Sequencing Introduced False Positive Rare Taxa Lead to Distorted Microbial Community Diversity, Assembly, and Interaction Interpretation in Amplicon Studies

Yangyang JIA

BGI-Shenzhen: BGI Group <https://orcid.org/0000-0002-9324-5052>

Shengguo Zhao

Chinese Academy of Agricultural Sciences Institute of Animal Science

Wenjie Guo

BGI-Qingdao

Ling Peng

BGI-Qingdao

Fang Zhao

BGI-Qingdao

Lushan Wang

Shandong University

Guangyi Fan

BGI-Qingdao

Yuanfang Zhu

BGI-Qingdao

Dayou Xu

BGI-Qingdao

Guilin Liu

BGI-Qingdao

Ruoqing Wang

BGI-Qingdao

Xiaodong Fang

BGI

He Zhang

BGI-Qingdao

Karsten Kristiansen

University of Copenhagen

Wenwei Zhang

BGI-Shenzhen: BGI Group

Jianwei Chen (✉ chenjianwei@genomics.cn)

<https://orcid.org/0000-0003-2332-4609>

Research

Keywords: Amplicon sequencing, Microbial rare taxa, Community assembly, Keystone species

Posted Date: October 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-965102/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Title: Sequencing introduced false positive rare taxa lead to distorted microbial community diversity, assembly, and interaction interpretation in amplicon studies

Yangyang Jia^{1,2,†}, Shengguo Zhao^{3,†}, Wenjie Guo², Ling Peng², Fang Zhao², Lushan Wang⁴, Guangyi Fan^{1,2}, Yuanfang Zhu², Dayou Xu², Guilin Liu², Ruoqing Wang², Xiaodong Fang¹, He Zhang^{1,2,5}, Karsten Kristiansen^{6,7*}, Wenwei Zhang^{1,2,*}, Jianwei Chen^{2,6,7*}

¹ BGI-Shenzhen, Shenzhen 518083, China

² BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China

³ State Key Laboratory of Animal Nutrition, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China

⁴ State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao, 266237, China

⁵ Department of Biology, Hong Kong Baptist University, Hong Kong SAR, China

⁶ Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Universitetsparken 13, 2100 Copenhagen, Denmark

⁷ Qingdao-Europe Advanced Institute for Life Sciences, BGI-Shenzhen, Qingdao 266555, China

[†] These authors contribute equally to this work.

* Correspondence: kk@bio.ku.dk, zhangww@genomics.cn, chenjianwei@genomics.cn

Abstract

Background: Due to the inherent scarcity of the microbial rare taxa, it is difficult to distinguish between *bona fide* rare taxa and potential false positives in metabarcoding amplicon sequencing studies. Although recently developed quality control and clustering or denoising algorithms could remove sequencing errors or non-biological artifact reads, no algorithm could remove high quality reads from sample-wise cross contaminations introduced by index misassignment.

Results: We thoroughly evaluated the rate of index misassignment of the mostly used NovaSeq 6000 and DNBSEQ-G400 sequencing platforms using both commercial and customized mock communities, and observed significant higher (5.68% vs 0.08%) fraction of potential false positive reads for NovaSeq 6000 compared to DNBSEQ-G400 in. Significant batch effects could be caused by the randomly detected false positive or false negative rare taxa. These false detections introduced by index misassignment could also lead to inflated alpha diversity for relatively simple samples while underestimated alpha diversity for complex samples. Further test using a set of cow rumen samples reported differential rare taxa by different sequencing platforms. Correlation test of the rare taxa detected by each sequencing platform demonstrated that the Novaseq 6000 platform detected rare taxa had a much lower possibility to be correlated with the physiochemical properties of rumen fluid. Community assembly mechanism and microbial network association analysis indicated that false positive or negative rare taxa detection could lead to distorted community assembly process and identification of even fake keystone species of the community, one of which was confirmed negative by PCR cloning and following Sanger sequencing.

Conclusions: We highly suggest proper positive and negative controls and technical replicate

45 settings, and proper sequencing platform selection in future amplicon studies, especially when
46 the microbial rare biosphere should be focused.

47

48 **Keywords:** Amplicon sequencing, Microbial rare taxa, Community assembly, Keystone
49 species

50

Introduction

Microbial communities in various environments are usually composed of a skewed abundance of microbes with a few highly dominant taxa and numerous rare taxa as revealed by the “long tail” of the rank-abundance curve [1,2]. The rare taxa, also known as the microbial “rare biosphere” [3], although exist in very low relative abundances, play important ecological roles in microbial communities. One of the most important roles is “seed bank” or “hidden backbone of microbial communities” in maintaining the stability and robustness of microbial communities [4]. For example, rare taxa were claimed to be the major driver of and played over-proportional role in soil multifunctionality [5]. Some rare taxa play crucial certain ecological functions in various biogeochemical process and human health [4]. For example, Michael Pester and colleagues demonstrated that *Desulfosporosinus*, despite detected with a relative abundance of less than 0.006%, had a fundamental role in sulfate reduction in a peatland ecosystem [6], and maintained high cellular activities under *in situ*-like conditions in lab [7]. Paul LE Bodelier and colleagues found the dynamics and consumption of methane in a wetland was correlated to the rare methane-oxidizing bacteria [8]. In addition, correlation network analysis of the potential interactions among the microbes revealed that some rare taxa were keystone species in a changing aquatic ecosystems [9]. Furthermore, several of the previous studies claimed that some of the rare microbes in various environments, such as ocean [10], and anaerobic digesters [11], even air [12] showed disproportionate high metabolic activities compared to their low relatively abundances.

Although rising interests of the microbial rare taxa have been observed in recent years, our knowledge of the rare fraction of microbial communities is still in its infancy. Because of current sequencing depth and the limitation of bioinformatic algorithms of shotgun sequencing [13,14], amplicon sequencing, normally targeting one or several of the highly variable region(s) of the small subunit ribosomal ribonucleic acid (SSU rRNA), still represents one of the most important and accessible ways in microbial study, especially for projects with thousands of samples [15,16]. However, the most challenging part in studying the rare biosphere is to distinguish between sequences from the *bona fide* rare taxa and sequence artefacts introduced by PCR or sequencing error or potential false positives. As reviewed by Michael D. J. Lynch and Josh D. Neufeld [13], sequencing errors introduced by low quality bases or ambiguous bases could normally be removed by setting stringent quality filtering threshold and clustering sequencing reads into Operational

Taxonomic Units (OTUs) with certain sequence identity [17]. Post-clustering curation algorithm was also developed to remove sequencing introduced errors [18]. Chimeric sequences generated during PCR could normally be removed by bioinformatic algorithms [19]. Recently developed denoising algorithms, including DADA2 [20], Deblur [21] and Unoise3 [22], facilitated the analysis of amplicon sequences at the exact sequence variant level [23], revealing the microbial community compositions with finer resolution, and could eliminate part of the sequencing errors during clustering.

However, despite the developments of algorithms and analyzing methods, all the high-throughput amplicon sequencing data filtering efforts made previously were focused on removing “artefacts”, that is non-biological sequences generated during the experimental process. None of the algorithms could remove potential index-hopping or misassignment of index caused sample-wise cross contaminations among samples pooled and sequenced in the same run, as they are high quality reads, not errors [24,25]. According to the Illumina official white paper [26] and other research articles, index misassignment could occur at a rate of 0.2~6% or even higher on various Illumina sequencing platforms, causing potential misinterpretation of the sequencing results, especially when clinical diagnose was made based on scarce mutations or the microbial rare biosphere was focused [25,27–29]. Platforms using different sequencing technologies could have different *pros* and *cons*. For example, frequency of index misassignment of the DNBSEQ platform, which used a combinatorial Probe-Anchor Synthesis method and DNA nanoball sequencing technology developed by MGI, was demonstrated to be as low as 0.0001 to 0.0004% [30]. Studies evaluating these different sequencing platforms in whole genome sequencing have been conducted by researchers worldwide [31,32]. In metagenomic studies, index misassignments have also been demonstrated to be an overlooked source of error in metabarcoding amplicon studies using pyrosequencing [24] or Illumina sequencing technology [33,34] for a decade. However, there are currently still very few studies evaluating how index misassignments could affect the results and data interpretation of high through-put amplicon sequencing in microbial ecology systematically [35,36], even when the rare taxa, the relative abundance of which was normally defined as lower than potential false positive rate of the sequencing technology, are focused.

To address these questions, in the present study, commercial and customized mock communities, and microbial communities with differential complexity from several typical

ecosystems were sequenced at two different mainstream sequencing platforms to show how index misassignment could interfere the interpretation and understanding of the rare taxa in various microbial communities. In addition, a real case study of cow rumen microbial communities further demonstrated that index-misassignment could lead to biased microbial compositions, community assembly, and ecological roles of the microbial rare biosphere. Finally, best practices for both experimental setting and data processing were suggested to eliminate potential false positives introduced by index misassignments in amplicon studies.

Results

Less batch effects and false positives on DNBSEQ-G400 platform

In order to evaluate the frequency of potential false positives that might be introduced by index misassignment in amplicon sequencing, commercial mock community ZymoBIOMICS™ Microbial Community DNA Standard with known composition (including eight bacterial members with their relative abundances ranging from 4.2% to 18.4%, **Table S1**) and two customized mock communities containing four or seven bacterial members (**Table S2, Table S3**) were subjected to amplicon sequencing of the 16S rRNA gene V4 region using both Illumina NovaSeq 6000 and MGI DNBSEQ-G400 platforms, respectively. Triplicate PCR reactions, library constructions, and sequencing runs were carried out to evaluate the reproducibility and potential batch effects of each sequencing platform (**Figure S1**). For the commercial mock community, a total of 17 (14, 15, 16 for each replicate respectively) unique OTUs were obtained by DNBSEQ-G400 platform, with 3 of them observed once and the other 14 OTUs consistently observed by all three technical replicates. In comparison, a total of 162 (92, 156, 66 for each replicate respectively) unique OTUs were obtained by Illumina NovaSeq 6000 platform, with 67 OTUs observed once, 95 OTUs twice and 57 OTUs observed by all three technical replicates. Significant batch effect was observed for NovaSeq platform with only 35% of the OTUs being consistently observed by all three replicates compared to 82% of DNBSEQ (**Figure 1A**). Comparison of the two customized mock communities revealed similar observations, with eight of eleven OTUs by NDBSEQ while 39 of 85 OTUs by NovaSeq platform for the 4Bac community; and nine of eleven by DNBSEQ while 42 of 83 OTUs by NovaSeq platform for the 7Bac community, consistently detected by all three technical replicates (**Figure S2, Table S2, Table S3**).

Taxonomic annotation of the OTUs revealed that all members of the mock community were consistently detected by all the technical replicates on both sequencing platforms, indicating successful detection of abundant microbial members in a community given enough sequencing depth. Despite successful detection of the expected members by both sequencing platforms, both platforms reported certain amount of unexpected OTUs, representing potential false positives. The number of unexpected OTUs for NovaSeq platform was almost two orders of magnitude higher than that of DNBSEQ platform. Relative abundances of the unexpected OTUs were up to 1.19% and 0.09%, accounting for a total of 5.68% and 0.08% reads for NovaSeq and DNBSEQ platform, respectively, for the commercial mock community (**Figure 1B, Table S1**). Similar trend was observed for both customized mock communities (**Table S2, Table S3**).

False positives might be introduced by index misassignment and could be not removed by routine QC process

We then carefully assessed whether the unexpected OTUs came from contaminations introduced during sample preparation or during sequencing. As aliquots from the same DNA sample was used for library construction and subsequent sequencing on different platforms, potential false positives detected by both platforms were more likely from the sample preparation process before library construction and sequencing, while potential false positives detected either of the platforms were more likely contaminations from the library construction or sequencing process of that platform. Comparison of unexpected OTUs detected by each of the platforms showed that all OTUs observed by DNBSEQ platform were consistently observed by NovaSeq platform, indicating that these unexpected OTUs were likely from the original sample, instead of from the sequencing process of DNBSEQ (**Figure 1B**). Sequence alignment of the nine unexpected OTUs detected by both platforms indicated that five of them might be different strains from the same species to their mock members with 97.18%~99.60% sequence identity (Otu1903, Otu0048, Otu0106, Otu0112 and Otu0060). The other four OTUs (Otu0010, Otu0062, Otu1666 and Otu1305) also had 91.70%~96.39% sequence identity to their mock members, but should not come from the mock members (**Figure 1C, Table S1**). Those OTUs from the same species of the mock members might be SNVs of the original mock bacteria, while there was a chance for the OTUs with relatively low sequence identity to mock members to be potential contaminants from the original sample or during aliquoting.

However, taxonomic annotation of the unexpected OTUs specifically detected by NovaSeq platform revealed a highly diverse spectrum of phylogeny, indicating a potential diverse contamination source during experimental process, including PCR amplification, library construction and sequencing. Moreover, the small fraction of shared OTUs by technical replicates also implied that the false positives were introduced likely in a random way. Amplicon sequencing of two customized mock communities consistently revealed more unexpected OTUs with diverse phylogeny detected by NovaSeq 6000 platform (Figure S2, Table S2, Table S3). Comparison of distinct samples sequenced in the same batch revealed considerable fraction of shared OTUs (37, 47, 37 OTUs were shared by five totally distinct samples sequenced in batch 1, batch 2 and batch 3, respectively) among totally different samples on NovaSeq platform (**Figure S3**).

In order to evaluate whether those potential contaminant reads could be removed *in silico* during data analysis, more stringent quality control process was used to filter the data before downstream clustering and statistical analysis. As index misassignments were supposed to happen at relatively low rate [25,27–29], higher threshold of minimum tags to be contained in an OTU was required to try to remove potential low-rate false positives. However, a raised threshold of even 50 could not remove all the potential contaminants (**Table S1, Table S2, Table S3**).

The rare taxa sub-community were more vulnerable to biases

Amplicon sequencing of mock communities with known community compositions confirmed the successful detection of abundant taxa by both sequencing platforms and at the same time frequently observed false positives likely introduced by index misassignments, especially for the NovaSeq 6000 platform. As real microbial communities occupying various ecosystems are much more complex than mock communities regarding both microbial composition and abundance distribution, we speculated that index misassignments might lead to more intriguing false positive and/or negative detections. In order to evaluate how index misassignments might affect community profiling with different microbial diversity, biological triplicate samples from several typical microbial ecosystems, including mice gut, surface seawater and mangrove sediment representing relatively low, moderate, and high diversity communities, were sequenced with technical triplicate on both DNBSEQ-G400 and NovaSeq 6000 platforms at comparable sequencing depth of around 60,000 reads. The number and fraction of clean reads after QC and number of obtained OTUs for

each sample were summarized in **Table S4**.

For these real samples from different microbial ecosystems, we speculated that: 1) there should be a difference in successful detection of abundant taxa with high relative abundance (RA) ($RA \geq 1\%$) versus moderate taxa (with $1\% > RA \geq 0.1\%$) and rare taxa (with $RA < 0.1\%$); 2) false positives introduced by random index misassignments would lead to significant batch effect of the rare taxa community; 3) false positives introduced by index misassignment would lead to inflated alpha diversity. Comparison of technical triplicates for each ecosystem revealed an average of 100% of abundant, 97.53% moderate and 68.93% rare taxa being consistently detected by more than one technical replicate by DNBSEQ-G400, while these fractions were 100%, 87.94% and 39.50% for NovaSeq platform (**Figure 2A, Figure S4A**). Comparison between sequencing platforms revealed higher fractions of sequencing platform specific taxa of the moderate and rare sub-communities, especially for NovaSeq platform. Although significant higher alpha diversity estimations were observed for seawater and mice gut sample for NovaSeq 6000 platform, significant lower alpha diversity was reported by NovaSeq 6000 platform for the mangrove samples, compared to the results of DNBSEQ-G400 platform (**Figure 2B**).

In order to further evaluate whether those platform specific OTUs were more likely false positives or “false negatives” missed by the other platform, we mapped the shotgun reads of the same samples to the OTU representative sequences of the mangrove (**Figure S4B**) [37] and mice gut (**Figure S4C**) (not published data) ecosystems. Much more OTUs not finding a hit in the shotgun dataset were found for NovaSeq platform, informing a higher fraction of potential false positives of NovaSeq platform. Higher reproducibility and less potential false positives could also be revealed by the clustering of samples based on beta diversity. Both weighted and unweighted UniFrac trees consistently showed that technical triplicates of the DNBSEQ-G400 results were grouped by their biological samples for all the tested ecosystems, while technical triplicates of more samples were grouped by sequencing batch or in a random way for NovaSeq sequencing results (**Figure S5**).

The rumen microbial community revealed by different sequencing platforms

Cow rumen ecosystem not only harbors a diverse microbes capable of digesting insoluble lignocellulosic biomass into accessible carbon and energy sources for their host, but also have complex connections with many of the host attributes and performances [38,39]. The rumen

microbiome are also valuable sources of genes coding cellulase and hemi-cellulases for industrial uses [40]. A lot of researches have been done trying to elucidate the microbial compositional and functional diversity, but most of the previous studies ignored the rare taxa of the microbial populations [35,36]. In order to study the ecological properties of rumen rare taxa, and evaluate the potential differences of the rare community revealed by different sequencing platforms, we sequenced a total of 47 cow rumen fluid samples using amplicon sequencing technology at both DNBSEQ-G400 and NovaSeq 6000 sequencing platforms with identical sequencing depth.

For both sequencing platforms, an average of around 60,000 high quality tags were generated and reads from different sequencing platforms were pooled before they were subjected to the QC and clustering processes. A total of 3043 OTUs were obtained and then split into three categories, namely abundant, moderate, and rare taxa, according to their averaged relative abundance being greater than 1% or between 1% and 0.1% or less than 0.1% across two sequencing platforms. As expected, among all the OTUs, only twelve were identified as abundant taxa, while 161 and 2870 taxa were identified as moderate and rare taxa respectively, consistent with previous studies demonstrating that rumen microbial communities were dominant by several abundant taxa and large amount of taxa with low relative abundances [41]. Identical abundant and moderate microbial taxa were revealed by DNBSEQ and NovaSeq, indicating relatively low sequencing platform effects regarding the abundant and moderate microbial sub-community membership. However, significant sequencing platform specific taxa were observed for the rare microbial population, especially for the NovaSeq sequencing platform. Of the 2870 rare OTUs, 1957 were consistently detected by both platforms, accounting for 68% of all the rare taxa detected; although only 24 rare OTUs were specifically detected by DNBSEQ, a much larger amount of 889 rare OTUs were specifically detected by NovaSeq, showing a much more diverse metacommunity reported by NovaSeq platform (**Figure 3A**).

The higher diversity of the metacommunity could be attributed to either higher alpha diversity in each sample or higher beta diversity among samples detected by NovaSeq platform. However, comparison of the alpha diversity revealed by different sequencing platform indicated significantly lower Chao I index for NovaSeq platform. But higher phylogenetic diversity was observed for NovaSeq dataset (although not significant), indicating that the rare taxa may spanned a wider range of the phylogenetic tree than DNBSEQ dataset although their Chao I index was significantly lower

(**Figure 3B**). Frequency analysis of the NovaSeq specific rare OTUs showed that more than 30% of them were detected only once across all the samples, consistent with the observed higher beta diversity (**Figure 4B, Figure S6**) and higher diversity of the metacommunity revealed by NovaSeq (**Figure 3A**). Taxonomic annotation of the OTUs revealed a much more diverse of phylogenetic distribution of the NovaSeq OTUs, including six phyla (Armatimonadetes, BRC1, Chloroflexi, *Deinococcus thermus*, Genmatinonadetes, candidate division WPS-2) not being detected by DNBSEQ platform, all of which were not commonly reported microbes of cow rumen system, and some of them were even aerobes (**Figure 3C**).

In order to further evaluate whether these NovaSeq specific rare taxa were *bona fide* rare taxa or false positive introduced during sequencing, we assessed the potential correlations between each rare taxa and a set of physiochemical properties of the rumen fluid. The hypothesis was that *bona fide* rare taxa in cow rumen should be correlated with the fermentation condition of their host with a higher probability than randomly introduced false positives. We tested the correlation between rare taxa detected by each sequencing platform respectively and a set of physiochemical parameters, including the relative concentration of NH_4^+ , acetate, propionate, butyrate and iso-butyrate. Consistently higher fractions of rare taxa reported by DNBSEQ platform were found to be significantly correlated with each of physiochemical parameters compared to rare taxa detected by NovaSeq platform (**Table S5**), further suggesting that the NovaSeq platform might detect higher fraction of false positive rare taxa than DNBSEQ.

Index misassignment could lead to biased community assembly mechanisms

The assembly of microbial communities in various ecosystems was simultaneously controlled by both stochastic and deterministic processes with each of them governing a differential fractions of the microbial community compositions in different ecosystems [1,42,43]. Understanding the mechanisms of microbial community assembly process is vital for active microbiome intervention for host health management [44]. However, how the large amount of potential false positives and missed *bona fide* rare taxa would influence the interpretation and understanding of the mechanism behind community assembly remained elusive. In order to answer this question, both Sloan's neutral community model [45] and the null assembly model [46,47] were applied to the cow rumen microbiome data sets generated on both sequencing platforms to test: 1) whether stochastic or

deterministic process dominated the assembly process of the cow rumen microbiome; 2) whether similar or distinct assembly mechanisms would be revealed by different sequencing platforms.

Neutral model gave relatively low coefficients of the neutral fit ($R^2 = 0.321$ for DNBSEQ; $R^2 = 0.360$ for NovaSeq) for microbial communities, and the coefficient revealed by both sequencing platforms was comparable (**Figure 4A**). Less than 55% of the OTUs distributed within the neutral prediction, indicating that the neutral process could only explain limited part of the microbial community assembly process in cow rumen. OTUs distributed above or below the neutral prediction representing microbes that were actively selected for or against by cow rumen environment, and these fractions were also comparable across sequencing platforms (**Figure 4A**). However, the estimated migration rate, m , indicating the probability that a random loss of an individual in a local community would be replaced by dispersal from the metacommunity, as opposite to reproduction within the local community, was larger on DNBSEQ-G400 than NovaSeq 6000 platform ($m = 0.221$ for DNBSEQ; $m = 0.079$ for NovaSeq).

In order to further discriminate between the deterministic and stochastic processes in cow rumen microbial community assembly, we calculated the β -nearest taxon index (β -NTI) for the microbial communities given by each sequencing platform. β -NTI quantifies the difference between the observed phylogenetic turnover between communities and that of the null communities. The fractions of community assembly process explained by stochastic process ($|\beta\text{-NTI}| < 2$), variable selection ($\beta\text{-NTI} \geq 2$) and homogeneous selection ($\beta\text{-NTI} \leq -2$) were calculate for each sequencing platform respectively (**Figure 4C**). Consistent with the results of Sloan's neutral model fitting, the distribution of β -NTI for both sequencing platforms indicated that the cow rumen microbial community assembly was simultaneously controlled by both stochastic and deterministic forces. Compared to the results of DNBSEQ platform, NovaSeq platform showed a wider range of distribution of the β -NTI values, and a small fraction of homogeneous selection process was observed by NovaSeq platform (**Figure 4C**). Removal of NovaSeq specific OTUs from the NovaSeq dataset returned similar β -NTI value distribution pattern as that of DNBSEQ platform (**Figure 4C**).

Differential keystone species were identified by DNBSEQ and NovaSeq sequencing platforms

Researchers often infer interactions among microbes based on the correlation coefficients of

their relative abundances distributed in a set of samples. The architectural or topological features of networks could provide us with invaluable insights into complex polymicrobial interactions and co-occurrence patterns, and could be used to identify microbes playing the most influential roles in the community, such as keystone species. Thus, in addition to inflated or underestimated alpha diversity and distorted community assembly mechanism, we assessed whether potential false positives may lead to misleading or even wrong interpretations of microbial interactions. Microbial interaction network for rumen microbial communities revealed by each sequencing platform was constructed based on the cooccurrence correlations (**Figure 5A**). Rare taxa contributed more than 90% of the nodes for each network, demonstrating potential important ecological roles of the rare taxa in cow rumen. The degree of nodes of both networks followed a power-law distribution, showing the property of scale-free networks. However, the microbial network based on NovaSeq platform was less integrated with significant lower node degree and stability under node attack compared with the network based on DNBSEQ platform (**Figure 5B**). Although microbes from Lachnospiraceae, Clostridiales, Bacteroidales and *Prevotella* were identified as keystone species by both sequencing platforms, differential OTUs from each of these taxa were identified (**Figure 5C**), which might be caused by minor sequencing biases or errors of different sequencing platforms leading to the formation of different OTUs with minor differences during clustering process using 100% sequencing identity algorithm. This minor within genus or even species difference would not lead to wrong interpretation of the ecological roles of these microbes. However, two of the keystone species with high node degree identified by DNBSEQ, including *Pseudobutyrvibrio xylanivorans* (Lachnospiraceae) and *Succiniclasticum ruminis* (Acidaminococcaceae), both of which were reported to play important ecological roles in rumen system [48,49], did not occupy a hub position in the interaction network based on NovaSeq sequencing results. On the contrary, several of keystone species identified by NovaSeq platform were with low degree (such as Porphyromonadaceae and Enterobacteriaceae) or not detected (*Nocardia coeliaca*, Otu0244) by DNBSEQ platform. While there were previous publications reporting the observation of Porphyromonadaceae, Enterobacteriaceae in cow rumen ecosystem [50–52], *Nocardia coeliaca* was reported as an aerobic, gram-positive bacterial and thus not a frequently observed microbe in cow rumen [53]. Considering the biological nature of *Nocardia coeliaca* and its specific detection by NovaSeq platform, we speculated it might be a false positive introduced by index misassignment.

Specific PCR primer pairs were designed according to its representative sequence (Otu0244_F1: AGGCGGTTTGTGCGGTCGTT, Otu0244_R1: TCGCTACCCACGCTTTCGTTCC; and Otu0244_R2: ACGCTTTCGTTCTCAGCGTCA) and used for PCR amplification of *Nocardia coeliaca* from original gDNA of the cow rumen samples. Although positive PCR amplification results were observed, Sanger sequencing of the positive clones returned low sequencing identity (~95%) to the target *Nocardia coeliaca* representative sequence, or mapped to other OTUs (Table S6). Taking the short fragment of target PCR product and the low sequence identities, we speculated that the positive PCR products might be from other microbes due to unspecific primer binding, but not *Nocardia coeliaca*.

Discussion

The past decade has seen a rising interest and understanding of the microbial rare biosphere, and both the functional and ecological roles of rare taxa in various ecosystems are being recognized [54–56]. However, due to the inherent scarcity of microbial rare taxa, and the fact that index misassignment caused sample-wise cross contamination could not be easily removed *in silicon* by bioinformatic approaches, the study of microbial rare taxa remains intriguing. In this study, we carefully evaluated the rate of potential index misassignment of two main stream sequencing platforms based on different sequencing technologies, and found significant higher fractions of unexpected reads for the NovaSeq 6000 platform in mock community test. Although the unexpected reads might be introduced during library construction from neighboring wells as described previously [35], neighboring well might not be enough to explain the observed high phylogenetic diversity and the large fraction of the unexpected taxa. Furthermore, the fraction of unexpected reads in the mock community test was consistent with previous report demonstrating up to 6% index misassignment on Illumina sequencing platforms [25,30]. Taking together, it might be reasonable to infer index misassignments might be the major cause of those observed false positives.

The presence of index misassignments introduced false positives harms the study of microbial rare biosphere and it should be of paramount importance to eliminate them before downstream biological and ecological interpretation. Tools and algorithms for potential contamination removal have been developed previously, such as Decontam [57] and PERFect [58] regarding amplicon sequencing data. While Decontam removes likely abundant contaminant taxa based on statistical

test but does not address the false positive issue of rare taxa, PERFect just removes rare taxa and demonstrates that rare taxa removal would not influence the overall statistical results of microbial communities [59], both of which were not suitable to remove index misassignments caused false positives, especially when rare taxa were focused. Moreover, arbitrarily removing reads with small number of copies in a dataset could harm the study of *bona fide* rare taxa, although the abundant taxa were demonstrated not likely to be influenced [59]. As index misassignments happens in a random way, we assume proper technical replication setting could partly alleviate the number of false positive OTUs, but caution still should be taken when focusing on the rare biosphere as some *bona fide* rare taxa might be missed and consistent detection in triplicates still did not guarantee the observation of *bona fide* rare taxa when NovaSeq sequencing platform was used and all technical replications were sequenced in the same run (**Figure 1, Figure S2, Figure S3**).

Amplicon sequencing test of microbial communities from three real ecosystems with differential complexity confirmed significant lower shared fraction of rare sub-community among technical replicates compared to moderate and abundant taxa for both sequencing platforms, clearly demonstrating the difficulty in confident rare taxa detection. Compared to DNBSEQ platform, NovaSeq platform reported even less shared fraction of rare taxa among technical replications and thus, gave significant lower run-to-run consistency. There are also previous works from other peer colleagues demonstrating significant but rarely considered run-to-run variations in microbial community studies using amplicon sequencing technology [54,60,61]. There would be a chance that the significant run-to-run variation, especially for NovaSeq 6000 detected rare taxa, were caused by index misassignment introduced random false positives. Lower mapping rate of the shotgun sequencing reads to NovaSeq OTUs for the mangrove and mice gut samples also support the possibilities of potential contaminations in the NovaSeq amplicon sequencing data set. Special caution should be taken when sequencing low biomass samples, as the low biomass samples were more vulnerable to index misassignment when pooled and sequenced with high biomass samples due to imbalanced index usage [36].

Index misassignment caused false positives could also explain the inflated alpha diversity estimation for relatively simple ecosystem but lower estimated alpha diversity for more complex samples of the NovaSeq platform, compared to that of the DNBSEQ platform. Because given certain sequencing depth, it is easier to recover all the microbial taxa in relatively simple communities, and

the successful detection of rare taxa would less likely be affected by index misassignment introduced false positives, leading to higher estimation of alpha diversity. However, for microbial community in complex ecosystems, such as mangrove sediment, more *bona fide* rare taxa would be contained, the successful detection of which could be more easily diluted out by highly frequent randomly introduced contaminations, leading to lower observed OTUs compared to DNBSEQ platform with scarcely happened index misassignments. Technical replications in this case could hardly improve the estimation of true alpha diversity as *bona fide* rare taxa's missing, like false positive's introduction, happens in a random way (**Figure 2B**). Furthermore, consistent observation of high rate of false positives potentially introduced by index misassignment using different samples in different sequencing batches suggested that this might be a systematic issue for the NovaSeq platform, instead of occasional case.

In addition to inflated or underestimated alpha diversity and biased microbial compositions, index misassignment introduced false positive rare taxa could affect the interpretation of microbial community assembly mechanism and identification of keystone species. Using a set of real samples from one of the classical anaerobic digestion systems, cow rumen, we evaluated whether similar or differential microbial community composition, assembly mechanism and keystone species would be revealed by different sequencing platforms. In congruent to what have been observed with the mangrove samples, lower Chao I diversity was observed by NovaSeq platform compared to DNBSEQ platform. Singleton and low frequent OTUs were observed in NovaSeq data set, leading to higher diversity of the metacommunity. Since all the samples were rumen fluid samples from cows with similar genetic background and under the same controlled husbandry regimes, diets and rearing conditions, considering the physiological phenomenon of rumination, there should not be large fraction of sample specific taxa [62], and some of the NovaSeq platform specifically detected taxa were not likely common resident of cow rumen systems. Taken together, the very large fraction of diverse singleton OTUs by NovaSeq platform might not be *bona fide* microbial taxa living in cow rumen ecosystem (**Figure 4B**), but more likely randomly detected taxa from other sources, in accord with the nature of index misassignments caused false positives [25,27–29]. The overall lower fraction of microbial taxa correlated with various key rumen physiochemical properties also suggested less overall confidence of the NovaSeq detected rare taxa.

Regarding community assembly mechanism, the overall low fit of neutral model by both

platforms was easy to understand as the cow rumen environment should exert certain selective pressure on its microbiome, which was in accord with previous study demonstrating that age and diet played an overall deterministic force over the entire microbial community after a stochastic microbial colonization at birth [62]. However, the migration rate m revealed by each sequencing platform were quite different. A migration rate $m = 1$ indicated an entirely open and highly coupled local and metacommunity, while a migration rate of 0 indicated an entirely isolated local community. With the drop of migration rate, the internal neutral dynamics increasingly act to dominant the dynamics of the local community until totally control and make the local community isolated [45,63]. Thus the extremely low migrate rate revealed by NovaSeq platform might not fit the case in our study as all the samples were from cows raised under the same controlled husbandry regimes, diets and conditions as discussed above, and there should be substantial exchange of the rumen microbiome considering the co-housing and the natural rumination process [62]. The low migration rate might be explained by the large fraction of singleton and very low frequent OTUs specifically detected by NovaSeq platform (**Figure 4B**), because the loss of singleton or low frequent OTUs could not or hardly be filled by an “immigrant” from the metacommunity, leading the local community more “isolated”. This was consistent with the high index misassignment of NovaSeq platform, as discussed above, causing both detection of false positives and miss detection of *bona fide* rare taxa in some of the samples, both of which could lead to higher fractions of singleton OTUs or OTUs with very low frequency (**Figure 4B**).

The assembly mechanism based on the overall distribution of β -NTI values was in accordance with the prediction using Sloan’s neutral model, indicating the microbial community was simultaneously controlled by both stochastic and deterministic forces. However, like differential migration rate was observed, slight but significant different mechanisms were observed between sequencing platforms. The NovaSeq platform specifically observed homogeneous selection process ($\beta\text{-NTI} \leq -2$) might be attributed to the homogeneously introduced potential false positives by index misassignments, as revealed by the NovaSeq specific OTUs with high frequency. The NovaSeq platform observed very large β -NTI values, representing variable deterministic selection process, might be caused by randomly introduced differential false positives from other non-relevant samples processed on the same sequencing lane, as revealed by the higher phylogenetic diversity. Although

very large β -NTI values was reported by NovaSeq dataset, NovaSeq revealed less overall fraction variable selection process compared to DNBSEQ, which might be because a fraction of differential *bona fide* rare taxa was missed by NovaSeq platform as revealed by the lower estimated alpha diversity compared to DNBSEQ.

Various microbes in a community interact with each other to communicate, cross-feed, recombine, and coevolve, in a way via which microbes form a complex interaction network and sustain their stability and robustness [64]. The cow rumen microbial interaction networks based on both sequencing platforms revealed that rare taxa occupied most of the nodes and some of them were even identified as keystone taxa, consistent with previous work demonstrating keystone species in a community were not necessarily to be dominant [9]. However, despite similar keystone species, differential taxa were identified by each sequencing platforms, which should be interpreted with cautions. For example, the NovaSeq identified keystone species, *Nocardia coeliaca*, was documented as an aerobic soil bacteria [53] and confirmed negative from the rumen genomic DNA sample using specifically designed PCR primer pairs. Keystone species could be the most influential microbes in a network interacting with most other microbes and essential for the stability and robustness of the microbial community [64]. Wrong identification of keystone species thus could lead to very miss-leading interpretation of the potential ecological roles of certain microbes and even wrong understanding entire microbial community.

Conclusions

In amplicon studies, although index misassignment would not have significant influence to the relative abundant taxa, it could lead to distorted features regarding the rare sub-community, including their composition, diversity, interaction network, assembly mechanisms and other properties to be found. Potential contaminants could also be introduced from any of the experimental processes, including extraction, PCR amplification, library construction and other processes. Properly set positive and negative samples, together with proper bioinformatic algorithms during data processing, could be used to eliminate the potential contaminants during normal experimental process before sequencing. Furthermore, in addition to careful normal quality control process, we suggest to set technical replicates and choose proper sequencing platforms with low potential index misassignment rate and sequence to enough sequencing depth to improve the accuracy of rare taxa

detection and downstream biological and ecological mechanisms interpretation. We also recommend researchers to cross validate the metabarcoding amplicon sequencing results using differential sequencing technologies when focusing on rare sub-community study.

Materials and Methods

Mock communities, typical sample collection and DNA extraction

The commercial mock microbial community, ZymoBIOMICS Microbial Community DNA Standard D6305, containing 8 bacteria strains was purchased from ZymoBIOMICS™ (Table S1). Two customized mock communities, one of which containing 4 bacterial strains (genomic DNA of *Bacillus halotolerans*, *Bosea robiniae*, *Streptomyces toxytricini* and *Nocardiopsis dassonvillei* mixed in a ratio of 2:1:1:1), and the other containing 7 bacterial strains (genomic DNA of *Photobacterium halotolerans*, *Vibrio parahaemolyticus*, *Vibrio natriegens*, *Bacillus aquimaris*, *Bacillus anthracis*, *Bacillus aryabhatai* and *Bacillus hwajinpoensis* mixed in a ratios of 10:10:10:10:1:1:1) were constructed. All the bacteria strains used in the customized mock communities were obtained from China National GeneBank (Qingdao), BGI-Qingdao, China.

Three mice fecal samples were selected from the deposited samples at China National GeneBank (Qingdao, China) in August, 2019, and total DNA was extracted using QIAamp DNA Stool Mini Kit (Qiagen, Hilden, Germany). Three surface seawater filter samples were collected from Wentai Fishery (Zhejiang, China) by filtering 2L of surface seawater in April, 2018, and the filter samples were used for total DNA extraction using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) [65]. Three mangrove rhizosphere topsoil samples were collected from East Harbour National Nature Reserve (Hainan, China) in April, 2018, and 0.5g of each soil sample was used to extract the total DNA using the PowerSoil DNA isolation kit (Mobio Labs, Inc., Solana Beach, CA, USA).

Cow rumen sample collection, DNA extraction and physiochemical parameters measurement

The lactating Holstein dairy cows with similar age and raised under the same controlled husbandry regimes, diets, and rearing conditions were selected at a commercial dairy farm (Yangling, Shanxi, China). Rumen fluid were collected from cows via esophageal tubing before morning feeding. Firstly, several hundreds of milliliters of rumen fluid were discarded to minimize saliva

contamination. Then the rumen fluid samples were filtered through four layers of cheesecloth, and stored at -80 °C before DNA extraction. Rumen fluid was centrifuged at 12,000×g for 10 min at 4 °C for supernatant collection. Total DNA was extracted from the centrifuged pellet using a method involving cetyltrimethylammonium bromide (CTAB) plus bead beating (Minas et al., 2011). Ruminal supernatant was used for volatile fatty acids analyzation by gas chromatography (Agilent 7890A, Wilmington, USA). The NH₃-N concentration in supernatant was determined using a Berthelot ammonia assay kit (Jiancheng, Nanjing, China).

PCR Amplification, library construction and sequencing

The universal primer pair for 16S rRNA gene V4 region 515F/806R (515F: GTGCCAGCMGCCGCGGTAA, 806R: GGACTACHVGGGTWTCTAAT) [66] were used for PCR amplification. A two-step PCR procedure was used to construct the amplicon libraries to be sequenced at the DNBSEQ-G400 sequencing platform as previously described [67]. Basically, for all samples and negative controls, the first-step PCR with zero to three random nucleotides inserted before each of the primer pairs to balance nucleotide proportion at each position for accurate base-calling was performed as follows: 95°C for 10 min, followed by 20 cycles at 98°C for 20 s, 58°C for 30 s, and 72°C for 30 s with a final extension at 72°C for 10 min. No band of target PCR product were observed for the negative controls. After the amplification, the primer with sample barcode and the DNBSEQ sequencer adapter was used for the second PCR amplification: 95°C for 5 min, followed by 15 cycles at 98°C for 20 s, 58°C for 30 s, and 72°C for 30 s with a final extension at 72°C for 10 min. After the two-step PCR amplification, the PCR products were verified using 1.5% agarose gel electrophoresis. The PCR products with target bands were mixed in equal mass, and 2% agarose gel was used for electrophoresis and gel cutting for purification, and then make DNA nanoballs (DNB) following the standard protocol of the DNBSEQ sequencing platform. All libraries were sequenced on DNBSEQ-G400 platform in the paired-end mode with 200 bp length reads at BGI-Qingdao (Qingdao, China).

For the Illumina amplicon sequencing library construction, about 10 ng DNA for each sample was used for the PCR amplification using the 515F/806R 16S rRNA gene primer pair. The PCR procedure was as follows: 98°C for 1 min, followed by 30 cycles of denaturation at 98°C for 10 s, annealing at 50°C for 30 s, and elongation at 72°C for 30 s with final extension at 72°C for 5 min.

PCR products with target bands were mixed in equal mass, and then the mixed PCR products were purified with GeneJET Gel Extraction Kit (Thermo Scientific). Sequencing libraries were constructed using Illumina TruSeq DNA PCR-Free Library Preparation Kit (Illumina, USA) following the manufacturer's protocol with index sequence and then libraries quality was assessed by Qubit 2.0 Fluorometer (Thermo Scientific). For all technical triplicate of mocks and typical samples, the qualified libraries were sequenced on an Illumina NovaSeq 6000 platform and generated 250 bp paired-end reads at Novogene Co., LTD (Beijing, China). Cow rumen samples were sequenced using the same strategies at Personalbio Co., LTD (Shanghai, China) with no technical replication as people do not set technical replications when sequencing large number of samples.

Quality control of reads and the bioinformatical process

An average of 50000~60000 reads were generated for each of the samples in the present study. Reads with adapter contaminations and low-quality reads (more than 20% base quality < Q20) in the raw data set were filtered out by SOAPnuke (v1.5.6) [68]. Paired-end high quality clean reads were merged into tags by FLASH (v1.2.11) [69] with parameters "--min-overlap 10 --max-mismatch-density 0.1". Reads generated from different sequencing platforms were combined and the denoising clustering algorithm unoise3 [22] was used to generate denoised OTUs by USEARCH (v10.0.240) [70] with default parameter of "-minsize 8", and generated the OTU abundance profiles [71]. OTU taxonomic assignment was carried out using syntax algorithm [72] against RDP training set (v18) with 0.8 confidence cutoff value. The phylogenetic tree of cow rumen OTUs was constructed by FastTree (v2.1.5) [73] and visualized by iTOL [74]. The alpha-diversity index including Shannon indices and Chao I indices, weighted and unweighted UniFrac beta-diversity distances were analyzed using QIIME (v1.9.1). The phylogenetic diversity index was calculated using R package "picante". All boxplots were visualized by R package "ggplot2". The venn diagrams were plotted by R package "venn" and software TBtools (v1.09856). To determine the potential false positive taxa, more than 150 Gb metagenomics sequencing data [37] of the three mangrove rhizosphere soil samples, and 15 Gb metagenomics sequencing data of mice gut samples (data not published) were mapped to the mangrove or mice gut represent OTUs, respectively, using Salmon [75] (v0.9.1). The OTUs with more than one metagenomic read mapped in average were

masked as OTUs consistently detected in shotgun sequencing.

PCR verification of *Nocardia coeliaca* in rumen fluid

Specific primer pairs for amplification of *Nocardia coeliaca* were designed based on the representative sequence of Otu0244 using primer premier (V6.0). One forward and two reverse primers were designed with their sequences as follows: Otu0244_F1: AGGCGGTTTGTGCGTCGTT, Otu0244_R1: TCGCTACCCACGCTTTCGTTCC; and Otu0244_R2: ACGCTTTCGTTCTCAGCGTCA. The genome DNA from three different rumen fluid samples were mixed by equal mass and used for PCR amplification to verify the presence or absence of *Nocardia coeliaca* in the cow rumen ecosystem. The representative sequence of Otu0244 was directly synthesized in Sangon Biotech (Shanghai, China), and was used as positive control for PCR amplification to test the efficiency of the designed primer pairs. PCR products were linked to pESI-T vector and transformed into DH5 α competent *E. coli* cells. Positive clones were then picked and sequenced using ABI 3730XL. Sequences of the positive clones were aligned to the entire OTU representative sequences set using blast to search for potential positive hits.

Statistical analysis

All statistical analysis was conducted in R environment (v3.4.1). The significance test of differential alpha diversity or phylogenetic diversity was assessed by the Wilcoxon-test, while difference between weighted and unweighted UniFrac distances was tested with PERMANOVA using “vegan” Pacakge. Potential correlation between cow rumen OTU relative abundances and the physiochemical properties of the rumen fluid was measured by “spearman” function with a significance cut-off of P-value < 0.05.

To compare the assembly mechanisms of cow rumen microbiomes revealed by DNBSEQ or NovaSeq sequencing platforms, both Sloan’s neutral model and the null model hypothesis were tested. The Sloan neutral community model prediction and statistics were performed by “MicEco” package in R [76], and the overall fitness of the model (R^2 value) and the proxy of dispersal limitation (estimated migration rate, m value) were calculated at the same time [45]. The beta Nearest Taxon Index (β NTI) values, an index representing the null assembly hypothesis, were calculated by “picante” package [77] in R. The β NTI value was used to infer the major driving

forces controlling the community assembly process. A $\beta\text{NTI} \geq 2$ is considered of variable selection of deterministic processes and $\beta\text{NTI} \leq 2$ is considered of variable selection of deterministic processes, while $|\beta\text{NTI}| < 2$ considered of stochastic processes. And to assess how NovaSeq specific OTUs, representing potential false positives, could influence the inference of community assembly process compared with DNBSEQ results, the NovaSeq specific OTUs were removed and the NovaSeq-without-specific OTU βNTI values were also calculated.

The co-occurrence networks of the microbiomes revealed by DNBSEQ and NovaSeq were preformed respectively using SparCC algorithm [78], and only the robust correlations with $|r| > 0.75$ and $P < 0.05$ were considered. The networks were visualized, and the node degrees were calculated by Gephi (0.9.2) [79]. The natural connectivity of networks was estimated by “attacking” nodes to confirm the robustness of the correlation networks [80].

Declarations

Ethics approval and consent to participate

The cow rumen experiment procedures were approved by the animal care committee of Institute of Animal Sciences of Chinese Academy of Agricultural Sciences (approval number: ISA2020-82). The Institutional Review Board of BGI (NO. BGI-R052-3-T1) provided approval for the mice gut samples.

Consent for publication

Not applicable.

Availability of data and material

The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA, <https://db.cngb.org/cnsa/>) of China National GeneBank DataBase (CNGBdb) with accession number CNP0002138 for the cow rumen samples and accession number CNP0002180 for the mock and typical samples.

Competing interests

YJ, WG, LP, FZ, GF, YZ, DX, GL, RW, XF, HZ, KK, WZ, JC are employees of BGI.

Funding

This work was partly supported by the National Natural Science Foundation of China (grand number 32100047), the Agricultural Science and Technology Innovation Program (grand number ASTIP-IAS12) and the State Key Laboratory of Animal Nutrition (grand number 2004DA125184G2108).

Authors Contributions

YJ, JC conceived the study, analyzed the data, and wrote the manuscript. SZ conceived and coordinated the cow rumen part of the work and revised the manuscript. WG, YZ, and RW contributed to the amplicon library construction and sequencing experiments. LP, FZ, DX, and GL contributed to the data analysis. LW, GF, XF, HZ, KK, and WZ read and revised the manuscript. The authors read and approved the final manuscript.

Acknowledgements

The authors thank China National GeneBank and GeneBank (Qingdao) for sequencing and experiment coordination.

Reference

1. Jia X, Dini-Andreote F, Falcão Salles J. Community Assembly Processes of the Microbial Rare Biosphere. *Trends Microbiol.* Elsevier Ltd; 2018. p. 738–47.
2. Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, et al. Global patterns in the biogeography of bacterial taxa. *Environ Microbiol.* NIH Public Access; 2011;13:135–44.
3. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proc Natl Acad Sci U S A.* National Academy of Sciences; 2006;103:12115–20.
4. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may be more: How the rare biosphere pulls ecosystems strings. *ISME J.* 2017;11:853–62.
5. Chen QL, Ding J, Zhu D, Hu HW, Delgado-Baquerizo M, Ma YB, et al. Rare microbial taxa as the

major drivers of ecosystem multifunctionality in long-term fertilized soils. *Soil Biol Biochem.* Elsevier Ltd; 2020;141.

6. Pester M, Bittner N, Deevong P, Wagner M, Loy A. A “rare biosphere” microorganism contributes to sulfate reduction in a peatland. *ISME J.* Nature Publishing Group; 2010;4:1–12.

7. Hausmann B, Pelikan C, Rattei T, Loy A, Pester M. Long-term transcriptional activity at zero growth of a cosmopolitan rare biosphere member. *MBio.* American Society for Microbiology; 2019;10:1–16.

8. Bodelier PLE, Meima-Franke M, Hordijk CA, Steenbergh AK, Hefting MM, Bodrossy L, et al. Microbial minorities modulate methane consumption through niche partitioning. *ISME J.* Nature Publishing Group; 2013;7:2214–28.

9. Xue Y, Chen H, Yang JR, Liu M, Huang B, Yang J. Distinct patterns and processes of abundant and rare eukaryotic plankton communities following a reservoir cyanobacterial bloom. *ISME J.* Nature Publishing Group; 2018;12:2263–77.

10. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci U S A.* National Academy of Sciences; 2011;108:12776–81.

11. Jia Y, Leung MHY, Tong X, Wilkins D, Lee PKH. Rare Taxa Exhibit Disproportionate Cell-Level Metabolic Activity in Enriched Anaerobic Digestion Microbial Communities. *mSystems.* American Society for Microbiology; 2019;4.

12. Zhou Y, Leung MHY, Tong X, Lai Y, Tong JCK, Ridley IA, et al. Profiling Airborne Microbiota in Mechanically Ventilated Buildings across Seasons in Hong Kong Reveals Higher Metabolic Activity in Low-Abundance Bacteria. *Environ Sci Technol.* American Chemical Society; 2021;55:249–59.

13. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. *Nat Rev Microbiol.* Nature Publishing Group; 2015;13:217–29.

14. Wang Y, Hatt JK, Tsementzi D, Rodriguez-R LM, Ruiz-Pérez CA, Weigand MR, et al. Quantifying the importance of the rare biosphere for microbial community response to organic pollutants in a freshwater ecosystem. *Appl Environ Microbiol.* American Society for Microbiology; 2017;83:3321–37.

15. He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med.* Nature Publishing Group; 2018;24:1532–5.

16. Liu T, Zhang AN, Wang J, Liu S, Jiang X, Dang C, et al. Integrated biogeography of planktonic

and sedimentary bacterial communities in the Yangtze River. *Microbiome*. BioMed Central Ltd.; 2018;6:1–14.

17. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol*. 2010;12:118–23.

18. Frøslev TG, Kjølner R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, et al. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat Commun*. Nature Publishing Group; 2017;8:1–11.

19. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. Oxford University Press; 2011;27:2194–200.

20. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. Nature Publishing Group; 2016;13:581–3.

21. Amir A, Daniel M, Navas-Molina J, Kopylova E, Morton J, Xu ZZ, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSYSTEMS*. 2017;2:1–7.

22. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. 2016;081257.

23. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*. PeerJ Inc.; 2018;6:e5364.

24. Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kauserud H. Don't make a mista(g)ke: Is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol*. Elsevier; 2012. p. 747–9.

25. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*. BioMed Central Ltd.; 2018;19:1–10.

26. Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis. Illumina. 2017;1–4.

27. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. *Nucleic Acids Res*; 2012;40.

28. Farouni R, Djambazian H, Ferri LE, Ragoussis J, Najafabadi HS. Model-based analysis of sample

index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing. *Nat Commun. Nature Research*; 2020;11:1–8.

29. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv. Cold Spring Harbor Laboratory*; 2017;125724.

30. Li Q, Zhao X, Zhang W, Wang L, Wang J, Xu D, et al. Reliable Multiplex Sequencing with Rare Index Mis-Assignment on DNB-Based NGS Platform. *BMC Genomics*. 2019;20.

31. Jeon SA, Park JL, Park S-J, Kim JH, Goh S-H, Han J-Y, et al. Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. *Genes Genomics* 2021 437. Springer; 2021;43:713–24.

32. Zhu K, Du P, Xiong J, Ren X, Sun C, Tao Y, et al. Comparative Performance of the MGISEQ-2000 and Illumina X-Ten Sequencing Platforms for Paleogenomics. *Front Genet. Frontiers*; 2021;0:1705.

33. Esling P, Lejzerowicz F, Pawlowski J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res. Oxford University Press*; 2015;43:2513–24.

34. Schnell IB, Bohmann K, Gilbert MTP. Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour. Blackwell Publishing Ltd*; 2015;15:1289–303.

35. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. Quantifying and Understanding Well-to-Well Contamination in Microbiome Research. *mSystems. American Society for Microbiology*; 2019;4.

36. Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. *Microbiome* 2021 91. *BioMed Central*; 2021;9:1–22.

37. Liao S, Wang Y, Liu H, Fan G, Sahu SK, Jin T, et al. Deciphering the Microbial Taxonomy and Functionality of Two Diverse Mangrove Ecosystems and Their Potential Abilities To Produce Bioactive Compounds. *mSystems. American Society for Microbiology*; 2020;5.

38. Xue M, Sun H, Wu X, Guan LL, Liu J. Assessment of rumen microbiota from a large dairy cattle cohort reveals the pan and core bacteriomes contributing to varied phenotypes. *Appl Environ Microbiol. American Society for Microbiology*; 2018;84:970–88.

39. Xue MY, Sun HZ, Wu XH, Liu JX, Guan LL. Multi-omics reveals that the rumen microbiome and its metabolome together with the host metabolome contribute to individualized dairy cow performance.

Microbiome. BioMed Central Ltd.; 2020;8:1–19.

40. Hess M. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science* (80-). 2011;463:463–7.

41. Zhao S, Min L, Zheng N, Wang J. Effect of Heat Stress on Bacterial Composition and Metabolism in the Rumen of Lactating Dairy Cows. *Anim* 2019, Vol 9, Page 925. Multidisciplinary Digital Publishing Institute; 2019;9:925.

42. Jiao S, Lu Y. Soil pH and temperature regulate assembly processes of abundant and rare bacterial communities in agricultural ecosystems. *Environ Microbiol.* Blackwell Publishing Ltd; 2020;22:1052–65.

43. Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, et al. Quantifying community assembly processes and identifying features that impose them. *ISME J* 2013 711. Nature Publishing Group; 2013;7:2069–79.

44. Wang J, Jia H. Metagenome-wide association studies: Fine-mining the microbiome. *Nat. Rev. Microbiol.* Nature Publishing Group; 2016. p. 508–22.

45. Sloan WT, Lunn M, Woodcock S, Head IM, Nee S, Curtis TP. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol.* 2006;8:732–40.

46. Webb CO, Ackerly DD, Kembel SW. Phylocom: Software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics.* Oxford Academic; 2008;24:2098–100.

47. Stegen JC, Lin X, Konopka AE, Fredrickson JK. Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J.* Nature Publishing Group; 2012;6:1653–64.

48. NO van G. *Succiniclasticum ruminis* gen. nov., sp. nov., a ruminal bacterium converting succinate to propionate as the sole energy-yielding mechanism. *Int J Syst Bacteriol.* *Int J Syst Bacteriol*; 1995;45:297–300.

49. DJ G, ME M, MC G, N S, MS R, MR T, et al. *Pseudobutyrvibrio xylanivorans* adhesion to epithelial cells. *Anaerobe.* *Anaerobe*; 2019;56:1–7.

50. Zehavi T, Probst M, Mizrahi I. Insights Into Culturomics of the Rumen Microbiome. *Front Microbiol.* *Frontiers*; 2018;0:1999.

51. Boggio GM, Meynadier A, Daunis-I-Estadella P, Marie-Etancelin C. Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency. *PLoS One.* *Public*

800 Library of Science; 2021;16.

801 52. Ozbayram EG, Akyol Ç, Ince B, Karakoç C, Ince O. Rumen bacteria at work: bioaugmentation
802 strategies to enhance biogas production from cow manure. J Appl Microbiol. John Wiley & Sons, Ltd;
803 2018;124:491–502.

804 53. Gordon RE, Barnett DA, Handerhan JE, Hor-nay Pang C. *Nocardia coeliaca*, *Nocardia*
805 *autotrophica*, and the *Nocardin* Strain. Int Assoc Microbiol Soc. 1974;24:54–63.

806 54. Yeh Y-C, Needham DM, Sieradzki ET, Fuhrman JA. Taxon Disappearance from Microbiome
807 Analysis Reinforces the Value of Mock Communities as a Standard in Every Sequencing Run.
808 mSystems. American Society for Microbiology; 2018;3.

809 55. Pascoal F, Magalhães C, Costa R. The Link Between the Ecology of the Prokaryotic Rare
810 Biosphere and Its Biotechnological Potential. Front Microbiol. Frontiers Media SA; 2020;11.

811 56. Pascoal F, Costa R, Magalhães C. The microbial rare biosphere: current concepts, methods and
812 ecological principles. FEMS Microbiol Ecol. Oxford Academic; 2021;97.

813 57. Davis NM, Proctor DiM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and
814 removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. BioMed
815 Central Ltd.; 2018;6:1–14.

816 58. Smirnova E, Huzurbazar S, Jafari F. PERFect: PERmutation filtering test for microbiome data.
817 Biostatistics. Oxford University Press; 2019;20:615–31.

818 59. Cao Q, Sun X, Rajesh K, Chalasani N, Gelow K, Katz B, et al. Effects of Rare Microbiome Taxa
819 Filtering on Statistical Analysis. Front Microbiol. Frontiers Media S.A.; 2021;11:607325.

820 60. Song Z, Schlatter D, Gohl DM, Kinkel LL. Run-to-run sequencing variation can introduce taxon-
821 specific bias in the evaluation of fungal microbiomes. Phytobiomes J. American Phytopathological
822 Society; 2018;2:165–70.

823 61. Sun X, Hu Y-H, Wang J, Fang C, Li J, Han M, et al. Efficient and stable metabarcoding sequencing
824 data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses. Gigabyte.
825 GigaScience Press; 2021;2021:1–15.

826 62. Furman O, Shenhav L, Sasson G, Kokou F, Honig H, Jacoby S, et al. Stochasticity constrained by
827 deterministic effects of diet and age drive rumen microbiome assembly dynamics. Nat Commun.
828 Nature Research; 2020;11:1–13.

829 63. Burns AR, Stephens WZ, Stagaman K, Wong S, Rawls JF, Guillemin K, et al. Contribution of

neutral processes to the assembly of gut microbial communities in the zebrafish over host development. ISME J. Nature Publishing Group; 2016;10:655–64.

64. Layeghifard M, Hwang DM, Guttman DS. Disentangling Interactions in the Microbiome: A Network Perspective. Trends Microbiol. Elsevier Ltd; 2017. p. 217–28.

65. Chen J, Chen Z, Liu S, Guo W, Li D, Minamoto T, et al. Revealing an Invasion Risk of Fish Species in Qingdao Underwater World by Environmental DNA Metabarcoding. J Ocean Univ China. Science Press; 2021;20:124–36.

66. Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. Nature Publishing Group; 2012;6:1621–4.

67. Zou K, Chen J, Ruan H, Li Z, Guo W, Li M, et al. eDNA metabarcoding as a promising conservation tool for monitoring fish diversity in a coastal wetland of the Pearl River Estuary compared to bottom trawling. Sci Total Environ. Elsevier; 2020;702:134704.

68. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience. Oxford University Press; 2018;7:1–6.

69. Magoč T, Salzberg SL. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics. Oxford University Press; 2011;27:2957–63.

70. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.

71. Zhang J, Liu YX, Zhang N, Hu B, Jin T, Xu H, et al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nat Biotechnol. Nature Publishing Group; 2019;37:676–84.

72. Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ. PeerJ Inc.; 2018;6:e4652.

73. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. Public Library of Science; 2010;5:e9490.

74. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. Oxford University Press (OUP); 2021;

75. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware

860 quantification of transcript expression. *Nat Methods*. Nature Publishing Group; 2017;14:417–9.

861 76. Russel J. Russel88/MicEco: v0.9.15. 2021;

862 77. Xun W, Liu Y, Li W, Ren Y, Xiong W, Xu Z, et al. Specialized metabolic functions of keystone

863 taxa sustain soil microbiome stability. *Microbiome*. BioMed Central Ltd; 2021;9:1–15.

864 78. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput*

865 *Biol*. 2012;8:e1002687.

866 79. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and

867 Manipulating Networks Visualization and Exploration of Large Graphs. *Proc. Int. AAAI Conf. Web*

868 *Soc. Media*. 2009 Mar.

869 80. Wu MH, Chen SY, Chen JW, Xue K, Chen SL, Wang XM, et al. Reduced microbial stability in the

870 active layer is associated with carbon loss under alpine permafrost degradation. *Proc Natl Acad Sci U S*

871 *A*. National Academy of Sciences; 2021;118.

872

Legend of figures and tables

Figure 1. Amplicon sequencing of the commercial mock community. (A) The number of shared OTUs among technical triplicates for both DNBSEQ and Novaseq platforms, and the number of shared unique OTUs detected by pooled technical triplicates by different sequencing platforms and the theoretical members of the mock community. (B) Sankey diagram showing the phylogeny and number of unique OTUs detected by each sequencing platform. For the expected OTUs, the height of the bars was proportional to the averaged relative abundance reported by two sequencing platforms. OTUs detected by both DNBSEQ and NovaSeq, or specifically detected by NovaSeq platform were indicated and color coded. (C) Phylogenetic tree of the expected members of the microbes and OTUs detected by both sequencing platforms.

Figure 2. Comparison of the amplicon sequencing results of three typical ecological systems between DNBSEQ and NovaSeq sequencing platforms. (A) Evaluation of the reproducibility of the amplicon sequencing results of DNBSEQ and NovaSeq sequencing platform in revealing the membership of microbes of abundant, moderate and rare taxa subcommunity from ecosystems with various complexity (DNB All, DNB 2, DNB 3 denotes number of all unique OTUs detected by DNBSEQ platform, consistently detected by at least 2 technical replicates, and consistently detected by all three technical replicates, respectively. Similar naming scheme was used for the NovaSeq platform). (B) Comparison of the alpha diversity of the overall community, abundant sub-community, moderate sub-community, and rare sub-community as revealed by DNBSEQ and NovaSeq sequencing platforms.

Figure 3. Characteristics of cow rumen microbial communities revealed by different sequencing platforms. (A) Number of shared abundant, moderate, and rare unique OTUs detected by different sequencing platforms in cow rumen. (B) Comparison of Chao I index, Shannon index, and the phylogenetic diversity of the overall community, abundant sub-community, moderate sub-community, and rare sub-community as revealed by DNBSEQ and NovaSeq sequencing platforms in cow rumen. (C) Phylogenetic distribution of OTUs detected by both or either of the DNBSEQ or NovaSeq sequencing platforms in cow rumen. The inner circle was color coded by phylogeny, and the outer circle was color coded according to whether the OTU was consistently detected by both sequencing platforms, or specifically detected by either of DNBSEQ or NovaSeq sequencing platform. Heatmap in the middle panel shows the relative abundance of different phylum revealed

by DNBSEQ or NovaSeq. Bar chart on the right panel shows the number of unique OTUs specifically detected by NovaSeq platform in different phylum.

Figure 4. Cow rumen microbial community assembly mechanism. (A) The fit of Sloan's neutral model to the cow rumen microbial communities revealed by each sequencing platform. The proportion of OTUs distributed within the Sloan's neutral model, at the upper or lower part were indicated. R^2 values indicate the overall fit of the model, and m values indicate the estimated migration rate. Dashed lines represent 95% confidence intervals around the model prediction. (B) Bar charts shows the OTUs' frequency distribution on each sequencing platform. More singleton OTUs and low frequent OTUs were observed on NovaSeq sequencing platform, which was largely driven by the OTUs specifically detected by NovaSeq platform. (C) The distribution of β NTI values between cow rumen microbial communities revealed by each sequencing platform. Each point represents a β NTI value. A $|\beta$ NTI| value of less than 2 (grey shaded region) indicates stochastic assembly processes; a β NTI value of less than -2 indicates a homogeneous selection event; and a β NTI value of greater than 2 indicates a heterogeneous selection event.

Figure 5. Cow rumen microbial interaction network revealed by DNBSEQ and NovaSeq. (A) The cooccurrence network of microbial communities in cow rumen. Each node represents an OTU and was color coded by both sub-community type and sequencing platform where appropriate. The size of the node is proportional to its node degree. Each line represents a potential correlation interaction, with blue lines indicating positive interaction while green lines indicating negative interaction. Only interactions with a correlation coefficient greater than 0.75 and significance of P smaller than 0.05 were plotted. (B) Node degree distribution and the Natural Connectivity change under node attack test of the microbial interaction network revealed by DNBSEQ and NovaSeq platforms. (C) Phylogeny of the top 20 nodes with the highest degree in the microbial interaction networks revealed by DNBSEQ and NovaSeq. The size of the solid circle is proportional to its node degree and color coded by the sub-community type. Solid or open asterisk indicates the platform by which the OTUs were identified as the top 20 nodes with the highest degree.

Legend of supplementary figures and tables.

Figure S1. Schematic presentation of the design and flow of this study. This study was designed to evaluate the potential harm of false positives introduced by index misassignment during the sequencing to the interpretation of metabarcoding or here amplicon sequencing studies. In part 1, mock communities with known microbial composition were used to confirm the presence of index misassignment introduced false positives and the rate of this misassignment on different sequencing platform; In part 2, samples from several typical ecosystems with various complexity were sequenced to test how index misassignment could affect the estimation of community diversity for different samples; In part 3, a case study using a set of cow rumen samples was conducted to evaluate how index misassignment caused false positives could affect the interpretation of microbial ecological roles and community assembly mechanisms.

Figure S2. UpSet graph showing the number of shared OTUs among the technical triplicates of DNBSEQ and NovaSeq sequencing platforms for customized mock community with four (A) or seven (B) microbes used in the current study.

Figure S3. Venn diagrams showing the shared OTUs among distinct samples sequenced within the same batch of sequencing run on the Novaseq platform. The large number of shared OTUs among distinct samples (including two customized communities) were likely sample wise cross contaminations caused by index misassignment.

Figure S4. Comparison of the amplicon sequencing results of three typical ecological systems between DNBSEQ and NovaSeq sequencing platforms. (A) Evaluation of the reproducibility of the amplicon sequencing results of DNBSEQ and NovaSeq sequencing platform in revealing the accumulated relative abundance of microbes of abundant, moderate and rare taxa subcommunity from ecosystems with various complexity (DNB All, DNB 2, DNB 3 denotes number of all unique OTUs detected by DNBSEQ platform, consistently detected by at least 2 technical replicates, and consistently detected by all three technical replicates, respectively. Similar naming scheme was used for the NovaSeq platform). (B) Cross-verification of the OTUs identified by amplicon sequencing on DNBSEQ and NovaSeq platforms and shotgun sequencing results of the same set of mangrove sediment samples. (C) Cross-verification of the OTUs identified by amplicon sequencing on DNBSEQ and NovaSeq platforms and shotgun sequencing results of the same set of mice gut samples.

Figure S5. Weighted (A) and unweighted (B) UniFrac distance-based clustering of the amplicon sequencing results revealed by DNBSEQ and NovaSeq sequencing platforms for samples from the three typical ecosystems.

Figure S6. Weighted (A) and unweighted (B) UniFrac distance-based beta diversity of the cow rumen microbial communities revealed by DNBSEQ and NovaSeq sequencing platforms.

Table S1. OTU table of the amplicon sequencing result of the ZymoBIOMICS™ Microbial Community DNA Standard.

Table S2. OTU table of the customized mock community with four bacteria.

Table S3. OTU table of the customized mock community with seven bacteria.

Table S4. Summary of the number of reads and clustered OTUs of samples from the three typical ecosystems.

Table S5. Correlation test of the relative abundance of OTUs and physiochemical properties of cow rumen fluid.

Table S6. Mapping results of the cloned sequences of specifically designed primers for *Nocardia coeliaca* to the entire OTU representative sequences.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Sequencing introduced false positive rare taxa lead to distorted microbial community diversity, assembly, and interaction interpretation in amplicon studies

Figure S1

(A)

(B)

Figure S2

Batch1

Batch2

Batch3

Figure S3

(A)

weighted_unifrac cluster tree

weighted_unifrac cluster tree

weighted_unifrac cluster tree

(B)

unweighted_unifrac cluster tree

unweighted_unifrac cluster tree

unweighted_unifrac cluster tree

46

Figure S6

Supplementary Files

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

- [SuppTables.xlsx](#)