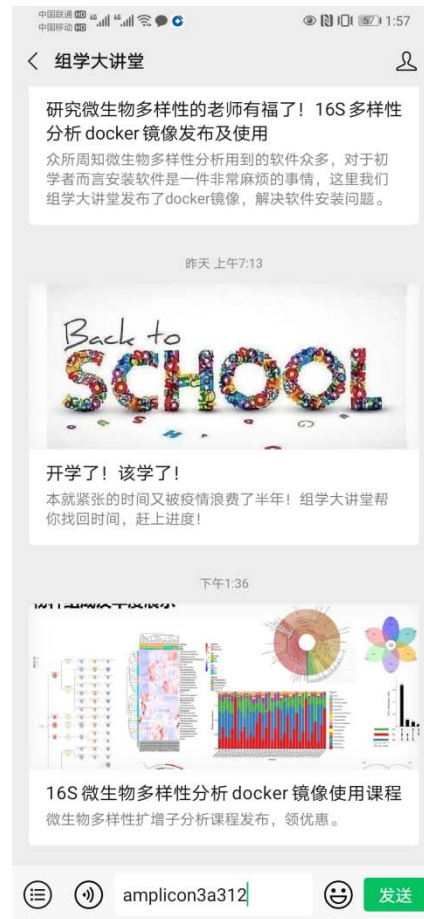


课程资料与代码领取方法

- 关注组学大讲堂服务号（微信扫码下方二维码关注）
 - 可获取课程更新、优惠、通知等
- 回复暗号：amplicon3a312
 - 即可获得资料下载地址





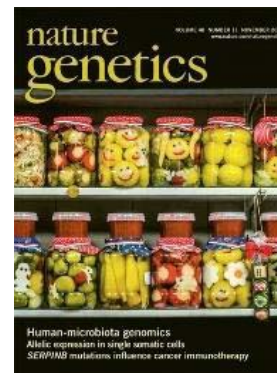
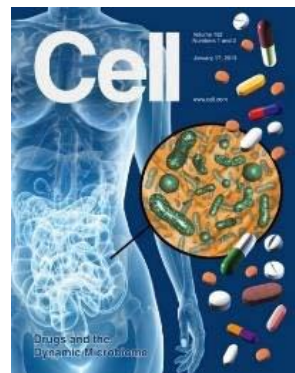
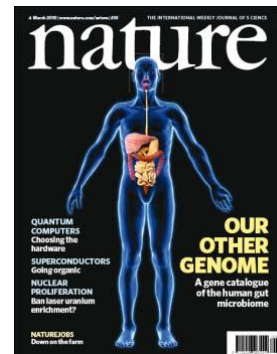
微生物多样性数据分析实操 (16S/ITS/18S) 高通量测序

--omicsgene



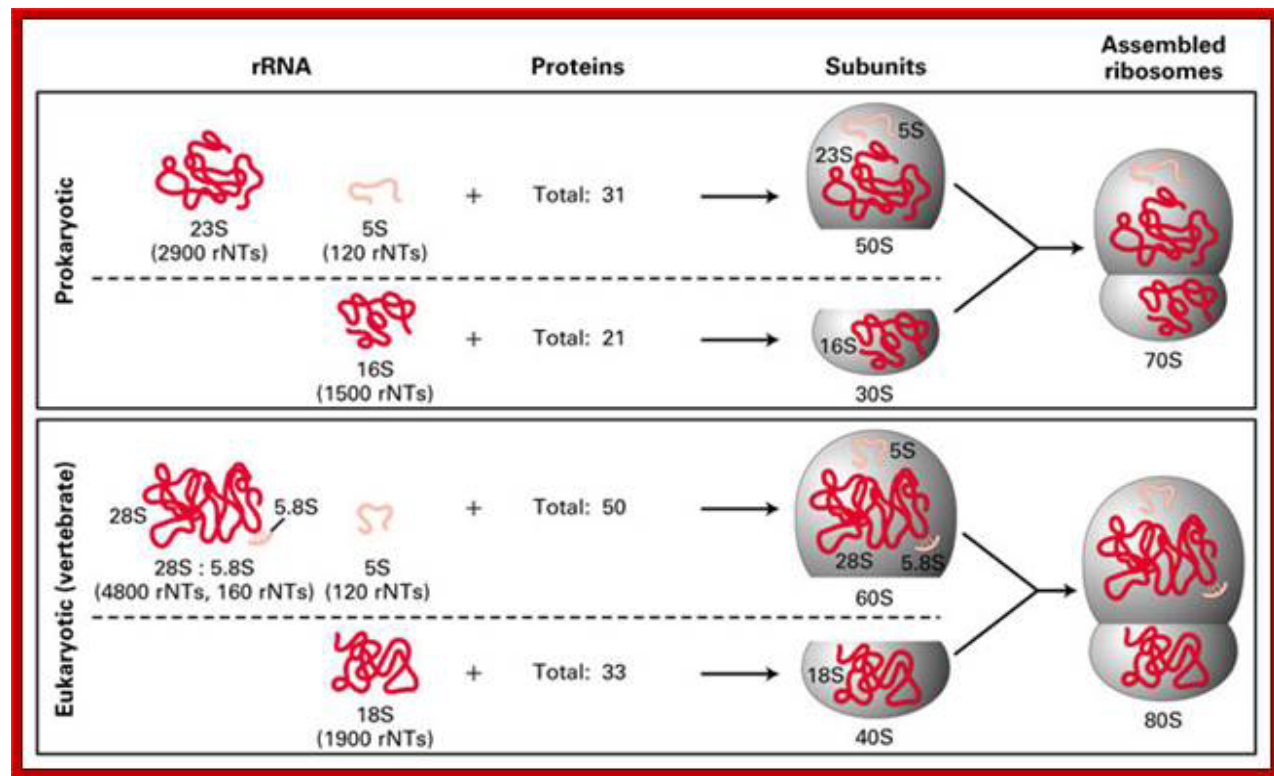
微生物多样性分析简介

微生物多样性测序应用范围



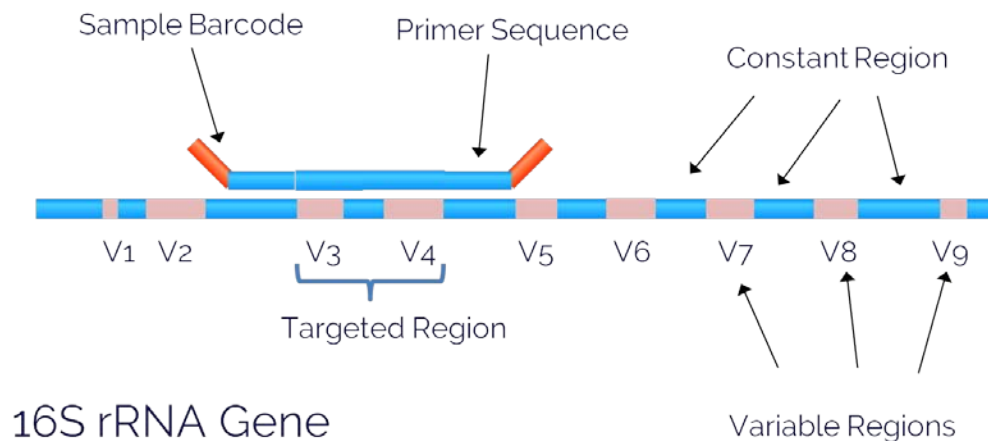
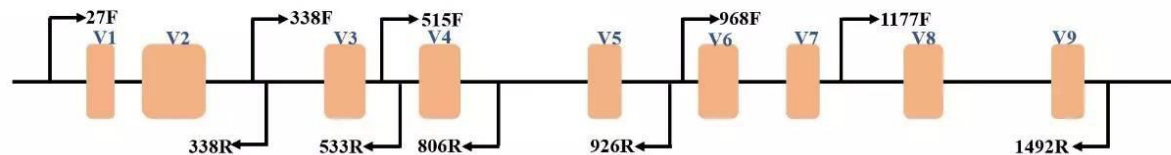
环境中的微生物-核糖体rRNA基因

- 原核 细菌
- 真核 真菌



原核16S 选择测序区域 V3+V4?

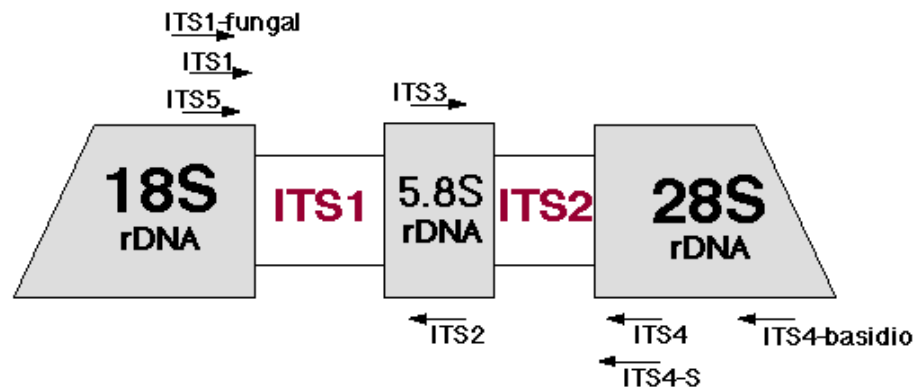
1-68	69-99	137-242	433-497	576-682	822-879	986-1043	1117-1173	1243-1294	1435-1465	1466-1452
	V1	V2	V3	V4	V5	V6	V7	V8	V9	



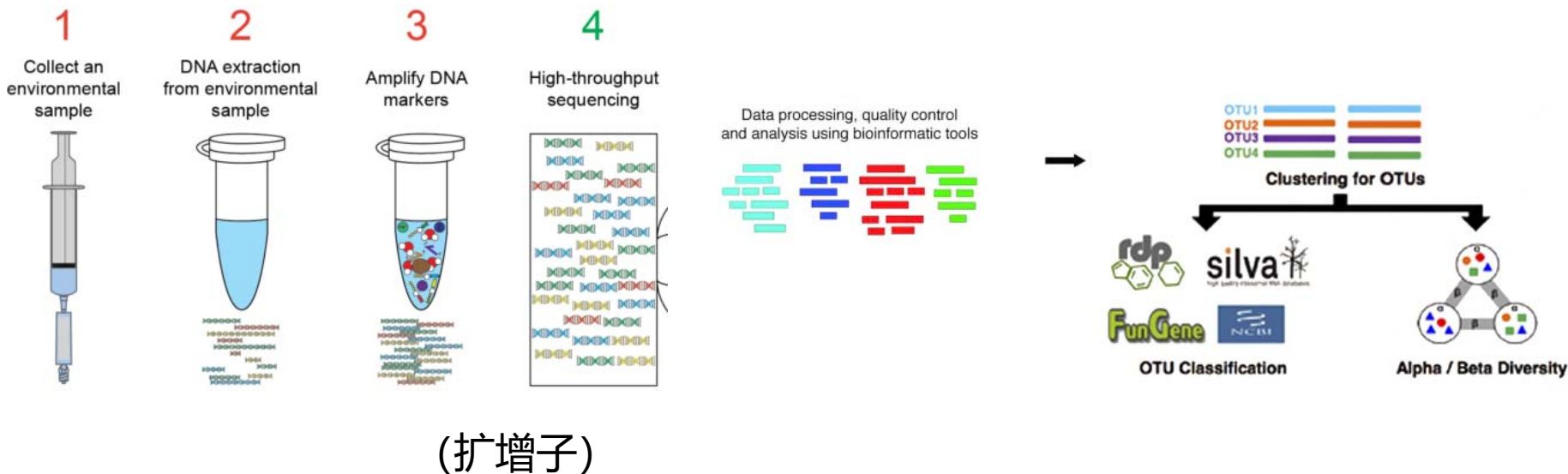
16S rRNA Gene

真菌ITS，真核18S

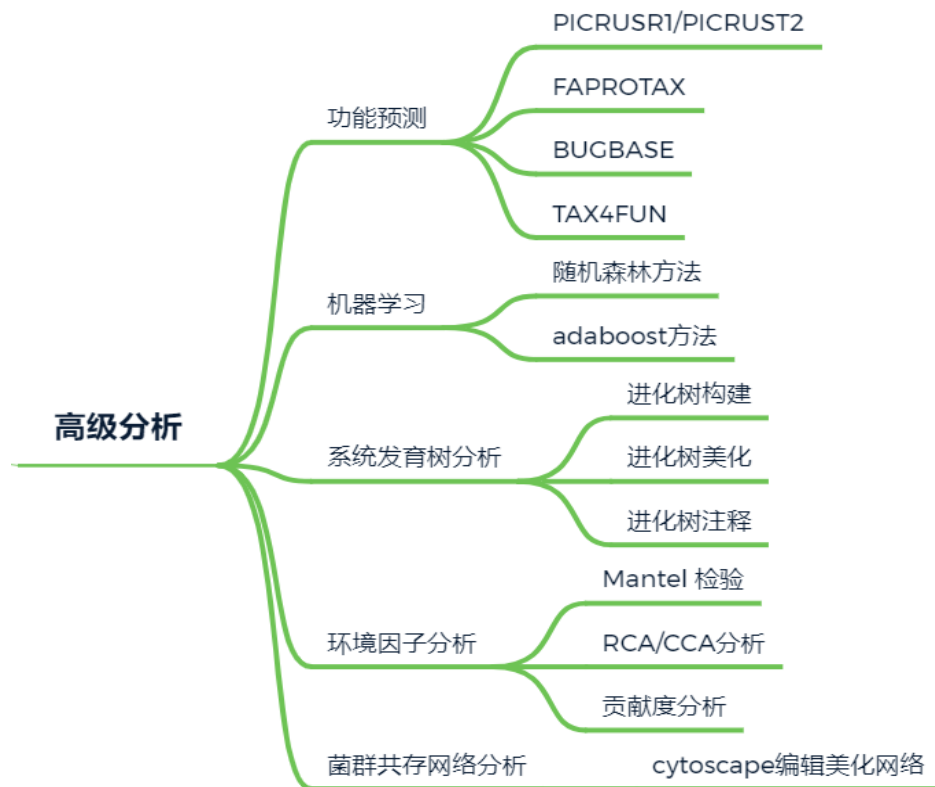
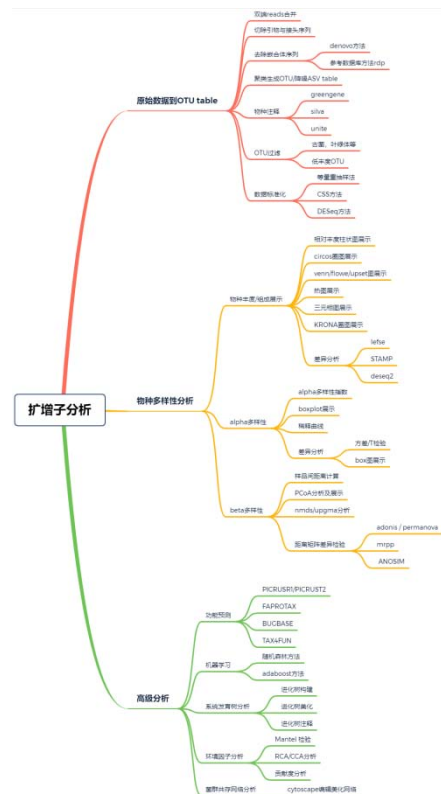
- ITS测序主要是针对真菌多样性的研究，注释准确度高；
- 18S测序针对的是真核微生物，注释到的范围广，但是就真菌来说精确度相对较低；
- 可根据研究目的合理选择测序方案。



微生物多样性分析主要步骤



微生物多样性分析主要内容



更详细微生物多样性分析原理与结果解读



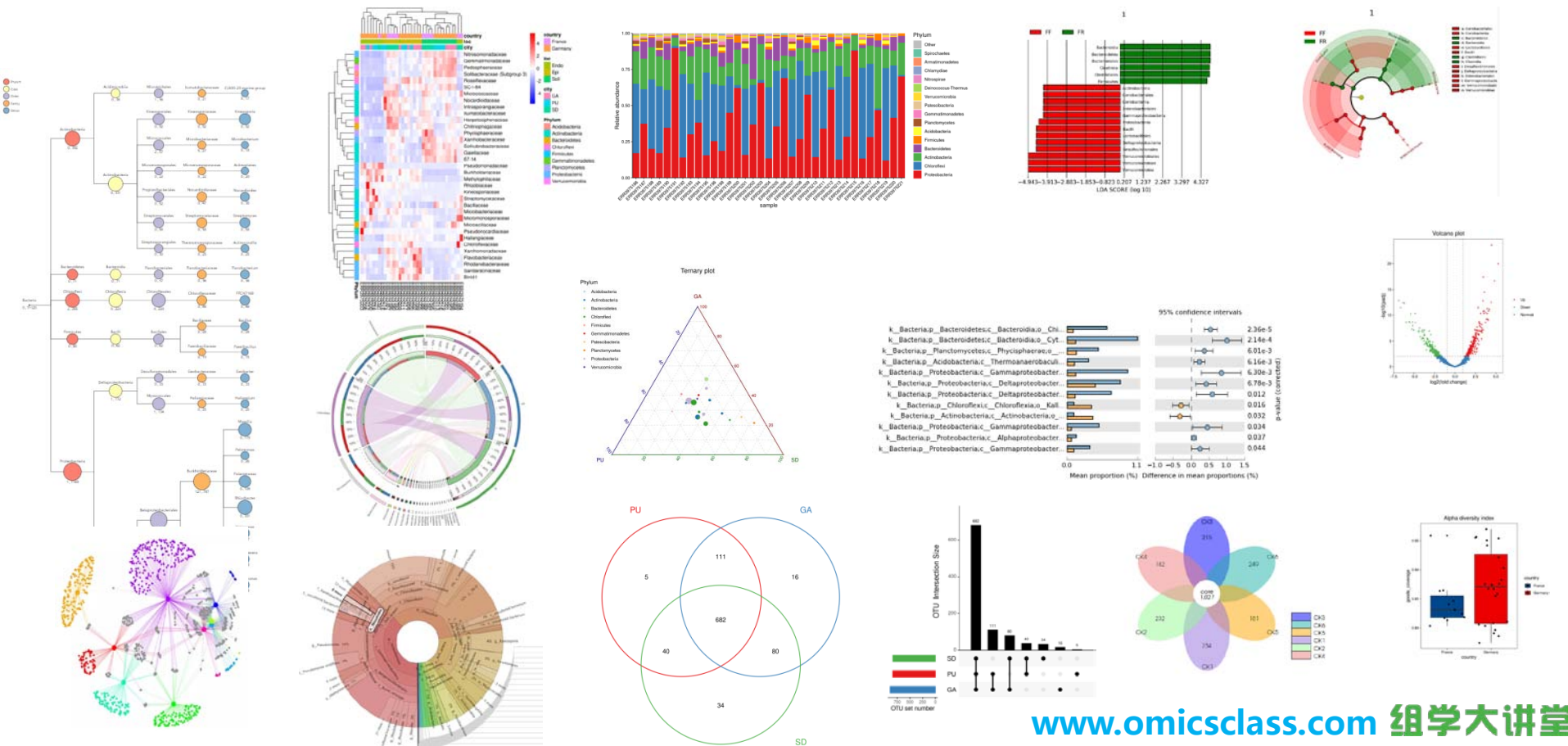
<https://study.163.com/course/introduction/1005831025.htm?share=1&shareId=1030291076>

课程特色

- 不需要为安装软件而烦恼
- 跨平台，任何操作系统都可以分析
- 学员可完全重现老师课上所有结果



不仅仅跑跑命令还有各种个性化绘图





数据分析环境搭建

计算机软件分析环境搭建

■ 练习用环境搭建 (windows练习)

- Win10+Docker Desktop:

- 笔记:

<https://www.omicsclass.com/article/1198>

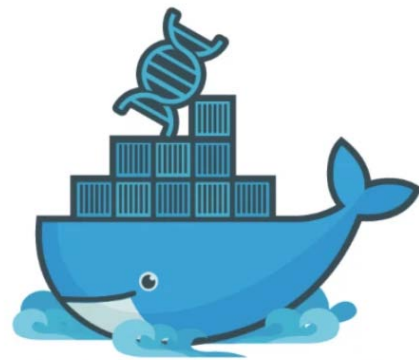
■ 真实数据Linux服务器环境搭建:

- Linux服务器+ Docker:

<https://www.omicsclass.com/article/1171>

■ Docker使用笔记：

- <https://www.omicsclass.com/article/1181>



docker for bioinformatics



计算机硬件建议

■ 系统软硬件建议：

- **练习测试**：内存8G以上，硬盘至少10G以上空余；win10系统+Docker
- **真实数据分析**：内存20G以上，CPU四核以上，存储1T以上；建议Linux服务器+docker





推荐学习生物信息基础课



Linux基础课



Python 编程基础



R 语言基础入门



R 语言绘图ggplot2等

扩增子分析docker镜像下载

- 老师系统环境：win10+docker desktop
- 下载扩增子分析镜像：
 - docker images #查看后台镜像
 - docker search omicsclass #搜索镜像
 - docker pull omicsclass/ampliseq-q1:v1.2 #下载镜像
- 启动镜像（Docker Desktop）：
 - docker run -it -m 3G --cpus 1 --rm -v D:/ampliseq-q1:/work omicsclass/ampliseq-q1:v1.2
- 启动镜像（Docker toolbox 需要设置共享D盘）：
 - docker run -it -m 3G --cpus 1 --rm -v /d/ampliseq-q1:/work omicsclass/ampliseq-q1:v1.2

分析数据介绍与准备

ARTICLES

<https://doi.org/10.1038/s41559-019-1063-3>

nature
ecology & evolution

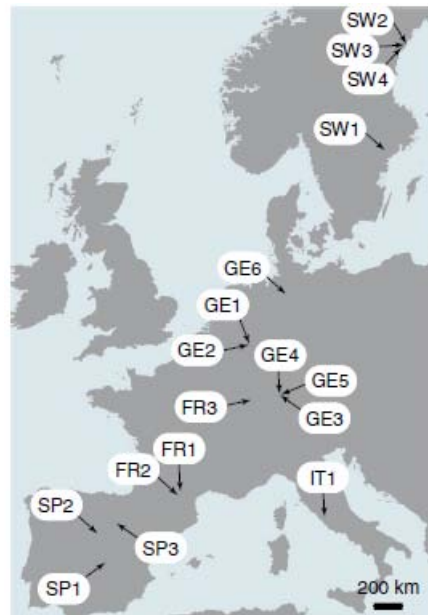
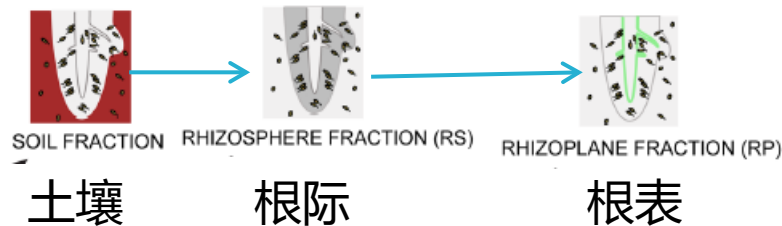
Root microbiota assembly and adaptive differentiation among European *Arabidopsis* populations

Thorsten Thiergart^{1,7}, Paloma Durán^{1,7}, Thomas Ellis², Nathan Vannier¹, Ruben Garrido-Oter^{1,3}, Eric Kemen⁴, Fabrice Roux⁵, Carlos Alonso-Blanco⁶, Jon Ågren^{2*}, Paul Schulze-Lefert^{1,3*} and Stéphane Hacquard^{1,3*}

■ 测序数据

■ 物种注释数据库

■ 下载测试数据及代码





原始数据到OTU表格

分析主要内容

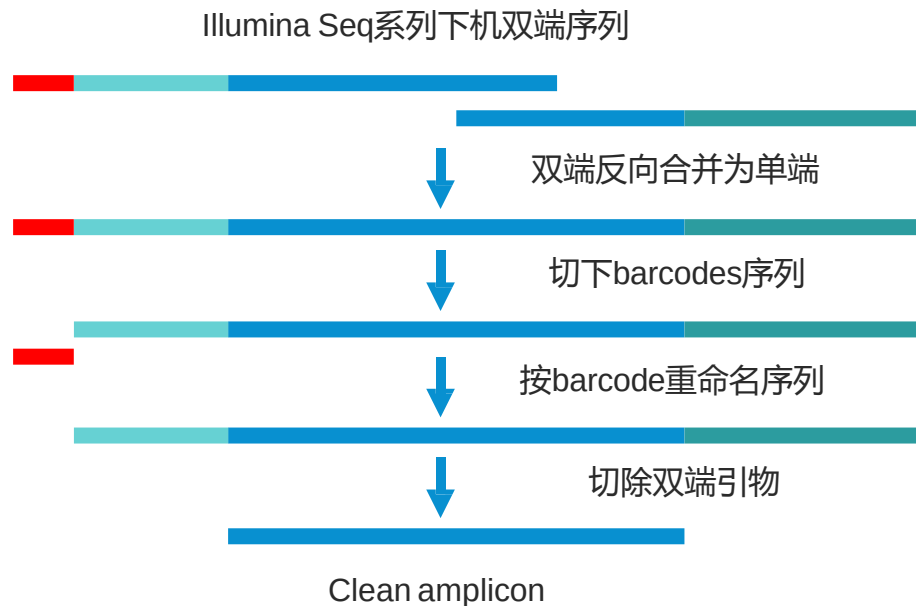
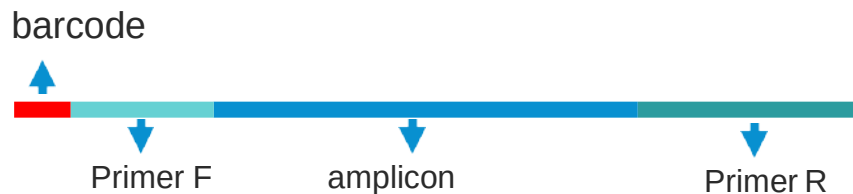
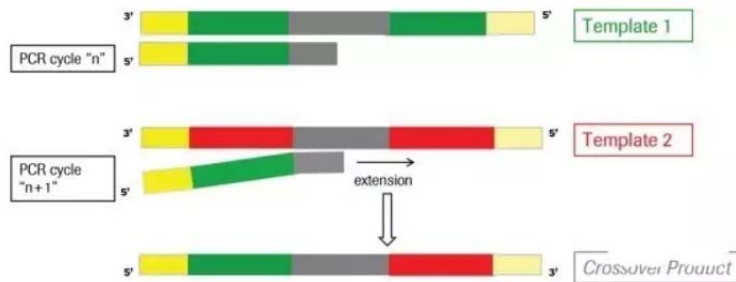


数据前期处理

- 双端测序数据合并
- Barcode 引物去除
- 去除嵌合体

扩增子结构与嵌合体

- 扩增子：
- 嵌合体



<https://www.omicsclass.com/article/8>

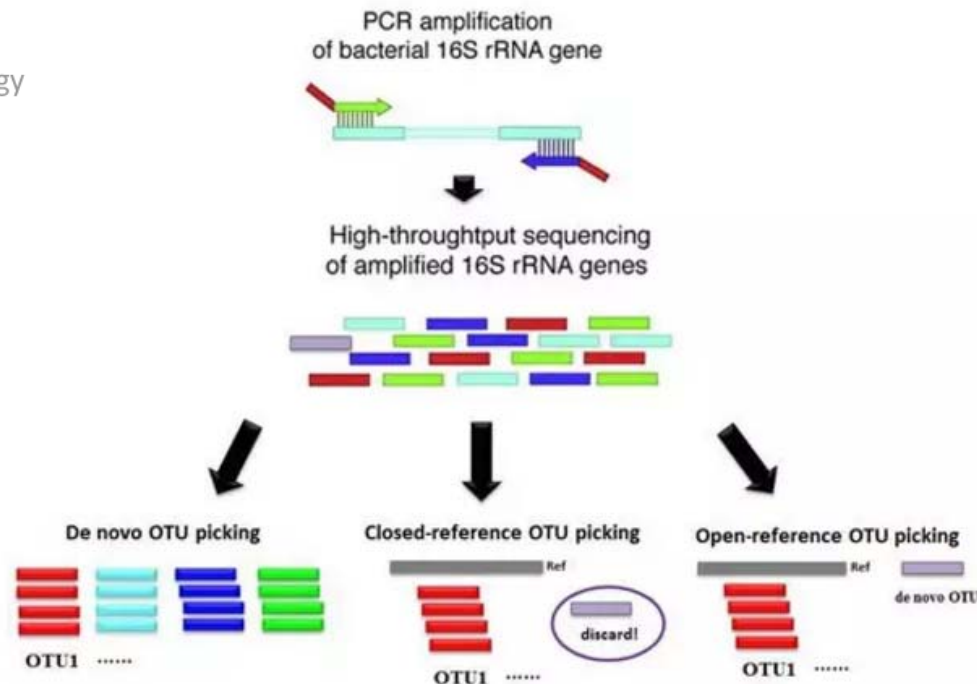
聚类生成OUT-分析流程及数据库选择



qiime1 聚类生成OTU的三种方法:



- de novo OTU 聚类
- open-reference OTU聚类
- closed-reference聚类

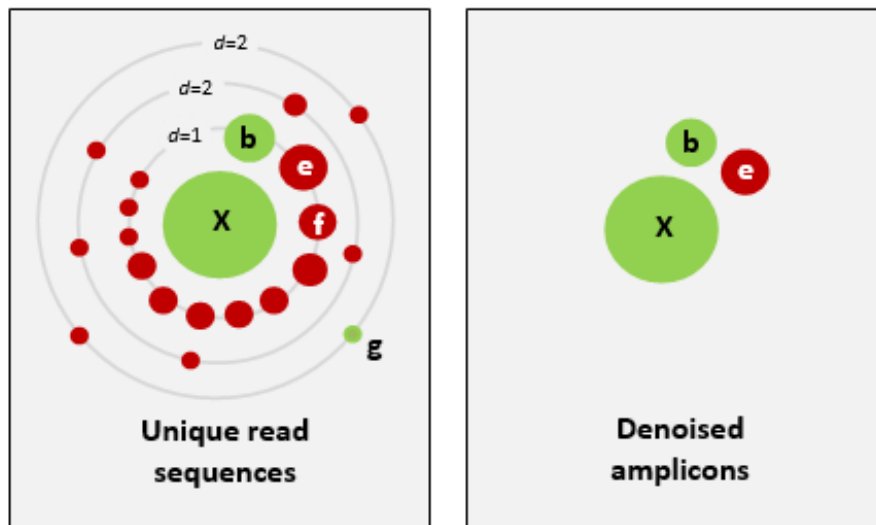


<https://www.omicsclass.com/article/8>

OTU/ASV概念

- OTU (Operational Taxonomic Units) 是人为给某一个分类单元 (品系, 种, 属, 分组等) 设置的同一标志。要了解一个样品测序结果中的菌种、菌属等数目信息, 就需要对序列进行聚类操作 (cluster)。通过聚类操作, 将序列按照彼此的相似性分归为许多小组, 一个小组就是一个OTU。
- 通常在97%的相似水平下聚类生成OTU, 选择每个聚类群中最高丰度序列作为代表性序列
- 近期讨论发现100%更合理, 即不聚类的ASV(Amplicon Sequence Variants), 更容易实现跨研究比较。

OTU / ASV分析原理



OTU table

Denoise

- Unoise3
- deblur
- data2

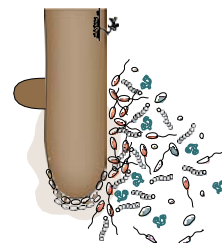
Feature(特征) table

标准化OTU Table

- 等量重抽样
- 相对丰度
- 其他: CSS, DESeq2 方法

	SampleA	SampleB
BacRed	6	8
BacGreen	3	4
BacBlue	1	6
BacPurple	0	2

SampleA



等量重抽样: 比较物种多样性

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

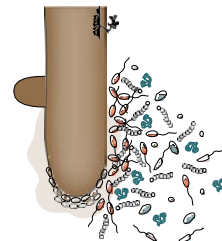
多样性指数: A的丰富度为3, 而B为4

相对丰度: 比较相比比例多少

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

样品间菌相关丰度存在差异

SampleB





物种多样性分析

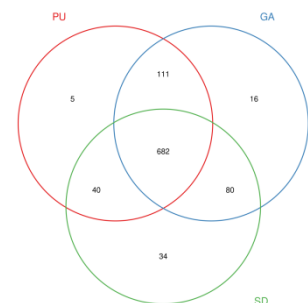
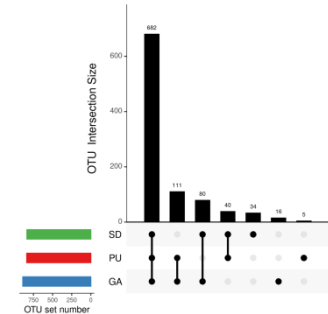
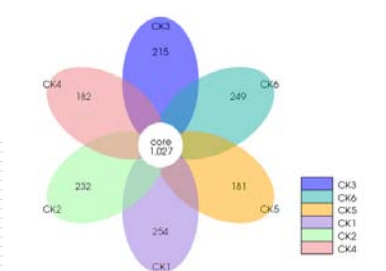
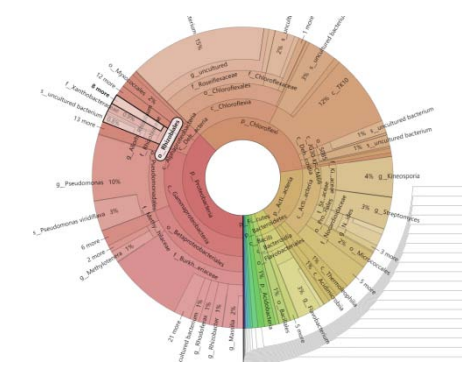
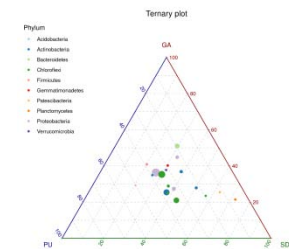
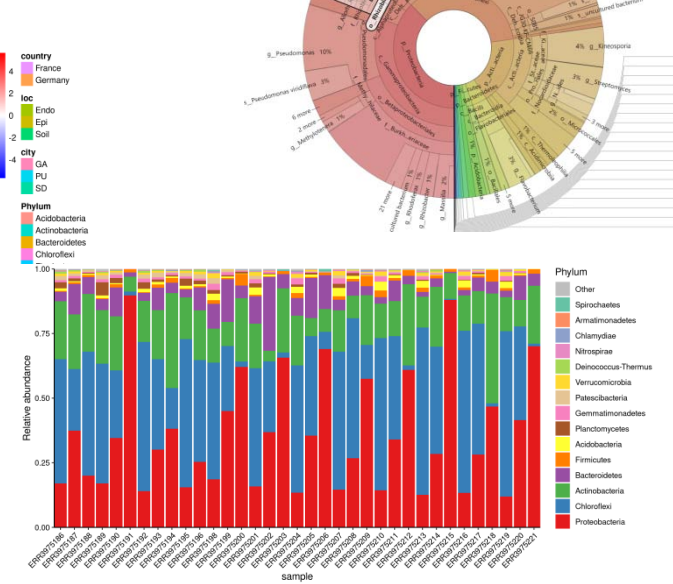
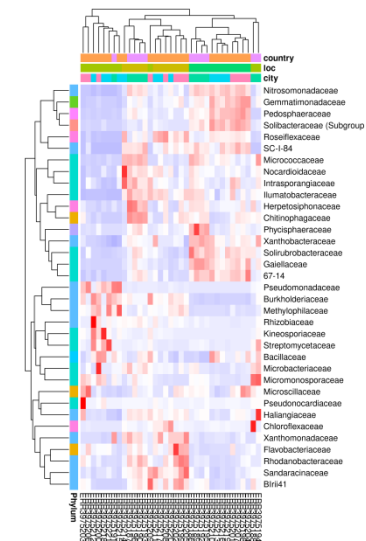
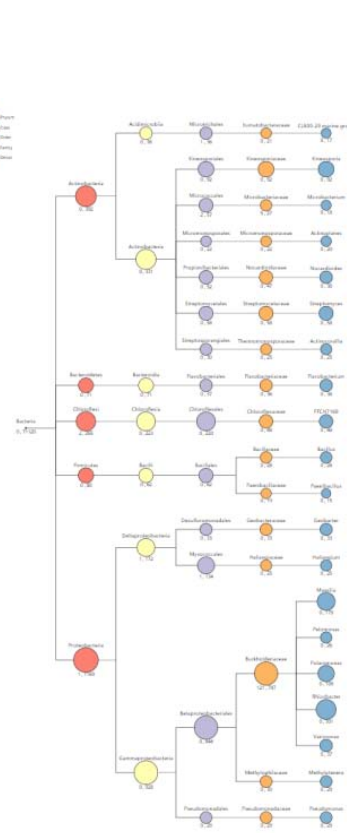
分析主要内容



物种丰度/组成展示



物种组成及丰度信息展示

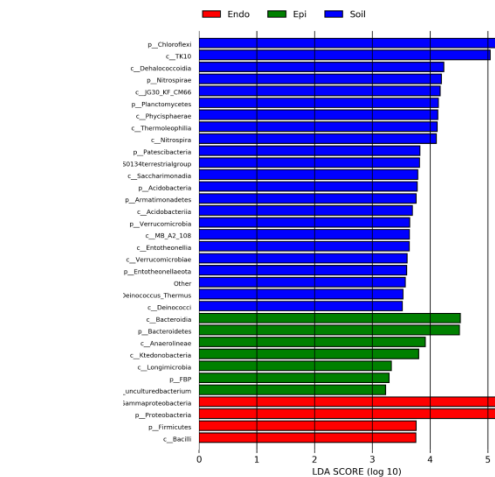
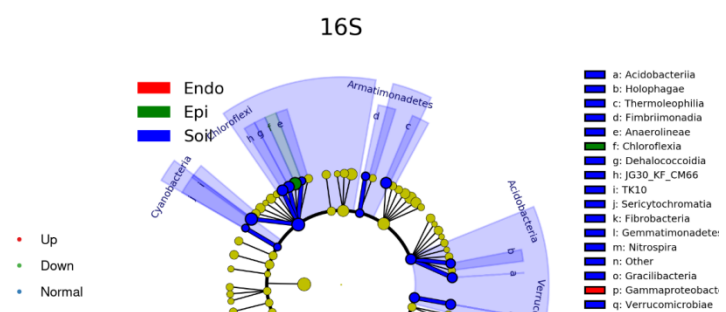




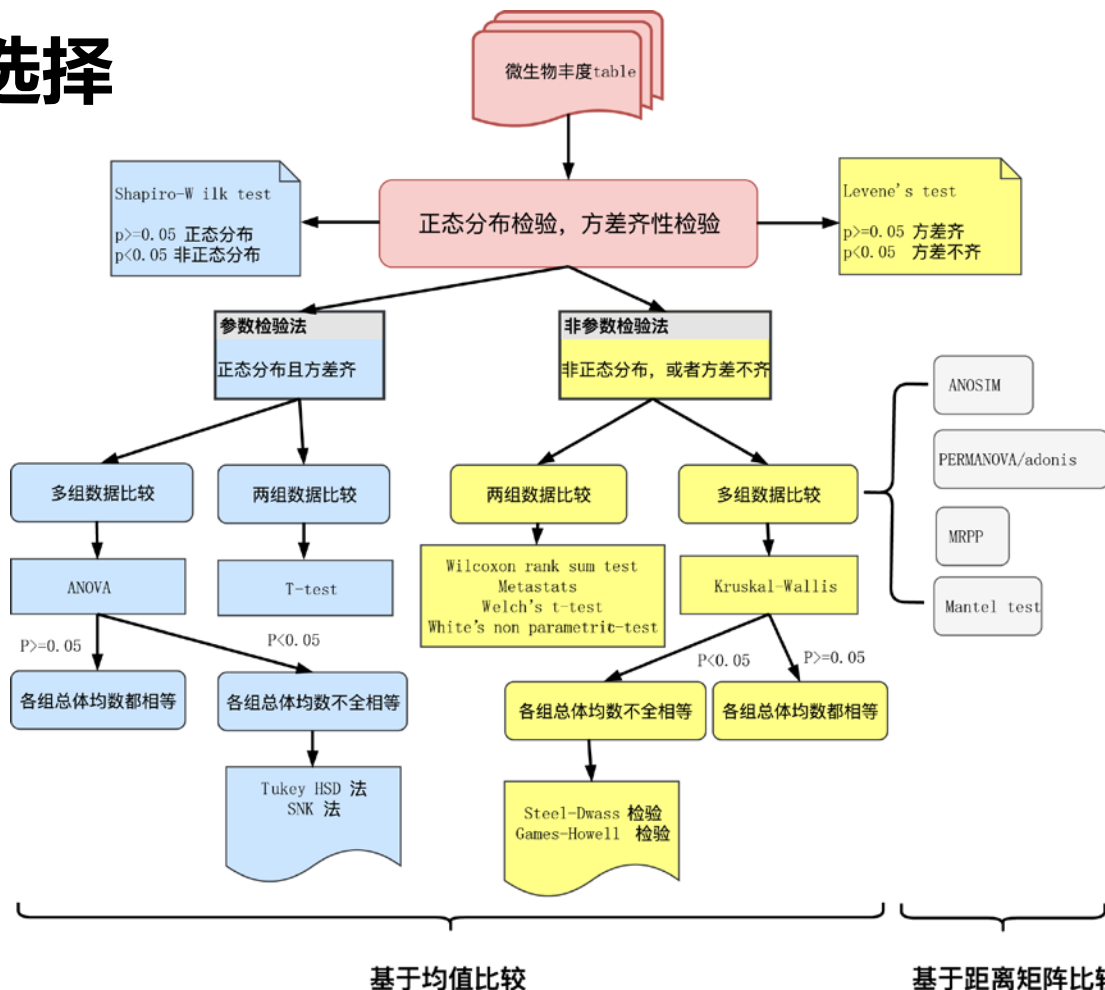
OTU物种丰度差异分析

- *lefse* LDA

- `deseq2`



比较方法的选择



Metastats结果

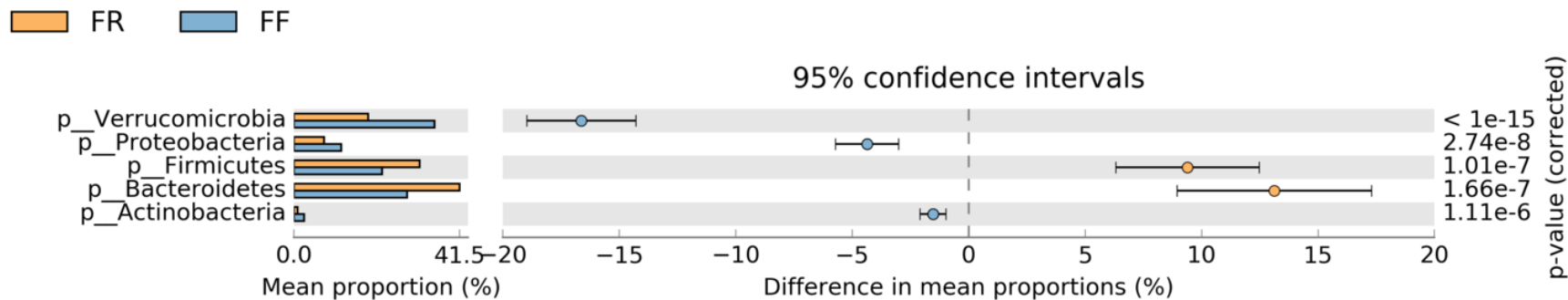
Level_name	Mean(group1)	Variance(group1)	Std.err(group1)	Mean(group2)	Variance(group2)	Std.err(group2)	P_value	Q_value
Bacilli	2.87E-04	2.33E-07	8.82E-05	2.81E-02	8.35E-04	4.41E-03	9.99E-04	1.17E-03
Bacteroidia	4.15E-01	1.09E-02	1.91E-02	2.84E-01	2.59E-03	7.76E-03	9.99E-04	1.17E-03
Clostridia	3.14E-01	4.68E-03	1.25E-02	1.92E-01	3.03E-03	8.39E-03	9.99E-04	1.17E-03
Coriobacteriia	9.59E-03	3.05E-05	1.01E-03	2.49E-02	2.90E-04	2.60E-03	9.99E-04	1.17E-03
Deltaproteobacteria	1.76E-02	4.89E-05	1.28E-03	4.54E-02	3.49E-04	2.85E-03	9.99E-04	1.17E-03
Verrucomicrobiae	1.86E-01	2.00E-03	8.17E-03	3.52E-01	3.06E-03	8.44E-03	9.99E-04	1.17E-03
Gammaproteobacteria	5.73E-02	9.22E-04	5.54E-03	7.31E-02	6.90E-04	4.01E-03	2.80E-02	2.80E-02



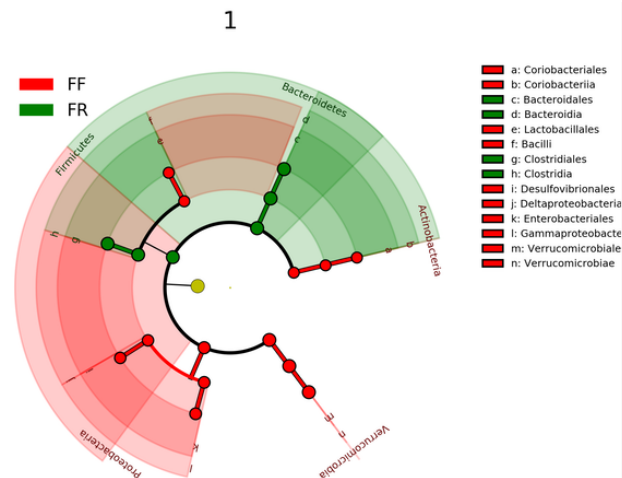
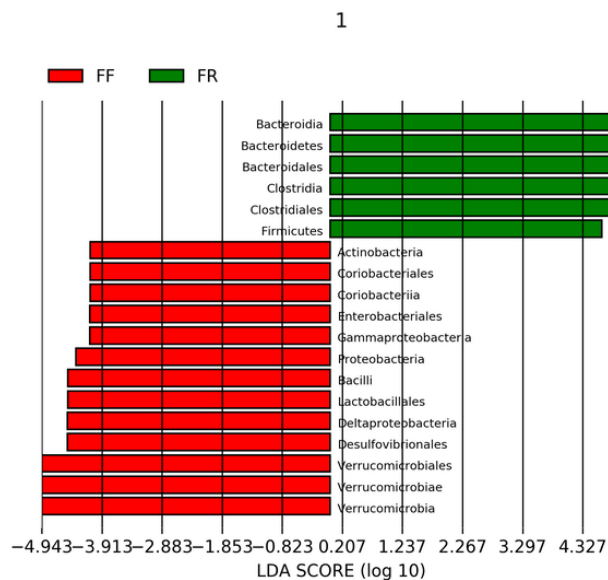
STAMP软件分析

参数检验法	两组比较	T-test
	多组比较	ANOVA
非参数检验法	两组比较	Welch' s t-test White' s non-parametric t-test
	多组比较	Kruskal-Wallis H-test

STAMP结果



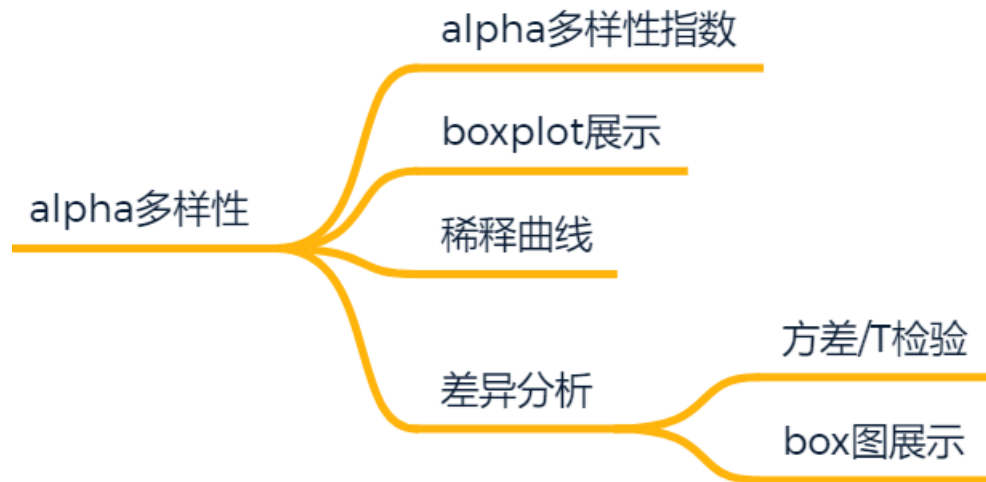
Lefse分析结果





ALPHA多样性分析

ALPHA多样性分析内容



alpha多样性指数

物种丰富度（有无）和均匀度（相对丰度）

■ 计算菌群丰富度（Community richness）的指数有：

Chao1：是用chao1 算法估计群落中含OTU 数目的指数，chao1 在生态学中常用来估计物种总数，由Chao (1984) 最早提出。

Chao1值越大代表物种总数越多，表明样品的丰富度越高。

Ace：是用来估计群落中含有OTU 数目的指数，同样由Chao提出(Chao and Yang, 1993)，是生态学中估计物种总数的常用指数之一。默认将序列量10以下的OTU都计算在内，从而估计群落中实际存在的物种数。**ACE指数越大，表明样品的丰富度越高。**

Species richness (Observed OTU, observed species)：物种丰富度指数为群落中丰度大于0的物种数之和，但只有物种种类信息，没有丰度信息；

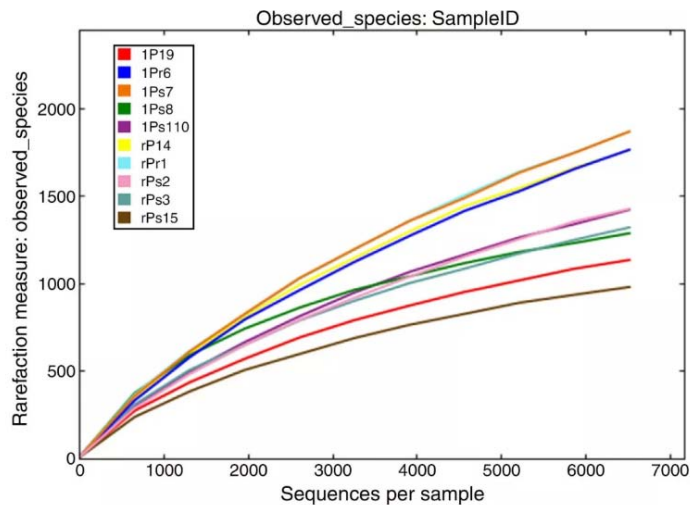
■ 计算菌群多样性（Community diversity）的指数有：

Shannon：(Shannon, 1948a, b) 综合考虑了群落的丰富度和均匀度。**Shannon指数值越高，表明样品的多样性越高。**

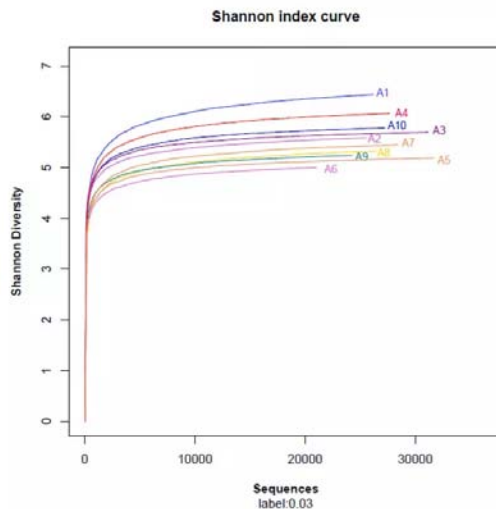
Simpson：用来估算样品中微生物的多样性指数之一，由Edward Hugh Simpson (1949) 提出，在生态学中常用来定量的描述一个区域的生物多样性。表示随机选取两条序列属于同一个分类（如OTUs）的概率（故数值在0~1之间，**Simpson 指数值越大，说明群落多样性越低。**

- 更多指数：<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html>

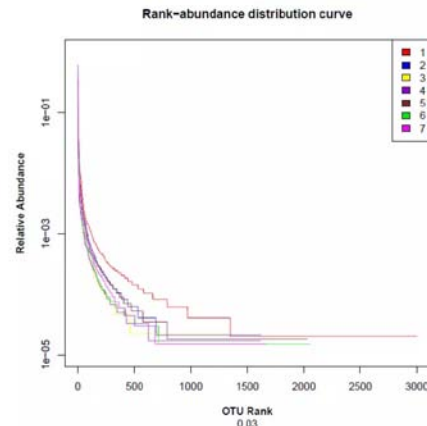
稀释曲线



注：横坐标代表随机抽取的序列数量；纵坐标代表观测到的OTU数量。样本曲线的延伸终点的横坐标位置为该样本的测序数量，如果曲线趋于平坦表明测序已趋于饱和，增加测序数据无法再找到更多的OTU；反之表明不饱和，增加数据量可以发现更多OTU。



注：与上圖一样，横坐标代表随机抽取的序列数量；纵坐标代表的是反映物种多样性的Shannon指数。



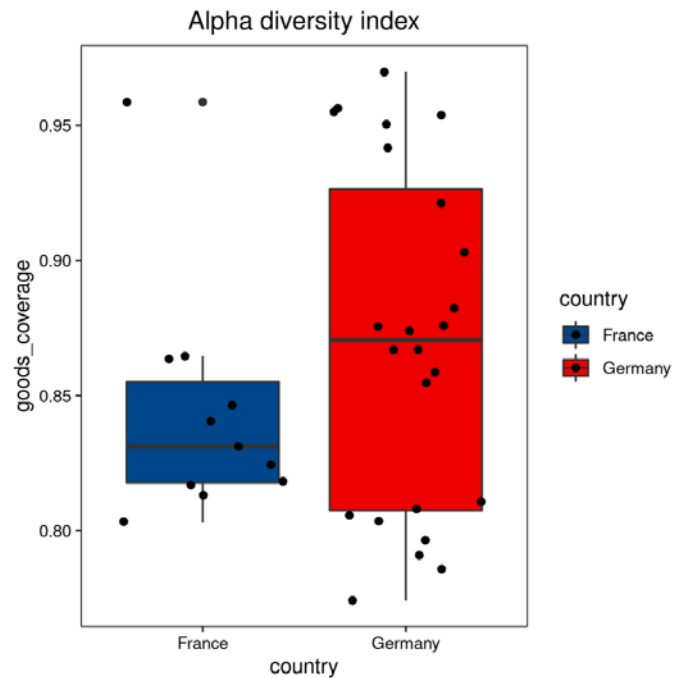
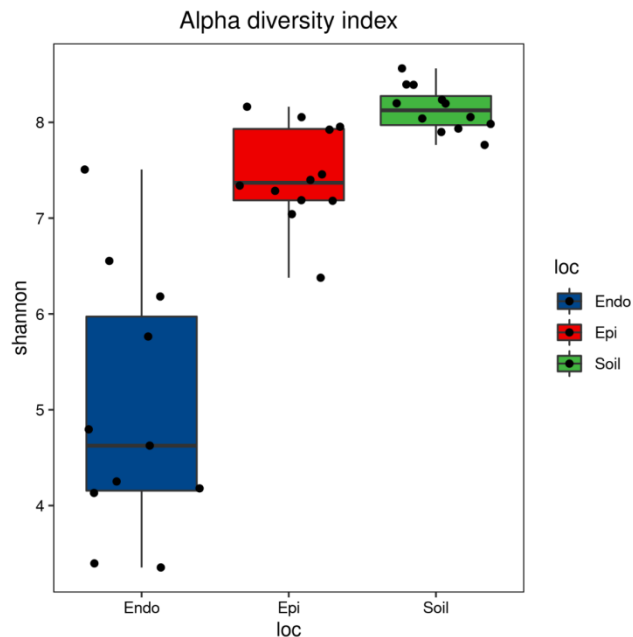
注：横坐标代表物种排序的数量；纵坐标代表观测到的相对丰度。样本曲线的延伸终点的横坐标位置为该样本的物种数量，如果曲线趋于平坦下降表明样本的物种多样性饱和，而曲线快速陡峭下降表明样本中的优势菌所占比例很高，多样性较低。

alpha多样性指数差异分析

■ T.test

■ ANOVA

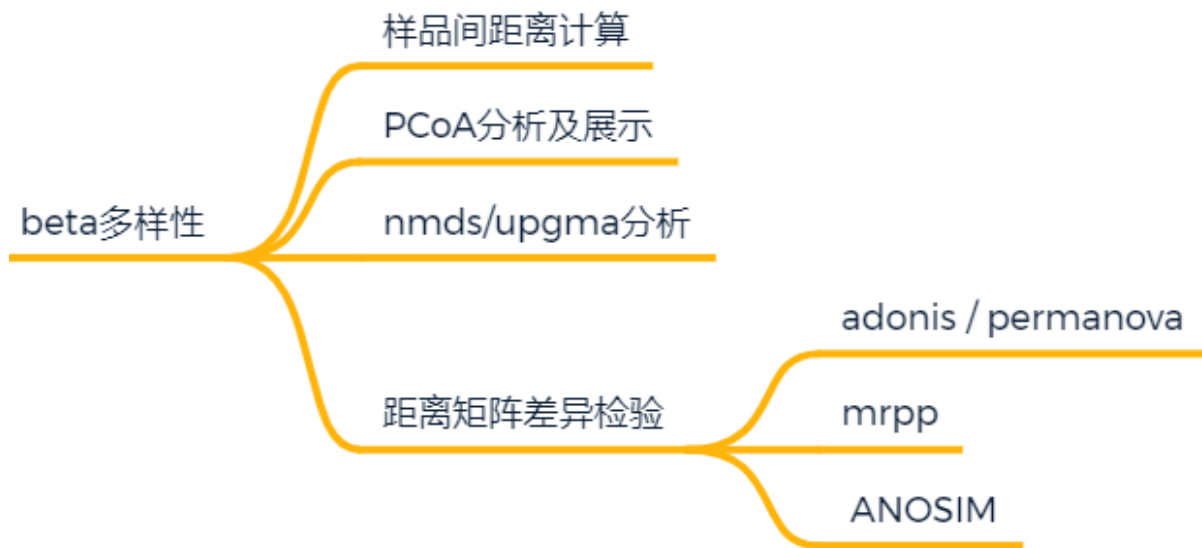
■ wilcox





Beta多样性分析

Beta多样性分析内容



排序方法总结

Beta多样性是生态学概念，专指不同组或生态位间物种组成的差异。常用排序降维的方法进行分析：

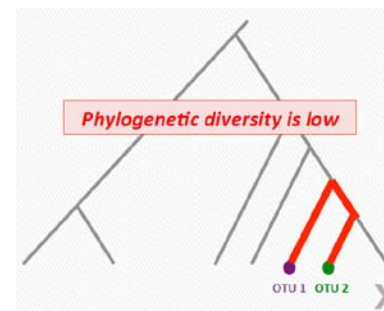
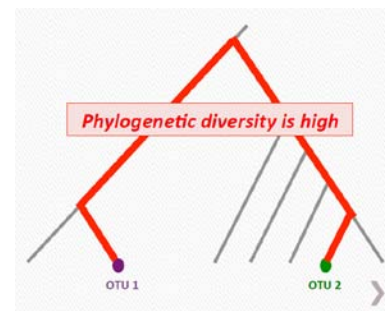
	Raw data based		distance based
	Linear model	Unimodal	
unconstrained ordination (indirect gradient analysis)	PCA	CA, DCA	PCoA, NMDS
constrained ordination (direct gradient analysis)	RDA	CCA, DCCA	CPCoA, db-RDA (CAP)

区别与联系：<https://www.omicsclass.com/article/148>

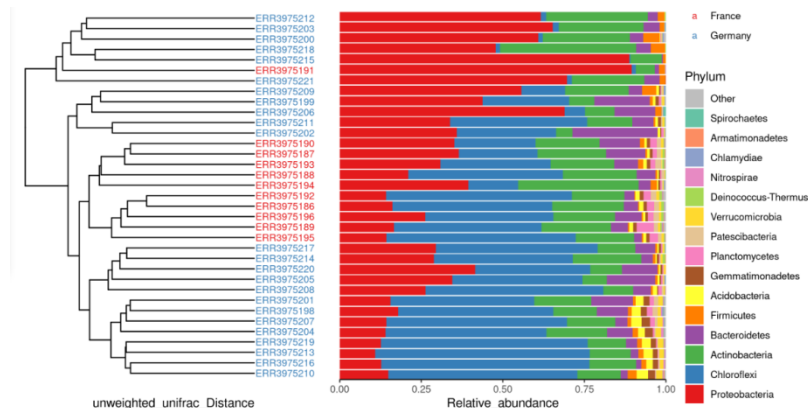
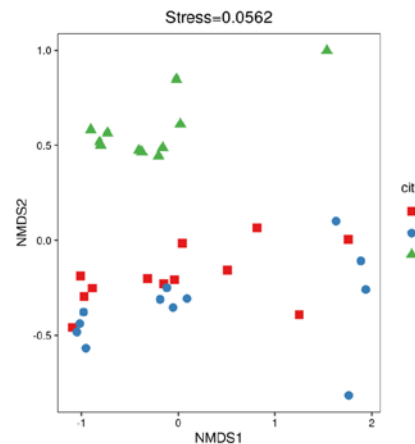
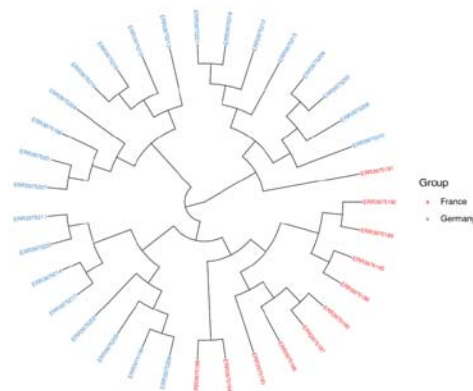
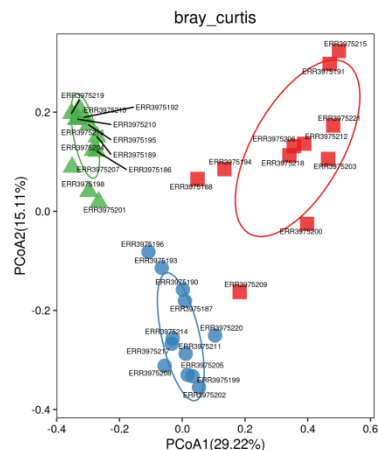
Beta多样性距离介绍

Beta 多样性：不同生态系统之间多样性比较：
物种丰富度（有无）和均匀度（相对丰度）

	Presence/Absence	Abundance
Without a phylogenetic tree	Jaccard	- Bray-Curtis - Euclidean (PCA)
With a phylogenetic tree	Unweighted UniFrac	Weighted UniFrac

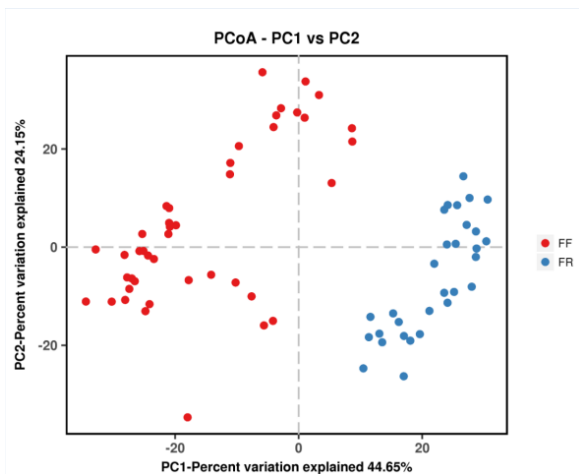


降维排序聚类分析 (PCA,PCoA,NMDS,UPGMA等)



PCoA图配合检验

Braycurtis PCoA

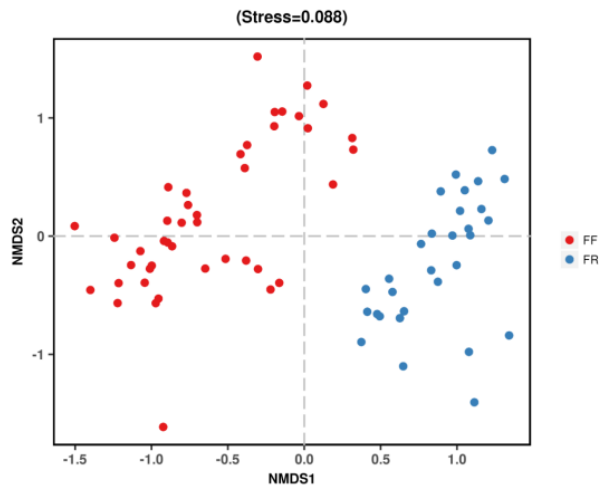


Adonis/PERMANOVA

R^2 0.59735

p-value 0.001

Braycurtis NMDS



ANOSIM

Statistic 0.7529785286381386

p-value 0.001

Adonis/PERMANOVA

■ PERMANOVA/adonis 多元方差分析

- Adonis又称置换多元方差分析或非参数多元方差分析。它利用距离矩阵（如Bray-Curtis, Euclidean等）对总方差进行分解，分析不同分组因素对样品差异的解释度，并使用置换检验对其统计学意义进行显著性分析。

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	
group_factor\$diet	2	0.715	0.35749	4.4185	0.11209	0.001	***
Residuals	70	5.6635	0.08091	0.88791			
Total	72	6.3785	1				

注：其中，group_factor\$*, *表示分组方案；Df，表示自由度；SumsOfSqs，总方差，又称离差平方和；Mean Sqs，表示均方（差），即SumsOfSqs/Df；F.Model，表示F检验值；R²，表示不同分组对样品差异的解释度，即分组方差与总方差的比值，R²越大表示分组对差异的解释度越高；Pr，表示P值，小于0.05说明本次检验的可信度高。

ANOSIM

- ANOSIM(analysis of similarities)相似性分析是一种非参数检验，用来检验组间（两组或多组）差异是否显著大于组内差异，从而判断分组是否有意义。利用两两样品间的距离，并对距离从小到大进行排序（秩次排序），然后进行检验。再用**置换检验是Permutation Test得到P值**：

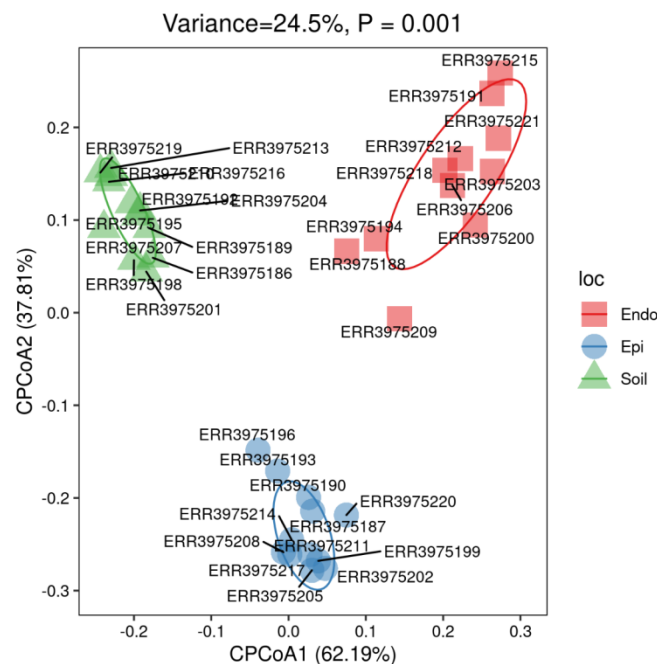
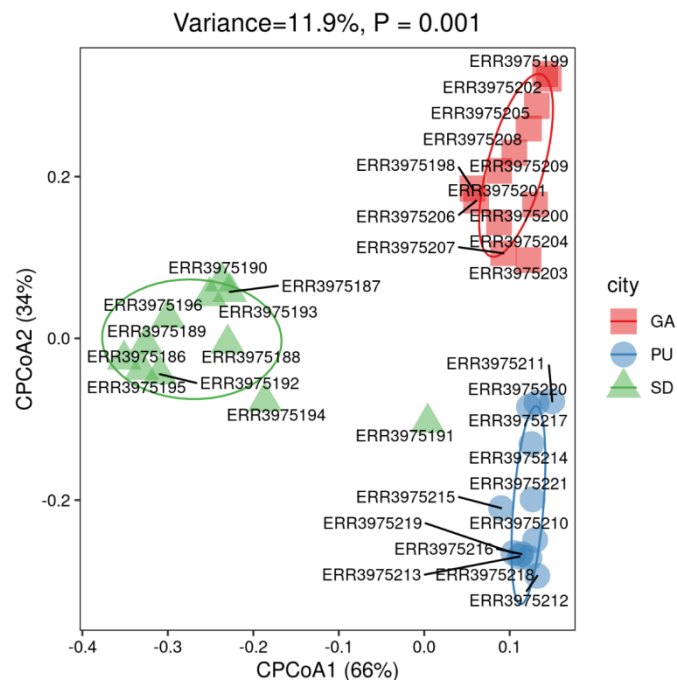
■

sample size	test statistic	p-value	number of permutations
73	0.161838	0.001	999

注：Statistic即为R值，范围为-1到+1。R值越接近1表示组间差异越大于组内差异，R值越小则表示组间和组内没有明显差异，并且当P值小于0.05时说明检验的可信度高。如果R=0，表明样本的分组效果等同于随机分配，各样本分组之间不具有可观测的统计学差异；如果R为负值，则表明组内差异超过了组间差异的大小，预示分组效果较差。P值则反映了ANOSIM分析结果的统计学显著性，**P值越小，表明各样本分组之间的差异显著性越高。**

限制性主成分分析CPCoA

- 寻找某一条件下，可最大限制解释这一条件的投影平面。





PICRUSt 细菌功能预测

PICRUSt 细菌功能预测原理

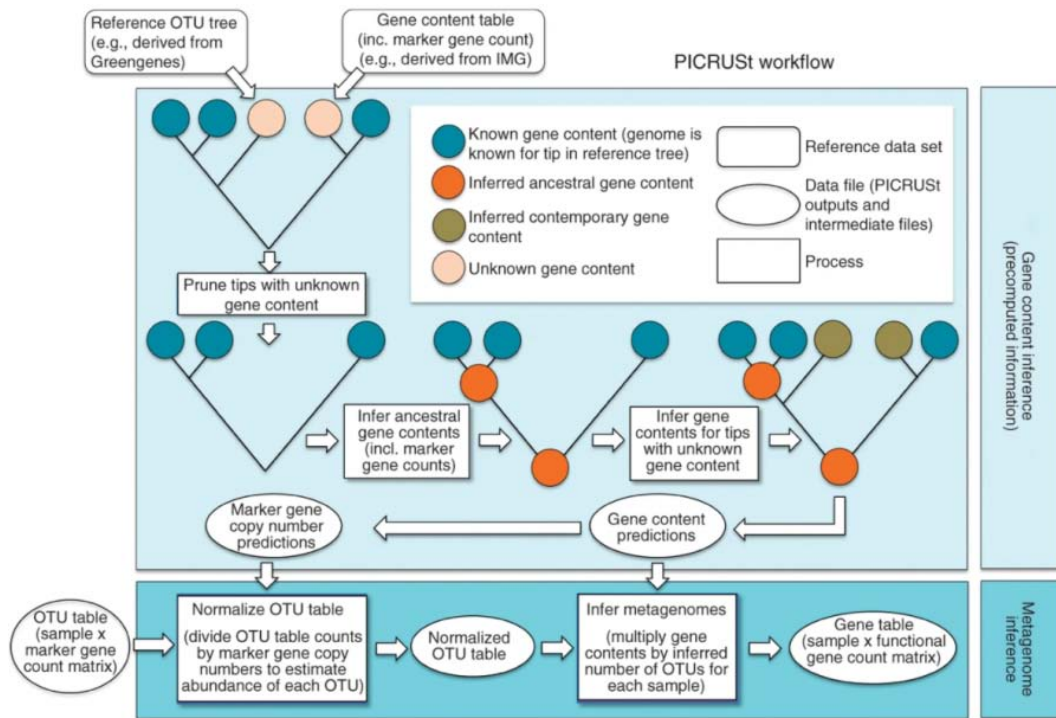
IMG微生物基因组数据库介绍: <https://www.omicsclass.com/article/1328>

■ PICRUSt全称为:

- Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

■ 分析原理:

- **构建“物种-基因”关系网:** 已有参考基因组细菌或古细菌中每个基因家族的基因数量 (通过IMG数据库获得) ;
- **预测未知细菌的功能:** Greengenes数据库含有物种进化树。通过未知物种的序列信息寻找其在进化树中的亲缘物种, 从而根据亲缘物种的基因信息预测未知物种的基因信息。
- **进行基因功能注释:** PICRUSt完成宏基因组的预测之后, 可以利用软件结合KEGG, COG和Pfam三大数据库进行注释, 从而赋予基因信息生物学意义。
- 核糖体rRNA基因拷贝数整理及标准化。



PICRUSt 分析原理详解: <https://www.omicsclass.com/article/1324>

(Langille *et al.*, 2013, Nature Biotechnology)

更多生物信息分析课程

- 1. 文章越来越难发？是你没发现新思路，基因家族分析发2-4分文章简单快速，学习链接：[基因家族分析实操课程](#)、[基因家族文献思路解读](#)
- 2. 转录组数据理解不深入？图表看不懂？点击链接学习深入解读数据结果文件，学习链接：[转录组（有参）结果解读](#)；[转录组（无参）结果解读](#)
- 3. 转录组数据深入挖掘技能-WGCNA，提升你的文章档次，学习链接：[WGCNA-加权基因共表达网络分析](#)
- 4. 转录组数据怎么挖掘？学习链接：[转录组标准分析后的数据挖掘](#)、[转录组文献解读](#)、[二代测序转录组数据自主分析](#)
- 5. [微生物16S/ITS/18S分析原理及结果解读](#)、[OTU网络图绘制](#)、[cytoscape与网络图绘制课程](#)
- 6. 生物信息入门到精通必修基础课：[linux系统使用](#)、[docker搭建生物信息分析环境](#)、[实验室linux生信分析平台搭建](#)、[linux命令处理生物大数据](#)、[perl入门到精通](#)、[perl语言高级](#)、[R语言画图](#)、[R语言快速入门与提高](#)、[python语言入门到精通](#)
- 7. 医学相关数据挖掘课程，不用做实验也能发文章：[TCGA-差异基因分析](#)、[GEO芯片数据挖掘](#)、[GEO芯片数据不同平台标准化](#)、[GSEA富集分析课程](#)、[TCGA临床数据生存分析](#)、[TCGA-转录因子分析](#)、[TCGA-ceRNA调控网络分析](#)
- 8. 其他: [NCBI数据上传](#)、[二代fastq测序数据解读](#)、
- 9. 测序数据分析：[基因组重测序数据分析](#)、[转录组数据分析](#)
- 10. 组学大讲堂全部生物生信数据挖掘课程可点击：[组学大讲堂视频课程](#)

课程不断更新中
更多组学大讲堂课程可扫描二维码直达

