

**16S\_Demo\_结题报告**

深圳市恒创基因科技有限公司

2018年01月01日

目录

[一 概述 3](#_Toc503170748)

[二 项目流程 4](#_Toc503170749)

[1 实验流程 4](#_Toc503170750)

[2 信息分析流程 4](#_Toc503170751)

[三 分析结果 5](#_Toc503170752)

[1 下机数据处理 5](#_Toc503170753)

[2 OTU及物种群落分析 6](#_Toc503170754)

[2.1 OTU聚类及注释分析 6](#_Toc503170755)

[2.2 OTU系统进化分析 8](#_Toc503170756)

[2.2.1特定属种水平之内OTU整体系统进化分析 8](#_Toc503170757)

[2.2.2特定物种系统进化分析 9](#_Toc503170758)

[3 Alpha Diversity分析 10](#_Toc503170759)

[3.1 Alpha Diversity指数统计分析 10](#_Toc503170760)

[3.2 组间比较Alpha diversity指数 12](#_Toc503170761)

[4 Beta Diversity分析 14](#_Toc503170762)

[4.1 Beta Diversity指数统计分析 14](#_Toc503170763)

[4.2 组间比较Beta Diversity指数 17](#_Toc503170764)

[5 OTU水平样品组间统计分析 18](#_Toc503170765)

[5.1 组间OTU差异显著性分析 18](#_Toc503170766)

[5.2样品共有OTU分析 19](#_Toc503170767)

[6 样品微生物组成功能预测分析 21](#_Toc503170768)

[6.1预测微生物群落功能 21](#_Toc503170769)

[6.2预测功能PCA分析 21](#_Toc503170770)

[6.3预测功能聚类分析 22](#_Toc503170771)

[6.4组间功能差异性分析 23](#_Toc503170772)

[7 关联统计分析 24](#_Toc503170773)

[7.1 CCA/RDA分析 24](#_Toc503170774)

[7.2 PERMANOVA分析 25](#_Toc503170775)

[7.3 物种相互作用network分析 27](#_Toc503170776)

[7.4 Source tracker分析 28](#_Toc503170777)

[四 分析方法及参考文献 29](#_Toc503170778)

[五 交付结果目录结构 37](#_Toc503170779)

[六 联系我们 37](#_Toc503170780)

# 一 概述

微生物世界是分子多样性最大的天然资源库，基于菌株水平的传统分离培养技术为人们认识微生物多样性提供了可能，但是据估计自然界中超过99%的微生物不能通过传统的分离培养技术获得其纯培养，从而导致环境微生物中的多样性基因资源难以被发现。许多重要的微生物我们还不能识别，随着微生物活性产物的广泛研究和深入开发利用，从环境微生物中筛选到新活性物质的几率将逐步下降。而如何开拓利用环境微生物新资源是微生物研究的重要课题。为此研究者们开发了多种以特定环境微生物为研究对象的高通量测序方法。细菌基因组相对较小，通常仅有一条环状DNA和质粒，通过高通量测序，可以了解其全部遗传信息。这也已经成为微生物研究的重要手段之一，为细菌的遗传进化、疾病预防与治疗、疫苗与抗生素的开发等提供重要的信息。

16S rRNA 测序技术是最常用的高通量测序依赖的组学技术之一，该技术着眼于对肠道微生物群落菌种组成的分析。细菌16S rRNA基因具有保守区与可变区间隔排列的特征，其中的可变区一般具有菌种特异性，并且可以反映细菌间亲缘关系的远近，因此通过分析可变区的序列即可得到各细菌的分类学特征。16S rRNA 基因序列包括9个可变区和10个保守区。传统分子生物学方法中有时也应用 Sanger法对单一菌种的16S rRNA进行测序和鉴定，而 16SrRNA 测序技术通过结合高通量测序技术的高通量优势和16S rRNA基因的菌种鉴定优势，实现了对复杂样品中混合菌种的分类学鉴定和精确定量。16SrRNA 测序技术的基本流程通过提取实验样品的DNA，并扩增16S rDNA某个可变区，采用高通量测序仪Miseq或Hiseq对其进行测序，通过生物信息学分析可以获得特定实验样品中细菌或古菌物种组成、物种丰度、系统进化、群落比较等诸多信息。

# 二 项目流程

## 1 实验流程

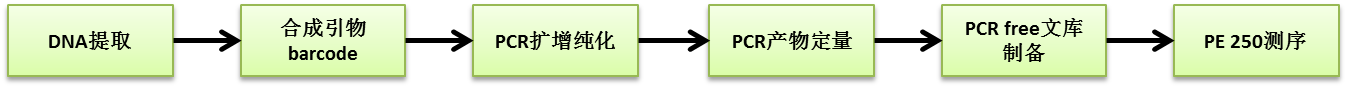


图2-1 实验工作流程图

## 2 信息分析流程

为了保证OTU(Operational Taxonomic Units)聚类及后续分析的准确性，首先会对原始数据(Raw reads)进行过滤处理，处理后的数据进行拼接、过滤，得到有效数据(Effective Tags)。然后基于有效数据进行OTU聚类和物种分类分析，形成OTU和其他物种分类等级的物种丰度谱。基于数据均一化后的OTU物种丰度谱，再对OTU进行丰度、多样性指数等分析，同时对物种注释在各个分类水平上进行群落结构的统计分析。还可以在以上分析的基础上，进行一系列的基于OTU、物种组成的聚类分析和统计比较分析，挖掘样品之间的物种组成差异，同时结合环境及临床因子等进行关联统计分析，寻找与其显著相关的物种群落。同时还可以基于16S数据，进行微生物群路的功能预测分析。



图2-2 信息分析流程图

# 三 分析结果

## 1 下机数据处理

采用Illumina MiSeq/HiSeq测序平台得到的下机数据(Raw Data)经过数据拆分、去引物序列、PE Reads拼接、Tags质量及长度过滤和截取以及去嵌合体后获得最终的有效序列(Effective Tags)，具体的数据产出统计结果见表3-1。

表3-1 数据预处理统计及质控

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample Name | Raw PE(bp) | Effectiv Tags(bp) | Base(bp) | AvgLen(bp) | Q20 | Q30 | Effective(%) |
| Sample1 | 37542 | 35866 | 9177946 | 255.9 | 87.71% | 82.27% | 95.54 |
| Sample2 | 38034 | 37002 | 9469669 | 255.92 | 87.15% | 81.62% | 97.29 |
| Sample3 | 48297 | 47182 | 12074726 | 255.92 | 87.66% | 82.22% | 97.69 |
| Sample4 | 54486 | 53363 | 13656221 | 255.91 | 88.95% | 83.86% | 97.94 |
| Sample5 | 55340 | 54479 | 13942096 | 255.92 | 86.99% | 81.38% | 98.44 |
| Sample6 | 49514 | 48494 | 12408340 | 255.87 | 87.55% | 82.08% | 97.94 |
| Sample7 | 37448 | 36366 | 9307842 | 255.95 | 88.80% | 83.67% | 97.11 |
| Sample8 | 34647 | 33211 | 8500181 | 255.94 | 87.84% | 82.54% | 95.86 |
| Sample9 | 56527 | 54264 | 13886391 | 255.9 | 87.44% | 81.98% | 96 |

说明：Raw PE：原始下机的PE reads数目；Effective Tags：过滤嵌合体后，最终用于后续分析的Tags序列数目；Base：最终 Effective Data的碱基数目，AvgLen：Effective Tags 的平均长度；Q20和Q30：Effective Tags中碱基质量值大于20(测序错误率小于1%)和30(测序错误率小于0.1%)的碱基百分比；Effective%：Effective Tags的数目与Raw data中的PE Reads数目的百分比。

结果目录：Result/1-Demux，存放样本的原始PE reads序列、拼接的序列、低质量及嵌合体过滤后的Demultiplexed序列以及相关的序列长度和数量的统计文件。目录下文件详细解释请见README文件。

重要：qzv文件是Qiime2的专属可视化文件格式，可以通过网页(支持浏览器为Mozilla Firefox和Google Chrome)<https://view.qiime2.org/>进行交互性查看(推荐)，如果电脑装有Qiime2软件也可通过命令$ qiime tools view \*.qzv进行查看。结果文件夹中只提取了部分主要信息用于产生报告，然而qzv文件中包含大量其他有用的交互式信息（比如可以选择不同的分类水平查询物种相对丰度），请用户自行查看。关于\*.qzv文件的查看后文不再反复描述，如果有问题请联系客服部门。

## 2 OTU及物种群落分析

### 2.1 OTU聚类及注释分析

为了研究样品物种组成及多样性信息，对所有样品的全部的有效序列进行聚类（默认选取相似度为97%），形成OTU(Operational Taxonomic Unit)。选取OTU的代表性序列，与核糖体RNA数据库(Greengenes Database [99%])进行比对获得物种注释信息。基于物种注释信息，去除注释为叶绿体、线粒体以及不能注释到界级别的OTU及其包含的序列。基于OTU的绝对丰度及注释信息，对每个样品在一共7个分类水平界门纲目科属种(Kingdom,Phylum,Class,Order,Family,Genus,Species)上的序列数目占总序列数的比例进行统计，可以有效的评估样本的物种注释分辨率（注释到属的比例越高表示样本的OTU注释效果越好）及样本的物种复杂度(注释到属的比例越低表示样本的物种复杂度越高)，图3-1同时统计了各分类水平的物种在每个样本中的相对丰度。图3-2主要结合OTU或者物种注释信息展示了门级别的物种相对分布情况，其他分类等级的物种分布情况请参见结果部分。

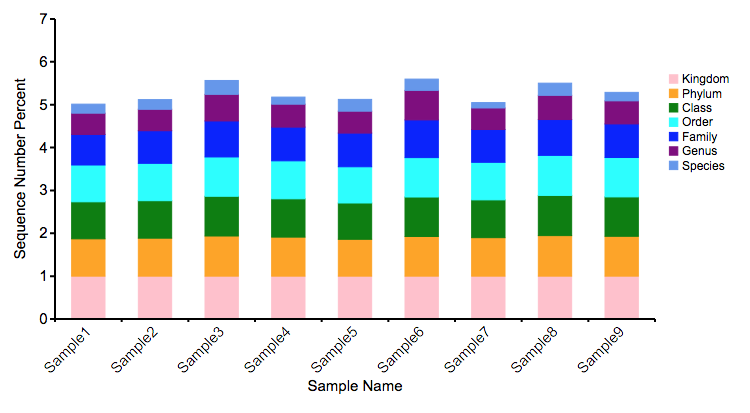


图3-1 每个样品在各分类水平上的序列构成柱形图

说明：横坐标(Sample Name)是样品名，纵坐标(Sequence Number Percent)表示注释到该水平的序列数目占总注释数据的比率，柱状图自上而下的颜色顺序对应于右侧的图例颜色顺序。每个分类水平最高值为1，代表100%的序列都得到了至少在这个级别的注释。

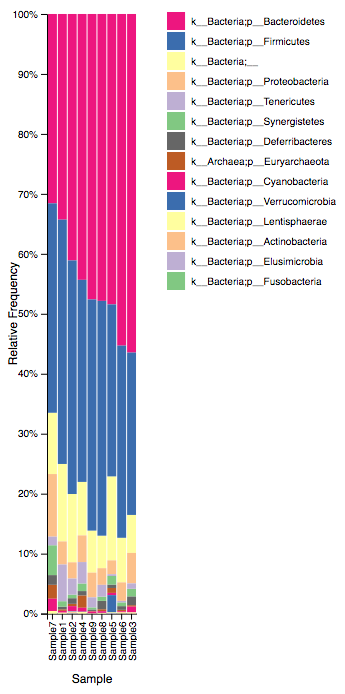


图3-2 门水平上的物种相对丰度柱形图

说明：横坐标(Sample Name)是样品名，纵坐标(Relative Abundance)表示相对丰度。其他分类水平的物种相对丰度图也都可以通过交互式网页打开，并对样本或者OTU在图片中呈现的顺序根据meta data或者相对丰度进行调节（强烈推荐）。具体查看方法将在本部分的结果目录部分和结果的README文件中详细阐述。

为了研究不同样品间的相似性，还可以通过对样品进行聚类分析从而构建样品的聚类树。选取关注的物种OTU（默认选取OTU绝对丰度大于1000），结合其系统发生关系数据，以及它们对应的OTU的相对丰度信息，实现OTU水平上的样品聚类（纵向聚类），以此考察不同样品或者分组间的相似或差异性（图3-3，此处为门水平分类的热图），并从分类信息和样品间差异两个层面进行聚类，寻找物种或样本的聚集规律。

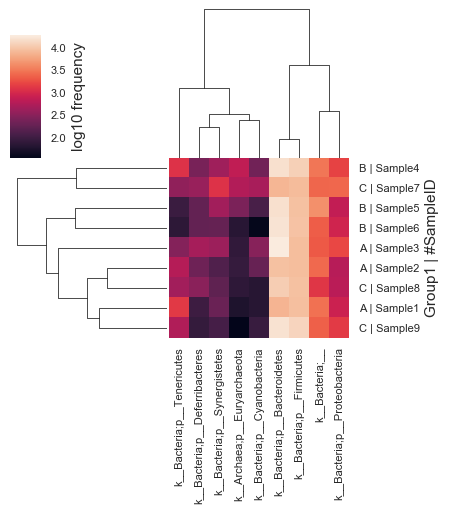


图3-3 门水平OTU丰度聚类图

说明：纵向为样品名称信息，同时也包括了分组信息。横向为OTU在该分类水平注释名称。图中上方的聚类树为OTU的系统发生关系及其物种注释，左侧的聚类树为样品聚类树，中间的热图是OTU的相对丰度热图，颜色与相对丰度的关系见图上方的刻度尺。

### 2.2 OTU系统进化分析

2.2.1特定属种水平之内OTU整体系统进化分析

基于OTU的物种系统进化关系及相对丰度信息，使用PhyloSeq（R Package）对该届水平和多个门水平之内选择的主要的OTU的物种注释结果进行可视化展示，可以很直观揭示该门水平内的OTU的物种在实验样品中分布和进化信息（详细请见结果部分）。如图3-4所示，分析样品在Bacteroidetes门下各个子分类的分布情况。

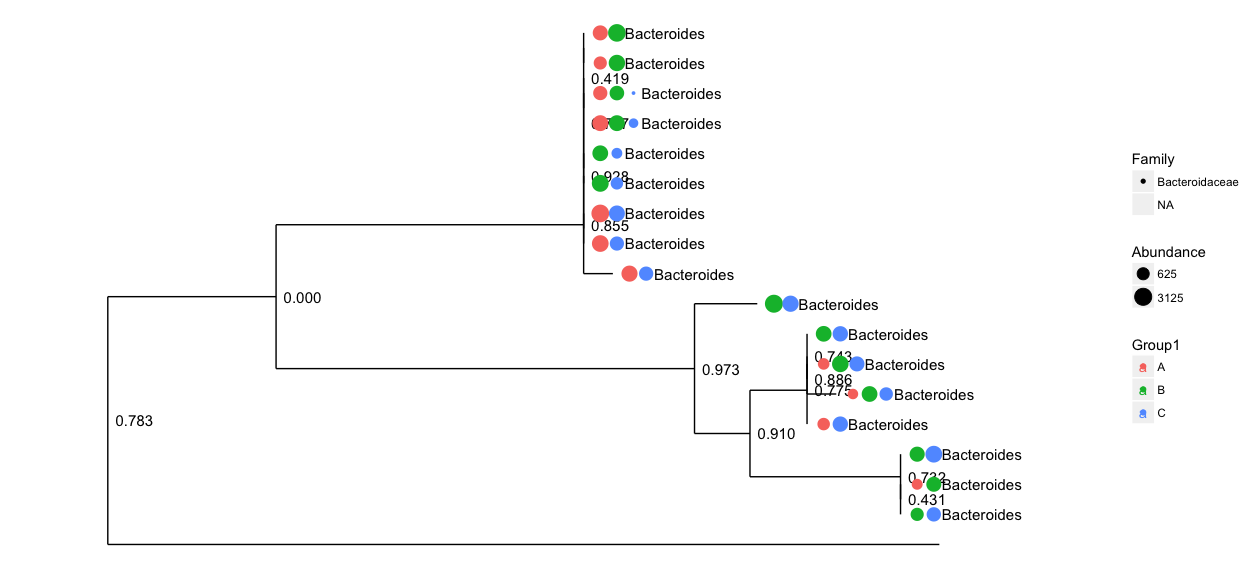


图3-4 特定属物种系统发生树

说明：分支末端不同颜色的注释表示不同的分组，形状代表不同的科水平分类，圆圈的大小表示该OTU在该分组的相对丰度大小，对应右侧图标；分支末端的节点的名称是该OTU在属分类水平的名字；数字代表在该内部节点的置信度，范围是0到1，数字越大代表这个节点的可信度越高。

2.2.2特定物种系统进化分析

选取关注的OTU代表序列进行系统进化分析(默认选取所有丰度大于1000的OTU进行分析)，同时可以结合OTU的相对丰度进行可视化展示（图3-5）,利用iTOL可以直观地展示研究环境中的物种群落组成多样性及其系统进化关系。详细的数据分析推荐用户使用iTOL网页直接上传打开iTOL分析文件。如果需要也可以基于单个样品的物种分类结果，筛选特别关注的物种进行样本内和样本间的物种分类树及物种相对丰度统计展示。

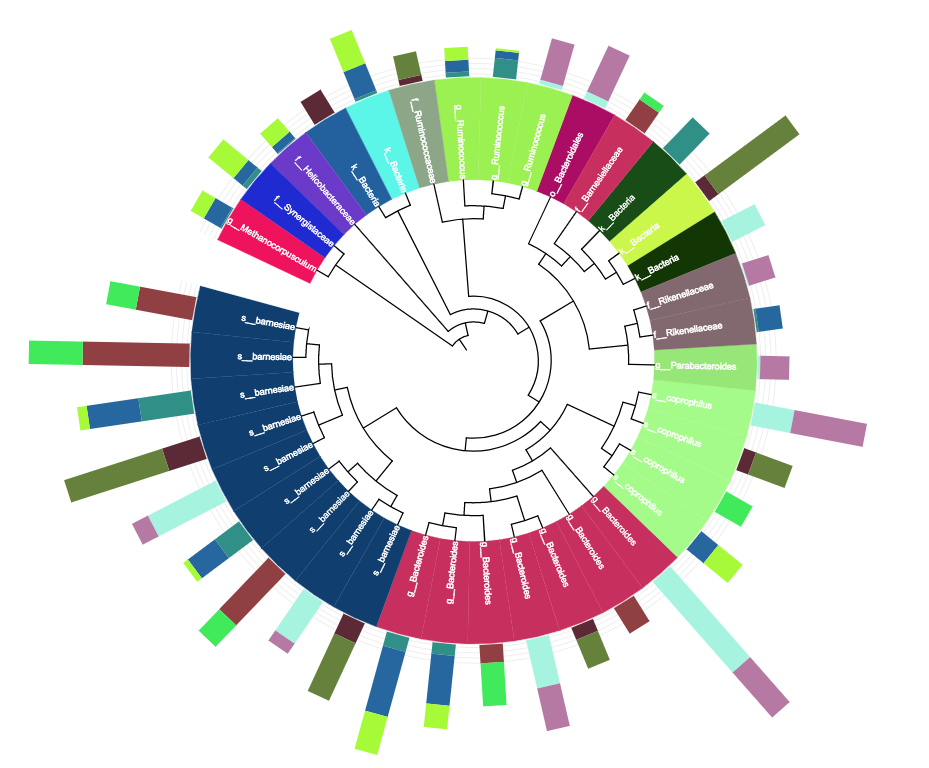


图3-5 主要OTU的系统发生关系及其物种注释

说明：圈图共有两层，由里往外：第一层是OTU代表序列构建的系统发育树，分支的颜色表示其对应的属名，每种颜色代表一个不同的分组；第二层是OTU的相对丰度分布，柱子的高度表示OTU的相对丰度大小，不同的颜色代表在不同的样品中的相对丰度，如果丰度为零则该样品没有对应颜色。

结果目录：结果目录：Result/2-OTUAnalysis，文件详细解释见README文件。

## 3 Alpha Diversity分析

### 3.1 Alpha Diversity指数统计分析

Alpha多样性（Alpha Diversity）是对某个样品中物种多样性的分析，包含样品中的物种组成的丰富度(Richness)和均匀度(Evenness)两个因素，通常用Observed OTU，Shannon, Evenness以及Faith’s Phylogenetic Diversity等指数来评估某个样本的物种多样性，指数越高，表明样本的多样性越复杂。Observed species指数是指样本中实际测定得到的OTU数量，衡量样品中OTU丰富度(Richness)的指数。Shannon指数,它的计算考虑到样品中的分类总数(Richness)，和每个分类所占的比例(Abundance)。物种均一度(Evenness)用来描述物种中的个体的相对丰富度或所占比例。Faith’s Phylogenetic Diversity是基于系统发生树来计算的一种多样性指数，它用各个样品中OTU的代表序列计算出构建系统发生树的距离，将某一样品中的所有代表序列的值加和，从而得到的数值。对不同的样本alpha diversity分析指数进行统计（默认提供97%阈值）（表3-2）。其中97%水平下聚类成为一个OTU的序列被认为可能是源自于同一个种(Species Boundary)的序列。

表3-2 Alpha Indices统计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | observed\_otus | chao1 | shannon | faith\_pd |
| Sample1 | 316 | 316 | 6.94 | 26.50 |
| Sample2 | 328 | 328 | 7.16 | 28.13 |
| Sample3 | 308 | 308 | 6.42 | 24.25 |
| Sample4 | 420 | 420 | 7.20 | 28.92 |
| Sample5 | 293 | 293 | 6.03 | 24.51 |
| Sample6 | 238 | 238 | 5.74 | 22.39 |
| Sample7 | 303 | 303 | 6.84 | 25.60 |
| Sample8 | 320 | 320 | 6.74 | 24.64 |
| Sample9 | 340 | 340 | 6.63 | 25.16 |

Alpha多样性稀释曲线(Rarefaction Curve)是从样品中随机抽取一定测序量的数据，统计它们所代表物种数目(也即是OTU数目)或多样性指数，以数据量与物种多样性来构建的曲线，以用来说明样品的测序数据量是否合理，并间接反映样品中物种的丰富程度。稀释性曲线图中，当曲线趋向平坦时，说明测序数据量渐进合理（图3-6为Faith’s Phylogenetic Diversity指数稀释曲线，其他指数的稀释曲线见结果部分）。

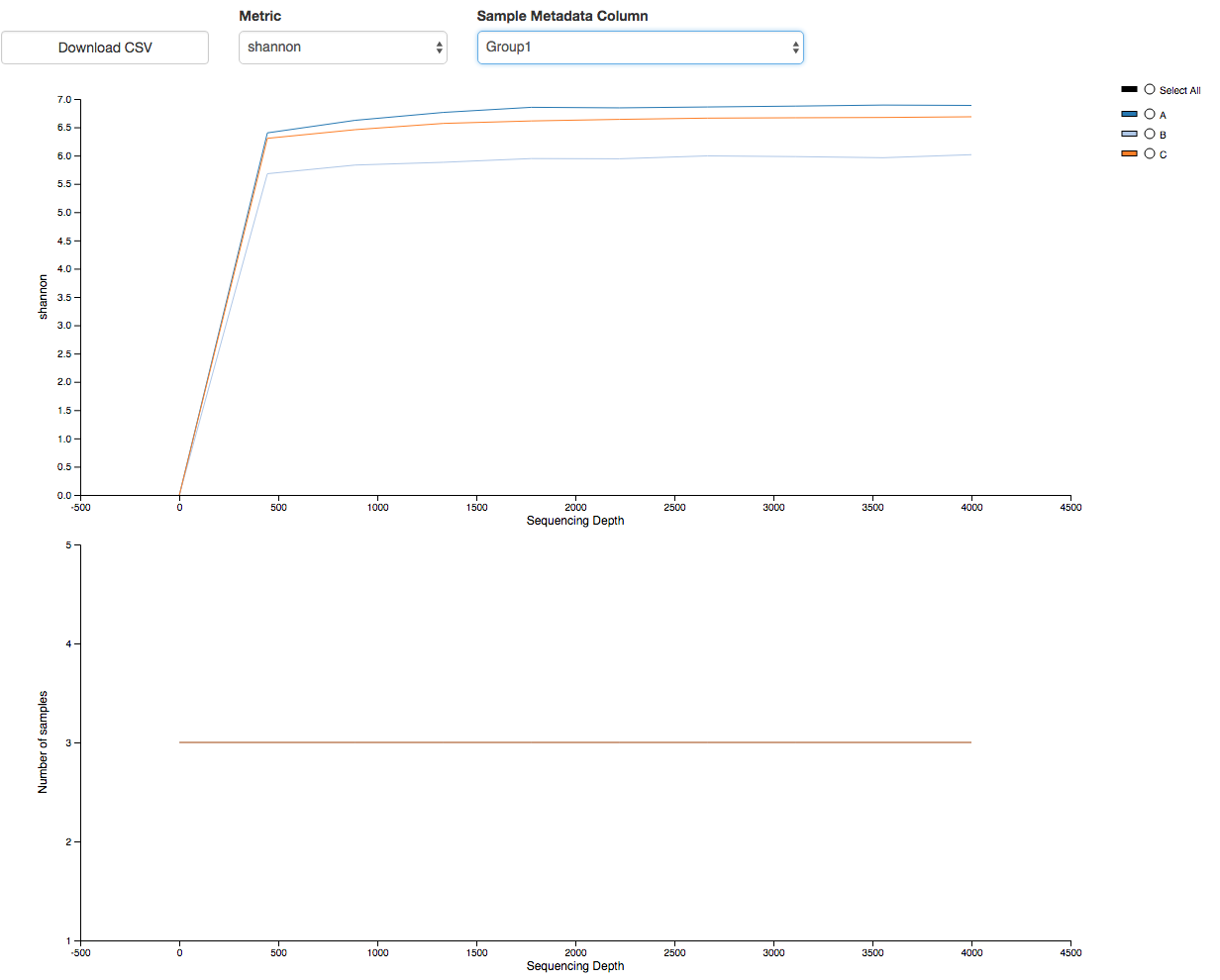


图3-6稀释曲线

说明：本图上部分为在Group1分类情况下Shannon Diversity指数稀释曲线，下半部分对应在相应抽取测序量时，各组仍含有的样品数量，用于查看分析是否合理。

### 3.2 组间比较Alpha diversity指数

在得到整体的Alpha多样性指数之后，接下来结合分组信息来用Kruskal-Wallis方法比较在各个样品分组之间Alpha多样性指数是否有显著性差异（图3-7和表3-3）。

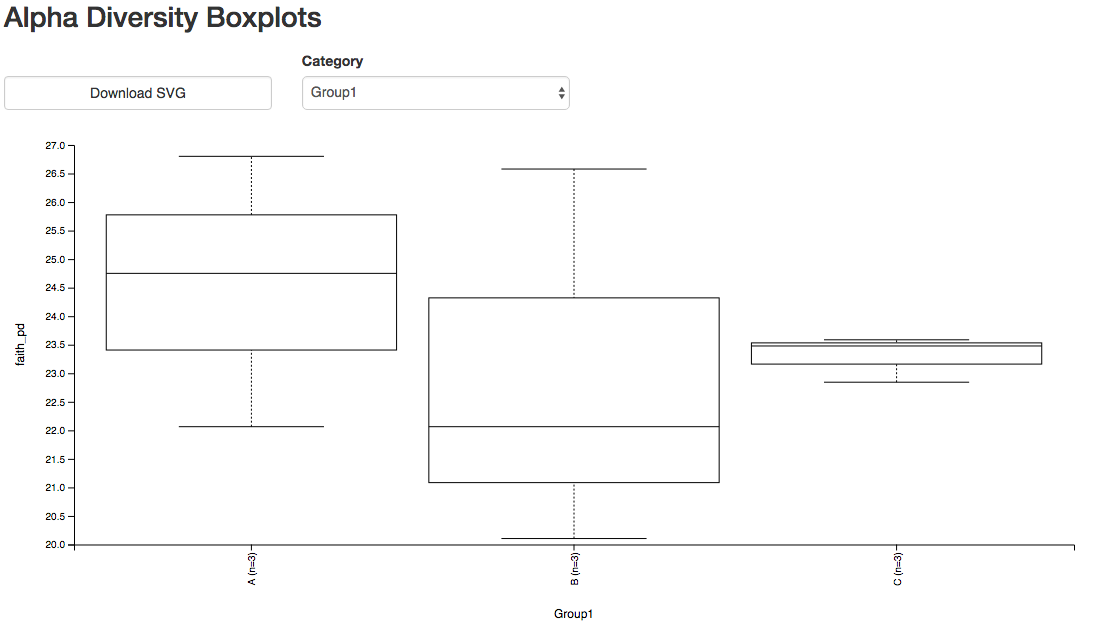
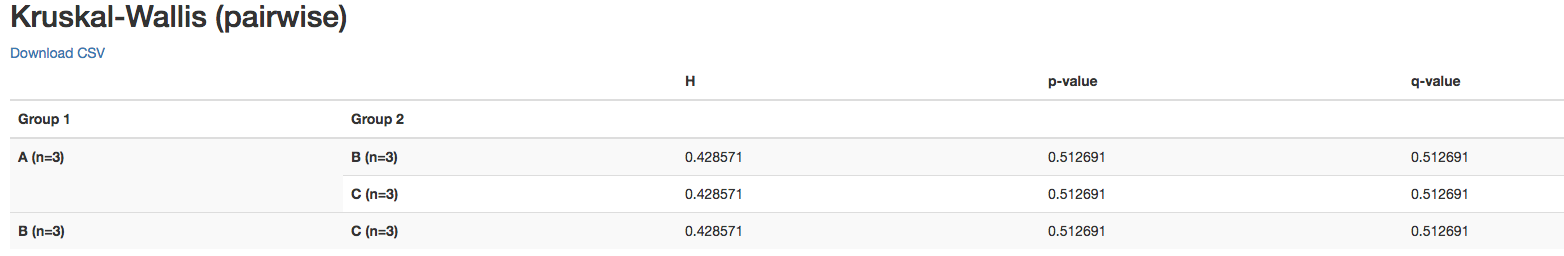


图3-7 各个分组的Alpha多样性指数

说明：横向为分组信息，纵向为Alpha多样性指数

表3-3各个分组的Alpha多样性指数详细数据（截图）



说明：为图3-7的详细数据说明表格

结果目录：Alpha diversity指数统计结果文件为Result/3-AlphaDiversity目录下的相关文件，文件详细解释见README文件。

## 4 Beta Diversity分析

### 4.1 Beta Diversity指数统计分析

Beta Diversity是对不同样品间的微生物群落构成进行比较。根据样本的OTU丰度信息计算Bray curtis，Weighted Unifrac和Unweighted Unifrac距离来评估不同样品间的微生物群落构差异。Bray curtis距离是生态学上反应群落之间差异性最常用的指标，只考虑了物种的丰度信息。Unweighted Unifrac距离是基于物种系统进化关系进行计算的样本间的距离，只考虑了物种的有无。Weighted Unifrac距离是结合OTU的丰度信息和系统进化关系获得的样本间的距离。Uweighted Unifrac距离对稀有物种比较敏感，而Bray curtis和Weighted Unifrac距离则对丰度较高的物种更加敏感。Weighted Unifrac距离，Unweighted Unifrac距离及Bray curtis距离作为Beta多样性距离是衡量两个样品间的相异系数的指标，其值越小，表示这两个样品在物种多样性方面存在的差异越小。图3-8为不同样品之间以上三种距离热图的展示。

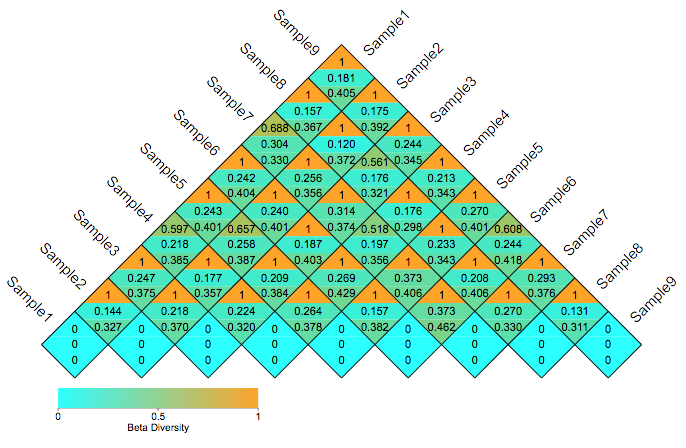


图3-8 Beta多样性指数热图

说明：热图方格的颜色代表样品两两之间的相异系数，相异系数越小的两个样品，物种多样性的差异越小。

主坐标分析（PCoA，Principal Co-ordinates Analysis），是一种与PCA类似的降维排序方法，从多维数据中提取出最主要元素和结构能够提取出最大程度反映样品间差异的三个坐标轴，从而将多维数据的差异反映在三维坐标图上，进而揭示复杂数据背景下的简单规律。区别在于PCA是基于样品的相似系数矩阵来寻找主坐标，而PCoA是基于距离矩阵来寻找主坐标。我们基于Bray curtis距离、Weighted Unifrac距离和Unweighted Unifrac距离来进行PCoA分析，并选取贡献率最大的主坐标组合进行作图展示，图中样品的距离越接近，表示样品的物种组成结构越相似（图3-9为基于Unweighted Unifrac距离进行的PCoA分析，其他距离展示见结果部分）。

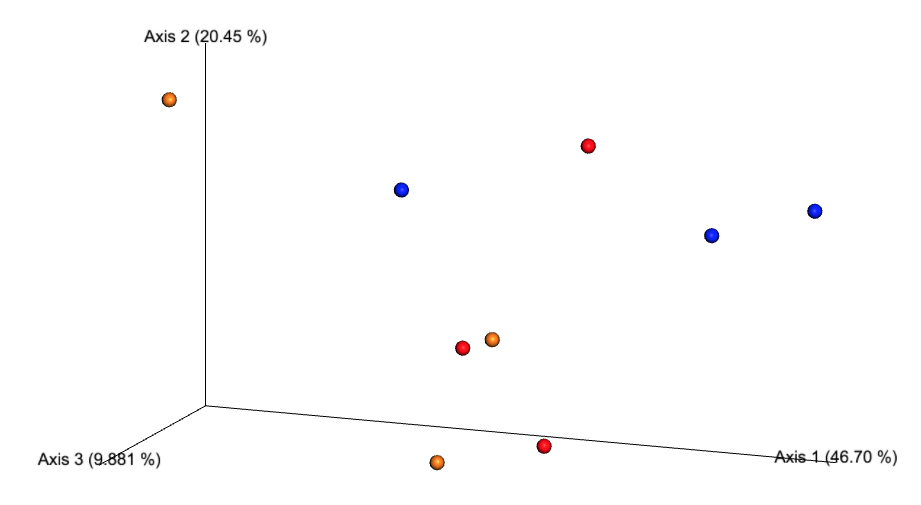


图3-9 3D PCoA分析

说明：横坐标（Axis 1）表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标（Axis 2）表示第二主成分，百分比表示第二主成分对样品差异的贡献值；Axis 3坐标表示第三主成分，百分比表示第三主成分对样品差异的贡献值；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。客户根据meta data的分类信息可以个性化展示其他分类的距离关系。

无度量多维标定法（NMDS ，Non-Metric Multi-Dimensional Scaling）统计是一种适用于生态学研究的排序方法，类似于PCoA，通过样本的分布，或者基因的分布，可以看出组间差异，组内差异等。NMDS包括一类排序方法，其设计目的是为了克服以前的排序方法（即线性模型，包括PCoA）的缺点，NMDS的模型是非线性的，能更好地反映生态学数据的非线性结构。非度量多维尺度法是一种将多维空间的研究对象(样本或变量)简化到低维空间进行定位、分析和归类, 同时又保留对象间原始关系的数据分析方法。适用于无法获得研究对象间精确的相似性或相异性数据，仅能得到他们之间等级关系数据的情形（图3-10为基于Unweighted Unifrac距离进行的NMDS分析，其他距离展示见结果部分）。

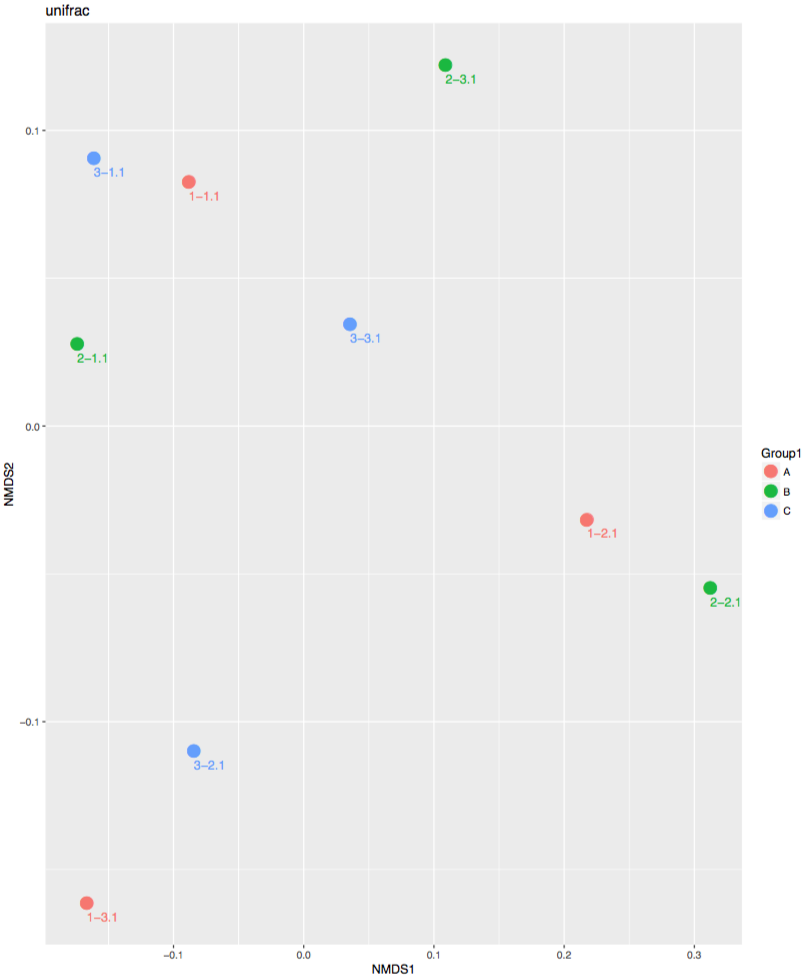


图3-10 2D NMDS分析

说明：图中的点代表样本，点与点之间的距离表示差异程度。

### 4.2 组间比较Beta Diversity指数

在得到整体的Beta Diversity多样性指数之后，接下来结合分组信息来用PERMANOVA(Permutational multivariate analysis of variance)方法和ANOSIM(Analysis of similarities)方法比较在各个样品分组之间的Beta Diversity多样性指数是否有显著性差异（图3-11和表3-4,为基于Unweighted Unifrac距离进行的PERMANOVA分析）。在有其他meta data的情况下，PERMANOVA 也可以用于分析环境因子或其他临床表型数据与微生物群落或物种之间是否显著相关，然后计算环境因子与微生物物种间的Spearman相关系数。

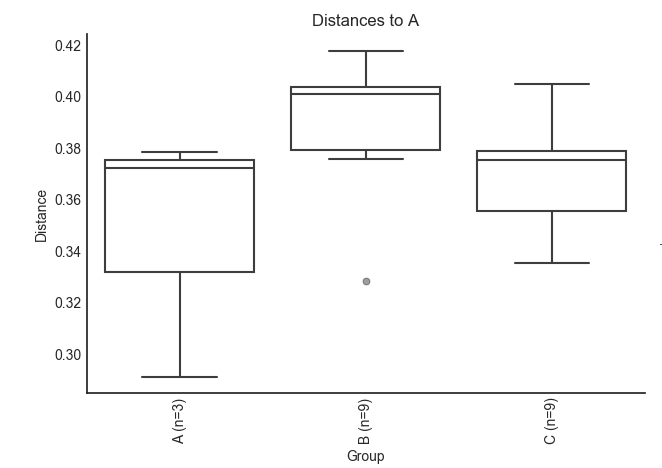
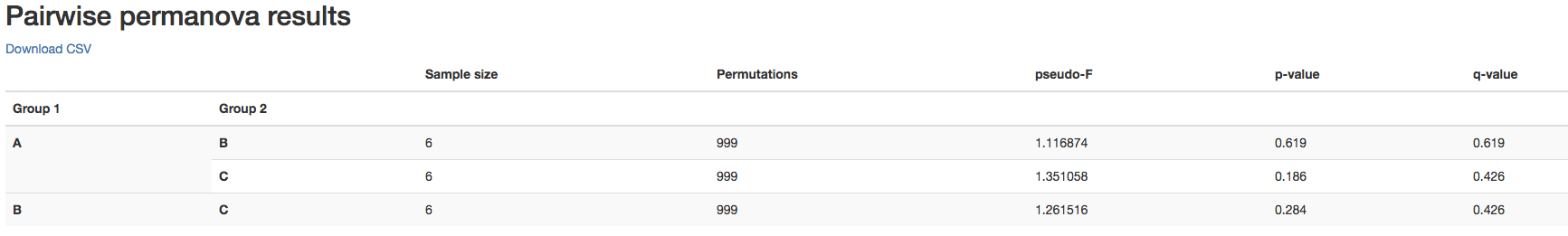


图3-11 各个分组的样品之间距离

说明：横向为分组信息，各个分组名字后面括号内的n值代表在组间的比较次数。纵向为比较的样品之间的距离。有多少个分组就会有多少个box plot，以上图仅展示了所有分组对其中第一个分组的样品之间的距离绘制的box plot，其他组的详细信息请查看结果文件。

表3-4各个分组的Beta多样性指数详细数据（截图）



说明：为图3-11的详细数据说明表格。

结果目录：Beta Diversity指数统计分析结果文件为Result/4-BetaDiversity目录下的相关文件，文件详细解释见README文件。

## 5 OTU水平样品组间统计分析

### 5.1 组间OTU差异显著性分析

ANCOM(Analysis of composition of microbiomes)分析被用于比较OTU在组间之间的显著性差异。如图3-12和表3-5所示，我们可以得知在本实验中有哪些种分类水平OTU在组间表现了丰度的显著差异。同时，各个其他的分类水平也都用ANCOM方法进行了分析，具体请见结果部分。

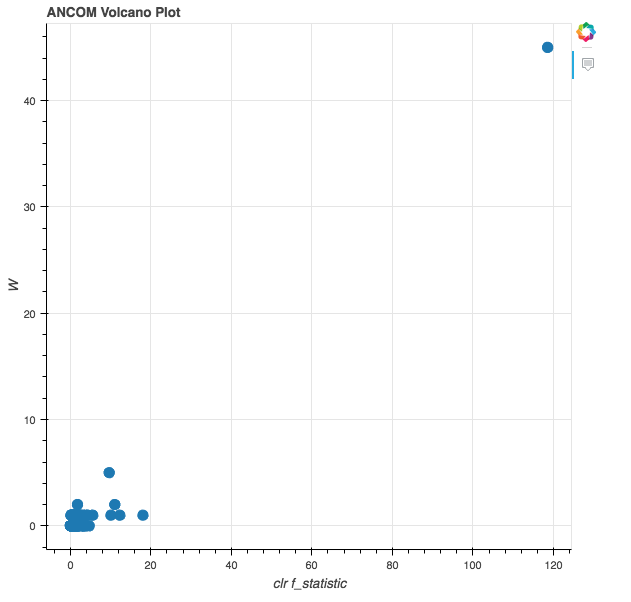
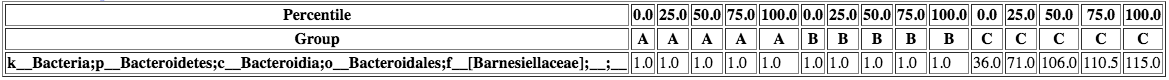


图3-12 OTU种水平的组间显著性差异图

说明：图中的每一个点都代表了一个比较的OTU。纵坐标（y）W值越高代表该OTU在组间的差异显著性越高。注意，ANCOM分析是基于假设只有少于25%的OTU在不同的组间会发生变化而大部分的OTU会相对稳定的分布（可应用于大多数扩增子分析的分组设定）。

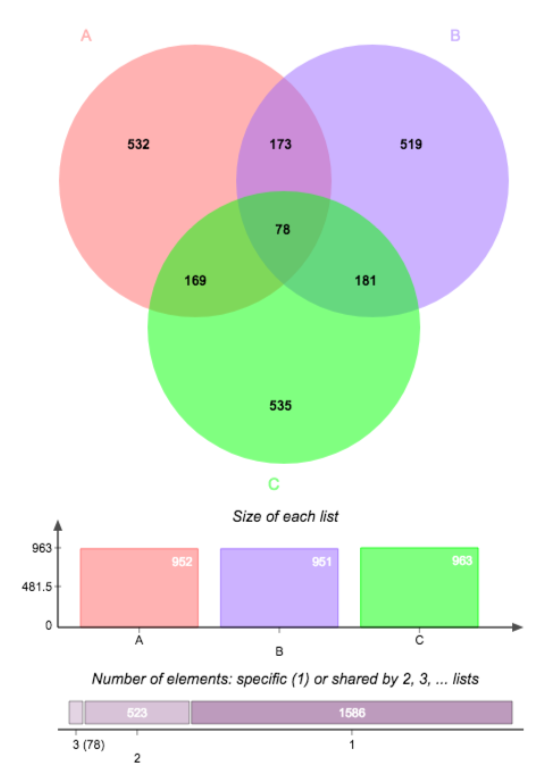
表3-5 OTU种水平的组间显著性差异详细数据（截图）



说明：为图3-12的详细数据说明表格。

### 5.2样品共有OTU分析

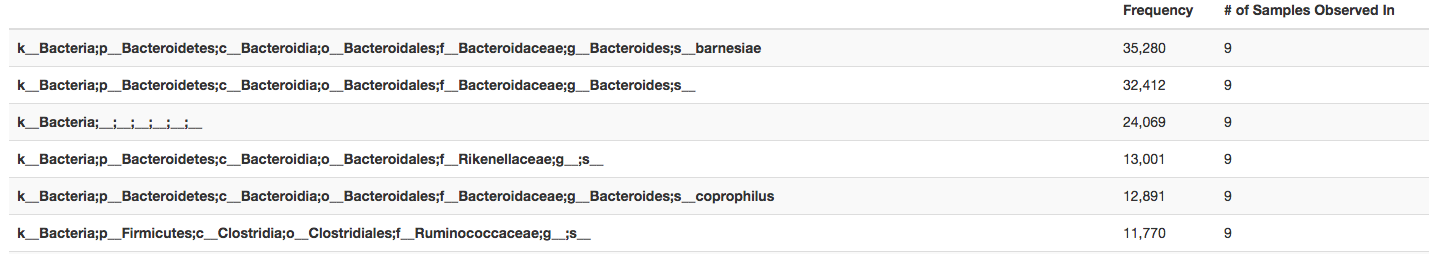
对于样本间，根据OTU是否存在来寻找样本之间的特有或共有的OTU，并绘制韦恩图（Venn Graph），样本数量要求2-5个。对于样本组间，根据OTU差异显著分析结果及其相对丰度，分析不同样品组之间特有或共有的OTU，每个样本组显著富集的OTU为其特有的OTU，不存在差异且相对丰度大于0的OTU为其共有OTU，并绘制韦恩图及三角图来展示结果（图3-13）。查看所有OTU在各个样本间的分布情况，详细数据如表3-6所示。



**图3-13共有或特有物种Venn图展示**

说明：最上方venn图显示的是不同样本（组）间共有或特有的OTU数量,每个圈即代表一个分组。中间柱形图显示不同样本（组）间的OTU总数量。底部条形图显示在不同样本（组）间的共享分布情况。

表3-6共享有OTU在样品中的分布（截图）



说明：表格中三列分别代表OTU注释名字，绝对丰度数量和拥有该OTU的样品数量。

结果目录：组间物种差异显著统计分析结果文件为Result/5-OTUComparision的相关文件，文件详细解释见README文件。

## 6 样品微生物组成功能预测分析

### 6.1预测微生物群落功能

PICRUSt的原理是基于已测细菌[基因组](http://www.oebiotech.com/RNA_seq.html)的16S rRNA序列，推断它们的共同祖先的基因功能谱，对Greengenes数据库中其它未测物种的基因功能谱进行推断，构建古菌和细菌域全谱系的基因功能预测谱，最后，将[测序](http://www.oebiotech.com/Projects/ngs.html)得到的菌群组成映射到数据库中，对菌群代谢功能进行预测。KEGG功能根据注释深度可以分为三级，表3-7展示的是产生的L1级别功能预测组成的柱形图 。

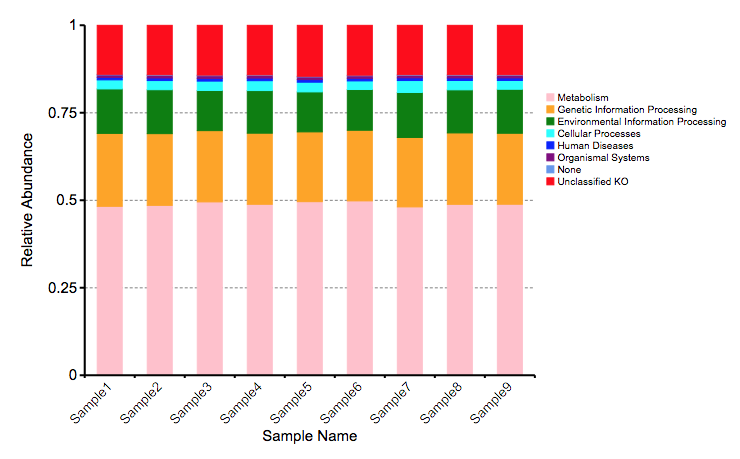


图3-14 KEGG L1微生物群落功能预测组成柱形图

说明：从上至下依次为KEGG结果展示，纵轴表示注释到某类型的功能的相对比例；横轴表示样品名称；各颜色区块对应的功能类别见右侧图例。

### 6.2预测功能PCA分析

基于已知物种的16S物种信息和KEGG功能信息，可以获得环境样本的KEGG功能种类及其相对丰度信息。基于KEGG功能及其代谢途径相对丰度，可以进行样本间KEGG代谢途径PCA分析，图3-32是基于KEGG KO功能级别的功能PCA图。

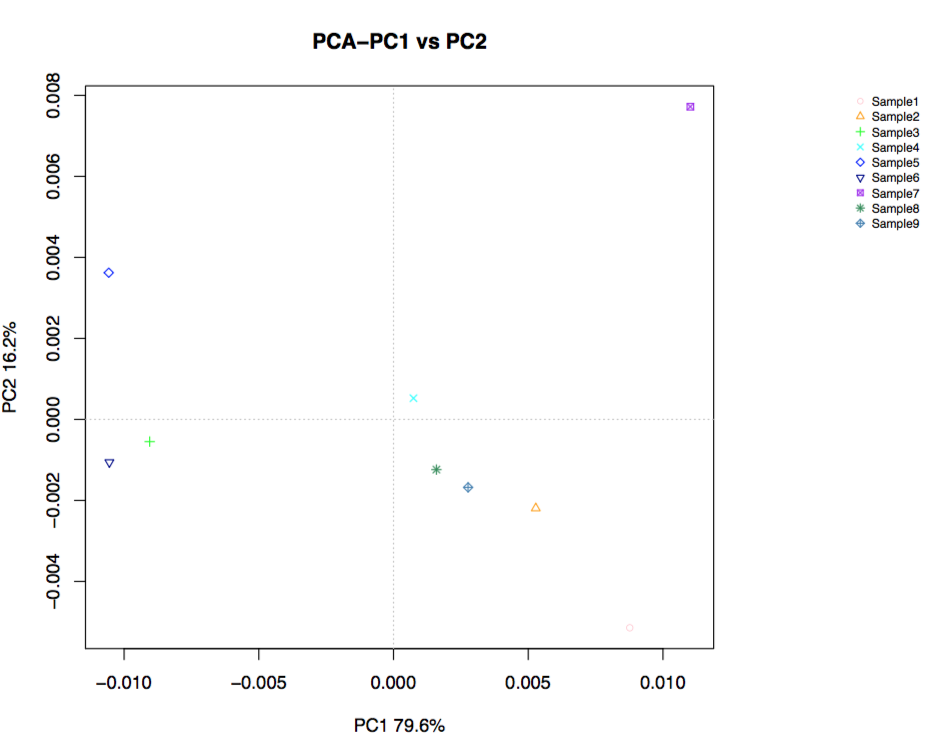


图3-15基于KEGG L1水平的PCA分析结果展示

说明：横坐标（x）表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标（y）表示第二主成分，百分比表示第二主成分对样品差异的贡献值；图中的每个点 表示一个样品，同一个组的样品使用同一种颜色表示。

### 6.3预测功能聚类分析

为了研究不同样品间的相似性，还可以通过对样品进行聚类分析，构建样品的聚类树。Bray-Curtis距离（http://en.wikipedia.org/wiki/Bray-Curtis\_dissimilarity）是系统聚类法中使用最普遍的一个距离指标，它主要用来刻画样品间的相近程度，它的大小是进行样品分类的主要依据。从基因在各样品中的丰度表出发，以Bray-Curtis矩阵进行样品间聚类分析（图3-16）。

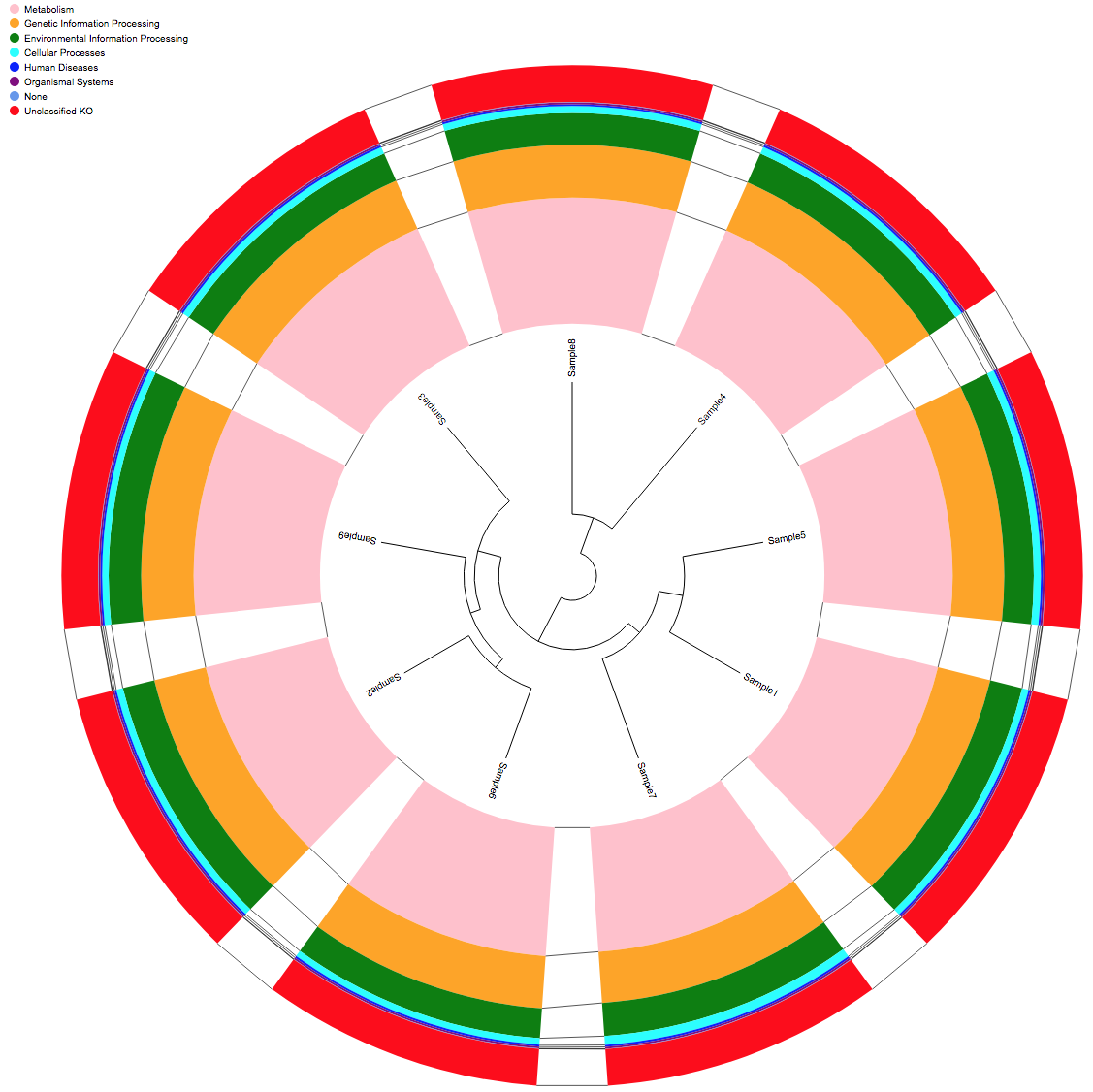
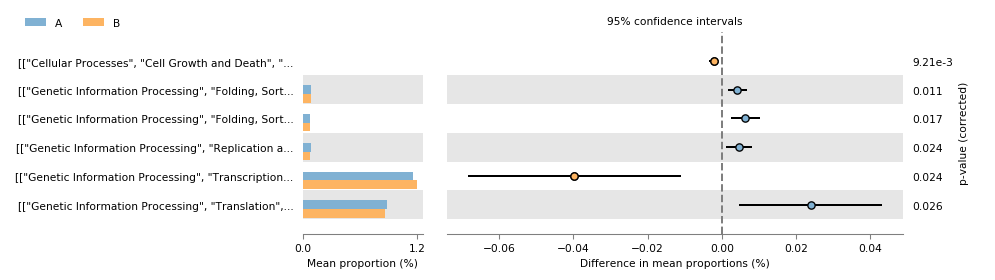


图3-16基于Bray-Curtis距离的第一层级聚类树

说明：中心是Bray-Curtis距离聚类树结构，外圈各层是各样品在第一层级上的功能相对丰度分布，各颜色区块对应的功能类别见左上角图例。

### 6.4组间功能差异性分析

在得到KEGG功能注释后，结合分组信息，STAMP软件在这里被用于分析在微生物群落在不同组间的功能是否有显著性差异。如图3-14所示，我们可以得知那些潜在的微生物群落功能在不同的组间是否有显著性差异，从而可以用于判断meta data或者不同分组对微生物群落的影响。截图所示是该比较组间在没有多重检验矫正P-value之前的结果，只是为了展示。推荐用户自己使用STAMP软件读入已配套生成的结果文件来进行个性化分析。

**图3-17 STAMP软件比较功能差异显著性(截图)**

说明：图标左边为组间有显著性差异得功能名称。中间的横向柱状图表示他们的相对丰度比例。右边的图水平轴表示平均相对丰度差异，纵向轴表示P-value。几个主要组间比较截图已经保存结果文件夹里。建议用户需要STAMP软件来对数据进行自定义分析。输入文件已经根据STAMP需要修改成可读的格式，具体请见README文件。

结果目录：组间物种功能差异显著统计分析结果文件为Result/6-FunctionAnalysis目录下的相关文件，文件详细解释见README文件。

## 7 关联统计分析

### 7.1 CCA/RDA分析

CCA/RDA分析是基于对应分析发展的一种排序方法，将对应分析与多元回归分析相结合，每一步计算均与环境因子进行回归，又称多元直接梯度分析。RDA是基于线性模型，CCA是基于单峰模型（图3-18）。该分析主要用来反映菌群与环境因子之间的关系，可以检测环境因子、样品、菌群三者之间的关系或者两两之间的关系，可得到影响样品分布的重要环境驱动因子。需要提供环境因子的数据，比如 pH值、温度值、临床因子等。

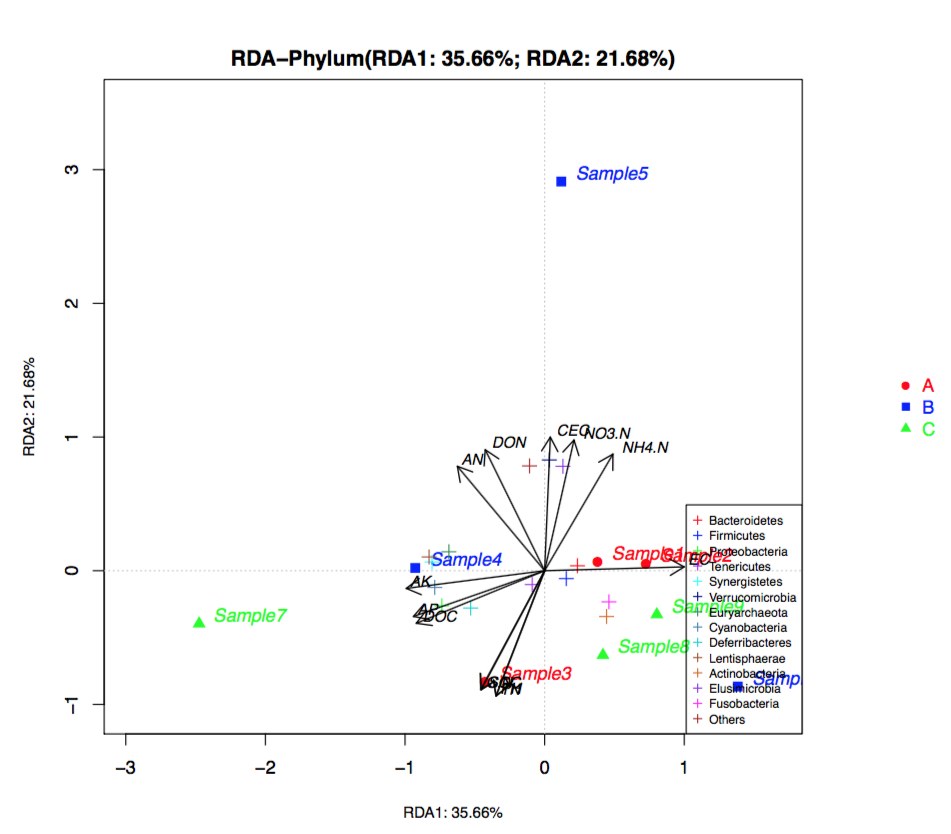


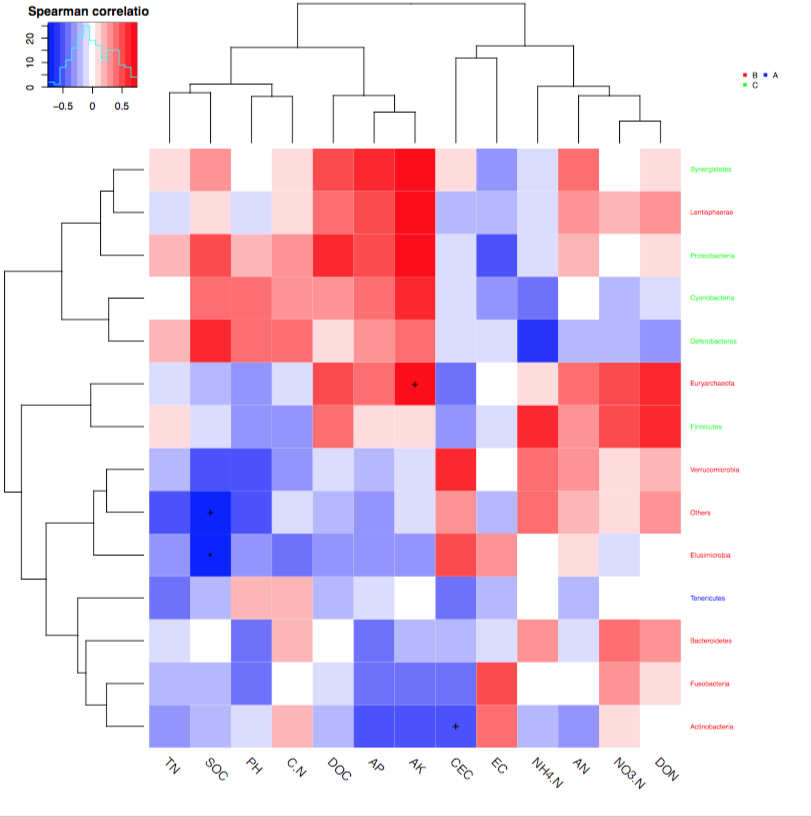
图3-18 CCA/RDA分析图

说明：在CCA/RDA排序图内，环境因子一般用箭头表示，箭头连线的长度代表某个环境因子与群落分布和种类分布间相关程度的大小，箭头越长，说明相关性越大，反之越小。箭头连线和排序轴的夹角代表某个环境因子与排序轴的相关性大小，夹角越小，相关性越高；反之越低。环境因子之间的夹角为锐角时表示两个环境因子之间呈正相关关系，钝角时呈负相关关系。

结果目录：CCA/RDA分析结果文件为/7-AssociationAnalysis/目录下的相关文件，文件详细解释见README文件。

### 7.2 PERMANOVA分析

PERMANOVA (Permutational multivariate analysis of variance)，可以用于分析环境因子或其他临床表型数据与微生物群落或物种之间是否显著相关，然后计算环境因子与微生物物种间的spearman相关系数，并用热图展示（图3-19）。该分析可以挑选出与某种环境因子或疾病显著相关的物种，要求每个样本组的样本数量不小于10个，同时提供环境因子的数据，比如 pH值、温度值、临床因子等。



**图3-19 微生物物种与表型之间的相互关系热图**

说明：横排代表的是物种，竖排代表的样本的表型信息，图中蓝色代表负相关，红色代表正相关，颜色越深代表相关性越高，图中的+表示p<0.05，\*表示p<0.01。右侧每一横排的是物种的ID，红色表示在某一个组中富集的物种，绿色代表再另外一个组中富集的物种。

结果目录：PERMANOVA分析结果文件为7-AssociationAnalysis/目录下的相关文件，文件详细解释见README文件。

### 7.3 物种相互作用network分析

基于物种样本间的相对丰度进行Spearman相关系数计算，获得样本内或样本组内的物种之间的相互关系，然后用展示物种间相互关系的可视化软件进行物种相互作用 network构建（图3-20）。该分析可以寻找相互拮抗或协同的物种，从而寻找环境样本中的微生物群体协同或相互抑制信息，例如针对某种疾病相关的微生物群落，就可以发现共同作用的微生物或相互抑制的微生物，为其治疗做指导。要求每个样本组的样本数量不少于30个。

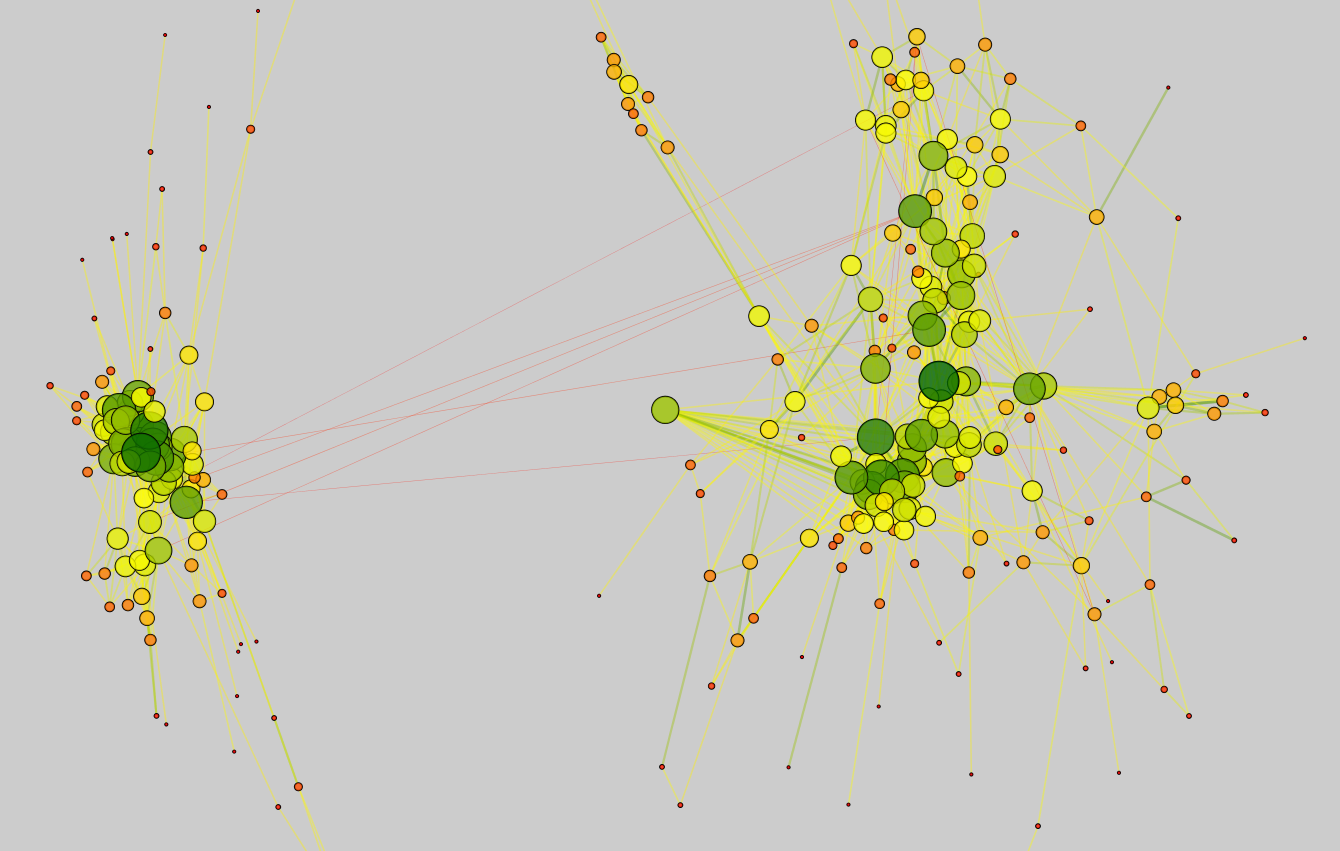


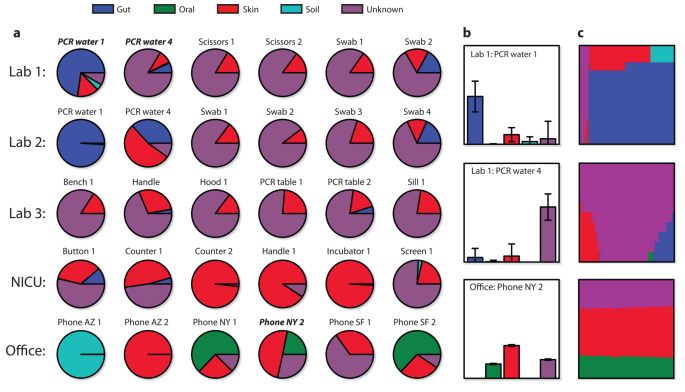
图3-20 物种相互作用network 图

说明： 样本组各自富集的物种的相互作用network图的比较，圆圈代表一个物种，大小代表其相对丰度，不同的颜色代表不同的物种分类，圆圈之间的线条代表这两个物种间的相关性，线条的颜色代表相关系数的大小。

结果目录：物种相互作用network分析结果文件为7-AssociationAnalysis/目录下的相关文件，文件详细解释见README文件。

### 7.4 Source tracker分析

使用基于贝叶斯算法的Source tracker分析，可以寻找样本中的微生物可能来源及其在样本中分布的比例，该方法适用于寻找环境样本或微生物感染的临床样本的微生物物种来源，需要采集可能会影响样本微生物组成的相关联的环境样本（图3-21）。



**图3-21 样本微生物来源分析**

说明：图中从上到下依次展示了样本的肠道、口腔、皮肤、土壤等微生物源

结果目录：Source tracker分析结果文件为7-AssociationAnalysis/目录下的相关文件，文件详细解释见README文件。

## 四 分析方法及参考文献

**1 下机数据处理**

**1.1分析方法**

采用Illumina MiSeq/HiSeq测序平台得到的下机数据(Raw Data) 存在一定的低质量数据，会干扰分析的结果，因此在进一步分析前，需要对下机数据进行预处理，具体处理步骤如下：

1）数据拆分：使用Perl脚本，根据Barcode序列将下机数据拆分为不同样品数据，并截去Barcode序列和PCR扩增引物序列；

2）PE Reads拼接：过滤read尾部质量值20以下的碱基，设置50bp的窗口，如果窗口内的平均质量值低于20，从窗口开始截去后端碱基，过滤质控后50bp以下的read；根据PE reads之间的overlap关系，将成对reads拼接(merge)成一条序列，最小overlap长度为10bp；拼接序列的overlap区允许的最大错配比率为0.2，筛选不符合序列；根据序列首尾两端的barcode和引物区分样品，并调整序列方向，barcode允许的错配数为0，最大引物错配数为2；

3）Tags去嵌合体序列：经过以上处理后得到的Tags序列与数据库(Gold database, <http://drive5.com/uchime/uchime_download.html>)进行比对(UCHIME Algorithm，<http://www.drive5.com/usearch/manual/uchime_algo.html>)检测[嵌合体序列](http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-Chimeras.cgi)，并最终去除其中的嵌合体序列，得到最终的有效数据(Effective Tags)。

**1.2参考文献**

1. Bokulich, N. A. et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods 10, 57–59 (2012).

2. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–6 (2010).

3. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–63 (2011).

4. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194–200 (2011).

5. Haas, B. J. et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 21, 494–504 (2011).

**2 OTU及物种群落分析**

**2.1 分析方法**

1）OTU聚类：使用Qiime2软件中的DADA2插件对所有样品的全部Effective Tags 序列进行质量控制并97%一致性(Identity)聚类，形成OTU；

2）OTU注释：选取OTU代表序列，使用Qiime2软件中的feature-classifier插件对Greengenes数据库(99%， 进行比对获得物种注释信息；

3）OTU聚类及物种注释统计：使用Perl脚本进行每个样品的Effective Tags数据，低频数的Tags数据、Tags注释数据、注释为叶绿体及线粒体数据及每个样本获得的OTU数量等信息进行统计；同时使用R软件进行OTU各个分类等级的注释比例和各个分类等级物种相对丰度的统计；

4）OTU进化分析：a）用PhyloSeq读取Qiime2生成的BIOM文件，OTU代表序列和有根树，结合分组信息对所关注样本的物种注释结果进行可视化，所用的R脚本也被保存在结果文件夹，用户如果需要，可以根据自己的需求来自定义需要的分析。b）用Qiime2对样品中主要的OTU进行系统进化树构建，然后使用Perl脚本，生成iTOL可读的注视文件，并在iTOL网页展示主要的OTU的进化关系和分布情况。

5）物种群落结构分析：使用Qiime2和R软件进行OTU各个分类水平的相对丰度热图绘制，同时进行样本间和物种间的聚类分析。

**2.2参考文献**

1. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–7 (2007).

2. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069–72 (2006).

3. Cole, J. R. et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141-5 (2009).

4. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590-6 (2013).

5. Kõljalg, U. et al. Towards a unified paradigm for sequence-based identification of fungi. Mol. Ecol. 22, 5271–7 (2013).

6. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217 (2013).

7. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39, W475-8 (2011).

8. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–3 (2016).

**3** **Alpha Diversity分析**

**3.1分析方法**

1）Alpha Diversity指数统计分析：使用QIIME2 diversity core-metrics-phylogenetic插件对OTU丰度表进行四种多样性指数（Observed OTUs, Shannon, PD whole tree和Evenness）的计算，并用QIIME2 diversity alpha-rarefaction插件进行四种多样性指数的稀释曲线数据计算及构图。

2）组间比较Alpha Diversity指数统计：使用diversity alpha-group-significance插件对感兴趣的分组进行比较分析。

**3.2参考文献**

1. Wang, Y. et al. Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. Appl. Environ. Microbiol. 78, 8264–71 (2012).

2. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–14 (2012).

3. Tuomisto, H. A consistent terminology for quantifying species diversity? Yes, it does exist. Oecologia 164, 853–60 (2010).

**4 Beta Diversity分析**

**4.1分析方法**

1）Beta Diversity指数统计分析：使用QIIME2 diversity core-metrics-phylogenetic插件对OTU丰度表进行四种距离指数（Jaccard distance, Bray-Curtis distance, unweighted UniFrac distance和weighted UniFrac distance）的计算。

2）组间比较Beta Diversity指数统计：使用Qiime2 emperor插件对感兴趣的分组进行比较分析。

**4.2参考文献**

1. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71, 8228–35 (2005).

2. Chen, J. et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28, 2106–13 (2012).

3. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2, 16 (2013).

4. Tuomisto, H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography (Cop.). 33, 2–22 (2010).

5. Adams, D. C. & Collyer, M. L. Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: what you shuffle matters. Evolution 69, 823–9 (2015).

6. Anderson, M. J. & Walsh, D. C. I. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecol. Monogr. 83, 557–574 (2013).

7. Kelly, B. J. et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics 31, 2461–8 (2015).

**5** **OTU水平样品组间统计分析**

**5.1分析方法**

1）组间群落结构差异显著性分析：基于均一化的OTU丰度表，使用Qiime2 composition ancom插件进行ANCOM分析显著性差异OTU；

2）样品共有OTU分析：基于均一化的OTU丰度表，使用Qiime2 feature-table summarize进行共有OTU分析，用MetaCoMET来对分组间共有OTU绘制Venn图。

**5.2参考文献**

1. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Health Dis. 26, 27663 (2015).

2. White, J. R., Nagarajan, N. & Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. 5, e1000352 (2009).

3. Segata, N. et al. Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60 (2011).

4. Wang, Y. et al. MetaCoMET: A web platform for discovery and visualization of the core microbiome. Bioinformatics. 32, btw507 (2016).

**6样品微生物组成功能预测分析**

**6.1预测微生物群落功能**

1）预测微生物群落功能：用Qiime2 vsearch插件获得close reference OTU之后，用PICRUSt所带的脚本normalize\_by\_copy\_number.py标准化所得的OTU表格，并用predict\_metagenomes.py根据KEGG数据库预测宏基因组功能，然后再用categorize\_by\_function.py对多预测功能按照级别分类汇总，生成L1，L2和L3三个结果文件。

2）组间功能差异性分析：使用microbiome\_helper功能包的biom\_to\_stamp.py脚本将所得的BIOM文件修改成STAMP接受的文件格式(.spf)。用STAMP读入.spf和分组信息文件来对数据进行组件功能差异性分析。

**6.2参考文献**

1. Langille, M. G. I. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31, 814–821 (2013).

2. Comeau, A. M., Douglas, G. M. & Langille, M. G. I. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems 2, e00127-16 (2017).

3. Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 30, 3123–4 (2014).

4. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584 (2016).

**7关联统计分析 （需要环境因子数据）**

**7.1分析方法**

1）CCA/RDA 分析：基于均一化的各个物种等级的丰度表，以及环境因子数据，首先使用R软件进行去趋势对应分析，即DCA(Detrended correspondence analysis)分析，根据梯度值确定线性模型（RDA）和单峰模型(CCA)哪个最合适（DCA分析结果中Axislength的前4个轴中最大的值如果大于4.0，应该选CCA，如果3.0-4.0之间，选RDA和CCA均可，如果小于3.0，RDA的结果要好于CCA），然后使用进行cca或rda分析并绘图；

2）PERMANOVA相关性分析：基于均一化的OTU丰度表，以及环境因子数据，使用R软件进行群落结构与各个环境因子之间的关联分析并绘制热图展示。

3）物种相互作用network分析：基于物种样本间的相对丰度,使用R软件进行相关系数计算，得出样本内或样本组内的物种之间的相互关系，然后使用cytoscape (<http://cytoscape.org/>)进行物种间相互作用 network关系构建及展示，要求每组样本至少10个样本；

4）Source tracker分析：基于均一化的OTU丰度表，使用sourcetracker软件寻找样本中的微生物可能来源及其在样本中分布的比例，该方法适用于寻找环境样本或微生物感染的临床样本的微生物物种来源，需要采集可能会影响样本微生物组成的其他样本；

**7.2参考文献**

1. Hill M, Gauch H. (1980) Detrended correspondence analysis: an improved ordination technique. Vegetatio. 42, 47-58.
2. Oksanen J, Minchin P. (1997) Instability of ordination results under changes in input data order: explanations and remedies. Journal of Vegetation Science. 8, 447-454.
3. Legendre P, Legendre L. (2012) Numerical Ecology. 3rd English ed. Elsevier.
4. McCune B. (1997) Influence of noisy environmental data on canonical correspondence analysis. Ecology. 78, 2617-2623.
5. Palmer M. (1993) Putting things in even better order: The advantages of canonical correspondence analysis. Ecology. 74,2215-2230.
6. TerBraak C. (1986) Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology. 67, 1167-1179.
7. Clarke K. (1993) Non-parametric multivariate analysis of changes in community structure. Australian Journal of Ecology. 18, 117-143.
8. Warton D, Wright T,Wang Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution. 3, 89-101.
9. Zapala M, Schork N. (2006) Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences. 103:19430-19435.
10. Qin J, Li Y, Cai Z, et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 490: 55-60.
11. Feng et al. (2015) Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nature communications 6: 6528.

## 五 交付结果目录结构

交付数据分为若干个子目录，分别是拆分的原始下机数据、过滤后数据、OTU分析结果、Alpha多样性分析结果、Beta多样性分析结果、样本组间比较分析、统计关联分析和功能预测分析。若样品数目少于3，则不做Beta多样性分析及后续统计比较分析。详情可查看交付目录下的README.txt。

|-- 0-RawData/ [拆分后的原始下机数据拆分结果]

|-- 1-Demux / [过滤后数据]

|-- 2-OTUAnalysis/ [OTU分析结果]

|-- 3-AlphaDiversity/ [Alpha多样性分析结果]

|-- 4-BetaDiversity/ [Beta多样性分析结果]

|-- 5-OTUComparision/ [样本组间OTU比较分析]

|-- 6-FunctionAnalysis/ [样本功能比较分析]

|-- 7-AssociationAnalysis/ [关联统计分析]

|-- FigureandTable/ [作图所需文件]

## 六 联系我们

地址：深圳市盐田区CBD壹海中心一区1栋702

邮编：518081

电话：0755-66814606

400-863-2010

邮箱：技术支持[support@tgs.org.cn](mailto:support@tgs.org.cn)

人力资源[career@tgs.org.cn](mailto:career@tgs.org.cn)

市场合作[marketing@tgs.org.cn](mailto:marketing@tgs.org.cn)

网址：<http://www.tgs.org.cn>