**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](mailto: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/>)

1. 数据库

UNITE (v8.2, <https://doi.org/10.15156/BIO/786372>) for ITS

silva (release 138, [https://www.arb-silva.de/documentation/](https://www.arb-silva.de/documentation/release-138/)) 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
2. 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. 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。

1. PCR扩增与PacBio SMRT测序
2. 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。

1. 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NanoDrop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。

1. 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。

1. 数据分析
2. 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。

1. 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。

1. 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。

1. 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)。

1. 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。

1. 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ITS2，--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。

1. 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\_28.06.2017.fasta --relabel\_keep --threads 25。重命名文件sample1\_out.full\_nonchimeras2.fasta为sample1.fasta用于后续分析。

1. 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。

1. 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

1. 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。

1. 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。

1. 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。

* 1. 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。

* 1. 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。

1. 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.](https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12073) *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.](http://www.ncbi.nlm.nih.gov/pubmed/19505945) *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.](https://www.mdpi.com/2079-7737/1/3/895/htm) *Biology* 1: 895–905.
4. Edgar, R.C., [MUSCLE: multiple sequence alignment with high accuracy and high throughput.](http://www.ncbi.nlm.nih.gov/pubmed/15034147) *Nucleic Acids Res* 2004; 32:1792-1797.
5. Edgar, R.C. (2010), [Search and clustering orders of magnitude faster than BLAST.](https://academic.oup.com/bioinformatics/article-abstract/26/19/2460/230188) *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.](http://www.ncbi.nlm.nih.gov/pubmed/21700674) *Bioinformatics* 27: 2194–2200.
7. Edgar, R.C. (2013) [UPARSE: highly accurate OTU sequences from microbial amplicon reads.](http://www.ncbi.nlm.nih.gov/pubmed/23955772) *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.](http://www.ncbi.nlm.nih.gov/pubmed/23060610) *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.](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.1993.tb00005.x) *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.](http://www.ncbi.nlm.nih.gov/pubmed/25371430) *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.](http://www.ncbi.nlm.nih.gov/pubmed/30442909) *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.](http://www.ncbi.nlm.nih.gov/pubmed/20224823) *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.](http://www.ncbi.nlm.nih.gov/pubmed/27781170) *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.](http://www.ncbi.nlm.nih.gov/pubmed/19801464) *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.](https://academic.oup.com/database/article-abstract/doi/10.1093/database/bau061/2634542) *Database* 2014: 1-21.
16. Stamatakis, A., (2014) [RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.](http://www.ncbi.nlm.nih.gov/pubmed/24451623) *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.](https://mycokeys.pensoft.net/article/4852/download/pdf/) *MycoKeys* 10: 1-43.
18. Tedersoo, L. and Lindahl, B. (2016). [Fungal identification biases in microbiome projects.](http://www.ncbi.nlm.nih.gov/pubmed/27348848) *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.](http://www.ncbi.nlm.nih.gov/pubmed/28906012) *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.](https://msafungi.org/wp-content/uploads/2019/03/February-2013-Inoculum.pdf) In: Innis MA, Gelfand DH, Sninsky JJWT, editors. PCR protocols: a guide to methods and applications. New York: Academic Press, p. 315–22.