We would like to thank the recommender and all three reviewers for helpful feedback. Based on your comments we have made numerous changes to the manuscript that we believe have greatly improved it. Please see our point-by-point responses below.

- Gavin Douglas and Morgan Langille

Recommender's comments

Generally, the reviewers and I found the manuscript relevant and interesting. I do agree with the first reviewer that occasionally there are distracting facts, of added value, that reduce the usability of the manuscript as a "guide for the novice". I do not suggest removing these but rather relocating them to a box. For example - the necessary traits of a marker gene are good to know, but realistically, most people embarking on the metabarcoding adventure will initially embrace known markers.

Thank you for the feedback – we have shifted several general topics to be in boxes. Specifically, the historical and extra information on marker genes and the 16S rRNA gene are now provided in Box 1. We have also moved our discussion on comparing taxonomic and functional stability to Box 2. Finally, we have shifted our descriptions of early DNA hybridization and 16S studies that investigated genomic vs phylogenetic diversity to be in Box 3.

With this respect, I feel that the paper can be made somewhat more concise.

As a primer - I suggest adding a glossary and to minimize abbreviations as much as possible.

We have added a glossary and removed many abbreviations, as mentioned in our response to reviewer 2 below.

Last, it is evident that the authors come from the field of human microbiome and so are most of the examples. I suggest adding a paragraph where this is specified clearly, explaining how the provided guidelines can be applied to microbial ecology in other types of environments (e.g. water, soil, biofilms, etc).

We have added this sentence to the final paragraph of the Background section:

> *"Although many of our examples are taken from the human microbiome literature, our key points and suggestions are relevant to research in any microbial environment."*

Recommender's in-text comments:

P2,L65-67: This is a very bold statement. The question is not whether all bacteria can be cultivated but also what is the effort needed to culture the majority of bacterial diversity and whether culturing one strain reflects the metabolic potential of the entire species/genus, something which may be better reflected in a genus / species-wide pangenome.

We agree that the other reasons you bring up are also important benefits of DNA sequencing vs. culturing. We have re-written the following paragraph to highlight those points:

> *"Despite these advances, DNA sequencing has several advantages over culturing. First, 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. DNA sequencing is also often a less time and labour-intensive method for assessing overall community diversity, although high-throughput culturing methods are becoming more common (Watterson et al. 2020). This is important, because high-throughput characterization of microbial communities is key to understanding microbial diversity, as closely related organisms can drastically differ in metabolic potential (Welch et al. 2002; Tettelin et al. 2005). For these reasons, DNA sequencing remains the predominant method for characterizing microbial communities, although it is well-complemented by culturing (Lau et al. 2016)."*

P2,L72L This is not, in my opinion, the main purpose or advantage of DNA sequencing. Abundance can be obtained by different FISH methods which in fact will give much more precise quantitative answers.
DNA sequencing allows provide a profile of the community with some quantitative aspects, it also allows us to generate hypotheses about processes and community functionality, diversity etc. Also as one reviewer mentioned, metagenomics also allows us to explore into the unknown,

We agree that methods like FISH are also useful for obtaining abundance information, but that paragraph is contrasting sequencing and culturing specifically. Regarding the second point - see our above changes to the paragraph, which now points out several other advantages of DNA sequencing as well.

P2,L87: This is a very - human oriented microbial ecology, whereas this review aims at a broader audience.

We agree that this was too specific, we have changed it to be:

> *"...between sample environments (e.g. locations, disease states, etc.)..."*

P2,L123-124: Not entirely. Taxonomic profiling is a tool to analyze diversity. For Alpha / Beta diversity we need to get a count of distinct entities and their relative (absolute) abundance. The functionality is derived from out inability to link between taxonomy/phylogeny and function a problem that does not exists in macro-ecology - one does not need the genome of a lion to know it's a top predator.

We agree and have changed the wording to "partially" rather than "entirely".

P3,L135: and the lack of proper databases

Added.

P3,L203-204: also known as SSU rRNA - I would rather refer to this and mention that it is 16S and 18S in Bac or Euk.

We would prefer to keep most of the discussion focused on the 16S specifically rather than SSUs in general, because most of the literature discussed is focused on the 16S specifically and in many cases is not applicable to the 18S (e.g., the specific biases across variable regions). However, we do mention that the 16S is an SSU and that the 18S is the homolog of the 16S in eukaryotes.

P4,L290-298: Not being from the field of human microbiome, I wonder if these differences have been assessed with FISH based counts.

This is an interesting question – we're also not familiar enough with the vaginal microbiome literature to know for sure whether FISH based counts are commonly used, but we do not believe so.

P4,L318: This is/was true - more recent approaches use 98.5 / 99 %. I suggest pointing here to the fact that the bigger problem is 16S does not correctly represent speciation. See the classical 16S vs DNA-DNA hybridization in Rosello-Mora and Amann 2001 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-6976.2001.tb00571.x

Thanks for the reference. We have changed the range to be "97-99%" rather than just "97%". We believe our discussion of DNA-DNA hybridization later in the manuscript covers your second point.

P5,L384: there are many similar examples from Bacteria as well

We agree that there are many such examples, but we think the 16S sequences of ***Haloarcula*** serve as a useful illustrative example.

P5,L402: rpoB was suggested as well a general marker to replace 16S
https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-019-1546-z

Thanks for pointing out this gene. We have added this text to highlight it:

> *Similarly, the gene rpoB, which encodes the DNA-directed RNA polymerase subunit beta, is another valuable prokaryotic marker gene, which provides comparable or better taxonomic resolution to the 16S rRNA gene (Case et al. 2007). Profiling rpoB can sometimes better identify relevant taxa in a community. For instance, it has been used for identifying a known nematode symbiont missed by standard 16S profiling with the V3-V4 region (Ogier et al. 2019).*

P8,L568: do you refer here to tertranucleotide? That would be 4-mer.
3-mer would more refer to codon-usage.

We have removed the reference to 3-mers here as many different k-mer sizes can be used.

P8,L580: contamination or strain heterogeneity (or heterozygosity) depending on phylogenetic distance between markers.

We now mention redundancy could represent strain heterogeneity as well.

> *"...the redundancy, which is potential evidence for contamination or strain heterogeneity in the genome..."*

P9,L730: As the authors to some extent advocate for lab work in the introduction, a solution to this problem in the form of cell counts (total / DAPI / FISH) to calibrate the data, should be mentioned and encouraged. The statistics are there to help when these data are unavailable.

We agree that there are great techniques being developed to better link absolute counts to relative abundance data. We now point to some of the recently developed methods for this purpose:

> *"An alternative solution is to leverage additional data to transform relative abundances to absolute abundances. This alternative data could be quantitative PCR data, flow cytometry data (Vandeputte et al. 2017) or spiked-in sequences of known abundances (Zemb et al. 2020). Different sample preparation protocols prior to DNA sequencing can also help retain information about differential absolute amounts of DNA across samples as well (Cruz et al. 2021). These are exciting approaches, but they have not been validated across many datasets and at the moment there is no consensus regarding which methods perform best."*

P12,L1067: here metagenome refers to marker gene as opposed to shotgun sequencing, I find the two terms too close to each other and many people use them as synonyms. I would change the title and the term to Marker-gene based prediction or phylogeny/taxonomy based predictions.

We actually intended metagenome to refer to the sense used as for shotgun sequencing because they are predicted metagenomes, rather than actual metagenomes. To clarify this point we have changed the section title to be "Marker-gene-based metagenome prediction methods"

P15,L1370: What is IBD? Not defined earlier and also does not appear again -can this be avoided?

IBD stands for "Inflammatory Bowel Disease", which we now write out.

P16, Figure 2 legend: Wouldn't this be obtained by looking at the diversity of the proteins between. i.e. highly expressed gene with high similarity in nucleotide and AA sequence vs high diversity. Taxonomic information may shed information on functional redundancy and who are the multiple players that carry out a similar function.

In principle this is true, but there are methodological challenges in performing an analysis like that. One major challenge is that proteins will have different rates of evolution which would make comparing relative AA diversity difficult without additional information. Horizontal gene transfer might also artificially make it seem like there are a more diverse set of contributing taxa then there are in actuality. Nonetheless, this is definitely a good idea and we think a bioinformatics pipeline along these lines could be valuable!

P16,L1461: consider to put in italics or underline to clarify that this is a name of a tool.

The tool phylogenize is now in italics throughout the manuscript.

P19,L1740-1743: I believe here the authors should highlight the development of enzymatic lysis protocols which are essential for long read sequencing which in turn may solve a lot of the issues associated with taxonomic annotation of shotgun metagenomic data

We have added these sentences to that paragraph:

"Common extraction protocols also often result in high rates of DNA fragmentation, which makes the extracted DNA less appropriate for long-read sequencing technologies. Updated extraction protocols based on robust enzymatic lysis have been developed to address this problem (Maghini et al. 2020)."

P20,L1873-1876: Such work was recently presented in the DNAqua conference. I am not sure where this was published. The conclusion was that dada2 is for now the best option whereas others remove too much or inflate biodiversity.
The abstracts of that meeting are here https://aca.pensoft.net/collection/244/. I am sorry but I cannot pinpoint to the exact talk.

Thanks for mentioning this work. We found similar findings in our own tool comparison paper (Nearing et al. 2018), but we disagree that this means it is the best option – that would require knowing whether taxa at low frequencies are true positives or not. However, even if a consensus is reached regarding DADA2 compared with other denoising tools, there remain many other areas in microbiome analysis whether there is no consensus regarding tool choice. Accordingly, we believe our current statements would still stand either way.

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

Reviewer 1

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.

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.

We acknowledge that there are many topics covered in the manuscript that might be challenging for newcomers to the field. However, we do respectfully disagree that it is inadvisable to provide a primer and discussion in the same article. There are many contentious issues in the microbiome field, and we believe highlighting the many unknowns and the disagreements in the field are just as important to explain to newcomers as are the basics of sequencing data types.

In addition, many of the revisions we have made now make the manuscript more appropriate as a primer. These additions include a glossary, distinct boxes for tangential content, and more figures.

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.

To make the presentation more homogenous we have moved several blocks of text to independent boxes. We have also added three new figures. One of these figures is now what used to be Figure 1d (i.e. a diagram of genome prediction based on a 16S sequence), which is now in the appropriate location of the article.

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).

Rather than dropping sections we have elected to move certain blocks to boxes (see above).

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.

We agree that this would be one way of presenting the manuscript, but we respectfully disagree that it would be better to remove the primer sections. We think that including these sections makes the

overall work more broadly useful. In other words, the major take-home messages regarding taxa-function integration and the reproducibility crisis in microbiome research will be more appreciable to newcomers after they have been introduced to the basics of data analysis in the field.

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.

We now specify that the tool was MetaPhlAn2.

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).

We have clarified the text in the section in several places (see below). At the beginning of the section we have specifically endeavoured to be clearer with our definitions. These are the key sentences that were added/edited:

> *"However, rather than individual gene sequences, research is typically focused on gene families, which are defined based on close sequence identity and/or similar functionality from the gene's eye view. Alternatively, the focus is sometimes on higher-order functional categories like pathways, which represent functionality of groups of potentially interacting gene families in reactions. To complicate matters further, there are several different functional databases and ontologies for annotating microbial functions. Ontologies are representations of information groupings and relationships of arbitrary entities (Thomas et al. 2007). In the context of functional annotation, different ontologies represent different ways of functionally grouping genes and also of defining higher-level and more general microbial functions."*

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.

Thanks for pointing us to epicPCR and for your more general suggestion. We have added numerous new references to address the points you and the other reviewers have brought up, which we believe has improved the breadth of referencing. We have also added this sentence that points the reader to

epicPCR if they are interested in linking taxa with small numbers of functions (rather than linking entire genomes with taxa, which is the key focus of our discussion on taxa and function integration):

> *"Approaches that expand on basic marker-gene sequencing, such as epicPCR (Spencer et al. 2016), can provide information on the presence of small numbers of genes linked to marker genes, but generally only limited genomic content can be gleaned from these methods."*

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.

Thanks for your suggestion. We have added a glossary which will help readers keep track of definitions. We have also added this sentence when introducing microbiomes:

> *As the suffix "biome" suggests, a microbiome refers to these constituent elements in a defined habitat (Berg et al. 2020).*

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).

Thanks for your suggestions, we have added these sentences:

> *"This method does have disadvantages however, for instance, it is difficult to distinguish between live, dormant, and dead cells using DNA sequencing (Carini et al. 2016). For researchers specifically interested in profiling active cells, leveraging alternative techniques such as metatranscriptomics, metaproteomics, and/or culturing, would be more appropriate."*

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.

We agree that the work of Willis and colleagues has resulted in useful findings on how (and whether) to analyze alpha diversity. Accordingly, we have added this sentence to point readers to this work:

> *"In addition, many diversity metrics rely on unrealistic assumptions and there has been a recent push to develop more robust methods (Willis 2019; Willis and Martin 2020)."*

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.

We agree entirely and have changed the wording to be "partially related". Our intended meaning was that when both taxa and functional data are available for the same communities that the reason that they are not jointly analyzed is often because it is methodologically difficult.

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.

Thanks for your feedback on this section. We have made several corrections (see below), including moving the historical background and motivation for marker-gene sequencing to be in Box 1.

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. »

Swapped in this sentence.

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

This sentence has been removed.

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).

Citation added.

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).

We have added this sentence:

> *"This is particularly true for less widely surveyed taxa such as archaea, which traditionally have been difficult to detect with 16S rRNA gene primers designed for bacteria (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).

Fixed all.

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

Fixed.

Shotgun metagenomics sequencing

Line 409 : including DNA viruses

Fixed.

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.

We have added these sentences to the first paragraph of this section which we believe hits on the key points you raise:

> *"This advantage also makes data analysis more challenging. This is particularly because sources of DNA that are not of interest, such as host DNA or contaminants (especially in low biomass samples), can often be substantial proportions of MGS datasets."*

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).

Thanks for pointing out these tools, we now cite them and have added this sentence:

- *"Several methods have been developed specifically for targeted assembly of this and other rRNA genes from MGS data (Miller et al. 2011; Pericard et al. 2018; Gruber-Vodicka et al. 2020)"*

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

We now mention Kaiju:

"In contrast, when reads are mapped against reference genomes in protein space, using a tool such as Kaiju (Menzel et al. 2016), this approach can provide useful taxonomic profiles."

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

Thanks for pointing out this paper. We now refer to it:

"One disadvantage of such methods is that taxonomic classifications can be highly dependent on the size of the database used (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.

We have added this sentence:

"For further details on metagenomics assembly and binning tools, readers can find recent reviews that describe the available bioinformatics tools (Breitwieser et al. 2019; Ayling et al. 2020)."

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

Changed.

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

Changed to now refer to transcriptome data specifically.

Characteristics of microbiome count data :

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

Changed wording from "count tables" to "abundance tables"

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

Removed this text and added Figure 2 to communicate this point.

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.

This is a good point and we now refer to data filtering with these lines:

> *"To compound these discrepancies, there are even disagreements regarding how to preprocess and filter datasets prior to statistical testing. For instance, microbial features with low prevalence or that are only found at low read depths are often discarded. Ad-hoc cut-offs for feature filtering, such as a minimum prevalence of 10%, are often used, but there is little consistency across studies. In addition, it has been suggested that filtering out rare features based on read depth can, at least under certain conditions, reduce statistical power (Schloss 2020)."*

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.

We have shifted some of the text to a separate box that helps cut down on the section length.

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.

We have changed the wording to be:

> *"...which are defined based on close sequence identity and/or similar functionality from the gene's eye view."*

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

These are clustered protein sequences that are part of the Universal Protein Resource database and are described in the preceding paragraph.

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 ».

We have changed the section title to be "Protein databases and ontologies for microbial genome functional annotation".

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

Fixed:

> *"Pathway reconstructions from this tool are based on the distinction between genes that are shared with multiple pathways from those that are unique to a single pathway."*

Line 1032 : philosophical perspective : really ?

Yes – you can see the cited paper (in Biology and Philosophy) for details.

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.

We have added in more specifics and cleared up the language in this and the preceding paragraph. Specifically, the transition between the paragraphs now reads like this:

> *This difficulty is largely because it is unclear which levels of granularity would be meaningful to compare between each data type. For instance, the gene and pathway perspectives of function represent two extremes of functional granularity. Many different functional ontologies exist as well for defining functional groups, as discussed in the main-text. Because taxa and functional data types are qualitatively different from each other and the choice of how to compare the two is based on somewhat arbitrary decisions on how to categorize them.*

This can be illustrated by comparing taxa and functions in the same communities based on different groupings of each data type. As described in the main-text, the sparsity and number of possible functional categories differs drastically across ontologies and sub-categories.

Metagenome prediction methods

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

We now cite these statements.

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.

We have shifted most of the historical account to be in Box 3 to make the section flow better.

Current state of the integration of taxonomic and functional data types:

I enjoyed reading this section.

Good to hear!

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

We now follow this statement with: "(e.g., in metagenome-assembled genomes)".

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

We now mention BURRITO in this sentence:

"However, it is becoming common to visualize stacked bar-charts of taxonomic contributors to functions of interest (see Figure 5 for examples), ***which can be created with tools such as BURRITO (McNally et al. 2018)."***

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).

Thanks for pointing us to these papers. We now cite all of them except for Davis et al. 2018, which we think is a little outside the scope of that section.

References

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

We would prefer to leave the DOI out of the references simply because we do not have them recorded for most cases and they would need to be added manually to our LaTeX file at this point.

References supplied by the reviewer:

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.

Thanks for your feedback. In addition to our rewrites described below we have changed several section titles, which we believe has helped make the manuscript clearer.

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.

We have re-arranged that section and clarified the particular type of marker-gene approaches that we discuss (see next point). The key change to this section was to move much of the preamble to a separate box.

- 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?

We only intended to introduce marker-gene sequencing for taxonomic profiling. Accordingly, we have changed the section title to be "Phylogenetic marker-gene sequencing" and added this sentence to clarify:

> *"These genes can be markers of a particular functionality (for example, Hug and Edwards 2013), but more commonly this approach is employed to taxonomically profile a community, which is the purpose that we will discuss."*

- 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.

We believe that "homolog" also works in that sentence and have swapped it in.

- 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?

We believe that moving the section referred to here to a separate box has greatly made the section 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.

We have written out our intended message more clearly in this paragraph, which now reads:

"We found that sequencing the V4-V5 region resulted in a higher proportion of Firmicutes and Bacteroides and a lower proportion of Actinobacteria, compared with the expected abundances. In contrast, sequencing the V6-V8 region resulted in a higher proportion of Proteobacteria and a lower proportion of Bacteroides. These biases highlight that choice of variable region can depend on which taxa are of interest. This is particularly true for less widely surveyed taxa such as archaea, which traditionally have been difficult to detect with 16S rRNA gene primers designed for bacteria (Bahram et al. 2019). The V4-V5 region was recently shown to be superior to region V6-V8 for studying 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 as it could be used to amplify the 16S rRNA gene of more archaea."

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).

We now highlight this advantage in the third sentence of the section:

"This has enabled the discovery of novel lineages, including previously unknown phyla (Spang et al. 2015), through analyzing MGS data."

- 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.

We agree the converse can be true and indeed we think this is reflected when we point out that eukaryotes and DNA viruses can be identified by MGS. However, we have added this line to clarify that this discrepancy between the numbers of phyla identified is not set in stone:

"This particular result is likely dependent on the taxonomic profiling approach used (see below)."

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

We now point out that MAGs could be the "individual members of the microbiome" that we referred to.

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"?

We think this is likely due to the reasons outlined in the beginning of the section: that there is substantial genomic variation between closely related prokaryotes and that 16S sequence identity is only approximately linked with genomic content.

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.

We have moved a panel from Figure 1 to this section to describe genome prediction based on 16S sequences in general (now Figure 3)

- 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?

We have added a new figure (Figure 5) which shows example stacked bar charts.

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...?

See below for our reply to reviewer #3 regarding a separate paper we have written that covers these topics in more detail.

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?

We have written out all of the less common abbreviations in our paper, which included: "ASVs", "CD", "CLR", "EC", "HMM", "HMP", "HSP", IBD", "ITS", "KOs", "MAGs", "mrcA", "mRNA", "OTUs", "SCFAs", "SSU", and "USCGs".

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

We believe both words would be correct in this case.

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

We have elected to italicize this tool name to make it stand out, as suggested by another reviewer.

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

Changed.

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.

Thanks for your feedback. The main reason this review does not focus on MAGs and 3rd generation sequencing technologies is we covered both topics in detail in another recent review that we published in GBE:

GM Douglas and MGI Langille. 2019. Current and Promising Approaches to Identify Horizontal Gene Transfer Events in Metagenomes. https://doi.org/10.1093/gbe/evz184

We now cite this paper in our revised manuscript and explain that readers can look there for further details:

> *"One final consideration is that several recent technologies have been developed that can lead to higher quality metagenome-assembled genomes. These include long-read sequencing technology (McCarthy 2010; Mikheyev and Tin 2014), Hi-C sequencing (Belton et al. 2012), optical mapping (Hastie et al. 2013), read clouds (Bishara et al. 2018), and single-cell metagenomics (Xu and Zhao 2018). We have previously discussed the utility of these specific technologies in the context of producing improved metagenome-assembled genomes in more detail (Douglas and Langille 2019)."*

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 diversity 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.

We agree, but nonetheless there is often more statistical power to detect diversity-level differences simply because there are fewer tests that need to be corrected for (whereas when testing for individual features there are often hundreds or thousands of tests).

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.

Now reads as:

> *"...the Earth Microbiome Project (Thompson et al. 2017), the Human Microbiome Project (Lloyd-Price et al. 2017), and the TARA Oceans investigations (Sunagawa et al. 2015)."*

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.

Agreed, but we believe it is a mistake to assume that all proteins are freely interacting across species in a community, which is the point we were trying to make.

Line 670: "relative abundances by the mean relative abundance"

- Should read geometric mean.

Added.

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."

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

We now refer to such approaches explicitly earlier in the section:

"...compositional approaches have been developed. For instance, several compositional correlation approaches have been developed (Friedman and Alm 2012; Kurtz et al. 2015; Schwager et al. 2017). One such approach is SparCC, which computes inter-taxon correlations while accounting for artifactual correlations that occur simply due to the interdependency between features in the same compositional dataset (Friedman and Alm 2012). Differential abundance approaches appropriate for compositional data analysis have also been developed, such as..."