequences of isolated microbes. This was painstaking work, as illustrated by the prediction in 1987 that future research groups could plausibly sequence on the order of one hundred 16S rRNAs a year (Woese 1987).

Thirty-four years later, through next-generation sequencing technology, insufficient availability of sequenced rRNA genes is no longer a common complaint. Databases such as SILVA contain enormous collections of sequenced SSU fragments; as of August 2020 SILVA contained 9,469,124 non-clustered, independent sequences (Quast et al. 2013). Software such as redbiom also enables unique 16S rRNA gene variants to be compiled from the growing number of 16S rRNA gene sequencing (hereafter referred to as 16S sequencing) studies (McDonald et al. 2019). These 16S datasets are produced to characterize and compare the relative abundances of prokaryotes across communities. However, despite the ubiquity of such datasets, they are non-trivial to analyze and interpret. There are numerous methodological reasons for this difficulty.

First, due to sequencing length constraints, only certain 16S rRNA gene variable regions are typically amplified and sequenced. Each variable region has particular strengths and limitations (Chen et al. 2019; Johnson et al. 2019). Along with our colleagues we have previously compared the biases between the amplified fragments from variable regions four and five and from regions six to eight (written as V4-V5 and V6-V8, respectively) on a mock community from the Human Microbiome Project (HMP) (Comeau et al. 2017). We found the V4-V5 region overrepresented Firmicutes and Bacteroides while drastically underestimating Actinobacteria. In contrast, the V6-V8 region overrepresented Proteobacteria and underrepresented Bacteroides. These biases highlight that choice of variable region can depend on which taxa are of interest. For example, region V4-V5 was recently shown to be superior to region V6-V8 for identifying archaea in the North Atlantic Ocean (Willis et al. 2019). In this case the authors were particularly interested in archaeal diversity so the V4-V5 region was more appropriate.

Typically, however, the taxonomic scope of interest and region biases in a particular environment are not clear and little or no rationale is given for the variable region selection. This is a problem, because analyses of the same communities with different variable regions can result in not only systematic biases in the raw data, but also in strikingly different biological interpretations. For example, key species that modulate human vaginal health are underrepresented or missing in V1-V2 sequencing datasets, such as *Gardnerella vaginalis*, *Bifidobacterium bifidum*, and *Chlamydia trachomatis* (Graspeuntner et al. 2018). Application of this region for profiling vaginal samples, instead of the more appropriate choice of the V3-V4 region, can result in entirely missing associations between vaginal health and the microbiome. Similarly, a comparison of the tick microbiome based on six sequenced 16S rRNA gene regions found a wide range of the number of prokaryotic families and in the Shannon diversity index for each individual tick (Sperling et al. 2017). The problem of such biases in variable region selection is beginning to recede as long-read technologies, such as that developed by Pacific Biosciences of California and Oxford Nanopore Technologies Limited, enable full-length 16S sequencing (Callahan et al. 2019; Johnson et al. 2019). However, it will remain an important issue for the foreseeable future as long as the microbiome is largely studied by short-read sequencing.

Regardless of the sequenced region, most reads originating from the same biological molecule will differ due to sequencing errors. Raw reads are either clustered based on sequence identity into operational taxonomic units (OTUs) or alternatively errors are corrected to produce amplicon sequence variants (ASVs). OTUs are typically clustered at 97% identity (Goodrich et al. 2014), which often results in merging different species into a single OTU (Mysara et al. 2017). This issue has long plagued 16S rRNA gene-based analyses. For instance, *Bacillus globisporus* and *Bacillus psychrophilus* are problematic cases because their 16S genes share 99.5% sequence identity, but are highly distinct at the genome level (Fox et al. 1992).

In contrast to clustering approaches, error-correcting approaches, referred to as denoising methods, theoretically can correct raw reads sufficiently well to produce exact biological molecules. Several different denoising approaches have emerged recently. DADA2 is the most sophisticated approach, which generates a different parametric error model for every input sequencing dataset (Callahan et al. 2016a). The raw sequencing reads are then corrected to generate ASVs based on this error model. Deblur and UNOISE3 are two other denoising tools that are based on rapidly clustering raw reads and using predetermined hard cut-offs related to the expected error rates to generate ASVs. We and other colleagues have evaluated the performance of these three tools and open-reference OTU clustering (which combines both de novo and reference-based clustering) and found that all three denoising methods result in

similar overall microbial communities (Nearing et al. 2018). In contrast, we found that open-reference OTU clustering resulted in a high rate of spurious OTUs compared to these methods. Nonetheless, there were important differences between the three methods, particularly in terms of richness and when profiling rare taxa (Nearing et al. 2018). A more recent independent validation based on a higher number of test datasets reached similar conclusions (Prodan et al. 2020).

In addition to 16S rRNA gene sequencing data, there are multiple marker genes appropriate for profiling eukaryotic diversity. As mentioned above, the 18S rRNA gene is the homolog of the 16S rRNA gene in eukaryotes and is widely used to profile that domain. However, fungi are more difficult to distinguish based on the 18S rRNA gene, because fungi lack several variable regions for this gene (Schoch et al. 2012). Instead, the internal transcribed spacer (ITS) region, although not strictly a marker gene, is more often amplified to study fungal communities, because it typically has more resolution to distinguish fungi than the 18S rRNA gene (Liu et al. 2015). This region is within the nuclear rRNA cistron of fungi genomes, which contains the 18S, 5.8S, and the 28S rRNA genes. The ITS regions encompasses the two intergenic regions, which have relatively high rates of insertions and deletions, and the 5.8S rRNA gene (Schoch et al. 2012). Only a single intergenic region is typically amplified, referred to as ITS1 or ITS2, which have better discriminatory resolution for the major phyla Basidiomycota and Ascomycota, respectively (Saroj et al. 2015).

Although the marker genes described above are the most commonly profiled loci, in many cases there are marker genes more appropriate for specific lineages. For example, several halophilic species of Haloarcula encode multiple 16S copies that can differ by more than 5% sequence identity within the same genome (Sun et al. 2013). Consequently, different marker genes are often used when building phylogenetic trees representing a single species or genera. The chaperonin-60 (*cpn60*) gene is one useful alternative prokaryotic marker gene, which is particularly useful for distinguishing taxa at resolutions below the genus level (Links et al. 2012). For example, the *cpn60* gene has been frequently profiled in vaginal microbiome samples, because variation at this locus can distinguish subgroups of *Gardnerella vaginalis* that cannot be distinguished based on the 16S rRNA gene alone (Jayaprakash et al. 2012). More generally, marker genes for specialized comparisons are often chosen to match the defining function of a given lineage. For example, the methyl coenzyme M reductase A (*mrcA*) gene and a nitrate reductase gene have been previously profiled to explore the diversity of methanogens (Hallam et al. 2003) and nitrogen-fixing microbes (Comeau et al. 2019), respectively.

Shotgun metagenomics sequencing

Shotgun metagenomics sequencing (MGS) is a qualitatively different method from marker-gene sequencing, because it involves sequencing all DNA in a community. This is a major advantage and means that MGS data can profile any taxa, including viruses and microbial eukaryotes. MGS approaches were first applied to study ocean water communities through a Fosmid cloning approach (Stein et al. 1996). Building upon such early studies, the potential for leveraging MGS was widely publicized by an investigation into the microbial diversity of the Sargasso Sea (Venter et al. 2004). This study identified 1.2 million previously unknown genes and many other microbial features that would be impossible to study with 16S rRNA gene sequencing. These and other related observations sparked an explosion of interest in profiling microbial communities with MGS approaches. This interest has culminated in the generation of enormous MGS datasets such as the ongoing work on the Earth Microbiome Project (Thompson et al. 2017) and the Human Microbiome Project (Lloyd-Price et al. 2017).

There are two main approaches for analyzing MGS data: read-based workflows and metagenomics assembly. Each of these approaches has strengths and weaknesses, but in both cases the generated profiles imprecisely reflect biological reality. For instance, the number of species identified by different read-based methods can vary by three orders of magnitude (McIntyre et al. 2017). The exact species relative abundances can also drastically differ across tools, as recently shown in a comparison of read-based methods applied to simulated datasets (Ye et al. 2019). Different approaches for metagenomic assembly will produce different assembled contigs and microbial profiles as well (Olson et al. 2019). Unsurprisingly, given this wide variation, there is also low concordance between 16S sequencing and MGS data taken from the same samples. For example, one comparison found that fewer than 50% of phyla identified in water samples based on 16S sequencing were also identified in the corresponding MGS profiles (Tessler et al. 2017). This wide variation in results highlights that any interpretation of MGS profiles, similar to 16S profiles, should be done cautiously. It

is crucial to appreciate that any approach will have important weaknesses and that the generated profile will only partially represent the actual microbial diversity.

With those important caveats in mind, an understanding of the different approaches is nonetheless important to give context to MGS data analysis. Read-based workflows involve little or no assembly of the reads and instead each read (or pair of reads) is treated independently. This is the most common approach for analyzing MGS data, particularly because it can be performed with low sequencing depth (Hillmann et al. 2018) and in complex communities (Zhou et al. 2015). However, an important disadvantage of this approach is that taxonomic and functional annotations are typically generated and treated as entirely independent data types (Figure 1a). It is also possible to map reads against a set of known reference genomes, which does link the two data types (Figure 1b). Although this is an invaluable approach when applied to genomes assembled from the study environment (see below), the results are typically near incomprehensible when reads are mapped against a database of thousands of genomes. Instead, the most common approach for generating taxonomic profiles is either based on a marker-gene or k-mer method.

Marker-gene approaches are based on the insight that specific genes can be used to identify the presence and relative abundance of certain taxa. An extreme example is to use solely the 16S rRNA gene for taxonomic classification (Hao and Chen 2012). More commonly, marker-gene approaches base classifications on many genes. For instance, PhyloSift (Darling et al. 2014) leverages 37 nearly universal prokaryotic marker-genes (Wu et al. 2013) in addition to eukaryotic and viral gene sets to make a combined set of approximately 800 (mainly viral) gene families for classification. Aligned reads are placed into a phylogenetic tree of reference sequences and taxonomic classification is performed based on summing the likelihood of each taxa based on each read placement (Darling et al. 2014). MetaPhlAn2 is a contrasting approach that instead bases taxonomic predictions on the presence of clade-specific marker genes, which are genes only found in that given lineage, and found in all members (Truong et al. 2015). This method has rapidly become the most popular marker-gene MGS approach. However, given that this approach is limited by the existence of robust clade-specific genes, it is not surprising that it tends to have low sensitivity (Tessler et al. 2017; Miossec et al. 2020), meaning that it misses taxa that are actually present.

In contrast, k-mer-based approaches are much more sensitive but have slightly lower specificity than marker-gene methods (Miossec et al. 2020). These approaches search for exact matches of short DNA sequences (k-mers) within reference genomes. An algorithm such as lowest-common ancestor is then performed to determine the likely taxonomic classification based on all matching genomes. Two common kmer-based approaches are kraken2 (Wood et al. 2019) and centrifuge (Kim et al. 2016), both of which match k-mers against a compressed database of reference genomes. In contrast to the marker-gene results, the main challenge of analyzing taxonomic profiles output by these methods is the high number of rare taxa of different ranks identified, some of which may be false positives. Summarizing the output profiles with an additional approach, such as the Bayesian abundance re-estimation tool Bracken (Lu et al. 2017) in the case kraken2 data, can help partially mitigate this problem.

Most functional read-based methods are based on a similarity search of reads against a database of known gene families. This is primarily done in protein space, because protein similarity matches are more informative and the database requirements are lower (Koonin and Galperin 2003). The common similarity searching tool BLASTX is prohibitively slow when scanning millions of reads, which has driven the development of faster alternatives like DIAMOND (Buchfink et al. 2015) and MMseqs2 (Steinegger and Söding 2017). These faster alternatives are leveraged by workflows implemented in software such as MEGAN (Huson et al. 2007) and HUMAnN2 (Franzosa et al. 2018) to identify gene family matches and output overall metagenome profiles. HUMAnN2 is a unique approach in that it first screens reads that map to reference genomes of taxa identified as present with MetaPhlAn2. This step enables a small subset of gene families to be linked directly to particular taxa. However, the vast majority of gene families typically have no taxonomic links and are only part of the community-wide metagenome. There are clear issues with the general approach implemented by these gene profiling approaches, as has been previously observed: "genes are expressed in cells, not in a homogenized cytoplasmic soup" (McMahon 2015).

Linking functional annotations to specific taxa by assembling raw reads is the ideal approach to resolve this problem, but this too comes with caveats. Most importantly, insufficiently high read depth, which depends on the complexity of a sample, can result in too few assembled contigs to sensibly analyze. Nonetheless, with sufficiently high read depth metagenome assembly can be a valuable way to

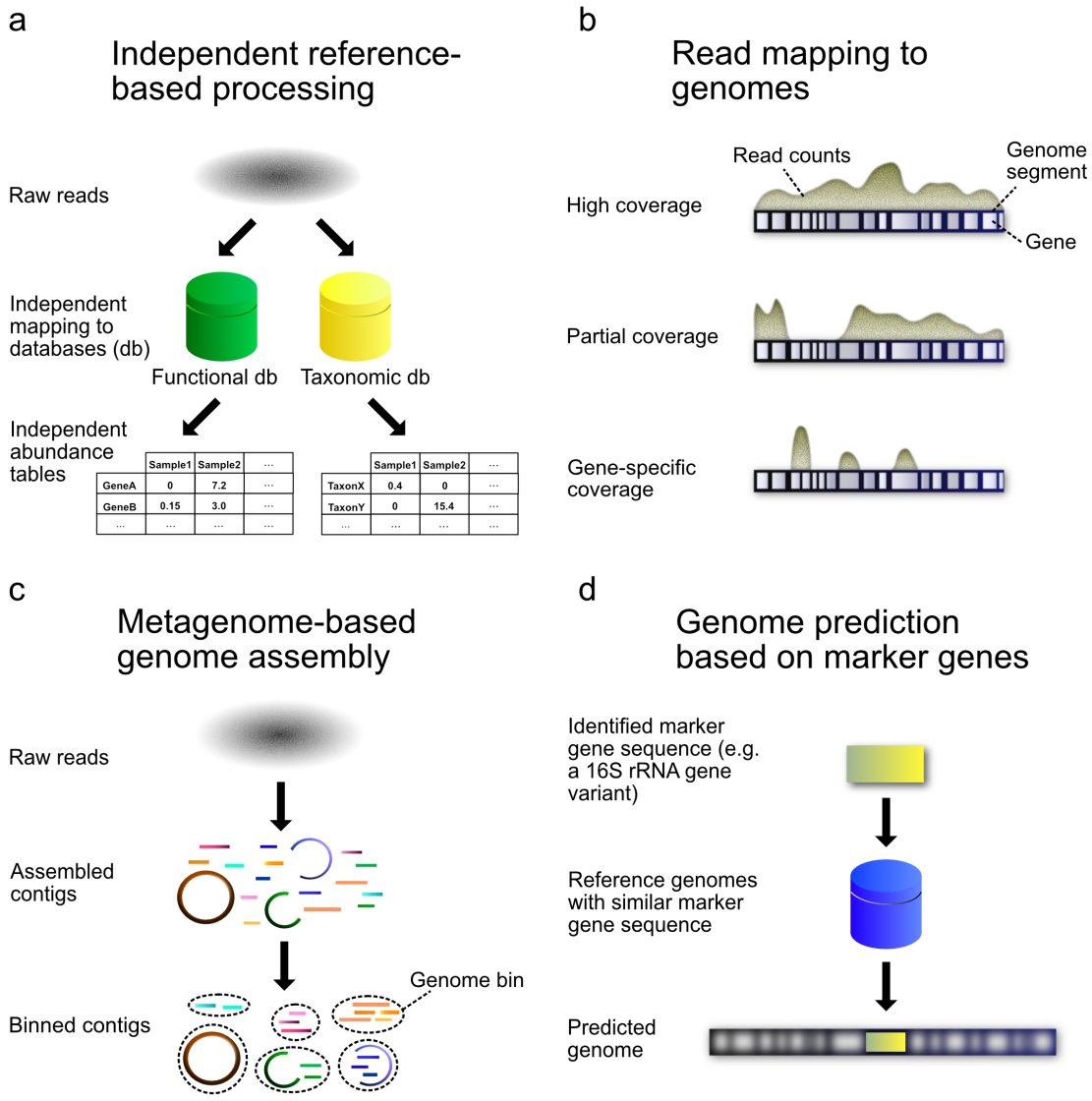

Figure 1: Key approaches for generating joint taxonomic and functional data from microbiome sequencing data. (a) Read-based processing of shotgun metagenomics data to generate functional and taxonomic abundance tables independently. (b) Read mapping to genome sequences can be used to infer the presence of a taxon based on read coverage. It can also be used to identify the presence of strains missing specific genes or of the inverse: a community containing specific genes from a genome while the rest of the genome is absent. Note that all of these inferences are best made in low complexity communities where there are few ambiguous read mappings, and where the possible set of genomes present is relatively well defined. This is particularly applicable when mapping reads against metagenome assembled genomes from the same dataset. (c) Metagenomics-based genome assembly involves assembling reads into contigs and then binning contigs into categories representing metagenome-assembled genomes. Missing from this diagram is the important quality control step, which is essential to follow-up metagenomics assembly. Also, this approach is best for profiling dominant organisms, and produces the best results when sequencing read depth is high and/or community complexity is low. (d) Genome prediction based on marker gene sequences is another method of producing joint taxa and function profiles, which in this case are explicitly linked, similar to assembling genomes. However, these approaches are highly biased towards the specific reference genomes used for prediction. In addition, they can only predict genome content to the level at which the chosen marker gene differs between closely related taxa. This is a major limitation as many strains of bacteria with highly divergent genome content have identical marker gene sequences.

leverage information about microbial communities (Figure 1c). There are many metagenome assembly tools available, such as MetaSPAdes (Nurk et al. 2017) and Megahit (Li et al. 2015). The resulting assembled contigs from these approaches are typically categorized (or "binned") into groups of contigs with similar characteristics. This binning is primarily performed by identifying contigs that are found at similar relative abundances across samples and/or that contain similar proportions of different k-mers (particularly 3-mers) (Ayling et al. 2019). These bins represent metagenome-assembled genomes (MAGs) that must undergo stringent checks to help evaluate the overall quality (Bowers et al. 2017). The key method for performing quality control on these genomes is to scan for known universal single-copy genes (USCGs), with a tool such as CheckM (Parks et al. 2015). The percentage of USCGs present provides an estimate of overall genome completeness. In contrast, the number of USCGs found in multiple copies can be used to calculate the redundancy, which is potential evidence for contamination in the genome.

Characteristics of microbiome count data

Regardless of the sequencing technology and workflow used for profiling a microbial community, the final product is typically a count table. This is true for many sequencing approaches, such as RNA sequencing, but there are several important differences. First, unlike in the case of RNA sequencing where there are typically a known number of genomic loci, novel taxa and functions are frequently identified in microbiome data. For instance, novel OTUs, ASVs, and contigs are frequently identified in taxonomic analyses. Similarly, 25-85% of proteins in MGS are novel microbial genes of unknown function (Prakash and Taylor 2012). Second, no statistical distribution fits microbiome data in all contexts. For example, many statistical distributions, including the negative binomial (Love et al. 2014), beta binomial (Martin et al. 2020), and Poisson (Faust et al. 2012) distributions have been proposed as appropriate fits to microbiome data. However, upon analysis with real data these and other distributions fit with inconsistent accuracy (Weiss et al. 2017; Calgaro et al. 2020). Last, microbiome count tables typically have high sparsity, meaning that there is a high proportion of features not found across many samples (Thorsen et al. 2016). These characteristics make microbiome data analysis challenging for all taxonomic analyses and most functional analyses (see Microbial functions section).

These challenges are exacerbated by the inherent compositionality of sequencing data. Compositional data refers to data that is constrained to an arbitrary constant sum (Aitchison 1982), such as the arbitrary number of raw sequencing reads output per sample. This characteristic means that the observed abundance of any given feature is dependent on the observed abundance of all other features. A simple example can help illustrate the implications of this characteristic. Imagine a microbe, microbe x, at low relative abundance in sample a and at high relative abundance in sample b. An observer might naively infer that there is more of microbe x in sample b than in sample a. However, there are many reasons this could be false. For instance, the absolute abundance of microbe x could be the same in each sample but the abundance of other microbes in general might be higher in sample a. This higher total microbial load would push the relative abundance of microbe x in sample a down. Depending on the total microbial cell count it is even possible that the absolute abundance of microbe x could be higher in sample a than in sample b, but that the relative abundance is simply lower. This example highlights a necessary consideration regarding microbiome sequencing data analysis: it only provides information on the relative abundances, or percentages, of features and does not provide insight on feature absolute abundances.

This important characteristic was not widely appreciated in the field until relatively recently, when researchers identified fatal issues with common approaches for analyzing microbiome data (Gloor et al. 2016, 2017). Standard differential abundance approaches, such as the t-test and Wilcoxon test, when applied to relative abundances, and microbiome-specific tools such as LEfSe (Segata et al. 2011) do not account for this compositionality. Common summary metrics for microbiome data, such as the UniFrac distance, also suffer from this problem (Gloor et al. 2017). This is a major issue, because ignoring this characteristic is known to lead to spurious discoveries with compositional data (Aitchison 1982; Jackson 1997; Fernandes et al. 2014).

Fortunately, there is active work in the field to resolve this issue and numerous compositional approaches have been developed. The focus has primarily been on developing novel correlation (Friedman and Alm 2012; Kurtz et al. 2015; Schwager et al. 2017) and differential abundance approaches, such as ALDEx2 (Fernandes et al. 2013, 2014) and ANCOM (Mandal et al. 2015). A common theme of these compositional approaches is that the data

is transformed based on the ratio of feature relative abundances to some reference frame (Aitchison 1982; Morton et al. 2019). This choice of reference frame varies substantially between approaches. For instance, ALDEx2 transforms relative abundances by the centred log-ratio (CLR) transformation (Fernandes et al. 2013), which essentially normalizes feature relative abundances by the mean relative abundance per sample. This approach transforms the original data but maintains the interpretation of individual features. In contrast, it has been suggested that analyses could instead be based on ratios between features (Morton et al. 2019), which converts the data type into comparisons of features rather than individual features.

There are no best-practices regarding approaches that compositionally transform individual features. More generally, differential abundance tests commonly produce widely different sets of significant taxa from each other (Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019). This wide variation is largely due to specific characteristics of microbiome count data. A large proportion of the variation in results is driven by high false discovery rates. Although many methods advertise that only approximately 5% of significant taxa are likely false positives, it has been estimated that for some methods the actual false discovery rate is substantially higher (Hawinkel et al. 2019). This particular validation observed this trend for several methods, including ANCOM (Mandal et al. 2015) and metagenomeSeq (Paulson et al. 2013), two microbiome-oriented methods that are otherwise considered conservative (Paulson et al. 2013; Weiss et al. 2017). In addition, a recent evaluation of differential abundance tools found that compositional methods are actually less robust than several non-compositional alternatives (Calgaro et al. 2020).

Given this wide variation in differential abundance tool performance and unclear best-practices, how is a microbiome researcher to proceed? One possible answer is that a change in expectations regarding the interpretability of microbiome data analysis is needed. In particular, analyses using ratios between the relative abundances of taxa has been shown to be robust, although it comes at the cost of interpretability (Morton et al. 2019). However, an important issue is how to determine which taxa should be the numerator and denominator of each ratio. One solution is to leverage the bifurcating structure of a clustered tree (Egozcue and Pawlowsky-Glahn 2011; Morton et al. 2017) or phylogenetic tree (Silverman et al. 2017) of features. Analyses can be focused on the ratios in relative abundances between features on the left-hand and right-hand of each node in the tree. Despite the potential of this approach, it is rarely used for standard microbiome analyses because it is unclear how to biologically interpret any differences in the values of these ratios across samples.

This discussion of microbiome data characteristics has focused on taxonomic features based on either 16S sequencing or read-based MGS data analysis. However, it is important to emphasize that count tables produced from MAGs do not resolve this issue. In fact, attempting to account for these challenging characteristics of microbiome count data and the links between taxa and function makes the analysis more difficult.

Microbial functions

To this point we have only discussed functional microbiome data in vague terms as referring to microbial gene abundances. When based on DNA sequencing data this information summarizes the functional potential, meaning the functions that are present, but not necessarily active in a community. However, rather than individual gene sequences, research is typically focused on gene families, which are gene clusters. Alternatively, the focus is sometimes on higher-order functional categories like pathways. To complicate matters further, there are several different functional ontologies, which are different frameworks for studying functions at different resolutions. Depending on which of these functional ontologies and sub-categories are analyzed, the characteristics of the data can drastically differ.

The Universal Protein Resource (UniProt) Reference Clusters (UniRef) database contains all protein sequences from the Swiss-Prot (manually curated) and TrEMBL (automated) databases clustered at either 50%, 90%, or 100% identity (Apweiler et al. 2004). The most recent versions of these clusters have been generated with the MMseqs2 algorithm (Steinegger and Söding 2018). As of June 30th, 2020, the 100% identity clusters (called UniRef100), corresponded to 235,561,514 unique protein sequences, which provides a detailed summary of almost all known protein sequences. Despite being clustered at lower identity thresholds, UniRef50 and UniRef90 nonetheless contain enormous numbers of protein clusters: 41,883,832 and 115,885,342, respectively.

The UniRef database contrasts with another common functional ontology, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto 2000; Kanehisa et al. 2016). KEGG is based on 23,530 individual gene families (as of

September 10th, 2020), which are called KEGG orthologs (KOs). The advantage of KOs is that the majority have well-described molecular functions that can be linked to higher-order KEGG pathways and modules. Accordingly, any analysis of KEGG data will likely result in less sparse count tables than the corresponding UniRef-based database, simply because KOs are shared across more taxa than UniRef clusters.

To illustrate this point, we and our colleagues have previously compared the taxonomic coverage of each function within these two functional ontologies and each sub-category (Inkpen et al. 2017). We found that all UniRef functions, including those in UniRef50 clusters, are on average found in a single domain and encoded by fewer than four species. In contrast, we found that KOs were encoded in 1.3 domains and 184.3 species on average. Similarly, the high-level KEGG modules and pathways were predicted to be potentially active in a mean of 1.7 and 2.5 domains and 671 and 1267.6 species, respectively (Inkpen et al. 2017). Based on these statistics, clearly a shift in the abundance of a UniRef cluster should not be treated the same as a KEGG function: the former corresponds to the activity of a small number of species while the latter could correspond to a large assemblage. This example highlights that the choice of how function is defined in a given analysis can have profound effects on the biological interpretation.

In addition to UniRef and KEGG, several other functional ontologies have been leveraged for microbiome analyses. Key examples of additional function types include: Clusters of Orthologous Genes (COGs) (Tatusov et al. 2000; Makarova et al. 2015), Enzyme Commission (EC) numbers, Protein families (Pfam) (Punta et al. 2012; Finn et al. 2014), and TIGRFAMs (Haft et al. 2003). These categories represent a range of approaches for defining gene families and functional categories.

The COG strategy for functional annotation was originally intended to phylogenetically classify proteins into groups of orthologs (Tatusov et al. 2000). This one-to-one approach of matching individual orthologs has now been expanded to allow for more complex relationships between genes, such as paralogs and horizontally transferred homologs (Makarova et al. 2015; Galperin et al. 2019). As of 2015, there were 4,631 independent COGs (Galperin et al. 2015). The COG framework is similar to that of the eggNOG database (Jensen et al. 2008), which is a more high-throughput, automated approach. However, the key advantage of the COG database is that orthologous genes are clustered into 26 interpretable functional categories, which are expanded from categories originally defined to functionally bin *Escherichia coli* genes (Riley 1993).

The EC number framework, which was developed in 1992 by the "International Union of Biochemistry and Molecular Biology", is a contrasting approach for functional annotation. Instead of focusing on orthologous genes, EC numbers specify particular enzyme-catalyzed reactions. An interesting characteristic of this database is that these reactions can be performed by non-homologous isofunctional enzymes (Omelchenko et al. 2010). As of August 12th, 2020, there were 6,520 EC numbers, which correspond to one of four levels of granularity. For example, the accession EC 3.5.1.2 corresponds to glutaminases, while the higher-level categories correspond to hydrolases (3.-.-.-), that act on carbon-nitrogen bonds other than peptide bonds (3.5.-.-), and that are in linear amides (3.5.1.-). One major advantage of EC numbers is that because they specify exact enzymatic reactions they are straight-forward to link into pathway ontologies based on reactions, such as MetaCyc pathways (Caspi et al. 2013).

The Pfam database categorizes protein families, which are protein regions that share sequence homology (Punta et al. 2012). Individual proteins with multiple domains can thus belong to multiple Pfam families. Each Pfam family is represented by a hidden Markov model (HMM), which models the likely amino acids at each residue and the likely adjacent amino acids based on curated alignments of representative protein sequence. This approach identified homologous protein regions, which are often hypothesized to have a shared evolutionary history, but not necessarily. As of May 2020, there were 18,259 Pfam families.

Lastly, TIGRFAMs are manually curated protein families, which are also identified based on HMMs, but also additional pertinent information (Haft et al. 2003). As of September 16th, 2014, there were 4,488 TIGRFAMs. The distinguishing feature for this database is that different information supplements each HMM. For instance, certain TIGRFAM are annotated based on species metabolic context and neighbouring genes, while others are based on validated functions from the scientific literature. This database has been less commonly analyzed in recent years and is best known as the annotation system for early large-scale metagenomics projects (Venter et al. 2004). Alternative approaches, such as the FIGfam protein database are now more commonly used than TIGRFAMs. FIGfams are based on a similar approach, but instead of being manually curated they are aggregated into isofunctional groups

based on shared roles in specific subsystems (Meyer et al. 2009).

A recurrent question thus far has been that given a range of comparable, or contrasting, bioinformatics options, how is one to proceed? Fortunately, in the case of selecting functional ontologies, the choice is much clearer than other bioinformatics areas. Each functional database typically excels for different purposes. For instance, UniRef is useful for identifying uncharacterized genes that may be of interest in an environment, but quickly becomes challenging to interpret and analyze in diverse communities.

In contrast, KEGG is useful for looking for shifts in well-described functions at a high level, which means this database is more robust to granular functional diversity. Due to also being more robust to granular functional diversity and because they are more interpretable, pathway-level functions are often of particular interest. For instance, obesity is associated with an enrichment of phosphotransferase systems involved in carbohydrate processing in human and mouse gut microbiomes (Turnbaugh et al. 2008, 2009). This straight-forward explanation quickly communicates the pertinent biological details, which might be lost by focusing on less granular levels.

However, it is worth noting that pathways identified based on DNA sequencing are merely theoretical reconstruction based on the identified individual gene families. Although there are several pathway reconstruction approaches, they all require some mapping from gene families or reactions to pathways. This mapping can be structured, meaning that optional and required contributors can be specified, or non-structured, meaning that all genes and/or reactions are treated equally.

The naïve approach for pathway reconstruction is to assume that a pathway is present if any gene or reaction involved is present in the community. This was the predominant approach used for pathway inference in early functional analyses (Moriya et al. 2007; Meyer et al. 2008) and in several pathway inference tools such as PICRUSt (Langille et al. 2013). Pathway abundance under this framework is calculated by summing the abundance of each contributing gene family. This approach errs towards avoiding missing the presence of a pathway, which is a concern in metagenomes as key genes may be missing due to mis-annotations. However, this approach comes at the cost of spurious annotations. Based on the naïve mapping approach the human genome was previously annotated as including the KEGG pathway equivalent of the reductive carboxylate cycle (Ye and Doak 2011). This pathway is restricted to autotrophic microbes and is similar to reversing the Krebs cycle. Consequently, several gene families are shared in both processes. Under the naïve mapping approach, the presence of genes involved in the Krebs cycle are also evidence for the predicted presence of this atypical microbial pathway in humans. Similarly, vitamin C biosynthesis would also be predicted in humans based on the naïve approach (Ye and Doak 2011). However, the *GLO* gene, which encodes the protein involved in the key last step of vitamin C biosynthesis in mammals, is pseudogenized in humans (Drouin et al. 2011), which makes vitamin C biosynthesis impossible.

The Minimal set of Pathways (MinPath) approach is an approach developed to address this issue (Ye and Doak 2011). This tool identifies the smallest set of pathways, based on maximum parsimony, that are required to explain the presence of a set of proteins. In this way, the approach is more conservative than naïve mapping and also accounts for incomplete protein sets. This method has been applied in numerous contexts, including for the "HMP Unified Metabolic Analysis Network 2" (HUMAnN2) (Abubucker et al. 2012; Franzosa et al. 2018) MGS gene family profiling and pathway reconstruction framework. This popular framework reconstructs pathways based on MinPath and infers pathway abundance based on different approaches, depending if the pathway mapping is structured. For unstructured mappings, the arithmetic mean of the upper half of individual gene family abundances is taken to be the pathway abundance (Abubucker et al. 2012). For structured mappings, the harmonic mean of the key (i.e. required) genes families is computed for pathway abundance (Franzosa et al. 2018). Both these approaches are motivated by the need to be robust to variable abundance in alternative gene families.

Although this approach for MGS pathway reconstruction is commonly performed, it is important to emphasize that it has not been universally accepted and there remains disagreement about best-practices. For example, "Evidence-based Metagenomic Pathway Assignment using geNe Abundance DAta" (EMPANADA) is a method that addresses the same issue as MinPath and HUMAnN2 in a different way (Manor and Borenstein 2017a). This method focuses pathway reconstruction on distinguishing genes that are shared with multiple pathways from those that are unique to a single pathway. Pathway support weightings are first given by the average abundance of gene families unique to each given pathway. The abundance of all shared gene families is then partitioned between all pathways according to their rel-

ative support values. Pathway abundances are then taken as the sum of the unique gene family relative abundances and the partitioned relative abundances of the shared gene families (Manor and Borenstein 2017a).

The exact reconstructed pathways and their respective abundances differ depending on whether naïve mapping, MinPath/HUMAnN2, or EMPANADA are used. Validating pathway reconstructions is challenging without a gold-standard comparison, particularly in metagenomes. Even in isolated genomes, as demonstrated by the above examples of the human pathway reconstructions, pathway reconstruction is non-trivial. However, the advantage in these cases is that experimental validation of pathway reconstructions is possible (Francke et al. 2005; Oberhardt et al. 2008). Such validations would be possible if predictions are based on individual members of a microbiome, but it is less clear what experiments could validate pathways predicted for an overall community. In MGS data pathways are typically inferred as though all gene families were free to interact with each other. In other words, they are inferred as though there was universal cross-feeding. All three approaches described above are intended to be used for such community-wide gene family profiles. However, as mentioned above, this assumption is invalid because clearly not all proteins and metabolites in the microbiome can freely interact (McMahon 2015). The implications of this assumption being invalid remain unclear, but nonetheless it is an important caveat when interpreting pathway reconstruction data based on community-wide MGS data.

This section would be incomplete without addressing the most common discussion regarding microbiome functional data: its ostensible high stability. Functional pathways are commonly at similar relative abundances across the same sample-types whereas taxonomic features, such as phyla, can substantially vary (Turnbaugh et al. 2009; Burke et al. 2011; HMP-consortium 2013; Louca et al. 2016). This functional consistency is often taken to be evidence of environmental selection for particular microbial functions (Turnbaugh et al. 2009; Louca and Doebeli 2017). However, the validity of comparing variation between these two data types is rarely discussed. We and our colleagues investigated this question from a philosophical perspective and concluded that any meaningful comparison of the relative variation between taxonomic and functional profiles is likely impossible (Inkpen et al. 2017). This difficulty is largely because it is unclear which levels of granularity would be meaningful to compare between each data type. In other words, each data type is qualitatively different from the other and the choice of how to compare the two is based on arbitrary decisions.

For instance, as discussed above, the sparsity and number of possible functional categories differs drastically across ontologies and sub-categories. We demonstrated how observations of functional and taxonomic stability are entirely dependent on how function and taxa are defined (Inkpen et al. 2017). We did this by comparing human stool sample profiles at each possible taxonomic rank and also each functional level for both the KEGG and UniRef functional ontologies. As expected, phyla were less stable across the samples than KEGG pathways, but more stable than UniRef50 protein clusters. However, this area remains an area of active debate. Others have also argued that taxonomic variability never unambiguously reflects functional variation, which they believe is strong evidence for functional conservation (Louca et al. 2018a). Nonetheless, this example demonstrates once again the common theme throughout this section: "function" has many meanings.

Metagenome prediction methods

Ideally, analyses of microbial functions are based on MGS data. However, predicted functions based on 16S rRNA gene (hereafter 16S) sequencing are often analysed instead. Metagenome prediction, predicting complete genomes for each individual ASV or taxon weighted by their relative abundance, when based on 16S data is much cheaper than performing MGS.

There are additional advantages of predicted metagenomes over actual MGS data. Namely, MGS is often prohibitively expensive for samples where host DNA overwhelms microbial DNA. The high read depths required to yield sufficient microbial read depths is infeasible in many cases (Gevers et al. 2014). Similarly, low-biomass samples are difficult to accurately quantify with MGS, but they can be profiled with PCR-based 16S sequencing. For example, applying MGS to profile human tumours is currently infeasible, but it is straight-forward to apply 16S sequencing (Nejman et al. 2020). In both cases, for host DNA contaminated and low-biomass samples, metagenome prediction based on 16S profiles is a useful alternative to MGS.

However, metagenome prediction suffers from important drawbacks. The key problematic assumption is that the marker gene used for predictions, typically

the 16S, is strongly associated with genome content. This broad assumption is correct: genera such as *Lactobacillus* and *Desulfobacter* can be easily distinguished based on the 16S and they are enriched for extremely different functions. Namely, *Lactobacillus* can often perform lactic acid fermentation whereas *Desulfobacter* can typically oxidize acetate to CO_2. Such comparisons of characteristic functions between distantly related taxa are uncontroversial. The difficulty arises when approaches attempt to predict entire genome contents for an entire community, including for closely related taxa.

This issue is highlighted by classic DNA hybridization experiments (Mandel 1966; Brenner 1973). These experiments were based on mixing single-stranded DNA from two organisms and recording the melting temperature required to separate the strands. Higher melting temperatures are required to break apart DNA that shares more complementary bases connected by hydrogen bonds. Accordingly, this approach provides a rough estimate of the genetic distance between different strains or species.

An early comparison of these genetic distances with 16S dissimilarity across 34 bacteria computed a linear correlation of 0.728 (Devereux et al. 1990). However, the relationship between these two metrics is not linear: many bacteria with highly similar 16S genes have hybridization rates much lower than 70% (Stackebrandt and Goebel 1994), which is the traditional cut-off for delineating species. This trend has been corroborated across diverse prokaryotes (Hauben et al. 1997, 1999; Kang et al. 2007). In addition, a meta-analysis of 16S gene sequencing and DNA hybridization data from 45 bacterial genera further clarified these observations (Keswani and Whitman 2001). This analysis established that 78% of the variability in hybridization rates could be accounted for by 16S similarity, based on a non-linear model. However, they also identified that a minority of hybridization rates were extremely poorly predicted by 16S similarity (Keswani and Whitman 2001).

These observations agree well with genomic comparisons of strains, which can drastically differ in genome content. For example, across 17 *E. coli* genomes there are 13,000 genes that are variably distributed and only 2,200 core genes (Rasko et al. 2008). This enormous range of genomic variation is not reflected at the 16S level, where *E. coli* strains are typically >99% identical (Suardana 2014). These genomic differences can translate to enormous variation at higher taxonomic levels as well. For instance, a comparison of the genomes from 11 *Yersinia* species found a range of genome sizes from 3.7 - 4.8 megabases (Chen et al. 2010). A closer comparison of three pathogenic species of *Yersinia* determined that they shared 2,558 protein clusters while 2,603 were variably distributed. These species-level differences are also not proportionally reflected by divergence in *Yersinia* species 16S genes, which are typically >97% identical (Ibrahim et al. 1993). These examples highlight that 16S similarity can be a poor predictor of genomic similarity. This issue is compounded when there are divergent 16S copies within the same genome, although typically these are >99.5% identical (Větrovský and Baldrian 2013).

Variation in gene content within a taxonomic lineage is a recurrent observation across microbial communities. Variably present genes are often linked to putative niche-specific adaptations (Wilson et al. 2005), such as genes affecting antibiotic resistance (Kallonen et al. 2017), carbohydrate catabolism (Arboleya et al. 2018), and wound healing (Kalan et al. 2019). Based on these and other observations, the understanding of bacterial genomic content has shifted from that of a static genome to a pan-genome, consisting of core and variable genes (Tettelin et al. 2005). Variably present genes are transmitted between genomes through horizontal gene transfer, which typically occurs between closely related organisms (Popa and Dagan 2011). However, horizontal gene transfer can also occur between distantly related organisms, such as between different bacterial phyla (Beiko et al. 2005; Kloesges et al. 2011; Martiny et al. 2013).

The high variability between bacterial genomes and extensive horizontal gene transfer highlights the major challenges facing metagenome prediction. Despite these challenges, interest in performing metagenome predictions has continued, supported by several observations. First, although there are important outliers, 16S sequence identity does logarithmically correlate well with the average nucleotide identity between genomes, with an R^2 of 0.79 (Konstantinidis and Tiedje 2005). Second, 16S sequence similarity does provide some information on the ecological similarity of bacteria (Chaffron et al. 2010). This was demonstrated by the fact that co-occurring environmental bacteria are more likely to have similar 16S sequences. In addition, overall differences in inferred KEGG pathway potential are strongly associated with 16S divergence (Chaffron et al. 2010). Last, within a given environment, such as the human gut, 16S divergence was shown to be particularly predictive of divergence in average gene content (Zaneveld et al. 2010).

Originally, metagenome prediction workflows were based on matching 16S sequences to reference

genomes. By taking the best matching genome or averaging across genomes with similar sequences, a predicted genome annotation can be acquired for all 16S sequences (Figure 1d). To infer the metagenome profile one must simply multiply the predicted genome annotations for each 16S sequence by the abundance of each 16S sequence in the metagenome. In addition to predicting microbial functions linked to Crohn's disease (Morgan et al. 2012), this approach has also been used to profile diet-related microbial functions across mammals (Muegge et al. 2011) and the functions of invasive bacteria within corals (Barott et al. 2012). Although bioinformatics tools for metagenome prediction are now typically used for performing this task, this 16S-matching approach is still used for custom analyses (Verster and Borenstein 2018; Bradley and Pollard 2020).

The first metagenome prediction tool to expand beyond this approach, and specifically intended for 16S sequencing data, was "Phylogenetic Investigation of Communities by Reconstruction of Unobserved States" (PICRUSt1) (Langille et al. 2013). This tool is based on leveraging classical ancestral-state reconstruction methods, which have been widely used in phylogenetics (Zaneveld and Thurber 2014). The crucial extension of this framework is to extend trait predictions from internal, or ancestral, nodes in a phylogenetic tree to tips with unknown trait values. This approach has been termed hidden-state prediction (HSP) (Zaneveld and Thurber 2014). We recently published a major update to PICRUSt, called PICRUSt2 (Douglas et al. 2020). The key improvement in PICRUSt2 is that predictions can be made for novel 16S sequences with this tool and custom databases can be more easily used for analyses.

PICRUSt1 introduced the step of normalizing relative abundances by the predicted number of 16S copies within each genome, which is intended to control biases in 16S sequencing due to copy number (Farrelly et al. 1995). Importantly, although 16S copy number correction has become a common step for metagenome prediction (Angly et al. 2014), accurately predicting 16S copy number is particularly challenging. An independent validation of several 16S copy number prediction methods, including PICRUSt1, identified poor agreement of predicted copy numbers against existing reference genomes (Louca et al. 2018b). In some cases, less than 10% of the variance in actual 16S copy number was explained by these predictions. In addition, these predictions were often only slightly correlated between prediction methods.

Since PICRUSt1 was published a number of similar metagenome prediction tools have been developed. All of these approaches aim to capture the shared phylogenetic signal in the distribution of functions across taxa. These tools include: PanFP (Jun et al. 2015), Piphillin (Iwai et al. 2016; Narayan et al. 2020), PAPRICA (Bowman and Ducklow 2015), and Tax4Fun2 (Wemheuer et al. 2020).

These metagenome prediction tools have primarily been validated by comparing how well the predicted gene family abundances they output correlate with the abundances of gene families identified in MGS data from the same samples. This approach generally identifies high correlations between the two profiles. For example, predicted KOs output by PICRUSt1 based on Human Microbiome Project (HMP) samples were highly correlated with the matching MGS-identified data (Spearman r = 0.82) (Langille et al. 2013). Importantly, a high Spearman correlation is actually expected by chance in these comparisons simply because many genes are common in most environments while others are usually absent or rare. Upon comparing to this expectation the predictions are still significantly better than expected by chance, but only slightly (Douglas et al. 2020). Nonetheless, based on this approach, we found that PICRUSt2 performed marginally better than other tools (Douglas et al. 2020). However, it is noteworthy that Piphillin, which represents a much simpler approach based on a nearest-neighbour approach, performed only slightly worse overall and better in some contexts.

An alternative approach for evaluating these methods is based on the concordance of differential abundance results between actual and predicted metagenomics profiles. When we conducted this analysis while validating PICRUSt2, we found that differential abundances tests on metagenome prediction tools agreed only moderately well with matching tests based on actual MGS data (Douglas et al. 2020). This is a crucial point to appreciate when analyzing metagenome prediction data; even though the overall predicted profiles might correlate with MGS profiles, the results from differential abundance testing might nonetheless be quite different. We also observed high variation across datasets in concordance between MGS and 16S-based predictions. In other words, differential abundance testing on predicted profiles resulted in fair agreement with MGS data on some datasets while disagreeing almost entirely on others. In addition, researchers performing independent work in this area have identified conflicting signals of how well individual metagenome prediction tools perform (Narayan et al. 2020; Sun et al. 2020). These observations might again reflect the high variation

across datasets in how well prediction profiles agree with MGS results.

Current state of the integration of taxonomic and functional data types

The above discussion has described the many faces of microbiome data types. Taxonomic and functional microbiome data are typically generated independently, but in some cases can be directly linked. Regardless of the exact processing workflow for these data types, we have yet to address one question: how are they integrated?

For independent taxonomic and functional data types this is largely done anecdotally. For example, this is commonly done in regards to the nine genera that are the primary producers of short-chain fatty acids (SCFAs) in the human gut (Moya and Ferrer 2016). SCFA levels have long had an ambiguous link with Crohn's disease (CD) (Treem et al. 1994), although they are typically negatively associated with disease activity (Venegas et al. 2019). Due to this association, there has been long-standing interest in identifying microbial taxa that are associated with altered SCFA levels. Accordingly, CD microbiome studies commonly hypothesize that shifts in the relative abundance of any known SCFA-producing taxa likely cause altered SCFA levels. For example, *Faecalibacterium prausnitzii* is a well-known commensal SCFA-producer in the human gut and is consistently found at lower levels in the CD patient microbiomes (Wright et al. 2015). Although potential links between lower levels of this species, in addition to other taxa such as *Roseburia* (Laserna-Mendieta et al. 2018), and SCFA levels are often discussed, this is rarely formally investigated.

More often, anecdotal links between function and taxa are based on observed associations between significant features. Several such cases have previously been noted as representative examples (Manor and Borenstein 2017b). For instance, *Propionibacterium acnes* has been identified as strongly correlated with NADH dehydrogenase levels in the skin microbiome (Oh et al. 2014). Consequently, this species was implicated as the likely cause for changes in NADH dehydrogenase levels. Similarly, *Bacteroides thetaiotaomicron* relative abundance has been identified as positively correlated with microbial genes involved with the degradation of complex sugars and starch in the infant gut (Bäckhed et al. 2015). Based on this observation, this species was hypothesized to be the key contributor to increased levels of these degradation genes. Such insights are valuable, but as previously discussed (Manor and Borenstein 2017b), these anecdotal links alone are not convincing evidence that particular taxa are the primary contributors to functional shifts.

Linked taxonomic and functional data alone is not sufficient to resolve this issue. There are substantial challenges facing the integration of these data types besides simply generating a combined format. For example, two massive datasets have recently been published as part of the next iteration of the Human Microbiome Project. Both datasets include numerous sequencing and profiling technologies, including 16S and MGS, from the stool and various body-sites of IBD (Lloyd-Price et al. 2019) and individuals with pre-diabetes (Zhou et al. 2019). However, in each case there was little integration of microbiome functional and taxonomic data types. Instead, these features were largely tested independently, despite the availability of links between the data types, and associations between top features were discussed (Lloyd-Price et al. 2019; Zhou et al. 2019).

In contrast to these examples, there have been calls for improved integration of these microbiome data types, which is rooted in a systems-level biology outlook (Greenblum et al. 2013). "Functional Shifts' Taxonomic Contributors" (FishTaco) is one bioinformatics method developed for this purpose, which quantifies taxonomic contributions to functional shifts (Manor and Borenstein 2017b). One major application of this approach is to distinguish two explanations for why a function might be at high relative abundance (Figure 2). First, a function might be higher in relative abundance simply because it hitchhiked on the genome of a taxon that bloomed for other reasons. In contrast, an alternative explanation might be that many taxa performing the same function gained a growth advantage and thus grew in relative abundance. FishTaco can also identify functions that have grown in relative abundance simply because microbes that do not encode it are at lower levels.

FishTaco works by first identifying significant shifts in functional abundances with a standard differential abundance test, typically a Wilcoxon test. Subsequently, a permutation analysis is undertaken, which consists of randomly shifting the relative abundance of a subset of taxa, while maintaining the rest. A large collection of such permutations is performed, which include permutations of single and multiple taxa in different replicates. Based on this approach an estimate of the relative contribution of each taxon to a functional shift can be estimated

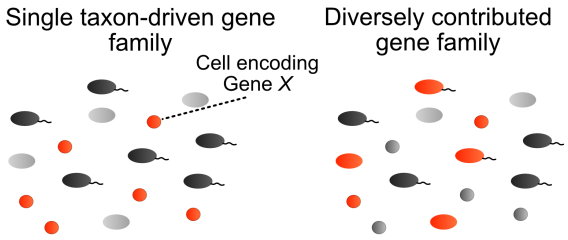

Figure 2: Two explanations for why a gene family might be at higher relative abundance that would be impossible to distinguish without joint taxonomic and functional data. Microbes encoding the gene of interest (Gene *X*) are indicated in red. This diagram contrasts how a gene family might be blooming due to a single taxon (left) versus a diverse set of taxa (right). The importance of distinguishing these scenarios is underappreciated: in the second case it is more likely the gene family itself that confers a growth or survival advantage in the environment. Note however that these are not the only two reasons why the relative abundance of a gene family might be at high levels in an environment.

(Manor and Borenstein 2017b). These relative contributions are then presented as stacked bar charts breaking down the direction and magnitude of each functional contribution. These visualizations help distinguish when a functional shift is due to the enrichment or depletion of taxa and also which sample grouping the shift occurred within. This approach was motivated by Shapley values, which were introduced in game-theory to summarize the contribution of each player in a multiplayer game (Shapley 1953). Specifically, FishTaco leverages a modified version of this approach that enables the contribution of individual features to be estimated in large datasets without exhaustively testing every possible permutation (Keinan et al. 2004).

FishTaco represents an important advancement in integration and improved interpretability of taxonomic and functional microbiome data. However, it nonetheless suffers from major limitations. First, although the taxonomic breakdown of contributors to a function is valuable, the FishTaco approach requires significant functions to be identified based on the relative abundance of individual gene families and pathways. This is done by systematically testing all functions across the entire metagenome, which is problematic when performed with a non-compositional approach like a Wilcoxon test. This approach also treats gene families under the bag-of-genes model, which is inappropriate, as discussed above. An improved method would conduct a compositionally sound analysis and integrate taxonomic information when identifying significant functions.

An alternative method is phylogenize, which does address each of these issues (Bradley et al. 2018; Bradley and Pollard 2020). This approach tests for significant associations between the presence of a taxa within a given sample grouping and the probability that a taxon encodes a given gene family. This is performed through phylogenetic linear regression, which accounts for the genetic similarity of co-occurring taxa that might trivially be due to a shared evolutionary history. A separate phylogenetic linear model is fitted for each gene family. The key distinction of this approach from a normal linear model is that instead of the residuals being independent and normally distributed, they covary so that phylogenetically similar microbes have higher covariance (Bradley et al. 2018). This overall approach was partially motivated by an attempt to address a similar problem by comparing the species and gene trees of gut and non-gut microbes (Lozupone et al. 2008). Based on simulated random data (i.e. data with no real functional shifts) the phylogenize authors demonstrated that performing standard linear models without controlling for phylogenetic structure results in false positive rates ranging from 20% - 68%. In contrast, controlling for phylogenetic structure with phylogenize resulted in a uniform P-value distribution and an appropriate false positive rate of 5%. One interesting feature is that phylogenize does not directly analyze relative abundances. Instead, the tool converts taxa relative abundance into one of three formats: (1) binary presence/absence across all samples, (2) overall prevalence within each sample grouping, (3) or the specificity within each sample grouping (Bradley et al. 2018).

Although phylogenize is undeniably an invaluable contribution to microbiome data analysis, it also has several limitations. First, information on taxa abundance is discarded entirely in favour of presence/absence data. From one perspective this is an advantage; eliminating taxa relative abundances enables phylogenize to circumvent compositionality issues. However, relative abundance data is often more important to investigate, because key taxonomic shifts might not be detected by presence/absence alone. In addition, phylogenize reports significant gene families for each phylum in a dataset. This is performed to reduce the memory usage and to enable phylum-specific rates of evolution for each function (Bradley et al. 2018). This focus on the phylum level makes the results difficult to interpret for two reasons. First, it is insufficiently broad, because it limits the potential to identify functions distributed

across multiple phyla that might be linked with a condition of interest. From another perspective, this focus on the phylum level is also not specific enough; although phylum-function associations are valuable they do not provide information on the relative contributions of lower-level taxa, such as species, to the association. Accordingly, there is room for improvement in both the statistical analysis and interpretation of the phylogenize approach.

Despite the availability of approaches for integrating functional and taxonomic data, they have yet to become a mainstay of microbiome analyses. However, it is becoming common to visualize stacked bar-charts of taxonomic contributors to functions of interest. This is typically performed on predicted metagenome output by PICRUSt or alternatively on HUMAnN2 output, although this could be performed with any linked taxa-function data. As discussed above, the HUMAnN2 pipeline includes a step for identifying particular strains in MGS dataset, which allows gene families to be linked to those strains (Franzosa et al. 2018). In some cases this approach enables complete links between taxa and function to be identified. For instance, *F. prausnitzii* was shown to be the obvious principal contributor to glutaryl-CoA biosynthesis in the HMP gut MGS samples (Franzosa et al. 2018). However, more commonly there are numerous taxonomic contributors to a single given function, and it is difficult to interpret which taxa are the key contributors by looking at visualizations alone. Nonetheless, even in the presence of many taxonomic contributors, the HUMAnN2 authors demonstrated that these visualizations can provide information about the diversity of taxa contributing to a function, termed the contributional diversity (Franzosa et al. 2018). This is most often quantified with the Gini-Simpson index, which is the complement of Simpson's evenness (Jost 2006).

Contributional diversity has been shown to be a useful approach for delineating housekeeping pathways encoded by many taxa, intermediate pathways, and those rarely encoded, which can correspond to opportunists or keystone species. For instance, *F. prausnitzii* has previously been linked with several human microbiome pathways identified through MGS that have intermediate contributional diversities (Abu-Ali et al. 2018). When present, this species tended to contribute the majority of all pathways it encoded.

This approach has also been valuable for profiling shifts in the contributions to microbial pathways over time, such as in the infant gut profiled with MGS (Vatanen et al. 2018). In this case, several microbial pathways, such as siderophore biosynthesis, were found to display decreasing contributional diversity with age. This is an interesting observation because siderophores are costly to produce but are highly beneficial in the human gut. In particular, siderophores can confer a strong benefit to multiple community members, including those that do not produce siderophores, by providing access to iron. Siderophores have previously been presented as microbial functions whose distribution is consistent with the Black Queen Hypothesis (Morris et al. 2012). This hypothesis states that adaptive gene loss may occur for functions that are costly to produce, provided that the function is provided by other community members. This hypothesis was discussed in the context of the infant microbiome as an explanation for why siderophore contributional diversity decreases over time (Vatanen et al. 2018): perhaps gene loss confers an adaptive benefit by avoiding the production of a costly metabolite. Although this is an interesting hypothesis, a less controversial interpretation of this result is simply that siderophores became less stably encoded over time in the profiled samples.

Related to this point, two additional metrics have also been developed to summarize the stability of taxonomic contributions to microbial functions (Eng and Borenstein 2018). More specifically, these metrics are intended to summarize functional robustness across samples, which is the stability in the relative abundance for a given function in response to taxonomic perturbation. This is performed by generating a taxa-response curve that describes the average change in functional relative abundances in response to taxonomic perturbations of different magnitudes. Two metrics are then computed based upon these curves: attenuation and buffering. Attenuation captures how rapidly a function shifts with increasing taxonomic perturbation magnitudes. In contrast, buffering represents how well functional shifts are suppressed at smaller taxonomic perturbation magnitudes.

Applying these metrics to PICRUSt-predicted metagenomes from 16S sequencing of human body sites, validated by a subset of MGS samples, yielded several novel perspectives. First, attenuation and buffering were conserved across body sites for microbial house-keeping pathways but varied for several others. For instance, robustness in the biosynthesis of unsaturated fatty acids varied substantially across body sites. In addition, human gut samples were found to have higher values of both attenuation and buffering than compared to vaginal samples. These trends were shown to be driven by more than simply lower richness in vaginal samples by subsampling

to comparable diversity levels across each body-site (Eng and Borenstein 2018). These observations are consistent with the controversial hypothesis that microbial communities may be under varying selection strengths for functional robustness, depending on the environment (Naeem et al. 1998; Ley et al. 2006).

The development of these metrics for summarizing functional contributions represent an important goal of microbiome research, which is to leverage sequencing data to yield novel biological insights. In contrast, another major goal is to answer a more practical question: how useful is microbiome data for classification and prediction tasks?

There is great interest in applying machine learning approaches to microbiome sequencing data (Knights et al. 2011). Most commonly this is performed with either Support Vector Machine or Random Forest (Breiman 2001) models. Applications of these and other machine learning approaches to microbiome data are primarily aimed at distinguishing samples from different environments or disease states (Zhou and Gallins 2019). Taxonomic features are the focus of most such microbiome-based machine learning approaches, which is true for both 16S (Duvallet et al. 2017) and MGS (Pasolli et al. 2016) data. However, on a growing number of occasions machine learning is focused on functional data types. For example, a recent MGS meta-analysis identified informative functional biomarkers across several human diseases by applying machine learning approaches to functional data types (Armour et al. 2019). Regardless of the data type, models trained on microbiome data typically have low generalizability across independent cohorts (Sze and Schloss 2016; Douglas et al. 2018), although there are exceptions.

One major exception is microbiome-based modelling of colorectal cancer, which in one investigation was shown to be generalizable across five independent datasets (Wirbel et al. 2019). This landmark study also systematically compared the utility of functional and taxonomic data types in these models and found them to be comparable overall. This finding is consistent with a past comparison of the classification performance of 16S-based taxa and predicted metagenome data (Ning and Beiko 2015). In the case of predicted metagenomes, which are based on 16S profiles, it is perhaps less surprising that they yield comparable classification performance. However, with MGS data in particular it might be possible to detect robust, informative functions that might be undetectable with taxonomy alone due to taxonomic variability (Doolittle and Booth 2017).

Despite this great interest in applying machine learning to different microbiome data types, there has been little focus on integrating across them. The aforementioned comparison of 16S-based taxa and predicted functions is one exception where a hybrid classification model of both data types was created (Ning and Beiko 2015). In this case, there was a small increase in classification performance for distinguishing nine human oral sub-locations. The original OTU and KO-based models yielded accuracies of 76.2% and 76.1%, respectively, while the hybrid model resulted in an accuracy of 77.7% (Ning and Beiko 2015). This result indicates that predicted functions may provide some additional information in combination with taxonomic data, but the consistency and biological significance of this small effect remains unclear. Further investigation into the integration of these data types within a machine learning context is needed to ensure that the highest-quality models possible are constructed.

Outlook

Herein we have described the unique characteristics of microbiome DNA data types and many of the approaches that have been proposed for their analysis. Throughout we have emphasized two ideas. First, increased integration of taxonomic and functional microbiome data types is needed. And second, there is often high variation in the results between microbiome data analysis pipelines.

Regarding the first point, we believe that several of the tools described above, such as FishTaco and phylogenize, largely solve the issue of how to jointly investigate taxa and functions. Increased usage and development of these and other related tools would greatly help with the interpretability of microbiome data.

One area where further development is particularly needed is in the context of classification models, where little work has been conducted to systematically link taxa and functions appropriately. One exception was a classification approach based on gene families that identified predictive genes and then subsequently identified metagenome assembled genomes within a given dataset enriched for these genes (Rahman et al. 2018). However, this approach still relied on follow-up analyses rather than integrating the data types. Instead, an improved approach could be based on explicitly leveraging the hierarchical nature of microbiome data types. This is because functional and taxonomic data types independently form clear hierarchical structures (e.g. Pathway - Gene and Phylum - Class - Order, etc.). The connection between taxa and gene families and pathways is

more complex, but nonetheless, links between groups of strains or ASVs and microbial functions can be defined. A modified machine learning framework that explicitly accounted for these relationships could result in more interpretable outputs.

Regardless of the specific tool, microbiome researchers should move towards more integration of taxonomic and functional data. It is odd to distinguish between functional and taxonomic datatypes in the first place: they are inextricably linked after all. The term "metagenome" itself is in some ways unfortunate as it implies that the genetic information for all organisms in a community can be simultaneously analyzed in a coherent way, without partitioning genes into genomes. This may be valid for high-level pathways but for generating hypotheses regarding specific gene families it is too often misleading. This perspective is becoming more common, as the availability of metagenome-assembled genomes increases (Frioux et al. 2020).

The other common thread throughout this manuscript has been that technical variation in microbiome data analyses means that making robust biological inferences, especially regarding specific microbial features, is challenging. Indeed, the lack of standardization in microbiome data analysis has previously been strongly criticized. An assessment of numerous papers attempting to define standard pipelines concluded that there was disturbingly little consensus (Pollock et al. 2018). This is true for many steps related to the processing, sequencing, and analysis of microbiome data. For instance, there have been contradictory results regarding the efficacy of different extraction protocols (Salonen et al. 2010). In particular, underrepresentation of Gram-positives has been observed (Maukonen et al. 2012), which may be partially resolved by using bead-beating extraction protocols (Guo and Zhang 2013). There is also substantial technical variation related to bioinformatics choices, which represent the final steps of a microbiome project. For example, as discussed above, the bioinformatics choices made when performing differential abundance testing on microbiome data can have severe impacts on any interpretations (Thorsen et al. 2016; Hawinkel et al. 2019).

We have encountered similar issues with our work, most strikingly when investigating pediatric Crohn's disease patients' microbiome profiles (Douglas et al. 2018). An important characteristic of these data was that 98% of the sequenced reads mapped to the human genome. This characteristic made taxonomic profiling of these data especially prone to false positives. In particular, an initial draft of our manuscript was based on profiles that included large proportions of viral-identified DNA and matches to certain eukaryotic parasites. We were initially excited about these observations, because the abundances of these non-prokaryotic taxa were discriminative for classifying patient disease state and treatment response. However, the exact taxa identified were peculiar: they were predominately represented by a range of plant-associated viruses and the eukaryotic genus *Plasmodium*, which is best known as including the causative agent for malaria, *Plasmodium falciparum*. Upon closer investigation it became clear that this signal was driven entirely by a difference in how reads were mapped to lineage-specific marker genes. Altering the parameter choice from local to global mapping entirely removed these taxa. This relatively small difference in parameter choice appeared to only affect our data and not more typical microbiome datasets, which we believe was due to the high proportion of human DNA in our data. Although this error was moderately embarrassing, it was more importantly an example of how easily a single parameter setting can result in starkly different biological interpretations. In this case the difference was driven by an option used for a single bioinformatics tool.

Such inconsistencies in microbiome analyses have previously been identified and been shown to make meaningful comparisons across studies challenging. For instance, associations between obesity and the human microbiome are commonly discussed as support for the utility of considering microbial links with human disease, despite inconsistencies across studies (Castaner et al. 2018; Muscogiuri et al. 2019). These inconsistencies are typically explained due to confounding variables that may differ between patient cohorts. Although this is a valid explanation, it is likely that technical variation, including in terms of bioinformatics analyses, also drives these inconsistencies. For instance, a meta-analysis of ten obesity human microbiome datasets identified only extremely weak signals when re-analyzing all datasets with a standardized approach (Sze and Schloss 2016). This finding greatly contrasts with how these studies were originally presented and again highlights how variation in bioinformatics can greatly affect how to biologically interpret microbiome data.

Similarly lower alpha diversity in stool microbiomes has been frequently linked with disease states (Mosca et al. 2016). These observations are intuitively reasonable as reduced alpha diversity could enable pathogens to bloom (Vincent et al. 2013) or represent differences in resource availability (Turnbaugh et al. 2009). However a re-analysis of data

from 28 studies representing ten diseases was unable to identify evidence for links between alpha diversity and disease states (Duvallet et al. 2017). The exceptions were diarrheal diseases and inflammatory bowel diseases.

Such inconsistencies across analyses on the same data are gradually coming to the forefront of the microbiome field (Allaband et al. 2019). Indeed, a recent plea for improved standardization has been made to enable better comparisons across studies (Hill 2020). This is a commendable goal, but given the diversity of opinions regarding best-practices (Callahan et al. 2016b; Knight et al. 2018; Schloss 2020), it is difficult to coherently recommend a single workflow for analyses at the moment. Accordingly, further work and benchmarking of different bioinformatics is needed to convincingly argue for best practices in microbiome data analysis.

Until a clear consensus is reached it is the responsibility of microbiome researchers to make the caveats and challenges facing this area clear to readers and newcomers to the field. This is crucial given the widespread interest in studying microbiomes through DNA sequencing; the number of microbiome sequencing-related publications continues to rapidly grow. This is in tandem with funding for these projects, which has steadily increased in the USA from at least 2007 to 2016 (NIH 2019). According to the US National Health Institute, there was US$766 million dollars invested in microbiome research in 2019, which was the 63rd most highly funded health-related research category out of 291. Although comparing across research categories of varying granularity is difficult, it is noteworthy that microbiome research was more highly funded than both breast cancer and Alzheimer's disease research. Importantly, an increased interest in microbiome research is warranted: recent technological developments are enabling improved investigations into microbial biology. However, as the monetary investment and research hours dedicated to microbiome research grows, it is crucial that scientists ensure the best use of these resources. Open discussions on the many contentious aspects of microbiome data analysis would help with this issue. Indeed, such clarifications by leaders in the microbiome field are starting become more common (Allaband et al. 2019). However, although these contributions are valuable, they do not adequately address the problem. In particular, instead of mentioning these issues in passing, inconsistencies between bioinformatics workflows should be emphasized more clearly for the benefit of the uninitiated.

Another practical improvement would be to normalize, and potentially require, explicit summaries of the effects of technical variation on any biological interpretations reported in microbiome studies. This is impossible to capture entirely, but it could be done by comparing how key results change depending on a subset of representative bioinformatics choices. For instance, researchers could compare how insights change depending on the combinations of denoising tools and differential abundance methods that they have applied when analyzing 16S data. Although these changes would result in increased workloads when conducting analyses and when communicating results, they would help ensure that any major biological findings are at least robust to a representative set of bioinformatics choices.

Regardless of which approach is taken to address these issues, the most important point is that action is needed on this front. The variation between bioinformatics methods is undeniable and unfortunately reflects a reproducibility crisis facing microbiome data analysis.

Acknowledgements

We would like to thank the following individuals for feedback on sections of this manuscript: Dr. Robert Beiko, Dr. Zhenyu Cheng, Dr. André Comeau, Casey Jones, Jacob Nearing, Dr. Laura Parfrey, Dr. Andrew Stadnyk, Chris Tang, and Dr. Robyn Wright.

Conflicts of interest

We declare that we have no conflicts of interest with the content of this article.

References

Abu-Ali GS et al. 2018. Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat. Microbiol. 3:356-366.

Abubucker S et al. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLOS Comput. Biol. 8:e1002358.

Aitchison J. 1982. The Statistical Analysis of Compositional Data. J. R. Stat. Soc. Ser. B. 44:139-177.

Allaband C et al. 2019. Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. Clin. Gastroenterol. Hepatol. 17:218-230.

Ambler RP et al. 1979. Cytochrome C2 sequence variation among the recognised species of purple nonsulphur photosynthetic bacteria. Nature. 278:659-660.

Angly FE et al. 2014. CopyRighter: A rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome. 2:11.

Apweiler R et al. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 32:D115-119.

Arboleya S et al. 2018. Gene-trait matching across the *Bifidobacterium longum* pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. BMC Genomics. 19:33.

Armour CR, Nayfach S, Pollard KS, Sharpton TJ. 2019. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. mSystems. 4:e00332-18.

Ayling M, Clark MD, Leggett RM. 2019. New approaches for metagenome assembly with short reads. Brief. Bioinform. 00:1-11.

Bäckhed F et al. 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 17:690-703.

Barott KL et al. 2012. Microbial to reef scale interactions between the reef-building coral *Montastraea annularis* and benthic algae. Proc. R. Soc. B Biol. Sci. 279:1655-1664.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. USA. 102:14332-14337.

Bowers RM et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35:725-731.

Bowman JS, Ducklow HW. 2015. Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal West Antarctic Peninsula. PLOS One. 10:e0135868.

Bradley PH, Nayfach S, Pollard KS. 2018. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. PLOS Comput. Biol. 14:e1006242.

Bradley PH, Pollard KS. 2020. Phylogenize: Correcting for phylogeny reveals genes associated with microbial distributions. Bioinformatics. 36:1289-1290.

Breiman L. 2001. Random Forests. Mach. Learn. 45:5-32.

Brenner DJ. 1973. Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. Int. J. Syst. Bacteriol. 23:298-307.

Buchfink B, Xie C, Huson DH. 2015. Fast and Sensitive Protein Alignment using DIAMOND. Nat. Methods. 12:59-60.

Bukin YS et al. 2019. The effect of 16s rRNA region choice on bacterial community metabarcoding results. Sci. Data. 6:190007.

Burke C, Steinberg P, Rusch DB, Kjelleberg S, Thomas T. 2011. Bacterial community assembly based on functional genes rather than species. Proc. Natl. Acad. Sci. USA. 108:14288-14293.

Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. 2020. Assessment of single cell RNA-seq statistical methods on microbiome data. Genome Biol. 191.

Callahan BJ et al. 2016a. DADA2: High resolution sample inference from amplicon data. Nat. Methods. 13:581-583.

Callahan BJ et al. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. 47:e103.

Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. 2016b. Bioconductor workflow for microbiome data analysis: From raw reads to community analyses. F1000 Res. 5:1492.

Caspi R et al. 2013. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 42:D459-D471.

Castaner O et al. 2018. The gut microbiome profile in obesity: A systematic review. Int. J. Endocrinol. 2018:4095789.

Chaffron S, Rehrauer H, Pernthaler J, Von Mering C. 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 20:947-959.

Chen PE et al. 2010. Genomic characterization of the *Yersinia* genus. Genome Biol. 11:R1.

Chen Z et al. 2019. Impact of Preservation Method and 16S rRNA Hypervariable Region on Gut Microbiota Profiling. mSystems. 4:e00271-18.

Comeau AM, Douglas GM, Langille MGI. 2017. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems. 2:e00127-16.

Comeau AM, Lagunas MG, Scarcella K, Varela DE, Lovejoy C. 2019. Nitrate Consumers in Arctic Marine Eukaryotic Communities: Comparative Diversities of 18S rRNA, 18S rRNA Genes, and

Nitrate Reductase Genes. Appl. Environ. Microbiol. 85:e00247-19.

Darling AE et al. 2014. PhyloSift: Phylogenetic analysis of genomes and metagenomes. PeerJ. 2014:243.

Devereux R et al. 1990. Diversity and origin of *Desulfovibrio* species: Phylogenetic definition of a family. J. Bacteriol. 172:3609-3619.

Doolittle WF, Booth A. 2017. It's the song, not the singer: an exploration of holobiosis and evolutionary theory. Biol. Philos. 32:5-24.

Douglas GM et al. 2018. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome. 6:13.

Douglas GM et al. 2020. PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. 38:685-688.

Drouin G, Godin J-R, Pagé B. 2011. The genetics of vitamin C loss in vertebrates. Curr. Genomics. 12:371-8.

Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. 8:1784.

Egozcue JJ, Pawlowsky-Glahn V. 2011. Exploring compositional data with the CoDa-dendrogram. Austrian J. Stat. 40:103-113.

Eng A, Borenstein E. 2018. Taxa-function robustness in microbial communities. Microbiome. 6:45.

Farrelly V, Rainey FA, Stackebrandt E. 1995. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl. Environ. Microbiol. 61:2798-2801.

Faust K et al. 2012. Microbial co-occurrence relationships in the Human Microbiome. PLOS Comput. Biol. 8.

Fernandes AD et al. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2:15.

Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. 2013. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLOS One. 8:e67019.

Finn RD et al. 2014. Pfam: The protein families database. Nucleic Acids Res. 42:D222-D230.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279-284.

Francke C, Siezen RJ, Teusink B. 2005. Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol. 13:550-558.

Franzosa EA et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods. 15:962-968.

Friedman J, Alm EJ. 2012. Inferring Correlation Networks from Genomic Survey Data. PLOS Comput. Biol. 8:e1002687.

Frioux C, Singh D, Korcsmaros T, Hildebrand F. 2020. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Comput. Struct. Biotechnol. J. 18:1722-1734.

Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin E V. 2019. Microbial genome analysis: The COG approach. Brief. Bioinform. 20:1063-1070.

Galperin MY, Makarova KS, Wolf YI, Koonin E V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43:D261-D269.

Gevers D et al. 2014. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 15:382-392.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: And this is not optional. Front. Microbiol. 8:2224.

Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. 2016. It's all relative: analyzing microbiome data as compositions. Ann. Epidemiol. 26:322-329.

Goodrich JK et al. 2014. Conducting a microbiome study. Cell. 158:250-262.

Graspeuntner S, Loeper N, Künzel S, Baines JF, Rupp J. 2018. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. Sci. Rep. 8:9678.

Greenblum S, Chiu HC, Levy R, Carr R, Borenstein E. 2013. Towards a predictive systems-level model of the human microbiome: Progress, challenges, and opportunities. Curr. Opin. Biotechnol. 24:810-820.

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

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31:371-373.

Hallam SJ, Girguis PR, Preston CM, Richardson PM, DeLong EF. 2003. Identification of Methyl Coenzyme M Reductase A (*mcrA*) Genes Associated with Methane-Oxidizing Archaea. Appl. Environ. Microbiol. 69:5483-5491.

Hao X, Chen T. 2012. OTU Analysis Using Metagenomic Shotgun Sequencing Data. PLOS One. 7:e49785.

Hauben L, Vauterin L, Moore ERB, Hoste B, Swings J. 1999. Genomic diversity of the genus *Stenotrophomonas*. Int. J. Syst. Bacteriol. 49:1749-1760.

Hauben L, Vauterin L, Swings J, Moore ERB. 1997. Comparison of 16S Ribosomal DNA Sequences of All *Xanthomonas* Species. Int. J. Syst. Bacteriol. 47:328-335.

Hawinkel S, Mattiello F, Bijnens L, Thas O. 2019. A broken promise: Microbiome differential abundance methods do not control the false discovery rate. Brief. Bioinform. 20:1-12.

Hill C. 2020. You have the microbiome you deserve. Gut Microbiome. 1:1-4.

Hillmann B et al. 2018. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems. 3:e00069-18.

HMP-consortium. 2013. Structure, Function and Diversity of the Healthy Human Microbiome. Nature. 486:207-214.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome Res. 17:377-386.

Ibrahim A, Goebel BM, Liesack W, Griffiths M, Stackebrandt E. 1993. The phylogeny of the genus *Yersinia* based on 16S rDNA sequences. FEMS Microbiol. Lett. 114:173-177.

Inkpen AI et al. 2017. The Coupling of Taxonomy and Function in Microbiomes. Biol. Philos. 32:1225-1243.

Iwai S et al. 2016. Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. PLOS One. 11:e0166104.

Jackson DA. 1997. Compositional data in community ecology: The paradigm or peril of proportions? Ecology. 78:929-940.

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. J. Clin. Microbiol. 45:2761-2764.

Jayaprakash TP, Schellenberg JJ, Hill JE. 2012. Resolution and characterization of distinct cpn60-based subgroups of *Gardnerella vaginalis* in the vaginal microbiota. PLOS One. 7:e43009.

Jensen LJ et al. 2008. eggNOG: Automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 36:D250-D254.

Johnson JS et al. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat. Commun. 10:5029.

Jones MB et al. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc. Natl. Acad. Sci. USA. 112:14024-14029.

Jost L. 2006. Entropy and diversity. Oikos. 113:363-375.

Jun SR, Robeson MS, Hauser LJ, Schadt CW, Gorin AA. 2015. PanFP: Pangenome-based functional profiles for microbial communities. BMC Res. Notes. 8:479.

Kalan LR et al. 2019. Strain- and Species-Level Variation in the Microbiome of Diabetic Wounds Is Associated with Clinical Outcomes and Therapeutic Efficacy. Cell Host Microbe. 25:641-655.

Kallonen T et al. 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res. 27:1437-1449.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28:27-30.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44:D457-D462.

Kang CH et al. 2007. Relationship between genome similarity and DNA-DNA hybridization among closely related bacteria. J. Microbiol. Biotechnol. 17:945-951.

Keinan A, Sandbank B, Hilgetag CC, Meilijson I, Ruppin E. 2004. Fair attribution of functional contribution in artificial and biological networks. Neural Comput. 16:1887-1915.

Keswani J, Whitman WB. 2001. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. Int. J. Syst. Evol. Microbiol. 51:667-678.

Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and accurate classificaton of metagenomic sequences. Genome Res. 26:1721-1729.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol. Biol. Evol. 28:1057-1074.

Knight R et al. 2018. Best practices for analysing microbiomes. Nat. Rev. Microbiol. 16:410-422.

Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. 2011. Human-associated microbial signatures: Examining their predictive value. Cell Host Microbe. 10:292-296.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA. 102:2567-2572.

Koonin E V, Galperin MY. 2003. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Kluwer Academic: Boston.

Kurtz ZD et al. 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Comput. Biol. 11:e1004226.

Langille MGI et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31:814-821.

Laserna-Mendieta EJ et al. 2018. Determinants of reduced genetic capacity for butyrate synthesis by the gut microbiome in Crohn's disease and ulcerative colitis. J. Crohn's Colitis. 12:204-216.

Lau JT et al. 2016. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. Genome Med. 8:72.

Ley RE, Peterson DA, Gordon JI. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 124:837-848.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31:1674-1676.

Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE. 2012. The Chaperonin-60 Universal Target Is a Barcode for Bacteria That Enables De Novo Assembly of Metagenomic Sequence Data. PLOS One. 7:e49755.

Liu J, Yu Y, Cai Z, Bartlam M, Wang Y. 2015. Comparison of ITS and 18S rDNA for estimating fungal diversity using PCR-DGGE. World J. Microbiol. Biotechnol. 31:1387-1395.

Lloyd-Price J et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 569:655-662.

Lloyd-Price J et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 550:61-66.

Louca S et al. 2018a. Function and functional redundancy in microbial systems. Nat. Ecol. Evol. 2:936-943.

Louca S, Doebeli M. 2017. Taxonomic variability and functional stability in microbial communities infected by phages. Environ. Microbiol. 19:3863-3878.

Louca S, Doebeli M, Parfrey LW. 2018b. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome. 6:41.

Louca S, Parfrey LW, Doebeli M. 2016. Decoupling function and taxonomy in the global ocean microbiome. Science. 353:1272-1277.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Lozupone C, Knight R. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl. Environ. Microbiol. 71:8228-8235.

Lozupone CA et al. 2008. The convergence of carbohydrate active gene repertoires in human gut microbes. Proc. Natl. Acad. Sci. USA. 105:15076-15081.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput. Sci. 3:e104.

Makarova KS, Wolf YI, Koonin E V. 2015. Archaeal clusters of orthologous genes (arCOGs): An update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. Life. 5:818-840.

Mandal S et al. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Heal. Dis. 26:27663.

Mandel M. 1966. Deoxyribonucleic Acid Base Composition in the Genus *Pseudomonas*. J. Gen. Microbiol. 43:273-292.

Manor O, Borenstein E. 2017a. Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome. Microbiome. 5:19.

Manor O, Borenstein E. 2017b. Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome. Cell Host Microbe. 21:254-267.

Martin BD, Witten D, Willis AD. 2020. Modeling microbial abundances and dysbiosis with beta-binomial regression. Ann. Appl. Stat. 14:94-115.

Martiny AC. 2019. High proportions of bacteria are culturable across major biomes. ISME J. 13:2125-2128.

Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in microorganisms. ISME J. 7:830-838.

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

McDonald D et al. 2019. redbiom: a Rapid Sample Discovery and Feature Characterization System. mSystems. 4:e00215-19.

McIntyre ABR et al. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 18:182.

McMahon K. 2015. "Metagenomics 2.0". Environ. Microbiol. Rep. 7:38-39.

Meyer F et al. 2008. The metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 9:386.

Meyer F, Overbeek R, Rodriguez A. 2009. FIGfams: Yet another set of protein families. Nucleic Acids Res. 37:6643-6654.

Miossec MJ et al. 2020. Evaluation of computational methods for human microbiome analysis using simulated data. PeerJ. 8:e9688.

Morgan XC et al. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 13:R79.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182-W185.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. MBio. 3:e00036-12.

Morton JT et al. 2017. Balance Trees Reveal Microbial Niche Differentiation. mSystems. 2:e00162-16.

Morton JT et al. 2019. Establishing microbial composition measurement standards with reference frames. Nat. Commun. 10:2719.

Mosca A, Leclerc M, Hugot JP. 2016. Gut microbiota diversity and human diseases: Should we reintroduce key predators in our ecosystem? Front. Microbiol. 7:455.

Moya A, Ferrer M. 2016. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. Trends Microbiol. 24:402-413.

Muegge BD et al. 2011. Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. Science. 332:970-974.

Muscogiuri G et al. 2019. Gut microbiota: a new path to treat obesity. Int. J. Obes. Suppl. 9:10-19.

Mysara M et al. 2017. Reconciliation between operational taxonomic units and species boundaries. FEMS Microbiol. Ecol. 93:fix029.

Naeem S, Kawabata Z, Loreau M. 1998. Transcending boundaries in biodiversity research. Trends Ecol. Evol. 13:134-135.

Narayan NR et al. 2020. Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. BMC Genomics. 21:56.

Nearing JT, Douglas GM, Comeau AM, Langille MGI. 2018. Denoising the Denoisers: An independent evaluation of microbiome sequence error- correction approaches. PeerJ. 2018:e5364.

Nejman D et al. 2020. The human tumor microbiome is composed of tumor type-specific intra-cellular bacteria. Science. 980:973-980.

NIH. 2019. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome. 7:31.

Ning J, Beiko RG. 2015. Phylogenetic approaches to microbial community classification. Microbiome. 3:47.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27:824-834.

Oberhardt MA, Pucha?ka J, Fryer KE, Martins Dos Santos VAP, Papin JA. 2008. Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. J. Bacteriol. 190:2790-2803.

Oh J et al. 2014. Biogeography and individuality shape function in the human skin metagenome. Nature. 514:59-64.

Olson ND et al. 2019. Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief. Bioinform. 20:1140-1150.

Omelchenko M V., Galperin MY, Wolf YI, Koonin E V. 2010. Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. Biol. Direct. 5:31.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25:1043-55.

Pasolli E, Truong T, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLOS Comput. Biol. 12:e1004977.

Paulson JN, Colin Stine O, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods. 10:1200-1202.

Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The Madness of Microbiome: Attempt-

ing To Find Consensus "Best Practice" for 16S Microbiome Studies. Appl. Environ. Microbiol. 84:e02627-17.

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. Curr. Opin. Microbiol. 14:615-623.

Prakash T, Taylor TD. 2012. Functional assignment of metagenomic data: challenges and applications. Brief. Bioinform. 13:711-727.

Prodan A et al. 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLOS One. 15:e0227434.

Punta M et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290-D301.

Quast C et al. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41:590-596.

Rahman SF, Olm MR, Morowitz MJ, Banfield JF. 2018. Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome. mSystems. 3:e00123-17.

Rasko DA et al. 2008. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 190:6881-6893.

Riley M. 1993. Functions of the gene products of *Escherichia coli*. Microbiol. Rev. 57:862-952.

Salonen A et al. 2010. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J. Microbiol. Methods. 81:127-134.

Saroj DB, Dengeti SN, Aher S, Gupta AK. 2015. ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. World J. Microbiol. Biotechnol. 31:189.

Schloss PD. 2020. Reintroducing mothur: 10 Years Later. Appl. Environ. Microbiol. 86:e02343-19.

Schoch CL et al. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. USA. 109:6241-6246.

Schwager E, Mallick H, Ventz S, Huttenhower C. 2017. A Bayesian method for detecting pairwise associations in compositional data. PLOS Comput. Biol. 13:e1005852.

Segata N et al. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12:R60.

Shade A. 2017. Diversity is the question, not the answer. ISME J. 11:1-6.

Shapley LS. 1953. A value for n-person games. In: Contributions to the Theory of Games, 2. Kuhn, HW & Tucker, W, editors. Princeton University Press: Princeton, NJ pp. 307-317.

Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife. 6:e21887.

Sperling JL et al. 2017. Comparison of bacterial 16S rRNA variable regions for microbiome surveys of ticks. Ticks Tick. Borne. Dis. 8:453-461.

Sprockett D et al. 2019. Treatment-specific composition of the gut microbiota is associated with disease remission in a pediatric Crohn's disease cohort. Inflamm. Bowel Dis. 25:1927-1938.

Stackebrandt E, Goebel BM. 1994. Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44:846-849.

Staley J, Konopka A. 1985. Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annu. Rev. Microbiol. 39:321-346.

Stein JL, Marsh TL, Wu KY, Shizuya H, Delong EF. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. 178:591-599.

Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. Nat. Commun. 9:2542.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35:1026-1028.

Suardana IW. 2014. Analysis of Nucleotide Sequences of the 16S rRNA Gene of Novel *Escherichia coli* Strains Isolated from Feces of Human and Bali Cattle. J. Nucleic Acids. 2014:475754.

Sun DL, Jiang X, Wu QL, Zhou NY. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl. Environ. Microbiol. 79:5962-5969.

Sun S, Jones RB, Fodor AA. 2020. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. Microbiome. 8:46.

Sze MA, Schloss PD. 2016. Looking for a signal in the noise: Revisiting obesity and the microbiome. MBio. 7:e01018-16.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33-36.

Tedjo DI et al. 2016. The fecal microbiota as a biomarker for disease activity in Crohn's disease. Sci. Rep. 6:35216.

Tessler M et al. 2017. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci. Rep. 7:6589.

Tettelin H et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA. 102:3950-13955.

Thompson LR et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 551:457-463.

Thorsen J et al. 2016. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 4:62.

Treem W, Ahsan N, M S, Hyams J. 1994. Fecal Short-Chain Fatty Acids in Children with Inflammatory Bowel Disease. J. Pediatr. Gastroenterol. Nutr. 18:159-164.

Truong DT et al. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods. 12:902-903.

Turnbaugh PJ et al. 2009. A core gut microbiome in obese and lean twins. Nature. 457:480-484.

Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI. 2008. Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome. Cell Host Microbe. 3:213-223.

Vatanen T et al. 2018. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat. Microbiol. 4:470-479.

Venegas DP et al. 2019. Short chain fatty acids (SCFAs) mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. Front. Immunol. 10:277.

Venter JC et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science. 304:66-74.

Verster AJ, Borenstein E. 2018. Competitive lottery-based assembly of selected clades in the human gut microbiome. Microbiome. 6:186.

(Větrovský T, Baldrian P. 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLOS One. 8:e57923.

Vincent C et al. 2013. Reductions in intestinal Clostridiales precede the development of nosocomial *Clostridium difficile* infection. Microbiome. 1:18.

Wang Y, Qian P, Ya S. 2013. Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes. Encycl. Metagenomics. https://doi.org/10.1007/978-1-4614-6418-1_772-1.

Watson EJ, Giles J, Scherer BL, Blatchford P. 2019. Human faecal collection methods demonstrate a bias in microbiome composition by cell wall structure. Sci. Rep. 9:16831.

Weiss S et al. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 5:27.

Wemheuer F et al. 2020. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. Environ. Microbiome. 15:11.

Wheeler NE, Barquist L, Kingsley RA, Gardner PP. 2016. A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. Bioinformatics. 32:3566-3574.

Willis C, Desai D, Laroche J. 2019. Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic. FEMS Microbiol. Lett. 366:fnz152.

Wilson GA et al. 2005. Orphans as taxonomically restricted and ecologically important genes. Microbiology. 151:2499-2501.

Wirbel J et al. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. 25:679-689.

Woese CR. 1987. Bacterial evolution. Microbiol. Rev. 51:221-271.

Woese CR et al. 1980. Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res. 8:2275-2294.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. USA. 74:5088-5090.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20:257.

Wright EK et al. 2015. Recent advances in characterizing the gastrointestinal microbiome in Crohn's disease: a systematic review. Inflamm Bowel Dis. 21:1219-1228.

Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as ?Markers' for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major

Subgroups. PLOS One. 8:e77033.
Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 178:779-794.
Ye Y, Doak TG. 2011. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Metagenomes. PLOS Comput. Biol. 5:e1000465.
Zaneveld JR, Lozupone C, Gordon JI, Knight R. 2010. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res. 38:3869-3879.
Zaneveld JRR, Thurber RLV. 2014. Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Front. Microbiol. 5:431.
Zhou J et al. 2015. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. MBio. 6:e02288-14.
Zhou W et al. 2019. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature. 569:663-671.
Zhou YH, Gallins P. 2019. A review and tutorial of machine learning methods for microbiome host trait prediction. Front. Genet. 10:579.
Zuckerkandl E, Pauling L. 1965. Molecules as documents of history. J. Theor. Biol. 8:357-366.

Reviewer 1

Dear Gavin Douglas and Morgan Langille,

In this review manuscript, you propose to deliver a detailed introduction to microbiome DNA sequence data types and analysis methods. You present marker-gene and shotgun DNA sequencing data types, discuss microbiome data characteristics and underscore the associated caveats. Then you present « the many-faceted concept of microbial functions ». You follow by a discussion on the problematic of functional annotation inferred from marker-gene data and you review the last development on the integration of taxonomy and function. Finally, you discuss reproducibility in microbiome research and provide an outlook with some personal experience.

The main strength of your manuscript is that as a reader I learned something because you deliver an interesting review and discussion on the integration of taxonomic and functional microbiome data, backed up by first-hand and authoritative experience. However, the main weakness is that your message is diluted by a lengthy and unclear explanation of some concepts that are not always directly linked to your main discussion point.

Therefore, I recommend a major revision of your manuscript.

Sincerely,

Nicolas Pollet

Major comments:

What is the audience ? The title says « A primer and discussion on DNA-based microbiome data and related bioinformatics analyses ». Since one aim is to deliver a primer, the reader is expected to be a non-expert, and therefore the discussion that follows is also expected to reach a non-expert in the field. Is this the case ? I don't think so. In fact, I am unsure of the efficiency to pursue the goal of fulfilling the role of a primer AND a discussion on microbiome data for a reader completely new to the field and for the complicated topics presented here. Since the reader could be misled by the title, I think you need to change it to better represent the content of the text.

Is the communication clear enough for a newcomer ? I think that you have to work on making your text more concise and more homogenous in terms of the depth of explanations. More and better iconography would help in this regard. The iconography should follow the main organization of the text : here you have six sections and only two figures. Figure 1 illustrates many aspects of the section on shotgun metagenomics, and figure 2 is an illustration on the integration of taxonomy and function. Figure 1d does not follow the text flow and I find this a bit strange.

What is the review message ? In my opinion, the discussion on the integration of taxonomic and functional data is the main message. I advise you to strengthen this aspect by dropping some sections (see below).

How to make the message clearer ? If you decide to follow the path of considering the integration of taxonomic and functional data as the main message to deliver, then the text could be reorganized to make this message stronger and clearer. I wonder if the sometime high level of details provided regarding marker-gene sequencing, shotgun metagenomics and the characteristics of microbiome count data is really helping the reader. The text would benefit from being way more concise an more equilibrated among sections. In my opinion, you should seriously consider to skip the "primer" sections on marker-gene sequencing, metagenomic sequencing, characteristics of microbiome count data and microbial functions.

I found that the discussion is the best part of the text, maybe because I am not a complete newcomer to the field. Your personal account is worthy, and maybe you could make it more precise (e.g. parameter choice from local to global using which tool ?). The last two sections are the most informative parts and in this regard.

Accuracy : The terminology about microbiome is sound and corresponds to what has been previously discussed in the literature (Marchesi & Ravel, 2015). I found that the terminology used in the section microbial function is not always clear and does not simplify the presentation of the associated concepts (Karp, 2000)(Thomas, Mi & Lewis, 2007)(Kotera et al., 2014).

Level of referencing : There are specific experimental approaches such as epicPCR that have been developed to tackle the integration of taxonomy and function; and this needs to be pointed out (Spencer et al., 2016). I think you should take a particular attention to be more homogeneous in the way you select the cited references.

Minor comments

Since the review aims to deliver a detailed introduction, I suggest to expand a bit the terminology and definition that you provide rapidly for the term microbiome (one sentence on line 31-33), and possibly include a text-box with definitions. Maybe the ecological suffix -biome that refers to biotic and abiotic factors characterizing a given microbiome environment would broaden the scope.

I fully understand that the topic is DNA-based sequencing for microbiome studies, but a pointer to RNA-based and protein-based sequencing would be a plus in the background, especially in the paragraph 45-67. In that same paragraph on culturing microbes, and given the theme of the integration between taxonomy and function, one possible additional point could be to discuss the discrimination of live, dormant and dead microorganisms (e.g. (Thomas, Mi & Lewis, 2007) (Jones & Lennon, 2010)(Carini et al., 2016)(Blazewicz et al., 2013).

In the background section presenting diversity analysis, I would like to underscore the work of Amy Willis and colleagues on modelling abundances as in my opinion it is an important advance in the analysis of diversity (Willis, 2019)(Willis & Martin). The purpose of this paragraph in the context of the review as a whole is unclear as it stands.

I do not agree with the assertion that the dichotomy between phylogenetic and functional profiling of microbiomes is « entirely related to methodological challenges » (line 123). We know that the genome of prokaryotic species varies in gene content because of horizontal gene transfer, gene duplication and other mechanisms (Puigbò et al., 2014). It has been shown through pangenome analysis that strain variation can be associated with different metabolic potential (Goyal, 2018) (Maistrenko et al., 2020). Therefore, it seems to me that the dichotomy between phylogenetic and functional profiling of microbiomes is one of their intrinsic characteristics. Indeed, you develop these points line 1131-1172.

Marker-gene sequencing

I advise to simplify the marker gene sequencing section if you want to keep it. While the paragraph from 149-202 are detailed and very informative, I am afraid that they depart from the global « granularity » of explanation and historical context provided on other aspects throughout the manuscript. This lengthen this section on marker genes comparatively to the other aspects developed in this review. And even if there are a lot of things to tell about 16S rRNA gene sequencing, many have already been told elsewhere in the literature.

While I typically enjoy reading historical perspectives, I found that these are exaggeratedly long and placed in the manuscript in a non-logical manner.

You copiously present 16S rRNA gene sequencing and this helps the reader for understanding the aspects on the integration of taxonomic and functional data. But you also consider other marker genes (and this is fine) and 18s rRNA gene sequencing for microeukaryote and fungi taxonomic profiling, but in a more concise manner. Yet the integration of taxonomic data obtained using such markers with shotgun sequencing data is not presented at all, and thus the reader does not benefit from this otherwise interesting piece of knowledge.

The sentence line 211 would benefit from some simplification such as :

« This is because if there are non-random substitutions within a single domain but random substitutions in the majority of other domains, there would likely be little effect on estimates of gene divergence. »

I do not understand the reason for presenting redbiom at this point line 250 ?

To further document your point on the limitations due to the use of short 16S amplicons (line 260-274), you could possibly cite the recent work of other groups such as (Abellan-Schneyder et al., 2021).

The point dealing with the use of classical bacteria 16S primer-pairs do characterize Archaea could be expanded as it is often a neglected limitation in taxonomic surveys (Raymann et al., 2017; Bahram et al., 2019).

The reference Fox et al 1992 is missing at line 235. I think it would be fair to reference deblur and UNOISE3 like it has been made for DADA2 software (line 336).

Very Minor : italicize latin names (e.g Haloarcula line 382)

Shotgun metagenomics sequencing

Line 409 : including DNA viruses

The impact of biomass and genome size as a limitation to MGS approach could be invoked (line 431). Also as a caveat emptor, the impact of host DNA and possible heterologous sequences on MGS data could be mentioned, (I wrote this sentence before reading your discussion !) and this would be a reflection of the discussion.

In the MGS data analysis section devoted to the generation of taxonomic profile (line 477-522) , I would like to point out the targeted assembly of rRNA sequences from shotgun data embodied in Emirge (Miller et al., 2011), phyloFlash (Gruber-Vodicka, Seah & Pruesse, 2020) and MATAM (Pericard et al., 2018).

I was surprised that the authors do not mention Kaiju as a read-based tool for taxonomic profiling (Menzel, Ng & Krogh, 2016).

On the impact of databases for k-mer based analysis (Nasko et al., 2018).

Line 560 : the citation of only these two assemblers is somehow partial, you could point to a review on metagenome assembly for the sake of comprehensiveness for the reader. Similarly the description of binning tools is very light in comparison to other aspects developed earlier. Here you could point to recent review papers on the subject.

Line 584 : maybe use « taxonomic profiling » instead of « profiling »

Line 586 : I guess that the authors are referring to transcriptome studies, the term RNA sequencing is maybe not so precise in this context.

Characteristics of microbiome count data :

Maybe at some point the word abundance table could be used.

Line 618-637 : Maybe a figure would be a better communication vector.

The impact of sequencing reads processing on the analysis of abundance tables is somehow skipped : there are different practices such as removing singletons, filtering on prevalence etc . This could be somehow mentioned as they impact downstream analysis.

Microbial functions

This section is quite lengthy in comparison to others and since it covers topics that are not specific to microbiome studies, I wonder if it hits the sweet spot.

Line 737 : « … focused on gene families, which are gene clusters. » It is not very clear what you are referring to in terms of gene cluster at this point.

Line 781 : I do not know what is a UniRef function.

What is described in this paragraph entitled microbial function is in fact a primer on protein databases and ontologies. I find therefore that the title is a bit misleading, maybe « Protein databases and ontologies for microbial genome functional annotation ».

Line 976 : this method focuses pathway reconstruction … please correct the sentence.

Line 1032 : philosophical perspective : really ?

Line 1060 : The whole presentation of this paragraph is somehow paradoxal : maybe the text could be more explicit on ontology and semantics in order to guide the analysis of « functional data » at a given level of an ontology (protein space, biochemical activity, pathway, evolutionary conservation.

Metagenome prediction methods

Line 1090-1097 : some references would be welcome here.

Lines 1101-1110 -1130: This historical account is perfect, but I wonder if the level of details provided is really needed to make the point that 16S diversity is not a perfect proxy of whole genome similarity.

Current state of the integration of taxonomic and functional data types

I enjoyed reading this section.

Line 1313: “in some cases can be directly linked” Please be more precise and provide an example or a reference.

Why the burrito software is not mentioned is unclear to me ?

Outlook

In my opinion, the paragraph 1726-1761 would benefit from citing additional recent references such as the MBQC study and a few others: (Sinha et al., 2017; Davis et al., 2018; McLaren, Willis & Callahan, 2019; Greathouse, Sinha & Vogtmann, 2019).

References

I suggest to use a style for references that includes a DOI.

Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K. 2021. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. mSphere 6. DOI: 10.1128/mSphere.01202-20.

Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. 2019. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. Environmental Microbiology Reports 11:487–494. DOI: 10.1111/1758-2229.12684.

Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. 2013. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. The ISME journal 7:2061–2068. DOI: 10.1038/ismej.2013.102.

Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nature Microbiology 2:1–6. DOI: 10.1038/nmicrobiol.2016.242.

Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6:226. DOI: 10.1186/s40168-018-0605-2.

Emerson JB, Adams RI, Román CMB, Brooks B, Coil DA, Dahlhausen K, Ganz HH, Hartmann EM, Hsu T, Justice NB, Paulino-Lima IG, Luongo JC, Lymperopoulou DS, Gomez-Silvan C, Rothschild-Mancinelli B, Balk M, Huttenhower C, Nocker A, Vaishampayan P, Rothschild LJ. 2017. Schrödinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. Microbiome 5:86. DOI: 10.1186/s40168-017-0285-3.

Goyal A. 2018. Metabolic adaptations underlying genome flexibility in prokaryotes. PLOS Genetics 14:e1007763. DOI: 10.1371/journal.pgen.1007763.

Greathouse KL, Sinha R, Vogtmann E. 2019. DNA extraction for human microbiome studies: the issue of standardization. Genome Biology 20:212. DOI: 10.1186/s13059-019-1843-8.

Gruber-Vodicka HR, Seah BKB, Pruesse E. 2020. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. mSystems 5. DOI: 10.1128/mSystems.00920-20.

Jones SE, Lennon JT. 2010. Dormancy contributes to the maintenance of microbial diversity. Proceedings of the National Academy of Sciences 107:5881–5886. DOI: 10.1073/pnas.0912765107.

Karp PD. 2000. An ontology for biological function based on molecular interactions. Bioinformatics (Oxford, England) 16:269–285. DOI: 10.1093/bioinformatics/16.3.269.

Kotera M, Nishimura Y, Nakagawa Z, Muto A, Moriya Y, Okamoto S, Kawashima S, Katayama T, Tokimatsu T, Kanehisa M, Goto S. 2014. PIERO ontology for analysis of biochemical transformations: effective implementation of reaction information in the IUBMB enzyme list. Journal of Bioinformatics and Computational Biology 12:1442001. DOI: 10.1142/S0219720014420013.

Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, Rodrigues JFM, von Mering C, Pedro Coelho L, Huerta-Cepas J, Sunagawa S, Bork P. 2020. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. The ISME Journal 14:1247–1259. DOI: 10.1038/s41396-020-0600-z.

Marchesi JR, Ravel J. 2015. The vocabulary of microbiome research: a proposal. Microbiome 3:31. DOI: 10.1186/s40168-015-0094-5.

McLaren MR, Willis AD, Callahan BJ. 2019. Consistent and correctable bias in metagenomic sequencing experiments. eLife 8. DOI: 10.7554/eLife.46923.

Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications 7:11257. DOI: 10.1038/ncomms11257.

Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biology 12:1–14. DOI: 10.1186/gb-2011-12-5-r44.

Nasko DJ, Koren S, Phillippy AM, Treangen TJ. 2018. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biology 19. DOI: 10.1186/s13059-018-1554-6.

Pericard P, Dufresne Y, Couderc L, Blanquart S, Touzet H. 2018. MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. Bioinformatics 34:585–591. DOI: 10.1093/bioinformatics/btx644.

Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biology 12. DOI: 10.1186/s12915-014-0066-4.

Raymann K, Moeller AH, Goodman AL, Ochman H. 2017. Unexplored Archaeal Diversity in the Great Ape Gut Microbiome. mSphere 2. DOI: 10.1128/mSphere.00026-17.

Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Abnet CC, Knight R, White O, Huttenhower C. 2017. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nature Biotechnology 35:1077–1086. DOI: 10.1038/nbt.3981.

Spencer SJ, Tamminen MV, Preheim SP, Guo MT, Briggs AW, Brito IL, A Weitz D, Pitkänen LK, Vigneault F, Juhani Virta MP, Alm EJ. 2016. Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. The ISME journal 10:427–436. DOI: 10.1038/ismej.2015.124.

Thomas PD, Mi H, Lewis S. 2007. Ontology annotation: mapping genomic regions to biological function. Current Opinion in Chemical Biology 11:4–11. DOI: 10.1016/j.cbpa.2006.11.039.

Willis AD. 2019. Rigorous Statistical Methods for Rigorous Microbiome Science. mSystems 4. DOI: 10.1128/mSystems.00117-19.

Willis AD, Martin BD. Estimating diversity in networked ecological communities. Biostatistics. DOI: 10.1093/biostatistics/kxaa015.

Reviewer 2

In this article, the authors propose an overview of the use of different approaches for microbiome data analyses, the questions that can be tackled using them, and their respective limitations. A particular focus is provided on the bioinformatics aspects, including an overview of the diversity of the most popular tools, and under which conditions/for which specific purposes they could be better used. Taxonomic and functional assignations tools are thoroughly discussed. And the crucial question of how to integrate the taxonomic and functional aspects. How marker-gene and shotgun metagenomic sequences (MGS) data are currently linked is exposed and the limitations of different approaches given. How the two approaches can lead to contradicting results, as well as the recurrent problem of reproducibility on microbiome data when using different bioinformatics pipelines are thoroughly discussed. Some interesting leads on how to make the field of microbiome biology more robust are given.

The paper is very well-written and thorough on several aspects, including by explaining the main trends in "DNA-based microbiome data" analyses. It is very interesting both from the level of technical details that are given, and from the fact that it does the synthesis of the current major pitfalls in microbiome studies. I particularly enjoyed the "Overview" and the last section on "Current state of the integration of taxonomic and functional data types". As such, beyond proposing a view of the current state-of-the-art, I think this primer paper should contribute to the reflexion on what good practices could be taken, and which approaches are the most promising in order to make discoveries in microbiome studies more robust and reliable in the near future.

In general, I thought the titles of the big sections could be improved to better reflect their content. A few sections might require a bit of rewriting for clarification, and I would like to raise some points that are listed below.

1) The section "marker-gene sequencing", where the case of the 16S rRNA amplicon sequencing is discussed at length (which is interesting!), is mostly dedicated to the particular task of characterizing the diversity within a community. However, it is only on lines 254-256 that the goal for using what is described as "robust marker genes" is introduced: "to characterize and compare the relative abundances of prokaryotes across communities."

I think the 1st page of the manuscript could be re-arranged, and clarified to explain the particular usage of marker-gene approaches that is exemplified here.

- At the beginning of the section there is a discussion on the definition of a "robust marker gene". But I believe this line of discussion depends on the goal of marker-gene sequencing – that should thus be introduced beforehand. Marker-gene approach can also be taken to question the presence of given metabolic processes in a particular environment. In which case, it is more important to fish for genes that are specifically involved in that process, leading even sometimes to multiply the set of probes to use in order to capture the diversity of the gene involved in the process of interest (some are paraphyletic for instance). In that case, the fact that the gene in question is a good molecular chronometer does not matter much, right? Or did I miss the point here?

- Line 156: a more general term would be “homolog”, as “ortholog” limits to vertically transmitted marker genes (excluding duplicated or laterally transferred genes for instance). Unless if it is explained beforehand that a desirable property of a marker gene could be to be vertically inherited? Or is the term “ortholog” used here to suggest a conserved biological function? Please clarify.

- In the end, I have the feeling that the first part of this 1st section kind of falls flat, as the authors write on lines 200-201: “Therefore, to select a robust marker gene one should adhere in some ways to the Goldilocks principle: some nucleotide conservation is needed, but not too much.” Maybe could this first part be shortened and be more straight-forward?

2) Lines 270-272: Please clarify what you mean by “V4-V5 region overrepresented Firmicutes … while drastically underestimating Actinobacteria”. Do you mean that these regions are not present from Actinobacteria? Or that the diversity is over-estimated in Firmicutes and under-estimated in Actinobacteria based on this region? Same comment for line 290-291 for V1-V2 region.

3) In the section “Shotgun Metagenomics Sequencing”, I felt like the topic of the contribution of MGS approach and MAG (metagenome assembled genomes) reconstruction to explore extant biodiversity was somehow missing (CPR, DPANN, Asgard archaea…). MGS helped to reveal novelties both at the taxonomic and functional level. As a conceptual advantage of the MGS approach, in spite of some biases highlighted by the authors, is that it is not needed to have an a priori of what is looked for. This is how some entire clades of archaea were missed by 16S approaches because of the probes being designed from known diversity (e.g. Raymann et al 2017, mSphere).

- On lines 443-447 an example is given for taxa represented in 16S data but not MGS. To be fair, the converse is also true. I don’t say the authors do not explicitly mention that there are caveats with both approaches, but this is one could be worth to be reminded.

- On lines 1004-1013, it could be added that techniques to bin MGS data as MAG could be a part of the solution.

4) On “the concordance of differential abundance results between actual and predicted metagenomics profiles” (lines 1882-1294), any lead on why the results are agreeing only “moderately well”?

5) Just a suggestion… Some figures could have been added to illustrate some parts of the text.

- On lines 1222-1225, the principle on which relies PICRUSt for inferring function is introduced. It could have been illustrated by a figure.

- On lines 1409-1412, “stacked barplots” are mentioned to be used to study functional shifts. Such a typical plot could have been borrowed from a published study for instance?

6) In the Discussion part, it would have been interesting to have the authors opinions on the role that could play new sequencing techniques in the future to help with some of the issues presented? For instance, on the advent of long-reads sequencing for MGS? Don’t you think it could eventually be a way to integrate taxonomic and functional analyses, by linking for instance 16S genes to big contigs, obtaining better quality MAGs, etc…?

7) Minor points and typos:

- A list of abbreviations should be included to help the reader. Otherwise, some of the less used abbreviations could be abandoned?

- Line 158: should it be “twice” instead of “double”?

- Line 1441 (and thereafter): maybe capitalize the tool name “phylogenize” to make it stand as a name in the text?

- Line 1445: “a taxa” => should be corrected by “a taxon”.

Reviewer 3

This review addresses many of the technical issues in the microbiome field. The text is very clear and concise, and it is very interesting for both initiated and uninitiated readers.

In general, the main point of the MS is the challenge of integrating taxonomic data with functional data. I agree that this is an issue but I feel in general the review downplay too much the binning/MAG approach dealing with this issue. I also missed in the text any discussion regarding long reads and how the 3rd generation sequencing methods could help with some of the limitations.

I have a few small comments that could improve the final version of the MS.

Line 107: There is often more statistical power to detect overall differences based on alpha and beta 108diversity metrics than to detect associations with individual features, but diversity-level insights are also less actionable (Shade 2017).

- However, often the difference of abundance in individual taxa/rank is larger than the difference in diversity indexes, especially in host-microbiome studies.

Line 422: This interest has culminated in the generation of enormous MGS datasets such as the ongoing work on the Earth Microbiome Project (Thompson et al. 2017) and the Human Microbiome Project (Lloyd-Price et al. 2017).

- Here another good and more recent example would be TARA oceans.

Line 548: “genes are expressed in cells, not in a homogenized cytoplasmic soup" (McMahon 2015).

- Agreed, however many ecological functions are performed in a collaborative way by consortiums.

Line 670: relative abundances by the mean relative abundance

- Should read geometric mean.

Line 723: This discussion of microbiome data characteristics has focused on taxonomic features based on either 16S sequencing or read-based MGS data analysis. However, it is important to emphasize that count tables produced from MAGs do not resolve this issue. In fact, attempting to account for these challenging characteristics of microbiome count data and the links between taxa and function makes the analysis more difficult.

- At the end of this, I would suggest a few lines about the network of co-abundances, for example using the SparCC tool.