bio-101 www.bio-protocol.org/e2003554 DOI:10.21769/BioProtoc.2003554

使用 QIIME 2 分析微生物组 16S rRNA 基因扩增子测序数据

Using QIIME 2 to Analysis Amplicon Sequencing of 16S rRNA Gene in Microbiome

刘永鑫 [1, 2, 3, #, *]，陈同 [4, #]，钱旭波 [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] 浙江中医药大学附属金华中医院儿科，金华，浙江；[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/同等贡献

引用格式：刘永鑫，陈同，钱旭波，白洋. (2021). 使用 QIIME 2 分析微生物组 16S rRNA 基因扩增子测序数据. Bio-101 e2003554. Doi: 10.21769/BioProtoc. 2003554.

How to cite: Liu, Y. X., Chen, T. Qian, X. B. and Bai, Y. (2021). Using QIIME 2 to Analysis Amplicon Sequencing of 16S rRNA Gene in Microbiome. Bio-101 e2003554. Doi: 10.21769/BioProtoc.2003554. (in Chinese)

摘要：QIIME 是目前微生物组扩增子分析领域使用最广泛的流程 (软件)，论文发表 10 年已经被引用超 2 万次。虽然 QIIME 的推出在微生物组数据分析领域具有里程碑意义，但是该流程已经无法满足快速发展的微生物组数据分析需求。全新开发的 QIIME 2 流程采用 Python 3 编写，它结合最新算法、提供交互式图表、插件可扩展性强、能更好地满足当前大数据和可重复分析的要求。然而，QIIME 2 无法在主流的 Windows 系统下直接运行，且用户说明文档长达 10 多万字，对缺少生物信息背景的研究人员来说，学会使用这样的流程仍然是个巨大挑战。本文将介绍软件安装方法和标准分析流程，方便同行快速上手使用 QIIME 2 流程；我们对使用过程的中间步骤和参数进行解读，帮助用户掌握参数优化方法，以获得更合理的结果；同时对软件安装和使用过程中的常见问题和解决方案进行总结。本文介绍的微生物组分析指标和方法具体包括数据导入导出、特征表生成、alpha 和 beta 多样性分析、物种组成分析、差异物种分析以及数据可视化

等。本文提供配套视频、分析代码、测序数据和预期结果，以方便同行学习和复现本文的分析过程。

关键词：微生物组，扩增子，QIIME 2, 16S rDNA，可视化

仪器设备

1. (可选) 推荐使用计算服务器 (操作系统：Linux 主流发行版本，如 CentOS 7+/Ubuntu 16.04+；CPU：4 核+；内存：16G+；硬盘：> 10 GB，且大于原始数据大小 3 倍) ，网络访问畅通。
2. 个人电脑推荐 Windows 10 系统，内存 8G+。先在应用商店中安装 Linux 子系统 (如 Ubuntu 20.04 LTS) ，然后安装 QIIME 2；也可使用 VirtualBox 虚拟机运行 QIIME 2 镜像，但效率较低不推荐使用；Mac 系统可直接安装 QIIME 2。(可选) Windows 用户远程访问服务器需安装 XShell 或 Putty 等终端类软件，Mac 使用系统内置终端即可远程访问计算服务器。

软件和数据库

1. QIIME 2 可运行的四种环境任选其一：Linux 服务器 (推荐，适合大数据) 、Windows 10 子系统 Ubutnu 20.04 LTS (推荐，适合小数据) 、Windows 中 VirtualBox 虚拟机中运行 Ubutnu 20.04 LTS (不推荐，小数据集且效率低) 、Mac 系统 (不推荐，兼容性问题较多)
2. 软件管理器 Miniconda3 Linux 64-bit (Python 3.8) ：https://conda.io/miniconda.html
3. QIIME 2 (Bolyen *et al.*, 2019) 发行版 2021.2：https://docs.qiime2.org/
4. GreenGenes 13.8 (McDonald *et al.*, 2011) 物种分类数据库：ftp://greengenes.microbio.me/greengenes_release/gg_13_5/gg_13_8_otus.tar.gz
5. 流程示例参考代码和结果文件详见：https://github.com/YongxinLiu/MicrobiomeProtocol/blob/master/e2.QIIME2/，如示例代码为 QIIME2_Pipeline.sh
6. (可选) 远程文件传输工具 FileZilla 客户端 3.49.1+：https://filezilla-project.org/
7. (可选) Windows 远程访问服务器终端工具 Xshell 6.0.0197p+：https://www.netsarang.com/zh/free-for-home-school/

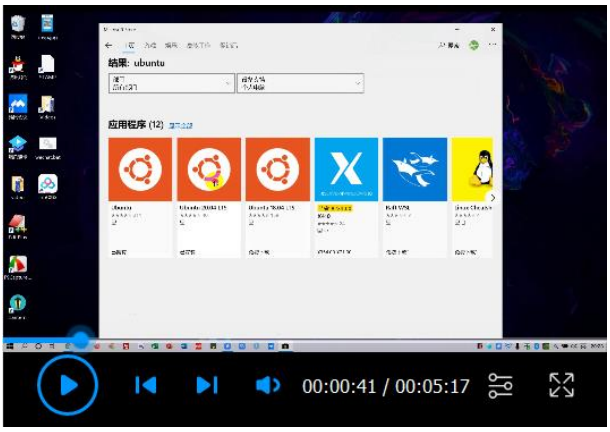
8. (可选) R 语言环境，下载适合自己系统的安装包 （版本：4.0.2+）：https://www.r-project.org/
9. (可选) R 语言开发环境 RStudio，用于执行流程，下载适合自己系统的安装包（版本：1.3.1056+）：https://www.rstudio.com/products/rstudio/download/#download

软件安装和数据库部署

QIIME 2 不支持在 Windows 系统下直接安装。我们主要介绍远程访问 Linux 服务器和 Windows 10 下安装 Linux 子系统并使用 QIIME 2 的两种方法，任选其一即可。

方法 1. 远程访问 Linux 服务器：Windows/Mac 用户安装 FileZilla 客户端，用于上传测序数据至服务器或数据中心，也可用于下载分析结果本地查看。Windows 用户安装 Xshell 用于远程访问服务器并开展分析，Mac 用户可使用系统自带 Terminal 中的 ssh 命令远程访问服务器。

方法 2. Windows 10 的 1609 以后的版本可以安装 Linux 子系统：开始→Microsoft Store→搜索“Ubuntu”→选择“Ubuntu 20.04 LTS”→安装。安装前的系统设置和常见问题请阅读《Windows10 安装 Linux 子系统 Ubuntu 20.04LTS》。安装成功后可以在开始中启动“Ubuntu 20.04 LTS”的命令行，也可选在 RStudio 中设置默认 Terminal 为“Bash (Windows Subsystem for Linux)”，打开新终端即可使用。

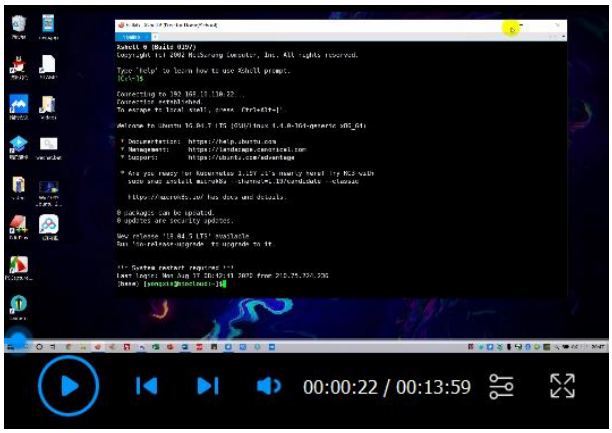


视频 1. Windows 10 中应用商店安装 Ubuntu 20.04 LTS

视频在投稿时已经上传，在线观看链接：https://v.qq.com/x/page/d3138i66xpy.html

在 Linux 系统下，以 Miniconda3 软件和 Python3 虚拟环境安装 QIIME 2 流程；然后下载 16S rRNA 基因数据库，建立物种分类器。

注：下文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。



视频 2. Conda 安装 QIIME 2 和训练分类器

视频在投稿时已经上传，在线观看链接：https://v.qq.com/x/page/s31384uumux.html

1. 安装 Miniconda3 Linux 64-bit (已安装请跳过)

   wget -c https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh

   bash Miniconda3-latest-Linux-x86_64.sh

   ~/miniconda3/bin/conda init

2. 下载 QIIME 2 流程安装软件列表

   直接 wget 下载，但有时无法下载，详见常见问题 1。

   wget -c https://data.qiime2.org/distro/core/qiime2-2021.2-py36-linux-conda.yml

3. Conda 新建环境安装 QIIME 2

   conda env create -n qiime2-2021.2 --file qiime2-2021.2-py36-linux-conda.yml

   进入工作环境

   conda activate qiime2-2021.2

4. GreenGenes 数据库下载并导入

   下载数据库文件 (greengenes)，无法下载或下载慢见常见问题 2

   wget -c ftp://greengenes.microbio.me/greengenes_release/gg_13_5/gg_13_8_otus.tar.gz

   解压

```
tar -zxvf gg_13_8_otus.tar.gz
```

使用 rep_set 文件中的 99_otus.fasta 数据和 taxonomy 中的 99_OTU_taxonomy.txt 数据作为参考物种注释。

导入参考序列

```
qiime tools import \
  --type 'FeatureData[Sequence]' \
  --input-path gg_13_8_otus/rep_set/99_otus.fasta \
  --output-path 99_otus.qza
```

导入物种分类信息

```
qiime tools import \
  --type 'FeatureData[Taxonomy]' \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path gg_13_8_otus/taxonomy/99_otu_taxonomy.txt \
  --output-path ref-taxonomy.qza
```

5. 训练分类器—全长 (通用)，耗时约半小时

```
time qiime feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads 99_otus.qza \
  --i-reference-taxonomy ref-taxonomy.qza \
  --o-classifier classifier_gg_13_8_99.qza
```

如果提示版本错误，详见常见问题 3。

6. (可选) 训练分类器—指定 V 区分类器

如果扩增了指定的 16S 区域，还可以构建特异区域的分类器，可进一步提高分类精度。常用 GreenGenes 13_8 按 99%聚类操作分类单元 (Operational taxonomic units, OTUs) 序列中的 V4 区域 (341F CCTACGGGNGGCWGCAG/805R GACTACHVGGGTATCTAATCC) 构建分类器。确定此处使用的引物与扩增引物保持一致。

本次使用与测试数据对应的 V5 (799F) - V7 (1193R) 引物为例进行提取序列，耗时约 9 分钟。

```
time qiime feature-classifier extract-reads \
  --i-sequences 99_otus.qza \
  --p-f-primer AACMGGATTAGATACCCKG \
  --p-r-primer ACGTCATCCCCACCTTCC \
```

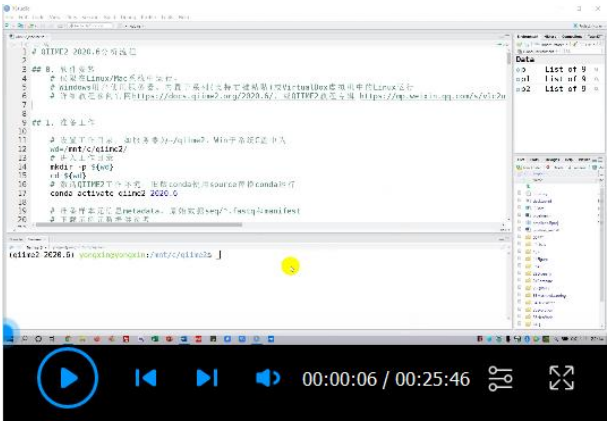
```
  --o-reads ref-seqs.qza
```

基于筛选的指定区段，生成实验特异的分类器，耗时约 8 分钟。

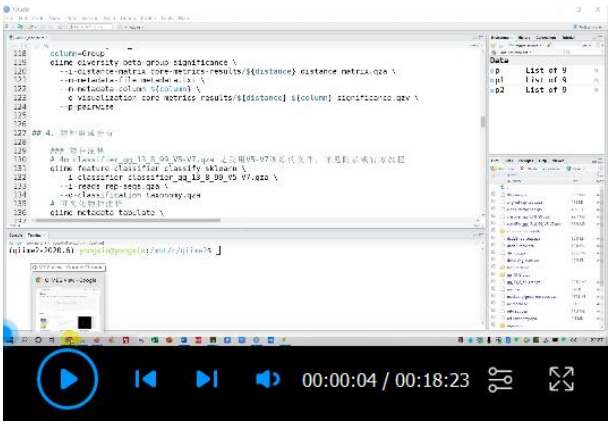
```
time qiime feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads ref-seqs.qza \
  --i-reference-taxonomy ref-taxonomy.qza \
  --o-classifier classifier_gg_13_8_99_V5-V7.qza
```

实验步骤

QIIME 2 分析流程演示和讲解见视频 3 和视频 4。



视频 3. QIIME 2 分析流程 1 - 数据导入、特征表、Alpha 和 Beta 多样性分析 https://v.qq.com/x/page/z3138q86ftr.html



视频 4. QIIME 2 分析流程 2 - 物种组成、差异比较、数据导入和导出 https://v.qq.com/x/page/p31384gityv.html

开始分析前，我们应处于项目所在目录（如 windows 下 C 盘的 qiime2 目录为/mnt/c/qiime2），并启动软件所在的 Conda 环境。

```
mkdir -p qiime2
cd qiime2
conda activate qiime2-2021.2
```

1. 准备和导入元数据和测序数据

下载示例元数据供编写自己课题对应的元数据时作为参考

```
wget -c http://210.75.224.110/github/MicrobiomeProtocol/e2.QIIME2/metadata.txt
```

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 对序列作为本教程的测试数据，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 搜索并手动逐个下载至 seq 目录。可选使用 awk 语言配合 wget 命令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
```

接下来根据样本名和测序文件位置手动编写样本文件的索引列表(manifest)，格式参考流程目录中示例 manifest 文件。可选使用 awk 语言根据 metadata 编写自动生成 mainfest 文件，实现全流程的可重复计算。

```
awk 'NR==1{print "sample-id\tforward-absolute-filepath\treverse-absolute-filepath"}
\
  NR>1{print $1"\t$PWD/seq/"$1"_1.fq.gz\t$PWD/seq/"$1"_2.fq.gz"}' \
  metadata.txt > manifest
```

数据导入 qiime2。

```
qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path manifest \
  --output-path demux.qza \
  --input-format PairedEndFastqManifestPhred33V2
```

导入 1G 大小的 fq 文件用时需 7 分钟，本文中使用测试数据仅需 34 秒。

注：QIIME 2 的起始文件为每个样本 1 个或 1 对文件。对于混池测序未拆分样本的原始数据，需要使用 QIIME (Caporaso et al., 2010) 中的脚本进行拆分，或要求测序服务商提供拆分后的单样本 fastq 格式测序文件。

2. 生成特征表和代表序列

DADA2 是基于 R 语言编写的扩增子分析流程，可以实现扩增子序列去除测序噪音、错误和嵌合体，并挑选扩增序列变体 (amplicon sequence variant, ASV) 和生成特征表 (feature table) 的功能 (Callahan *et al.*, 2016) 。

支持多线程加速，如测试平台 96 线程 (p) 环境下，可以使用--p-n-threads 参数指定。不同线程下的计算时间 (以分钟为单位，m) ：

0 (使用全部) /96 p, 34 m；

24 p, 44 m；

8 p, 77 m；

1 p, 462 m。

```
time qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux.qza \
  --p-n-threads 8 \
```

```
  --p-trim-left-f 29 --p-trim-left-r 18 \
  --p-trunc-len-f 0 --p-trunc-len-r 0 \
  --o-table dada2-table.qza \
  --o-representative-sequences dada2-rep-seqs.qza \
  --o-denoising-stats denoising-stats.qza
```

可使用 dada2 结果并导入主流程，或可选从外部其他流程结果的特征表和代表序列进行导入继续分析，详见常见问题 4。QIIME 2 结果 qza/qzv 文中结果导出方法见常见问题 5。

```
cp dada2-table.qza table.qza
cp dada2-rep-seqs.qza rep-seqs.qza
```

统计特征表。

```
qiime feature-table summarize \
  --i-table table.qza \
  --o-visualization table.qzv \
  --m-sample-metadata-file metadata.txt
```

结果 qzv/qaz 文件可上传到 https://view.qiime2.org/网站查看。QIIME 的结果 qza 为数据文件，qzv 为图表文件，本质上都是 zip 格式的压缩包，也可使用压缩软件解压查看内容，qzv 解压的目录中包括分析结果的网页报告和相关的图表文件。我们先观察每个样本的测序量的总体概述表 (表 1)，用于确定多样性分析的抽平标准化阈值，如本示例样本在特征表中使用的测序数据量最小值为 27,060，即选择最小值；如果最小值和第一分位数差别特别大，则需要结合样本测序量和样本量分布图 (图 1) 选择最小值和第一分位数间的合适数值，尽量保留足够样本量的前提下选择较大的抽平阈值以使用更多的测序数据，注意低于阈值的样本将不会参与多样性分析。更多抽平阈值的交互式选择详见网页中“交互样本细节 (Interactive Sample Detail) ”页面。

注：抽平阈值最小为 1,000 是基于早期 454 测序时代的标准，当前 Illumina 测序通量较大，最小值一般不小于 5,000，推荐 1 万，且越大越好。

表 1. 每个样本的测序量 (Frequency per sample)

分位数	测序量
最小值 (Minimum frequency)	27,060.0
第一分位数 (1st quartile)	28,581.75
中位数 (Median frequency)	30,867.0
第三分位数 (3rd quartile)	32,860.0
最小值 (Maximum frequency)	34,663.0
平均值 (Mean frequency)	30,779.22

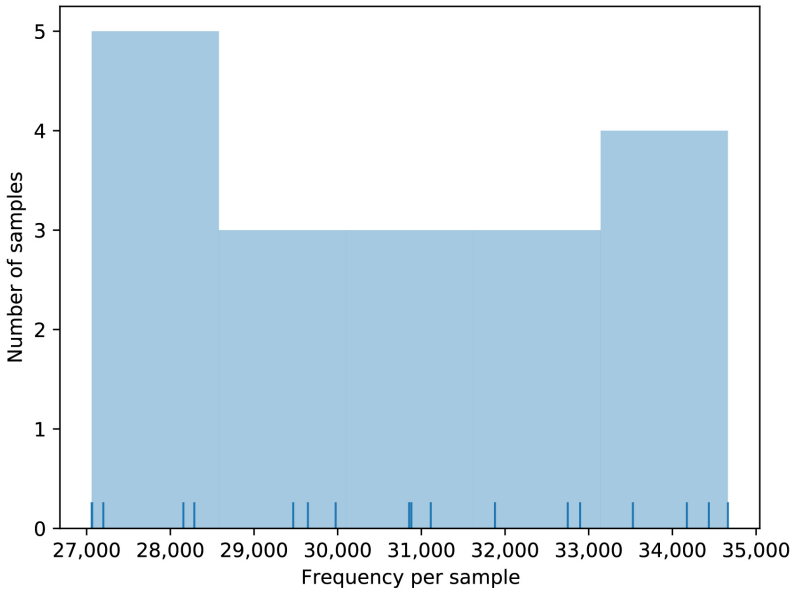


图 1. 样本可用测序量和样本量分布图

X 轴表示样本的可用测序量，轴须线为样本所在位置；Y 轴为样本量分布。本图可指导用户筛选合理的抽平阈值。

统计代表序列。

```
qiime feature-table tabulate-seqs \
  --i-data rep-seqs.qza \
  --o-visualization  rep-seqs.qzv
```

结果可上传至 https://view.qiime2.org/中查看（以 qzv 结尾的文件，我们接下来分析均默认在 QIIME 2 网页预览工具中查看，以后不再赘述），会显示序列长度统计，还有每个特征序列的 ID、长度和序列，其中序列可以点击跳转 NCBI BLAST 显示比对结果。

3. Alpha 和 Beta 多样性分析

构建进化树用于多样性分析

```
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences rep-seqs.qza \
  --o-alignment aligned-rep-seqs.qza \
  --o-masked-alignment masked-aligned-rep-seqs.qza \
  --o-tree unrooted-tree.qza \
  --o-rooted-tree rooted-tree.qza
```

多样性分析，低于重采样深度的样本将会丢弃，通常重采样深度会选择样本测序量最小值以保留较多样本，同时要兼顾保留总体测序量最大化。因此需要根据样本量、数据分布等实际情况选择适合的值尽量使数据利用率最大化，具体见上面自 table.qzv 结果中交互式筛选页面有辅助筛选工具可用。本研究数据分布较均匀，最小值即为最优阈值。最后会在 core-metrics-results 目录生成 4 种常用 Alpha 和 Beta 多样性结果。

```
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth 27060 \
  --m-metadata-file metadata.txt \
  --output-dir core-metrics-results
```

4. Alpha 多样性组间显著性分析和可视化

可选的 alpha 多样性指数有 faith_pd、shannon、observed_features 和 evenness。faith_pd 是综合物种间进化树信息的多样性指数（Faith 1992; Hamady *et al.*, 2010），shannon 是综合丰度和均匀度的指数，observed_features 是丰富度，evenness 是均匀度，中文简介进一步阅读《Alpha 多样性箱线图》，详细的介绍、计算方法和参考文献详见 scikit-bio 文档 (http://scikit-

bio.org/docs/latest/generated/skbio.diversity.alpha.html)

。此处以 observed_features 为例，在之前的版本中被称作 observed_otus。

```
index=observed_features
qiime diversity alpha-group-significance \
  --i-alpha-diversity core-metrics-results/${index}_vector.qza \
  --m-metadata-file metadata.txt \
  --o-visualization core-metrics-results/${index}-group-significance.qzv
```

结果中包括各种多样性指数分布的箱线图（图 2），和基于 Kruskal-Wallis 两两组较的 p-value 和 q-value（表 2），可以下载 svg 格式的矢量图和 tsv 格式的表格，还可以切换列（Column）探索不同分组方式下的分布和统计结果。

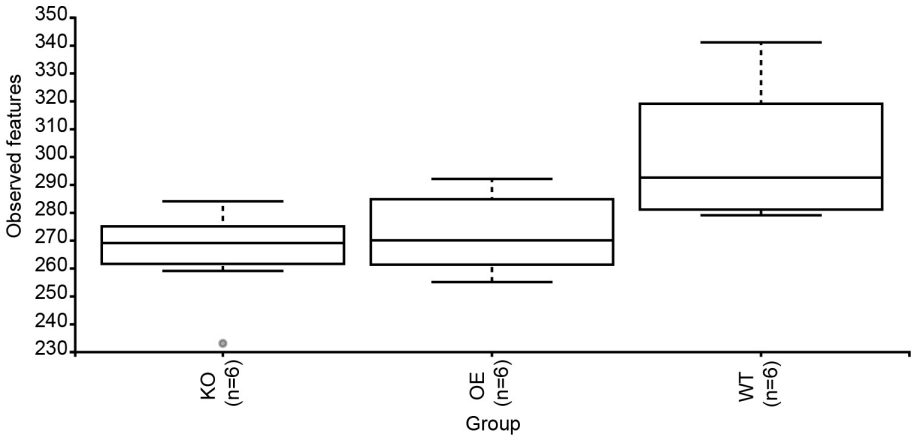


图 2. 各组 Alpha 多样性指数的箱线图分布

X 轴为分组，Y 轴为 observed features 多样性指数值，各组间统计的 p-value 值见表 2。

表 2. Alpha 多样性指数各组间比较的统计结果（Kruskal-Wallis pairwise）

Group 1	Group 2	H	p-value	q-value
KO (n=6)	OE (n=6)	0.231579	0.630356	0.630356
KO (n=6)	WT (n=6)	6.610329	0.010139	0.030417
OE (n=6)	WT (n=6)	3.705263	0.054241	0.081362

5. Alpha 多样性稀释曲线

max-depth 参数通常调置为样本测序量最大值，如果最大值为异常值（outlier），可以在第三分位数和最大值间选择合适的值，请参考 table.qzv 结果选择。

```
qiime diversity alpha-rarefaction \
  --i-table table.qza \
  --i-phylogeny rooted-tree.qza \
  --p-max-depth 34663 \
  --m-metadata-file metadata.txt \
  --o-visualization alpha-rarefaction.qzv
```

结果 alpha-rarefaction.qzv 中有 shannon, faith_pd 和 observed_otus 三种 Alpha 多样性指数可切换，以展示各样本组随测序深度 (Sequencing Depth) 增加对应多样性指数分布的箱线图，图例可点选实现控制分组显示开/关。

6. Beta 多样性组间显著性分析和可视化

可选的 beta 指数有 unweighted_unifrac、bray_curtis、weighted_unifrac 和 jaccard 。UniFrac 是结合特征间进化关系计算群落间距离的方法 (Lozupone *et al.*, 2010)，weighted 和 unweighted 分别是指是否考虑特征的丰度权重。Bray-Curtis (Beals 1984) 是一种生态学常用的距离计算方法。Jaccard 类似于非加权 Bray-Curtis 距离。中文简介进一步阅读《Beta 多样性 PCoA 和 NMDS 排序》，Unifrac 的详细的介绍详见 scikit-bio 文档 (http://scikit-bio.org/docs/latest/generated/skbio.diversity.beta.html) 。

指定 beta 多样性指数和分组用于减少计算量，因为置换检验较耗时。

```
distance=weighted_unifrac
column=Group
qiime diversity beta-group-significance \
  --i-distance-matrix core-metrics-results/${distance}_distance_matrix.qza \
  --m-metadata-file metadata.txt \
  --m-metadata-column ${column} \
  --o-visualization core-metrics-results/${distance}-${column}-significance.qzv \
  --p-pairwise
```

我们先查看 core-metrics-results 目录中的 weighted_unifrac_emperor.qzv，颜色选择分组 Group，此外还可以设置点的透明度、大小、形状等，确定图形样式后，在图右上角设置按钮选择“Save plot” — “SVG+Label”，可以保存主图和图例两个 SVG 格式的矢量图，发表时可使用矢量图编辑软件拼接 (图 3) 。

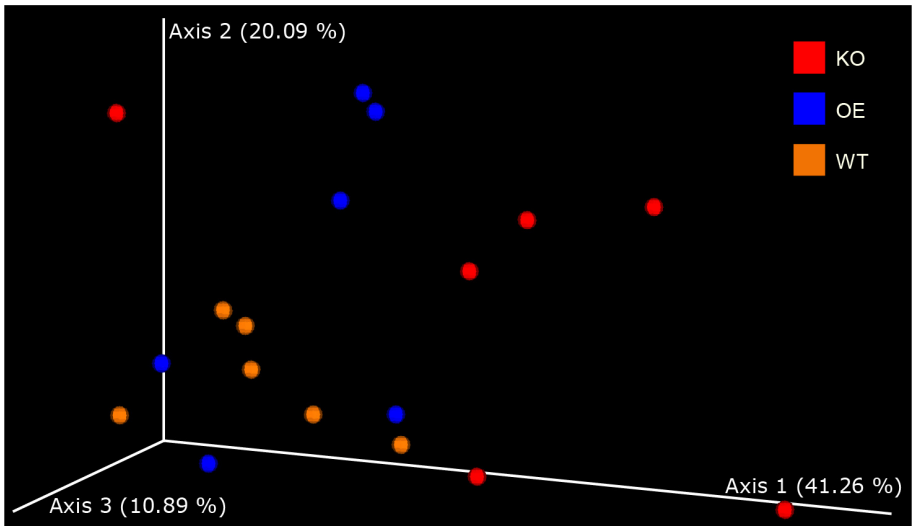


图 3. 基于 Weighted unifrac 距离的主坐标分析图

图中展示前三轴，括号中为解析率。图中点代表样本，颜色对应分组信息，对应关系见右上角图例。

再查看结果 weighted_unifrac-Group-significance.qzv，有组间距离分布图（图 4），还有组间成对 permanova 比较的结果（表 3），可以看到各组间均存在显著差异（p-value < 0.05, q-value < 0.05）。

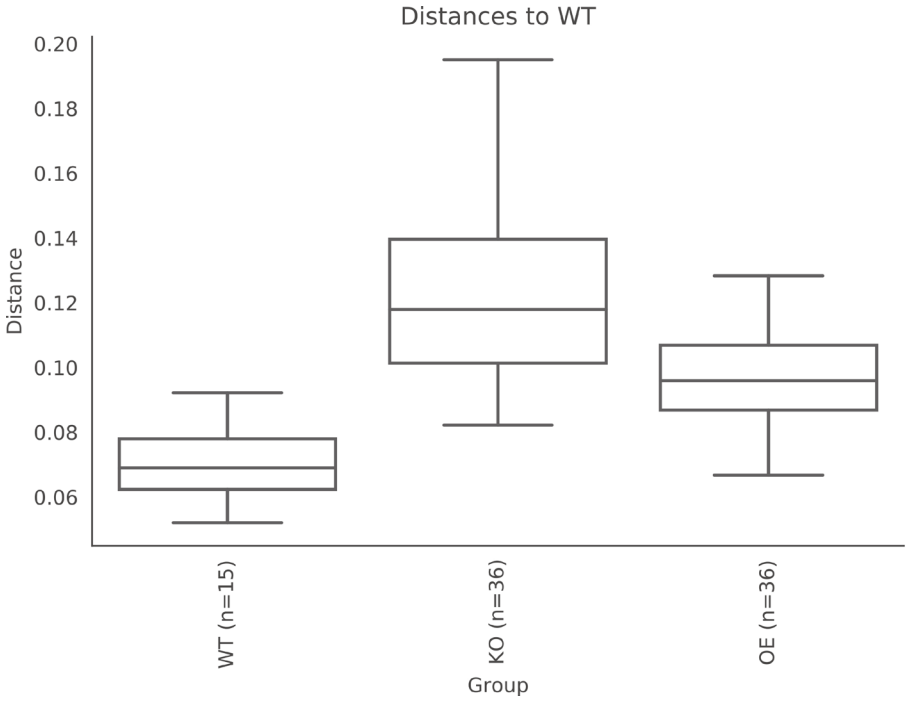


图 4. 相对于 WT 组样本的 Weighted Unifrac 距离分布箱线图

WT 组 6 个样本间非自身距离有 15 种 (6×5÷2)，而组间为 36 种 (6×6)。组间的显著性检验结果见表 3。

表 3. Beta 多样性指数各组间比较的统计结果 (PERMANOVA/ADONIS)

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
KO	OE	12	999	3.938714	0.008	0.012
KO	WT	12	999	4.114816	0.014	0.014
OE	WT	12	999	2.640708	0.008	0.012

7. 物种组成分析

物种注释使用 classifier_gg_13_8_99_V5-V7.qza，这是我用 V5-V7 训练的文件，也可以使用全长序列的训练集，耗时约 4 分钟，具体时间由输入数据和数据库大小决定。如果你使用的研究非此引物，可以使用全长训练集，同时推荐使用实验中采用的引物训练特异的分类器应用于此，详见数据库部署。

```
qiime feature-classifier classify-sklearn \
  --i-classifier classifier_gg_13_8_99_V5-V7.qza \
  --i-reads rep-seqs.qza \
  --o-classification taxonomy.qza
```

可视化物种注释

```
qiime metadata tabulate \
  --m-input-file taxonomy.qza \
  --o-visualization taxonomy.qzv
```

结果文件 taxonomy.qzv 为交互式网页表格，包括特征 ID、分类注释结果和置信度 3 列，可以实现排序和查找等功能。

堆叠柱状图展示 (图 5)

```
qiime taxa barplot \
  --i-table table.qza \
  --i-taxonomy taxonomy.qza \
  --m-metadata-file metadata.txt \
  --o-visualization taxa-bar-plots.qzv
```

结果文件 taxa-bar-plots.qzv 为交互式网页图片（图 5），如可以调整柱宽至中等，切换不同分类级别至 Level2，修改配色方案为 scheeDark2，按分组 Group 排序，再按 Proteobacteria 降序排列，然后保存 SVG 的柱状图和图例。也推荐保存表格的 csv 数据用于个性化绘图。

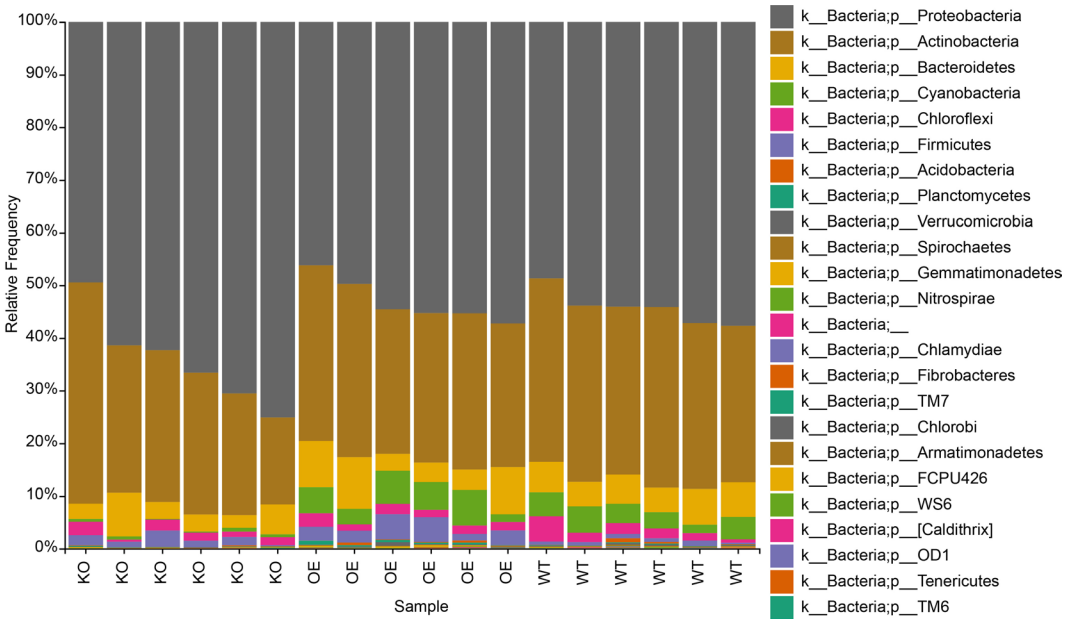


图 5. 堆叠柱状图展示样本门水平组成

图中按分组和变形菌门丰度依次排序。

8. 差异分析 ancom

格式化特征表，添加伪计数。

```
qiime composition add-pseudocount \
  --i-table table.qza \
  --o-composition-table comp-table.qza
```

ancom (Mandal *et al.*, 2015) 计算差异特征，指定分组类型比较，耗时约 7 m。

```
column=Group
time qiime composition ancom \
  --i-table comp-table.qza \
  --m-metadata-file metadata.txt \
  --m-metadata-column ${column} \
```

```
  --o-visualization ancom-${column}.qzv
```

结果 ancom-Group.qzv 用散点图显示显著差异的 ASV。本研究组间差异较小，仅有 1 个显著差异扩增序列变体 (amplicon sequence variants ASV) 为 b46f62815f5bc0f8c18c3c374acabe23，在 taxonomy.qzv 中查找其分类信息为 k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Comamonadaceae; g__Rubrivivax; s__。

扩增子差异分析也经常在属水平进行，结果有名称更利于与专业知识结合进行生物学意义的讨论和规律发现。下面演示按属水平合并，并采用 ancom 统计。

按属水平合并

```
qiime taxa collapse \
  --i-table table.qza \
  --i-taxonomy taxonomy.qza \
  --p-level 6 \
  --o-collapsed-table table-l6.qza
```

格式化特征表，添加伪计数

```
qiime composition add-pseudocount \
  --i-table table-l6.qza \
  --o-composition-table comp-table-l6.qza
```

计算差异属，指定分组类型比较

```
qiime composition ancom \
  --i-table comp-table-l6.qza \
  --m-metadata-file metadata.txt \
  --m-metadata-column ${column} \
  --o-visualization l6-ancom-${column}.qzv
```

结果 l6-ncom-Group.qzv 用散点图显示显著差异的属，本示例中只有 1 个显著差异的属位于图中右上角为 k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Methylocystaceae;g__Pleomorphomonas (图 6) 。

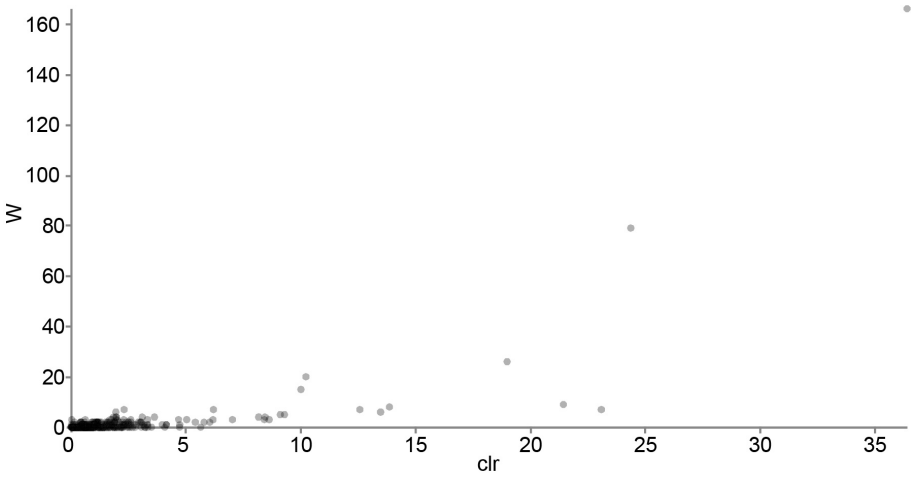


图 6. 属水平 ancom 差异分析散点图

常见问题

1. QIIME 2 流程安装软件列表无法下载

   QIIME 2 安装清单仅有几 KB 大小，包含有四百多个软件，通过 4 个 Conda 频道安装，但因保存在 Google 服务器导致国内下载困难。我提供了两个备用下载链接可选，一是 Github 中帮助文档的备用下载链接，可在 https://github.com/YongxinLiu/QIIME2ChineseManual 中下载；二是使用备用链接 http://210.75.224.110/github/QIIME2ChineseManual/2021.2/qiime2-2021.2-py36-linux-conda.yml 下载。

2. 参考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。

3. 分类器使用提示 scikit-learn 版本不兼容

   进行物种分类时，有时会显示如下错误：

   Plugin error from feature-classifier:

The scikit-learn version (0.21.2) used to generate this artifact does not match the current version of scikit-learn installed (0.22.1) . Please retrain your classifier for your current deployment to prevent data-corruption errors.

Debug info has been saved to /tmp/qiime2-q2cli-err-5ngzk2hm.log

可能是软件包升级导致版本不兼容，可重新运行软件安装和数据库部署中的 5 或 6 重新构建分类器即可。

4. 外部导入特征表和代表序列

其他常用扩增子分析流程如 Mothur (Schloss *et al.*, 2009) 、QIIME (Caporaso *et al.*, 2010) 、USEARCH (Edgar 2010) 、VSEARCH (Rognes *et al.*, 2016) 和 DADA2 (Callahan *et al.*, 2016) 等分析流程的特征表和代表序列结果也可以导入 QIIME 2 继续分析。需要准备特征表 (otutab.txt) 和代表序列 (otus.fa) 两个文件，示例在本文 github 中或以下链接下载。特征表通用的 BIOM 格式可以直接导入 QIIME 2 (Bolyen *et al.*, 2019) ，如果制表符分隔的纯文本格式需要使用 biom 命令转换为 BIOM 格式再导入 (McDonald *et al.*, 2012) 。

wget -c http://210.75.224.110/github/MicrobiomeProtocol/e2.QIIME2/otutab.txt

wget -c http://210.75.224.110/github/MicrobiomeProtocol/e2.QIIME2/otus.fa

转换文本为 Biom1.0，注意 biom --version 2.1.5/8 可以，2.1.7 可能报错

```
biom convert -i otutab.txt -o    otutab.biom \
  --table-type="OTU table" --to-json
```

导入特征表

```
qiime tools import --input-path otutab.biom \
  --type 'FeatureTable[Frequency]' --input-format BIOMV100Format \
  --output-path table.qza
```

导入代表序列

```
qiime tools import --input-path otus.fa \
  --type 'FeatureData[Sequence]' \
  --output-path rep-seqs.qza
```

5. 导出特征表、代表序列和物种注释

导出特征表为 biom 格式

```
qiime tools export \
  --input-path table.qza \
```

```
  --output-path feature-table
```

转换 biom 格式特征表为 tsv 格式

```
biom convert -i feature-table/feature-table.biom \
  -o feature-table/feature-table.txt \
  --to-tsv
```

删除多余注释行

```
sed -i '/# Const/d' feature-table/feature-table.txt
```

导出代表序列

```
qiime tools export \
  --input-path rep-seqs.qza \
  --output-path rep-seqs
```

导出物种注释

```
qiime tools export \
  --input-path taxonomy.qza \
  --output-path taxonomy
```

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)]。QIIME 2 软件流程的文章发表于 Nature Biotechnology 杂志（Bolyen *et al.*, 2019），官方最新版用户手册请访问：https://docs.qiime2.org/。本文作者负责流程中文用户手册的翻译工作，链接：https://github.com/YongxinLiu/QIIME2ChineseManual。此外，本流程在最近发表的综述中被推荐和概述（刘永鑫等, 2019; Liu *et al.*, 2020）。

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. Beals, E. W. (1984). Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Adv Ecol Res.* A. MacFadyen and E. D. Ford, Academic Press. 14: 1-55.

3. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., *et al.* (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37(8): 852-857.
4. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13(7): 581-583.
5. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5): 335-336.
6. Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19): 2460-2461.
7. Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol Conserv* 61(1): 1-10.
8. Hamady, M., Lozupone, C. and Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 4(1): 17-27.
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).
10. 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.
11. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. and Knight, R. (2010). UniFrac: an effective distance metric for microbial community comparison. *ISME J* 5: 169.
12. Mandal, S., Van Treuren, W., White, R. A., Eggesbo, M., Knight, R. and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 26: 27663.

13. McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. and Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* 1(1).
14. 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.
15. Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584.
16. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23): 7537-7541.
17. 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*. *Genomics Proteomics Bioinformatics* 15(1): 14-18.

请通过以下链接下载视频：

视频 1：

https://os.bio-protocol.org/doc/upprotocol/p3554/Abstract3554_20200822052233106/Win10%E4%B8%ADUbuntu20.04%E7%9A%84%E5%AE%89%E8%A3%85%E5%92%8C%E4%BD%BF%E7%94%A8%E6%96%B9%E6%B3%95.wmv

视频 2：

https://os.bio-protocol.org/doc/upprotocol/p3554/Abstract3554_20200822053044222/Conda%E5%

AE%89%E8%A3%85QIIME%202%E5%92%8C%E8%AE%AD%E7%BB%83%E5%88%86%E7%B1%BB%E5%99%A8.wmv

视频 3：

https://os.bio-protocol.org/doc/upprotocol/p3554/Abstract3554_20200822053823692/QIIME%202%E5%88%86%E6%9E%90%E6%B5%81%E7%A8%8B1%E6%95%B0%E6%8D%AE%E5%AF%BC%E5%85%A5%E3%80%81%E7%89%B9%E5%BE%81%E8%A1%A8%E3%80%81Alpha%E5%92%8CBeta%E5%A4%9A%E6%A0%B7%E6%80%A7%E5%88%86%E6%9E%90.wmv

视频 4：

https://os.bio-protocol.org/doc/upprotocol/p3554/Abstract3554_20200822054325371/QIIME%202%E5%88%86%E6%9E%90%E6%B5%81%E7%A8%8B2%E7%89%A9%E7%A7%8D%E7%BB%84%E6%88%90%E3%80%81%E5%B7%AE%E5%BC%82%E6%AF%94%E8%BE%83%E3%80%81%E6%95%B0%E6%8D%AE%E5%AF%BC%E5%85%A5%E5%92%8C%E5%AF%BC%E5%87%BA.wmv