bio-101 www.bio-protocol.org/e2003641 DOI:10.21769/BioProtoc.2003641

易扩增子：易用、可重复和跨平台的扩增子分析流程

EasyAmplicon: An Easy-to-use, Reproducible and Cross-platform Pipeline for Amplicon Analysis

刘永鑫 [1, 2, 3, #, *]，陈同 [4, #]，周欣 [5]，白洋 [1, 2, 3, 6, *]

[1] 中国科学院遗传与发育生物学研究所，植物基因组学国家重点实验室，北京；[2] 中国科学院大学，生物互作卓越创新中心，北京；[3] 中国科学院遗传与发育生物学研究所，中国科学院–英国约翰英纳斯中心植物和微生物科学联合研究中心，北京；[4] 中国中医科学院，中药资源中心，北京；[5] 中国科学院微生物研究所，真菌学国家重点实验室，北京；[6] 中国科学院大学现代农学院，北京

*通讯作者邮箱: yxliu@genetics.ac.cn; ybai@genetics.ac.cn

#共同第一作者

引用格式：刘永鑫，陈同，周欣，白洋. (2021). 易扩增子：易用、可重复和跨平台的扩增子分析流程. Bio-101 e2003641. Doi: 10.21769/BioProtoc. 2003641.

How to cite: Liu, Y. X., Chen, T., Zhou, X. and Bai, Y. (2021). EasyAmplicon: An Easy-to-use, Reproducible and Cross-platform Pipeline for Amplicon Analysis. Bio-101 e2003641. Doi: 10.21769/BioProtoc. 2003641. (in Chinese)

摘要：扩增子测序是目前微生物组研究中最广泛的使用手段，主流的分析流程有 QIIME、USEARCH 和 Mothur，但仍分别存在依赖关系过多导致的安装困难、大数据收费和使用界面不友好等问题。本文搭建了一套完整的扩增子分析流程命名为易扩增子(EasyAmplicon)，可以实现简单易用、可重复和跨平台地开展扩增子分析。流程计算核心采用体积小、安装方便、计算速度快且跨平台的软件 USEARCH，同时整合 VSEARCH 以突破 USEARCH 免费版的限制。分析选用 RStudio 的图形界面对流程代码文档管理和运行，实现命令行和/或鼠标点击操作方式开展扩增子可重复分析。同时流程还提供了数 10 个脚本，实现特征表过滤、重采样、分组均值等常计算，以及为常用软件如 STAMP、LEfSe、PICRUSt1/2 等提供标准的输入文件的功能。易扩增子可以从 https://github.com/YongxinLiu/EasyAmplicon 下载轻松部署于 Windows/Mac/Linux 操作系统，在普通个人电脑上 2 小时内可完成数十个样本的分析。

关键词：扩增子，分析流程，16S，ITS，USEARCH

仪器设备

个人电脑/服务器 (操作系统：Windows 10/Mac OS 10.12+/Linux Ubuntu 18.04+；CPU：2 核+；内存：8G+；硬盘：> 10 GB，且大于 10 倍原始数据大小)，网络访问畅通。

软件和数据库

1. R 语言环境，下载适合自己系统的 4.0.2 版：https://www.r-project.org/
2. R 语言开发环境，用于执行流程，下载适合自己系统的 RStudio 1.3.1056：https://www.rstudio.com/products/rstudio/download/#download
3. (可选) 仅 Windows 系统安装，提供 Git Bash 命令行环境的 GitForWidnows 2.28.0：http://gitforwindows.org/
4. 扩增子分析流程 USEARCH v10.0.240 (Edgar, 2010) https://www.drive5.com/usearch/download.html
5. 扩增子分析流程 VSEARCH v2.15.0 (Rognese *et al.*, 2016) https://github.com/torognes /vsearch/releases
6. 易扩增子流程 EasyAmplicon v1.10 (Zhang *et al.*, 2018 and 2019; Chen *et al.*, 2019; Huang *et al.*, 2019; Liu *et al.*, 2020; Qia *et al.*, 2020a and 2020b)：https://github.com/YongxinLiu/EasyAmplicon
7. 核糖体数据库 RDP v16 (Cole *et al.*, 2014)：http://rdp.cme.msu.edu/
8. (可选) 核糖体数据库 GreenGene 数据库(gg) 13_8 (McDonald *et al.*, 2011)：ftp://greengenes.microbio.me/greengenes_release
9. (可选) 核糖体数据库 SILVA v123 (Quast *et al.*, 2013)：http://www.arb-silva.de
10. (可选) 转录间隔区 (ITS) 数据库 UNITE v8.2 (Nilsson *et al.*, 2019)：https://unite.ut.ee/
11. (可选) Windows 版下载工具 wget：http://gnuwin32.sourceforge.net/packages/wget.htm

软件安装和数据库部署

注：以下的软件安装和使用均在 64 位 Windows 10 系统中演示，Linux/Mac 中不同的地方会有说明，流程代码提供有 Mac 版本 (pipeline_mac.sh)。

Windows 系统需要安装 GitForWindows (http://gitforwindows.org/) 提供 Git bash

环境支持常用 Shell 命令。Linux/Mac 系统自带 Bash 命令行工作环境。

以 64 位 Windows 10 系统为例，我们先安装 R/RStudio 软件，再把本流程 (EasyAmplicon/目录) 保存于 C 盘中，然后根据需要下载数据库至指定目录即完成部署。

注：代码行添加灰色底纹背景，其中需要根据系统环境修改的部分标为蓝色。

1. 流程运行环境 R 和 RStudio

   依次安装适合系统的最新版 R 语言 (https://www.r-project.org) 和 RStudio (https://www.rstudio.com/products/rstudio/download/)。注意操作系统用户名不要使用中文，否则会影响 R 语言使用。

2. 批量安装依赖 R 包

   流程会调用数百个 R 包，使用时可自动安装。但由于网络或系统等个性原因经常出现下载或安装失败，可以使用中根据提示手动安装缺失 R 包。本文推荐直接下载我们预编译好的 R 包合辑 (http://nmdc.cn/datadownload)，替换至 R 包所在目录即可，详见**常见问题 1**。

3. 易扩增子流程 EasyAmplicon

   访问 https://github.com/YongxinLiu/EasyAmplicon，选择 Code—Download ZIP 下载并解压，如保存于 C 盘并确保目录名为 EasyAmplicon。如在 RStudio 的 Terminal 中可使用 git 下载流程：

   ```
   git clone git@github.com:YongxinLiu/EasyAmplicon.git
   ```

4. (可选) 扩增子流程依赖软件

   EasyAmplicon 依赖的 Windows/Mac/Linux 版本软件已经保存于 EasyAmplicon 中的 win/mac/linux 目录中，如果出现问题，可按如下方法手动安装。

   USEARCH 下载页 https://www.drive5.com/usearch/download.html，选择适合自己系统的 10.0.240 版本 (不要下载最新版，因为有更多功能使用受限)，如 Windows 版本保存至 EasyAmplicon 目录中的 win 目录中，解压后改名为 usearch.exe。Linux/Mac 系统需下载到环境 linux/mac 目录，解压后改名为 usearch，并添加可执行权限 (`chmod +x usearch`)。VSEARCH 下载页面 https://github.com/torognes/vsearch/releases，选择适合自己系统的最新版下载，接下来操作与 USEARCH 类似。Windows 系统还需下载 wget 程序 (http://gnuwin32.sourceforge.net/packages/wget.htm) 至 win 目录。

5. (可选) 16S 核糖体基因物种注释数据库

   16S 扩增子测序分析，常用 RDP/SILVA/GreenGene 数据库进行物种注释，可以从上述数据库官网下载并整理为 USEARCH 使用的格式，此处推荐从 USEARCH 官网 (http://www.drive5.com/sintax) 下载 USEARCH 兼容格式的数据库。默认流程使用体积小巧的 RDP v16 数据库 (rdp_16s_v16_sp.fa.gz)，并已保存于 usearch 目录中。可选 GreenGenes 13.5 (gg_16s_13.5.fa.gz)和 SILVA (silva_16s_v123.fa.gz)数据库，可根据需要下载并保存于 usearch 目录中。此外，如果要开展 PICRUSt 和 Bugbase 功能预测分析，还需要使用 GreenGenes 数据库 13.5 中按 97%聚类的 OTU 序列 (已保存于流程 gg 目录中 97_otus.fas ta.gz)。可选手动下载 GreenGenes 官方数据库 (3641)，解压后选择其中的 97_otus.fasta 保存于 gg 目录下即可。

6. (可选) ITS 物种注释数据库

   如果研究真菌或真核生物采用转录间隔区 (Intergenic Transcribed Spacer) 测序，需要使用 UNITE 数据库，目前最新版已经保存于 usearch 目录 (utax_reference_dataset_all_04.02.2020.fasta.gz)。如流程中数据库没有及时更新，可在 UNITE 官网 (https://unite.ut.ee/) 下载最新版适合 USEARCH 的注释数据库。官方数据库存在格式问题，详细**常见问题 2**。

实验步骤

开始新项目分析前，我们需要在项目目录 (如 c:/test) 中准备三类起始文件：1. EasyAmplicon 流程中复制分析流程文件 (pipeline.sh)；2. 编写样本元数据 (metadata.txt)；3. seq 目录存放测序数据 (*.fq.gz)。然后使用 RStudio 打开 pipeline.sh，设置分析流程 (EasyAmplicon, ea) 和工作目录 (work directory, wd) 位置，添加依赖可执行程序至环境变量，并切换至工作目录。

注：用户请根据操作系统类型、软件和工作目录实际位置自行修改。

```
ea=/c/EasyAmplicon
wd=/c/test
PATH=$PATH:${ea}/win
cd ${wd}
```

1. 准备输入数据 (测试数据是流程正对照)

建议下载测试元数据和测序数据作为实验的正对照，首先完成全部流程分析，以确定流程部署成功。将来使用自己的数据出现问题，可以与测试数据分析中对应结果比较，以便确定问题产生的原因。

下载示例元数据用于参考编写格式 (表 1)。

```
wget -c http://210.75.224.110/github/EasyAmplicon/data/metadata.txt
```

表 1. 元数据格式示例

SampleID	Group	Date	Site	CRA	CRR
KO1	KO	2017/6/30	Chaoyang	CRA002352	CRR117575
KO2	KO	2017/6/30	Chaoyang	CRA002352	CRR117576
KO3	KO	2017/7/2	Changping	CRA002352	CRR117577

可用 Excel 编写，保存存为制表符分隔的的文本文件。

注意：有行列标题，行为样品名 (字母开头+数字组合)，列为分组信息 (至少 1 列，可多列)、地点和时间 (提交数据必须)、及其它属性，详见附表 1，或下载的 metadatat.txt 文件。

通常测序公司会返回原始数据，如 Illumina 双端测序的文件，每个样本有一对文件。本文使用的数据来自发表于 Science 杂志关于拟南芥根系微生物组研究的文章 (Huang *et al.*, 2019)，GSA 项目号为 PRJCA001296。为方便演示流程的使用，我们从中选取三个组 (每组包括 6 个生物学重复共 18 个样本)，并且随机抽取了 50,000 对序列作为软件的测序数据，该数据可以从中国科学院基因组研究所的原始数据归档库 (Genome Sequence Archive，GSA，https://bigd.big.ac.cn/gsa/) (Wang *et al.*, 2017) 中按批次编号 CRA002352 搜索并下载。本文使用 wget 根据样本元数据中批次和样本编号批量下载至 seq 目录，代码如下。

```
mkdir -p seq
awk '{system("wget -c ftp://download.big.ac.cn/gsa/"$5"/"$6"/"$6"_f1.fq.gz -O seq/"$1"_1.fq.gz")}' <(tail -n+2 metadata.txt)
awk '{system("wget -c ftp://download.big.ac.cn/gsa/"$5"/"$6"/"$6"_r2.fq.gz -O seq/"$1"_2.fq.gz")}' <(tail -n+2 metadata.txt)
```

awk 为 Linux 下的一种字符处理语言，可同时使用文本中的多个字段；使用

system 命令调用 wget，实现根据列表批量下载、改名的功能。

检查文件大小，确定是否下载完整或正常。

```
ls -lsh seq
```

数据库第一次使用需要解压

```
gunzip ${ea}/usearch/rdp_16s_v16_sp.fa.gz
gunzip ${ea}/gg/97_otus.fasta.gz
```

创建临时和结果目录，临时目录分析结束可删除

```
mkdir -p temp result
```

2. 合并双端序列并按样品重命名

依照实验设计采用 for 循环批处理样本合并 (图 1)。tail -n+2 去表头，cut -f 1 取第一列，即获得样本列表。vsearch 程序的--fastq_mergepairs 命令实现双端序列合并，接输入读长文件 1 (fq/fq.gz 均可)，--reverse 接读长文件 2，--fastqout 指定输出文件，--relabel 将序列按样本名进行重命名 (注：样本名后必须添加点，以分隔样本名和序列 ID，同时注意样本名中不允许有点)。本示例数据包括 5 万对读长的 18 个样本合并计算耗时约 2 min。Windows 下复制命令 Ctrl+C 为 Linux 下的终止命令，为防止异常中断，结尾添加&使命令转后台。注：如分析时提示文件质量值问题，详见**常见问题 3**；如输入文件为单端 FASTQ 文件，则只需序列改名即可，详见**常见问题 4**。

```
for i in `tail -n+2 result/metadata.txt | cut -f 1`;do
  vsearch --fastq_mergepairs seq/${i}_1.fq.gz --reverse seq/${i}_2.fq.gz \
  --fastqout temp/${i}.merged.fq --relabel ${i}.
done &
```

双端合并且重命名的序列，每条序列具有唯一且可识别样本的 ID。可以合并所有样品至同一文件，方便统一操作。

```
cat temp/*.merged.fq > temp/all.fq
```

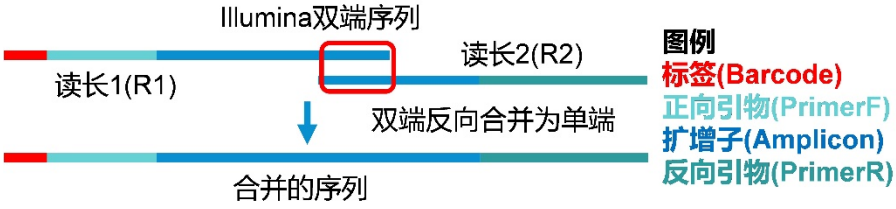

图 1. 典型扩增子测序双端序列合并结构图

3. 引物切除和质量控制

采用等长的方式切除引物，引物外侧如果有标签 (Barcode)，标签的长度需要计算在内 (图 1)。如本示例的结构为 10 bp 左端标签 + 19 bp 正向引物 V5 共计 29 bp，18 bp 反向引物 V7，分别使用--fastq_stripleft 和--fastq_stripright 传递给程序。注意：务必清楚实验设计中引物和标签长度，如果引物已经去除，可在下方参数处填 0 表示无需去除。此外，--fastq_maxee_rate 指定质量控制的错误率，0.01 代表要求错误率小于 1%。质控控制后，--fastaout 输出体积更小的无质量值 fasta (fa) 格式文件。

```
vsearch --fastx_filter temp/all.fq \
        --fastq_stripleft 29 --fastq_stripright 18 \
        --fastq_maxee_rate 0.01 \
        --fastaout temp/filtered.fa
```

4. 序列去冗余并挑选代表序列 (OTU/ASV)

序列去冗余可以使总数据量降低数量级 (降维)，减少下游计算资源消耗和缩短等待时间，更重要的是提供序列的出现频次对鉴定真实特征序列至关重要。扩增子分析中特征序列有可操作分类单元 (OTUs) 或扩增序列变体 (ASVs)。参数--miniuniqusize 设置使用序列的最小出现频次，默认为 8，此处设置为 10，推荐最小值为总数据量的百万分之一 (如 1 亿条序列至少需要过滤掉频次 100 以下的序列噪音)，可实现去除低丰度或测序噪音并极大地增加计算速度。-sizeout 输出频次，--relabel 设置输出序列前缀 (示例输出序列 ID: > Uni1; size = 17963)。

```
vsearch --derep_fulllength temp/filtered.fa \
        --output temp/uniques.fa \
        --relabel Uni --minuniquesize 10 --sizeout
```

按 97%聚类生成 OTUs 和去噪挑选 ASVs 是目前挑选特征序列的两种主流方法。通常两种分析方法的结果整体上比较类似，细节上会略有不同。下面按方法 1 或 2 分别介绍，用户可根据实际需求选择。如聚类或去噪运行中提示超过内存限制错误，表明数据低丰度和/或测序噪音较多，可增加上步--minuniquesize 参数 (如 30，50 或更大)，使输出非冗余序列数据小于 1 万条为宜，保证下游分析顺利且高效完

成。

方法 1. 使用 UPARSE (Edgar, 2013) 算法按 97%的序列相似度聚类 OTU

```
usearch -cluster_otus temp/uniques.fa \
        -otus temp/otus.fa \
        -relabel OTU_
```

此方法累计使用次数最多、分析速度快，适合大数据或 ASV 方法结果规律不明显时尝试。

方法 2. 使用 UNOISE 算法 (Edgar and Flyvbjerg, 2015) 去噪生成 ASV

此方法是当前的主流，推荐优先使用。类似于按 100%的序列相似度聚类，或不聚类的方法，详见方法原始文献 (Edgar and Flyvbjerg, 2015) 或宏基因组公众号推文《扩增子分析还聚 OTU 就真 OUT 了》。采用更先进的方法来鉴定测序过程中可能的错误，因此也需要消耗更多的计算时间。UNOISE 算法虽然慢于 UPARSE 算法，但也比同类去噪算法 Deblur 和 DADA2 分别快 10 倍和 100 倍 (Amir *et al.*, 2017)。此处-unoise3 去噪结果默认前缀为 Zotu，我们修改为主流使用的 ASV。

```
usearch -unoise3 temp/uniques.fa \
        -zotus temp/zotus.fa
sed 's/Zotu/ASV_/g' temp/zotus.fa > temp/otus.fa
```

(可选) 基于参考去嵌合。

全头 (*de novo*) 去嵌合时要求亲本丰度为嵌合体 16 倍以上防止造成假阴性，而参考数据库无丰度信息，只需 1 条序列在参考数据中没有相似序列且与 2-3 条序列相似即判定为嵌体，因此容易引起假阴性 (真实序列被当作假序列丢弃)，不推荐使用。如果必须要使用，由于已知序列不会被去除，选择越完整的数据库可降低假阴性率。

方法 1. VSEARCH 结合 RDP 数据库去嵌合 (快但容易假阴性)，推荐 SILVA 去嵌合 (silva_16s_v123.fa)，但计算极耗时 (本例用时 3 h，是 RDP 计算时间的 30 倍以上)。

```
vsearch --uchime_ref temp/otus.fa \
        -db ${ea}/usearch/rdp_16s_v16_sp.fa \
        --nonchimeras result/raw/otus.fa
```

Windows 系统下 vsearch 结果会添加了 windows 换行符^M 必需删除，否则会

出现换行混乱的问题。Mac/Linux 系统无须执行此命令。

```
sed -i 's/\r//g' result/raw/otus.fa
```

方法 2. 不去嵌合。

```
cp -f temp/otus.fa result/raw/otus.fa
```

5. 特征表生成和筛选

5.1 生成特征表

使用 vsearch 的--usearch_global 命令比对扩增子序列 (temp/filtered.fa) 至特征序列 (result/raw/otus.fa) 即可生成特征表，--threads 设置整数使用计算机可用的多线程加速计算。

```
vsearch --usearch_global temp/filtered.fa --db result/raw/otus.fa \
    --otutabout result/raw/otutab.txt --id 0.97 --threads 4
sed -i 's/\r//' result/raw/otutab.txt
```

5.2 按物种注释筛选特征表

基于物种注释，(可选) 可以筛选质体和非细菌和非古菌去除并统计比例。

使用 USEARCH 的 sintax 算法进行物种注释。选择 RDP 物种注释 (rdp_16s_v16_sp) 数据库具有体积小、分类速度极快 (本例耗时 15 s) 的特点，但缺少真核数据无法注释真核污染物来源详细。SILVA 数据库 (silva_16s_v123.fa) 可以更好地注释真核质体序列来源，但速度较慢 (本例耗时约 3 h)；还可选 Greengenes 数据库 (gg_16s_13.5.fa)，但此数据库自 13 年起再无更新。--sintax_cutoff 设置分类结果的可信度阈值，范围 0-1 之间，文章中常用 0.6/0.8，取值越大注释结果越可靠同时注释比例也越低。注意，结果中第三列方向正常应全为正向 (+)，如果全为反向 (-)，请参考**常见问题 5** 中方法将第 3 步结果序列取反向互补。

```
vsearch --sintax result/raw/otus.fa --db ${ea}/usearch/rdp_16s_v16_sp.fa \
    --tabbedout result/raw/otus.sintax --sintax_cutoff 0.6
```

为去除 16S rDNA 测序中的非特异性扩增和质体污染，我们编写了 R 脚本 otutab_filter_nonBac.R 实现选择细菌和古菌 (原核生物)、以及去除叶绿体和线粒体并统计比例，输出筛选后并按丰度排序的特征表。输入文件为原始特征表 (result/raw/otutab.txt) 和物种注释 (result/raw/otus.sintax)，输出筛选并排序的

特征表（result/otutab.txt）、统计污染比例文件（result/raw/otutab_nonBac.txt）和过滤细节（otus.sintax.discard），特征表的格式详见文件或附表 2。注：真菌 ITS 数据，请改用 otutab_filter_nonFungi.R 脚本，只筛选注释为真菌的序列。查看脚本帮助，请运行 Rscript ${ea}/script/otutab_filter_nonBac.R -h。

```
Rscript ${ea}/script/otutab_filter_nonBac.R \
--input result/raw/otutab.txt \
--taxonomy result/raw/otus.sintax \
--output result/otutab.txt\
--stat result/raw/otutab_nonBac.stat \
--discard result/raw/otus.sintax.discard
```

5.3 筛选特征表对应序的列和物种注释

特征表筛选后，对应的代表序列（otus.fa 或附表 3）、物种注释信息也需要对应进行筛选。

```
cut -f 1 result/otutab.txt | tail -n+2 > result/otutab.id
usearch -fastx_getseqs result/raw/otus.fa \
        -labels result/otutab.id -fastaout result/otus.fa
awk 'NR==FNR{a[$1]=$0}NR>FNR{print a[$1]}'\
        result/raw/otus.sintax result/otutab.id \
        > result/otus.sintax
sed -i 's/\t$/\td:Unassigned/' result/otus.sintax
```

此外，如果上述筛选方案不适合你的研究，如去除的 Chloroplast 为你研究的对象，可以跳过此步不进行筛选，运行 cp result/raw/otu* result/。

接下来对最终的特征表进行统计，结果有助于优化前面的分析方案，以及选择下游分析合适的参数。

```
usearch -otutab_stats result/otutab.txt \
      -output result/otutab.stat
cat result/otutab.stat
```

```
645539  Reads (645.5k)
    18  Samples
  2903  OTUs

 52254  Counts
 10183  Count  =0  (19.5%)
  8415  Count  =1  (16.1%)
  9497  Count >=10 (18.2%)

   832  OTUs found in all samples (28.7%)
  1155  OTUs found in 90% of samples (39.8%)
  2698  OTUs found in 50% of samples (92.9%)

Sample sizes: min 32086, lo 33267, med 36956, mean 35863.3, hi 37541, max 38967
```

图 2. USEARCH 的特征表统计结果示例

统计结果显示了总读长数量、样本量、特征数量，可以了解特征表的总体数据量和维度信息；接下来是特征表中单元格数量、为 0、1 和 > 10 的数量和比例，了解特征中频次分布情况；然后是特征在 100%、90%和 50%样品中发现的数量，展示了特征在样本中的流行频率；最后是样本测序量的分位数，对选择合适的重采样阈值非常有帮助。

我们根据特征表统计结果，将选择合适的参数对特征标进行等量重抽样方式的标准化，以减小由于样本测序量不一致引起的多样性差异，可实现更准确地多样性分析。重采样使用 otutab_rare.R 脚本调用 vegan 包 (Oksanen *et al.*, 2007) 实现，并计算了 6 种常用 alpha 多样性 (richness、chao1、ACE、shannon、simpson 和 invsimpson) 指数 (vegan.txt 或**表 2**)。重采样深度 (--depth) 参考特征表统计结果 (图 2) 选择，一般默认按最小值重采样。提高采样深度可以保留样本中更大的测序量，但也会剔除低于阈值的样本。因此如果样本测序量波动极大，尽量选择合适的阈值重采样以最大化保留测序量。

```
mkdir -p result/alpha
Rscript ${ea}/script/otutab_rare.R --input result/otutab.txt \
        --depth 32086 --seed 1 \
        --normalize result/otutab_rare.txt \
        --output result/alpha/vegan.txt
usearch -otutab_stats result/otutab_rare.txt \
        -output result/otutab_rare.stat
```

```
cat result/otutab_rare.stat
```

结果显示所有样本重采样后读长数量均为 32,086。这样特征表可以最大化减少测序量的影响，以便更准确评估多样性。

表 2. Alpha 多样性指数示例

SampleID	richness	chao1	ACE	shannon	simpson	invsimpson
KO1	2350	2692.008	2686.869	6.132835	0.990308	103.1788
KO2	2316	2664.35	2661.86	6.17406	0.991875	123.0733
KO3	1935	2278.252	2283.343	5.828452	0.989582	95.98662

由 otutab_rare.R 调用 vegan 包计算的 6 种常用 alpha 多样性指数，图中仅展示结果前 4 行作为示例。

6. Alpha 多样性计算

前面在特征表重采样标准化时，计算了 6 种常用 alpha 多样性指数。此外，USEARCH 的-alpha_div 命令可以快速计算 18 种 alpha 多样性指数 (alpha.txt)，各种指数的计算公式和描述详见：http://www.drive5.com/usearch/manual/alpha_metrics.html。这些结果我们常用于结合样品元数据开展组间比较，或箱线图展示组间异同。

```
usearch -alpha_div result/otutab_rare.txt \
        -output result/alpha/alpha.txt
```

由于测序数据深度对多样性影响较大，有时我们也关注不同测序样量多样性的变化，即可以判断组间差异是否在不同测序深度下稳定存在，同时确定测序量是否饱和并反映出结果较真实的多样性。USEARCH 的-alpha_div_rare 命令实现快速无放回百分数重采样计算各样本的丰富度 (richness/observed_feature，详见 alpha_rare.txt 或附表 4)。结果可进一步可视化为样本稀释曲线，或分组带误差棒的稀释曲线或箱线图。

```
usearch -alpha_div_rare result/otutab_rare.txt \
        -output result/alpha/alpha_rare.txt \
        -method without_replacement
```

Alpha 多样性丰富度指数相似只代表物种数量相近，然而其中的物种种类可能完全不同。我们需要制作记录每个组指大于定丰度的物种是否存在的数据格式 (表

3 和附表 5)，用于组间比较物种共有和特有的情况。可以使用 ImageGP 在线 (http://www.ehbio.com/ImageGP/) 选择维恩图 (Venn diagram)、集合图 (Upset view) 或桑基图 (Sankey diagram) 等方式展示。

表 3. 用于比较各组特征共有/特有的数据示例

特征 ID	分组 ID
ASV_1	All
ASV_1	KO
ASV_1	OE
ASV_1	WT
ASV_2	KO
…	…

我们通常结合元数据计算各组的丰度均值，如以 Group 列为分组信息计算原始计数的相对丰度并求组均值。

```
Rscript ${ea}/script/otu_mean.R --input result/otutab.txt \
    --design metadata.txt \
    --group Group --thre 0 \
    --output result/otutab_mean.txt
```

因为特征的数量较大，而且低丰度的特征是否存在偶然性较大，准确性不高且与测序噪音无法区分。因此筛选大于某一丰度阈值结果，可实现数据降维并保留数据的主体，然后用于组间比较共有和特有的情况。如以平均丰度> 0.1%为阈值，可选 0.5%或 0.05%，得到每个组中符合条件的特征 (表 3)。

```
awk 'BEGIN{OFS=FS="\t"}{if(FNR==1) {for(i=2;i<=NF;i++) a[i]=$i;} \
        else {for(i=2;i<=NF;i++) if($i>0.1) print $1, a[i];}}' \
        result/otutab_mean.txt > result/alpha/otu_group_exist.txt
```

7. Beta 多样性计算

Beta 多样性是群落整体结构的降维分析方法，需要基于特征表计算样本间的各种距离矩阵。常用的 Unifrac 算法 (Lozupone *et al.*, 2010) 考虑物种间的进化距离，这里我们使用 usearch 的-cluster_agg 命令基于代表序列获得进化树，然后再使用-beta_div 基于特征表和进化树计算多种矩阵矩阵，包括 bray_curtis, euclidean,

jaccard, manhatten, unifrac 等，每类矩阵还分为有或无 (binary) 权重两种 (附表 6 展示 Bray-Curtis 距离矩阵示例)。

```
mkdir -p result/beta/
usearch -cluster_agg result/otus.fa -treeout result/otus.tree
usearch -beta_div result/otutab_rare.txt -tree result/otus.tree \
        -filename_prefix result/beta/
```

8. 物种注释分类汇总

前在特征表筛选时已经对特征序列完成了物种注释，并根据注释进行筛选。物种注释存在命名混乱、分类级不完整和名称缺失等问题。我们先对格式进行调整方便开展分析。调整物种注释为特征 ID 和 7 级分类注释的两列格式 (表 4 和附表 7)。注意 7 级分类可能存在不完整的情况，可能是该特征没有相近种的报导，也可能是参考物种注释自身不完善。

```
cut -f 1,4 result/otus.sintax \
        |sed 's/\td/\tk/;s/:/__/g;s/,/;/g;s/"//g;s/\/Chloroplast//' \
        > result/taxonomy2.txt
```

表 4. 物种注释 2 列格式示例

特征 ID	分组 ID
ASV_1	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Thermomonosporaceae
ASV_2	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Pelomonas;s__Pelomonas_puraquae
ASV_3	**k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Pelomonas;s__Pelomonas_puraquae**

物种注释的 7 级分类，也提供了特征表多角度降维分析的可能，可以把特征表按物种注释信息合并为门、纲、目、科、属级别的表，以进行更容易与现在科学发现结合的讨论。首先将物种注释转化为 7 级分类的 8 列表格 (表 5 和附表 8)，其中缺失的分类级别填充为未分类 (Unassigned)。

```
awk 'BEGIN{OFS=FS="\t"}{delete a;
a["k"]="Unassigned";a["p"]="Unassigned";a["c"]="Unassigned";a["o"]="Unassigned";a["f"]="Unassigned";a["g"]="Unassigned";a["s"]="Unassigned";\
```

```
        split($2,x,";");for(i in x){split(x[i],b,"__");a[b[1]]=b[2];} \
        print $1,a["k"],a["p"],a["c"],a["o"],a["f"],a["g"],a["s"];}' \
        result/taxonomy2.txt > temp/otus.tax
sed 's/;/\t/g;s/.__//g;' temp/otus.tax|cut -f 1-8 | \
        sed '1s/^/OTUID\tKingdom\tPhylum\tClass\tOrder\tFamily\tGenus\tSpecies\n/' \
        > result/taxonomy.txt
```

表 5. 物种注释 8 列格式示例

OTUID	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Thermomonosporaceae	Unassigned	Unassigned
ASV_2	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pelomonas	Pelomonas_puraquae
ASV_3	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Rhizobacter	Rhizobacter_bergeniae

接下来对特征表按各级别进行分类汇总，获得门、纲、目、科和属水平上的特征表 (附表 9 门水平分类汇总表)，即可以直接用来绘制物种组成图，也可以进一步从更多角度分析组间差异或挖掘生物标志物。

```
for i in p c o f g;do
        usearch -sintax_summary result/otus.sintax \
        -otutabin result/otutab_rare.txt -rank ${i} \
        -output result/tax/sum_${i}.txt
done
sed -i 's/(//g;s/)//g;s/\"//g;s/\#//g;s/\/Chloroplast//g' result/tax/sum_*.txt
```

9. 有参分析和功能预测

一些功能注释数据库，如 PICRUSt (Langille *et al,* 2013)、BugBase (Ward *et al,* 2017) 等的输入文件必须是基于 GreenGenes 数据库生成的特征表。此处采用 vsearch 的--usearch_global 命令比对扩增子数据至 GreenGenes 数据库即可获得有参分析的特征表 (gg/otutab.txt)，该结果可作为主流功能预测软件的输入文件开展分析。

```
vsearch --usearch_global temp/filtered.fa --db ${ea}/gg/97_otus.fasta \
        --otutabout result/gg/otutab.txt --id 0.97 --threads 12
```

注意：如果使用 PICRUSt2 可直接使用第 4 步中的序列 (otus.fa) 和第 5 步中

的特征表 (result/otutab.txt) 作为输入文件开展分析。

10. 空间清理及数据提交

整个分析过程会占用原始数据大小数 10 倍的空间。项目结果分析结束可以删除整个 temp 文件夹。分析取得阶段性成果，也要及时清楚不常用的中间文件节省空间，如 fastq 文件。这个习惯对于团队共享存储时尤其重要，否则硬盘空间耗尽，所有任务都会立即停止。

```
rm -rf temp/*.fq
```

短期不用数据库压缩节省空间。

```
gzip ${ea}/usearch/rdp_16s_v16_sp.fa
gzip ${ea}/gg/97_otus.fasta
```

原始序列统计 md5 值，用于数据提交 (附表 10)。

```
cd seq
md5sum *_1.fq.gz > md5sum1.txt
md5sum *_2.fq.gz > md5sum2.txt
paste md5sum1.txt md5sum2.txt | awk '{print $2"\t"$1"\t"$4"\t"$3}' | \
  sed 's/*//g' > ../result/md5sum.txt
rm md5sum*
cd ..
cat result/md5sum.txt
```

11. STAMP 和 LEfSe 软件输入文件准备

STAMP 是常用的图型界面特征差异比较软件 (Parks *et al.*, 2014)，操作简单，同时支持主流操作系统 (Windows/Linux/Mac，在 Windows 中安装最方便)，软件可从其主页 https://beikolab.cs.dal.ca/software/STAMP 获取。我们提供了 format2stamp.R 脚本，可以基于特征表、物种注释信息快速获得界、门、纲、目、科、属、种、OTU/ASV 的 STAPM 输入格式兼容的特征表，同时提供按丰度过滤的参数。如按万分之一相对丰度过滤生成 STAMP 输入文件的代码如下：

```
Rscript ${ea}/script/format2stamp.R -h
mkdir -p result/stamp
Rscript ${ea}/script/format2stamp.R --input result/otutab.txt \
  --taxonomy result/taxonomy.txt --threshold 0.01 \
  --output result/stamp/tax
```

LEfSe 是常用的生物标记鉴定软件 (Segata *et al.*, 2011)，支持多组比较。其输入文件格式是整合了样本分组信息、界、门、纲、目、科、属层面相对丰度的结果。同时为了展示可读性的进化分枝树图形，还需要对特征表进行筛选。我们提供了 format2lefse.R 脚本，可以一步生成 LEfSe 要求的输入文件，同时提供按丰度过滤的参数。如按千分之一 (threshold) 丰度筛选以控制作图中的进化分枝数量有较好的可读性，代码如下：

```
mkdir -p result/lefse
Rscript ${ea}/script/format2lefse.R --input result/otutab.txt \
  --taxonomy result/taxonomy.txt --design result/metadata.txt \
  --group Group --threshold 0.1 \
  --output result/lefse/LEfSe
```

结果文件可以在软件官网 (http://huttenhower.sph.harvard.edu/galaxy) 或我们建立的国内备份站 ImageGP (http://www.ehbio.com/ImageGP) 开展在线分析。结果的进化分枝图中分枝过密和/或文字重叠严重，可进一步提高丰度阈值以减少分枝数量，反之同理。

常见问题

1. 软件、数据库下载慢或无法下载

由于国际带宽和站点的速度限制等原因，很多国外数据库下载缓慢甚至无法下载。宏基因组公众号团队建立了微生物组领域的扩增子和宏基因组常用软件和数据库的国内备份站点，方便同行下载和使用。站点 1. 国家微生物科学数据中心的数据下载页面—工具资源下载栏目 (http://nmdc.cn/datadownload) 即为宏基因组团队与中科院微生物所共同维护的站点之一，提供宏基因组常用软件、数据库的 FTP 下载链接。站点 2. 由刘永鑫的 GitHub 中《微生物组数据分析与可视化实战》专著的大数据下载页面 (https://github.com/YongxinLiu/MicrobiomeStatPlot/blob/master/Data/BigDataDownlaodList.md) 提供有常用资源下载百度云链接和 HTTP 下载链接。

R 包的批量安装，需要在 RStudio 中查看 R 包所在位置，然后替换下载的 R 包合辑，详细操作见—个人电脑搭建微生物组分析平台 (Win/Mac)。

2. ITS 物种注释数据库 UNITE 使用时报错

UNITE 数据库官方提供的数据格式有时存在错误。主要是分类级空缺的问题，我们使用 sed 命令对 utax8.2 数据库进行调整，示例如下。

```
sed -i 's/,;/,Unnamed;/;s/:,/:Unnamed,/g'
utax_reference_dataset_all_04.02.2020.fasta
```

3. 文件 Phred 质量错误—Fastq 质量值 64 转 33

Illumina 测序的 Fastq 格式文件中序列的质量值通常为 Phred33 格式，典型特点为以大写字母为主。有时测序服务提供商也会提供旧版 Phred64 格式的 Fastq 文件，直接使用会提示 Phred 编码错误，我们需要使用 vsearch 的--fastq_convert 转换 Phred64 为常用的 Phred33 格式。此外可选 fastp (Chen *et al.,* 2018) 实现格式转换。

```
vsearch --fastq_convert test_64.fq \
        --fastqout test.fq \
        --fastq_ascii 64 --fastq_asciiout 33
```

4. 单端序列改名

如果是单端测序数据或已经合并后的单端 FASTQ 序列样本文件，需要按样本名重命名每条序列，才能进行下游分析，否则将无法区分序列的样本来源。我们使用 vsearch --fastq_convert 命令中的--relabel 参数对序列按样本重命名，以 WT1 样本为例。

```
vsearch --fastq_convert test.fq --fastqout WT1.fq --relabel WT1.
```

5. Fasta 序列取反向互补

物种注释时发现序列全为反向 (-)，表明序列的方向有错误，可用 vsearch 的--fastx_revcomp 命令调整。

```
vsearch --fastx_revcomp filtered_test.fa --fastaout filtered.fa
```

致谢

本项目得到中国科学院青年创新促进会资助 (编号：2021092) [Supported by Youth Innovation Promotion Association CAS (No. 2021092)]。此分析流程在我之前的综述中被提及 (刘永鑫等，2019; Liu *et al.,* 2020)，本文是发表的详细使用方法和常见问题解决的经验。

参考文献

1. 刘永鑫，秦媛，郭晓璇，白洋. (2019). 微生物组数据分析方法与应用. *遗传* 41(9): 845-826.
2. Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A. and Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2(2): e00191-00116.
3. Chen, Q., Jiang, T., Liu, Y. X., Liu, H., Zhao, T., Liu, Z., Gan, X., Hallab, A., Wang, X., He, J., Ma, Y., Zhang, F., Jin, T., Schranz, M. E., Wang, Y., Bai, Y. and Wang, G. (2019). Recently duplicated sesterterpene (C25) gene clusters in *Arabidopsis thaliana* modulate root microbiota. *Sci China Life Sci* 62(7): 947-958.
4. Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17): i884-i890.
5. Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R. and Tiedje, J. M. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42(Database issue): D633-642.
6. Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19): 2460-2461.
7. Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10(10): 996-998.
8. Edgar, R. C. and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 31(21): 3476-3482.
9. Huang, A. C., Jiang, T., Liu, Y. X., Bai, Y. C., Reed, J., Qu, B., Goossens, A., Nutzmann, H. W., Bai, Y. and Osbourn, A. (2019). A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. *Science* 364(6440): eaau6389.
10. Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G. and Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31(9): 814-

821.

11. Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X. and Bai, Y. (2020). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* 11.
12. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J* 5(2): 169-172.
13. McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R. and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6(3): 610-618.
14. Nilsson, R. H., Larsson, K. H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glockner, F. O., Tedersoo, L., Saar, I., Koljalg, U. and Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 47(D1): D259-D264.
15. Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J. and Suggests, M. (2007). The vegan package. *Commun Ecol Pack* 10: 631-637.
16. Qian, X. B., Chen, T., Xu, Y. P., Chen, L., Sun, F. X., Lu, M. P. and Liu, Y. X. (2020a). A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. *Chin Med J (Engl)* 133(15): 1844-1855.
17. Qian, X., Liu, Y. X., Ye, X., Zheng, W., Lv, S., Mo, M., Lin, J., Wang, W., Wang, W., Zhang, X. and Lu, M. (2020b). Gut microbiota in children with juvenile idiopathic arthritis: characteristics, biomarker identification, and usefulness in clinical prediction. *BMC Genomics* 21(1): 286.
18. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41(Database issue): D590-596.
19. Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584.
20. Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N., Zhang, Q., Bai, Z., Dong, X., Chen, H., Sun, M., Zhai, S., Sun, Y., Yu, L.,

Lan, L., Xiao, J., Fang, X., Lei, H., Zhang, Z. and Zhao, W. (2017). GSA: Genome Sequence Archive*. *Genom Proteom Bioinf* 15(1): 14-18.

21. Ward, T., Larson, J., Meulemans, J., Hillmann, B., Lynch, J., Sidiropoulos, D., Spear, J. R., Caporaso, G., Blekhman, R., Knight, R., Fink, R. and Knights, D. (2017). BugBase predicts organism-level microbiome phenotypes. *bioRxiv*: 133462.
22. Zhang, J., Liu, Y. X., Zhang, N., Hu, B., Jin, T., Xu, H., Qin, Y., Yan, P., Zhang, X., Guo, X., Hui, J., Cao, S., Wang, X., Wang, C., Wang, H., Qu, B., Fan, G., Yuan, L., Garrido-Oter, R., Chu, C. and Bai, Y. (2019). *NRT1.1B* is associated with root microbiota composition and nitrogen use in field-grown rice. *Nat Biotechnol* 37(6): 676-684.
23. Zhang, J., Zhang, N., Liu, Y. X., Zhang, X., Hu, B., Qin, Y., Xu, H., Wang, H., Guo, X., Qian, J., Wang, W., Zhang, P., Jin, T., Chu, C. and Bai, Y. (2018). Root microbiota shift in rice correlates with resident time in the field and developmental stage. *Sci China Life Sci* 61(6): 613-621.
24. Parks, D. H., Tyson, G. W., Hugenholtz, P. and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30(21): 3123-3124.
25. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S. and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol* 12(6): R60.