Microbio diversity analysis method

目录

1. Experiment procedure

1.1. Sub-region Sequencing（Illumina）

DNA extraction

Microbial DNA was extracted using the HiPure Soil DNA Kits (or HiPure Stool DNA Kits) (Magen, Guangzhou, China) according to manufacturer's protocols.

PCR amplification

The 16S rDNA target region of the ribosomal RNA gene were amplified by PCR（95 °C for 5 min, followed by 30 cycles at 95 °C for 1 min, 60 °C for 1 min, and 72 °C for 1 min and a final extension at 72 °C for 7 min）using primers listed in the table [1]. PCR reactions were performed in triplicate 50 μL mixture containing 10 μL of 5 × Q5@ Reaction Buffer, 10 μL of 5 × Q5@ High GC Enhancer, 1.5 μL of 2.5 mM dNTPs, 1.5 μL of each primer (10 μM), 0.2 μL of Q5@ High-Fidelity DNA Polymerase, and 50 ng of template DNA. Related PCR reagents were from New England Biolabs, USA.

Illumina Novaseq 6000 sequencing

Amplicons were extracted from 2% agarose gels and purified using the AxyPrep DNA Gel Extraction Kit (Axygen Biosciences, Union City, CA, U.S.) according to the manufacturer's instructions and quantified using ABI StepOnePlus Real-Time PCR System (Life Technologies，Foster City, USA). Purified amplicons were pooled in equimolar and paired-end sequenced (PE250) on an Illumina platform according to the standard protocols. The raw reads were deposited into the NCBI Sequence Read Archive (SRA) database (Accession Number: SRP******).

1.2. Full length sequencing（PacBio）

DNA extraction and PCR amplification

Microbial DNA was extracted using the HiPure Soil DNA Kits (or HiPure Stool DNA Kits) (Magen, Guangzhou, China) according to manufacturer's protocols. The full length 16S rDNA were amplified by PCR（95 °C for 2 min, followed by 35 cycles of 95 °C for 30 s, 60 °C for 45 s, and 72°C for 90s, with a final extension 72°C for 10 min ） using primers 27F: AGRGTTYGATYMTGGCTCAG；1492 R: RGYTACCTTGTTACGACTT [44]（Listed in table 1）. The PCR reaction was carried out in a 50 μL reaction volume with TransGen High-Fidelity PCR SuperMix (TransGen Biotech, Beijing, China), 0.2 μM forward and reverse primers, and 5 ng template DNA.

PacBio sequencing

Amplicons were evaluted with 2% agarose gels and purified using the AxyPrep DNA Gel Extraction Kit (Axygen Biosciences, Union City, CA, USA) according to the manufacturer's instructions. Sequencing libraries were generated using SMRTbell TM Template Prep Kit (PacBio, Menlo Park, CA, USA) following manufacturer's recommendation. The library quality was assessed with Qubit 3.0 Fluorometer (ThermoFischer Scientific,USA) and FEMTO Pulse system (Agilent Technologies, Santa Clara, CA, USA). The libraries were sequenced on the PacBio Sequel platform. The raw reads were deposited into the NCBI Sequence Read Archive (SRA) database (Accession Number: SRP******).

Table 1. Primer information

type	region	primer name	primer sequence	product length	reference
16S	V1-V9 Full length	27F	AGRGTTTGATYNTGGCTCAG	~1465	[44]
		1492R	TASGGHTACCTTGTTASGACTT		
16S	V4	515F	GTGYCAGCMGCCGCGGTAA	~292	[40，41]
		806R	GGACTACNVGGGTWTCTAAT		
16S	V3-V4	341F	CCTACGGGNGGCWGCAG	~466	[1]
		806R	GGACTACHVGGGTATCTAAT		
16S	V4-V5	515F	GTGCCAGCMGCCGCGGTAA	~412	[2]
		907R	CCGTCAATTCCTTTGAGTTT		
16S	V5-V7	799F	AACMGGATTAGATACCCKG	~414	[3]
		1193R	ACGTCATCCCCACCTTCC		
16S	V4-V5	Arch519F	CAGCMGCCGCGGTAA	~416	[4]
		Arch915R	GTGCTCCCCCGCCAATTCCT		
18S	V4	528F	GCGGTAATTCCAGCTCCAA	~260	[5]
		706R	AATCCRAGAATTTCACCTCT		
ITS	ITS1	ITS1_F_KYO2	TAGAGGAAGTAAAAGTCGTAA	~366	[42]
		ITS86R	TTCAAAGATTCGATGATTCAC		
ITS	ITS1	ITS1-F	CTTGGTCATTTAGAGGAAGTAA	~321	[6]
		ITS2	GCTGCGTTCTTCATCGATGC		
ITS	ITS2	ITS3_KYO2	GATGAAGAACGYAGYRAA	~381	[6]
		ITS4	TCCTCCGCTTATTGATATGC		

2. Bioinformatics analysis

2.1. Quality control and clustering

Method1: Illumina + Usearch

2.1.1. Reads filtering

Raw data containing adapters or low quality reads would affect the following assembly and analysis. Thus, to get high quality clean reads, raw reads were further filtered according to the following rules using FASTP[7] (version 0.18.0):

1) Removing reads containing more than 10% of unknown nucleotides (N);

2) Removing reads containing less than 50% of bases with quality (Q-value)＞20.

2.1.2. Reads assembly

Paired end clean reads were merged as raw tags using FLASH[8] (version 1.2.11) with a minimum overlap of 10 bp and mismatch error rates of 2%.

2.1.3 Raw tag filtering

Noisy sequences of raw tags were filtered under specific filtering conditions[39] to obtain the high-quality clean tags. The filtering conditions are as follows：

1) Break raw tags from the first low quality base site where the number of bases in the continuous low quality value (the default quality threshold is ≤3) reaches the set length (the default length is 3 bp)；

2) Then, filter tags whose continuous high-quality base length is less than 75% of the tag length.

2.1.4 Clustering and chimera removal

The clean tags were clustered into operational taxonomic units (OTUs) of ≥ 97 % similarity using UPARSE[11] (version 9.2.64) pipeline. All chimeric tags were removed using UCHIME algorithm [10] and finally obtained effective tags for further analysis. The tag sequence with highest abundance was selected as representative sequence within each cluster.

Method2: Illumina +DADA2

The DADA2 R package [17] (version 1.14) implements a complete pipeline to turn paired-end fastq

files from the sequencer into merged, denoised, chimera-free, inferred sample sequences. In datail：

2.1.1 Filtering

Raw reads containing primers or unknown nucleotides (N bases) would affect the following assembly and analysis. Thus, to get clean reads, raw reads were filtered and truncated according to the following rules:

1) Removing reads containing unknown nucleotides (N);

2) Removing primer sequences.

2.1.2 Dereplication and denoising

Then, a dereplicated list of unique sequences and their abundances were output, as well as the consensus positional quality scores for each unique sequence by taking the average (mean) of the positional qualities of the component reads. These consensus scores are used by the error model. Considering that each amplicon sequencing sample had different error ratio, DADA2 used machine learning to construct the error model for reads denoising, by alternately estimating the error rate and learning the error model from the reference sample sequence until the learning model converges to the true error rate.

2.1.3 Merging

Then paired end denoised reads were merged as raw ASVs (amplicon sequence variants) with a minimum overlap of 12bp.

2.1.4 Chimera removal

Chimera sequences are identified and deleted by UCHIME algorithm[10]. After chimera removal, the denoised, chimera-free ASV sequences and their abundances were output.

Method3: PacBio+Usearch

2.1.1 Reads filtering

Raw reads were assigned to samples based on their unique barcode and truncated by cutting off the barcode sequence using the lima application (version 2.0.1) in the Pbbioconda package (Pacific Biosciences), followed by analysis of subreads to generate CCS (Circular Consensus Sequencing) reads using PacBio's open-source software suite SMRT Link (version 7.0) with the parameters as follows: minfullpass = 3, minPredictedAccuacy = 0.99. Primer sequences were trimmed using

cutadapt[45] (version2.10). The files generated by the PacBio platform were then used for amplicon size filtering to remove sequences outside the expected amplicon size (minlength 1.3 kb, maxlength 1.7 kb). Reads with same continuous base number more than 8 were considered as low quality reads and were removed.

2.1.2 Clustering and chimera removal

The retained clean reads were clustered into operational taxonomic units (OTUs) of ≥ 97 % similarity using UPARSE[11] (version 9.2.64) pipeline. All chimeric reads were removed using UCHIME algorithm [10] and finally obtained effective reads for further analysis. The sequence with highest abundance was selected as representative sequence within each cluster.

Method4: PacBio+DADA2

The DADA2 R package [17] (version 1.14) implements a complete pipeline to turn fastq files into denoised, chimera-free, inferred sample sequences. In datail：

2.1.1 Filtering

Raw reads were assigned to samples based on their unique barcode and truncated by cutting off the barcode sequence using the lima application (version 2.0.1) in the Pbbioconda package (Pacific Biosciences), followed by analysis of subreads to generate CCS (Circular Consensus Sequencing) reads using PacBio's open-source software suite SMRT Link (version 7.0) with the parameters as follows: minfullpass = 3, minPredictedAccuacy = 0.99. Primer sequences were trimmed using cutadapt[45] (version2.10). The files generated by the PacBio platform were then used for amplicon size filtering to remove sequences outside the expected amplicon size (minlength 1.3 kb, maxlength 1.7 kb). Reads with same continuous base number more than 8 were considered as low quality reads and were removed. Removing reads containing unknown nucleotides (N);

2.1.2 Dereplication and denoising

Then, a dereplicated list of unique sequences and their abundances were output, as well as the consensus positional quality scores for each unique sequence by taking the average (mean) of the positional qualities of the component reads. These consensus scores are used by the error model. Considering that each amplicon sequencing sample had different error ratio, DADA2 used machine learning to construct the error model for reads denoising, by alternately estimating the error rate and

learning the error model from the reference sample sequence until the learning model converges to the true error rate.

2.1.3 Chimera removal

Chimera sequences are identified and deleted by UCHIME algorithm[10]. After chimera removal, the denoised, chimera-free ASV sequences and their abundances were output.

2.2. Taxonomy annotation

Illumina

The representative OTU sequences or ASV sequences were classified into organisms by a naive Bayesian model using RDP classifier[15] (version 2.2) based on SILVA database [16] (version 132) or UNITE database[18] (version 8.0) or ITS2 database[19] (version update_2015), with the confidence threshold value of 0.8.

PacBio

Taxonomic classification were conducted by BLAST (version 2.6.0)[46], searching the representative OTU sequences or ASV sequences against the NCBI 16S ribosomal RNA Database (Bacteria and archaea) (http://www.ncbi.nlm.nih.gov) (version 202101) using the best hit with strict criteria (E value < e-5, query coverage ≥ 60% and the following identity thresholds: a hit with sequence identity ≥92% was considered to belong to the same species; with sequence identity ≥88% as indicators of belonging to the same genus; with sequence identity ≥85% as indicators of the same family; with sequence identity ≥80% as indicators of the same order; the classes were inferred when sequence identity ≥75%; the phylum were inferred when sequence identity ≥70%). If no BLAST hit was retained, the sequence was labeled unclassified.

An empirical statistical model with a receiver operating characteristic (ROC) curve was employed to determine the optimal thresholds for taxonomic classifications using method in the reference article[47].

2.3. Community composition analysis

The abundance statistics of each taxonomy was visualized using Krona[20] (version 2.6). The stacked

bar plot of the community composition was visualized in R project ggplot2 package [21] (version 2.2.1). Circular layout representations of species abundance were graphed using circos [22] (version 0.69-3). Heatmap of species abundance was plotted using pheatmap package (version 1.0.12) [23] in R project. Pearson correlation analysis of species was calculated in R project psych package[37] (version 1.8.4). Network of correlation coefficient were generated using Omicsmart, a dynamic real-time interactive online platform for data analysis (http://www.omicsmart.com) or igraph package[38] (version 1.1.2) in R project.

2.4. Indicator species analysis

Between groups Venn analysis was performed in R project VennDiagram package [12] (version 1.6.16) and upset plot was performed in R project UpSetR package [13] (version 1.3.3) to identify unique and common species or OTUs or ASVs. Species comparison between groups was calculated by welch's t-test and wilcoxon rank test in R project Vegan package[14] (version 2.5.3). Species comparison among groups was computed by tukey's HSD test and kruskal-wallis H test in R project Vegan package[14] (version 2.5.3). Biomarker features in each group were screened by LEfSe software[24] (version 1.0), randomforest package [25] (version 4.6.12) in R project, pROC package [26] (version 1.10.0) in R project, and labdsv package[27] (version2.0-1) in R project. Ternary plot of species abundance was plotted using R ggtern package[28] (version 3.1.0).

2.5. Alpha diversity analysis

Chao1, ACE, Shannon, Simpson, Good's coverage, Pielou's evenness index were calculated in QIIME[9] (version 1.9.1). PD-whole tree index was calculated in picante[43] (version 1.8.2). OTU/ASV rarefaction curve and rank abundance curves were plotted in R project ggplot2 package [21] (version 2.2.1). Alpha index comparison between groups was calculated by Welch's t-test and Wilcoxon rank test in R project Vegan package[14] (version 2.5.3). Alpha index comparison among groups was computed by Tukey's HSD test and Kruskal-Wallis H test in R project Vegan package[14] (version 2.5.3).

2.6. Beta diversity analysis

Sequence alignment was performed using Muscle[29] (version 3.8.31) and phylogenetic tree was constructed using FastTree[30] (version 2.1), then weighted and unweighted unifrac distance matrix were generated by GuniFrac package [31] (version 1.0) in R project. Jaccard and bray-curtis distance matrix calculated in R project Vegan package[14] (version 2.5.3). PCA (principal component analysis) was performed in R project Vegan package [14] (version 2.5.3). Multivariate statistical techniques including PCoA (principal coordinates analysis) and NMDS (non-metric multi-dimensional scaling) of (Un) weighted unifrac, jaccard and bray-curtis distances were generated in R project Vegan package[14] (version 2.5.3) and plotted in R project ggplot2 package [21] (version 2.2.1). Statistic analysis of Welch's t-test, Wilcoxon rank test, Tukey's HSD test, Kruskal-Wallis H test, Adonis (also called Permanova) and Anosim test was calculated in R project Vegan package[14] (version 2.5.3).

2.7. Function prediction

The KEGG pathway analysis of the OTUs/ASV was inferred using Tax4Fun[32] (version 1.0) or PICRUSt[33] (version 2.1.4). Microbiome phenotypes of bacteria were classified using BugBase[34] . FAPROTAX database (Functional Annotation of Prokaryotic Taxa) and associated software[35] (version 1.0) were used for generating the ecological functional profiles of bacteria. The Functional group (guild) of the Fungi was inferred using FUNGuild[36] (version 1.0). Analysis of function difference between groups was calculated by Welch's t-test, Wilcoxon rank test and Kruskal-Wallis H test, Tukey's HSD test in R project Vegan package[14] (version 2.5.3).

2.8. Environmental factor analysis

Redundancy analysis (RDA), canonical correspondence analysis (CCA), Variation partition analysis (VPA), mantel test and envfit test were executed in R project Vegan package[14] (version 2.5.3) to clarify the influence of environmental factors on community composition. Pearson correlation coefficient between environmental factors and species was calculated in R project psych package[37] (version 1.8.4). Heatmap and network of correlation coefficient were generated using Omicsmart, a dynamic real-time interactive online platform for data analysis (http://www.omicsmart.com).

Reference

[1] Guo M, Wu F, Hao G, et al. Bacillus subtilis improves immunity and disease resistance in rabbits[J]. Frontiers in immunology, 2017, 8: 354.

[2] Fazzini R A B, Levican G, Parada P. Acidithiobacillus thiooxidans secretome containing a newly described lipoprotein Licanantase enhances chalcopyrite bioleaching rate[J]. Applied microbiology and biotechnology, 2011, 89(3): 771-780.

[3] Beckers B., Beeck M. O. D., Thijs S., et al. Performance of 16s rDNA Primer Pairs in the Study of Rhizosphere and Endosphere Bacterial Microbiomes in Metabarcoding Studies. Frontiers in Microbiology,2016

[4] Teske A, Sørensen K B. Uncultured archaea in deep marine subsurface sediments: have we caught them all?[J]. The ISME journal, 2008, 2(1): 3.

[5] Cheung M K, Au C H, Chu K H, et al. Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing[J]. The ISME journal, 2010, 4(8): 1053.

[6] Toju H, Tanabe A S, Yamamoto S, et al. High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples[J]. PloS one, 2012, 7(7): e40863.

[7] Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor[J]. bioRxiv, 2018: 274100.

[8] Magoč T, Salzberg S L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27.21 (2011): 2957-2963.

[9] Caporaso, J. Gregory, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7.5 (2010): 335-336.

[10] Edgar, Robert C., et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27.16 (2011): 2194-2200.

[11] Edgar, Robert C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10.10 (2013): 996-998.

[12] Chen H, Boutros P C. VennDiagram: a package for the generation of highly-customizable Venn

and Euler diagrams in R[J]. BMC bioinformatics, 2011, 12(1): 35.

[13] Conway J R, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties[J]. Bioinformatics, 2017, 33(18): 2938-2940.

[14] Oksanen J, Blanchet F G, Kindt R, et al. Vegan: community ecology package. R package version 1.17-4[J]. http://cran. r-project. org>. Acesso em, 2010, 23: 2010.

[15] Wang, Qiong, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 73.16 (2007): 5261-5267.

[16] Pruesse, Elmar, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic acids research 35.21 (2007): 7188-7196.

[17] Callahan B J, Mcmurdie P J, Rosen M J, et al. DADA2: High-resolution sample inference from Illumina amplicon data[J]. Nature Methods, 2016, 13(7): 581-583.

[18] Nilsson R H, Larsson K H, Taylor A F S, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications[J]. Nucleic acids research, 2018, 47(D1): D259-D264.

[19] Ankenbrand M J, Keller A, Wolf M, et al. ITS2 database V: Twice as much[J]. Molecular Biology and Evolution, 2015, 32(11): 3030-3032.

[20] Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. Interactive metagenomic visualization in a Web browser. BMC bioinformatics 12.1 (2011): 385.

[21] Wickham, H. (2011). ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics, 3(2), 180–185. doi:10.1002/wics.147

[22] Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics[J]. Genome research, 2009, 19(9): 1639-1645.

[23] Kolde R, Kolde M R. Package 'pheatmap'[J]. R Package, 2015, 1(7).

[24] Segata, Nicola, et al. "Metagenomic biomarker discovery and explanation." Genome biology 12.6 (2011): 1.

[25] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[26] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves[J]. BMC bioinformatics, 2011, 12(1): 77.

[27] Roberts D W, Roberts M D W. Package 'labdsv'[J]. Ordination and Multivariate, 2016.

[28] Hamilton N E, Ferry M. ggtern: Ternary diagrams using ggplot2[J]. Journal of Statistical Software, 2018, 87(1): 1-17.

[29] Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput Nucleic Acids Res. 32(5):1792-1797.

[30] Price M N, Dehal P S, Arkin A P. FastTree 2–approximately maximum-likelihood trees for large alignments[J]. PloS one, 2010, 5(3): e9490.

[31] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities[J]. Appl. Environ. Microbiol., 2005, 71(12): 8228-8235.

[32] Aßhauer, Kathrin P., et al. "Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data." Bioinformatics 31.17 (2015): 2882-2884.

[33] Langille, Morgan GI, et al. "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." Nature biotechnology 31.9 (2013): 814-821.

[34] Ward T, Larson J, Meulemans J, et al. BugBase predicts organism level microbiome phenotypes[J]. BioRxiv, 2017: 133462.

[35] Louca S, Parfrey L W, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome[J]. Science, 2016, 353(6305): 1272-1277.

[36] Nguyen, Nhu H., et al. "FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild." Fungal Ecology 20 (2016): 241-248.

[37] Revelle W, Revelle M W. Package 'psych'[J]. The Comprehensive R Archive Network, 2015.

[38] Csardi M G. Package 'igraph'[J]. Last accessed, 2013, 3(09): 2013.

[39] Bokulich, Nicholas A., et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods 10.1 (2013): 57-59.

[40] Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field

samples. Environmental Microbiology, 18(5), 1403–1414.

[41] Apprill, A., McNally, S., Parsons, R., & Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquatic Microbial Ecology, 75(2), 129–137.

[42] Scibetta S, Schena L, Abdelfattah A, et al. Selection and Experimental Evaluation of Universal Primers to Study the Fungal Microbiome of Higher Plants[J]. Phytobiomes Journal, 2018, 2(4): 225-236.

[43] Kembel SW, Cowan PD, Helmus MR, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26(11):1463-1464.

[44] Singer, E., Bushnell, B., Coleman-Derr, D. et al. High-resolution phylogenetic microbial community profiling. ISME J 10, 2020–2032 (2016). https://doi.org/10.1038/ismej.2015.249

[45] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17:10. doi: 10.14806/ej.17.1.200

[46] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 5: 403-10.

[47] Gaby JC, Rishishwar L, Valderrama Aguirre LC, Green SJ, Valderrama-Aguirre A, Jordan IK, Kostka JE. 2018. Diazotroph community characterization via a highthroughput nifH amplicon sequencing and analysis pipeline. Appl Environ Microbiol 84:e01512-17.