bio-101 www.bio-protocol.org/exxxx DOI:10.21769/BioProtoc.xxxx

土壤宏转录组学样本前处理与数据分析

Sample Pretreatment And Data Analysis Of Soil Metatranscriptome

张丽燕 [1, 2]，连郑汉 [3]，褚海燕 [1, 2, *]

[1] 中国科学院南京土壤研究所，土壤与农业可持续发展国家重点实验室，江苏，南京，210008

[2] 中国科学院大学，北京，100049

[3] 广东美格基因科技有限公司，广州，510000

*通讯作者邮箱：hychu@issas.ac.cn

摘要：土壤宏转录组学是通过制备土壤 RNA 样本、RNA 测序、以及利用一系列生物信息学方法和平台搭建来完成土壤微生物组的转录过程分析，提供关于基因表达和土壤微生物组功能活性，从而获得微生物组关键代谢差异表征等信息的一门学科。关键点是针对土壤 RNA 样本，表征特定条件下执行各个代谢过程的微生物活性特征，极大地规避了因高通量 DNA 测序带来无法准确反映土壤微生物代谢活性的缺陷。**目的：**本实验以两种类型（酸性和碱性）湿地土壤样本为例，详述了利用市售 RNA 提取试剂盒进行的土壤宏转录组样本制备流程，为准确评价土壤 RNA 样本制备提供参考，同时给出了宏转录组数据分析流程，为从 RNA 水平分析土壤微生物表达活性提供思路。

关键词：土壤，宏转录组学，RNA，土壤微生物代谢活性

材料与试剂

1. 50 ml 离心管（Thermo Fisher Scientific, USA）
2. 酚氯仿异戊醇（配比 25：24：1, pH=8）
3. 琼脂糖
4. 电泳缓冲液（TAE）
5. RNA 提取试剂盒 RNA Mini Kit（Qiagen, Hilden, Germany)

仪器设备

1. 超微量紫外分光光度计 NanoDrop One（Thermo Fisher Scientific, MA, USA)
2. 台式高速冷冻离心机（Techcomp（Holdings), model: CT15RT）
3. 涡旋仪（Qiangen, catalog number: 13000-V1-15）

4. 水浴锅
5. 移液枪
6. 电泳仪

数据分析软件及平台

1. CD-HIT（http://weizhongli-lab.org/cd-hit/）
2. fastp（https://github.com/OpenGene/fastp）
3. SortMeRNA（https://github.com/biocore/sortmerna）
4. idba_tran（https://github.com/loneknightpy/idba）
5. prodigal（https://github.com/hyattpd/Prodigal）
6. CD-HIT（http://weizhongli-lab.org/cd-hit/）
7. RSEM（https://github.com/deweylab/RSEM）
8. bowtie2（http://bowtie-bio.sourceforge.net/bowtie2/index.shtml）
9. diamond 软件（http://www.diamondsearch.org）
10. 上述软件均在 Linux 操作系统下进行

土壤总 RNA 提取步骤

1. 取样过程

采集 0-20 cm 湿地土壤样品 5 克于 50 ml 的无菌离心管中，加入 RNA 酶抑制剂约 10 ml 使其浸没土壤样品并封存。低温运输并尽快于-80 °C 保存。

2. 注意事项

确保实验工作区无 RNase 污染并且整个操作过程戴橡胶手套。

3. RNA 制备

1). 加 2 g 土壤到 15 ml 磁珠管中（试剂盒提供）。

2). 依次向离心管内加入 2.5 ml Bead solution 溶液，0.25 ml SR1 溶液和 0.8 ml 的 IRS 溶液并漩涡混匀。

发生的反应：Bead Solution 是一种缓冲液可用来打散细胞和土壤颗粒；SR1 能帮助细胞裂解，可以破坏脂肪酸和几种微生物细胞膜相关的脂类；IRS 可以帮助去除腐殖质、细胞碎片和蛋白质等杂质。

3). 向磁珠管加入 3.5 ml 酚氯仿异戊醇溶液（pH=8），漩涡混匀直到分层消失。

4). 最大转速涡旋混匀 15 min。

发生反应：从 1 到 4 步的化学试剂和漩涡使细胞裂解，酚/氯仿/异戊醇使其最大程度裂解，溶解的细胞和试剂混合在一起，蛋白质降解只剩下核酸在溶液中。

5). 2,500 × g 离心 10 min。

发生反应：离心导致混合样品相分离。离心后能观察到三相，底下的有机相包括蛋白质和细胞碎片，中间相包括腐殖质和其他有机及无机物质，上层相包括所有的核酸。

6). 小心转移上层水相于一新 15 ml 离心管中（试剂盒提供）。

注：用枪头小心吸取上层水相，误触碰界面。

7). 加 1.5 ml SR3 到水相中，漩涡混合。4 °C 孵育 10 min， 2,500 × g 离心 10 min。

8). 将上清液转移到一新 15 ml 离心管中（不要碰到下面的沉淀物），加入 5 ml SR4 混匀，室温放置 30 min。

注：分层明显的界面处小心吸取上清，勿刺破（触碰）界面。

9). 2,500 × g 离心 30 min。

10). 倒出上清液，将离心管倒置在纸巾上 5 min。

注：依据土壤类型的不同，沉淀可能较大或颜色较深（Mettel, et al., 2010）。

11). 摇晃 SR5 溶液使其混合，加 1 ml SR5 溶液到离心管中，使沉淀再完全悬浮。

注：沉淀可能由于土壤样品的不同不易悬浮，可能需要将离心管放到 45 °C 的水浴池中 10 min 再悬浮，再漩涡混合，重复这样直到沉淀重悬浮。

12). 为每个 RNA 样品准备一个捕集柱。

12.1). 将捕集柱悬挂到离心管上。

12.2). 加 2 ml SR5 溶液到捕集柱上，使其重力流。允许 SR5 溶液完全流过捕集柱。

注：在加 RNA 样品前不要让捕集柱流干。

13). 将 11 步的 RNA 分离样加到捕集柱中，使其流过捕集柱。

14). 用 1 ml SR5 溶液洗涤捕集柱，流出液收集在 15 ml 离心管中。

反应：样品加到捕集柱上，核酸结合到柱基质上。捕集柱用 SR5 溶液洗涤确保未结合的污染物被去除掉。

15). 将捕集柱转移到一新 15 ml 离心管中，摇晃 SR6，然后加 1 ml SR6 溶液到捕集柱中使其流过捕集柱，洗提 RNA。

发生反应：SR6 溶液 RNA 洗提缓冲液是专有的盐溶液，它能使 RNA 流出而 DNA、剩下的细胞碎片和抑制剂依然留在捕集柱上。

16). 将洗提的 RNA 转移到 2.2 ml 离心管中，并加 1 ml SR4，至少倒置混合一次，-20 °C 静置 10 min。

17). 13,000 × g 离心 15 min。

18). 移去上清液，将 RNA 离心管倒置在纸巾上 10 min，风干颗粒物。

19). 加 100 µl SR7 溶液使 RNA 颗粒再悬浮。

注意事项

为防止 DNA 交叉污染，DNA 污染物的去除很重要，纯化的 RNA 应该直接用 PCR 检测。琼脂糖电泳缺乏检测复制片段表明缺乏检测到的交叉污染 DNA。如果检测到 DNA，需要进一步使用 DNase I 分离 RNA。

土壤 RNA 样本提取效果评价

实验借助市售土壤 RNA 提取试剂盒（RNA Mini Kit, Qiagen, Germany）比较两种不同类型（酸性、碱性）的湿地土壤样本总 RNA 提取效果，纯度和浓度测试结果如表 1 所示。OD260 代表核酸的吸光度，OD280 代表蛋白质的吸光度，OD230 代表其他杂质（多糖等）的吸光度。OD 是光密度值。一般来说，OD260/OD280 介于 1.9~2.0 说明 RNA 提取纯度高，污染小。OD260/OD280 < 1.7 时表明有蛋白质或酚污染；OD260/OD280 > 2.0 时表明可能有异硫氰酸残存。原核生物的核糖体 rRNA 主要由 23S、16S 和 5S rRNA 组成。实验室主要通过观察核糖体的 23S rRNA 和 16S rRNA 条带的亮度和片段形状来判定 RNA 的提取效果（Peano *et al.*, 2013），本实验中根据 OD260/OD280 介于 1.9~2.0 之间，16S rRNA 及 23S rRNA 两条标志性条带清晰，说明该方法提取 RNA 得到了比较纯的 RNA 样品，经公司评价，满足建库要求。将装有 RNA 样品的 EP 管用干冰密封好，送至美格基因进行宏转录组测序（测序平台为 Illumina Hiseq Xten）。

表 1. 部分样本 RNA 提取浓度和纯度, 1-4 代表酸性湿地四个土壤 RNA 样本，5-8 代表碱性湿地四个土壤 RNA 样本

SampleID	Concentration（ng/μl）	OD260/OD280	OD260/OD230	OD260	OD230	OD280
1	184.75	1.93	1.73	4.62	2.68	2.39
2	173.45	1.93	1.71	4.34	2.54	2.25
3	128.83	1.95	1.72	3.22	1.87	1.65
4	223.20	1.98	1.79	5.58	3.12	2.82
5	193.83	1.97	1.80	4.85	2.70	2.46
6	98.76	1.94	1.67	2.47	1.47	1.28
7	86.87	1.97	1.65	2.17	1.31	1.11
8	143.05	1.95	1.75	3.58	2.05	1.83

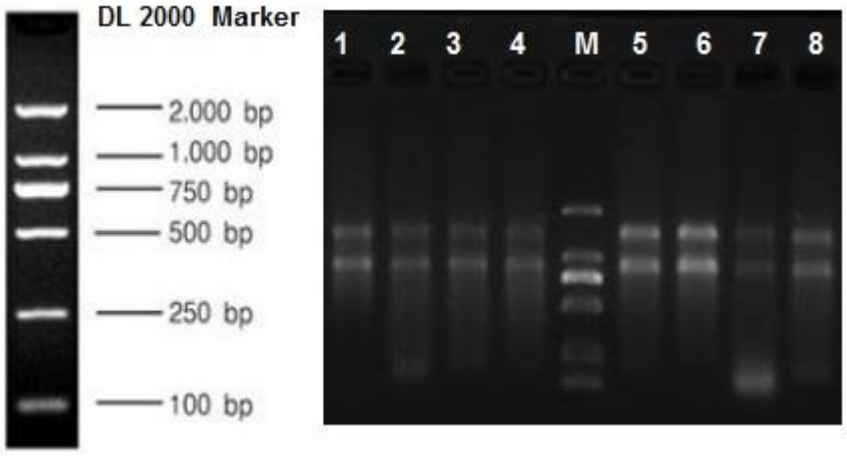

图 1. 两种类型湿地土壤样品试剂盒提取的总 RNA 琼脂糖凝胶电泳图。1-4 分别代表酸性湿地土壤 RNA，5-8 分别代表碱性湿地土壤 RNA，M 代表片段大小不同的核酸标记物。

宏转录组下机数据分析流程

1. 测序得到的原始数据（raw data）中包含接头序列及低质量碱基（Q<30）,首先经过 fastp（https://github.com/OpenGene/fastp）去接头（adapter）及过滤碱基质量后得到高质量序列（clean data）以便用于后续分析：

$ fastp -I $Sample_R1.fq.gz -I $Sample_R2.fq.gz -o $Sample_clean_R1.fq -O $Sample_clean_R2.fq -l 50 -q 30 -t 10

2. 高质量序列中仍包含大量的核糖体 RNA（rRNA），通过 SortMeRNA（https://github.com/biocore/sortmerna）过滤 clean data 中的 rRNA 序列：

$ sortmerna --ref $refdir/rRNA_databases/silva-arc-16s-id95.fasta, $refdir/index/silva-arc-16s-id95: \
$refdir/rRNA_db/silva-arc-23s-id98.fasta, $refdir/index/silva-arc-23s-id98: \
$refdir/rRNA_db/silva-bac-16s-id90.fasta, $refdir/index/silva-bac-16s-id90: \
$refdir/rRNA_db/silva-bac-23s-id98.fasta, $refdir/index/silva-bac-23s-id98: \
$refdir/rRNA_db/silva-euk-18s-id95.fasta, $refdir/index/silva-euk-18s-id95: \
$refdir/rRNA_db/silva-euk-28s-id98.fasta, $refdir/index/silva-euk-28s-id98: \
$refdir/rRNA_db/rfam-5s-database-id98.fasta, $refdir/index/rfam-5s-database-id98: \
$refdir/rRNA_db/rfam-5.8s-database-id98.fasta, $refdir/index/rfam-5.8s-database-id98 \
--reads $Sample_clean_R1.fq --reads $Sample_clean_R2.fq --fastx --other $Sample_rmRNA --aligned $Sample_aligned -a 30 -paired-out
$ unmerge-paired-reads.sh $Sample_rmRNA.fq $Sample_rmRNA_R1.fq $Sample_rmRNA_R2.fq

3. 通过 idba_tran（https://github.com/loneknightpy/idba）对转录组数据进行组装得到每个样本的转录组序列$Sample_assembly/contig.fa：

$ $IDBA/bin/fq2fa –merge $Sample_rmRNA_R1.fq $Sample_rmRNA_R2.fq $Sample_merge.fa
$ $IDBA/bin/idba_tran -r $Sample_merge.fa -o $Sample_assembly --pre_correction --mink 20 --maxk 60 --step 10 --num_threads 20

4. 组装得到的 Contig 中包含非 mRNA 信息，使用 prodigal（https://github.com/hyattpd/Prodigal）对蛋白质编码区进行预测：

$ prodigal -d $Sample_nul.fa -I $Sample_assembly/contig.fa -m -p meta

5. 预测到每个个体样品的基因序列后，利用 CD-HIT（http://weizhongli-lab.org/cd-hit/）对所有样品的基因进行聚类构建非冗余基因集（gene catalogue.fa）：

$ cat $Sample*_nul.fa > all_gene.fa
$ cd-hit-est -I all_gene.fa -c 0.95 -aS 0.9 -M 0 -o all_gene_nr -T 40

$ awk 'BEGIN{a=1}{if($0~/>/){print ">Unigene_"a;a+=1}else{print $0}}' all_gene_nr.fa >gene catalogue.fa

6. 样本中基因的表达量通过将 reads 比对到基因集（gene catalogue.fa）获得，通过将测序 reads 比对到基因序列上得到样品中基因表达量（sample_fpkm.txt）。即使用 RSEM（https://github.com/deweylab/RSEM）对 bowtie2（http://bowtie-bio.sourceforge.net/bowtie2/index.shtml）的比对结果进行统计，获得每个样本中每个基因的 reads counts，同时计算 FPKM。FPKM（全称 expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced）是每百万 fragments 中来自某一基因每千碱基长度的 fragments 数目，其同时考虑了测序深度和基因长度对 fragments 计数的影响，是对 read counts 进行标准化处理的目前最为常用的基因表达水平估算方法（Trapnell et al., 2010）。

$ rsem-calculate-expression -p 70 --bowtie2 --paired-end \
  $Sample_rmRNA_R1.fq $Sample_rmRNA_R2.fq rsem.ref $Sample
$ for i in `$Sample*.genes.results`; \
  do cut -f 1,7 $i >${i/gene.results/FPKM.txt}; \
  done
$ python merge_metaphlan_tables.py $Sample*.FPKM.txt > all_sample_fpkm.txt （脚本下载链接：https://github.com/biobakery/MetaPhlAn/blob/master/metaphlan/utils/merge_metaphlan_tables.py)

7. 使用 diamond 软件将非冗余的 Unigenes 序列与 NCBI-NR 数据库进行比对（设定阈值为 e-value≤1e-5），采取最近公共祖先 LCA（Lowest Common Ancestor）算法获得基因序列的物种注释信息：

$ diamond blastp –threads 20 -q Unigenes_pro.fa -d nr_*version.dmnd -o Unigenes_vs_nr_blt.txt --max-target-seqs 10 –evalue 1e-5 --outfmt 6 qseqid qlen sseqid stitle slen pident length mismatch gapopen qstart qend sstart send evalue bitscore
$ blast2lca -input Unigenes_vs_nr_blt.txt -o Unigenes_vs_nr_blt.tax -ms 50 -me 0.000001 -g2t gi_taxid-March2015X.bin

8. 通过将基因序列和特定数据库（如 KEGG）进行比对，完成基因功能注释。序列比对使用 diamond（https://github.com/bbuchfink/diamond）软件进行。

KEGG 数据库（https://pan.baidu.com/s/1jnulGNSQ3qDfoB3b76a0Dg，提取码：jzal）

```
$ diamond makedb –in meta.pep -d meta.dmnd
$ diamond blastp -d meta.dmnd -q Unigenes_pro.fa -k 1 -p 50 -f 6 -o Unigenes_vs_kegg_blt.txt
```

结果分析

通过和 KEGG 数据库中的 KO 数据库进行比对，得到不同层次的基因功能分类（图 2）。以一种酸性土壤样本为例，从 KO level 1 层次来看，土壤微生物的大部分活性基因与新陈代谢（Metabolism）相关，占总体基因表达量的 45.7%，其次为遗传信息加工（Genetic information processing，8.5%），环境信息加工（Environmental information processing，9.6%）和细胞过程（Cellular processes, 8.2%）。KO level 2 层次上的基因功能分类如图 2。 从图中可以看出，微生物新陈代谢相关的活动主要表现为能量代谢、碳水化合物代谢和氨基酸代谢；遗传信息加工主要表现在蛋白翻译；环境信息加工主要包括信号转导和膜转运；而细胞过程相关的基因主要参与了 cellular community-eukaryotes 和细胞迁移（cell motility）。

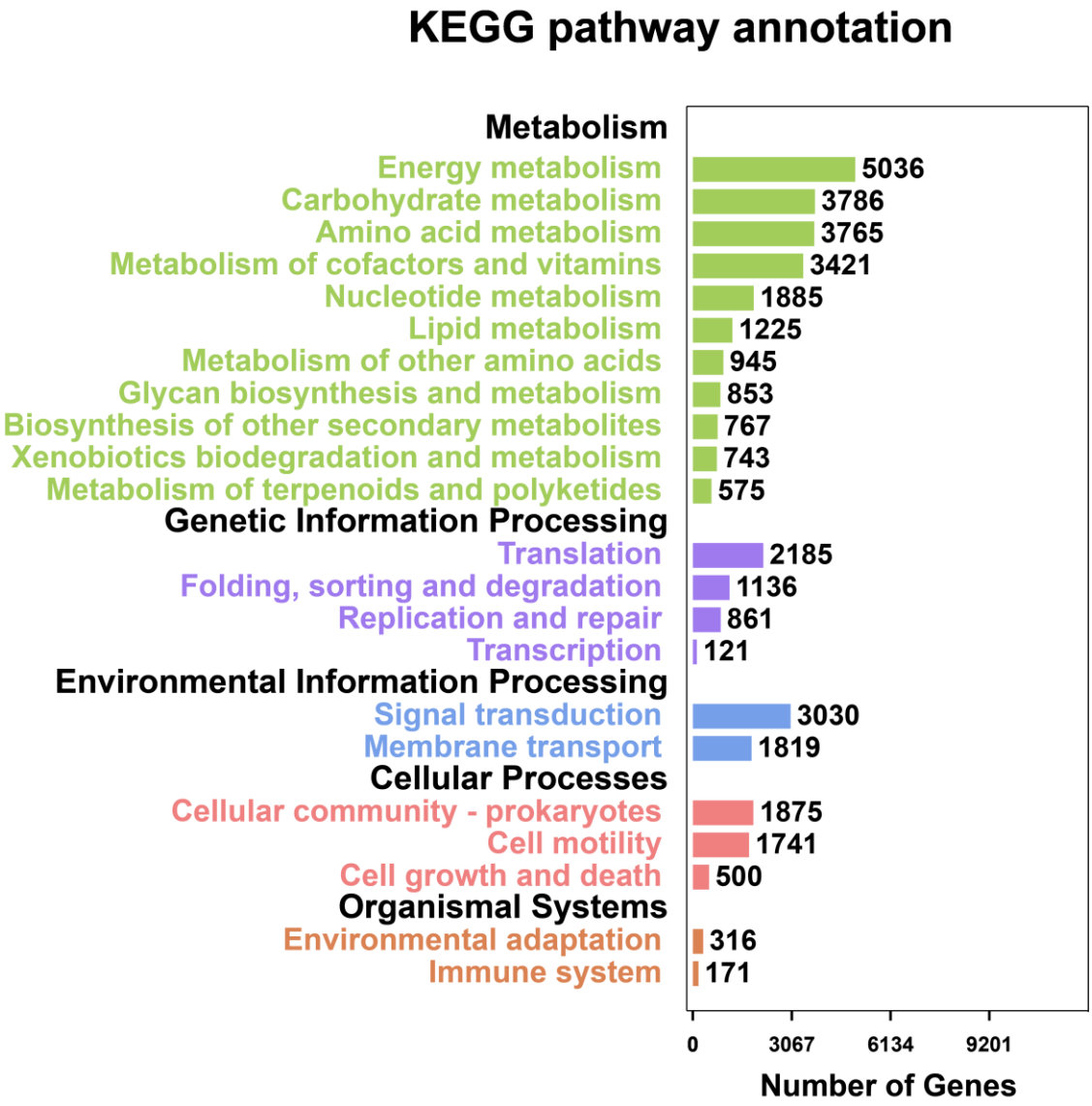

图 2.以酸性土样本为例的土壤微生物活性基因在 KEGG level2 水平上的功能分类

土壤宏转录组研究的优点

1). 当与室内培养控制实验和高通量测序结合使用时，可以估算群落中特定微生物正在进行的积极的转录过程。

2). 排除了死亡微生物细胞残体，休眠体的影响。

3). 能够捕捉土壤特定类群内部动态变化。

4). 直接评估土壤微生物活性，包括对于干扰或者暴露等情况的响应。

致谢

本实验得到国家自然科学基金项目（91951109）的资助。

参考文献

1. Mettel, C., Kim, Y., Shrestha, P.M. and Liesack, W. (2010). Extraction of mRNA from soil. *Appl Environ Microbiol* 76: 5995-6000.
2. Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Tagliabue, L., De Bellis, G.D. and Landini, P. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb Inform Exp* 3:1.
3. Torre, A., Metivier, A., Chu, F., Laurens, L.M., Beck, D.A., Pienkos, P.T., Lidstrom, M.E., and Kalyuzhnaya, M.G. (2015). Genome-scale metabolic reconstructions and theoretical investigation of methane conversion in *Methylomicrobium buryatense* strain 5G（B1). *Microb Cell Factories* 14: 188.
4. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A.,Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter,L. (2010).Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28: 5.