基于 PacBio SMRT 三代测序的红树林沉积物真菌群落的研究

Analysis of Fungal Community in Mangrove Sediments Based on PacBio SMRT Sequencing

张志锋[1]，李猛[1]*

[1]高等研究院，深圳大学，深圳，广东

*通讯作者邮箱：limeng848@szu.edu.cn

摘要：虽然二代测序可以在短时间内产生大量高质量的数据，但由于读长原因，测序结果并不能准确的鉴定到种水平，因而在环境微生物群落的鉴定上仍存在一定的局限性。以 circular consensus sequencing (CCS) 技术为基础的 PacBio SMRT 的三代测序技术，可以产生长度可达十至数十 kb 的高质量 DNA 数据，能够完整的覆盖细菌 16S rDNA，真菌 18S/28S rDNA 和 ITS 区域，甚至 18S rDNA+ITS+28S rDNA 区域全长，可以有效的解决注释精度的问题。在本研究中，通过红树林沉积物样品采集，DNA 提取，PCR 扩增，PacBio SMRT 测序和数据分析，最终获得高注释精度的真菌群落 OTU table。以此为基础，通过后续的生态学分析，对红树林真菌群落的多样性、组成、分布规律、影响因素、群落组装过程、物种相互作用关系等方面有深刻认识。分析方法部分同样适用于其他环境微生物群落 PacBio SMRT 三代扩增子测序数据的分析。

关键词：PacBio SMRT，真菌群落，扩增子测序

材料与试剂

1. DNeasy PowerSoil Kit (Qiagen, 12888-50/12888-100) 或 FastDNA Spin Kit for soil (MP Biomedicals, 116560)
2. PrimeSTAR GXL DNA Polymerase (Takara, R050) 或 Phusion High-Fidelity DNA Polymerase (Thermo Scientific, F537L)
3. 无菌超纯水

仪器设备

1. 沉积物采样器
2. 涡旋振荡器 (带适配器) /组织研磨仪 (MP Biomedicals, MP Fastprep-24 5G)
3. NanoDrop ND-2000c UV-Vis spectrophotometer (NanoDrop Technologies)
4. ProFlex™ PCR 仪 (ThermoFisher)

软件和数据库【可选】

1. 软件

pbccs (v4.02, https://github.com/PacificBiosciences/ccs)
BAM2fastx (https://github.com/pacificbiosciences/bam2fastx/)
lima (v1.11.0, https://github.com/pacificbiosciences/barcoding/)
flexbar (v3.0, https://github.com/seqan/flexbar)
mothur (v1.44.2, https://github.com/mothur/mothur/releases/tag/v1.44.2)
vsearch (v2.15.0, https://github.com/torognes/vsearch/releases/)
ITSx (v1.0.11, https://microbiology.se/software/itsx/)
usearch (v10.0.240, https://drive5.com/usearch/)
BLAST+ (v2.10.1, https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)
RAxML (v8.2.12, https://github.com/stamatak/standard-RAxML)
IQTree (v2.1.1, http://www.iqtree.org)
FastTree (v2.1.11, http://www.microbesonline.org/fasttree/)

2. 数据库

UNITE (v8.2, https://doi.org/10.15156/BIO/786372) for ITS
silva (release 138, https://www.arb-silva.de/documentation/) for 16S/18S rDNA
UCHIME reference dataset (v7.2, https://unite.ut.ee/repository.php)

实验步骤

1. 样品采集与处理

1.1 样品采集

根据实验目的设计合理的实验方案，选择要采集的红树林，确定采样位置。红树林样品的采集使用特制的沉积物样品采样器，沉积物样品的采集使用三点或五点取样法。对每个样点的 3-5 个重复样品进行等量混合以减少采样偏差。根据实

验需要可对样品按深度分层，使用无菌自封袋保存。样品需使用冰袋或干冰冷藏运输，实验室内存放于-40/-80 °C 冰箱。

1.2 DNA 提取

参考所使用试剂盒说明书进行。若使用 DNeasy PowerSoil Kit，可在机械破碎后增加 60 °C 水浴 30 min，可提高 DNA 产量。使用 NanoDrop 确定 DNA 质量和浓度，并 1%琼脂糖凝胶电泳检测。质量不好的 DNA 应重提或使用 DNA 纯化试剂盒进行纯化。

2. PCR 扩增与 PacBio SMRT 测序

2.1 Primers 选择

选择合适的 Primer 是研究微生物群落的最重要步骤。对于真菌群落的研究通常选用 ITS 区域。对于 ITS 全长的扩增，建议优先选用对真菌群落具有极高覆盖度的引物 ITS9Munngs/ITS4ngs (Tedersoo and Lindahl, 2016; Nilsson, *et al.*, 2019) 。若上述引物扩增效果较差，可根据情况选用 ITS1Fngs/ITS4ngs (White *et al.*, 1990; Tedersoo *et al.*, 2015) 或 ITS1F/ITS4 (White *et al.*, 1990; Gardes and Bruns, 1993) 。根据后续测序过程中的混样方案，在正反向引物上下游添加特异 barcode 信息，barcode 长度不小于 6 bp，以保证测序后数据的正确拆分。

2.2 PCR 扩增

PCR 扩增中嵌合体的形成与循环数，聚合酶的选择和初始模板质量有很大关系，特别是长片段扩增中更容易出现嵌合体。建议选择高活性高保真度的 DNA 聚合酶，延长延伸时间 (Tedersoo *et al.*, 2015) 。虽然循环数的增加容易导致嵌合体产生，但扩增效率也会随着片段长度的增加而下降，因而循环数的选择一定要慎重。ITS 片段扩增的循环数可以设置为 30-32 个，不超过 35 个，当片段长度增加时可以适当增加循环数 (Tedersoo *et al.*, 2015; Nilsson *et al.*, 2019) 。PCR 扩增前将 DNA 模板稀释到 5-10 ng/μl，以保证扩增过程中模板的一致性。由于三代测序所需 DNA 量较大，PCR 扩增体系可选择 30 μl，其中包含 1.5U polymerase，3 μl buffer，150 μM dNTPs，正反向引物各 0.12 μM，2-10 ng DNA 模板。PCR 扩增程序为，94 °C 5 min, 94 °C 30 s, 57 °C 30 s, 72 °C 1 min 20 s, 72 °C 10 min，其中 2-4 步为 32 循环。PCR 产物使用 Na

noDrop 和 1%琼脂糖凝胶电泳检测。每个样品至少设置 3 个重复，并将其等量混合，在以减少 PCR 偏差。同时每批次 PCR 都应设置空白对照。

2.3 PacBio SMRT 测序

此步骤主要由测序公司完成，简单步骤如下：将加有 barcode 的 PCR 产物按照预先设计的混样方案等量混合，然后将发卡测序接头连接到 PCR 文库并完成环化。使用 Enzyme Clean Up Kit 对测序文库进行纯化。将测序引物退火结合至 PCR 产物文库，并将 DNA 聚合酶结合测序模板。测序使用 PacBio Sequal 平台进行。

3. 数据分析

3.1 CCS (circular consensus sequencing) reads

PacBio SMRT 通过对环化连接的插入测序片段循环进行多次测序后比对校正纠错，获得高精度的保守序列 (CCS reads) 。当循环测序数达到 5 次时，CCS reads 的质量理论上可以达到 QC40，也就是错误率 0.01%。PacBio 平台产生数据格式为.bam，使用 pbccs 软件进行环化校正得到 CCS reads，命令为 ccs cell01.subreads.bam cell01.ccsreads.bam --minPasses 5 --reportFile ccs_report.txt。原始文件为 cell01.subreads.bam，输出文件为 cell01.ccsreads.bam，--minPasses 为循环数，--reportFile 为统计结果文件。

3.2 数据拆分 (demultiplex)

准备 barcode 文件 (Barcode.fasta) ，根据 barcode 信息，使用 lima 软件进行样品拆分：lima --ccs cell01.ccsreads.bam Barcodes.fasta split.bam --same --split-bam --split-bam-named -j 100。其中 split.bam 文件为输出文件，输出文件将以 split.xxx.bam 命名，--same 表示双端引物相同，--split-bam 表示按照 barcode pairs 拆分 bam 文件，--split-bam-named 表示按照 barcode 名称对拆分后输出的 bam 文件命名，-j 线程数。另外可先进行步骤 3.3 bam 转 fastq，将 cell01.ccsreads.bam 文件转换为 fastq 文件后，根据 barcode 序列，使用 flexbar 软件 (Dodt *et al.*, 2012) 进行拆分，命令为 flexbar -b Barcodes.fasta -r cell01.ccsreads.fastq -t cell01.ccsreads -bt ANY -be 0.1，-b 为 barcode 序列文件，-r 为需要拆分的 fastq 文件，-t 为输出文件前缀，-bt 为--barcode-trim-end，ANY 表示删除 barcode 两端序列，-be 为--barcode-error-rate。

3.3 bam 转 fastq

使用 BAM2fastx 软件进行。bam2fastq -o sample1 sample1.ccsreads.bam -u。-o 为输出文件前缀，sample1.ccsreads.bam 为输入 bam 文件，-u 表示输出文件不压缩。

3.4 质控过滤

使用 mothur (Schloss *et al.*, 2009) 进行质控过滤，首先将 fastq 拆分为序列文件及其对应的质量分数文件 fastq.info(fastq=sample1.fastq)，随后进行质控 trim.seqs(fasta=sample1.fasta,minlength=100,maxambig=0,maxhomop=12,qfile=sample1.qual,qwindowsize=50,qwindowaverage=20)。

3.5 文件及序列重命名

批量修改文件名以后使用 usearch (Edgar., 2010) 对序列按照文件名进行重命名，usearch11 -fastx_relabel sample1.fasta -prefix sample1- -fastaout sample1_relabel.fasta -keep_annots。-prefix 为重命名序列前缀，-fastaout 为输出文件名。

3.6 提取 ITS 序列

使用 ITSx (Bengtsson-Palme *et al.*, 2013) 进行提取。命令为 ITSx -i sample1_relabel.fasta -o sample1_out --cpu 4 --save_regions all --preserve T -E 1e-2。-i 输入文件名，-o 输出文件名，--cpu 核心数，--save_region 要保留的片段，all 表示保留所有片段，包括 18S，28S，ITS 全长，ITS1，5.8S 和 ITS 2，--preserve 序列名为原始序列名，-E，e-value。此步骤产生的 sample1_out.full.fasta 文件为 ITS 全长序列，用于后续分析。

3.7 嵌合体检测与删除

使用 vsearch (Rognes *et al.*,2016) 进行重头嵌合体检测 uchime_denovo (Edgar *et al.*, 2011) 和基于数据库的嵌合体检测 uchime_ref。uchime_denovo 命令：vsearch --uchime_denovo sample1_out.full.fasta --chimeras sample1_out.full_chimeras.fasta --nonchimeras sample1_out.full_nonchimeras.fasta --relabel_keep；uchime_ref 命令：vsearch --uchime_ref sample1_out.full_nonchimeras.fasta --chimeras sample1_out.ful_chimeras2.fasta --nonchimeras sample1_out.full_nonchimeras2.fasta --db uchime_reference_dataset_28.0

6.2017.fasta --relabel_keep --threads 25。重命名文件 sample1_out.full_nonchimeras2.fasta 为 sample1.fasta 用于后续分析。

3.8 长度筛选

绝大部分真菌 ITS 片段的长 300-900 bp，极少数担子菌可达到 1,100 bp (Schoch *et al.*, 2014; Nilsson *et al.*, 2019) 。为避免过长或过短序列干扰，使用 vsearch 进行片段长度筛选。vsearch --fastx_filter sample1.fasta --fastaout sample1.remian.fasta --fastaout_discarded sample1.discarded.fasta --fastq_maxlen 900 --fastq_minlen 300。

3.9 OTU 生成与代表序列挑选

主要使用 usearch，但由于免费版 32 位 usearch 限制，在处理大数据量时结合 vsearch 共同使用。此步骤可以分为以下步骤：

1) 计数与去重

usearch -fastx_uniques sample1.remian.fasta -fastaout uniques.sample1.fasta -sizeout

2) 去除单条序列

usearch -sortbysize uniques.sample1.fasta -fastaout desingl.uniques.sample1.fasta -minsize 2。-minsize 表示保留序列的最小条数。

注：以上两步可以在 vsearch 中合并为一步，vsearch --derep_fulllength sample1.fasta --sizein --fasta_width 0 --sizeout --output desingl.unique.sample1.fasta --minuniquesize 2 --threads 8。

3) OTU 聚类

usearch10 -cluster_otus desingl.uniques.sample1.fasta -otus sample1.otu.fasta -uparseout sample1.otu.txt -relabel OTU -minsize 2。-otus sample1.otu.fasta 输出即为 OTU 代表序列，usearch10 默认选择丰度最高的序列为代表序列。usearch10 默认使用 UPARSE 算法 (Edgar., 2013) 进行 OTU 聚类，序列相似度阈值为 97%，不能更改。若想更改相似度阈值，可选择 usearch10 以前的早期版本，通过-id 0.97 参数进行修改。常用的 OTU 聚类方法还有 CD-HIT 等方法 (Fu *et al.*, 2012) ，此处不再赘述。

4) OTU table 生成

vsearch --usearch_global sample1.fasta --db sample1.otu.fasta --id 0.97 --otutabout sample1.otutable.txt --threads 50。以步骤 3.9.3 产生的 OTU 代表序列的数据库 (-db) 对样品数据以 97%相似度 (-id) 为阈值进行聚类生成 OTU table。

3.10 代表序列注释

使用软件为 BLAST+，真菌 ITS 使用 UNITE 数据库。有研究表明早期的 UNITE 数据库中有许多序列都是错误鉴定的，将一些非真菌生物鉴定为真菌，主要为 Rozellomycota 和一些未知真菌。最新版的 UNITE 数据库分为真菌和真核两种，真菌数据库主要为真菌序列，真核数据库主要为真核序列，真核数据库相比于真菌数据库更加全面准确，因而建议使用真核数据库。构建数据库：makeblastdb -in UNITE_eukaryotes_all_04.02.2020.fasta -dbtype nucl -out unite_eukaryotes。-dbtype nucl 表示数据库类型为核酸序列。注释：blastn -max_target_seqs 10 -db unite_eukaryotes -out otu.rep.seq.euk.blast -query sample1.otu.fasta -num_threads 10 -outfmt "6 qseqid qlen qstart qend salltitles sseqid slen sstart send qcovs bitscore evalue pident"。-max_target_seqs 输出比对结果数量，-out 输出比对结果文建，-query 输入代表序列文件，-num_threads 核心数，-outfmt 输出文件格式，6 表示表格格式，后面内容为比对结果信息。为使注释结果更加可靠，采用 Tedersoo 等人 (2015, 2018) 使用的注释策略，对于注释在真菌界中的 OTU，以 90%、85%、80%和 75%的序列相似度分别作为属、科、目和纲的区分标准。

3.11 系统发育注释 (可选)

相比于二代测序而言，三代测序获得的更长的微生物 maker 基因序列可以获得更为精确的注释结果。但是由于人们目前对自然界微生物认识有限，现有数据库中许多微生物并未精确注释，如 UNITE 真核生物数据库中有许多序列被注释为“Eukaryota_kgd_Incertae_sedis”，同时有许多序列仅通过数十至上百 bp 的的 alignment 进行了注释，结果不够准确，而且有许多未知微生物的序列并未被包括在数据库中。因而我们提出了基于 blastn 注释结果的“基于系统发育分析的微生物鉴定”，以更准确的对未知真菌进行注释。简单来说就是使用 3.93 中获得 OTU 代表序列 (或去除其它真核序列的真菌 OTU 代表序列) 构建系统发育树。常用

构建系统发育树的方法有 Neighbor Joining (NJ)，Maximum Parsimony (MP)，Maximum Likelihood (ML) 和 Bayesian inference (BI)，几种方法各有优势。此处推荐使用 ML 方法构建系统发育树，推荐软件有 FastTree (号称速度最快的 ML 树构建软件，一般来简单观察系统发育关系，数据量特别大时使用) (Nguyen et al. 2015)，IQTree (一种精确快速的 ML 系统发育分析工具，bootstrap 计算速度是 RAxML 的 10-40 倍，大数据量时强烈推荐) (Price et al. 2010)，和 RAxML (最常用的系统发育树构建软件之一，支持多线程和向量指令运行，运行速度较快，数据量不是特别大时推荐使用) (Stamatakis 2014)，请根据情况酌情选用。序列比对使用 MUSCLE (Edgar 2004)：muscle -in input.ITS.fas -out aligned.input.ITS.muscle.fas；其中-in 为输如 fasta 序列；-out 为输出 alignment 文件。序列修剪使用 trimAl (Capella-Gutierrez et al. 2009)：trimal -in aligned.input.ITS.muscle.fas -out aligned.input.ITS.muscle.trim.phy -gt 0.1；-in 为输如 fasta 序列；-out 为输出 alignment 文件；-gt 序列中允许出现 gap 的部分。IQTree 构建系统发育树：iqtree -s aligned.input.ITS.muscle.trim.phy -m TESTONLY -nt 60 -bb 1000 -alrt 1000；-s 为输入 alignment 文件；-nt 为核心数，也可设为 AUTO 系统自动分配；-bb (ultrafast bootstrap approximation)重复抽样次数，默认 1000，大数据建议-bb，小数据可用-b；-alrt 是否启用 SH-aLRT 检验，可删除；-m，model，不提供时 iq-tree 自动选择；-o 可指定外群序列。FastTree 构建系统发育树：FastTree aligned.input.ITS.muscle.trim.phy > aligned.input.ITS.muscle.trim.tree；数据量特别大时使用。RAxML 构建系统发育树：raxmlHPC-PTHREADS-SSE3 -T 40 -f a -x 12345 -# 1000 -m GTRGAMMA -s ./BAE61387_aligned_sequences.fas.phy -n tre；若 CPU 支持也可选用 raxmlHPC-PTHREADS-AVX 或 raxmlHPC-PTHREADS-AVX2，可以极大提高运行速度；-T 线程数；-f 选择 RAxML 算法，a 为快速 bootstrap 分析；-x 随机数；-# bootstrap；-m，model，DNA 序列常用为 GTRGAMMA；-s 输入文件名；-n 输出文件后缀。构建完成系统发育树后，根据 OTU 所在的系统发育分枝对 OTU 进行相应的初步注释。

4. 数据统计与分析

基于步骤 3 所获得的代表性序列和 OTU table 对真菌群落的 α 多样性、β 多样性、生物地理学特征、分布与群落结构的影响因素，群落结构的组装过程、共现性关系等内容进行分析与可视化展示，此部分内容繁多，本文中不再赘述。

致谢

本工作由科技部基础资源调查专项 (2019FY100700) 、国家自然科学基金 (91851105、31970105) 及中国博士后科学基金 (2020M672779) 资助。本文分析方法已应用于待发表文章“Pacific Biosciences single-molecule real-time (SMRT) sequencing reveals high diversity of basal fungal lineages and stochastic processes controlled fungal community assembly in mangrove sediments”。

参考文献

1. Bengtssonpalme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., ... and Nilsson, R. H. (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol Evol* 4: 914–919.
2. Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; 25:1972-1973.
3. Dodt M, Roehr J, Ahmed R. Dieterich C. (2012) FLEXBAR—Flexible barcode and adapter processing for next-generation sequencing platforms. *Biology* 1: 895–905.
4. Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; 32:1792-1797.
5. Edgar, R.C. (2010), Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460-2461.
6. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27: 2194–2200.
7. Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10: 996–998.

8. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 23: 3150–3152.
9. Gardes, M., Bruns, T. D. (1993) ITS primers with enhanced specificity for basidiomycetes, application to the identification of mycorrihiza and rusts. *Mol Ecol* 2: 113–8.
10. Nguyen L, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 2015; 32:268-274.
11. Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P. and Tedersoo, L. (2019). Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat Rev Microbiol* 17: 95-109.
12. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490. doi:10.1371/journal.pone.0009490.
13. Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584.
14. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... and Weber, C. F. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537–7541.
15. Schoch, C. L., Robbertse, B., Robert, V., Vu, D., Cardinali, G., Irinyi, L., ... and Kirk, P. M. (2014). Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database* 2014: 1-21.
16. Stamatakis, A., (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
17. Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., ... and Bork, P. (2015). Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* 10: 1-43.
18. Tedersoo, L. and Lindahl, B. (2016). Fungal identification biases in microbiome projects. *Env Microbiol Rep* 8: 774-779.
19. Tedersoo, L., Tooming-Klunderud, A., Anslan, S. (2018) PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytol* 217: 1370–1385.

20. White, T. J., Bruns, T. D., Lee, S. B. and Taylor, J. W. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJWT, editors. PCR protocols: a guide to methods and applications. New York: Academic Press, p. 315–22.