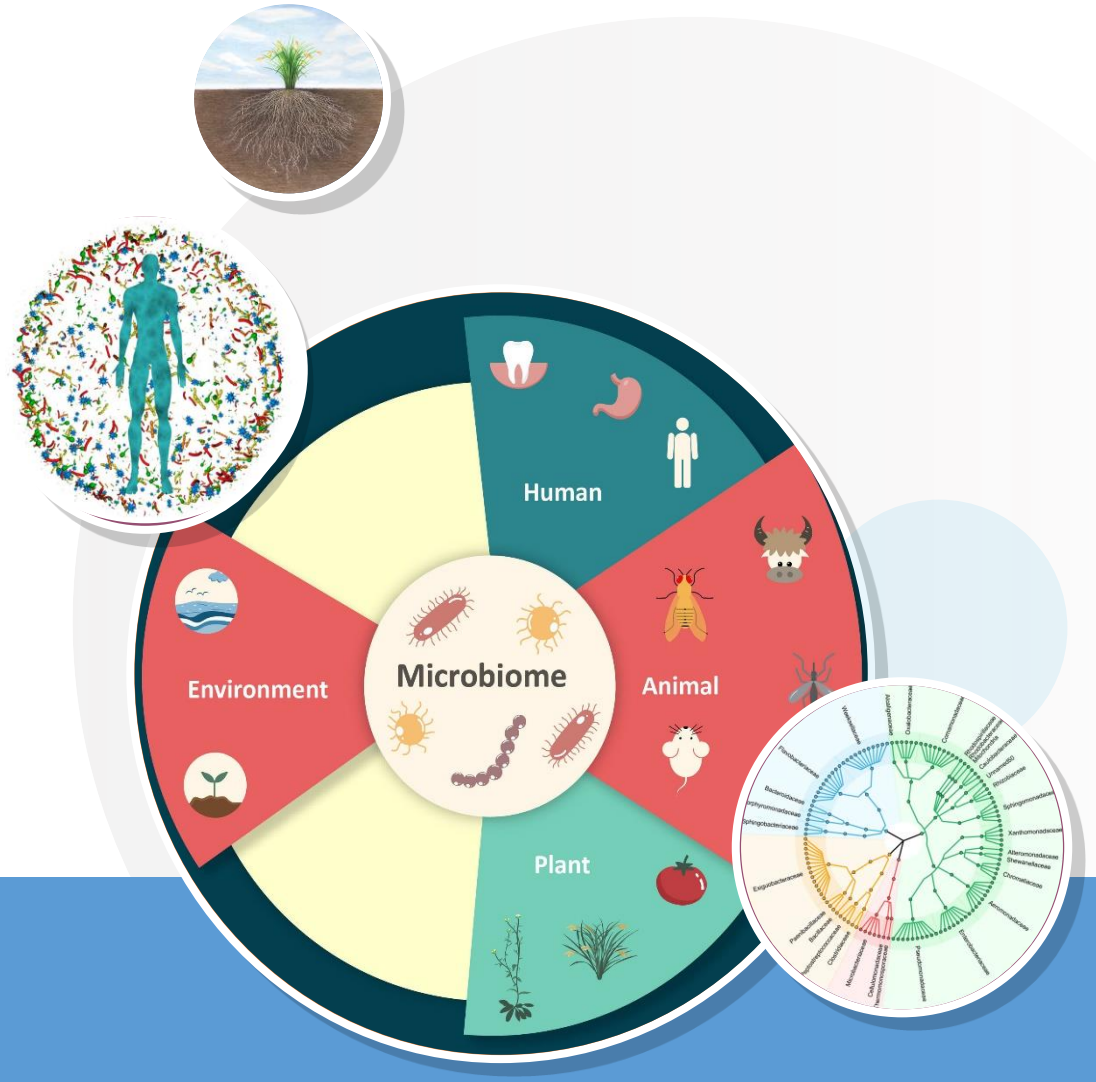


热心肠 热心肠研究院

宏基因组
meta-genome

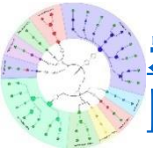
中国肠道大会 —《宏基因组》培训

2扩增子分析流程

刘永鑫
中科院遗传发育所 高级工程师
宏基因组公众号 创始人
2021年5月27日

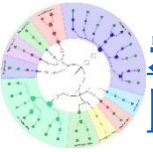
目录

- 分析的硬件要求
- 常用分析软件
- 扩增子测序简介
- 常用扩增子分析软件和数据库
- 扩增子分析的基本思路
- 易用的扩增子分析流程EasyAmplicon
- 最流行的扩增子分析流程QIIME 2

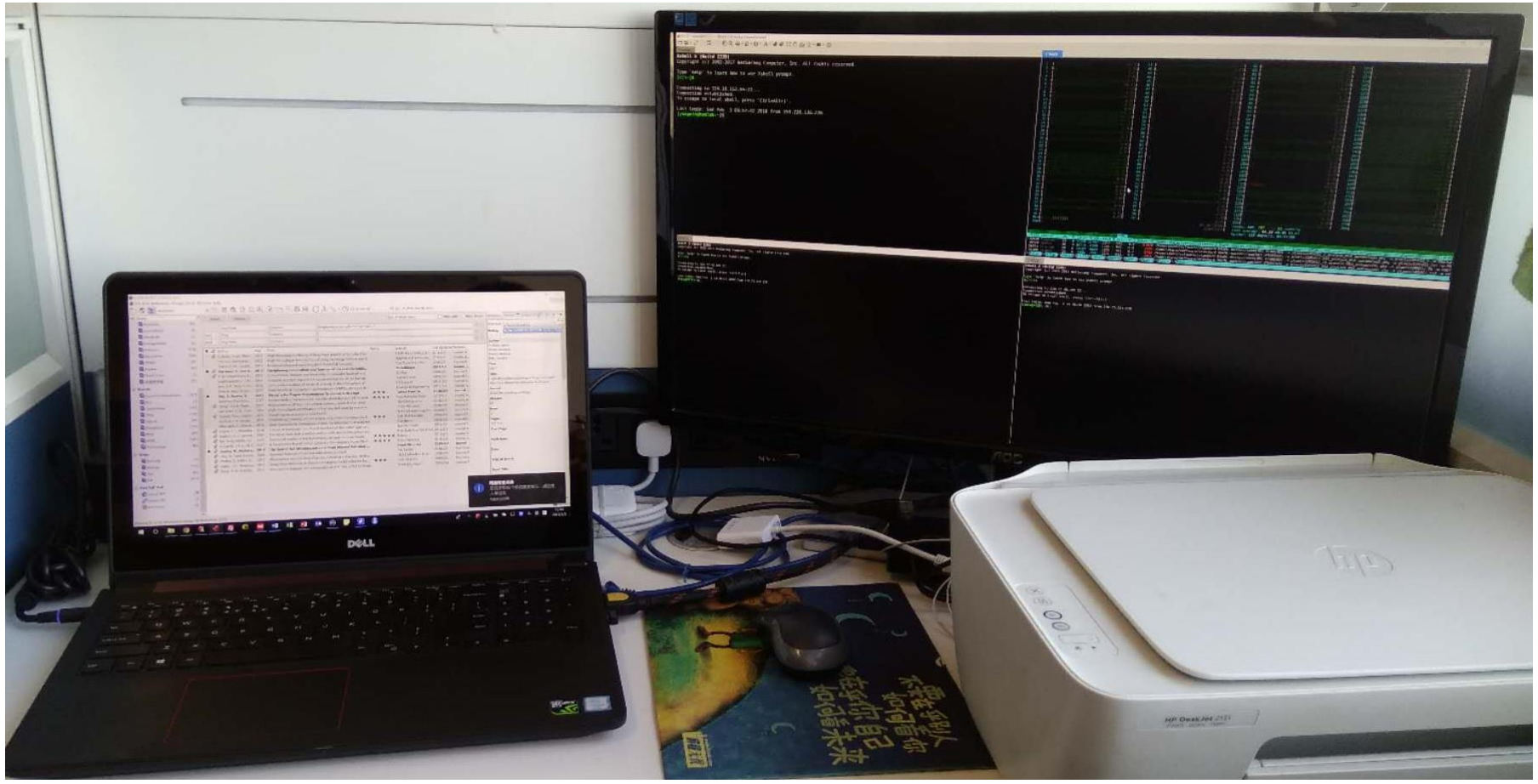


目录

- **分析的硬件要求**
- 常用分析软件
- 扩增子测序简介
- 常用扩增子分析软件和数据库
- 扩增子分析的基本思路
- 易用的扩增子分析流程EasyAmplicon
- 最流行的扩增子分析流程QIIME 2



硬件——笔记本+显示器



双显示器方便多任务管理、阅读文献和多图比较

宏基因组

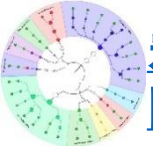
硬件——服务器/集群

服务器

集群

目录

- 分析的硬件要求
- **常用分析软件**
- 扩增子测序简介
- 常用扩增子分析软件和数据库
- 扩增子分析的基本思路
- 易用的扩增子分析流程EasyAmplicon
- 最流行的扩增子分析流程QIIME 2



宏基因组

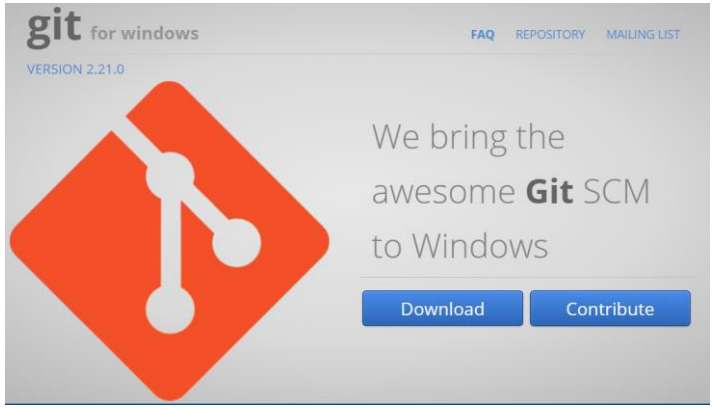
常用分析软件

- 数据分析环境Shell + R + IDE：GitForWindows、R、RStudio+R包

- 扩增子分析流程：USEARCH & VSEARCH，Win子系统+QIIME 2
- 辅助工具：序列工具seqkit、表格工具csvtk、并行管理rush
- 差异分析和可视化：STAMP
- 网络分析及可视化：Cytoscape、Gephi
- 图片排版：Adobe Illustrator
- 登录服务器：XShell 或PuTTY；上传下载文件：Filezilla 或 WinSCP

分析环境：GitForWindows(仅Win)

- 提供Windows下运行Shell命令的环境，可在RStudio的Terminal中使用
- 官网：http://gitforwindows.org/ ，点击Download下载最新版
- 在文件夹下安装预下载的 Git-2.30.2-64-bit.exe，按照默认参数右键管理员安装即可。（若不能调用，则设置环境变量C:\Program Files\Git\usr\bin）
- 具体使用见：Windows轻松实现linux shell环境：gitforwindows

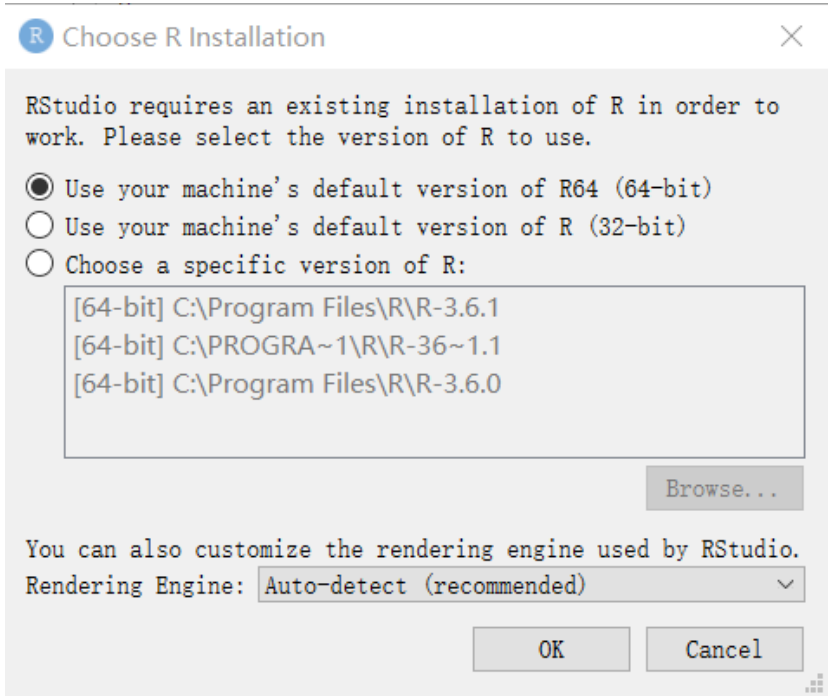


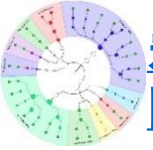
统计和可视化环境：R语言

- R语言是目前生物学、经济学等领域最流行的统计分析语言
- 官网：https://www.r-project.org/ 下载最新版：Downad CRAN - China Tsinghua - Download R for Windows(Mac) —— base —— Download R 4.x.x
- 双击安装程序，建议语言选择英文安装。注意：选择组件步可去掉32-bit，节约空间并减少选择。

R/Shell编程环境——RStudio

- 下载页面：https://www.rstudio.com/products/rstudio/download/#download
- 选择适合自己系统的版本(Windows 10/8或macOS 10.13+)，下载安装程序的最新版
- **右键使用管理员**身份安装
- 完成后打开时，如存在R多版本会提示选择
  - 系统允许下建议选第一项 -
  - “使用系统默认R64位版本”
  - 点击OK，默认为使用安装的最新版

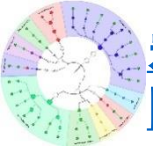




宏基因组

我常用的脚本和数据库

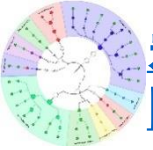
- 下载github仓库： https://github.com/yongxinliu/db
  - 方法1. 网页中点“Code”—— Download ZIP，下载后解压
  - 方法2. 命令行中 git clone git@github.com:YongxinLiu/db.git
- 如何使用脚本？
  - 将下载的db目录复制到windows的c盘(/c)，或Linux/Mac家目录(~)
  - 添加可执行程序至环境变量(以Windows中RStudio中Terminal为例)
  - export PATH=$PATH:/c/db/win/
  - 使用前设置目录变量，方便以后多次使用
  - sd=/c/db/script
  - R语言绝对路径使用R脚本
  - Rscript ${sd}/alpha_boxplot.R -h

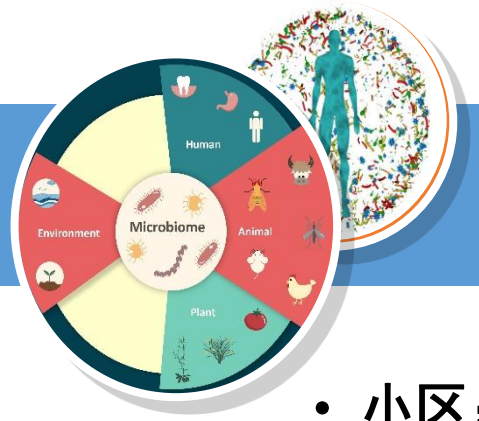


宏基因组

目录

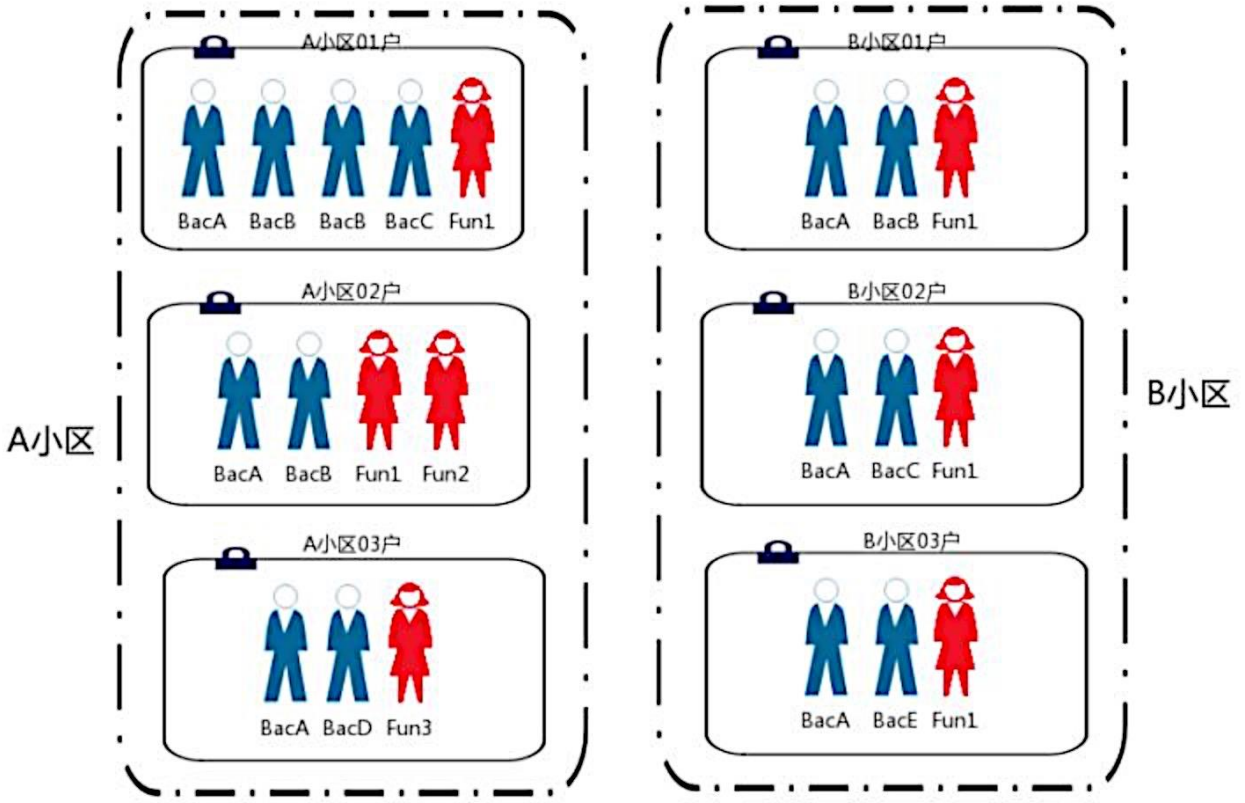
- 分析的硬件要求
- 常用分析软件
- **扩增子测序简介**
- 常用扩增子分析软件和数据库
- 扩增子分析的基本思路
- 易用的扩增子分析流程EasyAmplicon
- 最流行的扩增子分析流程QIIME 2

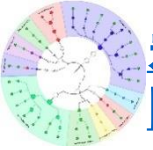




扩增子分析类似一次人口普查

- 小区：实验组
- 家户：样品
- 男生：样品中的细菌
  - BacA：北京人
  - BacB：山东人
  - BacB：山东人
  - BacC：东北人
  - “省份”这一规则进行分类
- 女生：样品中的真菌



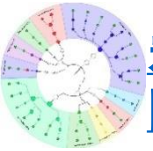


宏基因组

物种注释——相当于地址

- 界（Kingdom）、门（Phylum）、纲（Class）、目（Order）、科（Family）、属（Genus）、种（Species）
- 动物界、脊索动物门、哺乳纲、食肉目、熊科、大熊猫属、大熊猫
- 动物界、脊索动物门、哺乳纲、灵长目、人科、人属、智人种

- 国、省、市、县、镇、村、屯
- 中国、黑龙江省、哈尔滨市、五常县、冲河镇、三家子村、大排地屯



宏基因组

https://baike.baidu.com/item/%E7%95%8C%E9%97%A8%E7%BA%B2%E7%9B%AE%E7%A7%91%E5%B1%9E%E7%A7%8D

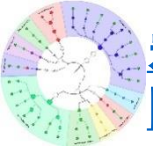
扩增子测序和分析流程

提取DNA

扩增
测序

质控、(聚类)
去噪、定量

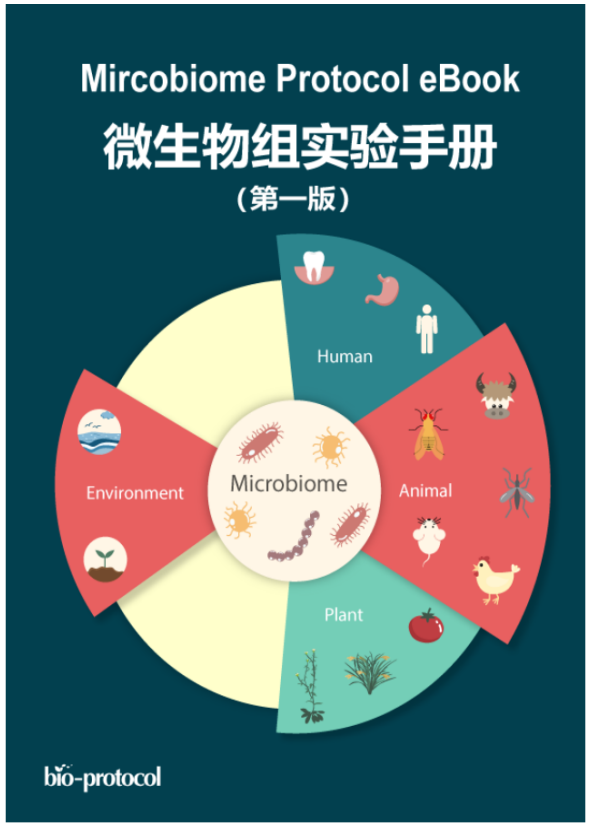
多样性分析



宏基因组

•Protein Cell：扩增子和宏基因组数据分析实用指南

《微生物组实验手册》



微生物组实验手册征稿中

Bio-protocol中国编辑部联合宏基因组公众号共同发起微生物组实验方法电子书(Microbiome Protocol eBook)项目，旨在促进微生物组领域国内外华人科研团队之间的交流和合作，倡导科研团队注重实验方法的沉淀、分享与传播。希望本电子书填补微生物组领域方法空白，解决实验和分析难重复的问题，推动实验标准化，为积累标准统一的数据和未来大数据整合分析做准备，助力微生物组学研究的发展。

eBook主题为微生物组(Microbiome)，包括培养组、扩增子、宏基因组、宏转录组、宏代谢组、单菌基因组、相关分子生物学和微生物学实验、以及微生物组学上下游相关实验和分析技术等。按研究对象分类主要包括人、动物、植物、环境、通用、土壤、水体、细菌、真菌、病毒等。按研究方法分类：主要包括样本制备、核酸提取、蛋白和代谢物提取、测序文库制备、微生物培养与鉴定、合成菌群、无菌实验、数据分析、微生物学常用实验和分析等。

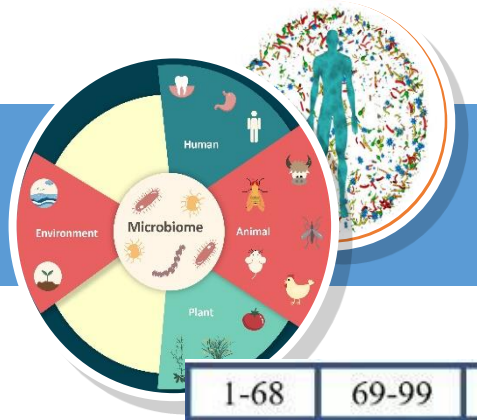
为了提高本实验手册的质量以及方法的多样化，我们诚邀更多国内外优秀华人同行参与本项目。欢迎您的来稿！ 征稿主页：https://bio-protocol.org/bio101/Special_Issue_info.aspx?siid=48
进展： https://kdocs.cn/l/cL8RRqHIL

科学顾问
朱永官 中科院城环境所
刘双江 中科院微生物所
朱宝利 中科院微生物所

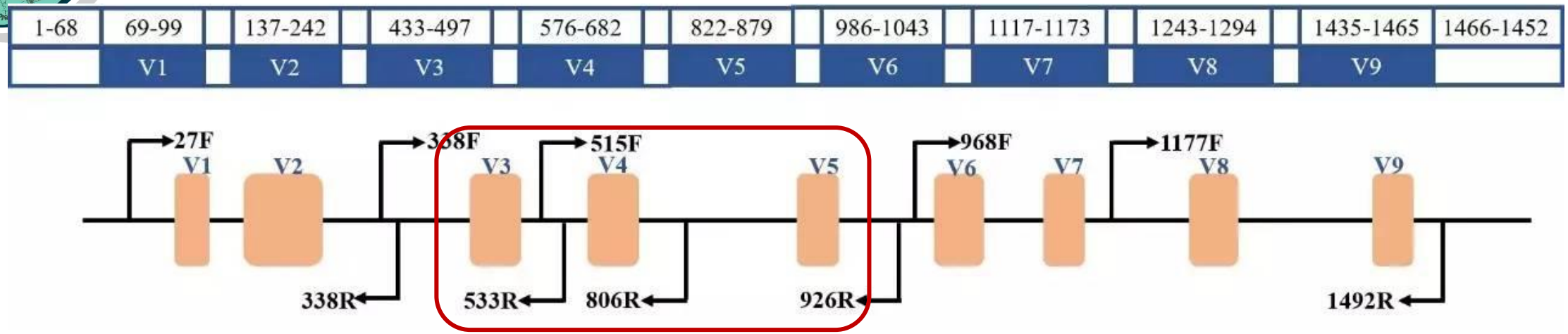
特邀主编
褚海燕 中科院南京土壤所
刘永鑫 中科院遗传发育生物所

特邀编委
朱怀球 北京大学 杨瑞馥 军事医学科学院
周宏伟 南方医科大学 成艳芬 南京农业大学
韦中 南京农业大学 邓晔 中科院生态中心
杨军 中科院城环境所 白洋 中科院遗传发育生物所
袁志林 中国林科院 李猛 深圳大学

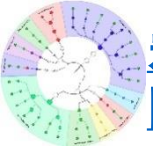
宏基因组



16S rDNA结构与引物的选择



- 所有活着的生物都有核糖体RNA
- rRNAs在蛋白翻译中起至关重要的作用
- rRNAs相对保守，且较少发生水平转移
- 有分子钟的特征，进化分析中非常有用



宏基因组

建库原理：主要以两轮PCR为主

Marker genes

第一轮PCR

扩增目标并标记样品barcode

barcode

双向Barcode大大增加混样数量

如48 X 96 = 4608个样本

第二轮PCR

标记文库Index和测序接头

Index

宏基因组

Microbiome：HiSeq平台16S扩增子文库构建方法

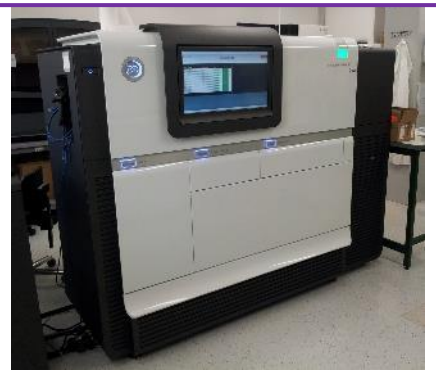
Plant Com：绝对定量检测宿主微生物组的HA-QAP技术简介* 全文解读

本领域常用测序平台

一代测序

Sanger

三代测序



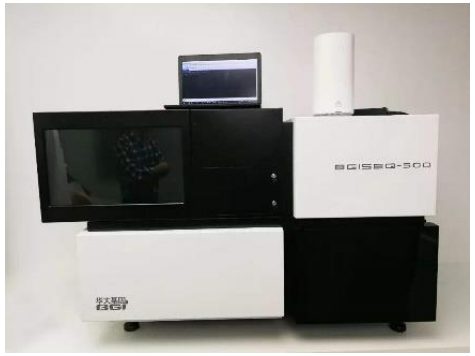
Pacific Biosciences

Nanopore

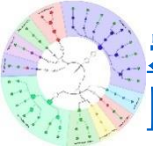
二代测序

Illumina *Seq

Ion Torrent



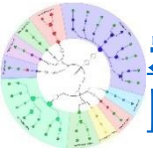
BGISEQ



宏基因组

扩增子混样测序

- MiSeq 数据产量~20M paired-end(PE) 300 bp读长, 共12 Gb, 可以多达200+样本测序(按平均10万条计算)，HiSeq2500单个Lane产量最高160M PE 250 bp, 80 Gb，上千样本混测，**NovaSeq6000，可产出800 M PE 250 bp reads，400Gb**，近万样本混测。
  - 样品的测序量由样品的复杂度决定，简单少测，复杂多测
- 序列唯一的DNA barcodes加入样品中，用于区分在同一个文库中测序的样品
- 扩增序列具有高度同质性
  - 需要将不同来源，或标记(marker)基因的扩增产物混合测序；或添加PhiX序列增加测序反应中的多样性



宏基因组

扩增子结构

barcode

Primer F　　amplicon　　Primer R

Primer：在16S/ITS/18S保守区设计的引物，用于扩增rDNA的部分高变区

Amplicon:扩增的部分 rDNA

Illumina Seq系列下机双端序列

双端反向合并为单端

切下barcodes序列

按barcode重命名序列

切除双端引物

Clean amplicon

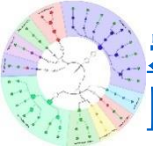
Xu-Bo Qian, **Tong Chen**, Yi-Ping Xu, Lei Chen, Fu-Xiang Sun, Mei-Ping Lu & **Yong-Xin Liu**. A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. ***Chin. Med. J.*** (**2020**).

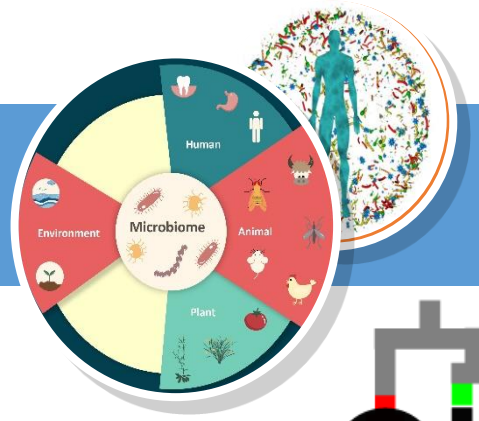
宏基因组

CMJ：人类微生物组研究设计、样本采集和生物信息分析指南

目录

- 分析的硬件要求
- 常用分析软件
- 扩增子测序简介
- **常用扩增子分析软件和数据库**
- 扩增子分析的基本思路
- 易用的扩增子分析流程EasyAmplicon
- 最流行的扩增子分析流程QIIME 2





扩增子分析软件和数据库

Quantitative Insights Into Microbial Ecology
qiime

USEARCH
Ultra-fast sequence analysis

mothur

GREENGENES
The 16S rRNA Gene Database and Tools

rdp

unite
community

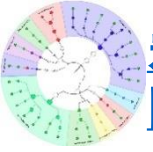
silva
high quality ribosomal RNA databases

16S、18S和ITS分析流程

16S、18S、ITS数据库

扩增子分析神器USEARCH 简介

微生物扩增子数据库大全
NAR：UNITE真菌鉴定ITS数据库



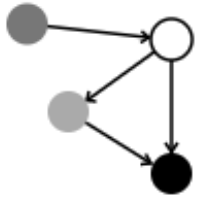
宏基因组

QIIME1/2———最流行

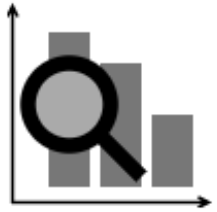
qiime2

QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible,

Learn more



Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!



Interactively explore your d beautiful visualizations that pro perspectives.

QIIME

About 40,800 results (**0.05** sec)

QIIME allows analysis of high-throughput community sequencing data

JG Caporaso, J Kuczynski, J Stombaugh, K Bittinger... - Nature ..., 2010 - nature.com

To the Editor: High-throughput sequencing is revolutionizing microbial ecology studies. Efforts like the Human Microbiome Projects 1 and the US National Ecological Observatory Network 2 are helping us to understand the role of microbial diversity in habitats within our ...

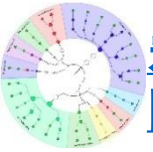
☆ 99 Cited by 24628 Related articles All 17 versions

Reproducible, interactive, scalable and extensible microbiome data science using **QIIME** 2

E Bolyen, JR Rideout, MR Dillon, NA Bokulich... - Nature ..., 2019 - nature.com

Conclusions Because of a clear trend toward more engagement and transparency with research participants, we should expect more research participants to exercise their HIPAA access right in coming years. The committee's recommendations ensure that researchers ...

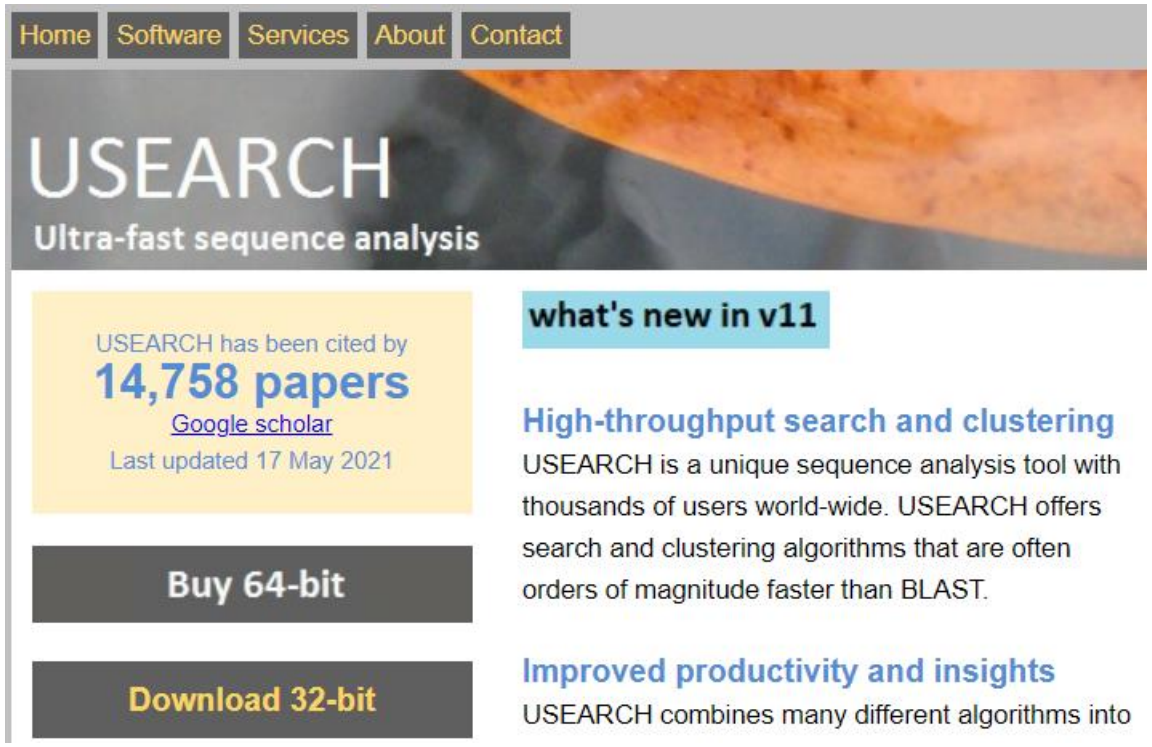
☆ 99 Cited by 2296 Related articles All 30 versions



宏基因组

- NBT：QIIME 2可重复、交互式的微生物组分析平台
- 1简介和安装Introduction&Install

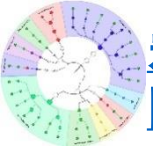
USEARCH——最好用



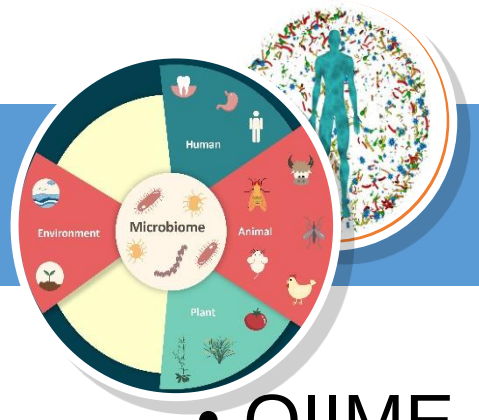
http://www.drive5.com/usearch/

- 由于USEARCH即好用，但收费，出现了模仿者VSEARCH，方便大家免费使用。
- 有多平台版本，轻松分析扩增子
- 想免费分析大数据的有福了

[HTML] **VSEARCH**: a versatile open source tool for metagenomics
T Rognes, T Flouri, B Nichols, C Quince, F Mahé - PeerJ, 2016 - peerj.com
Background **VSEARCH** is an open source and free of charge multithreaded 64-bit tool for processing and preparing metagenomics, genomics and population genomics nucleotide sequence data. It is designed as an alternative to the widely used USEARCH tool for which
☆ 99 Cited by 3115 Related articles All 22 versions »

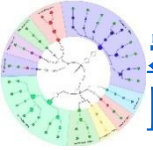


宏基因组



易扩增子(EasyAmplicon)

- 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
- 易扩增子实现简单易用、可重复和跨平台地开展扩增子分析
- 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免费版限制
- 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
- 提供数10个脚本，实现特征表过滤、重采样、分组均值等常用计算，并为STAMP、LEfSe、PICRUSt1/2等提供标准的输入文件

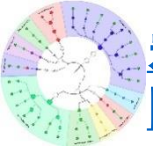


宏基因组

项目：https://github.com/YongxinLiu/EasyAmplicon
教程：MPB：易扩增子：易用、可重复和跨平台的扩增子分析流程

目录

- 分析的硬件要求
- 常用分析软件
- 扩增子测序简介
- 常用扩增子分析软件和数据库
- **扩增子分析的基本思路**
- 易用的扩增子分析流程EasyAmplicon
- 最流行的扩增子分析流程QIIME 2



数据分析的基本思想——三步走

大数据 → 大表 → 小表 → 图

```
@HISEQ:549:HLYNYBCXY:1:1101:1267:2220 1:N:0:CACTCAAT
TCGTCGCTCGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGTTGGGCG
+
DDDDDIHHHIIIIIIIIHIIHIIIIIIIIIIHIHIHIIIIIIIIIIIIIIIIIII
@HISEQ:549:HLYNYBCXY:1:1101:1887:2204 1:N:0:CACTCAAT
TACGAGTATGAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTCTA
+
DDDD@H<GHIIIIIIIIIIIIIIIIIIIIHIIIHIIIIIIIIIIGIIIIIIIIFH
@HISEQ:549:HLYNYBCXY:1:1101:2196:2168 1:N:0:CACTCAAT
TCGTCGCTCGAACAGGATTAGATACCCTGGTAGTCCACGCCATAAACGATGACAA
+
DDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIII
@HISEQ:549:HLYNYBCXY:1:1101:2025:2183 1:N:0:CACTCAAT
ATATCGCGAGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGCG
+
DDDD@E@HIGHIIIHHFHIIHIIIIFHHIIIHHGIHIIHIIICHDEHHIIIIHGH
@HISEQ:549:HLYNYBCXY:1:1101:2052:2198 1:N:0:CACTCAAT
CACGAGACAGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGGGTA
+
D@DD@H=?CCHHIIIIIIIIIIIIIIIIIIIIIIIIIIGI0CHIIIIIHIHIIH@
```

序列: 10^6 ~ 10^9

ID	WT6	WT3	OE4	WT2	OE3	WT1
OTU_265	18	18	6	11	20	15
OTU_36	63	77	57	194	155	163
OTU_102	20	44	18	77	18	43
OTU_49	106	92	25	137	76	65
OTU_270	9	5	22	5	22	5
OTU_1865		0	3	0	0	2
OTU_58	77	75	28	84	53	64
OTU_1110		6	3	3	2	2
OTU_30	100	142	78	111	124	145
OTU_51	87	79	21	38	42	102
OTU_1353		0	1	2	0	1
OTU_1137		0	1	0	3	0
OTU_18	166	150	126	318	130	265
OTU_4	498	343	189	804	224	626
OTU_3	459	690	340	1039	568	580
OTU_704	3	14	12	8	9	4
OTU_14	176	283	110	314	169	232

特征表: $10^{1\sim3}$ X $10^{3\sim5}$

Sample	berger_parker	buzas_gibson		chao1
WT6	0.042	0.0381	1388.9	0.992	0.817
WT3	0.0453	0.0425	1474.9	0.992	0.828
OE4	0.0359	0.0414	1476.4	0.993	0.828
WT2	0.0642	0.0244	1203.0	0.985	0.773
OE3	0.0426	0.0396	1716.9	0.991	0.807
WT1	0.0586	0.0293	1317.0	0.988	0.788
WT4	0.0518	0.0359	1353.2	0.991	0.813
OE5	0.0361	0.0441	1622.8	0.993	0.824
OE2	0.0466	0.0472	1733.3	0.992	0.827
OE6	0.0432	0.0523	1759.5	0.994	0.840
WT5	0.0435	0.0252	1181.6	0.987	0.776
OE1	0.0374	0.0524	1591.2	0.994	0.852
KO4	0.0558	0.0325	1474.1	0.990	0.796
KO1	0.0552	0.0409	1651.6	0.990	0.813
KO5	0.0732	0.025	1306.2	0.986	0.772
KO2	0.0509	0.0445	1675.3	0.992	0.825
KO3	0.0571	0.0329	1489.8	0.990	0.800
KO6	0.0518	0.0334	1215.9	0.991	0.813

统计表: 1~N X $10^{1\sim3}$

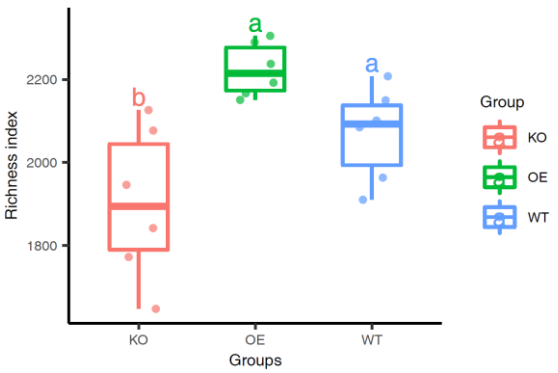


图: $10^{1\sim3}$个点和统计信息

宏基因组

Protein Cell：扩增子和宏基因组数据分析实用指南

刘永鑫：想学菌群生物信息分析-21分钟带你入门

扩增子分析流程

原始序列
(fastq)

实验设计
(mapping file)

输入

1) 序列预处理:
拆分样本、切除引物和标签、质控、去宿主/嵌合

2) 挑选OTU
代表性序列

3) 分类学注释

5) 多序列比对

4) 构建OTU表
(TXT / BIOM file)

6) 系统发育分析

重要中间文件

OTU 表

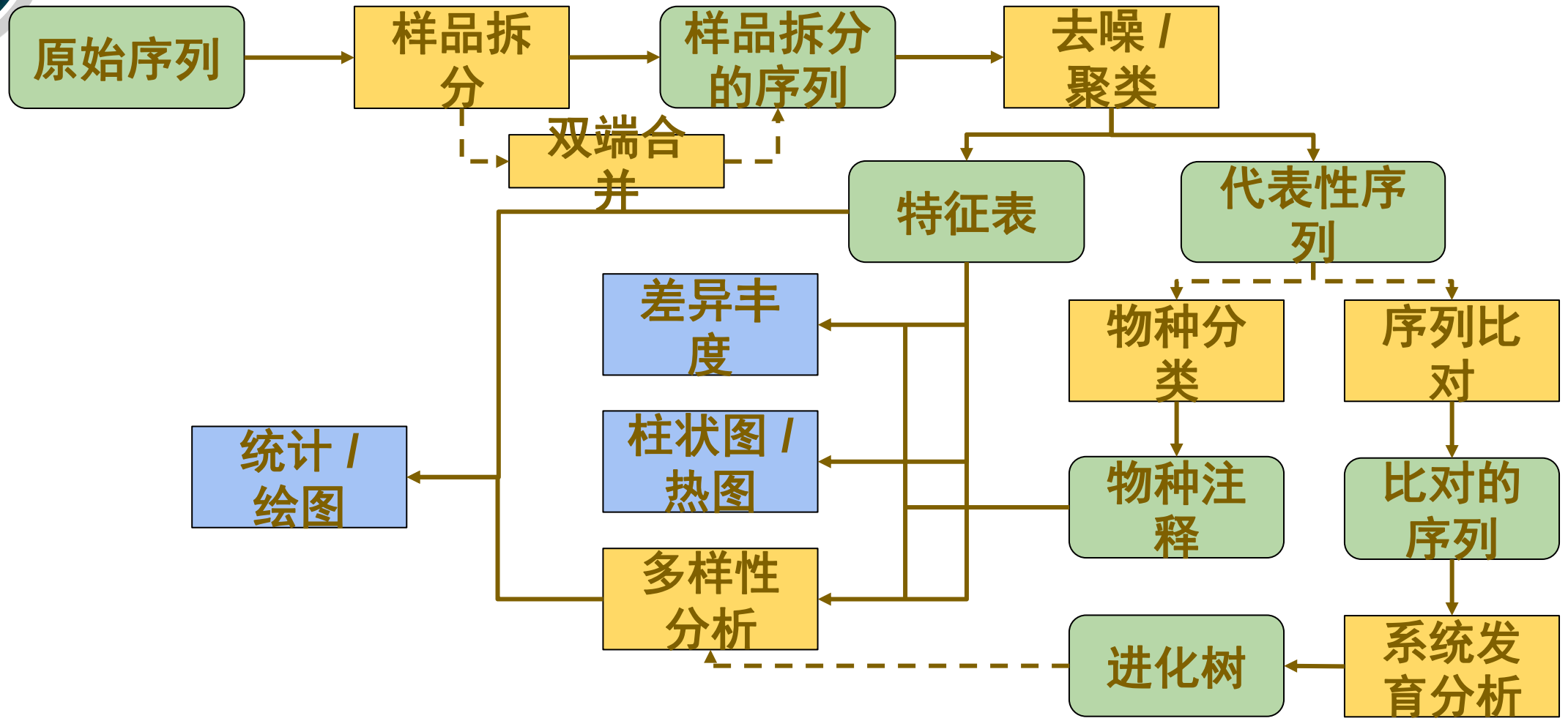
进化树

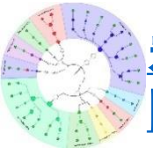
7) 下游分析和可视化
– 知识发现

输出

宏基因组

QIIME 2分析流程概述

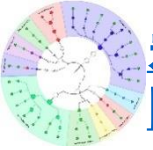




宏基因组

目录

- 分析的硬件要求
- 常用分析软件
- 扩增子测序简介
- 常用扩增子分析软件和数据库
- 扩增子分析的基本思路
- **易用的扩增子分析流程EasyAmplicon**
- 最流行的扩增子分析流程QIIME 2



1. 认识文件格式

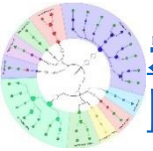
- 测序原始数据：seq/*.fq.gz

```
@HISEQ:549:HLYNYBCXY:1:1101:2135:2154 1:N:0:CAGGCGAT
ACGCTCGACAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTGTGCTGGGCGTCGGGGGGCTTGCCCCT
+
@DDDDHIIIIIIIHHIIIIGHIHICGHIIIIIIH<FHF?CHHIHHCGHHHHIFHCHE@G@EF?HHHHCHID/EEHCEHHI
```

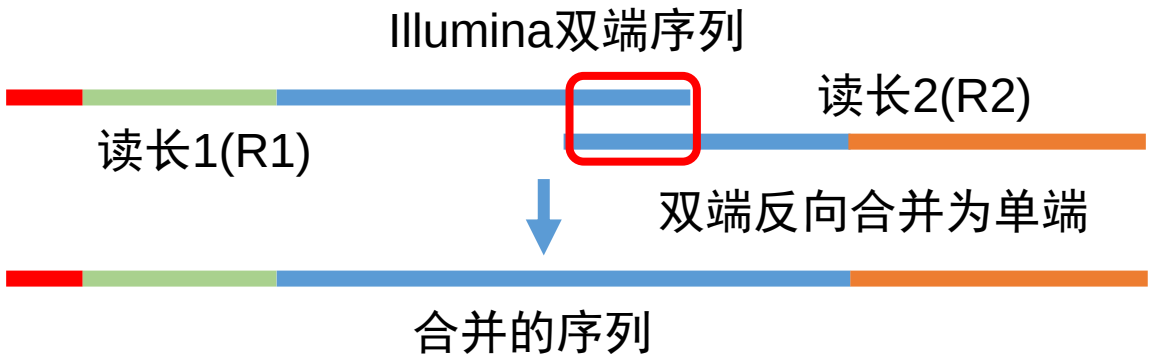
- 实验设计/样品信息：metadata.txt (制表符分隔文本文本，可用Excel或纯文本编辑器编写，如Editplus)

SampleID	Group	Date	Site	CRA	CRR	BarcodeSequence	LinkerPrimerSequence	ReversePrimer
KO1	KO	2017/6/30	Chaoyang	CRA002352	CRR117575	ACGCTCGACA	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
KO2	KO	2017/6/30	Chaoyang	CRA002352	CRR117576	ATCAGACACG	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
OE1	OE	2017/6/30	Chaoyang	CRA002352	CRR117581	TCTCTATGCG	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
OE2	OE	2017/6/30	Chaoyang	CRA002352	CRR117582	TACTGAGCTA	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC

- 注意事项：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。



2. 双端序列合并



图例
标签(Barcode)
正向引物(PrimerF)
扩增子(Amplicon)
反向引物(PrimerR)

R1

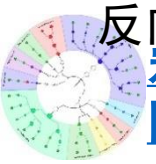
TCGTCGCTCGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGTTGGGCGCTAGGTGTGGGGGACATTCACGTTCTCCG
TGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGTTGAAACTCAAAGGAATTGACGGGGACCCGCGCA
AGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGTCTTGACATCCATGGAACCCTGCAGAGATGC

R2

ACGTCATCCCCACCTTCCTCCGGTTTGTCACCGGCGGTCTCCTTAGAGTTCCCAACTAAATGATGGCAACTAAGGACAAGGGTT
GCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTCTCATGGTTCCTTACGGC
ACCCCCGCATCTCTGCAGGGTTCCATGGATGTCAAGACCAGGTAAGGATCTTCGCGTGGCATCGAAGTAAAACACAGGCACC

R2_RC
反向互补

GGTGCCTGTGTTTTACTTCGATGCCACGCGAAGATCCTTACCTGGTCTTGACATCCATGGAACCCTGCAGAGATGCGGGGGTGC
CGTAAGGAACCATGAGACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAA
CCCTTGTCCTTAGTTGCCATCATTTAGTTGGGAACTCTAAGGAGACCGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGT



宏基因组

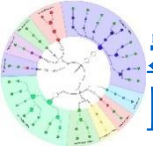
双端序列合并的实现

- 一条命令实现双端序列合并

```
vsearch -fastq_mergepairs seq/WT1_1.fq.gz -reverse seq/WT1_2.fq.gz \
        -fastqout temp/WT1.merged.fq -relabel WT1.
```

- 解释：扩增子分析 -序列合并 序列1 -反向序列 序列2 -输出 合并结果
- 小技巧，使用变量替换文件名可变部分，方便修改

```
i=WT1 # 如果你的文件名为human_skin_180910_beijing_1.fq
vsearch -fastq_mergepairs seq/${i}_1.fq -reverse seq/${i}_2.fq \
        -fastqout temp/${i}_merge.fq -relabel ${i}.
```



理解命令和命令行参数

盖个房子？

瓦匠 把砖 盖成房子

1. 谁能干：找人

瓦匠

2. 对谁干：材料

村东头砖厂

3. 结果：盖好的房子

你家马路对面的新房子

把双端测序文件按末端互相合并？

vsearch -fastq_mergepairs seq/WT1_1.fq -reverse seq/WT1_2.fq -fastqout temp/WT1_merge.fq

1. 谁能干：具体程序

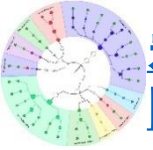
vsearch -fastq_mergepairs

2. 对谁干：输入文件

seq/WT1_1.fq -reverse seq/WT1_2.fq

3. 结果：输出文件

-fastqout temp/WT1_merge.fq



宏基因组

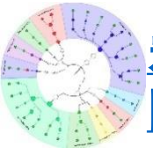
小技巧：循环批处理双端合并

- for循环实现处理实现数据中所有样品

```
for i in `tail -n+2 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 &
```

- cat合并所有样品至同一文件

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

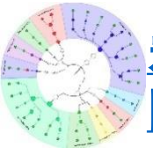


3. 切除扩增引物和质控

切除双端引物

Clean amplicon

- 先知道：barcode位置和大小
- 引物序列和长度(不清楚？谁建库问谁)
- 切除双端引物和barcodes，并质控错误率<1%
  - vsearch --fastx_filter temp/all.fq \
  - --fastq_stripleft 29 --fastq_stripright 18 \
  - --fastq_maxee_rate 0.01 \
  - --fastaout temp/filtered.fa
  - # 例如：本实验设计中左端为10bp barcode + 19 bp 5’ primer共29;
  - 右端3’ primer 18bp；错误率控制0.01即小于1%

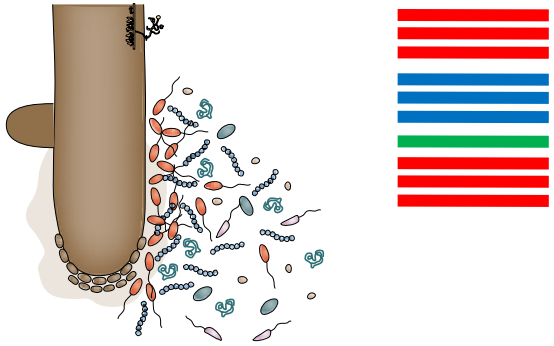


宏基因组

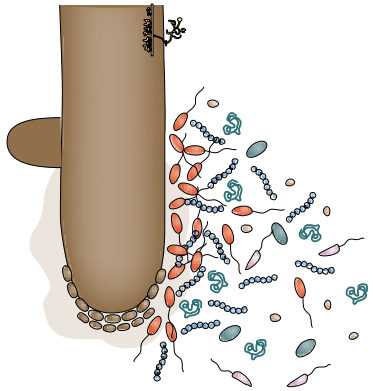
4.1 序列去冗余

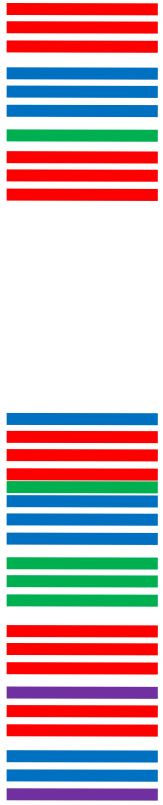
部分扩增子序列去冗余示例

SampleA



SampleB





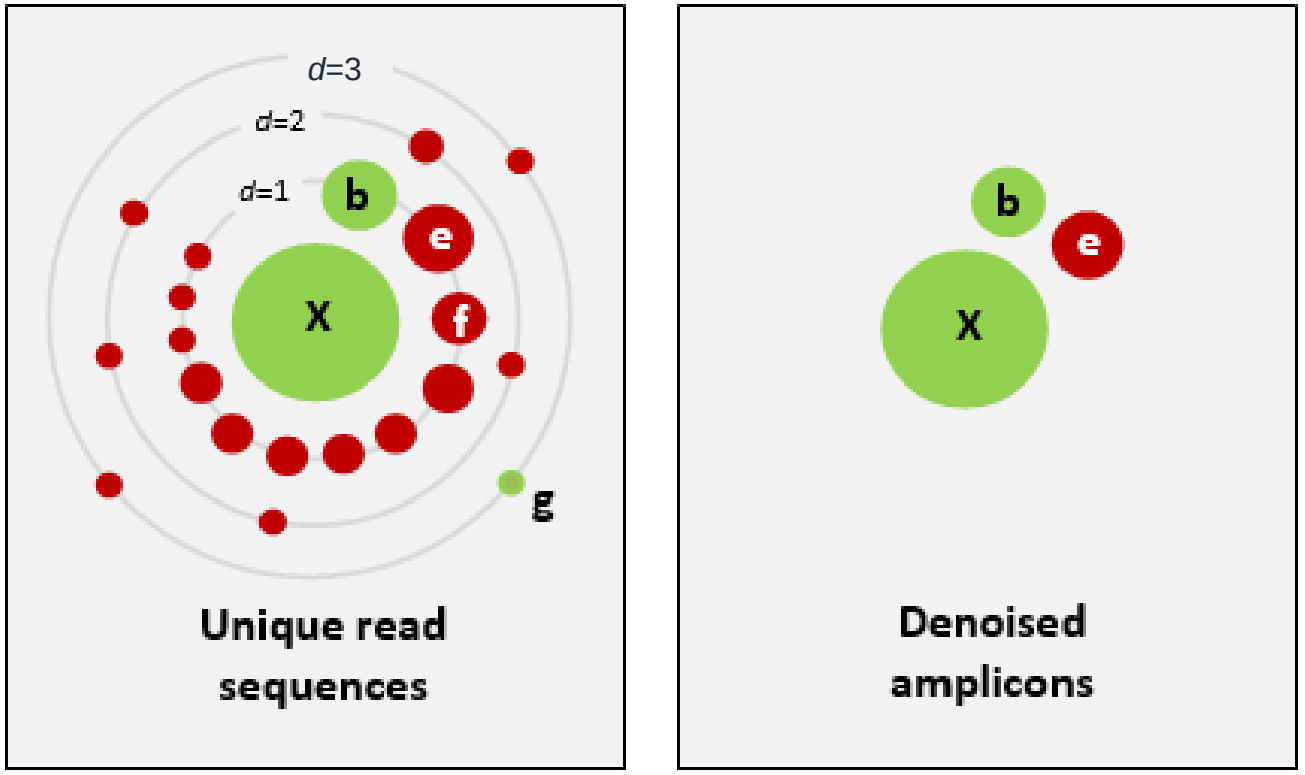
	SampleA	SampleB	Total
Red	6	8	**14**
Green	1	4	**5**
Blue	3	6	**9**
Purple	0	2	**2**
Total	10	20	**30**

去冗余数据量起码降1个数量级，减小下游分析工作量，也更适合基于丰度鉴定真实OTUs

```
# 去冗余实现
vsearch --derep_fulllength temp/filtered.fa \
  --output temp/uniques.fa --relabel Uni \
  --minuniquesize 8 --sizeout
```

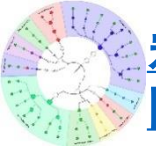
宏基因组

4.2 鉴定OTU/ASV原理



Cluster OTU VS Denoise ASV

Edgar, Robert C. "UPARSE: highly accurate OTU sequences from microbial amplicon reads." Nature methods 10.10 (2013): 996.
Edgar, Robert C., and Henrik Flyvbjerg. "Error filtering, pair assembly and error correction for next-generation sequencing reads." Bioinformatics 31.21 (2015): 3476-3482.
Callahan, Benjamin J., et al. "DADA2: high-resolution sample inference from Illumina amplicon data." Nature methods 13.7 (2016): 581.
Amir, Amnon, et al. "Deblur rapidly resolves single-nucleotide community sequence patterns." mSystems 2.2 (2017): e00191-16.



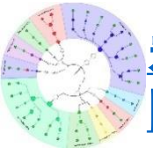
宏基因组

主流非聚类方法dada2,deblur和unoise3介绍与比较

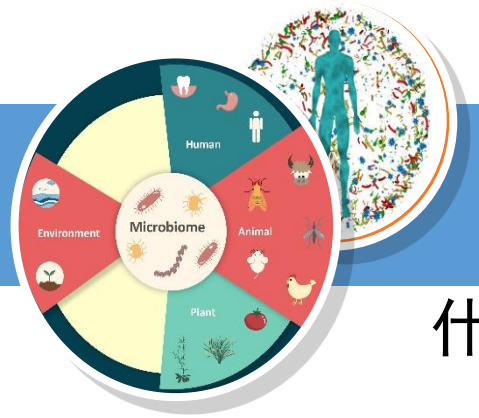
DADA2中文教程v1.8

4.2 鉴定OTU/ASV

- # 方法1. 97% UPARSE聚类OTU(快但容易被质疑方法旧，不同研究序列可能不一致比较困难)
  - usearch -cluster_otus temp/uniques.fa \
  - -otus temp/otus.fa -relabel OTU_
- **# 方法2. ASV非聚类去噪法 Denoise(相当于100%聚类) ——推荐**
  - usearch -unoise3 temp/uniques.fa \
  - -zotus temp/zotus.fa
  - # 修改序列名：格式调整 format OTU prefix方便下游分析
  - sed 's/Zotu/ASV_/g' temp/zotus.fa > temp/otus.fa

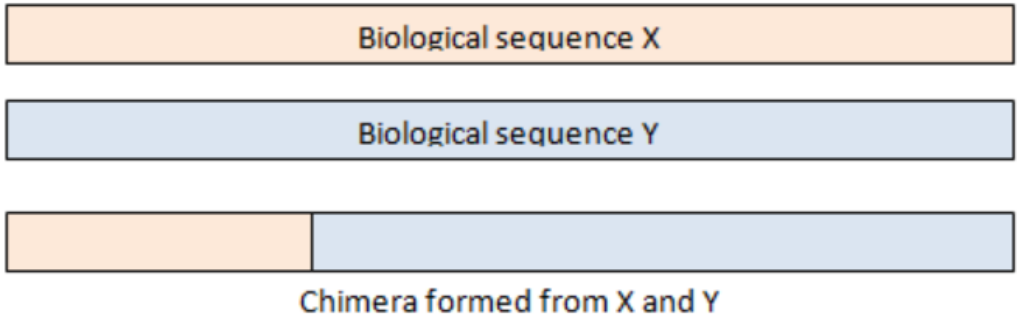


宏基因组

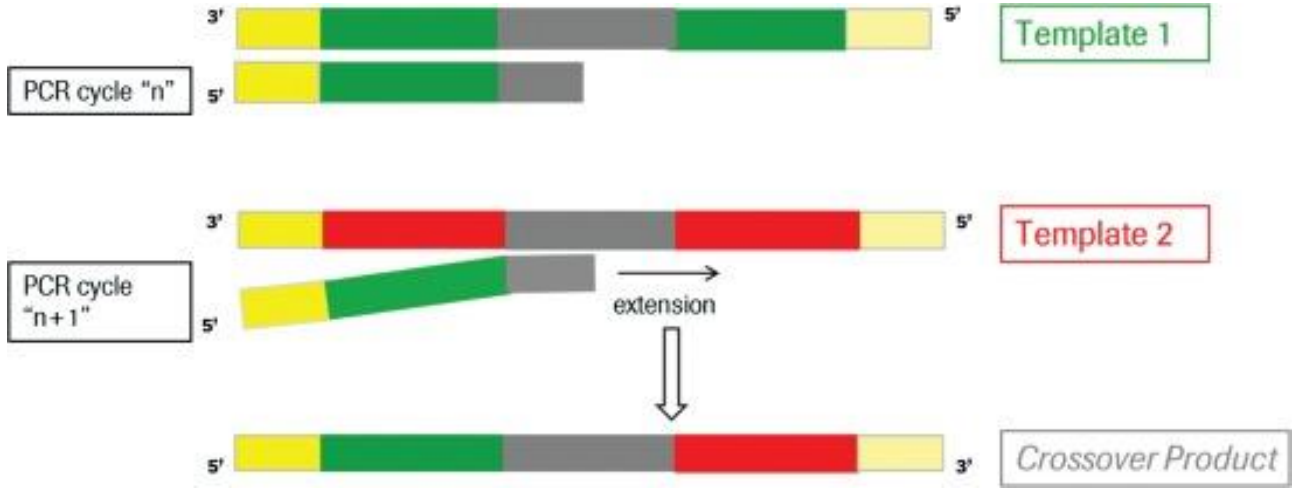


4.3 去除嵌合体

什么是嵌合体？



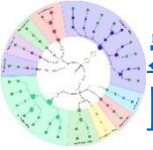
嵌合体如何产生的？



如何去除嵌合体？

无参De novo: unoise3或cluster_otus**内置de novo去嵌合体**

有参Reference: rdp、silva、greengene数据库选哪个呢？

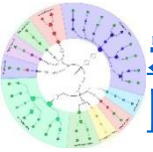


宏基因组

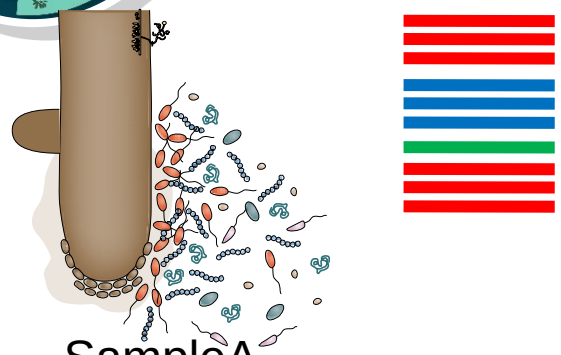
4.3 vsearch基于RDP/SILVA嵌合

- #方法1. vsearch使用RDP去嵌合(快15s但容易假阴性)，或SILVA去嵌合(silva_16s_v123.fa)，推荐(慢，耗时15m+，理论更好)
  - vsearch --uchime_ref temp/otus.fa \
  - --db ${db}/usearch/rdp_16s_v18.fa \
  - --nonchimeras result/raw/otus.fa
  - # 11m29s, 2.5G, Found 296 (8.9%) chimeras, 2962 (89.1%) non-chimeras
  - #Win用户注释：vsearch去嵌合后每行添加了windows换行符^M，需删除
  - sed -i 's/\r//g' result/raw/otus.fa

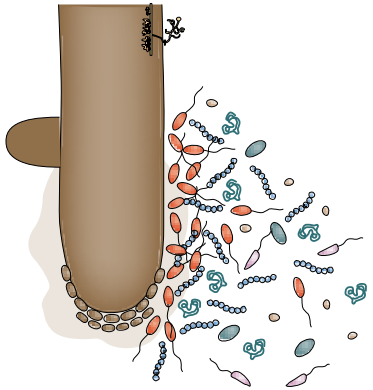
- # 方法2. 不去嵌合，请执行如下命令(发现已知菌被丢弃假阳性)
  - cp -f temp/otus.fa result/raw/otus.fa



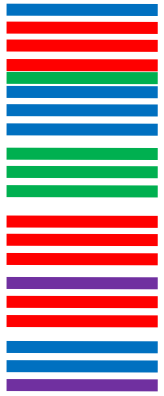
5. 生成特征表的原理



SampleA



SampleB



序列比对至代表序列计数

	SampleA	SampleB
BacRed	6	8
BacGreen	2	4
BacBlue	3	6
BacPurple	0	2

等量重抽样：比较物种多样性

	SampleA	SampleB
BacRed	6	4
BacGreen	2	2
BacBlue	2	3
BacPurple	0	1

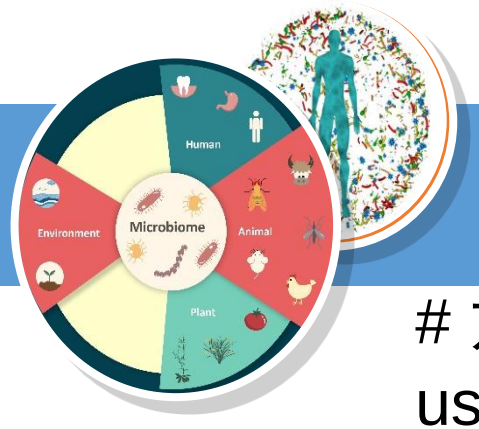
多样性指数：A的丰富度为3，B为4

相对丰度：比较比例差异

	SampleA	SampleB
BacRed	60%	40%
BacGreen	20%	20%
BacBlue	20%	30%
BacPurple	0	10%

与B比，A中Red高，Blue和Purple低

宏基因组



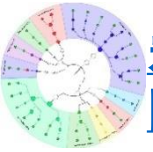
5.1 生成特征表

```
# 方法1. usearch生成特征表，小样本(<30)快
usearch -otutab temp/filtered.fa -otus result/raw/otus.fa \
    -otutabout result/raw/otutab.txt -threads 4
# 方法2. vsearch生成特征表
vsearch --usearch_global temp/filtered.fa --db result/raw/otus.fa \
      --otutabout result/raw/otutab.txt --id 0.97 --threads 4
# 224236 of 268019 (83.66%)可比对，耗时1m
```

#OTUID	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2	OE3	OE4	OE5	OE6	WT1	WT2	WT3	WT4
ASV_1	1113	1968	816	1372	1062	1087	1270	1637	1368	962	1247	1017	2345	2538	1722	2004
ASV_2	1922	1227	2355	2218	2885	1817	640	494	1218	1264	945	635	1280	1493	839	1115
ASV_3	568	460	899	902	1226	855	607	457	1058	1036	837	674	1041	1796	1019	1200
ASV_4	1433	400	535	759	1287	506	515	590	439	621	661	428	1123	1448	547	577
ASV_6	882	673	819	888	1475	1017	245	250	366	380	378	351	557	537	460	539
ASV_8	508	504	608	424	190	327	335	535	1578	780	507	516	634	763	553	1053
ASV_7	216	132	1232	367	1298	291	130	1208	834	508	195	220	799	919	547	215
ASV_9	344	801	354	444	270	551	293	442	637	392	552	398	588	325	439	430
ASV_10	360	363	689	760	1023	662	198	177	281	280	404	279	331	587	248	262
ASV_11	315	344	321	352	560	375	472	375	244	418	345	186	421	498	505	412

后续分析

- 5.2 物种注释-去除质体和非细菌/古菌并统计比例
- 5.3 等量抽样标准化——用于多样性计算
- 6. Alpha多样性指数计算
- 7. Beta多样性——样品间距离(差异)
- 8. 物种注释格式调整
- 9. 比对Greengene数据库(有参)生成OTU表
- 10. 项目空间清理
- 详见：MPB：遗传发育所刘永鑫等-易扩增子：易用、可重复和跨平台的扩增子分析流程



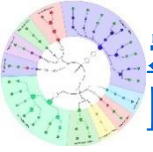
宏基因组

易扩增子(EasyAmplicon)

- - pipeline.sh # 流程脚本
- - pipeline_mac.sh # Mac版
- - result/ # 示例结果
- - result/Diversity-tutorial.Rmd
- 多样性分析交互脚本
- 使用有道云笔记Markdown阅读
- 每季度更新

项目：https://github.com/YongxinLiu/EasyAmplicon

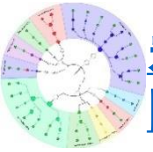
- 扩增子EasyAmplicon 1.11(2021.4)
- 21、背景介绍
- 22、扩增子16S分析流程
  - 1. 了解工作目录和起始文件
    - 1.1. metadata.txt实验设计文件
    - 1.2. seq/*.fq.gz原始测序数据
    - 1.3. pipeline.sh流程依赖数据库
  - 2. 合并双端序列并按样品重命名
  - 3. 切除引物与质控
  - 4. 去冗余挑选OTU/ASV
    - 4.1 序列去冗余
    - 4.2 聚类OTU/去噪ASV
    - 4.3 基于参考去嵌合
  - 5. 特征表和筛选
    - 5.1 生成特征表
    - 5.2 去除质体和非细菌
    - 5.3 等量抽样标准化
  - 6. Alpha多样性
    - 6.1. 计算多样性指数
    - 6.2. 计算稀释过程的丰富度变化
    - 6.3. 筛选高丰度菌
  - 7. Beta多样性
  - 8. 物种注释分类汇总
  - 9. 有参定量特征表
  - 10. 空间清理及数据提交
- 23、R语言多样性和物种分析
  - 1. Alpha多样性
    - 1.1 Alpha多样性箱线图
    - 1.2 稀释曲线
    - 1.3 多样性维恩图
  - 2. Beta多样性
    - 2.1 距离矩阵热图pheatmap
    - 2.2 主坐标分析PCoA
    - 2.3 限制性主坐标分析CPCoA
  - 3. 物种组成Taxonomy
    - 3.1 堆叠柱状图Stackplot
    - 3.2 弦/圈图circlize
    - 3.3 树图treemap/maptree
- 24、差异比较
  - 1. R语言差异分析
    - 1.1 差异比较
    - 1.2 火山图
    - 1.3 热图
    - 1.4 曼哈顿图
  - 2. STAMP输入文件准备
    - 2.1 命令行生成输入文件
    - 2.2 Rmd生成输入文件
  - 3. LEfSe输入文件准备
- 25、QIIME 2分析流程
- 31、功能预测
  - 1. PICRUSt功能预测
  - 2. 元素循环FAPROTAX
  - 3. Bugbase细菌表型预测



宏基因组

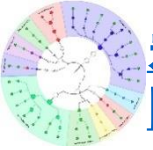
目录

- 分析的硬件要求
- 常用分析软件
- 扩增子测序简介
- 常用扩增子分析软件和数据库
- 扩增子分析的基本思路
- 易用的扩增子分析流程EasyAmplicon
- **最流行的扩增子分析流程QIIME 2**



软件安装(仅Linux系统)

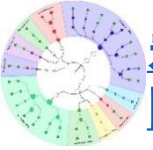
- 最新版qiime2 2021.4，更新较快，可以学最新版，每版差异不大
- Windows用户推荐**服务器**、**内置Linux子程序**、Virtualbox安装Linux等方式——适合小样本(*n*<100)，运行效率低
- Linux/Mac用户首选Conda虚拟环境安装，不成功或运行有问题推荐更稳定的Docker方式安装和运行(需管理员权限)
  - # 下载软件安装列表，无法下载时可以在public/linux或mac文件夹中找备份
  - wget https://data.qiime2.org/distro/core/qiime2-2021.2-py36-linux-conda.yml
  - # 创建虚拟环境并安装qiime2，防止影响其它已安装软件
  - conda env create -n qiime2-2021.2 --file qiime2-2021.2-py36-linux-conda.yml
  - # 激活QIIME2工作环境
  - conda activate qiime2-2021.2



文件索引：manifest文件编写

- 双端三列：样品名、正向绝对路径、反向绝对路径
- 制表符分隔，支持环境变量
  - sample-id　　forward-absolute-filepath　　reverse-absolute-filepath
  - KO1　　$PWD/seq/KO1_1.fq.gz　　$PWD/seq/KO1_2.fq.gz
  - KO2　　$PWD/seq/KO2_1.fq.gz　　$PWD/seq/KO2_2.fq.gz
  - KO3　　$PWD/seq/KO3_1.fq.gz　　$PWD/seq/KO3_2.fq.gz

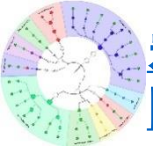
- 单端二列，没有第三列即可
  - sample-id　　absolute-filepath
  - KO1　　$PWD/seq/KO1_1.fq.gz



数据导入qiime2

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

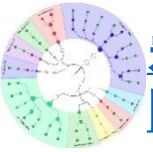
- 格式为双端33格式
- 导入1GB fq文件耗时7m，压缩后fq.gz文件用时34s



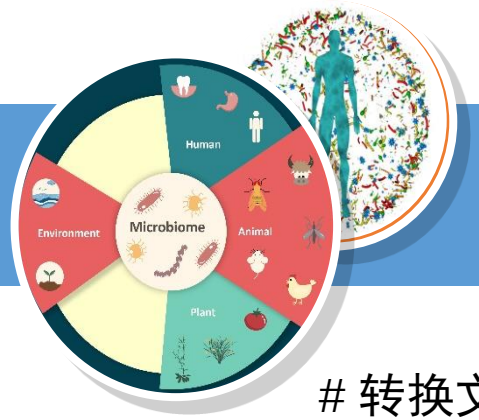
方案1. dada2生成特征表(R包慢)

- 从拆分好原始序列——特征表、代表序列
- --p-n-threads多线程加速，0为不限制；--p-trim-left截取序列起始5’端低质量、标签或引物；--p-trunc截取3’端低质量，必填项

```
qiime dada2 denoise-paired --i-demultiplexed-seqs demux.qza \
  --p-n-threads 12 \
  --p-trim-left-f 29 --p-trim-left-r 18 --p-trunc-len-f 0 --p-trunc-len-r 0 \
  --o-table table.qza \
  --o-representative-sequences rep-seqs.qza \
  --o-denoising-stats denoising-stats.qza
#支持多线程加速，90万条PE250数据，0/96p, 34m；24p, 44m；8p, 77m；1p, 462m，27万，8p, 9m
```



宏基因组

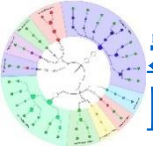


方案2：导入特征表和代表序列

```
# 转换文本为Biom1.0
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
```



宏基因组

特征表和代表序列统计

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

qzv结果三种查看方式：

1. 推荐在网站https://view.qiime2.org上查看
2. 解压查看data/index.html
3. 在支持图形界面的Linux环境中使用qiime tools view查看

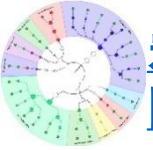
Table summary

Metric	Sample
Number of samples	18
Number of features	569
Total frequency	161,236

Frequency per sample

	Frequency
Minimum frequency	7,439.0
1st quartile	8,597.0
Median frequency	8,843.5
3rd quartile	9,474.75
Maximum frequency	10,298.0
Mean frequency	8,957.555555555555

table.qzv特征表基本统计信息
最小值用于多样性计算



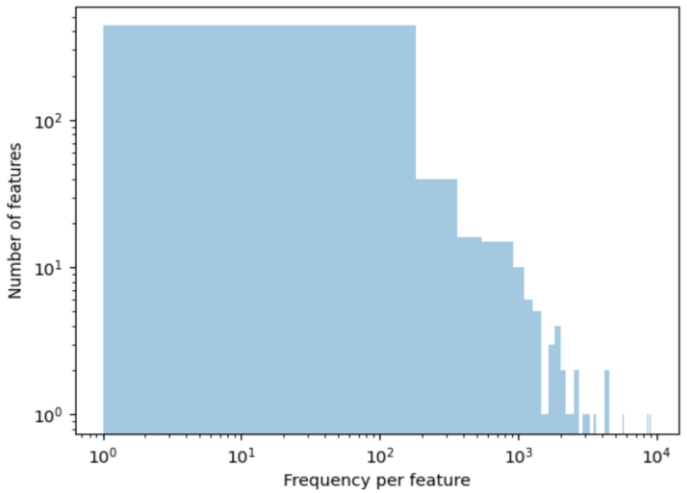
宏基因组

table.qzv：csv数据表+SVG矢量图

Frequency per feature

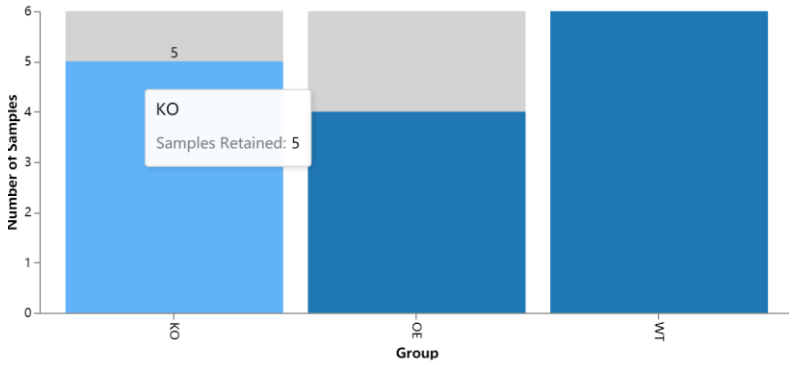
	Frequency
Minimum frequency	1.0
1st quartile	17.0
Median frequency	45.0
3rd quartile	154.0
Maximum frequency	9,089.0
Mean frequency	283.3673110720562

Frequency per feature detail (csv | html)



特征频率统计分位数和分布

Overview | Interactive Sample Detail | Feature Detail



Plot Controls

Save as SVG | Save as PNG | View Source

Metadata Category

Group

Sampling Depth

8026

(zero implies no even sampling)

Retained 120,390 (74.67%) features in 15 (83.33%) samples at the specifed sampling depth.

Sample ID	Feature Count
WT2	10298
KO5	10209
WT5	10012

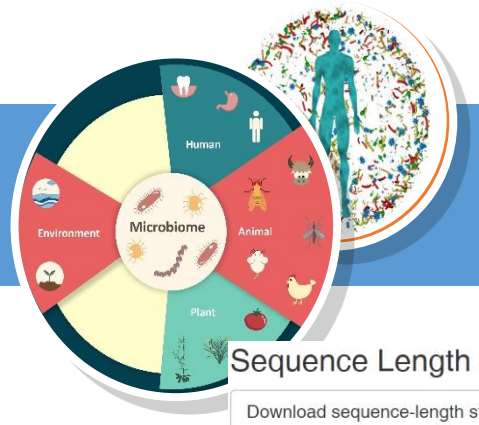
交互探索不同分组下数据量筛选
选择合适的子采样阈值

Overview | Interactive Sample Detail | Feature Detail

	Frequency	# of Samples Observed In
c40cc30ae45181d02b1f55e600a22236	9,089	18
16908489817e1cfdd163f13938103c93	8,641	18
583f789c1c9501c9386e498f18e5af76	5,708	18
2613542be210e163f03bf8393405f6aa	4,533	18
532871e414f8ab8a15a7b332e5e2f0e6	4,414	18
b2e2e66f74585fbaa04a53f2e374edda	4,233	18
918e42d19fae443929267607a0906a2e	4,199	18
2de532901d69f6c44c9411e154934189	3,480	18
2457d3a95c5d465e4a5c8247fa7b6976	3,139	17
7772ab45417ce7f94580ca640332faea	3,078	18
f409e1c0ea1ba1769ba10adf78402913	2,635	18
e2df26c03c0638a1af8a188baac28c9c	2,565	18
6479b61d7a9d6fc01f7c87fd2d411149	2,477	18
e18463edc0ad6dc5e4c56abe5da548e1	2,332	18
a79fe461438cb34c1377c73c03cd1bb1	2,094	18
b735fdbe550786eda1c3f9208a992768	2,057	17
f0051f2cca46b1be44ec19fb2d078ed5	1,887	18
e07996c6f212646e5fd459bff0c18110	1,862	15
4227a837a2b512ef8de9d7550762f54b	1,854	16

ASV数据量和样品中出现次数
高丰度和核心ASV

宏基因组



rep-seqs.qzv序列统计

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
569	356	415	376.38	59	5.07

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	360	372	373	377	379	380	385

*Values rounded down to nearest whole number.

Sequence Table

To BLAST a sequence against the NCBI nt database, click the sequence and then click the *View report* button on the resulting page.

Download your sequences as a raw FASTA file

Click on a Column header to sort the table.

Feature ID	Sequence Length	Sequence
c40cc30ae45181d02b1f55e600a22236	383	GTAGTCCACGCCGTAAACGGTGGGCGC
16908489817e1cfdd163f13938103c93	372	GTAGTCCACGCCCTAAACGATGTCAAC
583f789c1c9501c9386e498f18e5af76	372	GTAGTCCACGCCCTAAACGATGTCAAC

BLAST® » blastn suite Home Recent Results Saved Strategies Help

Format Request

Query Nucleotide Sequence (383 letters)
Database nt
Job title Nucleotide Sequence (383 letters)
Request ID 2W7EGBEG015 View report Show results in a new window
Format

Show: Alignment as HTML Old View Reset form to defaults
Alignment View: Pairwise
Display: Graphical Overview NCBI-gi CDS feature
Masking: Character: Lower Case Color: Grey
Limit results: Descriptions: 100 Graphical overview: 100 Line length: 60
Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
Enter organism name or id--completions will be suggested Exclude
Entrez query:

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Uncultured bacterium clone OTU_884 16S ribosomal RNA gene, partial sequence	679	679	100%	0.0	99%	MG275189.1
Uncultured bacterium clone 8-544 16S ribosomal RNA gene, partial sequence	670	670	100%	0.0	99%	KC554446.1
Uncultured bacterium clone OTU_1304 16S ribosomal RNA gene, partial sequence	628	628	100%	3e-176	96%	MG275597.1
Actinocorallia glomerata strain NBRC 15960 16S ribosomal RNA gene, partial sequence	628	628	100%	3e-176	96%	NR_112732.1
Actinocorallia longicatena strain IMSNU 22180 16S ribosomal RNA, partial sequence	628	628	100%	3e-176	96%	NR_042033.1

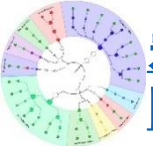
宏基因组

序列长度统计和序列列表

点序列跳转NCBI blast页
点View report探索序列来源

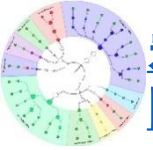
QIIME 2进一步学习

- 简明教程(5千字，1天入门)
  - MPB：使用QIIME 2分析微生物组16S rRNA基因扩增子测序数据(视频)
- 官方教程中文版(10万字，32节，半个月系统学习)
  - NBT：QIIME 2可重复、交互式的微生物组分析平台
  - 1简介和安装Introduction&Install
  - 2插件工作流程概述Workflow
  - 3老司机上路指南Experienced
- 英文原版(最新版见官网)
  - https://docs.qiime2.org/



总结

- 起始文件：测序数据(fq)、元数据(metadata)和参考数据库(RDP/Greengenes/SILVA用于16S, UNITE用于 ITS)
- 数据分析：双端合并、切除引物和质控、去冗余和生成特征(OTU/ASV)表、物种注释(代表序列与数据库比对)
- Alpha多样性：统计样品物种丰富度(richness/chao1)、均匀度(evenness/dominance)或两者(shannon/simpson)
- Beta多样性：计算样品距离矩阵常用物种距离(Bray-Cutis)、进化距离(Unifrac)，可进一步结合权重(Weighted)和无权重(Unweighted)
- EasyAmplicon：提供完整扩增子分析参考流程，简单、高效、免费、跨平台、与主流软件全兼容，每季更新，北京/线上提供收费培训服务
- QIIME 2: 更新及时、团队维护、在欧美提供收费培训服务



宏基因组(公众号、个人微信)

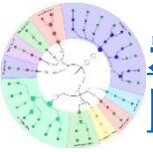
扫码关注宏基因组，获取专业学习资料

每天坚持学习进步一点点

$(1 + 0.01)^{356} = 37.8$

加我进入同行交流群

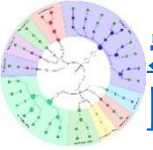
备注：姓名-单位-职称/年级-研究方向



宏基因组

参考资源

- 宏基因组公众号文章目录 生信宝典公众号文章目录
- 科学出版社《微生物组数据分析与可视化实战》——50+篇
- Bio-protocol《微生物组实验手册》计划——150+篇
- Protein Cell：扩增子和宏基因组数据分析实用指南
- CMJ：人类微生物组研究设计、样本采集和生物信息分析指南
- 加拿大生信网 https://bioinformatics.ca/ 宏基因组课程中文版
- 美国高通量开源课程 https://github.com/ngs-docs
- Curtis Huttenhower http://huttenhower.sph.harvard.edu/
- Nicola Segata http://segatalab.cibio.unitn.it/



宏基因组