Amplicon项目结题报告

华大基因有限公司

2018年8月31日

目录

一 综述

1 总体工作流程概述

DNA样品被接收后，对样品进行检测；检测合格的样品构建文库：回收目的Amplicon片段，用T4 DNA Polymerase、Klenow DNA Polymerase和T4 PNK将打断形成的粘性末端修复成平末端，再通过3' 端加碱基“A”，使得DNA片段能与3' 端带有“T”碱基的特殊接头连接；或者设计合成含有测序接头的双Index融合引物，以基因组DNA为模板，进行融合引物PCR，磁珠筛选目的Amplicon片段，最后，用合格的文库进行cluster制备和测序。用下机得到的数据进行相应的生物信息分析。

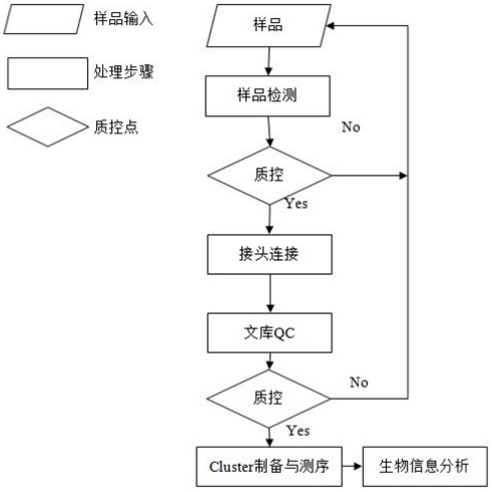


图1 总工作流程图

2 信息分析流程

下机数据经过数据过滤，滤除低质量的reads，剩余高质量的Clean data方可用于后期分析；通过reads之间的Overlap关系将reads拼接成Tags；在给定的相似度下将Tags聚成OTU，然后通过OTU与数据库比对，对OTU进行物种注释；基于OTU和物种注释结果进行样品物种复杂度分析以及组间物种差异分析。

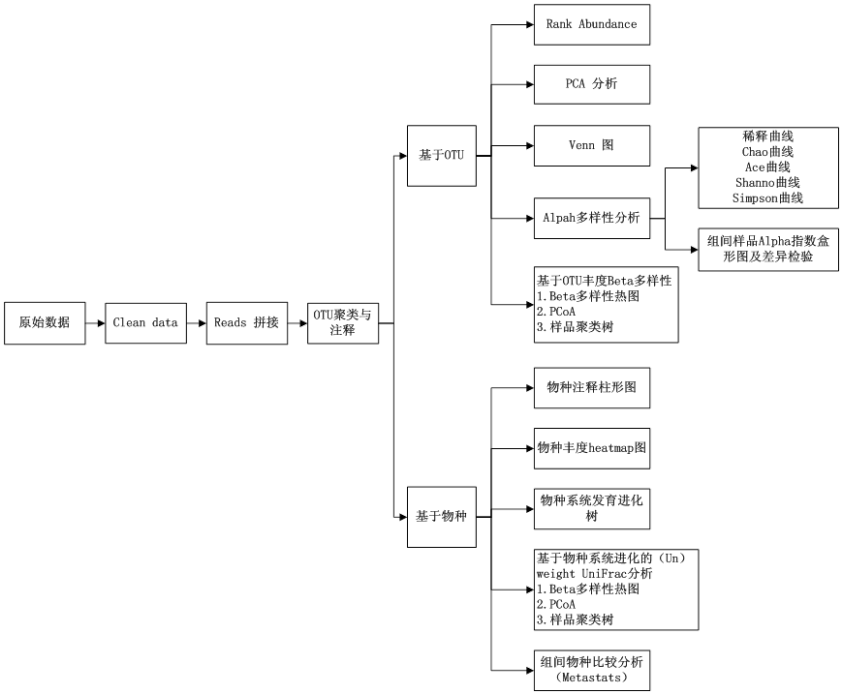


图2. Amplicon测序信息分析流程

1 数据统计

在进行数据处理过滤时[1]，使用内部撰写的程序对原始的测序数据进行如下处理，获得Clean Data，具体步骤如下：
1）采取按窗口去低质量的方法，具体操作如下：设置30 bp为窗口长度，如果窗口平均质量值低于20，从窗口开始截除read末端序列，移除最终read长度低于原始read长度75%的reads;
2）去除接头污染reads（默认adapter序列与read序列有15 bp的overlap，设置为15 bp，允许错配数为3）；
3）去除含N的reads；
4）去除低复杂度reads（默认reads中某个碱基连续出现的长度≥10，设置10 bp）。
如果样品通过barcode合并建库，得到Clean Data后，利用barcode序列通过内部撰写的程序对样品进行拆分。barcode序列与测序reads比对允许的错配个数为0 bp。

使用该方法，通过Illumina平台（Hiseq或者Miseq）进行Paired-end测序，下机数据经过去除低质量reads，每个样品数据产出详细统计结果见表1-1。

表1-1 样品测序数据统计

Sample Name	Reads Length (bp)	Raw Data (Mbp)	Adapter (%)	N base (%)	Ploy base (%)	Low Quality (%)	Clean Data (Mbp)	Data Utilization Ratio (%)	Raw Reads	Clean Reads	Read Utilization Ratio (%)
611A	250:250	33.65	0.000	0.076	0.003	5.285	30.32	90.11	67293*2	61194*2	90.94
612A	250:250	34.01	0.000	0.075	0.003	4.963	30.65	90.12	68023*2	61948*2	91.07
621A	250:250	32.29	0.000	0.094	0.001	3.215	30.18	93.47	64578*2	60787*2	94.13
622A	250:250	32.64	0.000	0.087	0.005	3.359	30.35	92.99	65278*2	61233*2	93.80
631A	250:250	33.19	0.000	0.065	0.001	4.002	30.57	92.10	66388*2	61636*2	92.84
632A	250:250	32.79	0.000	0.081	0.004	4.020	30.25	92.25	65584*2	60997*2	93.01
641A	250:250	32.00	0.000	0.095	0.002	2.823	30.26	94.58	63990*2	60834*2	95.07
642A	250:250	32.44	0.000	0.082	0.001	3.165	30.42	93.78	64878*2	61228*2	94.37
711A	250:250	33.31	0.000	0.089	0.002	4.643	30.17	90.58	66619*2	60945*2	91.48
712A	250:250	33.36	0.000	0.094	0.000	3.899	30.70	92.02	66727*2	61833*2	92.67
721A	250:250	33.85	0.000	0.081	0.002	4.394	30.78	90.92	67707*2	62069*2	91.67
722A	250:250	32.78	0.000	0.068	0.002	3.812	30.26	92.32	65557*2	60940*2	92.96
731A	250:250	31.86	0.000	0.094	0.001	2.126	30.57	95.96	63717*2	61378*2	96.33
732A	250:250	32.44	0.000	0.093	0.000	2.990	30.45	93.87	64877*2	61360*2	94.58
741A	250:250	32.78	0.000	0.077	0.001	3.066	30.80	93.95	65564*2	61994*2	94.55
742A	250:250	31.69	0.000	0.087	0.002	1.938	30.46	96.12	63380*2	61204*2	96.57

结果目录：BGI_results/Clean_Data/

2 序列拼接

序列拼接使用软件**FLASH**[2]（Fast Length Adjustment of Short reads，v1.2.11），利用重叠关系将双末端测序得到的成对reads组装成一条序列，得到高变区的Tags。拼接条件如下：
1）最小匹配长度15 bp；
2）重叠区域允许错配率为0.1。
去除没有overlap关系的reads。

Paired End Reads通过reads之间的overlap关系拼接成Tags，所有样品一共得到979025条Tags，平均每个样品61189条，SD值为421；Tag平均长度为252 bp，SD值为0 bp。每个样品Tags统计结果见表2-1。

表2-1 Reads拼接成Tags统计结果

Sample Name	Total Pairs Read Number	Connect Tag Number	Connect Ratio (%)	Average Length And SD	Tags Without Primer	Tag Utilization Ratio (%)	Average Length (bp) And SD
611A	61194	60978	99.65	252/0	-	-	-/-
612A	61948	61770	99.71	252/0	-	-	-/-
621A	60787	60636	99.75	252/0	-	-	-/-
622A	61233	61108	99.80	252/0	-	-	-/-
631A	61636	61466	99.72	252/0	-	-	-/-
632A	60997	60845	99.75	252/0	-	-	-/-
641A	60834	60704	99.79	252/0	-	-	-/-
642A	61228	61098	99.79	252/0	-	-	-/-
711A	60945	60774	99.72	252/0	-	-	-/-
712A	61833	61681	99.75	252/0	-	-	-/-
721A	62069	61928	99.77	252/0	-	-	-/-
722A	60940	60791	99.76	252/0	-	-	-/-
731A	61378	61205	99.72	252/0	-	-	-/-
732A	61360	61159	99.67	252/0	-	-	-/-
741A	61994	61821	99.72	252/0	-	-	-/-
742A	61204	61061	99.77	252/0	-	-	-/-

注：如果去除primer部分统计结果为“-”，说明样品下机数据不带引物，Tag不进行引物去除。

结果目录：BGI_results/Connect_Tags/

3 物种分类和丰度分析

将经过上述处理的Clean Tags进行OTU聚类，然后通过对OTU注释完成OTU的物种分类。

利用软件USEARCH（v7.0.1090）[3]将拼接好的 Tags聚类为OTU。其主要过程如下：
1) 利用UPARSE在97%相似度下进行聚类，得到OTU的代表序列；
2) 利用UCHIME（v4.2.40）[4]将PCR扩增产生的嵌合体从OTU代表序列中去除；
16S和ITS采取和已有的嵌合体数据库进行比对的方法去除嵌合体。18S采取Denovo的方法去除嵌合体
16S嵌合体数据库：gold database（v20110519）
ITS嵌合体数据库：UNITE（v20140703），分为ITS全长，ITS1和ITS2，按测序区域进行选择
3) 使用usearch_global方法将所有Tags比对回OTU代表序列，得到每个样品在每个OTU的丰度统计表[10]。

得到OTU代表序列后，通过RDP classifer（v2.2）软件将OTU代表序列与数据库 Greengene_2013_5_99 比对,进行物种注释，置信度阈值设置为0.6。
比对数据库：
16S（包括细菌与古菌）：Greengene（默认）： V201305; RDP Release 11_5 , 20160930
18S 真菌：Silva（默认）：SILVA_V132 , 20180410
ITS 真菌：UNITE（默认）：Version 7.2 , 20171201

对注释结果进行如下过滤：
1.去除没有注释结果的OTU；
2.去除注释结果不属于分析项目中的物种。例如，样品为细菌16S，如果OTU注释上古菌则去除。
剩余的OTU方可用于后期分析。

3.1 OTU及其丰度分析

3.1.1 OTU统计

拼接的Tags经过优化后，所有样品取选取样品中最少的Tags条数，在97%相似度下将其聚类为用于物种分类的OTU(Operational Taxonomic Units)，统计每个样品在每个OTU中的丰度信息，OTU的丰度初步说明了样品的物种丰富程度。16个样品共产生721个OTU，每个样品OTU统计结果见表3-1。

表3-1 样品OTU统计

Sample Name	Tag number	OTU number
611A	50909	409
612A	51137	424
621A	48911	439
622A	49230	474
631A	49455	446
632A	49138	456
641A	50208	427
642A	50733	396
711A	51404	393
712A	50357	398
721A	50267	427
722A	47904	456
731A	49744	491
732A	48934	465
741A	49158	497
742A	50922	439

注：Tags number：样品中能和OTU代表序列比对上，并且OTU具有注释结果的Tags的总数。

3.1.2 OTU Venn图分析

在97%的相似度下，得到了每个样品的OTU个数，利用Venn图可以展示多样品共有和各自特有OTU数目，直观展示样品间OTU的重叠情况。结合OTU所代表的物种，可以找出不同环境中的核心微生物。

根据每个样品OTU在每个样品的丰度文件，计算每个样品或组别具有的OTU（不考虑OTU丰度，只考虑OTU有无），通过R（v3.1.1）语言中的VennDiagram包做出Venn图，并给出样品间或组间共有与特有OTU ID文件。

图3-1为OTU venn图结果，图中不同颜色图形代表不同样品或者不同组别，不同颜色图形之间交叠部分数字为两个样品或两个组别之间共有的OTU个数。同理，多个颜色图形之间交叠部分数字为多个样品或组别之间共有OTU个数。Venn图容许2-5个样品或组别。

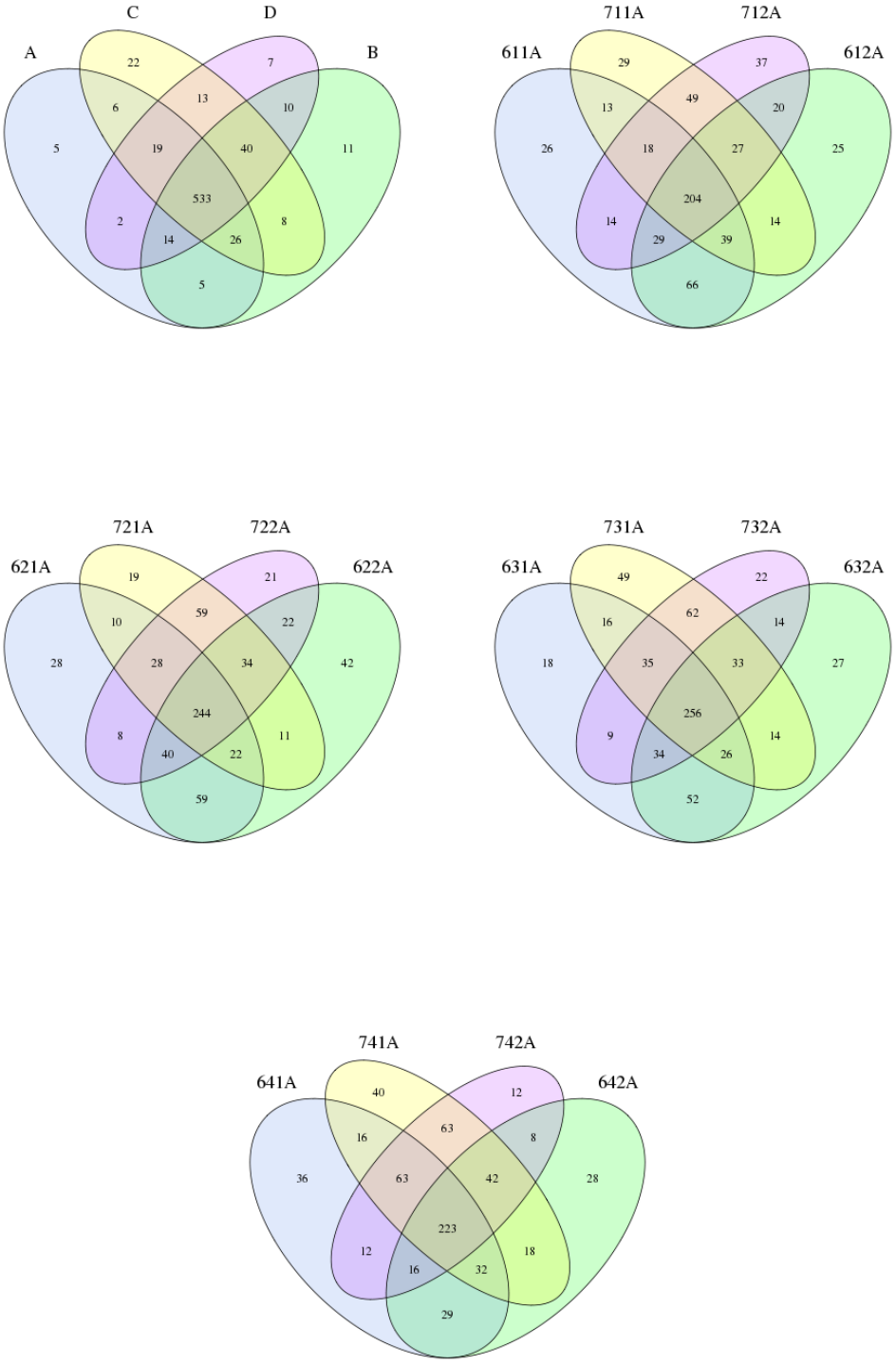


图3-1 OTU venn 分析

3.1.3 Core-Pan OTU 分析

Core-Pan OTU分析，统计样品间或者组间共有和特有OTU数目。

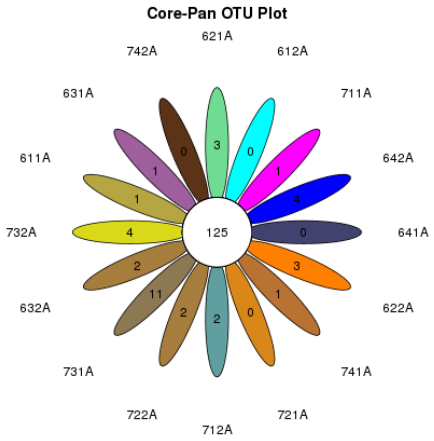


图3-2-1 共有和特有OTU

中间圆圈表示样品或组共有OTU个数，中间圆圈外边的椭圆表示这个样品或组有而其它样品或组都没有的OTU个数。

3.1.4 OTU PCA分析

PCA 分析(Principal Component Analysis)，即主成分分析，是一种分析和简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。通过分析不同样品OTU（97%相似性）组成可以反映样品的差异和距离，PCA运用方差分解，将多组数据的差异反映在二维坐标图上，坐标轴取能够最大反映方差值两个特征值。如果两个样品距离越近，则表示这两个样品的组成越相似。不同处理或不同环境间的样品可能表现出分散和聚集的分布情况，从而可以判断相同条件的样品组成是否具有相似性。

根据每个样品OTU在每个样品的丰度文件计算出每个OTU在每个样品的相对丰度，利用这个丰度信息进行OTU的PCA分析。通过R（v3.1.1）语言中ade4包进行统计与作图。

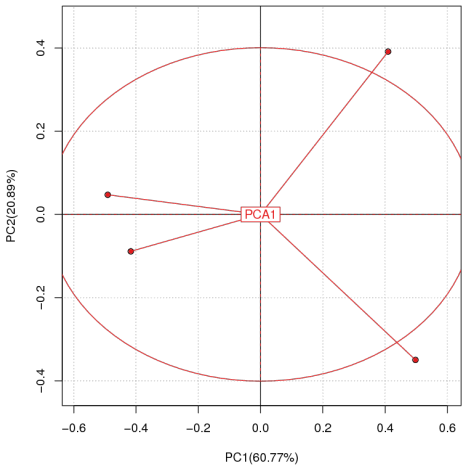


图3-3-1 基于OTU丰度的PCA分析(按Description分组)

横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值。图中点分别表示各个样品。不同颜色代表样品属于不同的分组。

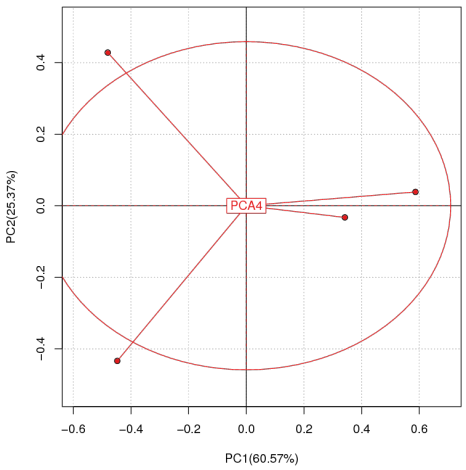


图3-3-2 基于OTU丰度的PCA分析(按Description分组)

横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值。图中点分别表示各个样品。不同颜色代表样品属于不同的分组。

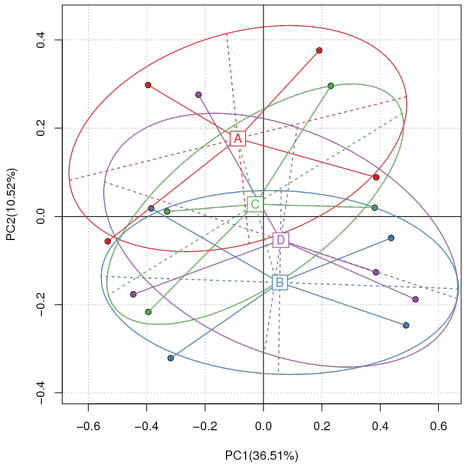


图3-3-3 基于OTU丰度的PCA分析(按Description分组)

横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值。图中点分别表示各个样品。不同颜色代表样品属于不同的分组。

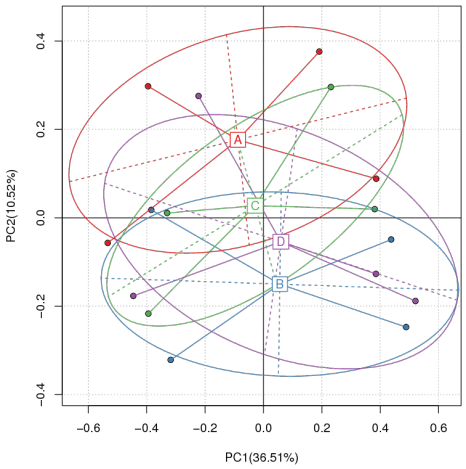


图3-3-4 基于OTU丰度的PCA分析(按Description分组)

横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值。图中点分别表示各个样品。不同颜色代表样品属于不同的分组。

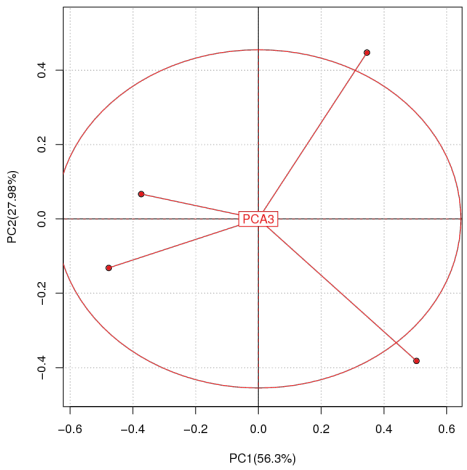


图3-3-5 基于OTU丰度的PCA分析(按Description分组)

横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值。图中点分别表示各个样品。不同颜色代表样品属于不同的分组。

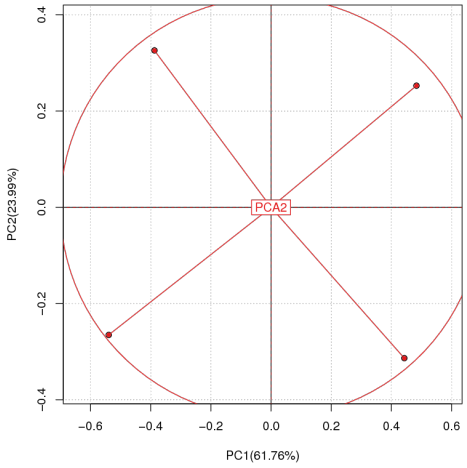


图3-3-6 基于OTU丰度的PCA分析(按Description分组)

横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值。图中点分别表示各个样品。不同颜色代表样品属于不同的分组。

3.1.5 物种累积曲线分析

物种累积曲线(species accumulation curves) 用于描述随着抽样量的加大物种增加的状况，是理解调查样地物种组成和预测物种丰富度的有效工具，在生物多样性和群落调查中，被广泛用于抽样量充分性的判断以及物种丰富度(species richness) 的估计。(样本数建议大于30个)

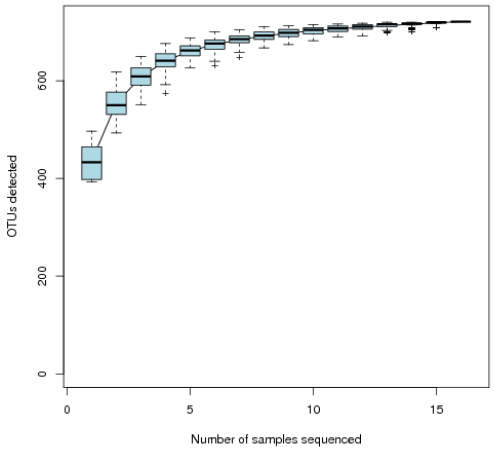


图3-4-1 物种累计曲线图

物种累计曲线图（Species Accumulation Analysis）,横坐标代表样品数目，纵坐标代表OTU数目（检测到的物种数），

3.1.6 OTU PLS-DA分析

PLS-DA 分析（Partial least squares discrimination analysis，PLS-DA）是一种用于判别分析的多变量统计分析方法，常用于来判断研究对象如何分类。与PCA相比，PLS-DA方法可将主成分分析和多元回归的功能相结合，达到循环 PCA的各个成分以使其之间的间隔达到最大化，从而加大各组观察结果之间的间隔的目的。 通过R软件 mixOmics 包作图。

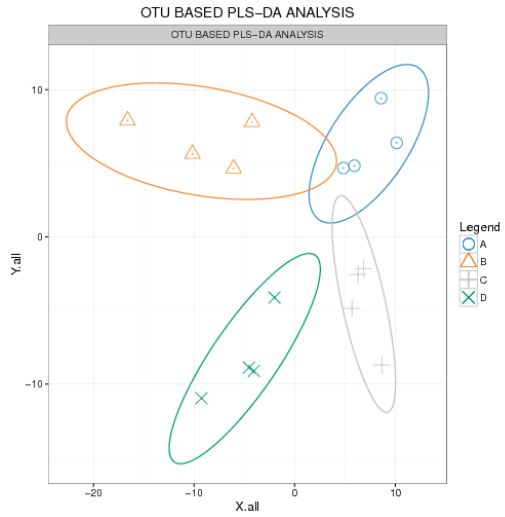


图3-5-1 基于OTU丰度的PLS-DA分析(按Description分组)

横纵坐标轴分别为贡献度最高的两个成分。图中每一个点代表一个样本，相同颜色的点来自同一个分组，两点之间距离越近表明两者的差异越小。

3.1.7 OTU Rank曲线

OTU Rank曲线是展现样品中物种多样性的一种形式，可以同时解释样品多样性的两个方面，即样品所含物种的丰富程度和均匀程度。样品中物种的丰富程度由曲线的横轴长度来反映，曲线越宽，说明样品中物种组成越丰富。样品中物种的均匀度由曲线纵轴的形状来反映，曲线越平坦，说明样品中物种组成的均匀度越高。

计算每个OTU在每个样品中的相对丰度，然后按丰度从大到小进行排列，以OTU的等级作为横坐标，以OTU的相对丰度作为纵坐标进行作图。使用R（v3.1.1）软件作图。

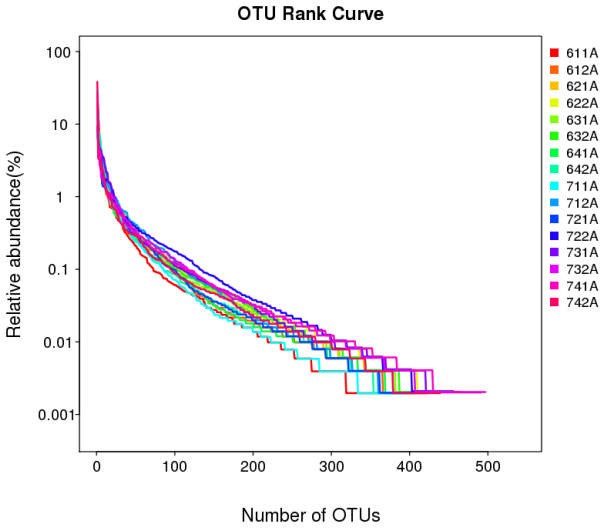


图3-6 OTU Rank 曲线图

横坐标为样品OTU丰度排位（由高至低），纵坐标为OTU丰度。

3.2 物种及其丰度分析

3.2.1 物种注释分析

通过与数据库进行比对，对OTU进行物种分类并分别在门、纲、目、科、属、种几个分类等级对各个样品作物种profiling面积图和柱状图。**图3-4-***展示了各样品在不同分类等级上的物种profiling。从图中可以直观看出不同物种在每个样品中所占的比例。门水平画所有物种的柱状图。从纲水平开始，将物种丰度在所有样品均低于0.5%的物种全部合并成Others。

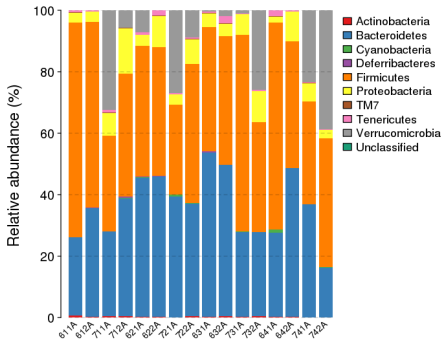


图3-4-1 样品Phylum分类水平中物种profiling柱状图

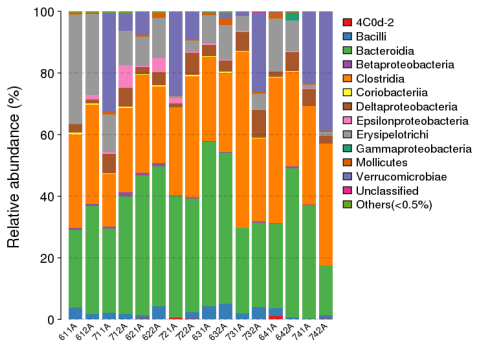


图3-4-2 样品Class分类水平中物种profiling柱状图

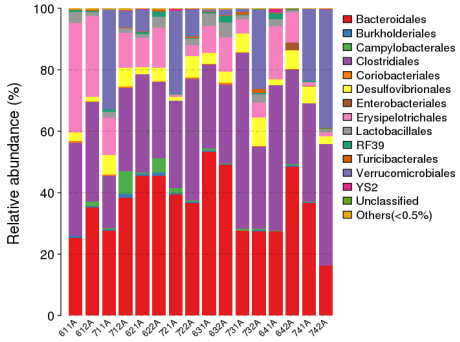


图3-4-3 样品Order分类水平中物种profiling柱状图

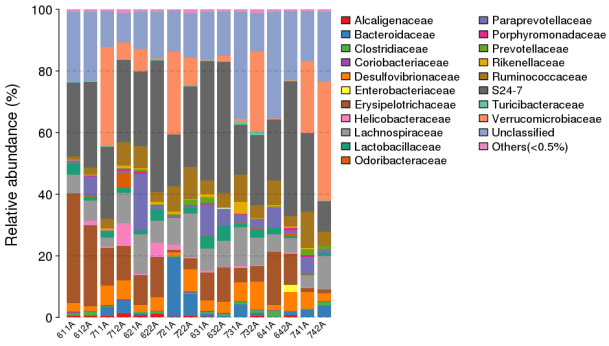


图3-4-4 样品Family分类水平中物种profiling柱状图

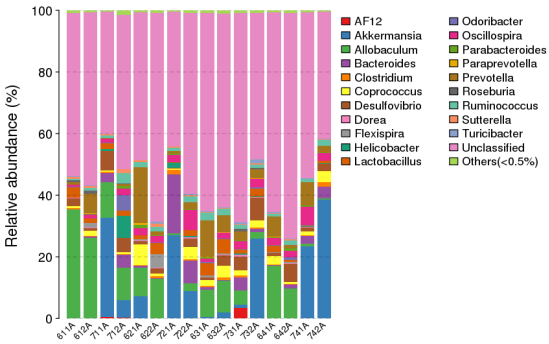


图3-4-5 样品Genus分类水平中物种profiling柱状图

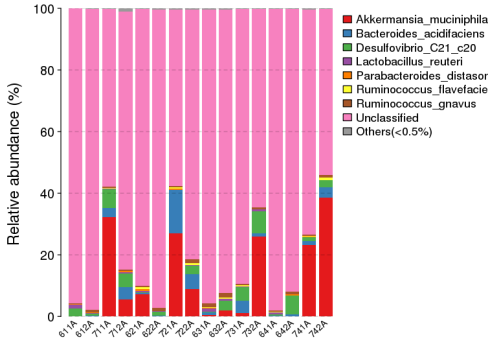


图3-4-6 样品Species分类水平中物种profiling柱状图

3.2.2 物种热图分析

Heatmap是以颜色梯度来代表数据矩阵中数值的大小并能根据物种或样品丰度相似性进行聚类的一种图形展示方式。聚类结果加上样品的处理或取样环境分组信息，可以直观观察到相同处理或相似

环境样品的聚类情况，并直接反映了样品的群落组成的相似性和差异性。本分析内容分别在门，纲，目，科，属，种分类等级进行heatmap聚类分析。纵向聚类表示所有物种在不同样品间表达的相似情况，距离越近，枝长越短，说明样品的物种组成及丰度越相似。横向聚类表示该物种在各样品丰度相似情况，与纵向聚类一样，距离越近，枝长越短，说明两物种在各样品间的组成越相似。门水平画所有物种的柱状图。从纲水平开始，将物种丰度在所有样品均低于0.5%的物种全部合并成Others。**图3-5-***展示了不同分类水平物种聚类热图。

根据每个物种在每个样品的相对丰度进行物种热图分析。由于物种的相对丰度差异可能较大，影响样品聚类。故对相对丰度进行了以10为底的log转化。如果样品中物种相对丰度为0，则取所有样品中物种丰度最小（非0）的值二分之一的log值代替。通过R（v3.1.1）语言中的gplots包进行作图，距离算法为euclidean，聚类方法为complete。

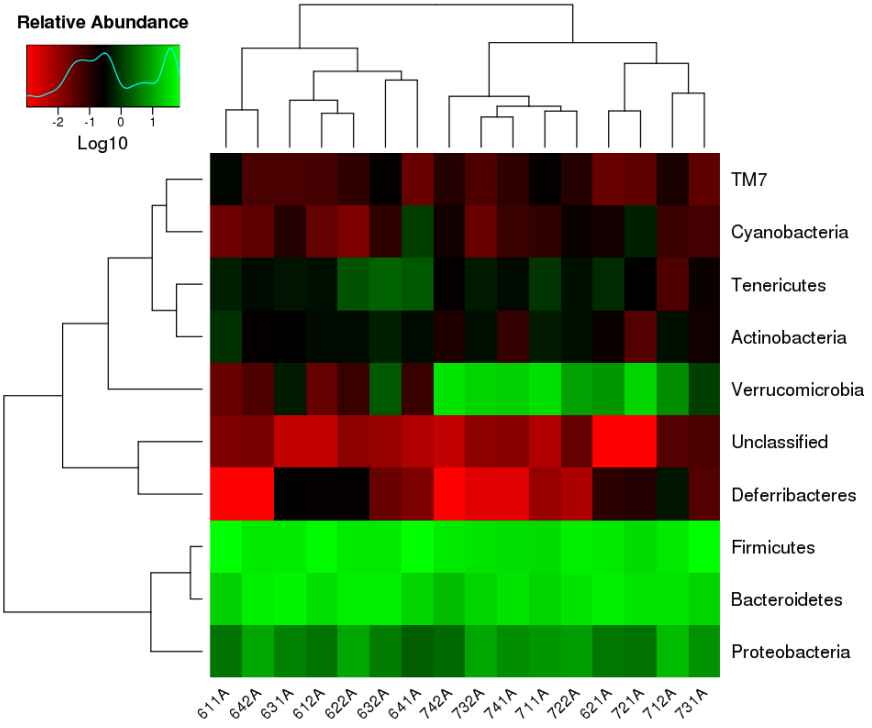


图3-5-1 Phylum水平物种丰度热图

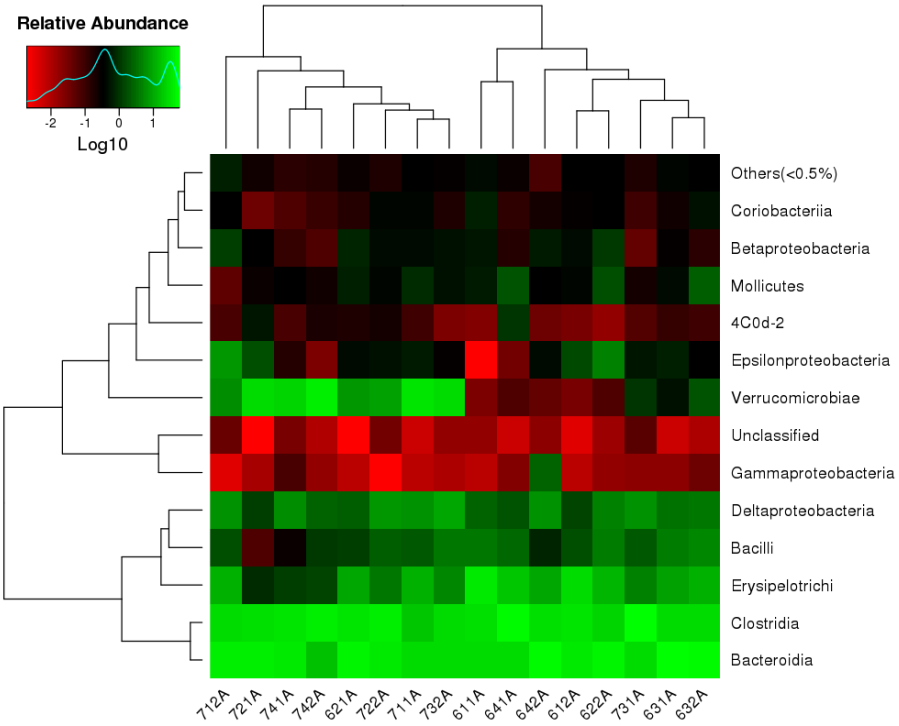


图3-5-2 Class水平物种丰度热图

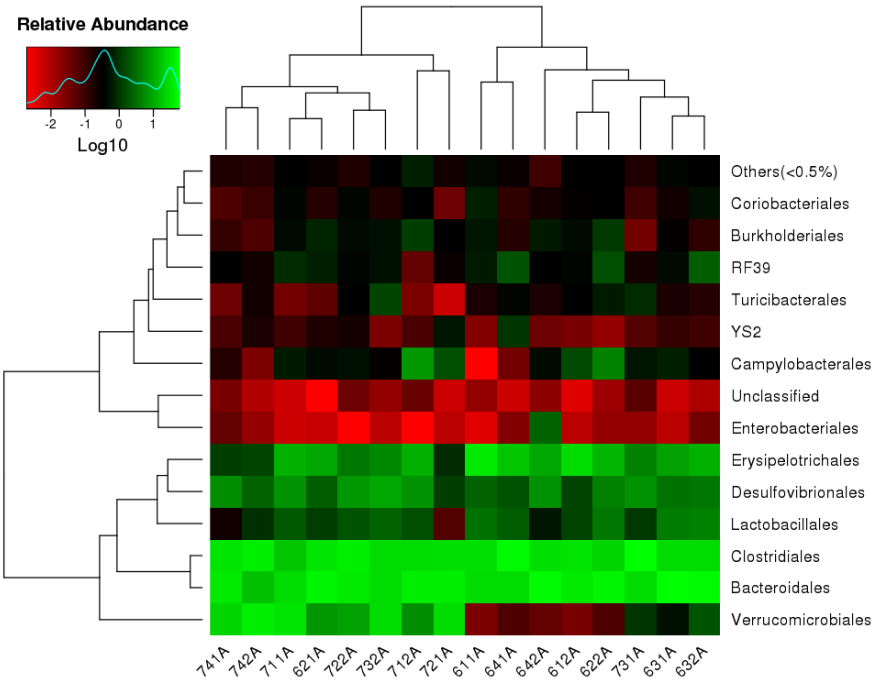


图3-5-3 Order水平物种丰度热图

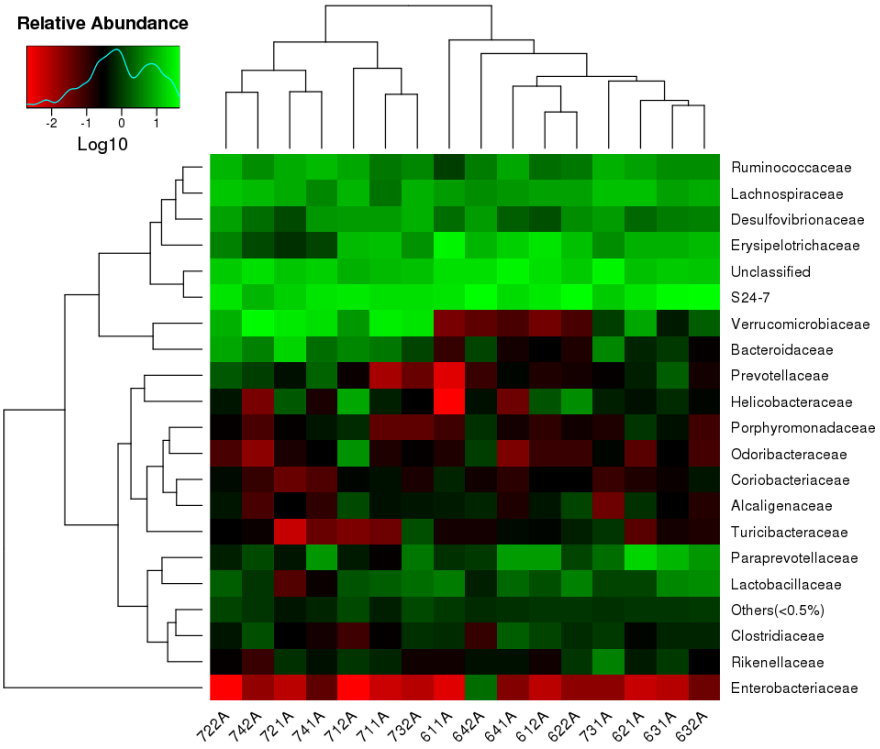


图3-5-4 Family水平物种丰度热图

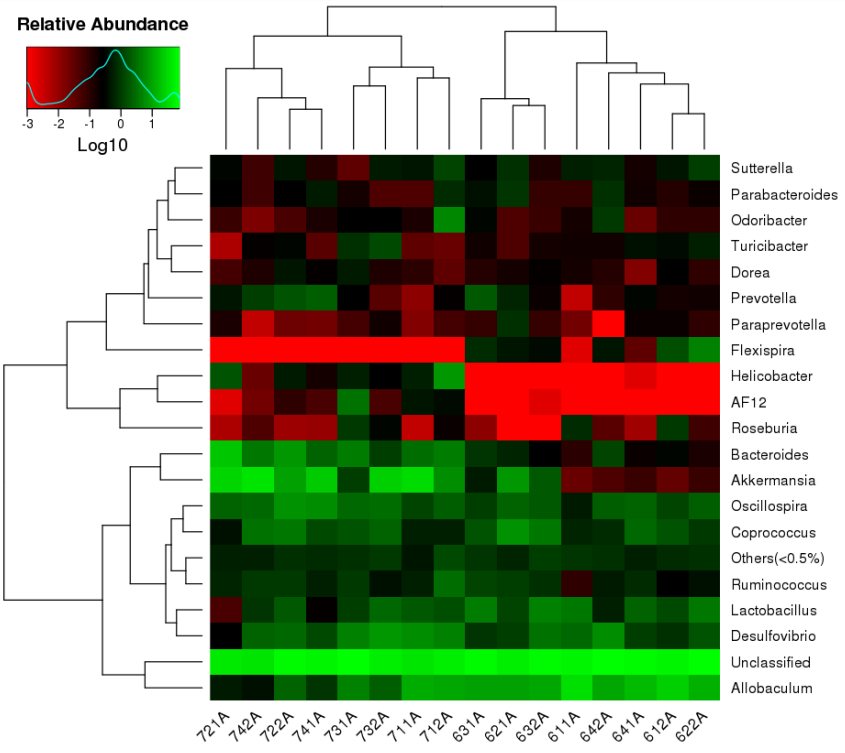


图3-5-5 Genus水平物种丰度热图

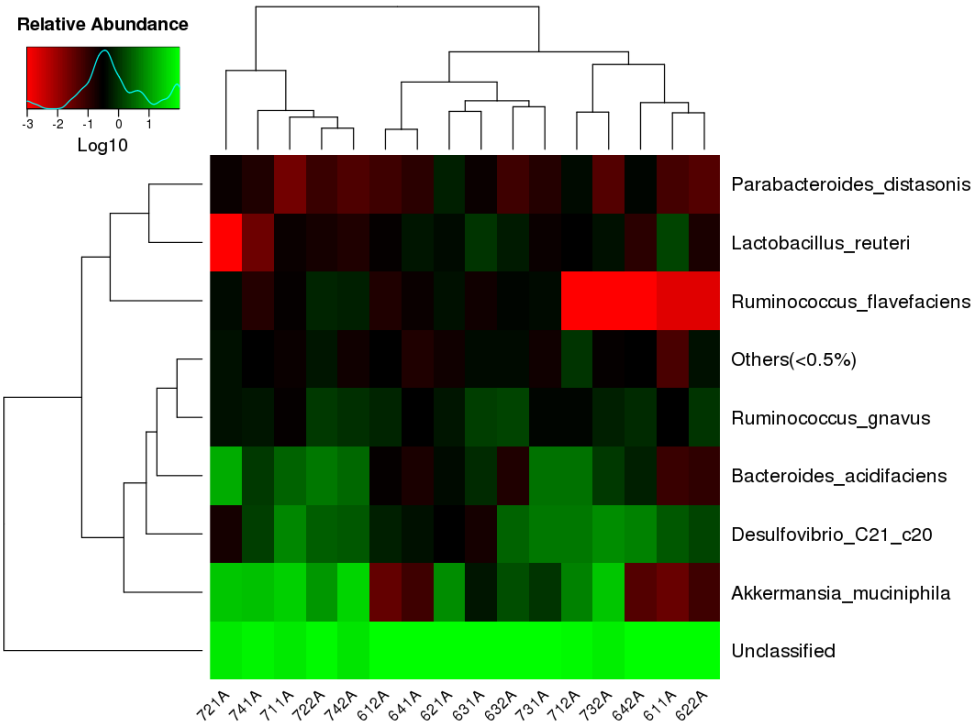


图3-5-6 Species水平物种丰度热图

3.2.3 物种系统进化分析

系统进化树是表示物种间发育关系的树状图，枝长的长短表示进化距离的差异。系统关系越近的物种，在进化树种距离越近。前面已经对样品中的物种以及丰度进行了分析。通过构建物种的进化树，可以更深一步了解样品中物种的进化关系。本分析在属的水平上，选取一条丰度最高的一条作为代表序列，构建物种系统进化树。

利用QIIME[5]软件中的align_seqs.py程序将OTU代表序列进行比对（16S与18S通过PyNAST算法与数据库为Silva_108_core_aligned_seqs进行比对，ITS通过MUSCLE软件进行比对），得到比对

好的OTU序列并利用make_phylogeny.py程序生成OTU的进化树，用于Beta多样性分析。 通过OTU的丰度文件，从比对OTU比对文件中挑选出每个属丰度最高的OTU的序列作为该属的代表序列，通过QIIME（v1.80）软件中的make_phylogeny.py，方法为“fasttree”构建系统进化树。最后通过R软件（v3.1.1）将系统进化树图形化。

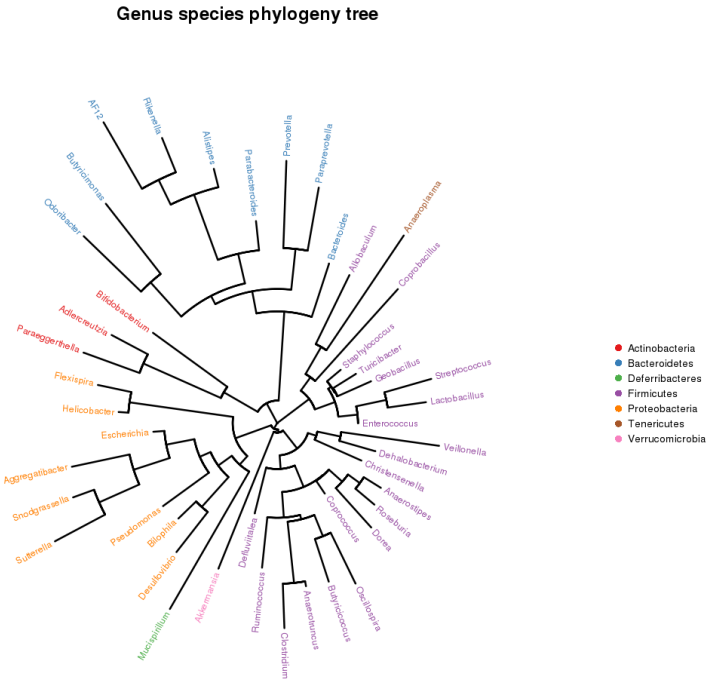


图3-6 物种系统进化树(相同颜色属名代表相同的门)

结果目录：BGI_results/OTU_Cluster_Taxonomy/

4 样品多样性分析

4.1 单个样品多样性分析

Alpha多样性（Alpha diversity）是对单个样品中物种多样性的分析[11]，包括observed species指数、chao指数、ace指数，shannon指数以及simpson指数等。前面4个指数越大，最后一个指数越小，说明样品中的物种越丰富。

其中，observed species指数、chao指数和ACE指数反映样品中群落的丰富度（species richness），即简单指群落中物种的数量，而不考虑群落中每个物种的丰度情况。这个3个指数对应的稀释曲线还可以反映样品测序量是否足够。如果曲线趋于平缓或者达到平台期时也就可以认为测序深度已经基本覆盖到样品中所有的物种；反之，则表示样品中物种多样性较高，还存在较多未被测序检测到的物种。

而shannon指数以及simpson指数反映群落的多样性（species diversity），受样品群落中物种丰富度（species richness）和物种均匀度（species evenness）的影响。相同物种丰富度的情况下，群落中各物种具有越大的均匀度，则认为群落具有越大的多样性。

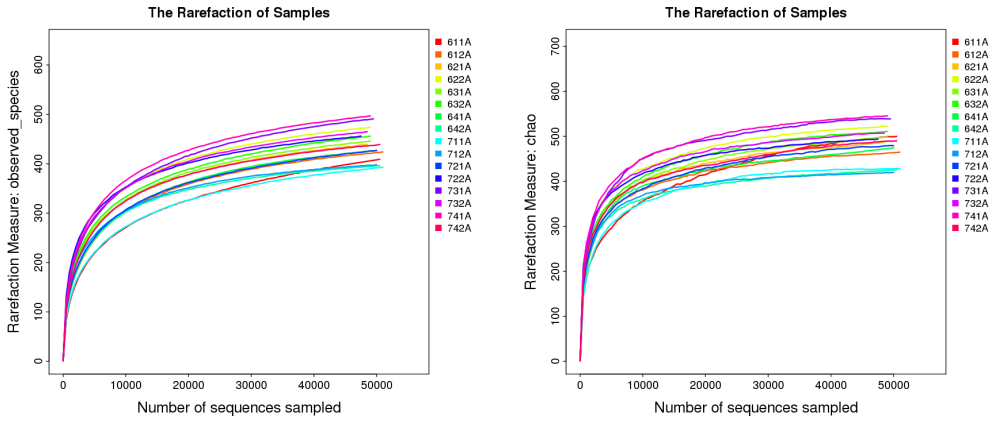
稀释曲线是利用已测得序列中已知的各种OTU的相对比例，来计算抽取n个（n小于测得Reads序列总数）Tags时各Alpha指数的期望值，然后根据一组n值（一般为一组小于总序列数的等差数列，本项目公差为500）与其相对应的Alpha指数的期望值绘制曲线。本分析通过mothur（v1.31.2）软件计算样品的Alpha多样性值并用R（v3.1.1）软件做出相应的稀释曲线图。每个指数的计算公式请参考：http://www.mothur.org/wiki/Calculators 。

如样品有提供分组信息，且每组样品个数不小于3，将对组间的Alpha多样性指数进行差异分析的检验。差异分析的检验方法为秩和检验，如果组数为2，采用两样品比较的Wilcoxon Rank-Sum Test（R中的wilcox.test）；如果组数大于2，采用多样品比较的Kruskal-Wallis Test（R中的kruskal.test）。最后利用Alpha多样性指数绘制盒形图。差异分析与作图均通过R软件（v3.1.1）进行。

表4-1 样品Alpha多样性统计结果

Sample Name	sobs	chao	ace	shannon	simpson	coverage
611A	409.000000	500.000000	496.095952	3.305848	0.107189	0.998212
612A	424.000000	464.738095	461.014661	3.606692	0.076133	0.998846
621A	439.000000	487.157895	480.589014	3.989761	0.051323	0.998753
622A	474.000000	522.372093	516.858605	4.208696	0.036014	0.998680
631A	446.000000	499.333333	491.166736	3.955057	0.055367	0.998686
632A	456.000000	510.886364	508.797702	4.130718	0.035861	0.998575
641A	427.000000	474.527778	460.443019	3.731458	0.063911	0.998825
642A	396.000000	426.100000	419.430355	3.777243	0.058066	0.999152
711A	393.000000	428.400000	435.857603	3.398250	0.119816	0.998833
712A	398.000000	419.794118	418.058471	4.311119	0.025888	0.999226
721A	427.000000	479.317073	472.313024	3.614581	0.098802	0.998687
722A	456.000000	493.657895	492.467038	4.469225	0.026079	0.998873
731A	491.000000	538.788462	535.474223	4.262088	0.030033	0.998573
732A	465.000000	507.775000	502.229190	3.998258	0.077648	0.998794
741A	497.000000	545.468085	541.751610	3.982309	0.069966	0.998617
742A	439.000000	489.833333	476.645198	3.447772	0.155148	0.998802

图4-1展示了Alpha多样性不同指数的稀释曲线结果。



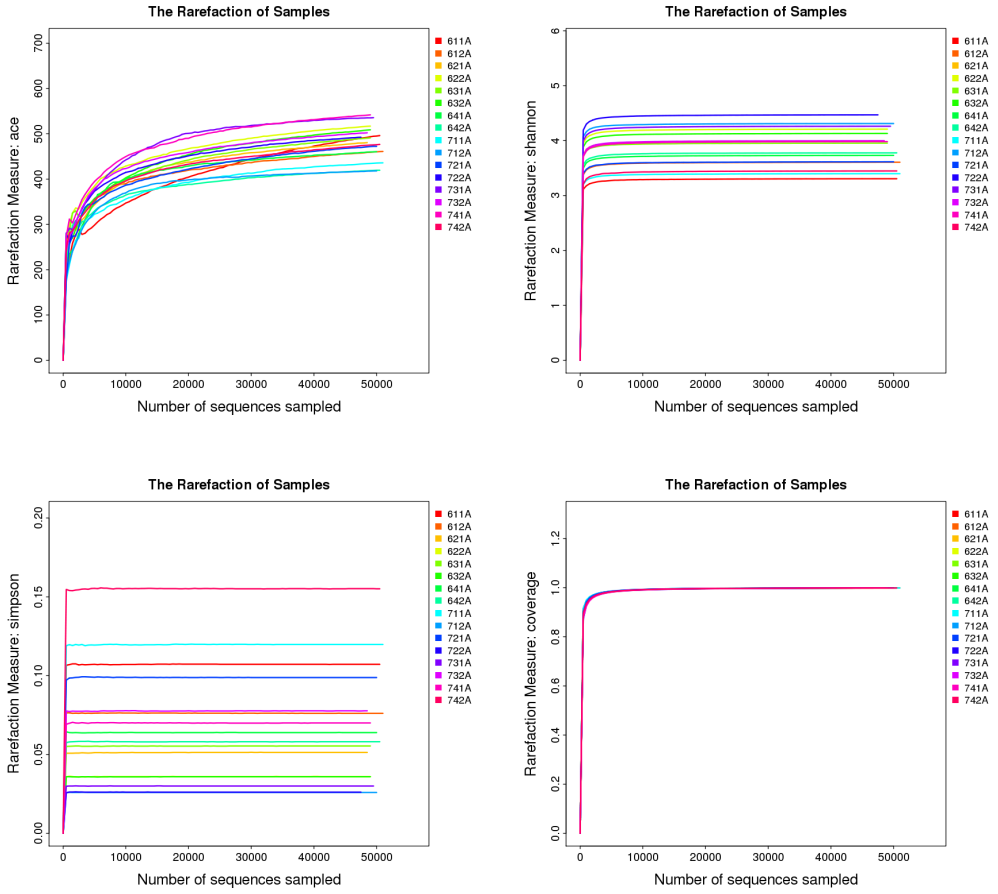


图4-1 表示样品物种多样性Alpha指数稀释曲线图

表4-2为Alpha多样性指数组间进行比较的结果，计算出Alpha多样性各组的均值与标准差。如果p值小于0.05，说明组间，至少有两组Alpha多样性存在显著差异，也就是组间物种多样性存在差异。

表4-2-1 样品Alpha多样性统计结果(按Description分组结果)

#Alpha	mean(A)	SD(A)	mean(B)	SD(B)	...	p-vaule
sobs	406.00000	13.73560	439.75000	42.24827	...	0.04547
chao	453.23305	36.76210	483.98230	49.17481	...	0.14175
ace	452.75667	33.84293	474.56755	50.85151	...	0.16493
shannon	3.65548	0.45485	3.73470	0.22020	...	0.17782
simpson	0.08226	0.04182	0.08677	0.04584	...	0.41605
coverage	0.99878	0.00042	0.99885	0.00022	...	0.35259

注：如果存在多组数据，只显示前面两组统计结果，详细结果请看附件。

图4-2为组间Alpha多样性盒形图，更直观显示组间Alpha多样性差异。盒形图可以显示5个统计量（最小值，第一个四分位数，中位数，第三个中位数和最大值，及由下到上的5条线），异常值以“o”标。

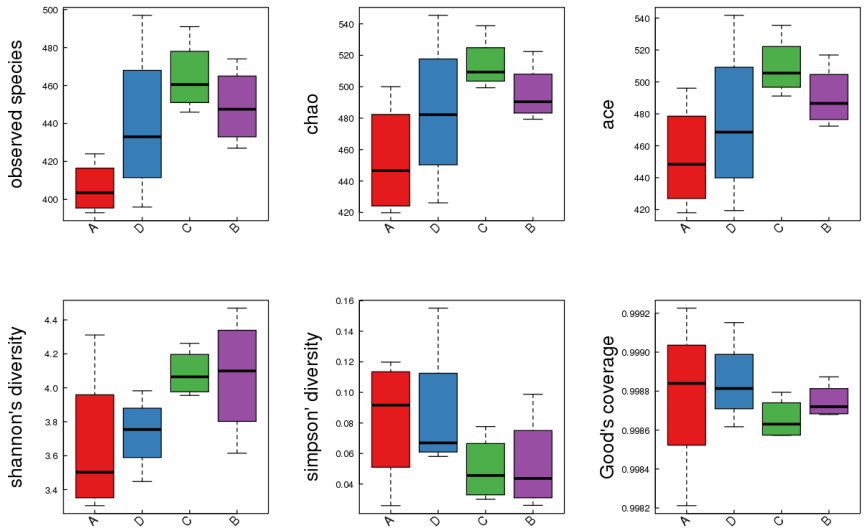


图4-2-1 组间Alpha多样性盒形图(按Description分组结果)

结果目录：BGI_results/Alpha_Diversity/

4.2 样品间多样性比较分析（n>=4）

与Alpha多样性分析不同，Beta多样性（Beta diversity）分析是用来比较一对样品在物种多样性方面存在的差异大小。分析各类群在样品中的含量，进而计算出不同样品间的Beta多样性值。Beta多样性通过QIIME（v1.80）进行。由于不同样品的测序深度不一样，需要对每个样品的序列数进行统一。每个样品按所有样品中序列数最少的样品的序列数随机抽取序列， 生成新的OTU table biom文件，并用该文件计算Beta多样性距离。

多种指数可以衡量Beta多样性，例如Bray-Curtis，weighted UniFrac，unweighted UniFrac，pearson等。常用的为Bray-Curtis，weighted UniFrac，unweighted UniFrac。

Bray-Curtis距离是反映两个群落之间差异性的常用指标。Bray-Curtis距离的计算不考虑序列间的进化距离，只考虑样品中物种存在情况。Bray-Curtis距离的值在0-1之间，值越大表示样品间的差异越大。

UniFrac是通过利用系统进化的信息来比较样品间的物种群落差异。其计算结果可以作为一种衡量beta diversity的指数，它考虑了序列间的进化距离，该指数越大表示样品间的差异越大。报告中给出的UniFrac结果分为加权UniFrac（weighted UniFrac）与非加权UniFirac（unweighted UniFrac）2种，其中weighted UniFrac考虑了序列的丰度，unweighted UniFrac不考虑序列丰度。

Beta多样性结果：
weighted_unifrac多样性结果：weighted_unifrac多样性结果
unweighted_unifrac多样性结果：unweighted_unifrac多样性结果
bray_curtis多样性结果：bray_curtis多样性结果

图4-3-*为Beta多样性矩阵heatmap，通过图形将Beta多样性数据进行可视化，并通过对样品进行聚类，具有相似beta多样性的样品聚类在一起，反映了样品间的相似性。

Beta多样性热图使用R（v3.1.1）软件中的NMF包的aheatmap进行作图。

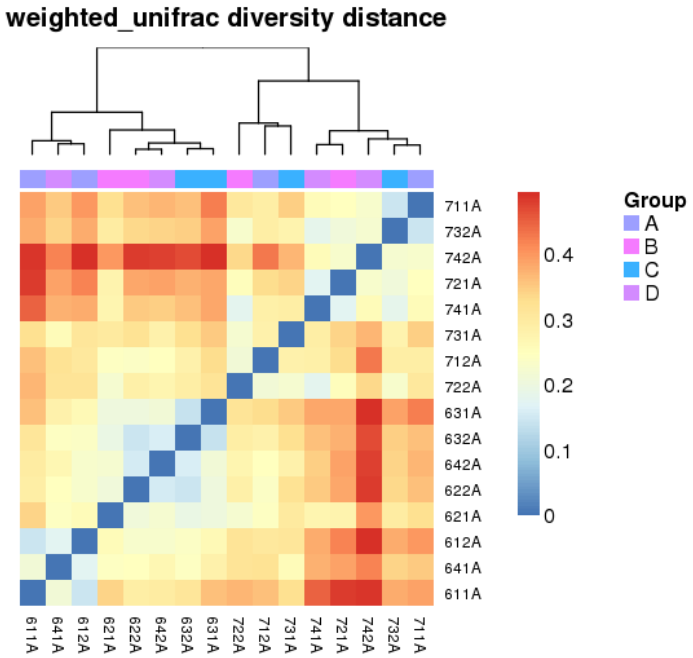


图4-3-1 Beta多样性heatmap(按Description分组，weighted_unifrac)

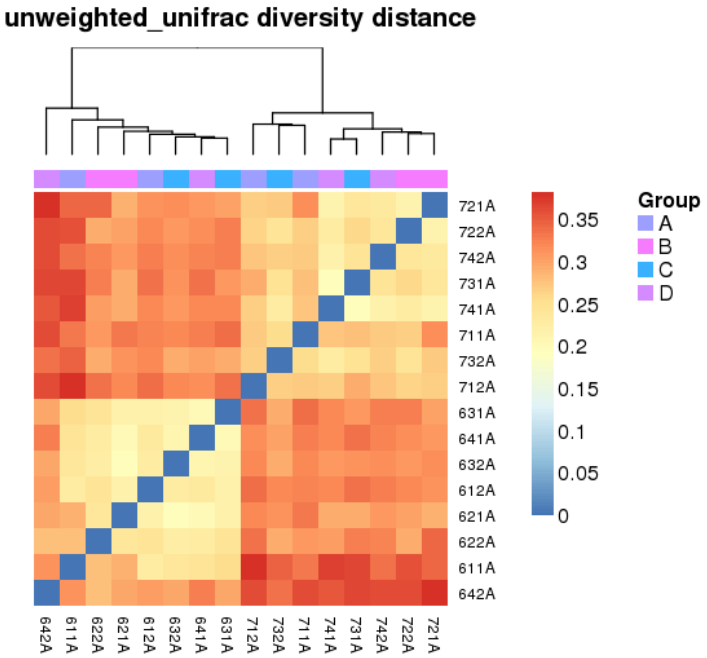


图4-3-2 Beta多样性heatmap(按Description分组，unweighted_unifrac)

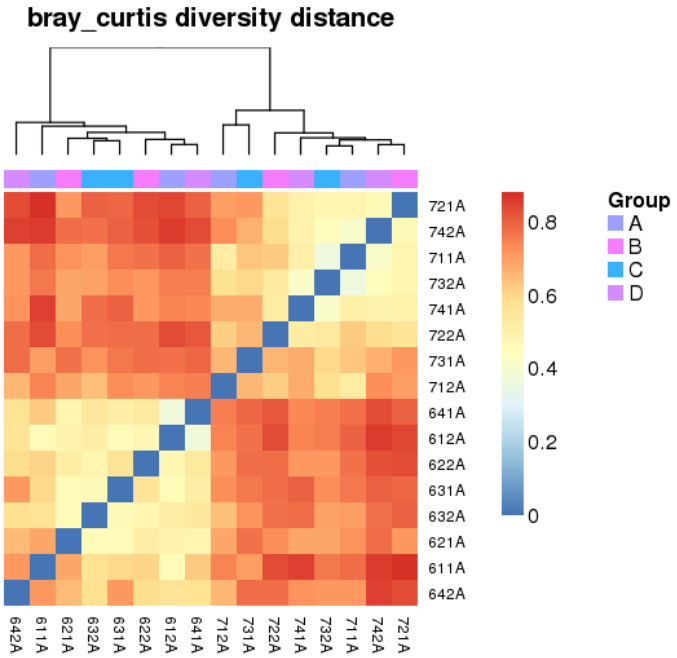


图4-3-3 Beta多样性heatmap(按Description分组，bray_curtis)

为了进一步展示样品间物种多样性差异，使用主坐标分析（Principal coordinates analysis，PCoA）的方法展示各个样品间的差异大小。如果两个样品距离较近，则表示这两个样品的物种组成较相似。结果为随机抽样100次计算的结果，如果样品的抽样重复性较好，样品颜色区域范围会比较小；反之，如果抽样重复性差，样品颜色部分区域范围较大。

PCoA分析通过QIIME（v1.80）[5]软件进行，采用迭代算法，分别在加权物种分类丰度信息和不加权物种分类丰度信息的情况下，使用所有样品中序列数最少的样品的序列数的75%进行抽样分析，迭代100次之后综合统计得到最终的统计分析结果表及PCoA展示图。

PCoA结果：
weighted_unifrac PCoA 2D结果：2D PCoA结果
weighted_unifrac PCoA 3D结果：3D PCoA结果
unweighted_unifrac PCoA 2D结果：2D PCoA结果
unweighted_unifrac PCoA 3D结果：3D PCoA结果
bray_curtis PCoA 2D结果：2D PCoA结果
bray_curtis PCoA 3D结果：3D PCoA结果

4.3 样品间物种组成聚类分析（n>=4）

根据各样品差异性的统计结果，对样品进行聚类分析并计算样品间距离，以判断各样品物种组成的相似性。所有样品的聚类分析结果如**图4-4-***所示，图中相同颜色的样品表示属于同一个分组。样品越靠近，枝长越短，说明两个样品的物种组成越相似。

聚类分析通过QIIME（v1.80）[5]软件进行，采用迭代算法，分别在加权物种分类丰度信息和不加权物种分类丰度信息的情况下，使用所有样品中序列数最少的样品的序列数的75%进行抽样分析，迭代100次之后综合统计得到最终的差异统计结果得到聚类树，以R（v3.1.1）画图。聚类方法为UPGMA（Unweighted Pair Group Method with Arithmetic mean）。

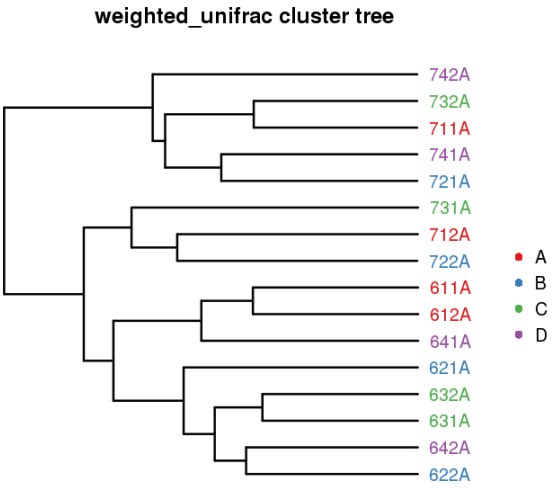


图4-4-1 样品聚类分析图(按Description分组，weighted_unifrac)

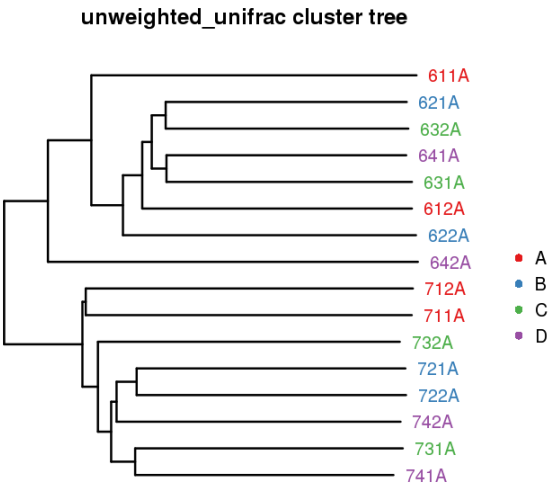


图4-4-2 样品聚类分析图(按Description分组，unweighted_unifrac)

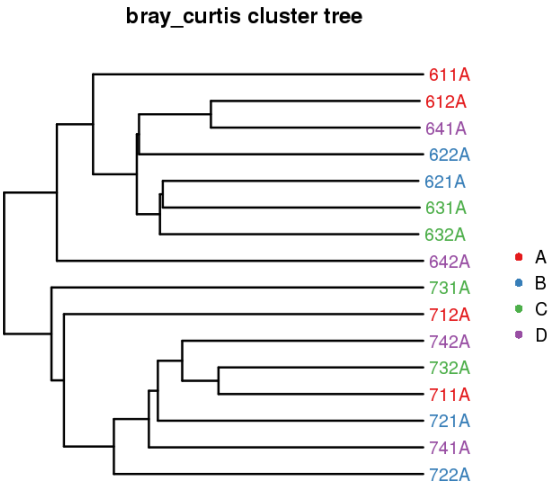


图4-4-3 样品聚类分析图(按Description分组，bray_curtis)

结果目录：BGI_results/Beta_Diversity/

5 样品组间显著性差异分析（组别>=2，每个组样品数>=3）

通过统计学的方法检验两组样品间微生物群落丰度的差异，并使用FDR(false discovery rate)评估差异的显著性。从检验结果中，可以筛选出导致两组样品组成差异的物种。本分析分别在门，纲，目，科，属，种分类等级进行组间显著性差异分析。

使用软件Metastats (http://metastats.cbcb.umd.edu/)(默认)或者R软件（秩和检验，Fisher's精确检验，卡方检验，t检验，方差检验）进行组间显著性差异分析。p值校正通过R（v3.1.1）包中的p.adjust进行，校正方法为"BH"（即Benjamini-CHochberg）[12]。

检验方法：kruskal.test

结果目录：BGI_results/Diff_Analysis/

6 LEFSE分析

LEfSe是一种用于发现高维生物标识和揭示基因组特征的软件。包括基因，代谢和分类，用于区别两个或两个以上生物条件（或者是类群）。该算法强调的是统计意义和生物相关性。让研究人员能够识别不同丰度的特征以及相关联的类别。通过生物学统计差异使其具有强大的识别功能。然后，它执行额外的测试，以评估这些差异是否符合预期的生物学行为。

使用软件Lefse软件实现,也可以使用在线LEFSE实现（https://huttenhower.sph.harvard.edu/galaxy/）。

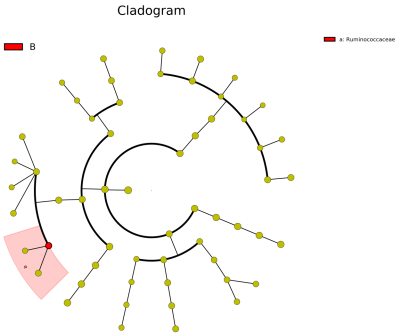


图6-1-1 LEFSE分析

图中为LEfSe聚类树，不同颜色表示不同分组，不同颜色的节点表示在该颜色所代表的分组中起到重要作用的微生物群，一个颜色圈点代表一个biomarker，右上角图例为biomarker名称。黄色节点表示的是在不同分组中没有起到重要作用的微生物类群。

结果目录：BGI_results/Diff_Analysis/LEFSe/

7 PICRUST分析

因为很多细菌自身具有不同数目的16SrRNA拷贝数，我们利用软件PICRUSt[3]对由16S测序样品中可能存在的各级KEGG通路及丰度值以及COG功能信息及丰度值进行预测。下表展示了KOs/COGs差异统计结果。

使用PIRCUST软件实现。

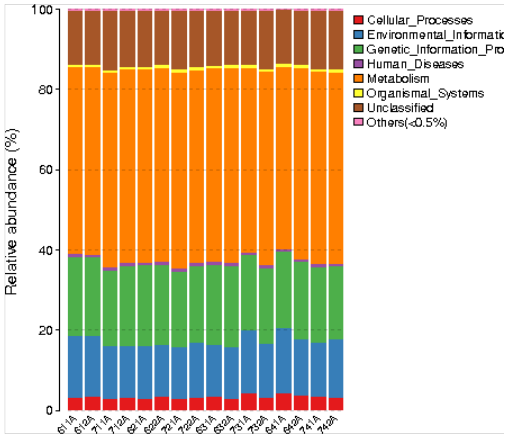


图7-1-1 PICRUST分析

KEGG通路分类柱状图。 图中同样颜色表示同样的一级通路。柱高表示相对丰度。

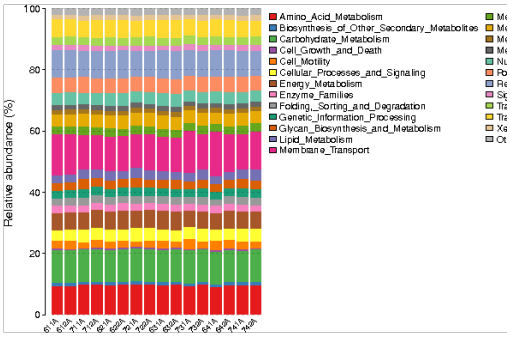


图7-1-1 PICRUST分析

KEGG通路分类柱状图。 图中同样颜色表示同样的一级通路。柱高表示相对丰度。

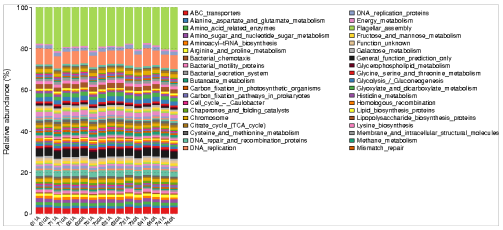


图7-1-1 PICRUST分析

KEGG通路分类柱状图。 图中同样颜色表示同样的一级通路。柱高表示相对丰度。

结果目录：BGI_results/Functional_Analysis/

8 信息挖掘推荐

基于16S，18S，ITS或功能基因的物种多样性和丰度分析主要应用于宿主肠道、土壤、水体等环境中，而绝大部分的研究均是比较不同环境或不同条件下物种组成和丰度上的差异性[13-17]，所以针对不同环境来源的样品，物种组成差异始终是分析的重点。

在物种组成和丰度差异上，结果**OTU_Cluster_Taxonomy**目录下中的OTU_table_for_biom.txt代表了不同样品中的物种多样性结果（即注释上了多少个物种，以及在什么水平上注释上的），更直观的结果参见**3.Taxa_summary，4.Species_heatmap**/中将注释上的物种在不同样品间相应水平进行比较，直观展示了不同样品在门、纲、目、科、属、种上的物种丰度，请关注能够明显显示样品间差异的分类水平，如"科"水平，进而进行相应科的背景的阐述和总结，从而关联物种组成差异和环境适应性。

在样品复杂度的分析上，结果**Alpha_Diversity**/目录和**Beta_Diversity**/下的各个文件是查看的重点，分别涵盖了单个样品内部的多样性，样品两两之间多样性的比较，以及样品间聚类的分析结果。

如果样品进行了分组分析，OTU丰度PCA分析，物种丰度热图以及Beta多样性热图，都可以直观反映出来组间有无差异。如果属于同一组别的样品在这些分析中均聚类在一起，说明了同一组别的样品在物种组成上具有相似性。最后**Diff_Analysis**/的结果从统计学的角度找出两个组别差异的物种。

三 参考文献

[1] Douglas WF, Bing M, Pawel G, Naomi S, Sandra O, Rebecca MB. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:6.

[2] T. Magoc and S. Salzberg. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21): 2957-63.

[3] Edgar, R.C. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads,Nature Methods. 10(10):996-8.

[4] Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 27:2194-2200.

[5] J Gregory C, Justin K, Jesse S. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 7:335-336.

[6] Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl. Acids Res. 41:D633-D642.

[7] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Gl?ckner FO. 2013.: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.

[8] DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-72.

[9] Abarenkov, Kessy; Nilsson, R. Henrik; Larsson, Karl-Henrik; Alexander, Ian J.; Eberhardt, Ursula; Erland, Susanne; H?iland, Klaus; Kj?ller, Rasmus; Larsson, Ellen; Pennanen, Taina; Sen, Robin; Taylor, Andy F. S.; Tedersoo, Leho; Ursing, Bj?rn M.; Vr?lstad, Trude; Liimatainen, Kare; Peintner, Ursula; K?ljalg, Urmas. 2010. The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytologist. 186(2), 281-285.

[10] Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 73:5261-5267.

[11] Patrick DS, Sarah LW et al. (2009). Introducing mothur: Open-Source, Platform- Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75(23):7537-7541.

[12] James RW, Niranjan N, Mihai P. 2009. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS computational biology.

[13] McCafferty J, Mühlbauer M, Gharaibeh RZ, et al. (2013) Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J. 7(11):2116-25.

[14] Zhao L, Wang G, Siegel P, et al. (2013) Quantitative genetic background of the host influences gut microbiomes in chickens. Sci Rep. 3:1163.

[15] Rubin BE, Gibbons SM, Kennedy S, et al. (2013) Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS One. 8(7):1:6.

[16] Mao Y, Xia Y, Zhang T. (2013) Characterization of Thauera-dominated hydrogen-oxidizing autotrophic denitrifying microbial communities by using high-throughput sequencing. Bioresour Technol. 128:703-10.

[17] Peng X, Yu KQ, Deng GH, et al. (2013) Comparison of direct boiling method with commercial kits for extracting fecal microbiome DNA by Illumina sequencing of 16S rRNA tags. J Microbiol Methods. 95(3):455-62.

四 常用数据格式介绍

1 FASTQ格式

Illumina测序原始数据下机文件为FASTQ格式，是序列格式常见的一种。FASTQ格式的序列一般包括四行。第一行由'@'开头，后面为序列的描述信息。第二行为序列。第三行由'+'开始，后面也可以跟序列的描述信息，一般省略。第四行为第二行碱基的质量评价，字符数与第二行相等。FASTQ文件举例如下：

read1的FASTQ文件 x1.fq中第一条reads：

@FC4290FAAXX:4:1:3:84#CAGATC/1

AGTTCCGGCGCACGGGTGAGTAACGCGTATCCAACCTTCCCCTTAGTAGGGCATAGCCCGGCGAAAGTCGGATTAATACTCTATGTTTTCCGTCGAGGACATCTGAAGTGGAACAAAGATT

+

CCCCFGCGGGFCCFGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGAFGFFGGFGGGGGGDGGGECBCFGDGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGEGGGGFGGGEFFEGFGFCF

read2的FASTQ文件 x2.fq中第一条reads：

@FC4290FAAXX:4:1:3:84#CAGATC/2

CCGCTGCTGCTGGCACGGAATTAGCCGGTCCTTATTCATCAGGTACCTACAAAAAGGGACACGTCCCTCACTTTATCCCCTGATAAAAGCAGTTTACAACCCATAGGGCCGTCATCCTGCA

+

CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGEDAFGGGGGGGGFEGGGFGGGGGGGGCFCGFFGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

2 FASTA格式

FASTA格式（又称为Pearson格式），是一种基于文本用于表示核苷酸序列或氨基酸序列的格式。在这种格式中碱基对或氨基酸用单个字母来编码，且允许在序列前添加序列名及注释。序列文件的第一行是由大于号">"或分号";"打头的任意文字说明（习惯常用">"作为起始），用于序列标记及对该序列的描述。从第二行开始为序列本身。如：

>Sample1_Tag1

TGAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGCAGGATGACGGTCCTATGGATTGTAAACTGCTTTTGTACGAGAAGAAAC

ACTCCTACGTGTAGGAGCTTGACGGTATCGTAAGAATAAGGATCGGCTAACTCC

Reads拼接结果及OTU代表序列结果均为FASTA格式。

3 OTU格式说明

OTU_taxonomy.xls 举例如下：

#OTUId	Abundance	Taxonomy
Otu1	1200	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;Acinetobacter_lwoffii
Otu2	900	Bacteria;Firmicutes;Clostridia;Thermoanaerobacterales;Thermodesulfobiaceae;Coprothermobacter
Otu3	4100	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas

第一列为OTU编号，第二列为OTU丰度，第三列为OTU注释上的物种分类信息。

OTU_table_for_biom.txt 举例如下：

#OTU ID	Sample1	Sample2	Sample3	taxonomy
Otu1	100	300	800	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter_lwoffii
Otu2	500	200	200	Bacteria; Firmicutes; Clostridia; Thermoanaerobacterales; Thermodesulfobiaceae; Coprothermobacter
Otu3	1000	1000	2100	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas

第一列为OTU编号，从第二列到倒数第二列为OTU所代表的各样品的tag数量，最后一列为OTU代表物种信息。

OTU_stat_detail.xls 举例如下：

Num samples: 3

Num total OTUs: 121

Num Singletons: 0

Num Non-singletons: 121

Num total sequences: 6200

Sample all OTU summary:

Min: 64

Max: 83

Mean: 73

Standard deviation: 7

Sample detail:

Sample name Tag number OTU number

Sample1 1200 64

Sample2 900 72

Sample3 4100 79

Num samples: 样品个数；Num total OTUs：OTU总数；Num Singletons: 丰度为1的OTU数目；Num Non-singletons: 丰度为1以上的OTU数目Num total sequences: Reads总数。

Sample all OTU summary:：所有OTU的信息摘要，包括最小OTU数目（Min），最大OTU数目（Max），平均OTU数目（Mean），标准差（Standard deviation）。

如果去除低丰度的OTU，该文件中还包括Non-singletons OTU的信息摘要。

Sample detail:每个样品OTU数目。第一列为样品名称；第二列为个样品中总的OTU数目；第三列为个样品中Non-singletons OTU的数目；第四列为Singletons OTU占总OTU的比例。

***sharedOTU.venn.xls** 举例如下：

group	shareOTU_num/uniuqeOTU_num	OTU_ID
Sample1-vs-Sample2	80	Otu1,Otu10,Otu100,Otu101,Otu102,Otu103...
Sample1-vs-Sample3	85	Otu1,Otu10,Otu100,Otu101,Otu102,Otu103...
Sample1	5	Otu113,Otu116,Otu118,Otu60,Otu98

该文件与韦恩图对应，为共有或特有OTU的统计文件。第一列为组名；第二列为组间共有或组特有的OTU数目；第三列为组间共有或组特有的OTU ID。

OTU_PCA.coordinate.xls 举例如下：

	Axis1	Axis2
Sample1	1.535	-3.381
Sample2	0.5531	-3.0168
Sample3	3.2705	3.4669

第一列为样品名称；第二列为x轴坐标（即PC1坐标）；第二列为y轴坐标（即PC2坐标）。

Genus.phylogeny.tree：此文件为Newick格式（进化树的常用格式），能被多种软件识别，例如FigTree，Treebest，PHYLIP等。

4 Alpha多样性结果文件格式说明

Alpha_diversity.detail.xls 举例如下：

label	group	sobs	chao	chao_lci	chao_hci	ace	ace_lci	ace_hci	...
0.03	Sample1	156.000000	521.500000	335.876875	898.676067	378.948085	283.807259	544.912564	...
0.03	Sample2	151.000000	297.176471	226.692999	433.292428	294.071360	231.482437	405.333926	...

第一列为差异水平；第二列为样品名称；第三列为observed species指数，反映样本中的OTU数；从第四列开始到最后一列，每3列为单位，第一列为95%置信区间计算得到的alpha多样性值，第二和第三列分别95%置信区间计算得到的最小值与最大值（即表中lci与hci）。

***.Alpha.test.reslut.xls** 举例说明：

#Alpha	mean(Group1)	SD(Group1)	mean(Group2)	SD(Group2)	...	p-vaule
sobs	191.00000	43.24350	682.50000	166.91016	...	0.00075
chao	292.44654	16.15483	898.39669	154.11534	...	0.00063
ace	382.62500	177.89841	854.80273	154.54333	...	0.00114
shannon	2.33207	4.34743	3.04424	4.27114	...	0.00066
simpson	0.18282	0.03691	0.15808	0.04751	...	0.00167

第一列为Alpha多样性指数名称；从第二列开始，两列为单位，分别为每组Alpha多样性注释的均值和SD值；最后一列为组间Alpha多样性注释差异检验p值。

5 Beta多样性分析文件格式说明

***Beta_diversity.xls** 举例如下：

	Sample1	Sample2	Sample3	Sample4
Sample1	0.0	0.529648285613	0.690040650407	0.570607988689
Sample2	0.529648285613	0.0	0.597119123365	0.449893955461
Sample3	0.690040650407	0.597119123365	0.0	0.708907741251
Sample4	0.570607988689	0.449893955461	0.708907741251	0.0

该文件为对称的矩阵，表中的数值为第一行与第一列两两样品计算出来的距离矩阵。

6 差异分析文件格式说明

*** differentially_abundant.xls** 举例如下：

Phylum	mean(Group1)	variance(Group1)	std.err(Group1)	mean(Group2)	variance(Group2)	std.err(Group2)	p-vlaue	FDR
Actinobacteria	0.0490001	4.8e-05	0.039710	0.073479	0.000135	0.0601112	0.653410	0.682357
Proteobacteria	0.014310	2e-06	0.008470	0.018354	1e-06	0.0069	0.685421	0.676623

第一例为物种名称；第二，三，四列分别为Group1平均丰度，方差与标准差；第五，六，七列分别为Group2平均丰度，方差与标准差；第八列为两组比较的p值；第九列为两组比较假发现率值。