扫描微信二维码

生工生物工程(上海)股份有限公司
Sangon Biotech (Shanghai) Co., Ltd.

生工® Sangon Biotech

地　址：上海市松江区香闵路 698 号
热　线：400-821-0268
总　机：021-57072168
传　真：021-57072170
邮　箱：sales@sangon.com
网　址：http://www.sangon.com

生工® Sangon Biotech

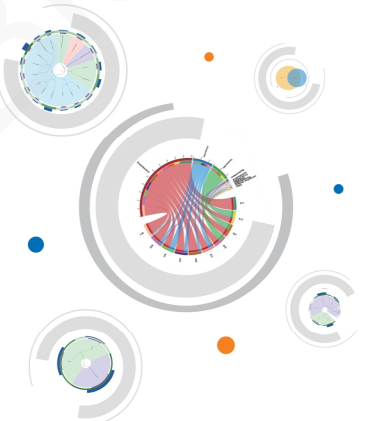
生工生物

宏基因组微生物分类测序

项目介绍

MICROBIAL CLASSIFICATION SEQUENCING BY 16S/ 18S/ ITS

▶ 第四版（V4.0）

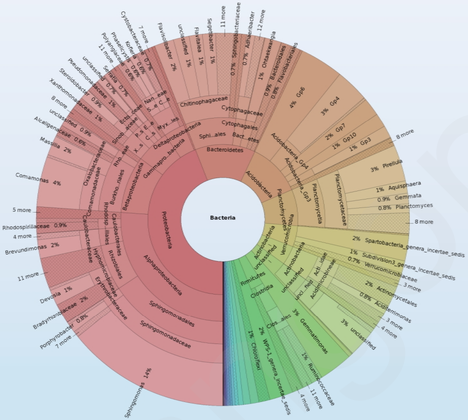


前言

微生物分类测序是生工生物工程（上海）股份有限公司高通量测序部的主营业务之一，仅在2019年，我们为多达2000位研究环境微生物的客户提供了该服务。随着技术的发展，我们也紧跟时代的步伐，不断更新着该服务的分析清单。本手册即是在这个背景下重新编写的。

本手册主要会给大家分享微生物分类测序的知识，涵盖了实验部分、分析部分、参考文献、写作模板等内容。帮助读者深入理解我们这项金牌服务的各项信息，助力您的科研工作快速取得成果。

本册内容如有不足之处，烦请读者海涵。欢迎读者朋友指正留言，联系邮箱：ngs@sangon.com。



目录 CONTENTS

目录 CONTENTS

生工生物 宏基因组微生物分类测序 项目介绍

目录 CONTENTS

www.sangon.com

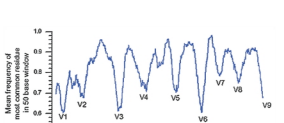
项目介绍

就测序目的而言，微生物分类测序的研究对象是特定的环境样品。我们要探究特定环境样品内的菌群多样性信息、不同样品间菌群丰度差异以及环境因子与环境内微生物多样性之间的关系、预测环境内功能基因的大致情况等。

那我们有什么办法能快速达成上述研究目标呢？

我们可以想象一下，如果环境样品内的每个微生物都有一张“身份证”，我们通过特定的手段，收集到它们的“身份证”，进行统一的查验、登记，然后再统计分析，这样就可以快速准确的达成上述目标。

以环境内的原核细菌为例，它还真有一张“身份证”，这张“身份证”就是所有的原核细菌都有的 16s rRNA 基因，该基因常用于做微生物的菌种鉴定。之所以会选择这个基因测序来鉴定菌种，一是由于它长度合适，约 1500 bp，二是因为该基因在不同种细菌内序列不一样，即在一定范围内存在若干个可变区（评见下图）。这样，这个基因序列就成为了一个合适的“身份证”。



上图横坐标为 16s rRNA 基因的 5’ -3’ 方向位置，纵坐标为序列的变异频率。纵坐标值越大，表明越保守，反之，则变异频率越高。根据变异频率不同，该基因分为 9 个可变区，编号 V1-V9 区。

那么，我们该用什么办法来收集这个“身份证”呢？其实办法很简单，就是设计通用引物（Universel Primer），以宏基因组 DNA 为模板，进行 PCR 扩增。此时的扩增产物就是一个“身份证”的集合。

根据上图可知，在相对保守的位置，可以设计一系列的通用引物（Universel Primer），用于扩增 16s rRNA 基因全长或者部分可变区。同时，我们认为，变异频率越高的区域，其序列组成多样性肯定高于变异频率越低的区域。为了测序结果尽可能比对更准确，我们应该选择变异频率更高的区域作为靶区域。

另外，这里额外说一下 illumina 旗下 2 个测序机器的参数。

测序机器	MiSeq	HiSeq2500
单次运行产出最大数据量	25 M	400 M
最大读长	2 × 300 bp	2 × 250 bp

备注：数据量单位 M 指 million Reads，意思是百万序列。

由于 illumina 平台最多能测 500-600 bp 的长度，去除一下低质量区域，保留一部分 overlap 重叠区，因此我们选择的扩增产物长度最好控制在 480 bp 以内，极限长度不超过 500 bp。

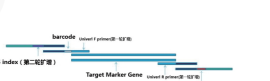
结合 16s 基因的可变区情况，比较恰当的区域应该包括 V3-V4，V6-V8，单独的 V3 区，单独的 V6 区等。

其它，研究真菌的标记基因有 18S 或者 ITS 区。还有一些特定的功能基因序列，例如 AmoA, nirH 等。

建库方法介绍

众所周知，illumina 平台高通量测序是要先连测序接头才可以测序的（这个过程称为建库）。构建文库的目的有 2 个。其一是为了使得不同的待测序 DNA 分子拥有相同的测序引物，方便机器识别起始测序位置等；其二，是为了区别不同用户的不同样品的数据。为了达到上述 2 个目标，接头上分别设计有 index 和 barcode 序列的，通过不同 index 及 barcode 的组合，可以同时对成百上千个不同的样品进行测序。

上面一节我们也讨论到，本项目的目标 DNA 其实是通用引物扩增得到的。同时，为了满足高通量测序建库的要求，我们优化并设计了 2 轮 PCR 的方法，将建库过程融入在 PCR 的过程中，这样可以高效快速的完成大量样品的检测。其基本原理图如下：



上述扩增产物经过纯化之后，即可上机测序。

实验流程

1 样品预处理

土壤和粪便等固体样品：

称取 200 mg-500 mg 混匀后的样品，放入灭菌的 2 mL 离心管中，加入 1X PBS 溶液，震荡混匀，10000 rpm 室温离心 3 min，弃置上层液体。倒置 2 mL 管于吸水纸上 1 min，直至没有液体流出。将样品管放入 55°C 金属浴 10 min，使残留液体完全挥发，保证后续实验操作。

污水和菌液等液体样品：

由于水样中单位体积的样品含有的细菌和真菌的数量比较少，提取之前必须先富集样品中的菌量。混匀原始样品，样品足够时，取 4 mL 液体样品，分多次加入灭菌的 2 mL 离心管中，10000 rpm 室温离心 3 min，弃置上层液体，倒置 2 mL 管于吸水纸上 1 min，直至没有液体流出。样品不足 4 mL 就全部离心。

植物组织样品和动物组织等难破碎的样品：

植物和动物样品需要提取的一般都是内部的内生菌群。直接提取菌群，细胞壁难以破裂，不易提取，需要把内部的菌群暴露出来。一般采取液氮研磨的方法将样品磨成粉末状，便于提取；较为柔软的、易破碎的样品用高温灭菌的剪刀剪碎后，放入 2 mL 离心管中，加入适量钢珠和裂解液，在破碎仪上进行破碎后进行提取。

污泥：

可以直接取适量样品放入灭菌的 2 mL 管中进行提取，含水较多的可以用适当的转速离心后，留取沉淀进行提取。

载体样品：

此类样品是客户已经富集菌群的样品，例如滤膜、碳毡、填料等块状、片状固体样品，为了使样品表面及内部的菌群能更好的与裂解液接触，使更多的细胞壁破裂，释放 DNA。操作过程中取适量样品，经液氮冷冻后，在组织破碎仪中进行破碎。液氮冷冻后难以破碎的样品使用高温灭菌的剪刀将其剪碎后加入裂解液，再进行常温破碎；另外活性炭、（沙）石子、碳刷等颗粒状、粉末状、丝状的载体样品可取适量样品，或者直接在样品管中加入适量的 1X PBS 溶液或者利用样品中原有的液体，强烈震荡，使样品表面菌体能够被洗到液体中，取适量洗过之后的液体 12000 rpm 离心 2min，留取沉淀进行提取。

食品：

为了提取该样品中的菌群，如：固体样品，使用高温灭菌的剪刀或刀片将其剪碎后混合均匀，取适量样品（不超过 0.5 g）放入 2 mL 样品管，在组织破碎仪中进行破碎后提取。液体样品：混合均匀后取适量样品离心留取适量沉淀后进行提取（一般不超过 0.5 g）。粉末、颗粒状样品：混合均匀后直接取不超过 0.5 g 的样品进行提取，如颗粒较大，在提取前做破碎处理。

菌体：

固体琼脂菌落：使用无菌枪头刮取表面菌落，转移适量样品至 2 mL 样品管中进行提取。液体菌群、菌液，混合均匀后吸取适量样品，12000 rpm 离心 2 min，留取沉淀的菌落进行提取。干燥的菌粉，菌剂等样品，可直接取样至 2 mL 样品管中进行提取。（此类样品取样量新鲜菌不超过 0.25 g，干燥粉末状的不超过 0.5 g）

内外生菌：

内生菌：将样品用 70% 乙醇溶液进行表面冲洗 3 遍，再使用 1X PBS 溶液冲洗 3 遍后，除干样品，使用液氮研磨或组织破碎仪破碎后进行提取。部分植物样品如土豆，萝卜等，可使用高温灭菌的刀片削去表面组织，直接取内部组织进行液氮研磨或破碎处理后进行提取。

外生菌：将样品放入适当的容器中如 6 mL- 60 mL 样品管，加入适量 1X PBS 溶液进行表面震荡冲洗，使表面菌群冲洗到 PBS 中，再收集 PBS 冲洗液，12000 rpm 离心 2 min 后小心去除冲洗液，留取沉淀进行提取。

备注：当客户样品达不到取样量时，则全部提取。

2 提取DNA

2.1 基因组提取

具体提取步骤参照 OMEGA 试剂盒 E.Z.N.A™ Mag-Bind Soil DNA Kit 的试剂盒使用说明书。

1)2 mL 离心管中加入 0.8 mLBuffer SLX Mlus, 震荡混匀 5 min。
2) 加入 80 μl Buffer DS 并且震荡混匀。
3) 恒温金属浴 70°C 裂解 10min。
4)13000 rpm 室温离心 5 min。
5) 吸取 600 μl 上层液体到新的 2 mL 离心管中，加入 200 μl Buffer SP2，震荡混匀。
6) 加入 100 μl HTR Reagent，混匀 10 s。冰浴 5 min, 13000 rpm 室温离心 5 min。
7) 吸取 400 μl 上清液体到新的 2 mL 离心管中。
8) 加入 450 μl Binding Buffer 和 40 μl Magsi Particles，震荡混匀，室温静置 2 min。
9)2 mL 离心管放置在磁力架上吸附 5 min，小心吸弃上层清液，并且移开。
10) 加入 500 μl Binding Buffer, 震荡混匀，室温放置 2 min。
11)2 mL 离心管放置在磁力架上吸附 5 min，小心吸弃上层清液，并且移开。
12) 加入 1000 μl Buffer PHB, 震荡混匀磁珠。放置在磁力架 5 min，弃上清。
13) 加入 1000 μl SPM Wash Buffer, 混匀磁珠。放置在磁力架 5 min，弃上清。
14) 重复步骤 13。
15)2 mL 离心管放在 55°C 烘箱 10 min, 使残留酒精完全挥发。
16) 加入 60 μl Elution Buffer 到离心管中，充分震荡混匀，65°C 金属浴 10 min。
17) 磁力架吸附 5 min，小心吸取上清 DNA 溶体到新的 1.5 mL 离心管中。

2.2 琼脂糖凝胶检测DNA完整性

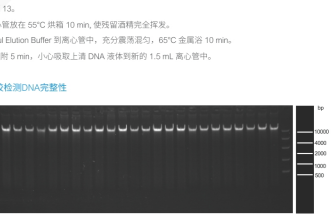


图 1 基因组 DNA 检测示意图

3 PCR扩增

第一轮扩增(以16s v3-v4区为例)

利用 Qubit3.0 DNA 检测试剂盒对基因组 DNA 精确定量，以确定 PCR 反应应加入的 DNA 量。PCR 所用的引物已经融合了测序平台的客户目标序列引物。

341F：ATGCGTAGCCGACCTGAGA

805R：CGTCAGACTTTCGTCCATTGC

体系按照如下进行：

2X Hieff® Robust PCR Master Mix	15 μl
Bar-PCR primer F	1 μl
Primer R	1 μl
PCR products	10-20 ng
H_2O	add to 30 μl

配置好的 PCR 体系按照如下反应条件进行 PCR 扩增：

94°C	3 min	
94°C	30 s	5 cycles
45°C	20 s	
65°C	30 s	
94°C	20 s	20 cycles
55°C	20 s	
72°C	30 s	
72°C	5 min	
10°C	……	

PCR 结束后进行第二轮扩增。

第二轮扩增

第二轮使用第一轮 PCR 产物进行扩增，引入 Illumina 桥式 PCR 兼容引物。

PCR 体系按照如下进行：

2X Hieff® Robust PCR Master Mix	15 μl
primer F	1 μl
Index-PCR-primer R	1 μl
PCR products	20-30 ng
H_2O	add to 30 μl

配置好的 PCR 体系按照如下反应条件进行 PCR 扩增：

94°C	3 min	
94°C	20 s	5 cycles
55°C	20 s	
72°C	30 s	
72°C	5 min	
10°C	……	

PCR 结束后，PCR 产物进行琼脂糖电泳检测，结果如下：

图 2 PCR 检测示意图

4 DNA 纯化回收

对于细菌和古菌扩增的 PCR 产物和正常扩增片段在 400 bp 以上的 PCR 产物，选用 0.6 倍的磁珠处理。对于真菌 PCR 产物和其他扩增片段小于 400 bp 的 PCR 产物，选用 0.8 倍的磁珠处理。

1) 取 25 μl PCR 产物中加入体积 0.6 倍（0.8 倍）的磁珠，震荡充分悬浮后放在磁力架上吸附 5 min，小心的用移液枪吸出上清。

2) 加入 30 μl 0.6 倍（0.8 倍）的磁珠洗涤液，震荡充分悬浮后放在磁力架上吸附 5 min, 小心吸出上清。

3) 加入 90 μl WashBuffer（或者 70% 乙醇），反向放置在磁力架上，使磁珠吸附到 PCR 管的另外一面，充分吸附后吸出上清。重复步骤一遍。

4) 将 PCR 管或 8 联管放在 55°C 金属浴 5 min，使里面的酒精完全挥发。

5) 加入 30 μl Elution Buffer 洗脱。

6) 将 PCR 管放在吸附架上 5 min，充分吸附，移出上清到干净的 1.5 mL 离心管中，定量备用。

定量混合

利用 Qubit3.0 DNA 检测试剂盒对回收的 DNA 精确定量，以方便按照 1:1 的等量混合后测序。

实验关键试剂

试剂 / 耗材
E.Z.N.A™ Mag-Bind Soil DNA Kit
Qubit3.0 DNA 检测试剂盒
2X Hieff® Robust PCR Master Mix
Hieff NGS™ DNA Selection Beads

仪器

仪器
台式高心机
漩涡混合器
混匀型干式恒温器
电泳仪电源
电泳槽
凝胶成像系统
Qubit® 3.0 荧光计
PCR 仪
移液器

数据分析部分

1 术语解释

Bp: base-pair，碱基对，读长的单位，每一个 bp 指一对互补的碱基。

Read: 读长，测序数据中每一条序列就是一个 read。

Raw_reads: 原始数据。

Clean_reads: QC 之后的数据。

Barcode: 标签序列，位于 reads 的开头，用于区分这一条 reads 属于哪一个样本。分配完毕之后 barcode 会被删除。

Fastq: 序列数据存储的标准格式之一，每 4 行为一条 read 的信息。包含测序 read 名，序列，正反链标示，序列质量值。

Fasta: 序列数据存储的标准格式之一，每两行为一条 read 信息。包含测序 reads 名和序列。通常在 QC 之后，以此格式保存数据。

Pair-end 测序： 双端测序，两端均测序，随后合并成一条 read。

质量评分： 指的是一个碱基的错误概率的对数值，即质量评分越高，错误概率越小。

QC: Quality control，即质量控制。

低复杂度序列： 即有大量简单重复的序列。

滑窗法： 检测一个窗口内的碱基质量值，如果满足条件则向前移动一个单位继续检测，如果不满足条件即做删除处理，随后继续移动到下一个单位进行检测，直到检测完所有的数据。

嵌合体： PCR 过程中，因为不同的模板混杂，错误产生的序列，这条序列并非真实存在。

OTU: operational taxonomic unit（操作单元分类）。要了解样品测序中群落分布信息，就需要对序列进行聚类（cluster），通过聚类，就可以根据序列的相似度分成很多序列的集合，每一个序列的集合就是一个 OTU。

RDP: Ribosomal Database Project。为了得到每个 OTU 对应的物种分类信息，采用 RDP classifier 贝叶斯算法对 97% 相似度水平的 OTU 代表序列进行分类学分析，并在界门纲目科属水平，统计各个样品的菌落组成。

Alpha 多样性： 是指一个特定区域或生态系统内的多样性，经常用物种丰富度来度量。

Beta 多样性： 不同生态系统之间多样性的比较，是物种组成沿环境梯度或者在群落间的变化率，用来表示生物种类对环境异质性的反应。

PCA 分析： 在多元统计分析中，主成分分析 PCA (Principal Component Analysis) 是一种简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中对方差贡献最大的特征，从而有效地找出数据中最“主要”的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构。

PCoA 分析： PCoA 分析 (Principal Co-ordinates Analysis) 是一种研究数据相似性和差异性的可视化方法。进过一系列的计算之后，选择主要的，排在前几位的特征值，对样本之间的关系进行描述。

NMDS 分析： 非度量多维尺度分析，是一种将多维空间的研究对象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。其特点是根据样品中包含的物种信息，以点的形式反映在多维空间上，而对不同样品间的差异程度，则是通过点与点的距离体现的，最终获得样品的空间定位点图。

RDA/CCA 分析： 是基于对应分析发展而来的一种排序方法，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回归，又称多元直接梯度分析。

ANOSIM 分析：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。

PERMANOVA 分析：置换多因素方差分析（permutational MANOVA）或非参数多因素方差分析（nonparametric MANOVA），又称 Adonis 分析。它利用平度量（如 Bray-Curtis）或度量距离矩阵（如 Euclidean）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，并使用置换检验对划分的统计学意义进行显著性分析。

PLS-DA 分析：偏最小二乘法判别分析，是多变量数据分析技术中的判别分析法，经常用来处理分类和判别问题。通过对主成分适当的旋转，PLS-DA 可以有效的对组间观察值进行区分，并且能够找到导致组间区别的影响变量。

菌群分型分析：菌群分型分析，主要通过统计聚类的方法研究不同样本优势菌群结构的分型情况，分型过程中一般不考虑环境因子等外部因素的影响。通过该分析，可以将优势菌群结构近似的不同样本聚为一类，主要适用于特定环境样本的菌群分型，如肠型（enterotypes），阴道分型（cervicotype），口腔分型等。

Mantel Test 分析：Mantel test 是检验两个矩阵相关关系的非参数统计方法。Mantel test 多用在生态学上检验群落距离矩阵（比如 UniFrac distance matrix）和环境变量距离矩阵（比如 pH，温度或者地理位置的差异矩阵）之间的相关性（Spearman 等级相关系数等）。Partial Mantel test 在控制矩阵 C 的效应下，来检验 A 矩阵的残留变异是否和 B 矩阵相关。该分析输入两个数值型矩阵，第三个控制矩阵可通过选择因子来确定。

Random Forest 分析：随机森林分析，属于机器学习算法，是一个包含多棵决策树的分类器，它的分类结果根据检测样本的各个维度上的属性，在不同的决策树上进行判定，综合考虑所有判定结果后给出最终分类，对于分类问题结果取概率最大值，回归分析则取概率均值，它可以高效快速挑选出对样本分类最为重要的物种类别（biomarker）。

ROC 曲线：ROC 曲线指受试者工作特征曲线 (receiver operating characteristic curve)，是反映敏感性和特异性连续变量的综合指标，通过构图法揭示敏感性和特异性的相互关系。ROC 曲线将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性为纵坐标、特异性为横坐标绘制成曲线，曲线下面积越大，诊断准确性越高。

Node: 网络图概念，每一个点就是一个 node。

Edge: 网络图概念，在 network 中，两点之间的连线就是 edge。

Degree: 联通性，某节点的联通性表示在网络中直接与该节点相连的节点数目，连通性越高表示该节点在整个网络中重要性越高，连通性非常高的节点也称为 Hub 节点。

Closeness Centrality: 紧密系数，也称为亲密系数，节点与网络中其它节点的距离都很短，即该点是整体的中心，该值越大说明节点越靠近网络的中心位置。

Betweeness Centrality: 介数中心性，该拓扑性质体现了某一个节点在与其他节点连接中所起的作用，该值越大，意味着该节点在保持整个网络紧密连接性中作用越重要。

Degree Centrality: 度中心性，是在网络分析中刻画节点中心性最直接的度量指标，一个节点的度越大就意味着该节点的度中心性越高，该节点在网络中就越重要。

Degree Distribution: 度分布，网络由一些节点和连接它们的边构成。每个节点连接的所有边的数量就是这个节点的度。度分布是对网络中节点度数的总体描述，度分布指的是网络图中节点度数的概率分布或频率分布。

Transitivity: 传递性，同一个节点的两个与之相连节点仍然互相连通的概率，也就是网络中节点形成三角形的概率。

Diameter: 网络直径，网络中两个节点之间的距离为连接两个节点的最短路径上的边数，网络中任意两个节点之间的距离的最大值称为网络的直径。

2 本项目使用软件与数据库

2.1 软件列表

SoftWare Name	Version
Cutadapt[1]	1.18
PEAR[2]	0.9.8
PRINSEQ[3]	0.20.4
Usearch[4,5]	11.0.667
RDP classifier[6]	2.12
NCBI Blast+[7]	1.43.0
Mothur[8]	3.8.31
Muscle[9]	3.8.31
BMGE[11]	1.12
Fasttree[12]	2.1.7
Krona[13]	2.7.1
ETE3[14]	3.1.1
GraPhlAn[15]	1.1.3
STAMP[16]	2.1.3
LefSe[17]	1.1.0
SparCC[18]	1.0.0
PICRUSt[19]	1.1.4
BugBase[20]	0.1.0
FAPROTAX[21]	1.2.1
FUNGuild[22]	1.0
R	3.6.0

R Package	Version
vegan	2.5-6
ggraph	2.0.0
Phyloseq[23]	1.30.0
gplots	3.0.1.1
Ape[24]	5.3
mixOmics[25]	6.10.6
Ade4[26]	1.7-13
clusterSim	0.48-3
VennDiagram[27]	1.6.20
UpSetR[28]	1.4.0
Circlize[29]	0.4.8
ggtern[30]	3.1.0
metagenomeSeq[31]	1.28.0
DESeq2[32]	1.26.0
randomForest	<4.6-14
pROC[33]	1.15.3
corrplot	0.84
dada2[34]	1.14.0

2.2 数据库细菌古菌 16S rRNA 数据库

RDP 数据库：默认数据库，http://rdp.cme.msu.edu/misc/resources.jsp

Silva[35] 数据库：http://www.arb-silva.de/

NCBI 16S 数据库：http://ncbi.nlm.nih.gov/

真菌 18S rRNA 数据库：

Silva 数据库：默认数据库，http://www.arb-silva.de/

NCBI 18S 数据库：http://ncbi.nlm.nih.gov/

真菌 ITS 数据库：

Unite[36] 数据库：默认数据库，http://unite.ut.ee/index.php

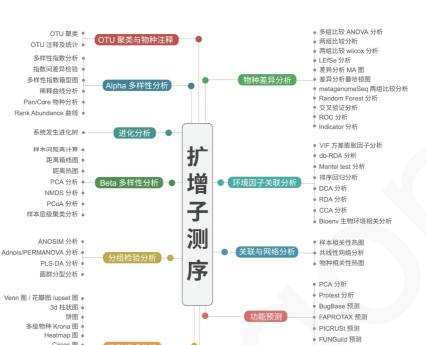
RDP 数据库：http://rdp.cme.msu.edu/misc/resources.jsp

功能基因数据库：

FGR：RDP 整理来源于 GenBank 的功能基因数据库，http://fungene.cme.msu.edu/

3 项目分析流程

illumina 测序得到的双末端序列，首先根据 overlap（重叠区）关系进行拼接，区分样本后对序列质量进行质控和过滤，然后进行 OTU 聚类分析和物种分类学分析。基于 OTU 可以进行多种多样性指数分析，以及对测序深度的检测；基于分类学信息，可以在各个分类水平上进行群落结构的统计分析。在上述分析的基础上，可以对多样本的群落组成和系统发育信息进行 Beta 多样性分析、分组检验分析、差异显著性检验、环境因子关联分析、关联与模型预测分析和功能预测等一系列深入的统计学和可视化分析。



4 分析结果展示

4.1 数据预处理

4.1.1 原始序列数据

Illumina Miseq™/HiSeq™得到的原始图像数据文件经碱基识别 (Base Calling) 分析转化为原始测序序列 (Sequenced Reads)，我们称之为 Raw Data 或 Raw Reads，结果以 FASTQ (简称为 fq) 文件格式存储，其中包含测序序列 (reads) 的序列信息以及其对应的测序质量信息。

FASTQ 格式文件中每个 read 由四行描述，如下所示：

```
@MISEQ03:113:000000000-AFJGE:1:1101:12409:1286 1:N:0:TCTACA
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@IDEEFFCFFHHFFCFFHHHHHHHHFAE0ECFFD/AEHH
```

其中第一行以 "@" 开头，随后为 Illumina 测序标识符 (Sequence Identifiers) 和描述文字 (选择性部分)；

第二行是碱基序列；

第三行以 "+" 开头，随后为 Illumina 测序标识符 (选择性部分)；

第四行是对应碱基的测序质量，该行中每个字符对应的 ASCII 值减去 33，即为对应第二行碱基的测序质量值。

Illumina 测序标识符详细信息详见下表：

MISEQ03	nstrument - unique identifier of the sequencer
113	run number - Run number on instrument
000000000-AFJGE	FlowCell ID - ID of flowcell
1	LaneNumber - positive integer
1101	ileNumber - positive integer
12409	X - x coordinate of the spot. Integer which can be negative
1286	Y - y coordinate of the spot. Integer which can be negative
1	ReadNumber - 1 for single reads; 1 or 2 for paired ends
N	whether it is filtered - NB: Y if the read is filtered out, not in the delivered fastq file, N otherwise
0	ontrol number - 0 when none of the control bits are on, otherwise it is an even number
TCTACA	Illumina index sequences

Miseq™/HiSeq™的碱基测序质量值 (Phred quality score, Qphred) 是测序错误率 (base-calling error probabilities, P) 的整数映射，映射关系为：$Q_{phred} = -10log_{10}(P)$。

Phred quality scores are logarithmically linked to error probabilities		
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

4.1.2 数据预处理

下机测序得到的是双端序列数据，且测序序列中含有 barcode 序列，以及测序时加入的引物和接头序列。首先需要去除引物接头序列，再根据双末端序列之间的 overlap 关系，将成对的 reads 拼接 (merge) 成一条序列，然后按照 barcode 标签序列识别并区分样品得到各样本数据，最后对各样本数据的质量进行质控过滤，得到各样本有效数据。

数据优化方法和参数：

1) 使用 cutadapt 去除 Read1 3' 端测序引物接头 TGGAATTCTCGGGTGCCAAGGAACTC，主要参数：-O 5 -m50；
2) 根据 PE reads 之间的 overlap 关系使用 PEAR 将成对 reads 拼接 (merge) 成一条序列；
3) 根据各样本 barcode 序列和引物序列从拼接后数据中分割出各样本数据，并校正序列方向；
4) 使用 PRINSEQ 切除 reads 尾部质量值 20 以下的碱基，设置 10 bp 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，过滤质控后的含 N 序列和短序列，最终过滤掉低复杂度的序列。

4.1.3 结果说明

结果目录：**1_QC/**

1_RawData: 各样本测序原始序列数据及统计结果。

2_CleanData: 各样本有效序列数据及统计结果。

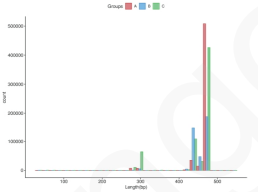


图 4.1.1 原始序列数据长度分布图

所有样本原始序列数据的序列长度和该长度对应的序列条数，不同的颜色代表不同的组别。

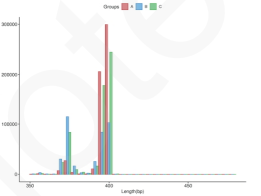


图 4.1.2 有效序列数据长度分布图

所有样本有效序列数据的序列长度和该长度对应的序列条数，不同的颜色代表不同的组别

表 4.1.3 各样本有效序列数据统计

Group	Sample	Barcode	SeqNum	BaseNum	MeanLen	MinLen	MaxLen
A	A11	CTCCTG	54885	21775695	396.75	351	440
A	A12	AATATC	67011	26613812	397.16	351	453
A	A21	TCTAGG	84099	33373637	396.84	350	457
A	A22	ATCGCA	54509	21688980	397.9	351	471
A	A31	GCCATC	68150	26544103	389.5	350	474
A	A32	TGGACG	47875	19011227	397.1	352	440
A	A41	GAAGGC	52883	21026553	397.61	353	447
A	A51	AGTGGC	60124	23850671	396.69	350	471
A	A52	TACCAC	68183	26967020	395.51	350	440
B	B11	GTCGGA	64397	24409400	379.05	350	451
B	B31	TGTGTT	66963	25957087	387.63	350	472
B	B32	CGATGT	67663	26139882	386.32	350	451
B	B41	ATGTCA	52273	20680865	395.63	353	438
B	B42	TTGCTC	68130	26286453	385.83	350	438

从左到右依次为样本的分组信息、样本名、区分样本使用的 barcode 信息、有效序列条数，碱基数，平均长度，最短序列长度，最长序列长度。

4.2 OTU 聚类

4.2.1 分析方法

OTU（Operational Taxonomic Units）是在系统发生学或群体遗传学研究中，为了便于进行分析，人为给某一个分类单元（品系，属，种，分组等）设置的统一标志。要了解一个样本测序结果中的菌种、菌属等数目信息，就需要对序列进行聚类（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。可根据不同的相似度水平，对所有序列进行 OTU 划分，通常对 97% 相似水平下的 OTU 进行生物信息统计分析。

OTU 聚类步骤如下：

1) 对各样本优化序列提取非重复序列，便于降低分析中间过程冗余计算量（http://drive5.com/usearch/ manual/cmd_-fastx_uniques.html）；
2) 所有样本去冗余序列合并后去除没有重复的单序列（http://drive5.com/usearch/manual/singletons.html）；
3) 按照 97% 相似性对非重复序列（不含单序列）进行 OTU 聚类，在聚类过程中去除嵌合体，得到 OTU 的代表序列（http://drive5.com/usearch/manual/cmd_cluster_otus.html）；
4) 将所有优化序列比对至 OTU 代表序列，选出与代表序列相似性在 97% 以上的序列，生成 OTU 表格（http://drive5.com/usearch/manual/pipe_otutab.html）。

软件：Usearch

4.2.2 结果说明

结果目录：**2_OTU_Taxa/OTU**

otu_table.xls: 各样本 **OTU** 中序列数统计表。

表 4.2.1 各样本有效序列数据统计

#OTU ID	A11	A12	A21	A22
OTU13	43282	21	4	0
OTU48	4735	6	4	0
OTU3	1056	5176	10	31
OTU5	1030	58	53	5
OTU9	263	619	380	65
OTU30	98	96	51	5
OTU19	139	161	121	26
OTU17	154	199	115	30
OTU37	78	43	22	2
OTU18	96	185	135	11
OTU372	56	0	0	0
OTU29	47	67	36	14
OTU2	144	101	288	175
OTU32	46	80	29	12

OTU ID 为 OTU 编号，其余列为各样本对应的每一个 OTU 的丰度（序列数）。

4.3 OTU 物种注释及统计

4.3.1 分析方法

为了得到每个 OTU 对应的物种分类信息，需要对 OTU 代表序列根据不同的扩增类型采用相应的数据库进行分类学分析，主要使用以下软件和算法：

RDP classifier:

RDP classifier 基于 Bergey's taxonomy，采用 Naïve Bayesian assignment 算法对每条序列在不同层级水平上计算其分配到此 Level 中的概率值。

SINTAX:

SImple Non-Bayesian TAXonomy，基于非朴素贝叶斯分类算法，使用 Kmer(默认 k=8, 放回式采样 32 kmers) 去计算和参考库共享 Kmer, 确定最佳 Hits, 迭代 100 次，确定每个 Level 的分类（出现次数最多的分类）和可信度（出现的频率）。

Blast:

利用 blastn 将序列与对应数据库进行比对，筛选出序列的最佳比对结果，并对比对结果进行过滤，默认满足相似度 >90% 且 coverage>90% 的序列被用来后续分类，不满足条件的序列则被归为 unclassified。

默认情况下：16S 使用 RDP classifier 比对 RDP 数据库；功能基因使用 Blast 比对 NCBI NT 数据库；18S 使用 Blast 比对 Silva 数据；ITS 使用 Blast 比对 UNITE 数据库。最终分别在各个分类水平(domain(域), phylum(门), class(纲), order(目), family(科), genus(属), species(种) 上统计各样本的群落组成。

4.3.2 结果说明

结果目录：**2_OTU_Taxa/**

Abundance/*_count.xls: 各分类学水平样本序列数统计表。

Abundance/*_abundance.xls: 各分类学水平样本序列数相对丰度百分比统计表。

OTU/otu_table_taxonomy.xls: OTU 分类学综合信息表，将 OTU 分析结果分类学信息结合得到的综合表。

OTU/otu_table.biom: OTU 分类学综合信息表 BIOM 格式文件。biom(Biological Observation Matrix) 格式是生物学样本中观察列联表的一种通用格式，具体信息参考 http://biom-format.org/。

OTU/otu_taxonomy.xls: OTU 分类学信息表。

表 4.3.1 OTU 分类学信息表

#OTU ID	taxonomy
OTU13	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g Vesperl1iibacter;
OTU48	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobac-teriales;f_Enterobacteriaceae;g Serratia;
OTU3	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g Vesperl1iibacter;
OTU5	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_Khizobiales;f_Hyllobacteriaceae;g Mesorhizobium;
OTU9	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;f_Methylobacteriaceae;g Methylobacterium;
OTU30	d_Bacteria;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g Ralstonia;
OTU19	d_Bacteria;p_Proteobacteria;c_Betaproteobacteria;o_Burkholde-riales;f_Comamonadaceae;g Delftia;
OTU17	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_Caulobacterales;f_Ca ulobacterace-ae;g Caulobacter;
OTU37	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomo-nadales;f_Pseudomonadaceae;g Pseudomonas;
OTU18	d_Bacteria;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Oxalobacteraceae;g Undibacterium;
OTU372	d_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g unclassified Bifidobacteriaceae:
OTU29	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;f_Rhizobialesincertaesedis;g_Phreatobacter;
OTU2	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobac-teriales;f_Enterobacteriaceae;g Escherichia Shigella;
OTU32	d_Bacteria;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g Ralstonia;

OTU ID 为 OTU 编号，分类学名称前的单个字母为分类等级的首字母缩写，以 "_" 隔开。分类学数据库中会出现一些分类学谱系中的中间等级没有科学名称，以 norank 作为标记。分类学比对后根据置信度阈值的筛选，会有某些分类谱系低于置信阈值，没有得到分类信息，在统计时以 Unclassified 作为没有分类信息的标记。

4.4 Rank-abundance 曲线

4.4.1 分析方法

Rank-abundance 曲线是分析多样性的一种方式。构建方法是统计单一样品中，每一个 OTU 所含的序列数，OTUs 按丰度（所含有的序列条数）由大到小等级排序，再以 OTU 等级为横轴，以每个 OTU 中所含的序列数（也可用 OTU 中序列数的相对百分含量）为纵轴做图。

Rank-abundance 曲线用于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。物种的丰富程度由曲线在横轴上的长度来反映，曲线越宽，表示物种的组成越丰富; 物种组成的均匀程度由曲线的形状来反映，曲线越平坦，表示物种组成的均匀程度越高。

软件：利用 **R** 制作曲线图。

4.4.2 结果说明

结果目录：**3 AlphaDiversity/RankAbundance/**

rank.xls: OTU Rank 表

rank_abundance.xls: OTU Rank Abundance 表

rank_abundance.pdf: RankAbundance 曲线图。

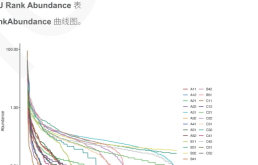


图 4.4.1 RankAbundance 曲线图

横轴为按 OTU 相对丰度含量等级降序排列的 OTU Rank 值，纵轴为样本在这个 OTU 下的丰度数值，每条曲线是一个样本。

4.5 Pan/Core 物种分析

4.5.1 分析方法

Pan and Core OTU 用于描述随着样本量的加大物种总量和核心物种量变化的情况。Pan OTU 指得是泛 OTU，是所有样本所包含的 OTU 的总和，用于观测随着样本数目的增加，总 OTU 数目的增加情况。Core OTU 指的是核心 OTU，是所有样本共有的 OTU 的数目，用于观测随着样本数目的增加，共有 OTU 数目的减少情况。(仅在样品量大于 10 个时分析)

Pan/Core 曲线图反映了持续抽样下新 OTU(新物种) 出现的速率。在一定范围内，随着样品量的加大，若曲线表现为急剧上升则表示群落中有大量物种被发现；当曲线趋于平缓，则表示此环境中的物种并不会随样品量的增加而显著增多。利用物种累积曲线可以作为对样品量是否充分的判断，曲线急剧上升表明样品量不足，需要增加抽样量; 反之，则表明抽样充分，可以进行数据分析。

软件：**R**。

4.5.2 结果说明

结果目录：**3_AlphaDiversity/Accumulation accumulation.xls: Pan/Core** 曲线表。

accumulation.pdf: Pan/Core 曲线图。

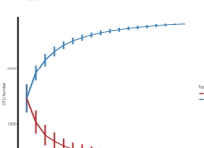


图 4.5.1 Pan/Core 曲线图

横轴表示观测的样本数目，纵轴表示在该数目下随机取样后的共有 / 核心 OTU 数目分布。

4.6 多样性指数分析

4.6.1 分析方法

群落生态学中研究微生物多样性，通过单样品的多样性分析 (Alpha 多样性) 可以反映微生物群落的丰度和多样性，包括一系列统计学分析指数估计环境群落的物种丰度和多样性。同时可以根据分组信息，运用统计学 T 检验的方法，检测每两组之间的指数值是否具有显著性差异。

计算群落分布丰度 (Community richness) 的指数有：

Sobs - the observed richness (http://www.mothur.org/wiki/Sobs)

Chao - the Chao1 estimator (http://www.mothur.org/wiki/Chao)

ACE - the ACE estimator (http://www.mothur.org/wiki/Ace)

计算群落分布多样性 (Community diversity) 的指数有：

Shannon - the Shannon index (http://www.mothur.org/wiki/Shannon)

Simpson - the Simpson index (http://www.mothur.org/wiki/Simpson)

Coverage - the Good's coverage (http://www.mothur.org/wiki/Coverage)

计算群落分布均匀度 (Community evenness) 的指数有：

Shannoneven - the Shannon index-based measure of evenness (Pielou's evenness index，J)

各指数算法如下：

Chao: 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。计算公式如下：

$$S_{chao1} = S_{obs} + \frac{n_1(n_1-1)}{2(n_2+1)}$$

其中，

S_{chao1} = 估计的 OTU 数

S_{obs} = 实际观测到的 OTU 数

n_1 = 只含有一条序列的 OTU 数目 (如 "singletons")

n_2 = 只含有两条序列的 OTU 数目 (如 "doubletons")

Ace: 用来估计群落中 OTU 数目的指数，由 Chao 提出，是生态学中估计物种总数的常用指数之一，与 Chao 1 的算法不同。计算公式如下：

$$S_{ACE} = \begin{cases} S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{n_1}{C_{ACE}}\hat{\gamma}^2_{ACE}, & \text{for } \hat{\gamma}_{ACE} < 0.80 \\ S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{n_1}{C_{ACE}}\hat{\gamma}^2_{ACE}, & \text{for } \hat{\gamma}_{ACE} \geqslant 0.80 \end{cases}$$

其中，

$$N_{rare} = \sum_{i=1}^{abund} in_i \quad C_{ACE} = 1 - \frac{n_1}{N_{rare}}$$

$$\hat{\gamma}^2_{ACE} = max\left[\frac{S_{rare}}{C_{ACE}}\frac{\sum_{i=1}^{abund} i(i-1)n_i}{N_{rare}(N_{rare}-1)} - 1, 0\right]$$

$$\hat{\gamma}^2_{ACE} = max\left[\hat{\gamma}^2_{ACE}\left\{1 + \frac{N_{rare}(1-C_{ACE})\sum_{i=1}^{abund} i(i-1)n_i}{N_{rare}(N_{rare}-C_{ACE})}\right\}, 0\right]$$

S_{rare} = 含有 "abund" 条序列或者少于 "abund" 的 OTU 数目

S_{abund} = 多于 "abund" 条序列的 OTU 数目

a_{bund} = "优势" OTU 的阈值，默认为 10

Shannon: 用来估算样品中微生物多样性指数之一，它与 Simpson 多样性指数常用于反映 alpha 多样性指数。**Shannon** 值越大，说明群落多样性越高。计算公式如下：

$$H_{shannon} = -\sum_{i=1}^{S_{obs}} \frac{n_i}{N} ln \frac{n_i}{N}$$

其中，

S_{obs} = 实际观测到的 OTU 数 ni = 第 i 个 OTU 包含的序列数

N = 所有个体数目，此处为序列总数

Simpson: 用来估算样品中微生物多样性指数之一，由 Edward Hugh Simpson (1949) 提出，在生态学中常用来定量描述一个区域的生物多样性。**Simpson** 指数值越大，说明群落多样性越低。计算公式如下：

$$D_{simpson} - \frac{\sum_{i=1}^{S_{obs}} n_i(n_i - 1)}{N(N-1)}$$

其中，

S_{obs} = 实际观测到的 OTU 数

ni = 第 i 个 OTU 包含的序列数

N = 所有个体数目，此处为序列总数

Coverage: 各样品文库的覆盖率，其数值越高，则样本中序列没有被测出的概率越低。该指数实际反映了本次测序结果是否代表样本的真实情况。计算公式如下：

$$C = 1 - \frac{n_1}{N}$$

其中，

n1 = 只含有一条序列的 OTU 数目 (如 "singletons")

N = 所有个体数目，此处为序列总数

Shannoneven: 是一个用于反映物种个体数目在群落中分配的均匀程度的指数。均匀度（evenness），是指一个群落或生境中全部物种个体数目的分配状况。反映的是各个物种个体数目分配的均匀程度。计算公式如下：

$$J' = \frac{H'}{H'_{\max}}$$

其中，

H' = 实际观察的 Shannon 多样性指数

H'max = ln(S)，为最大的物种多样性指数，其中 S 为群落中的总物种数软件：**mothur**。

4.6.2 结果说明

结果目录 **3_AlphaDiversity/AlphaIndex/ alpha_diversity_index.xls**:

表 4.6.1 Alpha 多样性指数统计表

Sample	Number	OTUs	Shannon	Chao	Ace	Simpson	Shannoneven	Coverage
A11	52855.0	159.0	0.90	176.55	178.93	0.68	0.18	1.00
A12	62084.0	240.0	2.45	303.59	266.17	0.12	0.45	1.00
A21	77931.0	523.0	2.34	557.70	546.60	0.17	0.37	1.00
A22	51768.0	267.0	1.33	448.86	626.49	0.36	0.24	1.00
A31	63375.0	588.0	5.09	700.8	697.63	0.02	0.80	1.00
A32	45886.0	137.0	1.01	160.33	167.02	0.58	0.21	1.00
A41	49257.0	134.0	1.25	183.29	228.48	0.46	0.26	1.00
A51	58281.0	214.0	1.94	284.91	245.68	0.29	0.36	1.00
A52	64339.0	140.0	1.94	149.0	144.59	0.29	0.39	1.00
B11	62508.0	224.0	2.70	265.35	313.06	0.25	0.50	1.00
B31	59296.0	848.0	5.45	947.0	877.93	0.01	0.81	1.00
B32	58782.0	1072.0	5.13	1105.46	1081.87	0.05	0.74	1.00
B41	49376.0	224.0	1.39	251.35	238.64	0.54	0.26	1.00
B42	64811.0	336.0	4.18	351.3	348.66	0.06	0.72	1.00

Sample: 样本名称。

Number: 样本的序列数。

OTUs: 样本聚类得到的 OTU 数目。

其余各列为多样性指数类型在各样本中所对应的数值。

***_t.test.xls:** 组间 Alpha 指数差异检验表格。

***_alpha_diversity_test.pdf:** 组间 Alpha 指数差异检验箱线图。

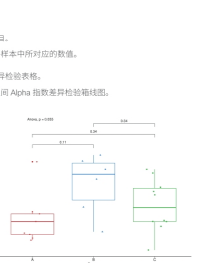


图 4.6.2 组间 Alpha 指数差异检验箱线图

横轴为分组名，纵轴为每组的 Alpha 指数。每两组之间使用 T 检验，组间比较使用 Anova 检验。

(以 Shannon 指数为例)

4.7 稀释曲线分析

4.7.1 分析方法

稀释曲线是从样本中随机抽取一定数量的序列，统计这些序列对应样本的 Alpha 多样性指数，以抽取的数据量为横轴，以 Alpha 多样性指数值为纵轴绘制曲线，根据曲线是否达到平缓来判断本次测序数据量是否足够。

软件：利用 **mothur** 做 rarefaction 分析，利用 **R** 软件制作曲线图。

4.7.2 结果说明

结果目录：**3_AlphaDiversity/Rarefaction**

***.xls:** Alpha 指数稀释性曲线表。

***_rarefaction.pdf:** Alpha 指数稀释性曲线图。

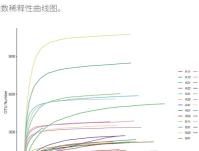


图 4.7.1 Alpha 指数稀释性曲线图

横轴为样本中随机抽取序列数，纵轴为所得相应的 Alpha 指数，每条曲线是一个样本。

（以 OTU 数目为例）

4.8 系统发生进化树

4.8.1 分析方法

在分子进化研究中，系统发生的推断能够揭示出有关生物进化过程的顺序，了解生物进化历史和机制，可以通过某一分类水平上序列间碱基的差异构建进化树。

软件：使用 **MUSCLE** 或者 **MAFFT** 对 OTU 序列进行多序列比对得到 alignment 文件，并使用 **BMGE** 进行多序列比对过滤，最后采用 **FastTree** 根据最大似然法 (approximately-maximum-likelihood phylogenetic trees) 构建进化树。

4.8.2 结果说明

结果目录：**2_OTU_Taxa/OTU/**

otus.tre: newick-formatted 树文件，newick 是树状标准格式文件，可被多种建树软件识别。

4.9 样本间距离计算

4.9.1 分析方法

样本间的物种丰度分布差异程度可通过统计学中的距离进行量化分析，使用统计算法 Euclidean，Bray-Curtis，Unweighted_unifrac，weighted_unifrac，计算两两两样本间距离，获得距离矩阵，可用于后续进一步的 beta 多样性分析和可视化统计分析。

样本距离分析提供多种计算方法，常见距离算法有 Bray-Curtis、Jaccard、UniFrac 等。Bray-Curtis 与 Jaccard 距离算法主要基于独立的物种分类单元（如 OTU、属等）进行计算，不考虑各物种之间的进化关系或关联信息。Jaccard 算法采用非加权的计算方法，主要考虑物种的有无，而 Bray-Curtis 算法采用加权的计算方法，同时考虑物种有无和物种丰度。同时，由于微生物极其多样，不同微生物彼此之间的系统发育关系往往千差万别，仅仅将群落中不同微生物成员视为相互独立的变量显然并不合理。因此，在比较不同群落样本之间的差异时，需要考虑两个群落成员之间的系统发育关系是否相似。基于这个思想，计算微生物群落样本间距离的 UniFrac 距离应运而生，通过比较两个群落各自独有的微生物成员之间系统发育关系的远近，更为客观地反映两个群落样本之间的相似程度。UniFrac 距离有非加权（Unweighted）和加权（Weighted）两种，前者仅仅考虑微生物成员在群落中存在与否，而不考虑其丰度高低；后者则兼顾群落成员之间的系统发育关系以及它们在各自群落中的丰度高低。两种距离算法侧重于不同的群落结构特征，究竟是由于群落成员的截然不同导致样本的差异，还是由于同一组成员在不同样本中丰度梯度的改变导致样本的差异。因此，合理运用非加权和加权两种 UniFrac 距离，可以较全面地揭示微生物群落数据背后隐含的生态学意义。

软件：使用 **phyloseq** 根据系统发生进化树得到样品间距离矩阵。并使用 **R** 的 **gplots** package 进行距离热图绘制，以便通过颜色直观的展现样本与样本之间的距离关系。

4.9.2 结果说明

结果目录：**4_BetaDiversity**

Distance/*_bray_distance.xls: 样品间的 Bray-Curtis 距离矩阵表。

Distance/otu_(w)unifrac_distance.xls: 基于系统发生进化树得到的样品间 Weighted(Unweighted) Unifrac 距离矩阵表。

DistHeatmap/*_*_distheatmap*.pdf: 样本距离热图。

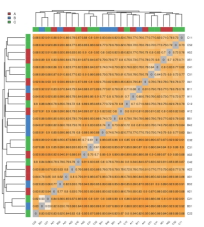


图 4.9.1 样本距离热图

颜色块代表距离值，颜色越灰表示样本间距离越近，每个方格中的数值代表横纵轴相对应样品之间的距离，范围在 0-1 之间，热图中对样本间做了聚类，通过聚类树示可看出样本间的距离关系。

（以在 OTU 水平上 Unweighted Unifrac 距离为例）

4.10 距离箱线图

4.10.1 分析方法

将不同分类或环境的多组样品的距离进行四分位计算绘制箱线图，比较不同样品组的组内和组间的距离分布差异。

软件：**R**。

4.10.2 结果说明

结果目录：**4_BetaDiversity/DistBoxplot**

dist_*_*_distviolin.pdf: 样本距离箱线图。

图 4.10.1 样本距离箱线图

横轴为分组信息，纵轴为距离值。

（以 OTU 水平上 Unweighted Unifrac 距离为例）

4.11 PCA 主成分分析

4.11.1 分析方法

PCA 分析（Principal Component Analysis），即主成分分析，是一种对数据进行简化分析的技术。这种方法可以有效的找出数据中最 “主要” 的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构。其优点是简单且无参数限制。通过分析不同样本群落组成可以反映样本间的差异和距离，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样品间差异的两个特征值。

如样本物种组成越相似，反映在 PCA 图中的距离越近。

软件：**R**。

4.11.2 结果说明

结果目录：**4_BetaDiversity/PCA**

***_pca_axis.xls:** 样本坐标表，样本降维后在各维度（主成分轴）的相对位置，由样本间距离矩阵通过 PCA 分析得来。分析只选取了矩阵特征值排在前两位的坐标数据。

***_pca_rotation.xls:** 物种主成分贡献度表，物种 /OTU 在主成分上的贡献度。

***_pca_importance.xls:** 主成分解释度表，记录了各维度解释结果的百分比。如果 PC1 值为 50%，则表示 x 轴的差异可以解释全面分析结果的 50%。

***_pca.pdf:** PCA 图。

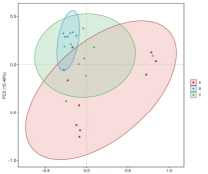


图 4.11.1 PCA 图

横轴和纵轴表示两个选定的主成分轴，百分比表示主成分对样本组成差异的解释程度值；横轴和纵轴的刻度是相对距离，无实际意义；不同颜色或形状的点代表不同分组的样本，两样本点越接近，表明两样本物种组成越相似。

（以 OTU 水平为例）

4.12 PCoA 主坐标分析

4.12.1 分析方法

PCoA (principal co-ordinates analysis) 是一种研究数据相似性或差异性的可视化方法，通过一系列的特征值和特征向量进行排序后，选择主要排在前几位的特征值，PCoA 可以找到距离矩阵中最主要的坐标，结果是数据矩阵的一个旋转，它没有改变样品点之间的相互位置关系，只是改变了坐标系统。通过 PCoA 可以观察个体或群体间的差异，如果样品距离越接近，表示物种组成结构越相似，因此群落结构相似度高的样品倾向于聚集在一起，群落差异很大的样品则会远远分开。

软件：**R** 的 **vegan** 的 package。

4.12.2 结果说明

结果目录：**4_BetaDiversity/PCOA**

***_pcoa_axis.xls:** 样本坐标表，样本在各维度的相对位置。

***_pcoa_values.xls:** 矩阵特征值，样本距离矩阵特征值。

***_pcoa*.pdf:** PCoA 主坐标分析图。

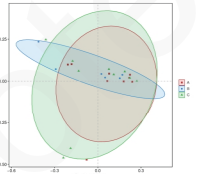


图 4.12.1 PCoA 主坐标分析图

横轴和纵轴表示两个选定的主坐标轴，百分比表示主坐标轴对样本组成差异的解释程度值；横轴和纵轴的刻度是相对距离，无实际意义；不同颜色或形状的点代表不同分组的样本，两样本点越接近，表明两样本物种组成越相似。

（以 OTU 水平上 Unweighted Unifrac 距离为例）

4.13 NMDS 非度量多维尺度分析

4.13.1 分析方法

NMDS (Non-metric Multidimensional Scaling, 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。其基本特征是将对象间的相似性或相异性数据看成点间距离的单调函数，在保持原始数据次序关系的基础上，用新的相同次序的数据列替换原始数据进行度量型多维尺度分析。换句话说，当资料不适合直接进行变量型多维尺度分析时，对其进行变量变换，再采用变量型多维尺度分析，对原始资料而言，就称之为非度量型多维尺度分析。其特点是根据样品中包含的物种信息，以点的形式反映在多维空间上，而对不同样品间的差异程度，则是通过点与点间的距离体现的，最终获得样品的空间定位点图。

软件：**R** 的 **vegan** package。

4.13.2 结果说明

结果目录：**4_BetaDiversity/NMDS**

***_nmds_axis.xls:** 样本坐标表，样本在各维度的相对位置。

***_stressplot.pdf:** Shepard 图，比较 NMDS 排序图内对象的距离与原始对象距离，其中 R^2 越大，NMDS 结果越合理。

***_nmds.pdf:** NMDS 非度量多维尺度分析图。

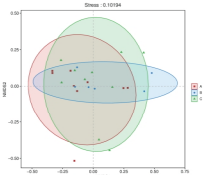


图 4.13.1 NMDS 非度量多维尺度分析图

不同颜色或形状的点代表不同分组的样本，两样本点越接近，表明两样本物种组成越相似。横纵轴表示相对距离，无实际意义。Stress：检验 NMDS 分析结果的优劣。通常认为 stress<0.2 时可用 NMDS 的二维点图表示，其图形有一定的解释意义；当 stress<0.1 时，可以认为是一个好的排序；当 stress<0.05 时，则具有很好的代表性。

（以 OTU 水平上 Unweighted Unifrac 距离为例）

4.14 样本层级聚类分析

4.14.1 分析方法

对距离矩阵进行层级聚类（Hierarchical clustering）分析，构建树状结构，得到树状关系形式用于可视化分析。

软件：使用 **R** 的 hclust 函数构建聚类树，并利用 ape package 绘制树状图。

4.14.2 结果说明

结果目录：**4_BetaDiversity/Tree**

***_*.tre:** newick-formatted 树文件。

***_*_tree.pdf:** 样本层级聚类树展示图。

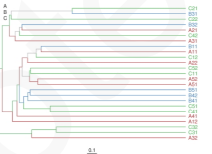


图 4.14.1 样本层次聚类树

树枝的长度代表样本间的距离，越相似的样本会越靠近，样本按分组以不同颜色区分。

（以 OTU 水平上 Unweighted Unifrac 距离为例）

4.15 Anosim 组间相似性分析

4.15.1 分析方法

相似性分析 (Anosim) 是一种非参数检验，用来检验组间（两组或多组）的差异是否显著大于组内差异，从而判断分组是否有意义。首先根据两两样品间的距离将所有距离从小到大进行排序，计算 R 值，之后将样品进行置换，重新计算 R* 值，R* 大于 R 的概率即为 P 值。

软件：**R** 的 **vegan** 的 package。

4.15.2 结果说明

结果目录：**5_GroupInfor/Anosim**

***_bray_anosim_result.xls:** Anosim 组间相似性分析结果表。

表 4.15.1 Anosim 组间相似性分析结果表

Comparison	Sample Size	Group Size	ANOSIM statistic R	P Value	Permutations
A vs B	15	2	0.34	0.01	999
A vs C	19	2	0.15	0.04	999
B vs C	16	2	0.01	0.38	999
Between	25	3	0.17	0.01	999

ANOSIM statistic R: R 值范围实际范围是（-1，1），但一般介于（0，1）之间，R>0，说明组间存在差异。R 等于 0 或在 0 附近，说明组间没有差异。R 偶尔也会 <0，这种情况是组内差异显著大于组间差异，这就说明我们的采样或者分组出现问题。P Value: 统计可信度，P 值越小，表明各样本分组之间的差异显著性越高。P<0.05 表示统计具有显著性。Permutations: 置换次数。（以 OTU 水平上 Bray-Curtis 距离为例）。

***_bray_anosim_boxplot.pdf:** Anosim 组间相似性分析箱型图。

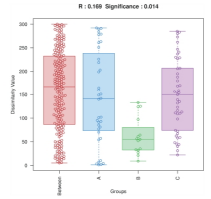


图 4.15.2 Anosim 组间相似性分析箱型图

总共有 N+1 个箱子，N 为分组数量。Between 对应的箱子代表组间差异的距离值，其他分别代表各自组内差异。

（以 OTU 水平上 Bray-Curtis 距离为例）

4.16 Adonis/PERMANOVA 分析

4.16.1 分析方法

Adonis 又称置换多因素方差分析（permutational MANOVA）或非参数多因素方差分析（nonparametric MANOVA），是一种对半度量或度量距离矩阵的离差平方和进行区分的非参数统计学方法。它利用距离矩阵对总方差进行分解，分析不同分组因素对样品差异的解释度，并使用置换检验对划分的统计学意义进行显著性分析。

软件：**R** 的 **vegan** 的 package。

4.16.2 结果说明

结果目录：**5_GroupInfor/Adonis**

***_bray_adonis_result.xls:** Adonis 分析结果表格文件。

表 4.16.1 Adonis 分析结果

Comparison	Sample Size	Group Size	F.Model	R2	P Value
A vs B	15	2	2.79	0.18	4.0e-03
A vs C	19	2	1.55	0.08	0.10
B vs C	16	2	1.36	0.09	0.11
Between	25	3	1.83	0.14	4.0e-03

comparison: 组间比较名称

F.Model：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

P Value：表示 P 值，小于 0.05 说明本次检验的可信性度高。（以 OTU 水平上 Bray-Curtis 距离为例）

4.17 PLS-DA 分析

4.17.1 分析方法

PLS-DA(Partial Least Squares Discriminant Analysis),即偏最小二乘法判别分析，是多变量数据分析技术中的判别分析法，经常用来处理分类和判别问题。通过对主成分适当的旋转，PLS-DA 可以有效的对组间观察值进行区分，并且能够找到导致组间区别的影响变量。

PLS-DA 采用了经典的偏最小二乘回归模型，其响应变量是一组反应统计单元间类别关系的分类信息，是一种有监督的判别分析方法。因无监督的分析方法（PCA）对所有样本不加以区分，即每个样本对模型有着同样的贡献，因此，当样本的组间差异较大，而组内差异较小时，无监督分析方法可以明显区分组间差异; 而当样本的组间差异不明晰，而组内差异较大时，无监督分析方法难以发现和区分组间差异。另外，如果组间的差异较小，各组的样本量相差较大，样本量大的那组将会主导模型。有监督的分析（PLS-DA）能够很好的解决无监督分析中遇到的这些问题。

与 PCA 分析的原理相同，PLS 利用偏最小二乘法对数据结构进行投影分析。但 PLS 与 PCA 数据有本质的不同，PCA 分析方法中只有一个数据集 X，所有分析都只是基于这个唯一的数据集，对应于一个多维空间。而 PLS 分析是建立在两个数据集 X 和 Y 基础上的，以此也就对应地存在两个多维空间。在利用投影方法计算 PLS 第一个主成分后，分别得到 X 和 Y 空间的两条轴线以及各个样本点在 X 和 Y 空间轴上的得分 t1、u1。对 X 和 Y 数据的关联分析就是将所有样本在 X 和 Y 空间第一个主成分轴上的得分 t1、u1 分别作相关分析，可以表示为 u1 = b1*t1, i 表示不同样本，n1 表示残差。对应的，经过第二个主成分计算可以得到 t2、u2, 有关系式 u2 = b2*t2 。

如果用 t1 、t2 作图，表示数据集 X 的 PCA 得分图，而如果用 t1、u1 作图就表示第一个主成分下数据集 X 与数据集 Y 相关性。与 PCA 的载荷图（变量分布散点图）相类似，PLS 可以用权重方式对 X、Y 数据集中的变量进行相关联，找出变量之间的关系。

PLS-DA 只需要一个数据集 X，但在分析时必须对样本进行指定分组，这样分组后模型自动加上另外一个隐含的数据集 Y，该数据集变量数等于组别数，赋值时把指定的那一组规定为 1，其他所有值均为 0。其他计算方法与上述 PLS 方法相同。这种模型计算的方法强行把各组分广泛约束，有利于发现组间的异同点。

软件：使用 **R** 的 **mixOmics package**，根据物种丰度矩阵和样本分组数据构建 PLS-DA 判别模型。

4.17.2 结果说明

结果目录：**5_GroupInfor/PLSDA**

***_plsda_sites.xls:** 样本坐标表，样本降维后在各维度的相对位置。分析只选取了矩阵特征值排在前两位的坐标数据。

***_plsda_rotation.xls:** 物种主成分贡献度表，物种 /OTU 在主成分上的贡献度。

***_plsda_importance.xls:** 分组主成分贡献度表，各分组在主成分上的贡献度。

***_plsda*.pdf:** PLS-DA 分析图。

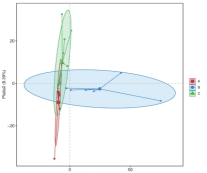


图 4.17.1 PLS-DA 分析图

不同颜色或形状的点代表不同环境或条件下的样本组，横轴、纵轴的刻度是相对距离，无实际意义。PLSDA1、PLSDA2 分别代表对于两组样本微生物组成发生偏移的疑似影响因素，需要结合样本特征信息归纳总结，假如 A 组和 B 组样本在 PLSDA1 轴的方向上分离开来，则可分析为 PLSDA1 是导致 A 组和 B 组分开（可以是两个地点或酸碱不同）的主要因素，同时验证了这个因素有较高的可能性影响了样本的组成。

（以 OTU 水平为例）

4.18 菌群分型分析

4.18.1 分析方法

菌群分型分析，主要通过统计聚类的方法研究不同样本优势菌群结构的分型情况，分型过程中一般不考虑环境因子等外部因素的影响。通过该分析，可以将优势菌群结构近似的不同样本聚为一类，主要适用于特定环境样本的菌群分型，如肠型（enterotypes），阴道分型（cervicotype），口腔分型等。

通常根据菌群在所选分类水平上的相对丰度，计算 Jensen-Shannon Distance（JSD）等距离，并进行 PAM (Partitioning Around Medoids) 聚类，通过 Calinski-Harabasz (CH) 指数计算最佳聚类 K 值，然后采用 Between-class analysis (BCA，K≥3) 或 principal coordinates analysis (PCoA，K≥ 2) 进行可视化。

软件：使用 **R** 的 **ade4**、**cluster**、**clusterSim** package。

4.18.2 结果说明

结果目录：5_GroupInfor/Enterotype

"_Enterotype_summary.xls: 各组不同分型的个体数目表，各组样本中不同分型的样本数目。

"_Enterotype.xls: 样本的分型分类表，每个样本所属的分型。

"_type".xls: 菌群分型样本丰度表，各分型中的物种在各个样本中的丰度。

"_Enterotype_bar.pdf: 各组分型组成柱状图。

"_CH-index.pdf: CH 指数图。

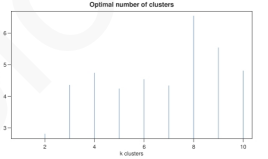


图 4.18.1 CH 指数图

横轴表示聚类的个数（K 值），纵轴表示 CH 指数的大小；一般 CH 指数越大，聚类效果越佳，故选取 CH 指数最大时的聚类个数进行分型。

（以 OTU 水平为例）

"_pcoa_axis_infor.xls: 菌群分型 PCoA 样本坐标图。

"_pcoa.pdf: 菌群分型 PCoA 图。

"_bac_axis_infor.xls: 菌群分型 BCA 样本坐标图。

"_bca.pdf: 菌群分型 BCA 图。

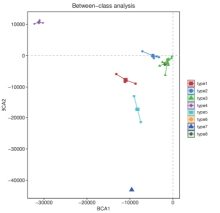


图 4.18.2 菌群分型 BCA 图

不同颜色或形状的点代表不同分型的样本，两样本点越接近，表明两样本组成越相似。

（以 OTU 水平为例）

4.19 物种分布韦恩图

4.19.1 分析方法

VENN 图可以用来统计样本中共有的和独有的 OTU 的数目，直观的展现出环境样品的 OTU 数目组成相似性及重叠情况。

当样本数小于 10 个时，提供 UpSet 图展示结果；当样本数大于 5 个时，提供花瓣图展示结果。

使用软件：**R** 的 **VennDiagram**、**UpsetR** package。

4.19.2 结果说明

结果目录：**6_Taxonomic/VENN/**

***_core_spec.xls:** 样本间共有或特有的物种分类表格。

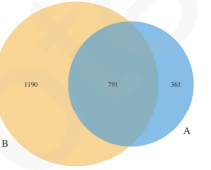


图 4.19.1 物种分布韦恩图或花瓣图

不同样品（组）用不同颜色表示，图中数字代表特异或共有的物种数。

（以 OTU 水平为例）

4.20 相对丰度图

4.20.1 分析方法

使用统计学的分析方法，观测样本在不同分类水平上的群落结构。将多个样本的群落结构分析放在一起对比时，还可以观测其变化情况。根据研究对象是单个或多个样本，结果可能会以不同方式展示。通常使用较直观的饼图或柱状图等形式呈现。

如无特殊说明，默认将在所有样本中丰度占比均小于一定比例（1%）的物种归为 others，其余的作为优势物种进行分析。

使用软件：**R**。

4.20.2 结果说明

结果目录：6_Taxonomic/Barplot/

*_barplot.xls: 优势物种统计表。

Barplot/*_barplot.pdf: 优势物种相对丰度柱状图。

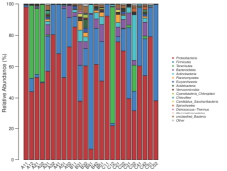


图 4.20.1 相对丰度柱状图

横轴为各样品的编号，纵轴为物种相对丰度比例。颜色对应此分类学水平下各物种名称，不同色块宽度表示不同物种相对丰度比例。

（以 phylum 门水平为例）

Pie/*_*_pie.xls: 各样本优势物种相对丰度饼状图

Pie/*_*_pie.xls: 各样本优势物种相对丰度饼状图

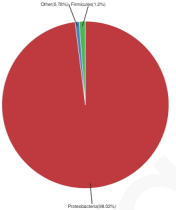


图 4.20.2 相对丰度饼状图

某一样本在选定分类学水平下的微生物群落组成。不同颜色表示不同的物种，饼图积表示该物种所占百分比。

（以 phylum 门水平为例）

4.21 共线性关系图

4.21.1 分析方法

共线性关系图是一种描述样本与物种、功能之间对应关系的可视化图图，该图不仅反映了每个样本的优势物种组成比例，同时也反映了各优势物种在不同样本之间的分布比例。

软件：**R** 的 **circlize** package。

4.21.2 结果说明

结果目录：**6_Taxonomic/Circos**

***_circos.xls**: 优势物种统计表。

***_circos.pdf**: 共线性关系图。

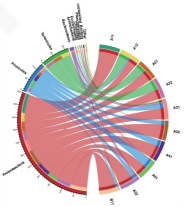


图 4.21.1 共线性关系图

右边半圆表示样本的物种丰度组成情况，左边半圆表示在物种在不同样本中的分布比例情况。圆圈从外到内：第一、二彩色圈：右半部分圆圈表示不同样本对应的物种组成，不同颜色表示不同物种，长度代表某一物种在该样本中的丰度比例（第二圈内显示的百分比）；左半部分圆圈表示不同样本在优势物种中的分布比例，不同颜色表示不同样本，长度代表该样本在某一物种中的分布比例（第二圈内显示的百分比）；第三圈：圈内的彩色条带，一端连接样本（右边半圆），条带端点宽度表示物种在该样本中的丰度，另一端连接物种（左边半圆），条带端点宽度表示该样本在相应物种中的分布比例，圆外数值表示相应物种的丰度数值。（为了显示效果，仅显示前 10 个样本和丰度最高的前 10 个物种分类；

（以 phylum 门水平为例。）

4.22 丰度 3D 柱状图

4.22.1 分析方法

丰度 3D 柱状图可以更立体的观察所有样本中的优势物种或功能分布情况。

软件 Python 的 matplotlib package。

4.22.2 结果说明

结果目录：**6_Taxonomic/ThreeD**

***_3dbarplot.xls:** 优势物种统计表。

***_3dbarplot.pdf:** 相对丰度 3D 柱状图。

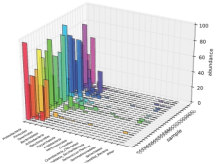


图 4.22.1 相对丰度 3D 柱状图

横轴代表菌落，纵轴代表相对丰度比例，Z 轴代表样本。

（以 phylum 门水平为例）

4.23 优势物种 Heatmap 图

4.23.1 分析方法

Heatmap 可以用颜色变化来反映二维矩阵或表格中的数据信息，它可以直观的将数据值的大小以定义的颜色深浅表示出来。常根据需要将数据进行物种、功能或样本间相似性聚类，将聚类后的数据表示在 heatmap 图上，可将高峰度和低丰度的物种或功能分块聚集，同过颜色梯度及相似程度来反映多个样本在各分类水平上群落组成或功能的相似性和差异性。

软件：**R** 的 **circlize** package。

4.23.2 结果说明

结果目录：**6_Taxonomic/Heatmap**

***_select.xls:** 优势物种统计表

***_heatmap*.pdf:** 丰度 Heatmap 热图

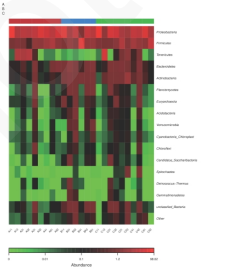


图 4.23.2 丰度 Heatmap 热图

每一行代表一个物种，每一列代表一个样本，每一小格的颜色代表物种在样本中的相对丰度，颜色越深（红），表示物种的丰度越高，颜色越绿，表示物种的丰度越低。在图上的上方有颜色块，来自同一组的样本颜色相同。

（以 phylum 门水平为例）

4.24 物种分布气泡图

4.24.1 分析方法

气泡是可用于展示三个变量之间的关系。它与散点图类似，绘制时将一个变量放在横轴，另一个变量放在纵轴，而第三个变量则用气泡的大小来表示。物种分布气泡图可以很明显的看出样本间物种丰度的差异。

4.24.2 结果说明

结果目录：**6_Taxonomic/Ballon**

***_ballon.xls:** 优势物种统计表。

***_ballon.pdf:** 物种分布气泡图。

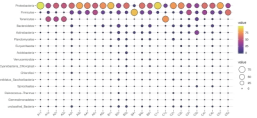


图 4.24.1 物种分布气泡图

横轴为样本，纵轴为物种分类信息，丰度信息通过气泡大小和颜色来进行表示，颜色越黄，气泡越大，代表物种丰度越高。

（以 phylum 门水平为例）

4.25 样本聚类树与柱状图组合分析

4.25.1 分析方法

根据样品中相似程度进行排布，并绘制对应样本聚类树状图反应样本中功能组分柱状图。

4.25.2 结果说明

结果目录：**6_Taxonomic/TreeBarplot**

***_treebarplot.pdf:** 样本聚类树与柱状图组合分析图。

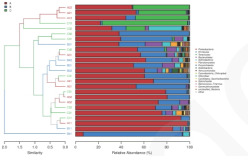


图 4.25.1 样品聚类树与柱状图组合分析图

左侧是相似性树状图，样本间差异越小，样本更会处在相近的同一个分支，样本颜色按分组信息区分。右侧柱状图，展示样本中的物种分布，不同颜色代表不同物种。

（以 phylum 门水平为例）

4.26 单样品多级物种组成图

4.26.1 分析方法

单样品多级物种组成图可以将单个样本在域、门纲目等分类学水平的注释结果，通过多个同心圆由内向外直观地展现出来。

软件：**Krona**。

4.26.2 结果说明

结果目录：**6_Taxonomic/Krona**

Krona.html: 样品多级物种组成图。

图 4.26.1 单样品多级物种组成图

从最里圈往外圈看，依次为域、门、纲、目、科、属等水平的物种组成。

（仅为示例图）

4.27 分类学系统组成树

4.27.1 分析方法

根据每个样本或多个样本的分类学比对结果，选出优势物种的分类，从整个分类系统上了解测序的环境样本中优势微生物的进化关系和丰度差异。

软件：**python** 的 **ete3** package。

4.27.2 结果说明

结果目录：**6_Taxonomic/ETE3**

***_ete3.pdf:** 单样本分类系统组成树状图。

图 4.27.1 单样品分类系统组成树状图

图中不同颜色代表不同分类层级，从左至右依次为分类层级，圆圈大小代表物种丰度，支点附近分别为该分类名称和其对应丰度数值。

（以 genus 属水平为例）

*_ete3.pdf: 所有样本分类学系统组成树状图。

图 4.27.2 所有样本分类学系统组成树状图

比对到不同样品在某分支上的丰度差异，通过带颜色的饼状图呈现。不同颜色代表不同的样品，颜色的扇形面积越大，说明在该分支上该样品的丰度越高。支点附近分别为该分类名称和其对应平均丰度数值。

（以 genus 属水平为例）

4.28 分类和系统发育信息可视化

4.28.1 分析方法

根据每个样本的分类学比对结果，选出优势物种的分类，结合物种丰度信息，以环状树状图显示。

软件：**GraPhlAn**。

4.28.2 结果说明

结果目录：**6_Taxonomic/GraPhlAn**

***_graphlan.pdf**: GraPhlAn 绘制的分类和系统发育信息可视化图

图 4.28.1 单样品分类和系统发育信息可视化图

图居中的为所有样本中丰度大于 1% 的物种进化分类树，不同的颜色代表不同的门，点的大小代表丰度大小。平均丰度值大于 1% 的结点用紫色方块表示，反之用橙色倒三角标出。外围环为热力图，颜色越深代表丰度越高；最外环为柱状图，柱子越高代表丰度越高。

（以 genus 属水平为例）

4.29 Ternary 三元相图

4.29.1 分析方法

Ternary 三元相图是用一个三角形描述三个变量之间不同属性的比率关系，在分析中可以根据物种分类或功能信息对三个样品的物种或功能组成进行比较分析，通过三角图可以直观的显示出不同物种或功能在样品中的比重和关系。

软件：**R** 的 **ggtern** package。

4.29.2 结果说明

结果目录：**6_Taxonomic/Ternary**

***_ternary.xls:** T 优势物种统计表。

***_ternary.pdf:** Ternary 三元相图。

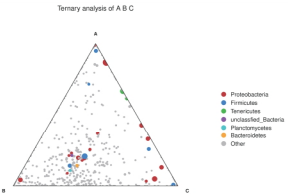


图 4.29.1 Ternary 三元相图

三个角分别代表三个样本（组），彩色圆代表优势门水平的物种分类，三角图中的圆圈代表某一个门水平下包含的所有当前水平的物种分类，圆圈大小代表物种的平均相对丰度。

（以 genus 属水平为例）

4.30 两组样本 Welch’ s t-test 分析

4.30.1 分析方法

STAMP 差异分析用于比较两组样本之间物种或功能的丰度，通过此分析可获得显著性差异物种或功能以及该物种或功能更趋向何种环境条件下的样本。

当比较对象为组时，采用 Welch’ s t-test。最后将检验得到的 pvalue 值采用 FDR 做 Multiple test correction 得到 qvalue 值。

软件：**STAMP**。

4.30.2 结果说明

结果目录：**7_Different/STMAP**

***_vs_*.*_diff.xls:** Welch’ s t-test 差异分析结果表。

***_vs_*.*_diff_sign.xls:** Welch’ s t-test 显著差异分析结果表。

***_vs_*.*_ExtendErrorBar.pdf:** Welch’ s t-test 差异分析结果图。

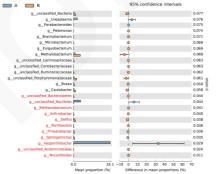


图 4.30.1 差异分析结果图

左图所示为不同物种分类在两组样本中的丰度比例，中间所示为 95% 置信度区间内，物种分类丰度的差异比例，最右边的值为 p 值，p 值 <0.05，表示差异显著，红色标识。

（仅列出 p 值最低的 25 个，以 genus 属水平为例。）

4.31 ANOVA 方差分析

4.31.1 分析方法

方差分析（Analysis of Variance，ANOVA），又称“变异数分析”，用于两个及两个以上样本组均数差别的显著性检验。由于各种因素的影响，研究所得的数据呈现波动状。造成波动的原因可分成两类，一是不可控的随机因素，另一是研究中施加的对结果形成影响的可控因素。方差分析是从观测变量的方差入手，研究诸多控制变量中哪些变量是对观测变量有显著影响的变量。

软件：**STAMP**。

4.31.2 结果说明

结果目录：**7_Different/ANOVA/**

***_anova.xls:** ANOVA 分析结果表格文件。

***_anova_sign.xls:** 显著 ANOVA 分析结果表格文件。

表 4.31.1 ANOVA 分析结果

	Eta-Squared	p-value	q-value
g_unclassified Acidimicrobidae	0.45	1.4e-03	0.84
g_Enhydrobacter	0.37	5.7e-03	1.0
g_Methanobacterium	0.37	6.1e-03	1.0
g Arthrobacter	0.37	6.5e-03	0.99
g_unclassified Verrucomicrobiaceae	0.35	9.0e-03	1.0
g_Mesorhizobium	0.35	9.4e-03	0.96
g_Nocardioides	0.33	0.01	1.0
g_Alloprevotella	0.33	0.01	0.98
g_unclassified Bacteroidetes	0.32	0.01	1.0
g_Smaragdicoccus	0.31	0.02	0.99
g_unclassified Coriobacteriaceae	0.31	0.02	0.91
g_unclassified Bacillales	0.31	0.02	0.84
g_Ruminococcus	0.31	0.02	0.80
g_Clostridium IV	0.31	0.02	0.80

Eta-Squared: 关联强度 (correlation ratio)，因变量的变异被自变量解释的百分比。

p-value: 统计可信度，P<0.05 表示统计具有显著性。

q-value: FDR 矫正后的 p-value 值。（以 genus 属水平为例）

Eta-Squared: 关联强度 (correlation ratio)，因变量的变异被自变量解释的百分比。

p-value: 统计可信度，P<0.05 表示统计具有显著性，q-value: FDR 矫正后的 p-value 值。（以 genus 属水平为例）将 ANOVA 方差分析得到的显著物种分类采用 Scheffe 算法进行 Post Hoc 检验。

*_*_postHoc.xls: Post Hoc 分析结果表

*_*_postHoc.pdf: Post Hoc 分析结果图

图 4.31.2 Post Hoc 分析结果图

左图所示为不同组中在某个物种分类中的丰度比例，不同颜色代表不同的组，中间所示为 95% 置信度区间内组间差异比例，最右边的值为 p 值，p 值 <0.05，表示差异显著，组间比较用红色标记。

（以 genus 属水平为例）

4.32 Wilcox 秩和检验分析

4.32.1 分析方法

Wilcoxon rank-sum test，也叫曼 - 惠特尼检验 (Mann-Whitney U test)，是两组独立样本非参数检验的一种方法。其原假设为两组独立样本来自的两总体分布无差异，同过对两组样本平均秩的研究来实现判断两总体的分布是否首在差异，该分析可以对两组样本的物种、基因或者功能进行显著性差异分析，并对 p 值进行假发现率 (FDR) q 值。

软件：**R**

4.32.2 结果说明

结果目录：**7_Different/Wilcox**

***_vs_*_wilcox.xls:** Wilcox 秩和检验差异分析结果表。

***_vs_*_wilcox_sign.xls:** Wilcox 秩和检验显著差异分析结果表。

表 4.32.1 Wilcox 秩和检验分析结果

Feature ID	Freq1(A)	Freq2(B)	pValue	qValue	Difference between means	95% lower CI
g Abiotrophia	6.4e-04	0	0.27	0.46	9.6e-04	0
g Acetoanaerobium	0	5.9e-03	0.09	0.43	-0.01	-0.01
g Achromobacter	9.8e-03	4.2e-03	0.94	0.99	1.5e-04	-0.00
g Acidphilium	0.03	0.20	0.95	0.99	-0.09	-0.21
g Aciditerrimonas	0.02	0.01	0.92	0.99	8.3e-06	-1.2e-05
g Acidithiobacillus	6.9e-03	0	0.50	0.66	4.1e-05	0
g Acidocella	0.02	0.04	1	1	-2.3e-05	-6.2e-05
g Acidovorax	0.02	0.03	0.68	0.86	-0.02	-0.05
g Acinetobacter	1.28	1.63	0.01	0.29	-1.11	-1.93
g Actinobacillus	9.94	6.2e-03	0.25	0.46	15.00	-0.00
g Actinocatenispora	0	2.6e-03	0.28	0.46	15.00	-7.0e-05
g Actinomyces	4.1e-03	0.07	0.14	0.46	-0.08	-0.16
g Actinomycetospora	2.3e-04	0	0.50	0.66	4.3e-05	0
g Actinotalea	2.2e-03	0.14	1	1	-2.1e-05	-8.7e-05

Freq：物种平均相对丰度。

pvalue：两组样本丰度的秩和是否有显著差异的概率值。

qvalue：矫正后的 pvalue。

Difference between means：组间平均丰度差异。

95% lower CI：置信区间下限。

95% upper CI：置信区间上限。（以 genus 属水平为例）

4.33 LEfSe 差异物种判别分析

4.33.1 分析方法

LEfSe (Linear discriminant analysis Effect Size, 线性判别分析及影响因子) 是用于发现不同生物条件或环境下的两组或多组样本中最能解释组间差异的基因或功能特征，以及这些特征对组间差异的影响程度。LEfSe 适用于多层次的生物学标识和特征的发现和解释。运用统计学方法进行差异特征发现和显著性检验，软件首先使用非参数系数的 Kruskal-Wallis (KW) sum-rank test 检测组间丰度显著差异的特征，如果组间有相关联的子分组，则再进一步使用 (unpaired) Wilcoxon rank-sum test 对上一步的差异特征在子分组中的差异一致性检查，最后运用 LDA 判别分析估计这些差异特征对组间区别的影响大小。

软件：**LEfSe**。

4.33.2 结果说明

结果目录：**7_Different/LEfSe**

***_lefse.xls:** LEfSe 输入文件。

***_lefse_result.xls:** LEfSe 分析得到的显著 Marker LDA 值表格。

***_lefse_bar.pdf:** LEfSe 分析得到的显著 Marker LDA 值柱状图。

***_lefse.pdf:** LEfSe 差异判别分析环形树状图。

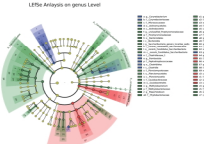


图 4.33.1 LEfSe 组间差异判别分析环形树状图

不同颜色表示不同分组，树枝中不同颜色节点表示在该颜色对应分组中的起到重要作用的微生物类群，黄色节点表示均未起到重要作用的微生物类群。英文字母表示的物种名称在右侧图例中进行展示。如果图中颜色一致，则表明没有找到显著的 Maker。

（以 genus 属水平为例）

4.34 metagenomeSeq 差异分析

4.34.1 分析方法

metagenomeSeq 是用 R 开发的一个包，其基本思想是先将数据标准化。然后用零膨胀高斯分布 (Zero-inflated Gaussian distribution) 处理测序深度带来的影响，最后基于线性模型找出差异。主要用于两组样本和多组样本的差异比较，并且每组样品个数是否一致对结果并无影响。

软件：**R** 的 **metagenomeSeq** package。

4.34.2 结果说明

结果目录：**7_Different/Metagenomeseq**

***_vs_*_*_coeff_detail.xls:** metagenomeSeq 差异分析结果表。

***_vs_*_*_coeff_detail_sign.xls:** metagenomeSeq 显著差异分析结果表。

表 4.34.1 metagenomeSeq 差异分析结果

	Samples in group A	Samples in group B	Counts in group A	Counts in group B	oddsRatio	lower
OTU6	6	4	55439	22	1	0.06
OTU1199	7	3	210	4	3.20	0.23
OTU196	0	3	0	220	0	0
OTU10	5	3	20203	4	1.23	0.10
OTU727	1	5	1	224	0.04	4.8e-04
OTU328	1	5	3	166	0.04	4.8e-04
OTU608	2	0	97	0	Inf	0.12
OTU887	0	2	0	29	0	0
OTU1220	2	0	63	0	Inf	0.12
OTU4	7	3	43584	4	3.20	0.23
OTU802	2	0	50	0	Inf	0.12
OTU1223	2	0	81	0	Inf	0.12
OTU636	2	0	52	0	Inf	0.12
OTU610	3	4	86	18	0.28	0.12

Samples_in_group_*：组内丰度大于 0 的样本个数。

Counts_in_group_*：组内物种丰度之和。

oddsRatio：比值比。lower/upper：置信区间的下限和上限。fisherP：精确检验的 P 值，主要含义是样品按有无分组对丰度分组是否有影响，或者有关联。fisherAdjP：假发现率 q 值。logFC：Log2(Fold Change) 值。pvalues：两组的差异 P 值。

adjPvalues：两组的假发现率 q 值。（以 OTU 水平为例）

*_vs_*.otu_manhattan.pdf: metagenomeSeq 显著差异分析结果曼哈顿展示图。

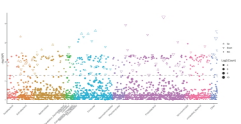


图 4.34.2 metagenomeSeq 显著差异分析结果曼哈顿展示图

不同颜色表示不同的门，点的大小代表平均丰度值。根据 Log2FC 绝对值大于 1 和 pvlaue 小于 0.05 将 OTU 分为上调、下调和非显著三类，用不同的形状标出。

4.35 DESeq2 差异分析

4.35.1 分析方法

在转录组数据分析中，常用 DESeq2 进行差异表达基因的分析。作为延伸，我们也常将这种方法应用在分类测序数据分析中，以寻找两组数据间的差异丰度微生物类群。

软件：R 的 DESeq2 package。

4.35.2 结果说明

结果目录：**7_Different/DESeq2**

***_vs_*.*_deseq2.xls:** DESeq2 差异分析结果表。

***_vs_*.*_deseq2_sign.xls:** DESeq2 显著差异分析结果表。

表 4.35.1 DESeq2 差异分析结果

	Counts(A)	Counts(B)	baseMean	log2FoldChange	lfcSE	stat
OTU4	43584	4	15751.23	-16.74	2.45	-6.85
OTU3	105475	98	11714.66	-10.84	1.67	-6.50
OTU10	20203	4	5876.89	-14.96	2.69	-5.55
OTU8	23467	5	4133.44	-13.54	2.55	-5.32
OTU13	52434	27	1089.32	-9.49	1.94	-4.90
OTU6	55439	22	4583.17	-11.94	2.44	-4.90
OTU34	3137	1	492.84	-12.27	2.73	-4.49
OTU16	16462	840	2603.67	-6.47	1.51	-4.30
OTU1199	210	4	43.92	-8.12	2.03	-4.00

表 4.35.1 DESeq2 差异分析结果

	Counts(A)	Counts(B)	baseMean	log2FoldChange	lfcSE	stat
OTU12	31628	689	1808.38	-6.57	1.67	-3.94
OTU36	676	10912	679.33	4.48	1.45	3.09
OTU14	7675	852	332.95	-4.22	1.38	-3.05
OTU165	116	0	6.62	-7.39	2.53	-2.92
OTU75	2	1784	28.60	7.88	2.71	2.91

Counts：组内物种丰度之和。

baseMean：经过标准化后的丰度均值。log2FoldChange：差异倍数，已进行了 log2 转化。pvalues：两组的差异 P 值。padj：两组的假发现率 q 值。（以 OTU 水平为例）

***_vs_*.otu_manhattan.pdf:** DESeq2 显著差异分析结果 OTU 水平上曼哈顿展示图。

***_vs_*.*_MA.pdf:** DESeq2 显著差异分析结果 MA 展示图。

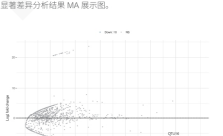


图 4.35.2 DESeq2 显著差异分析结果 MA 展示图

横轴为经过 Log2 转化后的标准化丰度均值，纵轴为经过 Log2 转化后差异倍数，根据 Log2FC 绝对值大于 1 和 padj 小于 0.05 将物种分为上调、下调和非显著三类，用不同的颜色标出。

（以 OTU 水平为例）

4.36 随机森林分析

4.36.1 分析方法

Random Forest 分析，即随机森林分析，属于机器学习算法，是一个包含多棵决策树的分类器，它的分类结果根据检测样本的各个维度上的属性，在不同的决策树上进行判定，综合考虑所有判定结果后给出最终分类，对于分类问题结果取概率最大值，回归分析则取概率均值，它可以高效快速挑选出对样本分类最为重要的物种类别（biomarker）。

软件：R 的 randomForest package。

4.36.2 结果说明

结果目录：**7_Different/RandomForest**

***.*_summary.txt:** RandomForest 分析参数。

***.*_feature_importance_scores.xls:** 各物种对组间差异的贡献值统计表。

***.*_feature_importance_scores.filter.xls:** 筛选后的各物种对组间差异的贡献值统计表，默认挑选贡献值靠前的 30 个。

***.*_importance_table.xls:** 根据贡献度筛选后的物种丰度信息。

***.*_feature_importance_scores.filter.pdf:** 筛选后的各物种对组间差异的贡献值条形图。

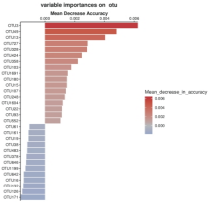


图 4.36.1 筛选后的各物种对组间差异的贡献值条形图

物种重要性排序图，纵轴为重要性衡量标准（比如物种），横轴等于物种的重要性测量值 / 标准差值；纵轴对应按重要性排序后的物种名称。
（以 OTU 水平为例）

4.37 交叉验证分析

4.37.1 分析方法

交叉验证（Cross validation），是一种统计学上将数据样本切割成较小子集的实用方法。先在一个子集上做分析，而其它子集则用来做后续对此分析的确认及验证。一开始的子集被称为训练集。而其它的子集则被称为验证集或测试集。其中最常见的为 k-folder cross-validation，它指的是将所有数据分成 k 个子集，每个子集均做一次测试集，其余的作为训练集。交叉验证重复 k 次，每次选择一个子集作为测试集，并将 k 次的平均交叉验证识别正确率作为结果。所有的样本都被作为了训练集和测试集，每个样本都被验证一次。对随机森林方法筛选出的关键 OTU 的组合进行遍历，以期用最少的 OTU 数目组合构建一个错误率最低高效分类器。

一般地，对随机森林分析筛选出的关键 OTU，按照不同组合进行 10 倍交叉验证分析，找出能够最准确区分组间差异的最少的 OTU 组合，再做进一步的分析，如 ROC 分析等。

软件：R 的 randomForest package。

4.37.2 结果说明

结果目录：**7_Different/RFCV**

***.*_cv_error.xls:** 交叉分析结果表。

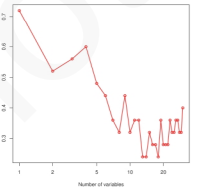


图 4.37.1 交叉分析结果图

横轴表示不同数量的 OTU 组合，纵轴表示该数量物种组合下分类的错误率。物种组合数越少，且错误率越低，则该物种组合被认为是能够区分组间差异的最少的物种组合。
（以 OTU 水平为例）

4.38 ROC 曲线分析

4.38.1 分析方法

接收者操作特征曲线（Receiver operating characteristic curve，ROC 曲线）也是一种有效的有监督学习方法。ROC 分析属于二元分类算法，用来处理只有两种分类的问题，可以用于选择最佳的判别模型，选择最佳的诊断界限值。

可依据专业知识，对疾病组和参照组测定结果进行分析，确定测定值的上下限、组距以及截断点 (cut-off point)，按选择的组距间隔列出累积频数分布表，分别计算出所有截断点的敏感性 (Sensetivity)、特异性和假阳性率 (1- 特异性：Specificity)。以敏感性为纵轴代表真阳性率，(1- 特异性) 为横轴代表假阳性率，作图绘成 ROC 曲线。ROC 曲线越靠近左上角，诊断的准确性就越高。亦可通过分别计算各个试验的 ROC 曲线下的面积 (AUC) 进行比较，哪一种试验的 AUC 最大，则哪一种试验的诊断价值最佳。

软件：R 的 pROC package。

4.38.2 结果说明

结果目录：**7_Different/ROC**

***_vs_*.*_ROC_AUC.xls:** ROC 分析 AUC 结果表。

***_vs_*.*_ROC_*-se.xls:** 物种 ROC 分析结果表。

***_vs_*.*_ROC_*.pdf:** 物种 ROC 分析结果图。

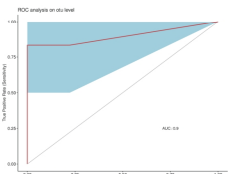


图 4.38.1 物种 ROC 分析结果图

横轴为假阳性率 false positive rate（FPR）：Specificity，纵轴为真阳性率 true positive rate（TPR）：Sensetivity。最靠近左上角的 ROC 曲线的点是错误最少的最好阈值，其假阳性和假阴性的总数最少。ROC 曲线下的面积值在 1.0 和 0.5 之间。在 AUC>0.5 的情况下，AUC 越接近于 1，说明诊断效果越好。AUC 在 0.5~0.7 时有较低准确性，AUC 在 0.7~0.9 时有一定准确性，AUC 在 0.9 以上时有较高准确性。AUC=0.5 时，说明诊断方法完全不起作用，无诊断价值。AUC<0.5 不符合真实情况，在实际中极少出现。

（以 OTU 水平为例）

4.39 Indicator 分析

4.39.1 分析方法

指标性物种是指一定区域范围内能指示生长环境或某些环境条件的生物种、属或群落，而 Indicator 分析常被用来筛选各个样本组中的指标性物种。

软件：R 的 indicspecies package。

4.39.2 结果说明

结果目录：**7_Different/Indicator**

***.*_indicator_statistical.xls:** Indicator 分析统计结果表。

***.*_Indicator.xls:** Indicator 分析结果表。

***.*_Indicator.pdf:** Indicator 分析气泡图。

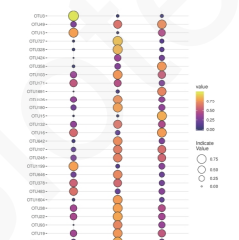


图 4.39.1 Indicator 分析气泡图

横轴表示样本分组信息，气泡的大小表示每个物种在样本组中的 indicator 值大小，即该物种在该分组中的指示性大小。

（以 OTU 水平为例）

4.40 VIF 方差膨胀因子分析

4.40.1 分析方法

影响样本菌群组成的环境 / 临床因子很多，但其中有很多环境 / 临床因子之间具有较强多重共线性（相关）关系，会影响后续的相关分析，所以在进行环境 / 临床因子关联分析前，可以对环境 / 临床因子进行筛选，保留多重共线性较小的环境 / 临床因子，进行后续研究。VIF (Variance Inflation Factor，方差膨胀因子) 分析目前常用的环境 / 临床因子筛选方法。VIF 表达式为：$VIF_i = 1/(1-R_i^2)$。其中 R_i^2 代表模型中与其它自变量相关的第 i 个自变量的方差比例，用于衡量第 i 个自变量与其它自变量间的共线性关系。VIF 值越大，表明自变量间的多重共线性关系越严重。通常认为 VIF 值大于 10 的环境因子是无用的环境因子。过滤掉 VIF 大于 10 的环境因子，进行多次筛选，直到选出的环境因子对应的 VIF 值全部小于 10 为止。

软件：R 的 vegan package。

4.40.2 结果说明

结果目录：**8_Environment/VIF**

***_vif_filter.xls:** 筛选后环境因子的 VIF 值。

env.*_viffilter.xls: 筛选后环境因子表。

***_vif_raw.xls:** 筛选前环境因子的 VIF 值。

表 4.40.1 筛选前环境因子的 VIF 值

	VIF
pH	1154.47
Eh	95.69
FerrousIron	3.30
TrivalentIron	5.77
TC	1517.45
TIC	868.07
TOC	260.06
TNb	9.57
CH4	100.25
CO2	64.64
DecompositionRate	416.07

从左至右依次为环境因子及其 VIF 值。(以 OTU 水平为例)

4.41 Bioenv 生物环境相关分析

4.41.1 分析方法

Bioenv 分析通过计算样本群落结构的距离矩阵和环境因子的距离矩阵，计算两个距离之间的相关系数，挑选出最佳的环境因子组合；默认情况下，计算群落结构的距离矩阵时，使用 Bray-Curtis 距离；计算环境因子的距离矩阵时，使用 Euliodean 欧式距离，计算相关性，则采用 spearman 相关系数。

软件：**R** 的 **vegan package**。

4.41.2 结果说明

结果目录：**8_Environment/Bioenv**

env.*_bioenvfilter.xls: 筛选后环境因子表。

***_bioenv.xls:** 环境因子组合相关性表格。

表 4.41.1 环境因子组合相关性表格

variables	size	correlation
DecompositionRate	1	0.13
TNb DecompositionRate	2	0.11
TC TNb DecompositionRate	3	0.08
TC TOC TNb DecompositionRate	4	0.06
pH TNb CH4 CO2 DecompositionRate	5	0.03
pH TIC TNb CH4 CO2 DecompositionRate	6	0.02
pH TC TIC TNb CH4 CO2 DecompositionRate	7	5.7e-03
pH TC TIC TOC TNb CH4 CO2 DecompositionRate	8	-0.01
pH Eh TC TIC TOC TNb CH4 CO2 DecompositionRate	9	-0.03
pH Eh TrivalentIron TC TIC TOC TNb CH4 CO2 DecompositionRate	10	-0.04
pH Eh FerrousIron TrivalentIron TC TIC TOC TNb CH4 CO2 DecompositionRate	11	-0.06

从左至右依次为不同的环境因子组合及其与群落结构的相关性值。(以 OTU 水平为例)

4.42 DCA 去趋势对应分析

4.42.1 分析方法

对应分析（Correspondence Analysis，CA，或称 Reciprocal Averaging，RA），是一种单峰非约束排序方法，在生态学数据分析中常使用它计算样方与物种之间的对应关系。CA 排序图中，有时会产生“弓形效应”，这是由第一轴和较高轴之间的非线性相关性引起的。与此对应，在 CA 的基础上衍生了去趋势对应分析（Detrended Correspondence Analysis，DCA）以解决这个问题。

进行排序分析之前，首先要判断是选择线性模型（PCA 和 RDA）还是单峰模型 (CA 和 CCA) 的排序方法。一般来说，如果物种分布变化大，选择单峰模型效果比较好，反之，线性模型也是不错。可以通过 DCA 分析来判断，如果 DCA 排序前 4 个轴中最大值超过 4，选择单峰模型排序更合适。如果是小于 3，则选择线性模型更好。如果介于 3-4 之间，单峰模型和线性模型都可行。

软件：**R** 的 **vegan package**。

4.42.2 结果说明

结果目录：**8_Environment/DCA**

***_dca.xls:** DCA 分析表格。

表 4.42.1 DCA 分析表格

	DCA1	DCA2	DCA3
Eigenvalues	0.45	0.41	0.32
Decorana values	1.02	1.62	1.11
Axis lengths	3.03	3.66	2.47

Axis length：对每个 DCA 轴 "梯度长度" 的评估。(以 OTU 水平为例)

4.43 RDA 分析

4.43.1 分析方法

冗余分析（redundancy analysis，RDA）是一种回归分析结合主成分分析的排序方法，也是多响应变量（multiresponse）回归分析的拓展。从概念上讲，RDA 是响应变量矩阵与解释变量之间多元多重线性回归的拟合值矩阵的 PCA 分析，更准确地说，RDA 是一种直接梯度分析技术 (direct gradient analysis technique)，它总结了一组解释变量 "冗余"（即 "解释"）的响应变量分量之间的线性关系。RDA 是基于线性模型，可以检测环境因子，样品，群落结构三者之间的关系或者两两之间的关系。

软件：**R** 的 **vegan package**。

4.43.2 结果说明

结果目录：**8_Environment/RDA**

***_rda_axis_infor.xls:** 维度解释度表，记录了各维度解释结果的百分比。

***_rda_samples_axis.xls:** 样本坐标表，样本降维后在各维度的相对位置。

***_rda_species_axis.xls:** 物种坐标表，物种降维后在各维度的相对位置。

***_rda_env_axis.xls:** 环境因子坐标表，环境因子降维后在各维度的相对位置。

***_rda_envfit.xls:** 环境因子 envfit 分析结果，记录了环境因子和排序轴的相关性、决定系数和显著性检验 p 值。

***_rda.pdf:** RDA 分析结果图。

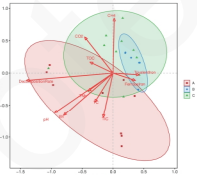


图 4.43.1 RDA 分析图

箭头分别代表不同的环境因子在平面上的相对位置，箭头越长，说明其作用越大，箭头与样本 - 中心连线之间的夹角代表了样本与环境因子之间的相关关系：为锐角时表示两个物种分类之间呈正相关关系，钝角时呈负相关关系。样本（或者群落结构）对环境因子箭头的连线做投影，投影点距离箭头越近，说明该环境因子对样本（或者群落结构）产生的影响越大。

(以 OTU 水平为例)

4.44 CCA 分析

4.44.1 分析方法

典范对应分析（canonical correspondence analusis, CCA），是基于对应分析（CA）发展而来的一种排序方法，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。其基本思路是在对应分析的迭代过程中，每次得到的样方排序坐标值均与环境因子进行多元线性回归。CCA 要求两个数据矩阵，一个是物种数据矩阵，一个是环境数据矩阵。首先计算出一组样方排序值和种类排序值（同对应分析），然后将样方排序值与环境因子用回归分析方法结合起来，这样得到的样方排序值即反映了样方种类组成及生态意义重要值对群落的作用，同时也反映了环境因子的影响。再用样方排序值加权平均求种类排序值，使种类排序坐标值也简介地与环境因子相联系。CCA 是一种基于单峰模型的排序方法，样方排序与对象排序对应分析，而且在排序过程中结合多个环境因子，因此可以把样方、对象与环境因子的排序结果表示在同一排序图上。

软件：**R** 的 vegan package。

4.44.2 结果说明

结果目录：**8_Environment/CCA**

***_cca_axis_infor.xls:** 维度解释度表，记录了各维度解释结果的百分比。

***_cca_samples_axis.xls:** 样本坐标表，样本降维后在各维度的相对位置。

***_cca_species_axis.xls:** 物种坐标表，物种降维后在各维度的相对位置。

***_cca_env_axis.xls:** 环境因子坐标表，环境因子降维后在各维度的相对位置。

***_cca_envfit.xls:** 环境因子 **envfit** 分析结果，记录了环境因子和排序轴的相关性、决定系数和显著性检验 **p** 值。

***_cca.pdf: CCA** 分析结果图。

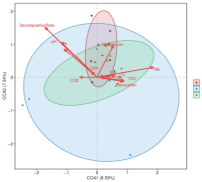


图 4.44.1 CCA 分析图

箭头分别代表不同的环境因子在平面上的相对位置，箭头越长，说明其作用越大，箭头与样本 - 中心连线之间的夹角代表了样本与环境因子之间的相关关系；为锐角时表示两个物种分类之间呈正相关关系，钝角时呈负相关关系。样本（ 或者群落结构，功能）对环境因子箭头的连线做投影，投影点距离箭头越近，说明该环境因子对样本（或者群落结构，功能）产生的影响越大。

（以 OTU 水平为例）

4.45 dbRDA 分析

4.45.1 分析方法

RDA 分析是一种约束性对应分析方法，常采用欧氏距离（Euclidean distances）进行分析。但是欧氏距离并不适用于一些数据类型，采用 db-RDA 分析可以解决数据类型的限制，并用于分析物种与环境因子之间的关系。

db-RDA（distance-based redundancy analysis）是一个三步分析过程：1）计算距离矩阵；2）进行 PCoA 分析；3）利用 PCoA 计算获得的特征值进行 RDA 分析。

db-RDA 分析，和 PCoA 分析类似，但是一种加入了环境因子约束性的分析。

软件：**R** 的 **vegan package**。

4.45.2 结果说明

结果目录：**8_Environment/dbRDA**

***_*_db-rda_axis_infor.xls:** 维度解释度表，记录了各维度解释结果的百分比。

***_*_db-rda_samples_axis.xls:** 样本坐标表，样本降维后在各维度的相对位置。

***_*_db-rda_env_axis.xls:** 环境因子坐标表，环境因子降维后在各维度的相对位置。

***_*_db-rda.pdf:** dbRDA 分析结果图。

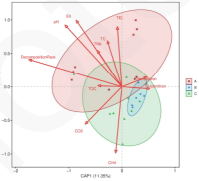


图 4.45.1 dbRDA 分析图

箭头分别代表不同的环境因子在平面上的相对位置，箭头越长，说明其作用越大，箭头与样本 - 中心连线之间的夹角代表了样本与环境因子之间的相关关系；为锐角时表示两个物种分类之间呈正相关关系，钝角时呈负相关关系。样本（ 或者群落结构，功能）对环境因子箭头的连线做投影，投影点距离箭头越近，说明该环境因子对样本（或者群落结构，功能）产生的影响越大。

（以 OTU 水平 Bray-Curtis 距离为例）

4.46 Mantel Test 分析

4.46.1 分析方法

Mantel test 是检验两个矩阵相关关系的非参数统计方法。Mantel test 多用在生态学上检验群落距离矩阵和环境变量距离矩阵（比如 pH，温度或者地理位置的差异矩阵）之间的相关性（Spearman 等级相关系数等）。

Partial Mantel test 在控制矩阵 C 的效应下，来检验 A 矩阵的残留变异是否和 B 矩阵相关。该分析输入两个数值型矩阵，第三个控制矩阵可通过选择因子来确定。

软件：**R** 的 **vegan package**。

4.46.2 结果说明

结果目录：**8_Environment/Mantel**

***_*_env_*_mantantel.xls:** Mantel test 分析结果表格。

表 4.46.1 Mantel test 分析结果

Mantel statistic r	Significance	permutations
0.11	0.06	999

Mantel r statistic : r 统计值，该值越接近 1 表明两矩阵越正相关；越接近 -1 表明两矩阵越负相关；0 表示两矩阵不相关。

permutations : 置换检验的次数。（以 OTU 水平 Bray-Curtis 距离为例）

4.47 排序回归分析

4.47.1 分析方法

线性回归（Linear Regression）是利用数理统计中回归分析，来确定一个或多个自变量和因变量之间关系的一种统计分析方法。环境因子排序回归分析，根据 PCoA 分析结果，以各样本在 PC1 轴上的分值为 x 轴，以该样品对应的环境因子（如 pH、温度等）为 y 轴做散点图，并进行线性回归（Linear Regression），标注 R2，可用于评价二者间的关系。其中 R2 为决定系数，代表变异被回归直线解释的比例。

软件：**R**

4.47.2 结果说明

结果目录： **8_Environment/LinearModel**

***_*_lm.xls:** 排序回归分析结果表格。

表 4.47.1 筛选前环境因子的 VIF 值

Aixs	Environmental	Corrlation	P value
PC1	pH	0.47	0.02
PC1	Eh	0.48	0.01
PC1	FerrousIron	-0.20	0.35
PC1	TrivalentIron	-0.23	0.28
PC1	TC	0.15	0.48
PC1	TIC	0.08	0.70
PC1	TOC	0.21	0.32
PC1	TNb	0.23	0.27
PC1	CH4	-0.00	0.99
PC1	CO2	0.22	0.30
PC1	DecompositionRate	0.71	6.3e-05
PC2	pH	-0.18	0.38
PC2	Eh	-0.18	0.39
PC2	FerrousIron	-0.03	0.89

Environmental : 环境因子名称。

Corrlation: 决定系数，代表变异被回归直线解释的比例

P value: 相关性分析检验 p 值。（以 OTU 水平为例）

***_*_lm_*_*.pdf:** 环境因子排序回归结果图。

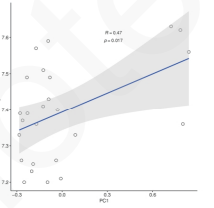


图 4.47.2 环境因子排序回归结果图

横轴为 Beta 多样性排序轴或 Alpha 多样性指数，纵轴为环境因子（如 pH、温度等）。

（以 OTU 水平为例）

4.48 物种与环境因子相关性热图

4.48.1 分析方法

相关性热图通过相关性数值可视化展示样本中不同的物种与环境变量之间的关系，评估微生物分类与环境变量之间的相关性。基本输出是一个矩阵，表示群落中每个微生物分类与每个环境因子变量之间的相关系数，可以用热图直观的展现数值矩阵。

软件：**R**

4.48.2 结果说明

结果目录： **8_Environment/Envcorr**

***_select.xls:** 参与相关性分析的物种分类表格。

***_env_pearson_correlation.xls:** 物种与环境因子相关性矩阵。

***_env_pearson_pvalue.xls:** 物种与环境因子相关性 P 值矩阵。

***_envcorr_heatmap_dendrogram.pdf:** 物种与环境因子相关性热图。

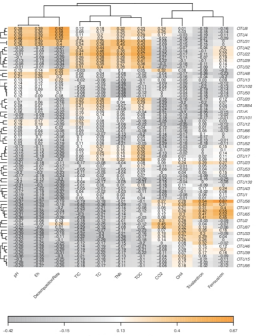


图 4.48.1 物种与环境因子相关性热图

横轴和纵轴分别为环境因子和物种。
（以 OTU 水平为例）

4.49 物种相关性热图

4.49.1 分析方法

相关性分析是用于分析微生物间相互作用关系的经典方法，可甄别出微生物群落间具有显著相关性、强相关、正相关、负相关的各项。分析时选取丰度高于 1% 的物种进行双侧检验。

软件：使用 **SparCC** 计算群落 /OTU 间的相关性系数和 p 值，并使用 **R** 的 **corrplot** package 绘制相关矩阵热图。

4.49.2 结果说明

结果目录：**9_Network/Corrlation**

***_sparcc_correlation.xls:** 物种相关性矩阵。

***_sparcc_pvalue.xls:** 物种相关性 P 值矩阵。

***_sparcc_corrplot.pdf:** 物种相关性热图。

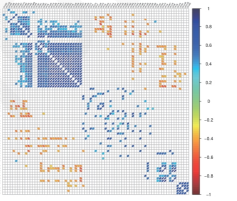


图 4.49.1 物种相关性热图

图中椭圆大小代表相关性系数绝对值大小（相关性绝对值越大，椭圆越小），左斜的为正相关，右斜的为负相关，颜色随右侧色阶变化。图中仅显示 p 值小于 0.05 和相关性绝对值大于 0.8 的结果。
（以 OTU 水平为例）

4.50 共表达网络图

4.50.1 分析方法

物种相关性网络通过计算物种之间的相关性，构建出物种相关性网络，网络图中的节点都是 species-node 节点，当物种与物种之间的相关系数符合某一阈值的时候，物种与物种之间就有一条连线。最后利用图论知识对建立的生物相关性网络进行分析，通过计算网络的节点度分布，网络的直径，网络的平均最短路径，以及节点联通性（Degree），紧密系数（Closeness Centrality），介数中心性（Betweeness Centrality）等属性，来获得物种和样本的组内或组间的相关信息，更加全面高效地挖掘山复杂数据中包含的信息。

软件：**R** 的 **ggraph package**。

4.50.2 结果说明

结果目录：**9_Network/CoNetwork**

***_sparcc_co-occurrence_network_property.xls:** 网络图属性文件。

***_sparcc_co-occurrence_node_infor.xls:** 网络图节点属性文件。

***_sparcc_edge.xls:** 网络图边 edge 信息文件。

***_sparcc_node.xls:** 网络图节点 node 信息文件。

***_sparcc_co-occurrence.graphml:** 网络图 GraphML 格式文件，可用 Cytoscape 等网络图工具打开。

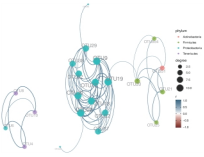


图 4.50.1 物种相关性热图

图中节点的大小表示物种连通度 Degree 大小，不同颜色表示不同的门；连线的颜色表示正负相关性；线的粗细表示相关性系数的大小，线越粗，表示物种之间的相关性越高；线越多，表示该物种与其他物种之间的联系越密切。图中仅显示 p 值小于 0.05 和相关性绝对值大于 0.8。

(以 OTU 水平为例)

4.51 样本相关性热图

4.51.1 分析方法

生物学重复是任何生物学实验所必须的，高通量测序技术也不例外。样本间相关性是检验实验可靠性和样本是否合理性的重要指标。相关系数越接近 1，表明样本之间的相似度越高。

软件：**R** 的 **gplots package**。

4.51.2 结果说明

结果目录：**9_Network/SampleCorr**

***_sample_spearman_correlation.xls:** 样本相关性矩阵。

***_sample_spearman_pvalue.xls:** 样本相关性 P 值矩阵。

***_samplecorrplot_corrplot.pdf:** 样本相关性 Corrplot 图。

***_samplecorr_heatmap_dendrogram.pdf:** 样本相关性热图。

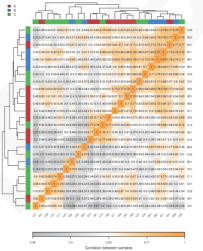


图 4.51.1 样本相关性热图

颜色块代表相关性指数值，颜色越灰表示样本间相关性指数越低，颜色越美则相关性指数越高

4.52 PICRUSt 功能预测分析

4.52.1 分析方法

PICRUSt（Phylogenetic Investigation of Communities by Reconstruction of Unobserved States）是由美国哈佛大学的 Curtis Huttenhower 课题组开发的菌群代谢功能预测工具，通过将现有的 16S rRNA 基因测序数据与代谢功能已知的微生物参考基因组数据库相对比，从而实现对细菌和古菌代谢功能的预测。预测过程中还考虑了不同物种 16S rRNA 基因拷贝数的差异，并对原始数据中的物种丰度数据进行校正，使预测结果更准确可靠。

PICRUSt 分析的总体思路如下：

1) 先根据已测微生物基因组的 16S rRNA 基因全长序列，推断它们的共同祖先的基因功能谱；

2) 对 Greengenes 16S rRNA 基因全长序列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；

3) 将测序得到的 16S rRNA 基因序列数据与 Greengenes 数据库比对，寻找每一条测序序列的“参考序列最近邻居”，并归为参考 OTU；

4) 根据“参考序列最近邻居”的 rRNA 基因拷贝数，对获得的 OTU 丰度矩阵进行校正；

5) 最后，将菌群组成数据“映射”到已知的基因功能谱数据库中，实现对菌群代谢功能的预测。

PICRUSt 能将 16S rRNA 基因序列在 3 种功能谱数据库中进行预测，即 KEGG、COG 和 Rfam。其中，KEGG 数据库的核心为生物代谢通路分析数据库（KEGG PATHWAY Database，http://www.genome.jp/kegg/pathway.html），其中将代谢通路归为 6 大类，包括代谢（Metabolism）、遗传信息处理（Genetic Information Processing）、环境信息处理（Environmental Information Processing）、细胞进程（Cellular Processes）、生物体系统（Organismal Systems）和人类疾病（Human Diseases），每一类代谢通路又被进一步划分为多个等级。目前，第二等级一共包括 45 种代谢通路子功能，第三等级即对应代谢通路图，而第四等级则对应代谢通路上各个 KO（KEGG orthologous groups，KEGG 直系同源基因簇）的具体注释信息。COG（Clusters of Orthologous Groups，https://www.ncbi.nlm.nih.gov/COG/）数据库是由 NCBI 维护的直系同源基因数据库，是指不同个体中由于物种形成（Speciation）的进化历程而产生的的同源基因，这些基因来源于共同祖先；因此，在进化历程中，直系同源基因通常都保留了相同或相似的功能特性。

根据 PICRUSt 的预测结果，可以获得每样本对应于各功能谱数据库的注释信息，以及预测得到的功能类群的丰度矩阵。再根据丰度矩阵绘制热图。

软件：**PICRUSt**。

4.52.2 结果说明

结果目录：**10_Function/PICRUSt prediction/**

cog_nsti.xls: COG 功能预测的 NSTI 值

prediction/cog_predictions.xls: COG 功能预测结果表格

prediction/cog_category_predictions.xls: COG 二级分类预测结果表格

prediction/ko_nsti.xls: KEGG 功能预测的 NSTI 值

prediction/ko_l2_predictions.xls: KEGG 二级分类预测结果表格

prediction/ko_l3_predictions.xls: KEGG 三级分类预测结果表格

prediction/ko_predictions.xls: KEGG 功能预测结果表格

表 4.52.1 功能预测结果表

KO	A11	A12	A21
K01365	0	0	0
K01364	0	0	0
K01361	31	34	72
K01360	1	1586	4
K01362	3873	17464	18226
K02249	1	1	5
K05841	6	10	56
K05844	849	2161	6515
K05845	755	548	977
K05846	1611	1389	2711
K05847	799	3321	1520
K00508	763	2335	2053
K00500	847	474	711
K00507	358	2201	1343

从左至右依次为功能 ID 及其在每个样本的丰度信息。（以 KEGG 功能注释为例）

4.53 BugBase 分析

4.53.1 分析方法

BugBase 是一款分析微生物组样品表型的工具，可对微生物群落根据七类表型进行分类：革兰氏阳性（Gram Positive）、革兰氏阴性（Gram Negative）、生物膜形成（Biofilm Forming）、致病性（Pathogenic）、移动元件（Mobile Element Containing）、氧需求（Oxygen Utilizing，包括 Aerobic、Anaerobic、facultatively anaerobic）及氧化胁迫耐受（Oxidative Stress Tolerant），这些信息可以帮助更好地了解微生物与疾病的关系。

软件：**BugBase**。

4.53.2 结果说明

结果目录：**10_Function/BugBase**

normalized_otus: BugBase 标准化的 OTU 表

thresholds: 包含分析使用的阈值 thresholds_used.txt 和不同阈值下的结果 variances.txt 数据，以及 9 表型在不同阈值下相对丰度变化

otu_contributions: 9 种表型或功能预测结果表 contributing_otus.txt，和 9 种表型按实验组比较的结果堆叠柱状图和物种颜色方案图例 PDF 版。

predicted_phenotypes: 主要有 9 种表型或功能预测结果表 predictions.txt，和 9 种表型按实验组比较箱线图，和相关组间统计信息。

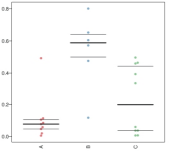


图 4.53.1 表型比较箱线图

横轴为组名，纵轴为相对丰度。

4.54 FAPROTAX 分析

4.54.1 分析方法

FAPROTAX 取词自 Functional Annotation of Prokaryotic Taxa，是 Louca 等人为解析微生物群落功能于 2016 年创建的基于原核微生物分类的功能注释数据库。FAPROTAX 是基于目前对可培养菌的文献资料手动整理的原核功能注释数据库，其包含了收集自 4600 多个原核微生物的 80 多个功能分组（如硝酸盐呼吸、产甲烷、发酵、植物病原等）的 7600 多条功能注释信息。FAPROTAX 预测的功能主要集中在海洋、湖泊中微生物的功能，特别是硫、碳、氢、氮的循环功能。

软件：**FAPROTAX**。

4.54.2 结果说明

结果目录：**10_Function/FAPROTAX/prediction**

FAPROTAX_table.xls: 基于 FAPROTAX 数据库预测菌群的功能表格。

表 4.54.1 FAPROTAX 数据库预测菌群的功能表

group	A11	A12	A21
methanotrophy	0	0	0
acetoclastic methanogenesis	0	0	0
methanogenesis by disproportionation of methyl groups	2	0	10
methanogenesis using formate	0	0	0
methanogenesis by CO2 reduction with H2	5	8	29
methanogenesis by reduction of methyl compounds with H2	1	1	2
hydrogenotrophic methanogenesis	6	9	31
methanogenesis	34	38	73
methanol oxidation	297	720	418
methylotrophy	298	721	420
aerobic ammonia oxidation	5	3	72
aerobic nitrite oxidation	0	0	29
nitrification	5	3	101
sulfate respiration	25	5	18

从左至右依次为功能 ID 及其在每个样本的丰度信息。

4.55 功能丰度热图

4.55.1 分析方法

Heatmap 可以用颜色变化来反映功能的丰度信息，可以直观的将功能丰度值用定义的颜色深浅表来。

软件：**R** 的 **gplots package**。

4.55.2 结果说明

结果目录：**10_Function/PICRUSt/**

Heatmap cog_select.xls: 优势 COG 功能统计表。

cog_heatmap.pdf:

ko_select.xls: 优势 KEGG 功能统计表。

ko_heatmap.pdf: KEGG 功能预测热图。

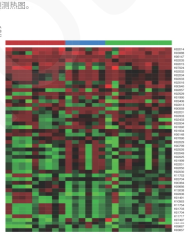


图 4.55.1 功能预测热图

功能丰度热图，用功能丰度矩阵绘制，图中每一列代表一个样本，行代表功能，颜色块代表功能丰度值，颜色越红表示丰度越高，颜色越绿反之丰度越低。

（以 KEGG 功能注释为例）

4.56 功能 PCA 图

4.56.1 分析方法

在多元统计分析中，主成分分析 PCA (Principal Component Analysis) 是一种简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中对方差贡献最大的特征，从而有效地找出数据中最“主要”的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构。

软件：**R**。

4.56.2 结果说明

结果目录：**10_Function/PICRUSt/PCA**

***_pca_axis.xls:** 样本坐标表，样本降维后在各维度（主成分轴）的相对位置，由样本间距离矩阵通过 PCA 分析得来。分析只选取了矩阵特征值排在前两位的坐标数据。

***_pca_rotation.xls:** 功能主成分贡献度表，功能在主成分上的贡献度。

***_pca_importance.xls:** 主成分解释度表，记录了各维度解释结果的百分比。如果 PC1 值为 50%，则表示 x 轴的差异可以解释全面分析结果的 50%。

***_pca.pdf:** PCA 图。

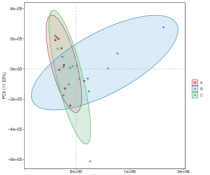


图 4.56.1 功能 PCA 图

横轴和纵轴表示两个选定的主成分轴，百分比表示主成分对样本组成差异的解释度值；横轴和纵轴的刻度是相对距离，无实际意义；不同颜色或形状的点代表不同分组的样本，两样本点越接近，表明两样本物种组成越相似。

（以 KEGG 功能注释为例）

4.57 Procrustes 分析

4.57.1 分析方法

Procrustes 分析（Procrustes analysis）是一种用来分析形状分布的方法。数学上来讲，就是不断迭代，寻找标准形状（canonical shape），并利用最小二乘法寻找每个样本形状到这个标准形状的仿射变化方式。普氏分析可基于不同多元数据集的排序构型（≥2组），通过平移、旋转、缩放等转换方式，实现最大叠合（maximal superimposition），用于不同数据集的对比分析。排序方法可选择 PCA、PCoA 等。

软件：**R** 的 **vegan package**。

4.57.2 结果说明

结果目录：**10_Function/PICRUSt/Procrustes**

***_*_pca_protest_axis.xls:** Procrustes 分析结果表。

***_*_pca_protest.pdf:** Procrustes 分析结果图。

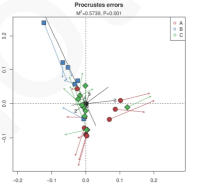


图 4.57.1 Procrustes 分析结果图

图中不同颜色代表不同样本或者不同的 group 中的样本，样本间相似度越高则在图中越聚集。箭头代表物种丰度到功能丰度之间的偏移量。

（以 KEGG 功能注释和 OTU 物种分类 PCA 为例）

5. 参考文献

[1] Martin M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads *EMBnet J.* 17, 10–12. [DOI]

[2] Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics.* 2014;30(5):614-620. [PubMed]

[3] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics.* 2011;27(6):863-864. [PubMed]

[4] Edgar RC. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods.* 2013;10(10):996–998. [PubMed]

[5] Edgar RC. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv.* 2016. [DOI]

[6] Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology.* 2007;73(16):5261-5267. [PubMed]

[7] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 1997;25(17):3389-3402. [PubMed]

[8] Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: Open-Source, PlatformIndependent,Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology.* 2009;75(23):7537-7541. [PubMed]

[9] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research.* 2004;32(5):1792-1797. [PubMed]

[10] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772–780. [PubMed]

[11] Criscuolo, A., Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol.*2010;10:210. [PubMed]

[12] Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, ed. *PLoS ONE.* 2010;5(3):e9490. [PubMed]

[13] Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics.* 2011;12(1):385. [PubMed]

[14] Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution.* 2016;33(6):1635-1638. [PubMed]

[15] Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ.* 2015;3:e1029. [PubMed]

[16] Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics.* 2014;30(21):3123-3124. [PubMed]

[17] Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. *Genome Biology.* 2011;12(6):R60. [PubMed]

[18] Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol.* 2012;8(9):e1002687. [PubMed]

[19] Langille MGI, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology.* 2013;31(9): 814-821. [PubMed]

[20] Ward T, Larson J, Meulemans J, et al. BugBase predicts organism-level microbiome phenotypes. *bioRxiv* 2017; [DOI]

[21] Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. *Science.* 2016;353(6305):1272–1277. [PubMed]

[22] Nguyen, Nhu H., et al. FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology* 2016 Apr 1;20:241-8. [DOI]

[23] McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One.* 2013;8(4):e61217. [PubMed]

[24] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics.* 2019;35(3):526-528. [PubMed]

[25] Rohart F., Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol.* 2017;13(11):e1005752. [PubMed]

[26] Bougeard, S., Dray, S. Supervised Multiblock Analysis in R with the ade4 Package. *Journal of Statistical Software.* 2018;86(1):1-17 [DOI]

[27] Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics.*2011;12:35. [PubMed]

[28] Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics.* 2017;33(18):2938–2940. [PubMed]

[29] Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. *Bioinformatics.* 2014;30(19):2811–2812. [PubMed]

[30] Hamilton, N., Ferry, M. ggtern: Ternary Diagrams Using ggplot2. *Journal of Statistical Software, Code Snippets.* 2018;87(3):1-17 [DOI]

[31] Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial markergene surveys. *Nat Methods.* 2013;10(12):1200–1202. [PubMed]

[32] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNAseq data with DESeq2. *Genome Biol.*2014;15(12):550. [PubMed]

[33] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.*2011;12:77. [PubMed]

[34] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*.2016;13(7):581–583. [PubMed]

[35] Quast C, Pruesse E, Yilmaz P, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*.2013;41(Database issue):D590-D596. [PubMed]

[36] Urmas Kõljalg, R. Henrik Nilsson, Kessy Abarenkov, Leho Tedersoo, et al. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*2013;22,5271–5277. [PubMed]

生工® Sangon Biotech

高通量测序之微生物分类测序手册

▪ 客户使用生工高通量发表的部分相关文献

期刊	Impact Factor (2018)	文章标题
Bioresource Technology	6.669	Freezingthawing pretreatment coupled with biological process of thermophilic Geobacillus sp. G1 Acceleration on waste activated sludge hydrolysis and acidification
The Royal Society of Chemistry	3.538	Decolorization enhancement by optimizing azo dye loading rate in an anaerobic reactor
Scientific Reports	4.011	Effect of electrode position on azo dye removal in an up-flow hybrid anaerobic digestion reactor with built-in bioelectrochemical system
ENERGY & FUELS	3.021	Effects of Ammonia on Anaerobic Digestion of Food Waste Process Performance and Microbial Community
Emerging Microbes & Infections	6.212	Fecal bacterial microbiome diversity in chronic HIV-infected patients in China
Scientific Reports	4.011	The musk chemical composition and microbiota of Chinese forest musk deer males
Bioresource Technology	6.669	Bacterial community and nitrate removal by simultaneous hetero-trophic and autotrophic denitrification in a bioelectrochemically-as-sisted constructed wetland
Bioresource Technology	6.669	Enhancement of microbial nitrogen removal pathway by vegetation in Integrated Vertical-Flow Constructed Wetlands (IVCWs) for treating reclaimed water
Bioresource Technology	6.669	Performance and microbial communities of completely autotrophic denitrification in a bioelectrochemically-assisted constructed wetland system for nitrate removal
Bioresource Technology	6.669	Phosphorus removal performance and biological dephosphorization process in treating reclaimed water by Integrated Vertical-flow Constructed Wetlands (IVCWs)
Bioresource Technology	6.669	Acute and persistent toxicity of Cd(II) to the microbial community of Anammox process
Bioresource Technology	6.669	Full nitration-denitration versus partial nitration-denitration-anammox for treating high-strength ammonium-rich organic wastewater

Food and Chemical Toxicology	3.775	Kiwifruit seed oil prevents obesity by regulating inflammation, thermogenesis, and gut microbiota in high-fat diet-induced obese C57BL6 mice
Frontiers in Microbiology	4.259	Non-isoflavones Diet Incurred Metabolic Modifications Induced by Constipation in Rats via Targeting Gut Microbiota
Chemosphere	5.108	Performance of the nitrogen removal, bioactivity and microbial community responded to elevated nortioxacin antibiotic in an Anammox biofilm system
Biotechnology for Biofuels	5.452	The diversity of hydrogen-producing bacteria and methanogens within an in situ coal seam
Water Research	7.913	Upgrading liquor-making wastewater into medium chain fatty acid: Insights into co-electron donors, key microflora, and energy harvest
Emerging Microbes & Infections	6.212	2018-Changes in intestinal microbiota in HIV 1 infected subjects-following cART initiation influence of CD4 T cell count
Bioresource Technology	6.669	Anaerobic bioaugmentation hydrolysis of selected nitrogen heterocyclic compound in coal gasification wastewater
Bioresource Technology	6.669	Effects of a pulsed electric field on nitrogen removal through the ANAMMOX process at room temperature
Journal of Environmental Management	4.865	Effcient nitrogen removal from synthetic domestic wastewater in a novel step-feed three-stage integrated anoxicoxic biological aerated alter process through optimizing influent flow distribution
Water Research	7.913	Field tests of cubic-meter scale microbial electrochemical system in a municipal wastewater treatment plant
Bioresource Technology	6.669	Mechanism of process imbalance of long-term anaerobic digestion of food waste and role of trace elements in maintaining anaerobic process stability
Bioresource Technology	6.669	Specific quorum sensing signal molecules inducing the social behaviors of microbial populations in anaerobic digestion
Chemical Engineering Journal	8.355	Synergistic degradation on aromatic cyclic organics of coal pyrolysis wastewater by lignite activated coke-active sludge process

Bioresource Technology	6.669	Synergistic degradation on phenolic compounds of coal pyrolysis wastewater (CPW) by lignite activated coke-active sludge (LAC-AS) processInsights into succession of microbial community unde
Science of the Total Environment	5.589	Enhanced anaerobic degradation of selected nitrogen heterocyclic compounds with the assistance of carboxymethyl cellulose
Bioresource Technology	6.669	Enhanced biodegradation of coal gasification wastewater with anaerobic biofilm on polyurethane (PU), powdered activated carbon (PAC), and biochar
Bioresource Technology	6.669	Comparative investigation on carbon-based moving bed biofilm reactor (MBBR) for synchronous removal of phenols and ammonia in treating coal pyrolysis wastewater at pilot-scale
Bioresource Technology	6.669	Enhanced biofilm formation and denitrification in biofilters for advanced nitrogen removal by rhamnolipid addition
animals	1.832	Effect of Dietary Supplementation of Lactobacillus Casei YYL3 and L. Plantarum YYL5 on Growth, Immune Response and Intestinal Microbiota in Channel Catfish

▪ 写作模板

- Materials and methods
- Sample collection

客户自行描述

- DNA extraction

Total community genomic DNA extraction was performed using a E.Z.N.A.Soil DNA Kit (Omega, USA),following the manufacturer’ s instructions. We measured the concentration of the DNA using a Qubit 2.0(life,USA) to ensure that adequate amounts of high-quality genomic DNA had been extracted .

- 16S rRNA gene amplification by PCR

Our target was the V3–V4 hypervariable region of the bacterial 16S rRNA gene. PCR was started immediately after the DNA was extracted.The 16S rRNA V3–V4 amplicon was amplified using KAPA HiFi Hot Start Ready Mix (2×) (TaKaRa Bio Inc., Japan). Two universal bacterial 16S rRNA gene amplicon PCR primers (PAGE purified) were used: the amplicon PCR forward primer(CCTACGGGNGGCWGCAG) and amplicon PCR reverse primer (GACTACHVGGGTATCTAATCC). The reaction was set up as follows: microbial DNA (10 ng/μl) 2 μl; amplicon PCR forward primer (10 μM) 1 ; amplicon PCR reverse primer (10 μM) 1 μl; 2X KAPA HiFi Hot Start Ready Mix 15 μl (total 30 μl). The plate was sealed and PCR performed in a thermal instrument (Applied Biosystems 9700, USA) using the following program: 1 cycle of denaturing at 95 ° C for 3 min, first 5 cycles of denaturing at 95° C for 30 s, annealing at 45° C for 30 s, elongation at 72° C for 30 s, then 20 cycles of denaturing at 95° C for 30 s, annealing at 55° C for 30 s, elongation at 72° C for 30 s and a final extension at 72° C for 5 min. The PCR products were checked using electrophoresis in 1 % (w/v) agarose gels in TBE buffer (Tris, boric acid, EDTA) stained with ethidium bromide (EB) and visualized under UV light.

- 16S gene library construction,quantification,and sequencing

We used AMPure XP beads to purify the free primers and primer dimer species in the amplicon product.Samples were delivered to Sangon BioTech (shanghai) for library construction using universal Illumina adaptor and index. Before sequencing, the DNA concentration of each PCR product was determined using a Qubit® 2.0 Green double-stranded DNA assay and it was quality controlled using a bioanalyzer (Agilent 2100, USA). Depending on coverage needs, all libraries can be pooled for one run.The amplicons from each reaction mixture were pooled in equimolar ratios based on their concentration. Sequencing was performed using the Illumina MiSeq system (Illumina MiSeq, USA), according to the manufacturer’ s instructions.

- Sequence processing

After sequencing, data were collected as follows:(1)The two short Illumina readings were assembled by PEAR(v0.9.6) software according to the overlap and fastq files were processed to generate individual fasta and qual files, which could then be analyzed by standard methods.(2)Sequences containing ambiguous bases and any longer than 480 base pairs (bp) were dislodged and those with a maximum homopolymer length of 6 bp were allowed (Köchling et al. 2015).And sequence short than 200bp were removed.(3)All identical sequences were merged into one.(4)Sequences were aligned according to a customized reference database.(5) The completeness of the index and the adaptor was checked and removed all of the index and the adaptor sequence.(6) Noise was removed using the Pre.cluster tool.Chimeras were detected by using Chimera UCHIME. All the software was in the mothur package. We submitted the effective sequences of each sample to the RDP Classifier again to identify archaeal and bacterial sequences. Species richness and diversity statistics including coverage, chao1, ace,simpson, and shannonever were also calculated using mothur.The modified pipeline is described on the mothur website. Finally, all effective bacterial sequences without primers were submitted for downstream analysis (Kozich et al. 2013).

联系我们

总 部
地址：上海市松江区香闵路 698 号
邮编：201611
电话（总机）：400-821-0268，021-37772168
传真：400-821-0268 按 9
Email: sales@sangon.com

投诉与建议
电话：400-821-0268 按 3
Email: mbts@sangon.com

合成测序服务网点

地区	引物合成网点联系方式	测序网点联系方式
上海	电话：021-57072171/72/73/74 邮箱：synth@sangon.com	电话：021-57072160/61/62 邮箱：shseq@sangon.com
北京	电话：010-81767585/86 传真：010-81767586 邮箱：beijing@sangon.com	电话：010-81767529/79 邮箱：bjseq@sangon.com
武汉	电话：027-65522298 邮箱：whsynth@sangon.com	电话：027-87002907 邮箱：whseq@sangon.com
广州	电话：020-38452026 传真：020-32207701 邮箱：gz_synth@sangon.com	电话：020-38455693/38452693 邮箱：gzseq@sangon.com
成都	电话：028-64259944 邮箱：cdsynth@sangon.com	电话：028-64259946 邮箱：cdseq@sangon.com
南京	电话：025-85383702 邮箱：njsynth@sangon.com	电话：025-85383701 邮箱：njseq@sangon.com
郑州	电话：0371-63313093 或 0371-61652655 邮箱：zzsynth@sangon.com	电话：0371-61171352 邮箱：zzseq@sangon.com
青岛	电话：0532-68012226 邮箱：qdsynth@sangon.com	电话：0532-68012178 邮箱：qdseq@sangon.com
昆明		电话：15021124412 邮箱：kmseq@sangon.com
长春		电话：13636536956 邮箱：ccseq@sangon.com
西安		电话：021-57072160/18602066713 邮箱：xaseq@sangon.com
长沙		电话：18373141571 邮箱：csseq@sangon.com
杭州		电话：021-57072160 / 61 / 62 邮箱：hzseq@sangon.com
厦门		电话：17602185336 邮箱：xmseq@sangon.com

生工生物全国销售网点联系方式 (按省份首字母排列)

省份	网点	手机	办公电话	邮箱	传真
安徽	合肥	18917713933	0551-65428048	anhui@sangon.com	0551-65428048
北京	北京	18917713568	010-82363780	bjorder@sangon.com	010-82363790
福建	福州	18917713433		fuzhou@sangon.com	
	厦门	18917713608	0592-2181892	xiamen@sangon.com	
甘肃	兰州	18917713838	0931-8310565	lanzhou@sangon.com	0931-8310565
广东	广州	13318787115	020-32206684	guangzhou@sangon.com	020-32207701
	深圳	18917713663	0755-86011411	shenzhen@sangon.com	0755-86011411
广西	桂林	18917713348		guilin@sangon.com	
	南宁	18917713345	0771-3821595	nanning@sangon.com	
贵州	贵阳	18917714277		guiyang@sangon.com	
海南	海口	13111904256	0898-66862960	haikou@sangon.com	
河北	石家庄	18917713638	0311-85046606	shijiazhuang@sangon.com	
河南	郑州	18917713330	0371-56690353	zhengzhou@sangon.com	
黑龙江	哈尔滨	18917713822	0451-83331061	haerbin@sangon.com	
湖北	武汉	18917713883		wuhan@sangon.com	
湖南	长沙	17752882112	0731-84556676	changsha@sangon.com	
吉林	长春	18917710639	0431-88541636	changchun@sangon.com	0431-88541636
江苏	南京	18917713993	025-86667569	nanjing@sangon.com	
	苏州	13616278094		suzhou@sangon.com	021-37772170
	无锡	18917714878		hemin@sangon.com	
	徐州	13953192492	0531-82951640	xuzhou@sangon.com	0531-82941640
	扬州	18917713633		yangzhou@sangon.com	
江西	南昌	18917713866	0791-86853779	nanchang@sangon.com	
辽宁	大连	18917713477	0411-39759235	dalian@sangon.com	
	沈阳		024-23412941	shenyang@sangon.com	
内蒙古	呼和浩特	18917713400	0471-2250562	neimenggu@sangon.com	0471-2250562
青海	西宁	18917713848	0971-8814295	qinghai@sangon.com	
山东	济南	13953192492	0531-82951640	jinan@sangon.com	0531-82941640
	青岛	18053256633	0532-82716990	qingdao@sangon.com	
山西	太原	18917713299		taiyuan@sangon.com	
陕西	西安	18917713699	029-82497082	xian@sangon.com	
上海	上海	18917713773	021-64746299	shanghai@sangon.com	
		18917713798			
四川	成都	18180498155	028-87434681	chengdu@sangon.com	
天津	天津	18917713466	022-27460687	tianjin@sangon.com	
新疆	乌鲁木齐	18917713877	0991-4338172	wulumuqi@sangon.com	
云南	昆明	18917713411	0871-65170776	kunming@sangon.com	
浙江	杭州	18917713636	0571-88497358	hangzhou@sangon.com	
	宁波	15267878330		ningbo@sangon.com	
	温州	18917713948		wenzhou@sangon.com	
重庆	重庆	18917713833	023-81363286	chongqing@sangon.com	

网上订购，更多产品信息，请点击 www.sangon.com