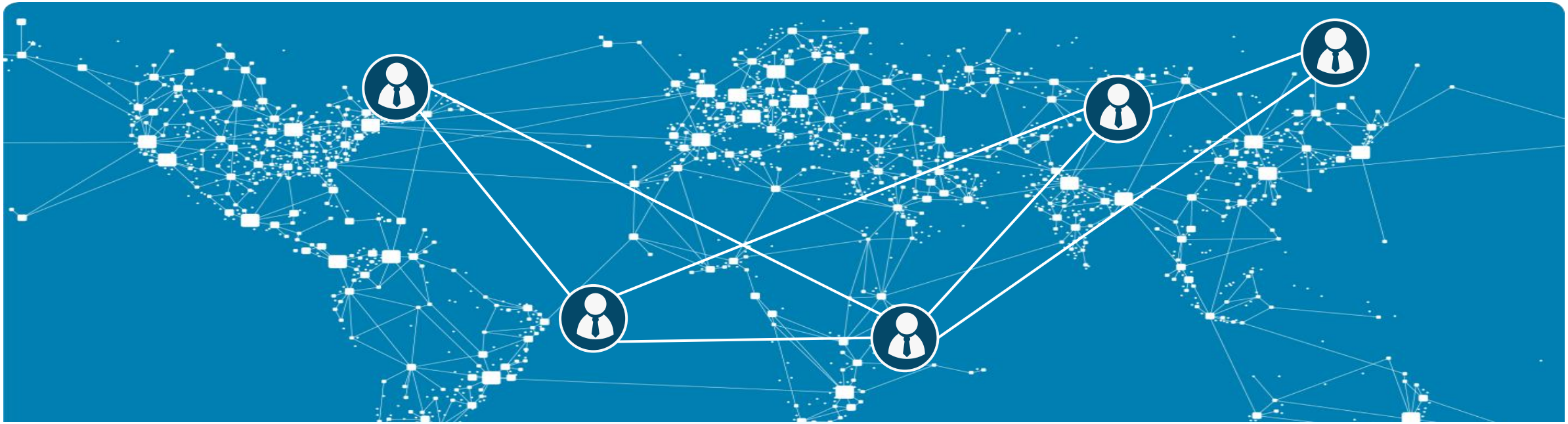
2022微生物组—扩增子16S分析研讨会第14期



22测序数据到特征表

易生信

2022年1月8日

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 $10^{1\sim3}$

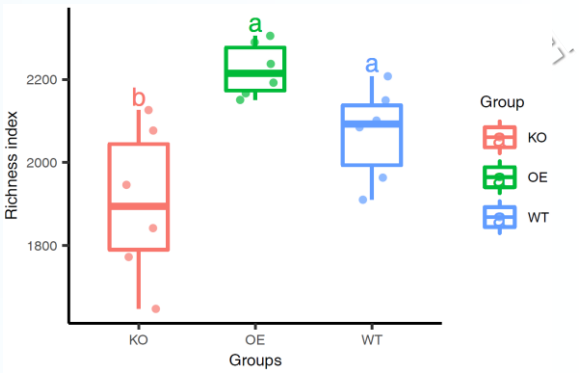


图: $10^{1\sim3}$个点和统计信息

[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]

扩增子分析基本思路

16S rRNA gene

USEARCH

物种组成

	Sample 1	Sample 2	Sample 3
OTU 1	4	0	2
OTU 2	1	0	0
OTU 3	2	4	2

PICRUSt

FAPROTAX

	Sample 1	Sample 2	Sample 3
K00001	20	15	18
K00002	1	2	0
K00003	4	5	4

功能组成

R
STAMP
LEfSe

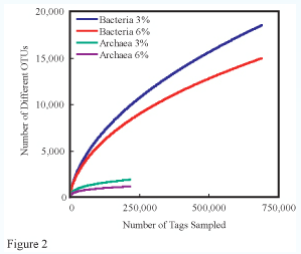
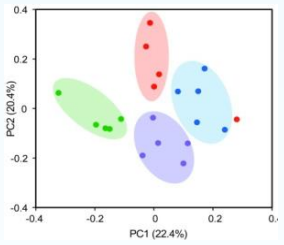
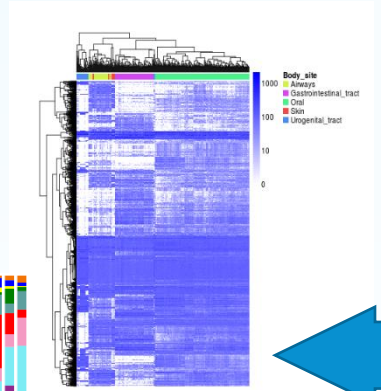
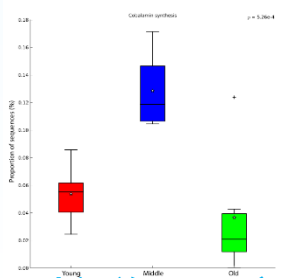
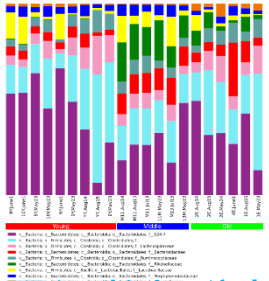


Figure 2









R
STAMP
LEfSe

Protein Cell：扩增子和宏基因组数据分析实用指南

熟记此图，胸中有丘壑

扩增子分析流程——实战分析大纲

一. 原始数据到特征表

二. LEfSe生物标记鉴定

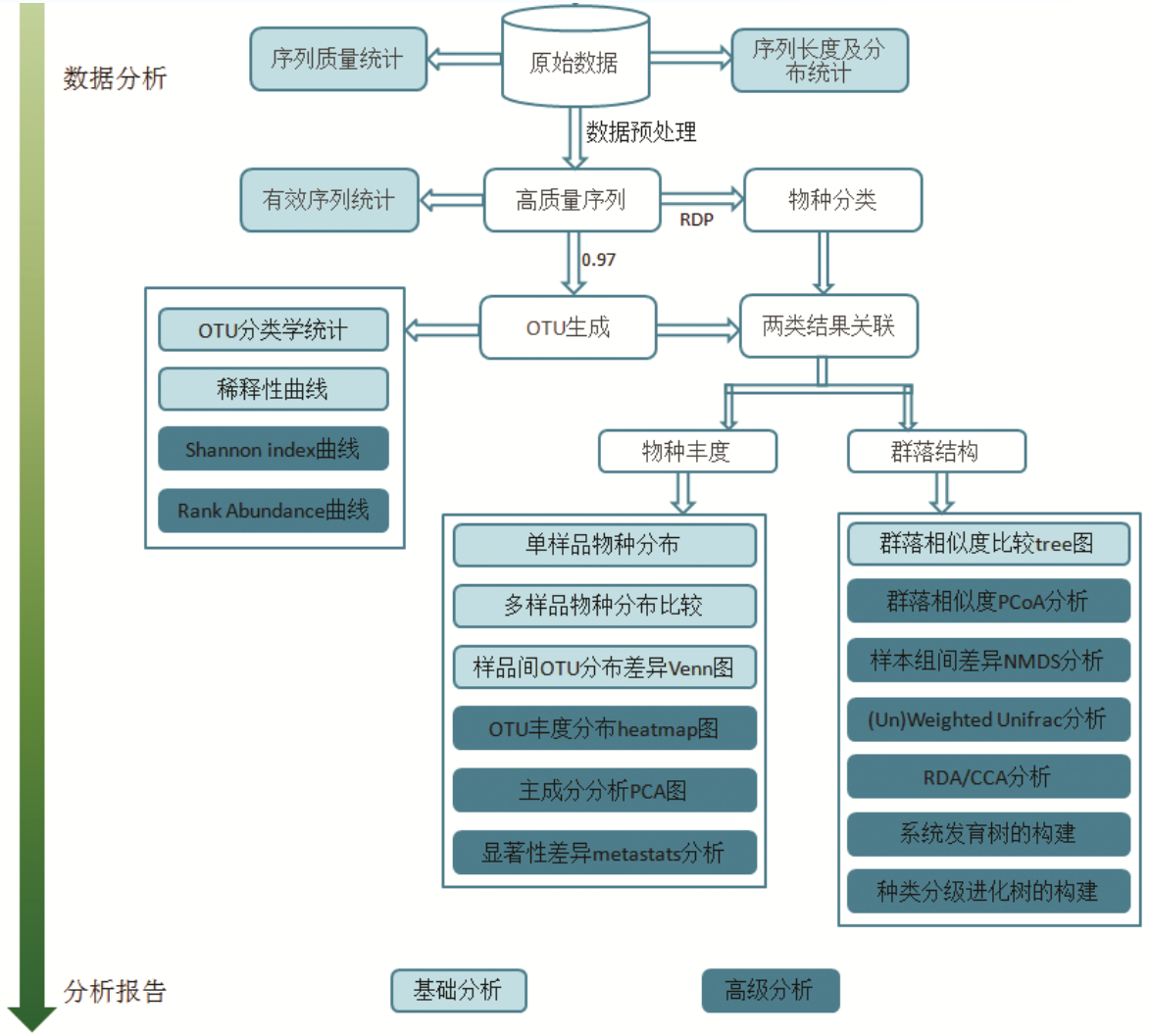
三. STAMP统计差异物种或功能

四. R语言物种多样性分析

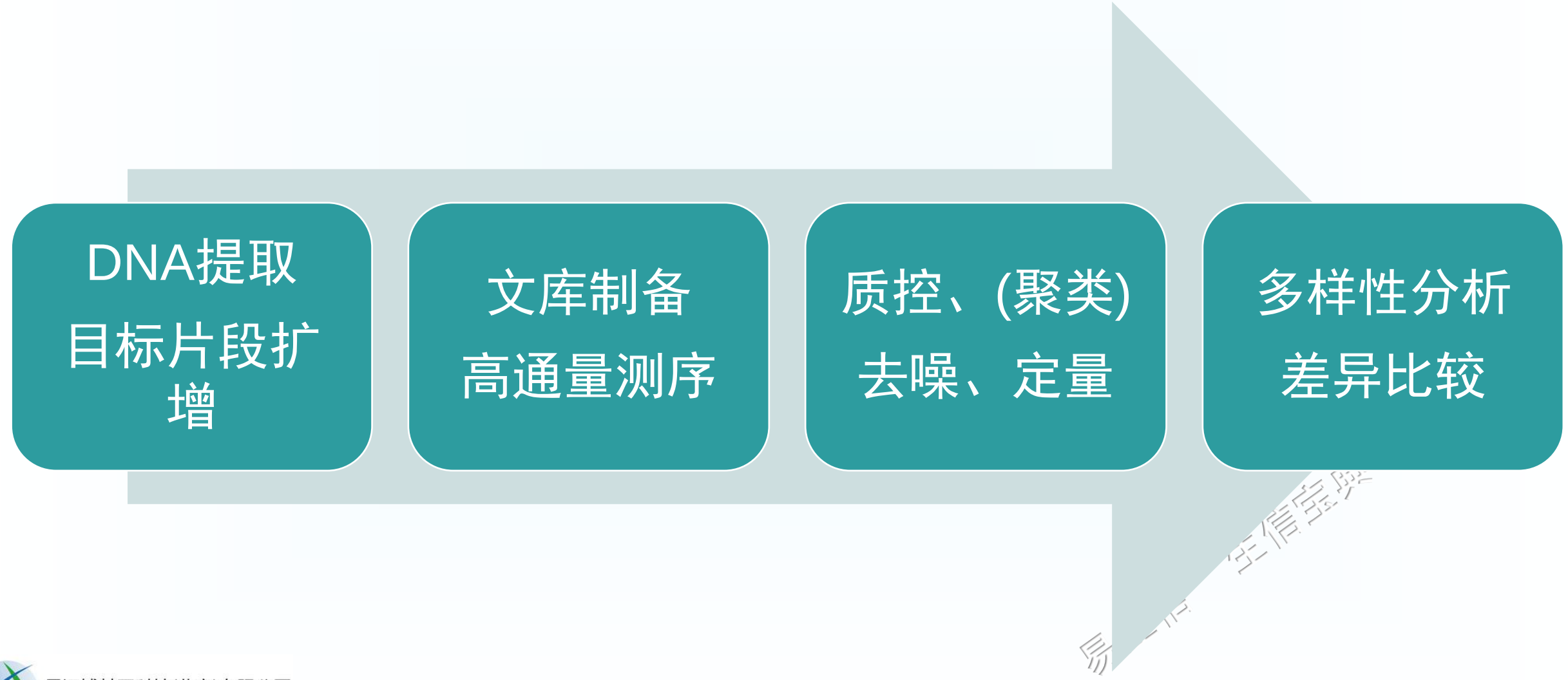
五. R语言差异比较

六. PICRUSt等功能预测

七. 机器学习分类与回归



扩增子实验和分析的基本流程



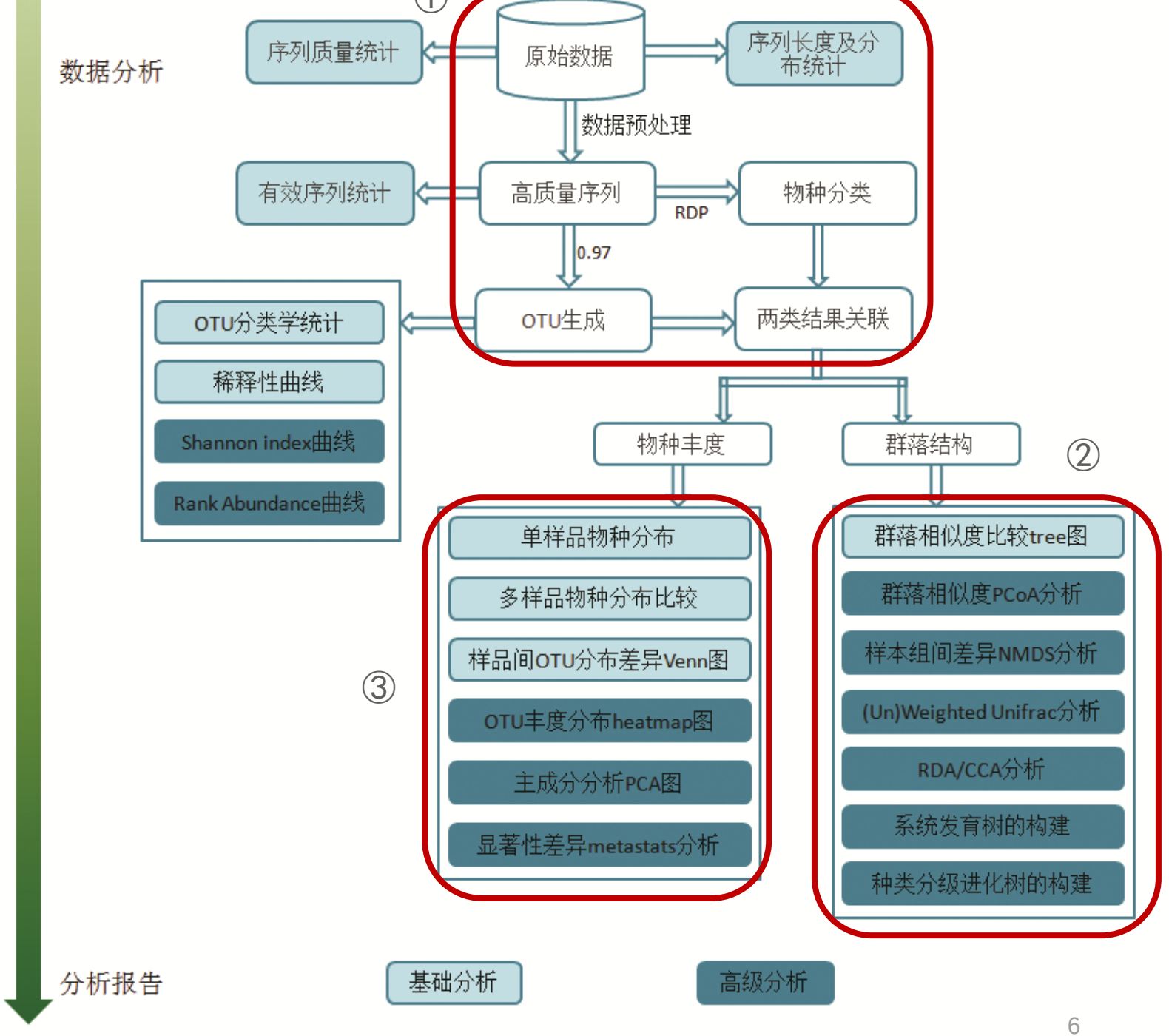
扩增子分析流程和常见结果

可回答的3个科学问题：

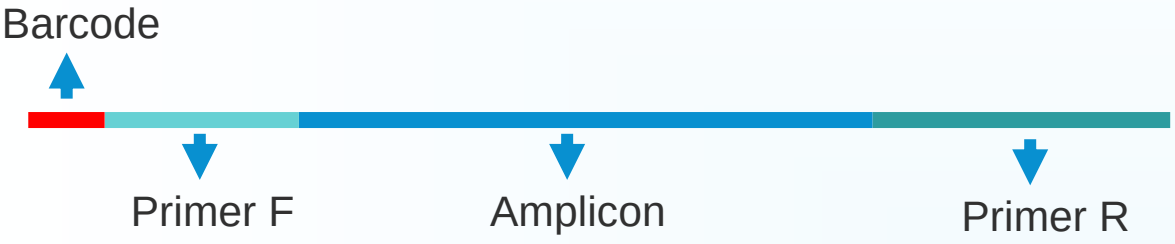
1. 样品中有什么？
测序数据→Feature表+物种注释

2. 组间差异是否存在？
Alpha多样性，PCoA，CCA

3. 差异具体是什么？
差异ASV/属/门



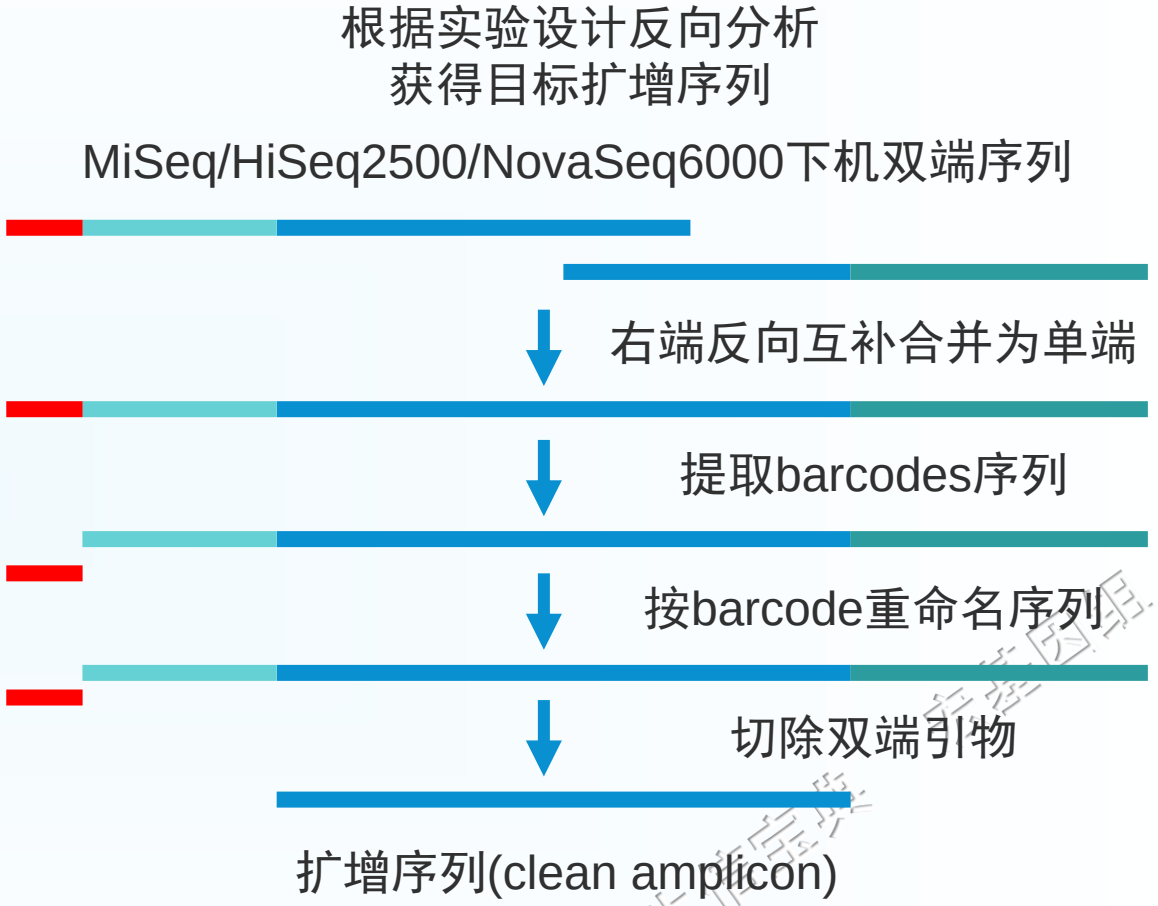
扩增子测序常用结构的模式图



Barcode：样品标签，用于混池测序后区分序列来自那个样本。过百个样本混合于同一文库中一般需要双端加Barcode

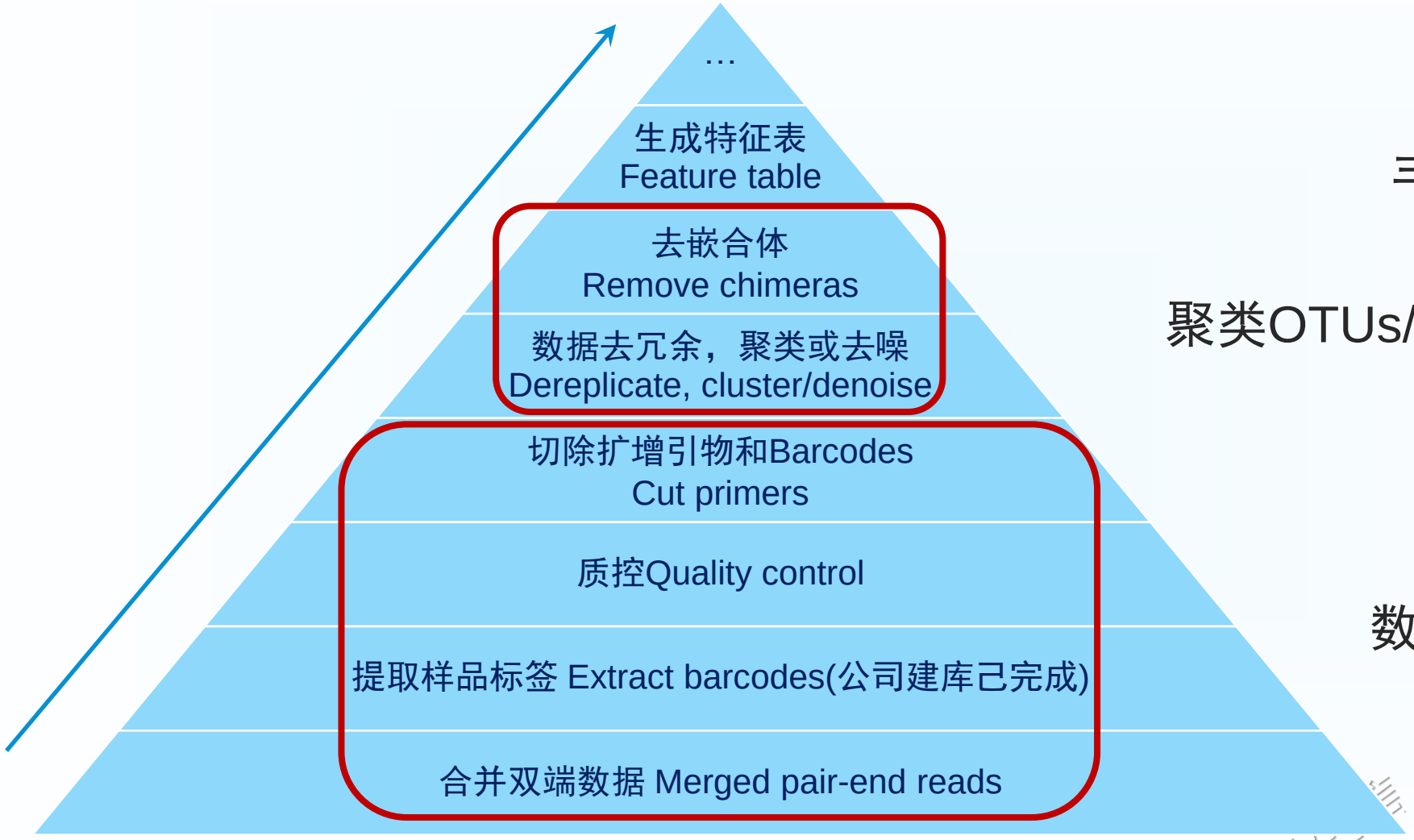
Primer：在16S/ITS/18S保守区设计的引物，用于扩增rDNA(rRNA基因)的部分高变区

Amplicon: 扩增的rDNA片段



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.***, doi: https://doi.org/10.1097/CM9.0000000000000871 (**2020**).

从原始序列到特征表



丰度矩阵

聚类OTUs/去噪ASVs

数据预处理

CMJ：人类微生物组研究设计、样本采集和生物信息分析指南

分析前准备：RStudio中几个设置

- 选择Tools菜单 —— Global Options 选项

设置默认工作目录：解决找不到工作目录

General —— Default working directory —— 选择 C:/amplicon

选择编码格式：解决中文乱码问题

Code —— Saving —— Default text encoding —— 选择UTF-8

镜像选择：加速安装包下载

Packages —— CRAN mirror —— 选择Beijing，另有Hefei / Guangzhou / Lanzhou / Shanghai可选

选择Bash：解决命令行类型不是~

Terminal —— New terminal open with —— 选择Git bash (如果没有请关闭所有软件后，重装git和Rstudio再试)

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
+
@DDDDHIIIIIIHHIIIIGHIHICGHIIIIIIH<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编辑或编程工具如Editplus等纯文本编辑器）

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
KO3	KO	2017/7/2	Changping	CRA002352	CRR117577	ATATCGCGAG	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
KO4	KO	2017/7/2	Changping	CRA002352	CRR117578	CACGAGACAG	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
KO5	KO	2017/7/4	Haidian	CRA002352	CRR117579	CTCGCGTGTC	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
KO6	KO	2017/7/4	Haidian	CRA002352	CRR117580	TAGTATCAGC	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。

序列格式详解fasta&fastq，样品命名 注意事项 实例

小技巧：测序数据、元数据统计

- seqkit统计测序数据，获得格式、数量、总/最小/平均/最大长度

```
seqkit stat seq/KO1_1.fq.gz
file              format  type  num_seqs    sum_len  min_len  avg_len  max_len
seq/KO1_1.fq.gz  FASTQ   DNA     15,000  3,750,000      250      250      250
```

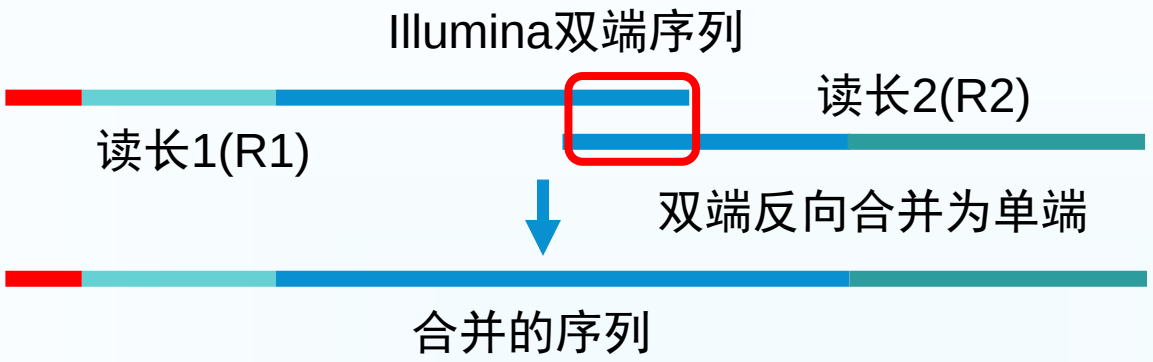
- 统计RDP数据库，2.1万条，总长30M

```
seqkit stat ${db}/usearch/rdp_16s_v18_sp.fa.gz
File              format  type  num_seqs    sum_len  min_len  avg_len  max_len
rdp_16s_v18 FASTA   DNA     21,195  30,743,088      455  1,459    1,968
```

- csvtk统计元数据

```
csvtk -t stat result/metadata.txt
file                  num_cols  num_rows
result/metadata.txt         15        18
```

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

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/${i}_1.fq.gz -reverse seq/${i}_2.fq.gz \
        -fastqout temp/${i}_merge.fq -relabel ${i}.
```

数据分析的基本思路——如何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

除输入输出外，其它参数为具体描述

vsearch -fastq_mergepairs seq/WT1_1.fq -reverse seq/WT1_2.fq -fastqout temp/WT1_merge.fq -relabel WT1.

- **样品名中不允许有点(.)**
- -relabel WT1. # 改序列名
- 原始序列名：@HISEQ:549:HLYNYBCXY:1:1101:1267:2220 1:N:0:CACTCAAT
- 新序列名：@WT1.1
- 两点好处：节省空间，方便识别。

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
```

小技巧：并行处理双端合并

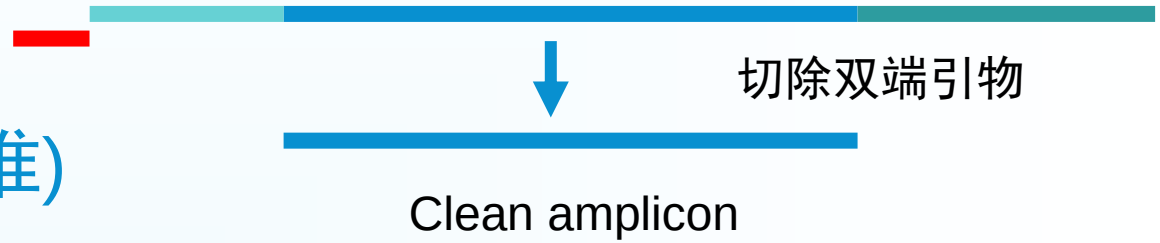
- rush并行计算任务管理程序(https://github.com/shenwei356/rush)，相当于Linux下的parallel，有效利用多线程，在计算资源允许条件下可成倍提高工作效率
- time统计运行时间，tail+cut提取样本列表，rush -j 2允许2个程序运行

```
time tail -n+2 result/metadata.txt | cut -f 1 | \
    rush -j 2 "vsearch --fastq_mergepairs seq/{}_1.fq.gz \
     --reverse seq/{}_2.fq.gz \
     --fastqout temp/{}.merged.fq --relabel {}."
```

- 在18个样本的2任务并行用时4s，比for循环8s快一倍。由于计算中受硬盘读写限制，机械硬盘上j任务建议2-4，SSD硬盘可3-7。

3. 切除扩增引物和质控

- 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
```

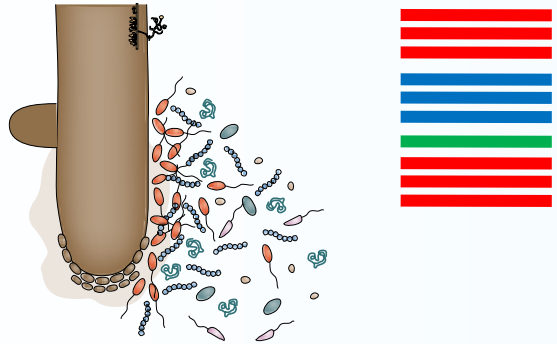
例如：本实验设计中左端为10bp barcode + 19 bp 5’ primer共29;

右端3’ primer 18bp；错误率控制0.01即小于1%

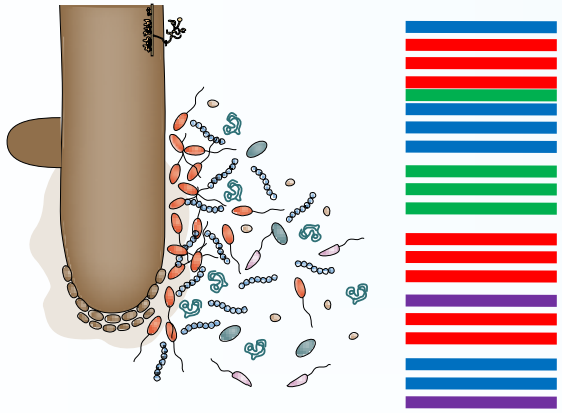
4.1 序列去冗余

部分扩增子序列去冗余示例

SampleA



SampleB



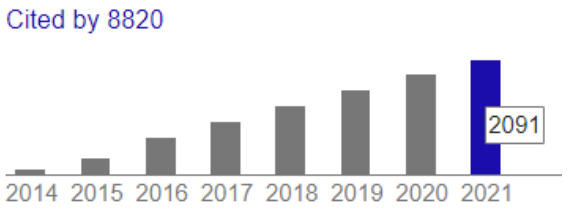
	SampleA	SampleB	Total
Red	6	8	**14**
Green	1	4	**5**
Blue	3	6	**9**
Purple	0	2	**2**
Total	10	20	**30**

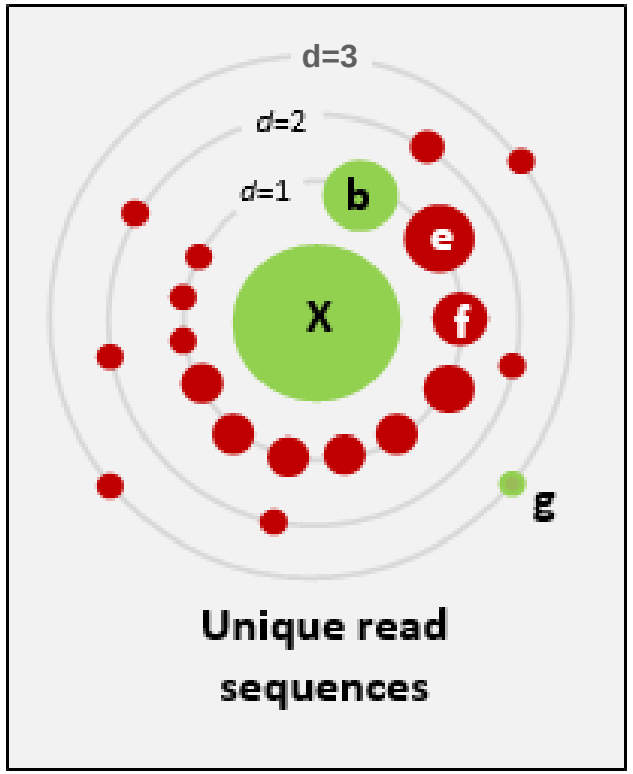
去冗余数据量起码降1个数量级，减小下游分析工作量，也更适合基于丰度鉴定真实OTUs

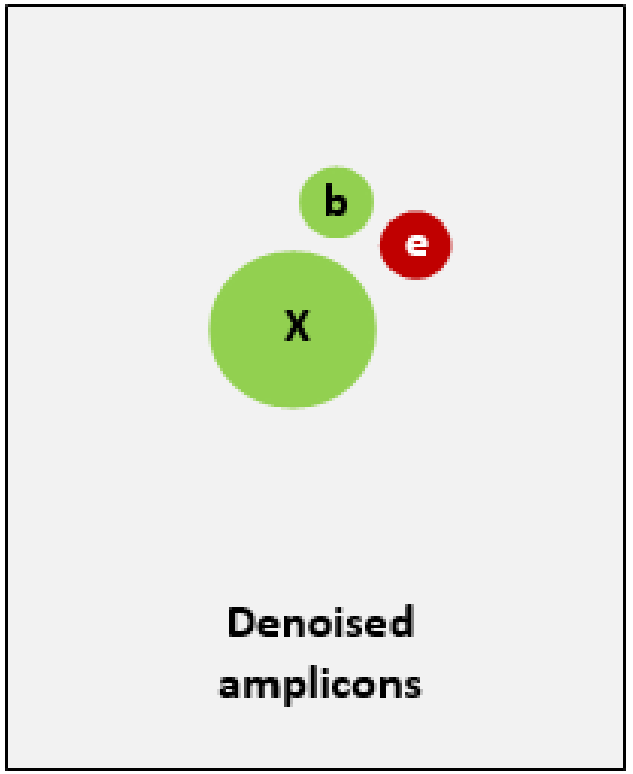
```
vsearch --derep_fulllength temp/filtered.fa \
  --output temp/uniques.fa --relabel Uni \
  --minuniquesize 8 --sizeout
# 去冗余控制最小序列频率(2~999, 1/M)，加速下游分析效率，推荐控制特征在3千~1万间
```

4.2 鉴定OTU/ASV原理

UPARSE

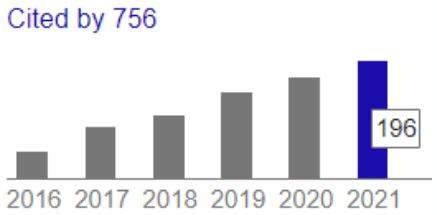




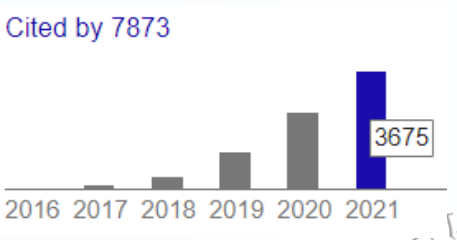


Cluster OTU VS Denoise ASV

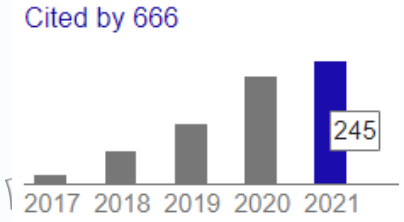
unoise3 (u/vsearch)



dada2 – R/qiime2



deblur - EMP



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.

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(Amplicon Sequences Variant)

- # 方法1. 97% UPARSE聚类OTU(快但容易被质疑方法旧，序列不为真实序列不可比较)

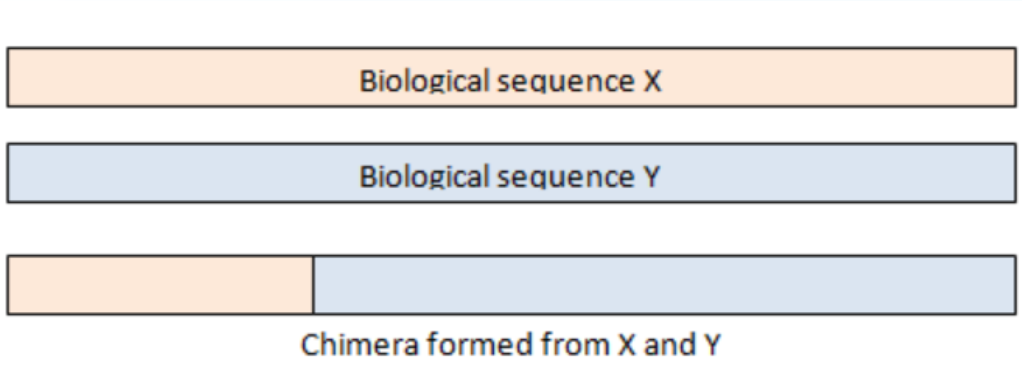
```
usearch -cluster_otus temp/uniques.fa \
  -otus temp/otus.fa -relabel OTU_
```

- **# 方法2. ASV非聚类去噪法 Denoise(相当于100%聚类) ——推荐**

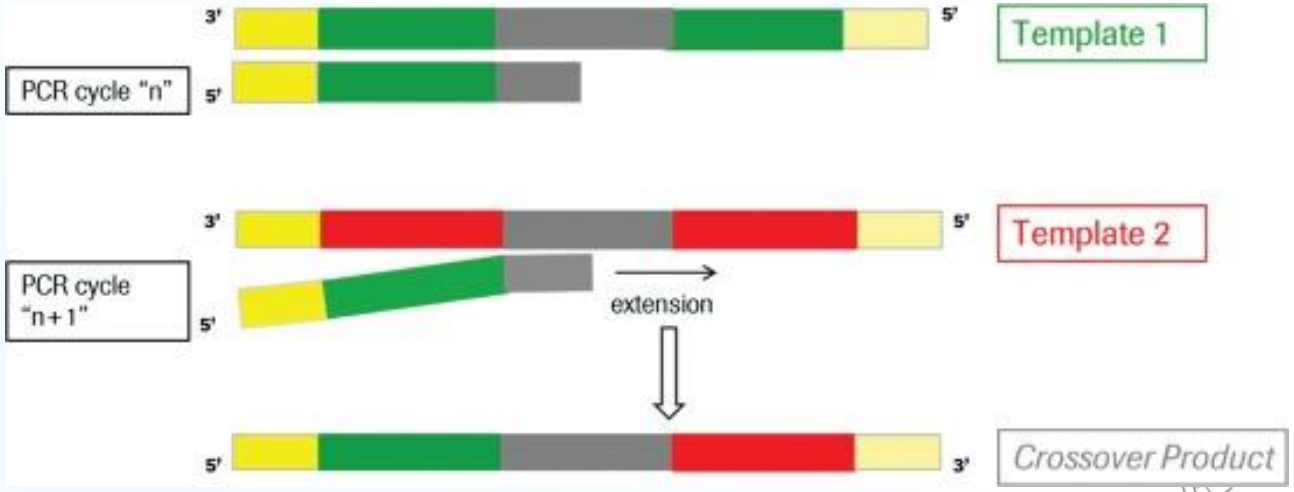
```
usearch -unoise3 temp/uniques.fa \
  -zotus temp/zotus.fa
# 修改序列名：格式调整 format OTU prefix方便下游分析
sed 's/Zotu/ASV_/g' temp/zotus.fa > temp/otus.fa
```

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？

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
```

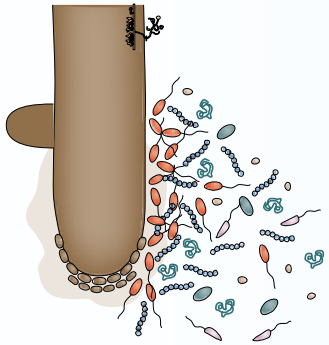
- # 方法2. 不去嵌合，请执行如下命令(发现已知菌被丢弃假阳性用户)

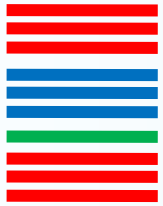
```
cp -f temp/otus.fa result/raw/otus.fa
```

5. 生成特征表的原理

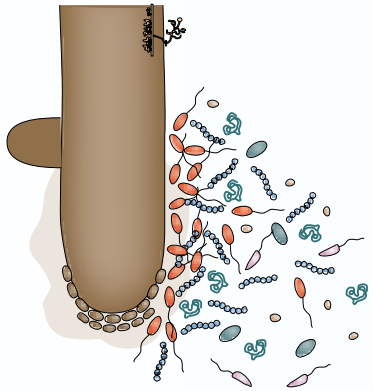
SampleA

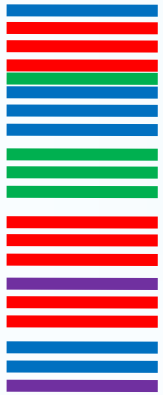
PCR后挑单克隆测序





SampleB





统计序列多样性(OTU/ASV表)

	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

5.1 生成Feature表 Create Features table

```
# 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	WT5	WT6
ASV_1	1113	1968	816	1372	1062	1087	1270	1637	1368	962	1247	1017	2345	2538	1722	2004	1439	1558
ASV_2	1922	1227	2355	2218	2885	1817	640	494	1218	1264	945	635	1280	1493	839	1115	1489	1170
ASV_3	568	460	899	902	1226	855	607	457	1058	1036	837	674	1041	1796	1019	1200	1205	768
ASV_4	1433	400	535	759	1287	506	515	590	439	621	661	428	1123	1448	547	577	1112	922
ASV_6	882	673	819	888	1475	1017	245	250	366	380	378	351	557	537	460	539	495	492
ASV_8	508	504	608	424	190	327	335	535	1578	780	507	516	634	763	553	1053	457	514
ASV_7	216	132	1232	367	1298	291	130	1208	834	508	195	220	799	919	547	215	580	857
ASV_9	344	801	354	444	270	551	293	442	637	392	552	398	588	325	439	430	754	512
ASV_10	360	363	689	760	1023	662	198	177	281	280	404	279	331	587	248	262	524	281
ASV_11	315	344	321	352	560	375	472	375	244	418	345	186	421	498	505	412	325	383

5.2 物种注释-去除质体和非细菌/古菌并统计比例

```
# 物种注释
vsearch --sintax result/raw/otus.fa --db ${db}/usearch/rdp_16s_v18.fa \
    --tabbedout result/raw/otus.sintax --sintax_cutoff 0.6
Rscript ${db}/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
```

按筛选后Feature表重新筛选代表序列和物种注释

```
#按筛选后特征表筛选对应序列
    cut -f 1 result/otutab.txt | tail -n+2 > temp/otutab.id
    usearch -fastx_getseqs result/raw/otus.fa -labels temp/otutab.id \
        -fastaout result/otus.fa
 #过滤特征表对应序列注释
    awk 'NR==FNR{a[$1]=$0}NR>FNR{print a[$1]}'\
        result/raw/otus.sintax temp/otutab.id > result/otus.sintax
#补齐末尾列
    sed -i 's/\t$/\td:Unassigned/' result/otus.sintax


#方法2. 觉得筛选不合理可以不筛选
    # cp result/raw/otu* result/
```

5.2 Feature表简单统计 Summary Features table

usearch -otutab_stats result/otutab.txt -output result/otutab.stat

cat result/otutab.stat # 统计信息如下：

218931 Reads (218.9k)
18 Samples
1612 OTUs

29016 Counts
6577 Count =0 (22.7%)
5655 Count =1 (19.5%)
3945 Count >=10 (13.6%)

403 OTUs found in all samples (25.0%)
573 OTUs found in 90% of samples (35.5%)
1450 OTUs found in 50% of samples (90.0%)

Sample sizes: **min 10912, lo 11546**, med 12318, **mean 12162.8**, hi 12566, **max 13679**

5.3 等量抽样标准化——用于多样性计算

```
#使用vegan包进行等量重抽样，输入reads count格式Feature表result/otutab.txt
#可指定输入文件、抽样量和随机数
mkdir -p result/alpha
Rscript ${bin}/script/otutab_rare.R --input result/otutab.txt \
    --depth 10000 --seed 1 \
    --normalize result/otutab_rare.txt \
    --output result/alpha/vegan.txt
```

```
1] "The input feature table is result/otutab.txt"
[1] "Samples size are:"
  KO1   KO2   KO3   KO4   KO5   KO6   OE1   OE2   OE3   OE4   OE5   OE6   WT1
11218 12318 13279 13063 13679 12413 11403 11256 11570 11546 11885 10912 12557
[1] "Rarefaction depth 10000. If depth set 0 will using sample minimum size 10912"
[1] "Random sample number: 1"
[1] "Calculate six alpha diversities by estimateR and diversity"
    richness    chao1      ACE  shannon   simpson invsimpson
KO1     1209 1473.014 1479.111 5.847877 0.9895511   95.70349
[1] "Name of rarefaction file result/otutab_rare.txt"
[1] "Output alpha diversity filename result/alpha/vegan.txt"
```

6. Alpha多样性指数计算

- 6.1 计算样品内的丰富度(richness)、均匀度(evenness)

- alpha_div命令基于标准化OTU表计算14种指数

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

- 6.2稀释抽样：1%-100%抽样一百次(richness)

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

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

Alpha多样性维基百科 https://en.wikipedia.org/wiki/Alpha_diversity
图表解读1箱线图：Alpha多样性 http://mp.weixin.qq.com/s/CkHVLzDVosKzoxFEIUpNWw
QIIME分析方法：http://mp.weixin.qq.com/s/FGvj_jLiGiiMcHZq4dQq5w

多样性相同 ≠ 物种相同

6.3. 筛选各组高丰度菌用于比较

```
#按组求均值，需根据实验设计metadata.txt修改组列名
#输出为特征表按组的均值-一个实验可能有多种分组方式
Rscript ${bin}/script/otu_mean.R --input result/otutab.txt --design metadata.txt \
    --group Group  --thre 0 --output result/otutab_mean.txt
    #如以平均丰度频率高于0.1%为筛选标准，得到每个组的OTU组合
awk 'BEGIN{OFS=FS="\t"}{if(FNR==1) {for(i=2;i<=NF;i++) a[i]=$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
    # 结果可以直接在http://www.ehbio.com/ImageGP绘制Venn、upSetView和Sanky
```

7. Beta多样性——样品间距离(差异)

- 物种距离： Bray-Curtis(Z-scores)、Euclidean

Beals, Edward W. "Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data." *Advances in ecological research*. Vol. 14. Academic Press, 1984. 1-55.

Cited by 891

- 进化距离：Unifrac，考虑进化关系

Catherine Lozupone, Rob Knight. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. ***Applied and Environmental Microbiology*** 71 (12):8228-8235, doi: doi:10.1128/AEM.71.12.8228-8235.2005

Cited by 6404

- 无权重 ： Unweighted(binary)，只考虑有无

标准化OTU表

	A	B
Bac1	6	4
Bac2	0	2

$$\sqrt{(6-4)^2+(0-2)^2}$$

	A	B	Taxonomy
Bac1	6	4	E. Coli 1
Bac2	0	2	E. Coli 2

非权重OTU表

	A	B
Bac1	1	1
Bac2	0	1

样品间距离矩阵计算

```
mkdir -p result/beta/

# 基于OTU构建进化树 Make OTU tree

usearch -cluster_agg result/otus.fa -treeout result/otus.tree

# 生成5种距离矩阵：bray_curtis，euclidean，jaccard，manhatten，
unifrac，又有非权重版本(_binary)

usearch -beta_div result/otutab_rare.txt -tree result/otus.tree \
  -filename_prefix result/beta/
```

8. 物种注释格式调整

USEARCH物种注释文件 result/otus.sintax: 特征分类和置信度，特征筛选后结果

ASV_1	d:Bacteria(1.00),p:"Actinobacteria"(1.00),c:Actinobacteria(1.00),o:Actinomycetales(1.00),f:	+	d:Bacteria,p:"Actinobacteria",c:Actinobacteria,o:Actin
ASV_2	d:Bacteria(1.00),p:"Proteobacteria"(1.00),c:Betaproteobacteria(1.00),o:Burkholderiales(1.(	+	d:Bacteria,p:"Proteobacteria",c:Betaproteobacteria,o
ASV_3	d:Bacteria(1.00),p:"Proteobacteria"(1.00),c:Gammaproteobacteria(0.87),o:Pseudomonad	+	d:Bacteria,p:"Proteobacteria",c:Gammaproteobacter

可以使用调为以下标准格式，这里使用Shell命令的sed, awk等命令组合调整

标准两列注释文件 result/taxonomy2.txt

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__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Rhizobacter;s__Rhizobacter_bergeniae
ASV_4	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Rhizobacter
ASV_6	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales

标准8列注释文件 result/taxonomy.txt

OTUID	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	Bacteria	Actinobac	Actinobac	Actinomyc	Thermom	Unassigne	Unassigned
ASV_2	Bacteria	Proteobac	Betaprote	Burkholde	Comamor	Pelomona	Pelomonas_puraquae
ASV_3	Bacteria	Proteobac	Gammapr	Pseudom	Pseudom	Rhizobact	Rhizobacter_bergeniae

8.物种注释结果分类汇总

```
# 统计门纲目科属，使用 rank参数 p c o f g
mkdir -p result/tax
for i in p c o f g;do
  usearch -sintax_summary result/sintax.txt \
  -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. 比对Greengene数据库(有参)生成OTU表

- Greengene数据库是最旧、最准，支持最广泛的数据库，基于它的OTUs表可进行功能预测(PICRUSt)和形态学预测(Bugbase)
- 生成OTU表 Create OTUs table

```
usearch -otutab temp/filtered.fa -otus ${db}/gg/97_otus.fasta \
    -otutabout result/gg/otutab.txt -threads 4
```

#OTU ID	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2	OE3	OE4	OE5	OE6	WT1	WT2	WT3	WT4	WT5	WT6
57759	81	43	414	143	436	78	47	437	287	175	63	76	258	310	186	81	199	275
810167	477	136	194	258	494	184	206	199	140	206	216	132	356	473	194	199	395	292
1134692	326	270	363	367	601	415	109	129	145	165	168	163	199	255	178	227	190	199
546343	3	2	4	0	1	8	5	2	1	2	1	5	0	0	3	0	1	2
48487	135	147	97	89	13	305	109	74	91	174	77	92	127	39	100	98	53	63
940737	651	415	875	775	1075	631	256	168	414	437	353	238	446	562	265	383	510	438
827300	2	0	1	1	0	0	2	0	0	1	2	0	0	0	0	0	1	0

OTU表统计

- OTU表统计

usearch -otutab_stats result/gg/otutab.txt -output gg/otutab.stat
cat gg/otutab.stat # 显示文件全部内容，适合小文本文件

```
214459  Reads (214.5k)
    18  Samples
  4623  OTUs

 83214  Counts
 58415  Count  =0  (70.2%)
 11216  Count  =1  (13.5%)
  3542  Count >=10 (4.3%)

   328  OTUs found in all samples (7.1%)
   423  OTUs found in 90% of samples (9.1%)
  1111  OTUs found in 50% of samples (24.0%)

Sample sizes: min 10849, lo 11438, med 11935, mean 11914.4, hi 12427, max 13247
```

10. 项目空间清理

```
#删除中间大文件
    rm -rf temp/*.fq

# 分双端统计md5值，用于数据提交
    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
    cd ..
    cat result/md5sum.txt
```

```
KO1_1.fq.gz     cda4f2efd86d52415405036adfce1c03     KO1_2.fq.gz     4634f5cc458c361888d3d9ee18ad1876
KO2_1.fq.gz     9328f79a2cf3326427d48e545b43db39     KO2_2.fq.gz     26163adcb1432565e68351e25f0070a3
KO3_1.fq.gz     da4dc7513a6535b57ed1eeccaec73536     KO3_2.fq.gz     27da8e494e076db8adab321117c26a37
KO4_1.fq.gz     98e9e6e78757b1b4a0d98b597eaf9b14     KO4_2.fq.gz     042325519f3e5b6a1a1d0fe8da572e74
KO5_1.fq.gz     ec39201b9781bf9f3f72580ee3468600     KO5_2.fq.gz     2963eee1181bbdcbf3cdac92add02fa1
KO6_1.fq.gz     744fcfe4705974330fd7aac61c16ca0a     KO6_2.fq.gz     adf68ea71887728431252eb9b61f0d3b
```

三个重要结果文件(result)

- **特征表**：样本与Feature(OTU/ASV)对应reads count矩阵 otutab.txt

- **代表序列**：每个OTU中选择的代表性或ASV序列，无参为最高丰度，有参按97%聚类选择中心序列 otus.fa

- **物种注释**：每个OTU/ASV的物种注释，一般包括域/界、门、纲、目、科、属种。但其中有很多为未注释(unassigned/unclassified) taxonomy.txt

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(如RDP)比对，确定分类层级和置信度
- 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)

进一步阅读

- 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》——30+篇
- Bio-protocol《微生物组实验手册》计划——200+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
- 扩增子图表解读 分析流程 统计绘图
- QIIME2中文教程-把握分析趋势
- 扩增子16S分析专题研讨论会——背景介绍

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识