Current Microbiology (2018) 75:1240–1246
https://doi.org/10.1007/s00284-018-1516-y

Necessary Sequencing Depth and Clustering Method to Obtain Relatively Stable Diversity Patterns in Studying Fish Gut Microbiota

Fanshu Xiao[1] · Yuhe Yu[2] · Jinjin Li[3] · Philippe Juneau[4] · Qingyun Yan[1]

Received: 24 March 2018 / Accepted: 22 May 2018 / Published online: 25 May 2018

Abstract

The 16S rRNA gene is one of the most commonly used molecular markers for estimating bacterial diversity during the past decades. However, there is no consistency about the sequencing depth (from thousand to millions of sequences per sample), and the clustering methods used to generate OTUs may also be different among studies. These inconsistent premises make effective comparisons among studies difficult or unreliable. This study aims to examine the necessary sequencing depth and clustering method that would be needed to ensure a stable diversity patterns for studying fish gut microbiota. A total number of 42 samples dataset of *Siniperca chuatsi* (carnivorous fish) gut microbiota were used to test how the sequencing depth and clustering may affect the alpha and beta diversity patterns of fish intestinal microbiota. Interestingly, we found that the sequencing depth (resampling 1000–11,000 per sample) and the clustering methods (UPARSE and UCLUST) did not bias the estimates of the diversity patterns during the fish development from larva to adult. Although we should acknowledge that a suitable sequencing depth may differ case by case, our finding indicates that a shallow sequencing such as 1000 sequences per sample may be also enough to reflect the general diversity patterns of fish gut microbiota. However, we have shown in the present study that strict pre-processing of the original sequences is required to ensure reliable results. This study provides evidences to help making a strong scientific choice of the sequencing depth and clustering method for future studies on fish gut microbiota patterns, but at the same time reducing as much as possible the costs related to the analysis.

Introduction

Although microorganisms have existed on Earth for over billions of years and almost evolved into every conceivable niche on the planet, there are still many unanswered questions about the microbial cosmos [1]. Fortunately, in the past decade microbiomics ushered in a new era for microbiome research [2]. The arrival of 'microbiome spring' and the increasing findings on microorganisms benefited greatly from the continuous development of sequencing and related analysis methods. The metagenome-based sequencing tools improved our ability to address the issues involving microbial diversity, function, interaction, evolution and dynamics in relation to environmental variations.

The first-generation sequencing (i.e. Sanger sequencing) has been considered as one of the most influential innovations in biological studies [3], which predominated the sequencing market for nearly 30 years. From the first 16S rRNA gene clone library that was directly sequenced from environmental samples [4], the Sanger sequencing greatly enhanced our ability to understand the culture-independent microbial world [5]. That strategy for microbial studies

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00284-018-1516-y) contains supplementary material, which is available to authorized users.

✉ Qingyun Yan
yanqy6@mail.sysu.edu.cn

[1] Environmental Microbiomics Research Center, Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, and the School of Environmental Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

[2] Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China

[3] Qilu Normal University, Jinan 250013, China

[4] Département des Sciences biologiques - GRIL - TOXEN, Laboratory of Aquatic Microorganism Ecotoxicology, Université du Québec à Montréal, Succ. Centre-Ville, C.P. 8888, Montreal, QC H3C 3P8, Canada

remained popular until 2005 when the second-generation sequencing appeared (i.e. high-throughput sequencing) [6], permitting another leap in our understanding of environmental microbiomes. This new era is characterized by high throughput, low sequencing errors and much lower costs than the Sanger sequencing strategy. Although the third-generation sequencing (i.e. single-molecule sequencing) has been developed and already routinely used to sequence microbial genomes [7], the second-generation sequencing will continue its important role in culture-independent studies of microbial communities. Among all the molecular markers (genes) that are used for evaluating microbial diversity, the 16S rRNA gene is one of the most commonly used [8]. The 16S rRNA gene analyses benefit the developed standard procedures and are relatively inexpensive. Furthermore, the availability of huge 16S rRNA gene datasets and reference sequences also makes comparisons among different studies reliable and effective.

However, the results based on the 16S rRNA gene sequencing may partly depend on the applied analysis procedures. For example, Fierer et al. [9] found sequencing depth may affect the microbial diversity evaluation, and therefore suggested to rarefy all the samples to the same sequencing depth before downstream analysis. Another example is the UPARSE clustering method, used for generating OTUs, which was found to be closer to the expected number of species in a community than other methods [10]. Therefore, how to select a reasonable sequencing depth for addressing the issues of interest is especially important, and this choice may optimize the budget needed to analyse as many samples as possible. Although Lundin et al. [11] suggested that 1000 sequences/sample can give equally good results as the ones obtained with more than 15,000 sequences for estimating the beta diversity trends of water or sediment community, how deeply fish gut microbial communities need to be sequenced was never experimentally tested and need further confirmation.

The fish-dependent gut ecosystems are considerably different from the surrounding natural environments (such water or sediment). Indeed, environmental microbial community is mainly determined by the environmental conditions, and gut microbes colonized in the fish can be also significantly affected by the fish's development, feeding and health [12–14]. Therefore, understanding the community assembly and turnover rules in the gut ecosystems should be more complicated than that found in the natural environmental ecosystems. A dataset of fish gut microbiota was used to address the question of which sequencing depth is sufficient and which clustering method can optimally reveal the microbial patterns to obtain stable results in analysing 16S rRNA gene sequences. The ultimate aim of the present work is to provide guidance in further studies to improve our understanding of diversity patterns of the fish gut microbiota.

Materials and Methods

Sampling

The gut samples ($n = 42$) of carnivorous *Siniperca chuatsi* were collected as described previously [13]. In brief, the hatched larvae were sampled at the first day post-hatching (dph) and then sampled every 5 days from 3 to 23 dph. After that, the sampling was paused and only adult individuals (500–800 g) were collected from Poyang Lake, China. The larval and juvenile individuals collected were transported to laboratory with in situ water and immediately dissected under aseptic conditions to get the gut samples, and the captured adult fishes were kept at −20 °C and transported to laboratory until following procedures. The collection, preservation and research of wild animal and endangered species are approved by national regulations "China biodiversity conservation strategy and action plan". All protocols involved in the animal experiment were approved by the Institutional Animal Care and Use Committee of Institute of Hydrobiology, Chinese Academy of Sciences (Approval ID: Keshuizhuan 08529).

DNA Extraction and PCR Amplification

Genomic DNA of the collected gut microbiota was extracted using the PowerFecal® DNA Isolation Kit (Mo Bio, CA, USA) according to the manufacturer's instructions. The DNA concentrations and quality were determined with a ND-1000 spectrophotometer (NanoDrop, DE, USA) and all the samples were diluted to the same concentration for subsequent PCR amplification. The primers 515f (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806r (5′-GGACTACHVGGGTWTCTAAT-3′) targeting the V4 region of 16S rRNA gene were used to analyse the community patterns of gut microbiota according to Wu et al. [15]. All PCRs were conducted in triplicates for each sample, and in each of 25 μl PCR mixture contains 1× Buffer II, 0.4 μM of each primer, 0.5 U AccuPrime™ Taq (Invitrogen, CA, USA) and 10 ng genomic DNA. PCRs were performed using the following conditions: 1 min at 94 °C, followed by 10 cycles of 20 s at 94 °C, 25 s at 53 °C, and 45 s at 68 °C, with a post-amplification extension of 10 min at 68 °C. The products were purified with Agencourt® Ampure® XP (Beckman, CA, USA) according to the manufacturer's instructions. The purified DNA was then applied as template to perform the second PCR amplification using 20 cycles with the same primer set but the reverse primer contains appropriate adapters and different barcodes. PCR products were visualized using 1% agarose gels stained with ethidium bromide, and negative

controls were always performed to make sure there was no contamination.

Sequencing Analysis

The concentration of each PCR product was quantified with a PicoGreen dsDNA Assay Kit (Invitrogen, CA, USA). All the 42 samples were then equally combined and followed by gel purification using a QIAquick Gel Extraction Kit (Qiagen, CA, USA). The purified DNA was re-quantified with PicoGreen, and then combined with other similarly prepared DNA libraries for sequencing at the Institute for Environmental Genomics (Norman, OK, USA) using a MiSeq platform (Illumina, CA, USA). Quality filtering and processing of sequencing reads were conducted on our Galaxy pipeline (http://zhoulab5.rccc.ou.edu/root) as previously described methods [13]. After trimming the primers and delete the sequences containing uncertain Ns, the high-quality sequences with 245–260 bp were kept to generate OTU table using the method (97% cutoff) of UPARSE and UCLUST, respectively. OTUs were generated based on the clustering results, and taxonomic annotation of individual OTUs was achieved based on representative sequences using RDP's 16S Classifier 2.5 [15]. To explore the possible effect of sequencing depth on the community patterns, all the samples were rarefied to the same sequencing depth by resampling from 1000 to 11,000 before downstream analysis. Therefore, we obtained 11 sub-datasets for each clustering method to compare the possible effects from the sequencing depth. The raw data of the high-throughput sequencing have been submitted to the Sequence Read Archive (SRA) of NCBI with the run accession ID SRR6144368-SRR6144409.

Statistical Analysis

The bacterial diversity patterns that generated under different clustering methods (UPARSE and UCLUST) with wide ranges of sequencing depths (1000–11,000) were evaluated by the following statistical methods: (i) comparison of the alpha and beta diversity trends; (ii) significance tests of alpha diversity were performed through an analysis of variance (ANOVA) with least significant difference (LSD) to examine whether differences were significant or not; (iii) DCA ordination on the basis of community composition to show the general community patterns; and (iv) nonparametric tests including multiple-response permutation procedure (MRPP) with Bray–Curtis distances for comparing community differences. All statistics were performed using R packages of 'vegan' and 'picante' (R Foundation for Statistical Computing, Vienna, Austria).

Results and Discussion

For sequencing analysis of the 16S RNA gene, it was advanced that a more realistic image of the microbial community occurred when more sequences were obtained [16]. Therefore, in an ideal world, where the sequencing cost (always related to the sequencing depth) is not a limit, all studies theoretically should get as many sequences as possible for better understanding of the communities. On the other hand, the clustering method of UPARSE [10] is considered to be more reliable in generating OTU by testing the mock communities, but the UCLUST method is also widely used due to its advantages of higher speed, improved sensitivity and clustering at lower identities [17]. This study want to examine the necessary sequencing depth and clustering method to obtain stable diversity patterns in studying fish gut microbiota.

Interestingly, the sequencing depth (resampling 1000–11,000 sequences per sample) and clustering methods (UPARSE and UCLUST) did not bias our estimates of alpha and beta diversity patterns. The bacterial diversity estimates of Shannon index and Pielou's evenness across the three investigated stages (i.e. 1–13 dph, 18–23 dph and adult) showed no significant variation at different sequencing depth (Fig. 1). Moreover, the UPARSE and UCLUST methods showed similar trends in the alpha diversity patterns. Briefly, when fish open their mouth to feed at the early stage, various kinds of bacteria found in the surrounding environment can easily immigrate into the new gut habitat. Therefore, the initial stage (1–13 dph) showed relatively high Shannon diversity (~3.0), and then it decreased to ~2.0 at the stage of 18–23 dph and was constant thereafter. Analyses of trends in Pielou's evenness also showed the same tendency as that of Shannon index, i.e. the stage of 1–13 dph showed relatively higher diversity than the other two stages. Regarding the bacterial OTUs that classified into different taxa, we can see that the Proteobacteria is always the largest phylum regardless of the clustering methods and sequencing depth. The Proteobacteria accounted for ~34% (UPARSE) or ~43% (UCLUST) of the detected OTUs, followed by the phyla of Firmicutes, Bacteroidetes, Actinobacteria and Cyanobacteria. We noticed also that approximately 10% of the OTUs cannot be classified into any known phylum (Fig. S1). The beta diversity patterns as visualized in DCA ordination indicated that the communities were always grouped according to the fish developmental stage and showed almost identical patterns for all the sub-datasets (Fig. 2, Fig. S2). Furthermore, the dissimilarity test confirmed that the differences of gut microbiota among the host developmental stages are significant (Table 1).

Our results indicated that an overall image of the community diversity patterns among fish gut samples does

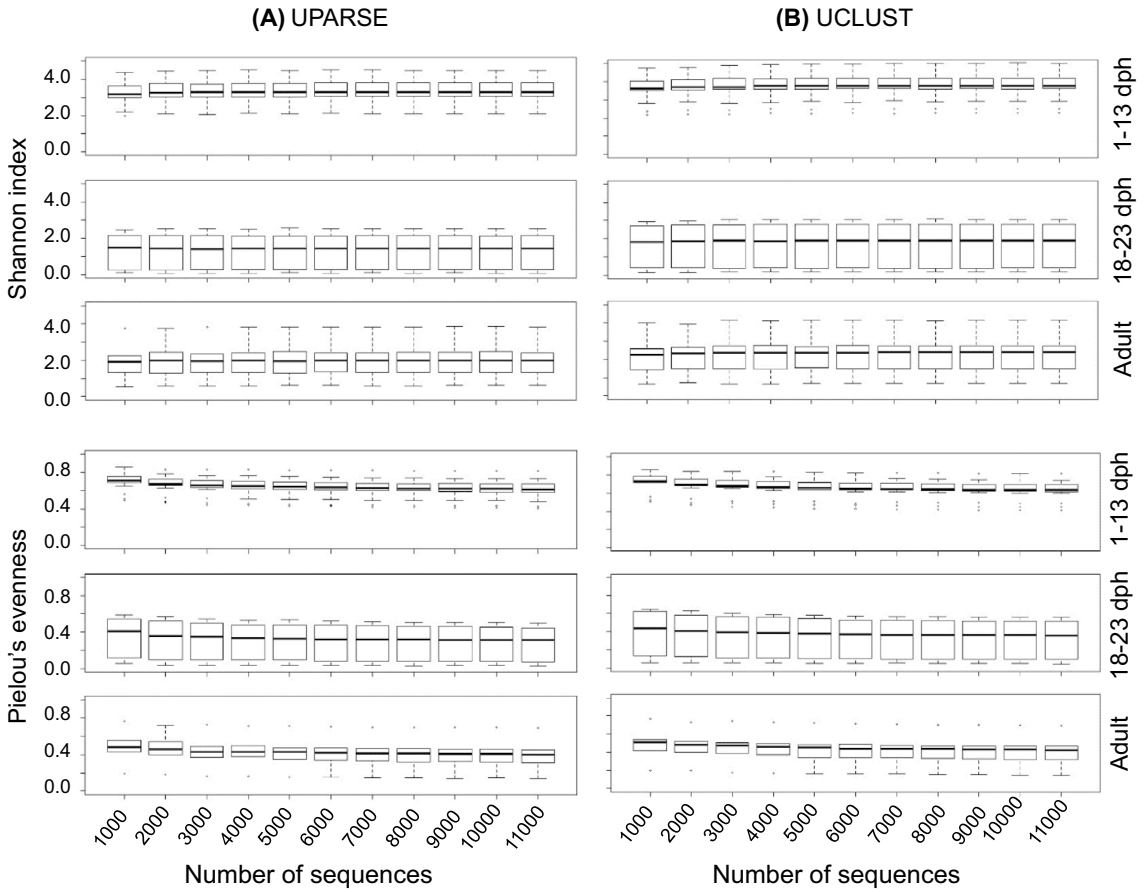

Fig. 1 Clustering (UPARSE and UCLUST) and sequencing depth (from 1000 to 11,000) do not significantly affect the alpha diversity of the investigated fish gut microbiota

not necessarily need high sequencing depth. For example, 1000 sequences per sample seem adequate, since we obtained almost similar community patterns as an evaluation done with more than ten times of sequences (11,000), and the diversity patterns remain consistent in the analyses by resampling 1000 to 11,000 (Fig. 2, Fig. S2). Our finding about this minimum sequence requirement for the fish gut microbial ecosystem is consistent with the ones in soil [18], water and sediment [11] ecosystems. Lundin et al. [11] found that 1000 denoised sequences/sample can explained up to 90% the trends in beta diversity among water and sediment samples. By analysing a diverse array of 25 environmental samples and three "mock communities", Caporaso et al. [16] found that 2000 sequences per sample are sufficient to recapture the same relationship among samples that were observed with 3.1 million reads per sample. Therefore, it is really not necessary to acquire a large number of sequences if the aim of a study is not to follow the rare taxa.

Actually, in most microbial ecology studies, the trend among samples with different spatial or temporal scales is the major interest rather than the absolute number of species. Moreover, there is still a lack of method to detect all the microorganisms in an environment [19] or even in a very simple community. For ecological studies, we generally use dominant taxa to represent how microbial community differs among samples across the environmental, temporal or geographic gradients. In most of the abundance-weighted analysis, the overall microbial community composition is mainly reflected by the dominant and common taxa. Therefore, if a study only aims to explain the overall diversity patterns rather than to focus on the related issues of rare taxa, keeping a shallow sequencing depth (e.g. 1000 sequences per sample), but including as much samples as possible should be a more reasonable study design [20] than to sequence deeply for a very limited number of samples. To the best of our knowledge, an exhaustive census of a microbial community in an ecosystem is usually impossible. Therefore, community diversity must be inferred from representative samples taken from the ecosystem [19]. However, the estimation of diversity is sample-size dependent [21] and sensitive to sample coverage [22]. Consequently, with a fixed budget for revealing diversity pattern of microorganisms in an ecosystem, keeping relatively shallow sequencing depth

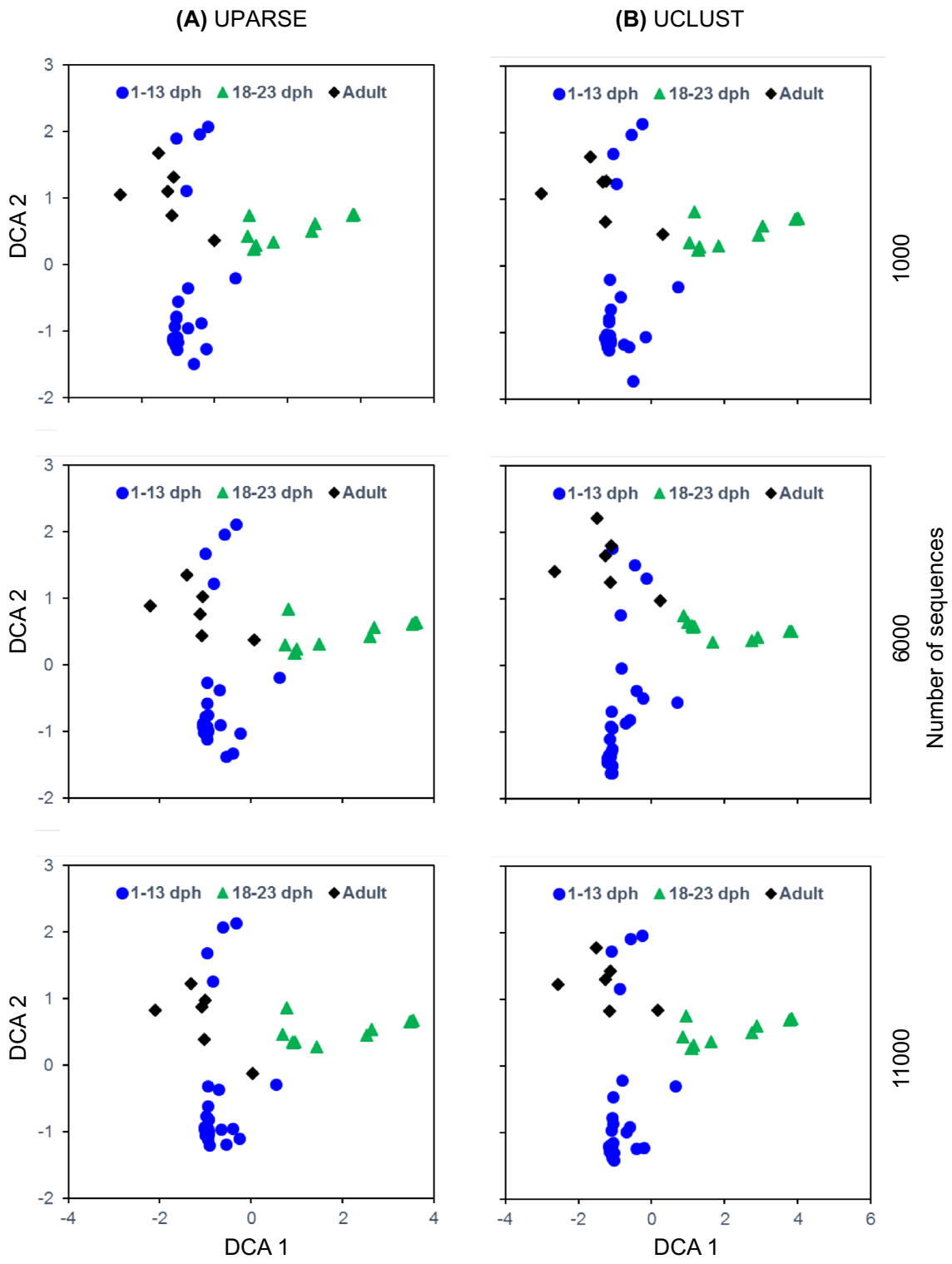

Fig. 2 DCA ordinations showing the bacterial community patterns are similar at different sequencing depths (only the resampling of 1000, 6000 and 11,000 sequences are presented) and independent of clustering methods (UPARSE and UCLUST)

Table 1 Summary of Bray–Curtis distance-based dissimilarity test determined by the MRPP

	1–13 dph vs. 18–23 dph		1–13 dph vs. adult		18–23 dph vs. adult	
	Delta	*P*	Delta	*P*	Delta	*P*
UPARSE						
1000	0.652	0.001	0.628	0.001	0.714	0.003
2000	0.653	0.001	0.625	0.001	0.710	0.001
3000	0.650	0.001	0.624	0.001	0.710	0.001
4000	0.650	0.001	0.623	0.001	0.709	0.002
5000	0.646	0.001	0.619	0.001	0.708	0.001
6000	0.646	0.001	0.620	0.001	0.708	0.001
7000	0.645	0.001	0.619	0.001	0.708	0.001
8000	0.645	0.001	0.618	0.001	0.708	0.001
9000	0.646	0.001	0.620	0.001	0.708	0.001
10,000	0.645	0.001	0.619	0.001	0.708	0.001
11,000	0.645	0.001	0.618	0.001	0.707	0.001
UCLUST						
1000	0.678	0.001	0.660	0.001	0.726	0.001
2000	0.670	0.001	0.648	0.001	0.719	0.001
3000	0.667	0.001	0.648	0.001	0.718	0.002
4000	0.663	0.001	0.644	0.001	0.720	0.004
5000	0.663	0.001	0.644	0.001	0.718	0.003
6000	0.663	0.001	0.642	0.001	0.717	0.001
7000	0.662	0.001	0.641	0.001	0.718	0.002
8000	0.661	0.001	0.640	0.001	0.717	0.001
9000	0.661	0.001	0.640	0.001	0.716	0.004
10,000	0.661	0.001	0.641	0.001	0.717	0.003
11,000	0.660	0.001	0.639	0.001	0.716	0.002

on multi-samples could be more efficient than analysing limited samples with deep sequencing.

Besides the sequencing depth as discussed above that may affect biodiversity estimates, the choice of the clustering method that group sequences to generate OTUs can also significantly affect the diversity estimates [23]. Although various clustering methods are available to generate OTUs for the 16S rRNA gene analysis, previous studies found that the UPARSE was more precise and produced more consistent OTU numbers than other methods [10, 24]. Therefore, in our most recent studies we always used the UPARSE to generate OTUs in analysing fish gut microbiota [13, 25] and bacterioplankton diversity [26]. However, here we also wanted to know the possible effect of the clustering methods (UPARSE and UCLUST) under different sequencing depth, since this assessment should be valuable to guide the selection of optimal clustering method for particular study. Interestingly, these two clustering methods did not generate significant differences about the diversity trends at any sequencing depth (Figs. 1, 2). This may be mainly due to the same pre-processing [13, 26] applied for the sequences, as the pre-processing really found to have larger impact than the clustering methods themselves [27]. Consequently, strict pre-processing of sequences is suggested to get reliable results regardless of any of the clustering method used. Although we should acknowledge that a suitable sequencing depth may differ case by case, the present study showed that for future studies on fish gut microbiota patterns, the choice of the sequencing depth and clustering method can be compatible with a strong scientific strength and relative low expenses related to the analyses.

Acknowledgements We thank Dr. Liyou Wu and Yujia Qin from the University of Oklahoma for help with sequencing and analysis tools. This work was supported by the National Natural Science Foundation of China (31672262), the Hundred Talents Program through Sun Yat-sen University (38000-18821107) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-06210).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethics Approval The collection, preservation and research of wild animal and endangered species are approved by national regulations "China biodiversity conservation strategy and action plan". All protocols involved in the animal experiment were approved by the Institutional Animal Care and Use Committee of Institute of Hydrobiology, Chinese Academy of Sciences (Approval ID: Keshuizhuan 08529).

References

1. Gibbons SM, Gilbert JA (2015) Microbial diversity—exploration of natural ecosystems and microbiomes. Curr Opin Genet Dev 35:66–72
2. Sharma P, Brahma V, Sharma A, Dubey RK, Sidhu GS, Malhotra PK (2015) Microbiomics: an approach to community microbiology. In: Barh D, Khan M, Davies Sarwar E, (eds) PlantOmics: the omics of plant science. Springer India, New Delhi, pp 633–653
3. Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111
4. Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173:4371–4378
5. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol R 68:669–685
6. Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8
7. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33:623–630
8. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821
9. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL et al (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci USA 109:21390–21395
10. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998
11. Lundin D, Severin I, Logue JB, Ostman O, Andersson AF, Lindstrom ES (2012) Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity? Environ Microbiol Rep 4:367–372
12. Bolnick DI, Snowberg LK, Hirsch PE, Lauber CL, Knight R, Caporaso JG et al (2014) Individuals' diet diversity influences gut microbial diversity in two freshwater fish (threespine stickleback and Eurasian perch). Ecol Lett 17:979–987
13. Yan Q, Li J, Yu Y, Wang J, He Z, Van Nostrand JD et al (2016) Environmental filtering decreases with fish development for the assembly of gut microbiota. Environ Microbiol 18:4739–4754
14. Xiong J, Zhu J, Dai W, Dong C, Qiu Q, Li C (2017) Integrating gut microbiota immaturity and disease-discriminatory taxa to diagnose the initiation and severity of shrimp disease. Environ Microbiol 19:1490–1501
15. Wu L, Wen C, Qin Y, Yin H, Tu Q, Van Nostrand JD et al (2015) Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. BMC Microbiol 15:125
16. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ et al (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108:4516–4522
17. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461
18. Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N (2013) Temporal variability in soil microbial communities across land-use types. ISME J 7:1641–1650
19. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS (2013) Robust estimation of microbial diversity in theory and in practice. ISME J 7:1092–1101
20. Prosser JI (2010) Replicate or lie. Environ Microbiol 12:1806–1810
21. Soetaert K, Heip C (1990) Sample-size dependence of diversity indices and the determination of sufficient sample size in a high-diversity deep-sea environment. Mar Ecol Prog Ser 59:305–307
22. Brose U, Martinez ND, Williams RJ (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology 84:2364–2377
23. Bachy C, Dolan JR, López-García P, Deschamps P, Moreira D (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. ISME J 7:244–255
24. Flynn JM, Brown EA, Chain FJJ, MacIsaac HJ, Cristescu ME (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecol Evol 5:2252–2266
25. Li X, Zhou L, Yu Y, Ni J, Xu W, Yan Q (2017) Composition of gut microbiota in the gibel carp (*Carassius auratus gibelio*) varies with host development. Microb Ecol 74:239–249
26. Yan Q, Stegen JC, Yu Y, Deng Y, Li X, Wu S et al (2017) Nearly a decade-long repeatable seasonal diversity patterns of bacterioplankton communities in the eutrophic Lake Donghu (Wuhan, China). Mol Ecol 26:3839–3850
27. May A, Abeln S, Crielaard W, Heringa J, Brandt BW (2014) Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations. Bioinformatics 30:1530–1538