项目名称：16srRNA sequencing data report

目录

1.项目整体流程概况

项目信息

项目编号	MD2022040717 25W27R
合同号	Mbnj202204018
开题单号	MbPA202204379
项目名称	中医药25个多样性云平台分析服务合同
项目类别	微生物群落多样性组成谱分析
项目类型	标准分析
完成日期	2022-04-29 13:44:54

分析流程

1) 首先对高通量测序的原始下机数据根据序列质量进行初步筛查；对问题样本进行重测、补测。

2) 通过质量初筛的原始序列按照index和Barcode信息，进行文库和样本划分，并去除barcode序列。

3) 按照QIIME2 dada2分析流程或Vsearch软件的分析流程进行序列去噪或OTU聚类。

4) 对各样本(组)在不同物种分类学水平的具体组成进行展示，了解整体概况。

5) 根据ASV/OTU在不同样本中的分布，评估每个样本的Alpha多样性水平，并通过稀疏曲线反映测序深度是否合适。

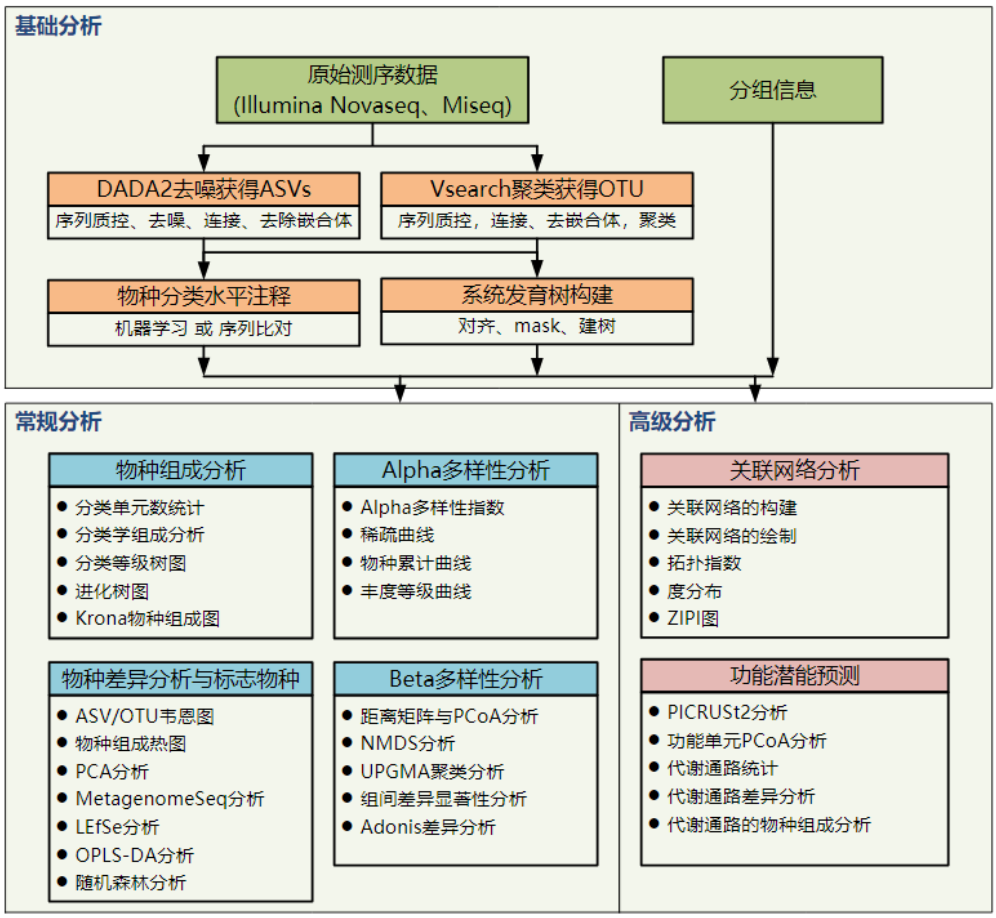
6) 在ASV/OTU层面，计算各样本的距离矩阵，并通过多种非监督的排序、聚类手段，结合相应统计学检验方法，衡量不同样本(组)间的beta多样性差异 及差异显著性。

7) 在物种分类学组成层面，通过各种非监督、监督的排序、聚类和建模手段，结合相应统计学检验方法，进一步衡量不同样本(组)间的物种丰度组成差异， 并尝试寻找标志物种。

8) 根据物种在各样本中的组成分布，构建关联网络，计算拓扑指数，并尝试找出关键物种。

9) 根据16S rRNA、18S rRNA和ITS基因测序结果，还可预测样本的菌群代谢功能，找出差异通路，并获得特定通路的物种组成。

10) 在以上结果的基础上，绘制具备论文发表水准的图表，并进行统计检验分析，分析结果文档可参见附录。


基础分析
原始测序数据
(Illumina Novaseq、Miseq)
分组信息
DADA2去噪获得ASVs
序列质控、去噪、连接、去除嵌合体
Vsearch聚类获得OTU
序列质控，连接、去嵌合体，聚类
物种分类水平注释
机器学习 或 序列比对
系统发育树构建
对齐、mask、建树
常规分析
物种组成分析
分类单元数统计
分类学组成分析
分类等级树图
进化树图
Krona物种组成图
Alpha多样性分析
Alpha多样性指数
稀疏曲线
物种累计曲线
丰度等级曲线
物种差异分析与标志物种
ASV/OTU韦恩图
物种组成热图
PCA分析
MetagenomeSeq分析
LEfSe分析
OPLS-DA分析
随机森林分析
Beta多样性分析
距离矩阵与PCoA分析
NMDS分析
UPGMA聚类分析
组间差异显著性分析
Adonis差异分析
高级分析
关联网络分析
关联网络的构建
关联网络的绘制
拓扑指数
度分布
ZIPI图
功能潜能预测
PICRUSt2分析
功能单元PCoA分析
代谢通路统计
代谢通路差异分析
代谢通路的物种组成分析

2.项目分析结果

2.1 序列处理分析

序列去噪或聚类

此步骤主要有DADA2和Vsearch两种方法可选。

DADA2方法(Benjamin et al., 2016)主要进行去引物，质量过滤，去噪(denoise)，拼接和去嵌合体等步骤。它不再以相似度聚类，只进行去重(dereplication)或者说相当于以100%相似度聚类。使用DADA2质控后产生的每个去重的序列称为ASVs (amplicon sequence variants)，或称为特征序列(对应于OTU代表序列)，而这些序列在样本中的丰度表称为特征表(对应于OTU表)。以DADA2为代表的去噪生成特征序列的方法是目前主流分析平台(QIIME2和USEARCH)所力推的。在QIIME2中有这样的一段描述“The features produced by clustering methods are known as operational taxonomic units (OTUs), which is Esperanto for suboptimal, imprecise rubbish.”，认为以OTUs聚类为基础建立的分析方法是不理想、不准确的https://docs.qiime2.org/2019.7/tutorials/overview/)。因此默认选择DADA2进行分析。

尽管如此，以上方法目前尚未实现与所有的扩增子类型适配，因此我们依旧保留了基于OTU聚类的Vsearch (Rognes et al., 2016)方法作为备选。Vsearch方法主要包括去引物，拼接，质量过滤，去重，去嵌合体，聚类等步骤。Vsearch软件是专门针对独立学者Edgar开发的 USEARCH (Edgar et al., 2011)软件所研发的一款开源的64位免费分析软件。在Vsearch文章中，作者提出该软件的聚类和去嵌合体准确率均优于USEARCH的uparse算法。功能基因项目默认选择Vsearch方法进行分析。

分析结果:

每样本测序量统计表

SampleID	Input	Filtered	Denoised	Merged	Non-chimeric	Non-sing
Con1	116853	110081	106875	89723	53292	5167
Con2	106618	100078	98117	88424	61272	6031
Con3	93329	87176	84738	72437	49980	4886
Con4	92515	86866	84879	74310	50795	4986
Con5	96654	91019	88002	72119	47707	4628
HF1	130183	122955	119438	98377	59710	5703
HF2	120387	113684	110125	90013	55961	5376
HF3	129263	120443	116864	96205	58486	5647

SampleID	Input	Filtered	Denoised	Merged	Non-chimeric	Non-sing
HF4	124028	116161	112613	89689	58642	5578
HF5	124881	117253	113510	88969	57477	5480

1. dada2: 表中第一列为样本的ID；第二列为原始数据中能同时匹配到正向和反向引物的序列量；第三列为去除低质量序列后的数据量；第四列为去噪后的序列数据量，即有效序列量；第五列为拼接后的序列量，第六列为去除嵌合体后序列量，即为高质量序列量；第七列为去除singleton后的序列量。
2. vsearch: 表中第一列为样本的ID；第二列为原始数据中能同时匹配到正向和反向引物的序列量；第三列为拼接后的序列量；第四列为去除低质量序列后的数据量；第五列为聚类后去除嵌合体后序列量，即为高质量序列量；使用了FrameBot的功能基因项目，第六列为FrameBot校正后的序列量；最后一列为去除singleton后的序列量。

序列长度分布

获得了ASV特征序列或OTU代表序列之后，我们可以对其长度分布进行统计，以便检查这些序列的长度是否和测序目的片段的长度范围相当，是否存在异常长度的序列等。

分析软件：自编perl脚本。

分析步骤：对样本中所包含的高质量序列的长度分布进行统计。注意，这里的统计包含了singleton序列。

物种分类学注释

物种的注释过程，其本质是与参考序列数据库进行比对，以及对比对结果进行打分判定的过程。因此，数据库的选择及其重要，一个好的物种注释数据库应该尽可能涵盖所有待测的序列的物种，同时，尽可能减少其他非待测的物种；这样的好处在于一方面提高真阳性，一方面降低假阳性，这相当于提高了物种注释分辨率。

值得注意的是，由于微生物种类繁多，而目前的数据库中的微生物序列还无法完全覆盖测序序列(以unassigned标记)，又或是参考序列缺少准确的物种信息(unidentified, uncultivated, uncultured, incertae sedis)；加之测序读长的限制，一些具体的种、属之间无法鉴别分开(unclassified)。因此，在实际分析过程中，并非所有特征序列都能获得种、属水平的注释信息。

分析软件： QIIME2 (2019.4)

数据库：

1）对于细菌或古菌的16S rRNA基因，默认选用Greengenes数据库（Rlease 13.8,http://greengenes.secondgenome.com/)(DeSantis et al,2006),也可选用Silva数据库(Release132,http://www.arb-silva.de(Quast et al., 2013)；

2) 对于真核微生物18S rRNA基因，默认选用本地化的nt(2019.8下载，ftp://ftp.ncbi.nih.gov/blast/db/)数据库，也可选用Silva数据库(Release132)；

3) 对于真菌ITS序列的， 默认选用UNITE数据库(Release 8.0，https://unite.ut.ee/)(Koljalg et al., 2013)；

4)对于功能基因或其他需求， 我们使用本地化的nt或nr(2019.8下载，ftp://ftp.ncbi.nih.gov/blast/db/)数据库进行注释。

分析步骤：

1. 对于前三类数据库，采用QIIME2的classify-sklearn算法(Bokulich et al., 2018) (https://github.com/QIIME2/q2-feature-classifier)：对于每个ASVs的特征序列或每个OTU的代表序列，在QIIME2软件中使用默认参数，使用预先训练好的Naive Bayes分类器进物种注释。

2. 对于nt或nr数据库，采用BROCC算法(Nilsson et al., 2006) (https://github.com/kylebittinger/q2-brocc#the-brocc-algorithm)：首先使用blastn或balstx，将序列与nt或nr数据库中的核酸或蛋白序列进行比对；再调用brocc.py脚本，依据推荐的参数获取注释信息。需要注意的是，根据上文提及的数据库选择的原则，这里的nt或nr数据库一般都不是指整个数据库，而是根据测序目的物种和/或基因名称的Accession号或Gi号进行了限定的数据库子集。

构建系统发育树

微生物多样性组成谱分析中的许多分析都会用到系统发育树，因此，我们在获得了ASV特征序列或OTU代表序列之后，需要构建这些序列的系统发育树，以获得序列间的遗传距离或亲缘关系。构建系统发育树的方法很多，但基本可以归结为以下几类：UPGMA法，邻接法（neighbor joining），最大简约法（most parsimonious），最大似然法（Maximum Likelihood）以及贝叶斯法（Bayesian inference）。其中又以后两者较常见。这里默认采用最大似然法中的FastTree进行构建，此外，qiime2也支IQtree和RaxML的方法。

需要注意的是，对于ITS序列，由于它在多序列对齐(multiple sequence aligments)时的质量极差，若直接de-novo构建系统发育，其效果也不佳；对于部分功能基因序列，由于其核酸水平的序列差异性较大，不适合核酸水平的de-novo建树，因此，我们默认不对ITS序列、功能基因序列进行建树。

分析软件：QIIME2 (2019.4)

分析步骤：使用“qiime phylogeny align-to-tree-mafft-fasttree”的分析流程，调用mafft (Katoh, 2002)进行多序列对齐，并mask无系统发育信息的部分，再调用FastTree (Price et al., 2009)构建了系统发育树，生成无根树文件；再以中心点基础构建有根树文件。

ASV/OTU表抽平

前面的分析步骤中，已经产生了ASV/OTU的丰度表，而后续的部分分析步骤需要各样本在同一测序深度水平下进行，因此需要对该表格进行一定的转化处理。可用稀疏(Rarefaction)的方法，它通过从每个样本中分别随机抽取一定数量的序列以到达统一的深度，从而预测各样本在该测序深度下，所能观测到的ASVs或OTUs及其相对丰度(Heck et al., 1975; Kemp and Aller, 2004)，因此，这一过程也称为抽平。

分析软件：QIIME2 (2019.4)

分析步骤：使用qiime feature-table rarefy功能，抽平深度设为最低样本序列量的95%。

2.2 物种组成分析

分类单元数统计

通过对抽平后的ASV/OTU表格进行统计，可以获得每个样本中的微生物群落在各分类水平的具体组成表。通过该表，首先可以计算不同样本在各分类水平所含有的分类单元的数目。

分析软件：自编perl脚本。

分析步骤：依据序列物种分类学注释的结果以及选择的样品，统计这些样本的物种注释结果中域、门、纲、目、科、属、种七个分类水平各自含有的分类单元的数量。注意，统计的数目中已去除了包含unclassifed，uncultured，uncultivated，unknown，metagenome等字段的分类单元。

分类学组成分析

物种组成的堆叠柱状图或简称柱状图是最常用的表征多样本物种组成情况的手段。通过对去除singleton后的特征表进行统计，实现各样本在门、纲、目、科、属、种六个分类水平上的组成分布的可视化，并以柱状图呈现分析结果。

分析软件：QIIME2 (2019.4)；自编perl脚本等。

分析步骤：调用“qiime taxa barplot”命令，进行绘图。

分类等级树图

以上方法都是在分别选取一个分类水平展示相应分类单元的组成情况，那么如何才能同时展示所有分类单元的组成情况呢?

最早，我们采用层级树图配合饼图的形式展示每个水平下的组成情况（详见线下旧版报告），这种形式对于物种组成较为复杂的项目，往往无法通过有限的画幅展示样本的分类学全貌。

近来，我们关注到了Science杂志上一篇题为“Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome”的文章(Carrión etal., 2019)，它的第一幅主图使用了圈堆积图(circle packing chart)的形式来展示微生物群落的分类学构成，非常直观且生动，试想再辅以饼图，就可以同时展示不同分类学单元在不同分组中的组成比例了。

分析软件：R语言，ggraph，ggplot2包等。

分析步骤：以树图(Treemap; 圈堆积图是其中一种具体形式)的形式，绘制微生物分类等级树并将每个ASV或OTU的分组的丰度数据以饼图的形式添加到图中。此外，也可选定具体的水平学水平，通过对ASV或OTU圆点的不同着色来强调该水平下微生物的分类学组成情况。

进化树图

除了通过树图的形式来展示分类学组成形成，还可以使用ggtree作进化树来展示各ASV/OTU在进化树中的位置,以及相互间的进化距离，并通过热图和柱状图等反映它们的组成与丰度、分类学等信息。

分析软件：R语言，ggtree，phyloseq等R包。

分析步骤： 我们推荐在进行绘图之前，先对ASV特征序列或OTU代表序列进行合并、过滤以简化图形，主要分三种方式：1) 按照分类单元进行合并（采用R语言phyloseq包tax_glom功能），即具有相同分类单元注释的tips合并为一个tip（选择其中一个tip作为代表）；2) 按照进化距离进行合并(采用R语言phyloseq包tip_glom功能)，即进化距离小于一定阈值的tips合并为一个tip（选择其中一个tip作为代表）；3) 按照丰度对物种进行排序，在进化树图中保留排名前n的物种。当然，也可以不合并分支直接作图。相关绘图方法可参考ggtree的绘图说明文档（https://yulab-smu.github.io/treedata-book）。

Krona物种组成图

此外，还可以使用Krona软件(https://github.com/marbl/Krona/wiki)进行群落分类学组成的交互展示(Ondov et al., 2011)。

2.3 Alpha多样性分析

Alpha多样性指数

Alpha多样性指数生态学家使用alpha多样性和beta多样性指数，分别表征物种在生境内和生境间的多样性，以综合评价其总体多样性(Whittaker，1972；Whittaker，1960)。这里我们先分析alpha多样性。alpha多样性是指局部均匀生境下的物种在丰富度(richness)、多样性(diversity)和均匀度(evenness)等方面的指标，也被称为生境内多样性(within-habitat diversity)。

为了能较为全面的评估微生物群落的alpha多样，本流程以Chao1 (Chao, 1984)和Observed species指数表征丰富度，以Shannon (Shannon, 1948a, b)和Simpson (Simpson, 1949)指数表征多样性，以Faith’s PD (Faith, 1992)指数表征基于进化的多样性，以Pielou’s evenness (Pielou, 1966)指数表征均匀度，以Good’s coverage (Good, 1953)指数表征覆盖度。对这些alpha多样性指数的计算方法可参见附表，或详见http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html#module-skbio.diversity.alpha。

分析软件：QIIME2 (2019.4)，R语言，ggplot2包

分析步骤：使用未抽平的ASV/OTU表，调用“qiime diversity alpha-rarefaction”命令，设置参数“--p-steps 10 --p-min-depth 10 --p-iterations 10”，即最小抽平深度为10，参数“--p-max-depth”设为全体样本中最低测序深度样本序列量的95%，再在这一深度与最小深度之间均匀选取10个深度值，每个深度值抽平10次，计算所选的alpha多样性指数。选取最大抽平深度时的得分平均值作为alpha多样性指数。

对于有分组的样本(每组样本数≥3)，可使用R脚本将上表中数据绘制成箱线图，以直观地展示不同样本组之间的alpha多样性差异，并可通过Kruskal-Wallis秩和检验和dunn’test作为事后检验，验证差异的显著性(两组样本时Kruskal-Wallis检验与wilcoxon检验等价)。

分析结果:

指数表

Sample	Chao1	Faith_pd	Goods_coverage	Observed_species	Pielou_e
B1	4075.44	187.266	0.976741	3616.4	0.81019
B2	4231.56	188.697	0.975856	3788.5	0.809007
B3	4294.58	161.644	0.975192	3844.9	0.790415
B4	3451.49	160.62	0.982175	3146.3	0.773015
B5	4491.07	177.636	0.973972	3983.4	0.805588
Con1	2730.64	160.762	0.986556	2525.4	0.74545
Con2	1935.63	107.013	0.988939	1673.9	0.567071

Sample	Chao1	Faith_pd	Goods_coverage	Observed_species	Pielou_e
Con3	2341.04	143.575	0.988747	2144.3	0.714993
Con4	2021.28	129.765	0.990003	1842.7	0.672376
Con5	2576.29	164.819	0.988747	2425.1	0.769006

1. 指数表：表中第一列为样本名，之后各列分别对应每个样本在相应测序深度下的alpha多样性指数的计算结果。
2. 事后检验表：行列名为分组名，其间单元格中的数值为相应的两个组之间事后检验的显著性。

稀疏曲线

显然，alpha多样性指数的大小是与使用的ASV/OTU表的抽平深度有关，为了探究样本alpha多样性随抽平深度的变化趋势，可绘制稀疏曲线(Rarefaction Curve)。稀疏曲线是生态学领域的一种常用方法，通过从每个样本中随机抽取一定数量的序列(即在不超过现有样本测序量的某个深度下进行重抽样)，可以预测样本在一系列给定的测序深度下，所可能包含的物种总数及其中每个物种的相对丰度(Heck et al., 1975; Kemp and Aller, 2004)。因此，通过绘制稀疏曲线，还可以在相同的测序深度下，比较不同样本中ASV/OTU数的多少，从而在一定程度上衡量每个样本的多样性高低。

分析软件：QIIME2 (2019.4)

在2.3.1分析过程中由“qiime diversity alpha-rarefaction”命令产生alpha-rarefaction.qzv文件(关于QZV文件的介绍与使用请见附录)。通过将该文件拖入https://view.qiime2.org/，可实现可视化。

物种积累曲线

物种累积曲线(Species accumulation curves)与稀疏曲线类似，用于衡量和预测群落中物种丰富度随样本量扩大而增加的幅度，被广泛用于判断样本量是否足够并估计群落丰富度(Chao and Shen, 2004)。一般而言，在样本量较少时，随着新样本的加入，将有较大可能性发现大量的新物种，此时曲线将呈现急剧上升的形态；当样本量已经较大时，此时群落中的ASV/OTU总数将不再随着新样本的加入而显著增加，曲线也将趋于平缓。因此，可以利用物种累积曲线判断样本量是否足够大：若曲线始终保持上升趋势，则表明样本量不足，需要扩大采样规模；反之，则表明样本量已足以反映群落的物种组成。

使用R脚本，对ASV/OTU丰度矩阵中每个样本所对应的ASV/OTU总数绘制Specaccum物种累积曲线。

丰度等级曲线

与稀疏曲线不同，丰度等级曲线(Rank abundance curve)将每个样本/分组中的ASV/OTU按其丰度大小沿横坐标依次排列，并以各自的丰度值为纵坐标，用折线或曲线将各ASV/OTU互相连接，从而反映各样本中ASV/OTU丰度的分布规律(详见https://en.wikipedia.org/wiki/Rank_abundance_curve)。对于微生物群落样本，该曲线可以直观地反映群落中高丰度和稀有ASV/OTU的数量。

将各样本/各分组的ASV/OTU按其丰度从大到小沿横坐标依次排列后，将丰度值经Log2对数转换（Log10转化, 百分比转化或不转化）后的值作为纵坐标，在R软件中编写脚本绘制各样本或各分组的丰度等级曲线。

2.4 Beta多样性分析

距离矩阵与PCoA分析

Beta多样性指数聚焦于不同生境间多样性的比较，也就是样本间的差异。事实上，每一个物种(ASV/OTU)在两个样本之间差异，即是反映这两个样本间群落差异的一个维度。而由于群落中物种的数量往往非常巨大，样本间群落的差异往往就会是多维度的，进而难以进行直接比较。此时，就需要相关算法对这个多维数据进行降维。不同的Beta多样性距离，其实就是采用了不同的算法(利用样本中物种(ASV/OTU)的丰度信息，特征/代表序列间的进化关系等)，来将多维度的物种数据降成一维数据——样本差异距离，从而在不同的角度表征这两个样本间的群落差异。常用的beta多样性距离如Jaccard距离(Jaccard，1908)，Bray-Curtis距离(Bray and Curtis, 1957)，unweighted UniFrac距离(Lozupone and Knight, 2005)和weighted UniFrac距离(Lozupone and Knight, 2005)等。对这些距离的介绍可参见附表。

每一对样本都可以计算这样的差异距离，便形成了样本差异距离矩阵(distance matrix)。随着样本数的增加，这个距离矩阵的维度也不断增大。这时就需要使用排序(ordination)的方法来将这些样本在一个可视的低维空间(通常是二维)重新排列，使得空间内样本点之间的距离能够最大程度地反映距离矩阵中的样本差异距离。常见的排序方法包括非约束排序和约束排序(典范排序)。

主坐标分析(Principal coordinates analysis，PCoA)便是一种最经典的非约束排序(Classical Multidimensional Scaling，cMDScale)分析方法(Ramette, 2007)。它通过将样本距离矩阵经过投影后，在低维度空间进行展开，并最大限度地保留原始样本的距离关系。PCoA以样本距离为整体考虑，相比于主成分分析(Principal components analysis，PCA)，更符合生态学数据特征，因此作为排序分析手段，更为推荐使用。

分析软件：QIIME2 (2019.4)； R语言，ape包等

分析步骤：使用抽平后的ASV/OTU表，依据树文件的有无调用“qiime diversity core-metrics-phylogenetic”或“qiime diversity core-metrics”命令，计算Jaccard，Bray-Curtis，unweighted UniFrac和weighted UniFrac等四种距离矩阵或Jaccard和Bray-Curtis等两种距离矩阵，并对这些距离矩阵做 PCoA分析，输出QZV文件。将QZV文件拖入https://view.qiime2.org/相应区域可实现可视化。同时，可使用R脚本在R中进行PCoA分析输出样本点的PCoA坐标，并将其绘制成二维散点图。

NMDS分析

非量度多维尺度分析(NMDS)与上述PCoA分析类似，也是通过对样本距离矩阵作降维分解，简化数据结构，从而在特定距离尺度下描述样本的分布特征。与PCoA分析不同，NMDS分析不依赖于特征根和特征向量的计算，而是通过对样本距离进行等级排序，使样本在低维空间中的排序尽可能符合彼此之间的相似距离的远近关系(而非确切的距离数值)。因此，NMDS分析不受样本距离的数值影响，仅考虑彼此之间的大小关系，对于结构复杂的数据，排序结果可能更稳定。NMDS结果的应力值(Stress)越小越好，一般认为当该值小于0.2时，NMDS分析的结果较可靠(Legendre, 1998)。

分析软件：R语言，vegan包等。

分析步骤: 使用R脚本对4.1节分析所获得的bray-curtis距离矩阵（默认）进行NMDS分析，并通过二维排序图展示微生物群落的组成差异。

层次聚类分析

如果说排序(ordination)的目的在于寻找数据的连续性(通过连续的排序轴展示数据的主要趋势)，那么聚类分析的目的则是在于寻找数据的间断性。Beta多样性聚类分析中多采用层次聚类(Hierarchical clustering)的分析方法，以等级树的形式展示样本间的相似度，通过聚类树的分枝长度衡量聚类效果的好坏。与排序分析相同，聚类分析可以采用任何距离评价样本之间的相似度。常用的聚类分析方法包括非加权组平均法(Unweighted pair-group method with arithmetic means，UPGMA)、单一连接法(Single-linkage clustering)和完全连接法(Complete-linkage clustering)等等。

分析软件：R语言，vegan，ape，ggtree包等

分析步骤：使用R语言stat包的uclust函数，默认对bray-curtis距离矩阵采用UPGMA算法（即聚类方法为average）进行聚类分析，并使用R脚本ggtree包进行可视化。

组间差异分析

PCoA、NMDS等排序分析只是一种探索分析手段，并不是统计检验；对于排序图中呈现的分布规律，又或是聚类分析中所呈现的分组规律，我们需要使用检验手段进行验证。常用的有PERMANOVA，anosim和permdisp等，这里做简要介绍如下：

PERMANOVA (Permutational multivariate analysis of variance)分析(McArdle and Anderson, 2001)是基于置换检验的多元方差分析，它假设组内的样本比与组间的样本更相似。换句话说，它测试每个组的组内样本距离是否与组间的样本距离不同。通常使用R语言vegan包的adonis函数或adnois2函数调用该算法，因此有时也直接称为adonis分析。

Anosim (Analysis of similarities)分析(Clarke, 1993; Warton et al., 2012)。是一种判别两组或多组样本之间是否存在显著差异的非参数检验方法。该方法认为如果两组样本在物种组成上确实不同，那么这些组之间的组成差异应该大于组内的不同。检验统计量R取值在-1和1之间，小于0时表示组内差异大于组间差异，大于0时表示组间差异大于组内差异。该方法使用差异矩阵中不同差异值的排名顺序来检验组间的差异是否显著大于组内差异，因而在方法上与 NMDS 相对应，适合与NMDS搭配使用。

需要注意的是PERMANOVA分析(adnois)和anosim分析不仅对组间的分布的差异产生响应，它们也在一定程度上对离散效应产生响应，即该分析的显著性也可能来自组内样本距离的变化趋势在组间显著不同而导致(如需了解更多相关信息，请参考(Anderson and Walsh, 2013))。而相比之下adnois对离散效应更不敏感，更加稳健，因此是差异检验的首选。

一方面我们可以通过观察排序图来判断PERMANOVA和anosim分析的显著性是否来自样本的组间差异分布。另一方面，我们也可以使用Permdisp(Permutation test of multivariate homogeneity of groups dispersions)分析(Anderson et al., 2006)来检验分组离散性的多变量均质性。这个分析手段不关注样本组间的差异，而是检验不同样本组的组内方差之间是否存在显著差异，也就是组内样本的离散程度在不同组间是否有差异，因而可以作为以上两种方法的补充。

分析软件：python语言，scikit-bio包；R语言，vegan包等。

分析步骤：使用4.1节分析所获得的bray-curits距离矩阵（默认）文件，通过python的scikit-bio包进行“permanova”（默认检验方法）组间差异分析，设置置换检验次数设为999。当检验方法选为“adonis”时，则通过R的vegan包计算该分组方案对距离矩阵方差的解释度(R2)及显著性(P)，并设置置换检验次数设为999。

2.5 物种差异分析与标志物种

ASV/OTU韦恩图

为研究不同的样本(组)间有哪些物种是共有的，哪些是独有的，使用花瓣图或韦恩图(Venn;https://en.wikipedia.org/wiki/Venn_diagram)来进行群落分析。使用ASV/OTU丰度表制作韦恩图，根据其在各样本(组)间的有无情况分别统计各个集合的成员数，也就是各个分组独有的，以及组间共有的ASV/OTU的个数(注意不是丰度值)

分析软件：R脚本，VennDiagram包。

分析步骤：默认采用ASV/OTU在所有样本中的丰度数据，每个分组作为一个集合，按样品的分组情况统计每个集合中的成员（ASV/OTU），并计算各个集合之间的关系。

物种组成热图

为了进一步比较样本间的物种组成差异，实现对各样本的物种丰度分布趋势的展示，可以使用热图进行物种组成分析。我们默认使用平均丰度前50位的属的丰度数据绘制热图。热图的横、纵坐标可以按照特定的顺序进行的排列，如样本的采集时间等进行排序；也可根据样本间或者样本间的相关性绘制聚类树排序，即绘制聚类热图。

分析软件：R语言，pheatmap包等

分析步骤：使用R脚本计算各样本以及各分类单元的聚类结果，以交互图的形式呈现

PCA分析

在Beta多样性分析一章中，我们提到了PCoA相比于PCA更适合于生态数据的排序，这主要是由于微生物数据一般来说都非常复杂，物种(ASV/OTU)数远多于样本数，且常呈“长尾分布”，即高丰度物种较少，而大多数物种的丰度都极低；而PCA分析是基于物种丰度矩阵(欧式距离；PCoA分析则是基于样本距离矩阵)做降维处理的(Ramette, 2007)，因此，物种数越多(维度越高)，其损失的信息就越多(特别是那些低丰度物种的差异信息)。然而相比于样本中的ASV/OTU数目，在特定分类水平下(如属)的物种数就减少很多了；这时，对于物种组成相对简单的研究样本(不建议特别复杂的样本使用，如土壤等)，我们或许可以尝试使用PCA分析，这一最常见的排序分析方法来展示样本(组)间的物种丰度组成差异。

PCA分析设法将样本间的众多物种差异特征重新做线性组合，导出少数几个相互无关的的综合指标，使其尽可能多地保留原本的信息；这些指标便称为主成分。将主成分对原始数据中样本差异的解释比例排序，便得到了第一、二、三……主成分。依据各样本在这些主成分指标上的得分对其进行排序，便可以量化展示样本间物种组成的差异程度。

分析软件：R语言

分析步骤：使用R脚本计算各样本以及各分类单元的主成分坐标得分或载荷值，并以交互图的形式呈现。

MetagenomeSeq分析

样本(组)间的物种组成的差异并不是意味着所有物种组分的差异，而往往是一部分组分的差异分布。而这些差异的组分，又具体表现在不同的分类水平上。一般而言，对于本身环境类型迥异或是时空分布距离极大的样本(组)，那么它们可能在门、纲等水平就已经体现出了显著的差异；而对于本身环境类型相同、时空分布相近的样本(组)，它们之间的组成差异可能就仅限于ASV/OTU水平或是种、属水平，而在门、纲等分类水平上不具有或少有显著的差异性。所以，我们可以首先尝试寻找样本组间在统计上具有显著差异的ASV/OTU，再尝试找出这些差异ASV/OTU在不同分类水平上是否具有富集的趋势。这里使用了metagenomeSeq方法对样本组进行两两比较(默认)。该方法避免了数据稀疏(Rarefaction)过程对结果准确性的影响，特别适用于具有稀疏性的微生物组成数据。我们进一步通过曼哈顿图(Manhattan plot)展示metagenomeSeq的分析结果。使用曼哈顿图展示差异ASV/OTU并结合分类学注释，相比于其他方式，不仅可以展示数据全貌，又能快速找到目标ASV/OTU，同时又可以获知目标的具体分类位置和显著程度，并尝试发掘差异物种的分类学特征或规律。因此，近些年曼哈顿图在扩增子测序、宏基因组等领域，开始受到广泛关注(Zgadzaj, R., et.al., 2016)。这种展示更适合于物种组成复杂的样本，如土壤、底泥等。

分析软件：R脚本，metagenomeSeq包等。

分析步骤：使用未抽平的ASV/OTU表，选取上调组和对照组，按照metagenomeSeq的教程示例，调用fitFeatureModel函数使用zero-inflated log-normal model对每个ASV/OTU的分布进行拟合，并使用该模型的拟合结果判别差异的显著性。

LEfSe分析

EfSe (LDA Effect Size)分析是一种将非参数的Kruskal-Wallis以及Wilcoxon秩和检验，与线性判别分析(Linear discriminant analysis，LDA)效应量(Effect size)相结合的分析手段(Segata et al., 2011)。与metagenomeSeq分析类似，LEfSe分析也是一种差异分析方法；但LEfSe分析可以直接对所有分类水平同时进行差异分析；同时，LEfSe更强调寻找分组之间稳健的差异物种,即标志物种(biomarker)。它的一大特点是，不仅局限于对不同样本分组中的群落组成差异进行分析，更可以深入到不同的子分组(Subgroup)中，挑取在不同子分组中表现一致的标志微生物类群，目前在微生物扩增子分析、宏基因组分析等领域已获得了广泛的应用，且特别适用于医学研究中寻找生物标记物。

LEfSe的分析结果包括两部分，分别是显著差异物种LDA值分布柱状图，用以展示每个组内显著富集的物种(注意不显示显著下调的)及其重要性程度；物种分类学分枝图(Cladogram)，用以展示在各组样品中标志物种的分类学层次分布。

分析软件：Python LEfSe包，R语言，ggtree包等。

分析参数：

1. 比较策略：one-against-all(less strict)，对于通过Kruskal-Wallis检验呈显著多组差异的物种，找出其丰度最高的组（组的丰度为组内样本丰度的中位值），若该组的丰度比其他任意一组的丰度都高，则成为待验差异物种；all-against-all(more strict)，对于通过Kruskal-Wallis检验呈显著差异的物种，若所有分组的丰度都不同（同为0也认为相同），则成为待验差异物种。

2. Wilcoxon检验：在找出待验差异物种后，可选择使用Wilcoxon检验组间差异的显著性，检验的策略与比较策略保持一致，显著的物种即为差异物种；也可选择不使用检验，直接将该物种未作为差异物种。如果组内样本较少（如只有3-5个），可考虑跳过Wilcoxon检验的步骤。注：已对LEfSe参数进行调整，最小Wilcoxon样本要求设置为3（亦为最低分析要求），即只要选择了该Wilcoxon检验，该检验步骤都会发生。

3. LDA阈值：对找出的差异物种进行LDA分析来估算每个差异组分（差异物种）丰度对组间差异的效应量大小，可设置一定阈值，只有通过该阈值的差异物种才认为是标志物种（biomarker）

4. 二级分组：对于一些研究，除了分组对样本组分存在影响外，还存在着时间、空间、批次效应等不可避免的因素或影响。为了去除这些效应的影响，或是验证在这些效应下标志物种依旧稳健，可以尝试添加二级如：一级分组是性别，二级分组是年龄；一级分组是土壤类型，二级分组是所处的地理区域。

5. 二级分组的比较和检验方式：当添加了二级分组后，前面的比较步骤和检验步骤，默认只在一级分组中具有相同二级分组名的子样本组中进行，即为一次子比较（比较策略依旧有两种）和子检验（检验方式依旧与比较方式保持一致）；但只有所有的子比较和子检验的结果一致且显著，才能认为该物种为差异物种；更严格的方式则是不考虑二级分组名是否相同，将一级分组中的每个二级分组都与其他以及分组中的二级分组进行比较（比较策略依旧有两种）和检验（检验方式依旧与比较方式保持一致）。一般情况下，推荐使用前一种策略，如：比较肠道微生物差异，一级分组为性别（男、女），二级分组为年龄（老、中、青）；宽松的比较和检验方式是，分别比较老年、中年、青年中男性样本和女性样本的物种差异，结果一致的再分别进行检验，显著的即为差异物种。当然，如果是要找到男女肠道的最稳健标志物种，即不论男女间的年龄是否存在大跨度，都是保持差异的物种，那么这时可以采用严格的比较和检验策略。

6. 丰度过滤阈值：在进行LEfSe分析之前，先会进行数据的归一化，这里是将样本的组成数据都转化为每个样本总丰度为1000000；之后，可以通过设置比例阈值，去掉部分低丰度物种，0.001即代表去掉丰度比例低于0.001的物种。

7. 数据处理：可将物种在各样本中的丰度平均值作为该物种的丰度值，用于进行过滤或可视化（分支图中圆点大小）；也可将物种在各组中的丰度（组的丰度为组内样本丰度的平均值）的平均值作为该物种的丰度值。

OPLS-DA分析

寻找标志物的过程，其实就是模式识别的过程。之前介绍的PCA、层次聚类等分析方法，其实就属于无监督的模式识别方法。这里我们再介绍有监督的模式识别方法，即通过某种已知的样本关系(比如样本的来源，实验的分组)，尽可能地按照它们的变化规律提取原始数据中与之相关的变化模式，而不关注其它无关的数据信息。比如在LEfSe分析中使用的LDA分析就是一种基于监督的模式识别方法。(比如样本的来源，实验的分组)，尽可能地按照它们的变化规律提取原始数据中与之相关的变化模式，而不关注其它无关的数据信息。比如在LEfSe分析中使用的LDA分析就是一种基于监督的模式识别方法。

PLS-DA (Partial Least Squares Discriminant Analysis)分析是目前组学数据分析中最常使用的模式识别方法方法之一。它通过寻找物种丰度矩阵和给定的分组信息的最大协方差，从而在新的低维坐标系中对样本排序。OPLS-DA (Orthogonal Partial Least Squares Discriminant Analysis)分析是组学数据分析中另一常用方法，它是将正交信号校正(orthogonal signal correction, OSC)与PLS结合并对PLS进行修正，以过滤无直接关系的变化，使生成的结果更清晰明了。如果说，PCA分析在排序空间中展示的是

样本间的总物种丰度组成差异，那么PLS-DA和OPLS-DA分析则是在排序空间中只展示样本在组间的物种丰度组成差异的部分。对于组内方差存在显著差异的样本组，OPLS-DA的效果可能优于PLS-DA (Mahadevan et al., 2008)

需要注意的是，PLS-DA或OPLS-DA更常见于代谢组学数据的分析中，对于一些微生物数据常出现的组间差异小、组内变异大(噪声大)的情况，二者的表现可能并不理想。此时，或许需要更为稳健和准确的手段，如下一小节提到的随机森林分析。因此，该方法并不是首选，但提供了一种分析的可能性。

分析软件：R语言，muma包等。

分析步骤：使用R脚本计算各样本以及各分类单元的主成分坐标得分或载荷值，并以交互图的形式呈现

随机森林分析

LDA和OPLS-DA分析其实都可以归类为有监督的机器学习算法，当然在使用中，我们更强调了它们的物种重要性指标，即寻找标志物种。而寻找标志物种的目的之一就在于建立样本分类器。QIIME2就介绍了许多适用于多种组学数据的机器学习方法(machine-learning methods)，包括：ExtraTreesClassifier、LinearSVC、AdaBoostClassifier、SVC、GradientBoostingClassifier、KneighborsClassifier和RandomForestClassifier。这里我们选用默认的随机森林(Random Forests) (Breiman, 2001)算法进行分析。随机森林是一种基于决策树(Decision tree)的经典高效的机器学习算法，属于非线性分类器(Non-linear classifier)，能够深入挖掘变量之间复杂的非线性相互依赖关系，对于经常呈现离散、不连续分布的微生物群落数据而言尤其适用，近几年已有研究证明这一算法能够对微生物群落样本进行有效、稳健且准确的分类(Yatsunenko et al., 2012)。

分析软件：qiime2（2019.4）。

分析步骤：默认使用未抽平的ASV/OTU表，或使用据此产生的分类水平在门、纲、目、科、属的分类单元绝对丰度表，调用q2-sample-classifier中的“classify_samples_ncv”函数进行随机森林分析以及巢式分层交叉检验。当最大分组样本数不少于12时，进行10倍交叉检验(10-fold cross-validations)；当最大分组样本数小于12且大于等于7时，进行5倍交叉检验；最大分组样本数小于7时，设置交叉检验倍数为该值减2。

2.6 关联网络分析

关联网络的构建

识别微生物成员之间相互关系，通常是通过相关性分析来实现的。但是，标准方法计算相关值可能会因为相对丰度的换算而产生偏差，从而导致错误的关联。SparCC方法避免了上述问题，能够根据微生物群落的组成数据估计成员之间的相关值。

构建相关性矩阵之后需要确定一个阈值，以从随机噪声干扰中分离出具有生物学意义的相关值。随机矩阵理论（RMT）是一种将复杂系统中的非随机属性与随机噪声区分开的有效方法，因而，在此使用随机矩阵理论来确定相关性阈值。

分析软件：R脚本，SparCC，igraph包，RMThreshold包等

分析步骤：默认采用ASV/OTU在本项目所有样本中的丰度数据，过滤掉序列总数少于10，出现样本数少于5或者[总样本数*0.2]的ASV/OTU,采用SparCC算法，构建相关性矩阵，使用随机矩阵理论确定相关性数值的过滤阈值，再采用igraph构建关联网络数据。

关联网络的绘制

在微生物生态学的关联网络图中，每个点可称为一个node(或vertex)，它可以代表群落中的一个ASV/OTU,又或是一个分类单元；两个点之间的连接线可称为edge，代表所连接的两个点之间的正相关或负相关的分布趋势。通过关联分析的方法，寻找特定微生物群落在时空变化，环境过程驱动下所呈现的共现(Co-occurrence)或互斥(Co-exclusion)的固有模式。

分析软件：R语言，igraph，ggraph包。

分析步骤：采用igraph的induced_subgraph功能，依据节点(ASV/OTU)的丰度，提取平均丰度前100(默认值)的节点构建优势物种子网络，再使用ggraph包进行可视化。默认展示去掉当前关联网络的负相关数据以构建共现网络(即去掉负相关连线)，再采用igraph的“multi-level modularity optimization algorithm”算法对该共现网络进行模块化切割。；同时生成.gml文件可导入gephi软件(Bastian, 2009)和Cytoscape软件(Shannon et al., 2003)供其他个性化的操作。

拓扑指数

微生物生态学研究中，常常会使用到关联网络的拓扑指数，这里我们计算了常见的网络水平和节点水平拓扑指数以供此类研究需求。

分析软件：R语言，igraph包等。

分析步骤：使用R包igraph，计算构建的共现网络的拓扑特征(Csardi, 2006)。这些拓扑学指数的具体含义与计算方法详见下表。为表征各分组/样本间微生物相关网络的拓扑学特征差异，依据每个分组/样本中检出的ASV/OTU，使用R包igraph中的subgraph函数过滤共现网络以获得对应于各样本的子网络；再使用R包igraph，计算各样本子网络的图形水平拓扑学指数。

度分布

在自然界或人类社会中，绝大多数的网络都可以基于他们的拓扑结构分为随机网络、非随机网络、无尺度网络(或无标度网络)、小世界网络等。网络中某个节点与其他节点的连接数，称为这个节点的度(degree)。在随机网络中，节点的度服从泊松(poisson)分布。小世界网络的度分布与随机网络相似，但它具有更高的高集聚系数(Clustering Coefficient)。而在无尺度网络中，节点的度服从幂律分布(power-law)：多数节点可通过较短路径联系。现实世界里，一部分的微生物网络的度常呈现幂律分布规律，属于无尺度网络，而另一部则符合小世界网络的特征。不同的网络类型具有不同的群落结果特点：如小世界网络非常稳定，而无尺度网络则同时显现出对于随机故障的鲁棒性和对于针对性攻击的脆弱性。 因此，可以通过节点度的分布初步判断网络的类型。

分析软件：R脚本，igraph包等。

分析步骤：使用R包igraph，根据Erdos Renyl模型以及当前网络的节点数目和边数目，构建随机网络。再分别计算当前网络、随机网络中每个节点的相连节点数(度)。最后，分别计算当前网络、随机网络中不同度对应的节点数量。

ZIPI图

常用的指示关键物种(hubs or Keystone species)的拓扑指数有：betweenness centrality score、PageRank score等等。而利用Zi、Pi 值可将network中的节点(ASV/OTU)分成了4部分，分别为peripherals, connectors, module hubs和 network hubs。在生态学研究中，peripherals代表了微生物网络中的一些specialists，module hubs和connectors主要代表一些接近generalists的物种，network hubs则代表微生物网络中的super-generalists。详细说明请见参考文献(Deng et al., 2012)。

分析软件：R脚本，igraph包

分析步骤：使用R，根据当前共现网络的模块化切割结果，计算当前网络中每个节点的Zi、Pi score值。Zi值指的是within-module connectivity，而Pi指的是among-module connectivity。再根据Zi、Pi score值确定每个节点在关联网络中的角色(Deng et al., 2012)。

2.7 功能潜能预测

PICRUSt2分析

PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States)是一款基于样本中的标记基因序列丰度来预测样本功能丰度的软件(Gavin M. Douglas, et al., preprint)。这里的功能是指基因家族，如KEGG同源基因、EC酶分类号等。

相比其初代版本(Langille et al., 2013)，PICRUSt2包括了以下改进和特征：

1) 使用软件自带的参考基因组数据，除16S rRNA序列，还可以使用18S rRNA序列和ITS序列进行功能预测；也可以使用用户自定义参考基因组数据，从而在理论上支持各类扩增子序列的功能预测。

2) 适配OTU序列和丰度数据的同时，也适配于ASV序列和丰度数据。

3) 参考基因组数据库比原来扩大了10倍以上。

4) 使用Castor (Louca S. & Michael D., 2018)中的隐藏状态预测算法。

5) 增加了MetaCyc (Caspi, R., et al., 2008)代谢通路预测，可与宏基因组数据的结果直接比较。

6) 使用MinPath (Ye Y. & Doak T.G., 2009)推断代谢通路，使得预测更严谨。

分析软件：PICRUSt2等。

分析步骤：PICRUSt2的一般分析流程如下图所示，这里我们只做简要描述,详细请参见PICRUSt2官网(https://github.com/picrust/picrust2/wiki)。

1）首先将已知微生物基因组的16S rRNA(或18S rRNA、ITS,下同)基因序列进行对齐(align)，构建进化树，并推断它们的共同祖先的基因功能谱。此步骤该软件已经完成的。

2）将16S rRNA特征序列与参考序列对齐，从而构建新的进化树。

3）使用Castor隐藏状态预测算法，依据进化树中参考序列所对应的基因家族拷贝数，推测特征序列的最近序列物种，进而获得其基因家族拷贝数。注意，在计算每条序列的最近序列物种索引(NTSI)时，默认如果序列的NTSI>2，将在后续分析中被排除。

4）结合各样本特征序列的丰度，计算各样本的基因家族拷贝数。注意，这里我们使用分层处理，即对每个特征序列的基因家族，添加序列的物种信息并分层输出结果(即不同特征序列的功能单元不合并处理)，以便实现功能与物种的对应分析。

5）最后，将基因家族“映射”到各类数据库中，默认使用MinPath推断代谢通路的存在，进而获得各样本中代谢通路的丰度数据。

PICRUSt2能将16S rRNA基因序列在多个功能数据库中进行预测，包括MetaCyc (https://metacyc.org/)，KEGG (https://www.kegg.jp/)、COG (https://www.ncbi.nlm.nih.gov/COG/)、Pfam (http://pfam.xfam.org/)和TIGRFAM (http://tigrfams.jcvi.org/cgi-bin/index.cgi)等，我们默认只提供最常用的MetaCyc、KEGG和COG数据库注释结果；同时，它还能使用18S rRNA基因和ITS基因序列在MetaCyc数据库(目前不支持其他数据库)进行代谢通路预测。

为了便于后续比较分析，我们默认分别使用每个样本的KO和EC的丰度总和的百万分之一，对代谢通路丰度文件(即path_abun_unstrat.tsv和path_abun_strat.tsv)以及功能单元文件(即pred_metagenome_unstrat.tsv和pred_metagenome_strat.tsv)进行归一化，使丰度值的单位为“每

百万功能单元”。当然作为个性化的选项，我们也可以使用物种丰度总和(剔除NTSI>2的物种)进行归一化。需要注意的是，不同样本的代谢通路丰度之和可能不相等，但一般无需再进行归一化处理。

功能单元PCoA分析

由于功能单元(EC/KO/COG)的数量往往也非常巨大，难以直接比较，我们同样也可以使用样本差异距离矩阵(默认采用Bray-Curtis距离)结合主坐标分析将样本功能差异在低维度展开。

分析软件：R语言，vegan、ape包等。

分析步骤：使用归一化后的功能单元丰度表（每个样本丰度和为1M），使用R脚本在R中计算距离矩阵并进行PCoA分析输出样本点的PCoA坐标，并将其绘制成二维散点图。

代谢通路统计

获得的功能单元，可以依据代谢通路数据库和一定的计算方法，获得代谢通路的丰度值。KEGG数据库、MetaCyc数据以及COG数据都是常用的数据库。

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。

MetaCyc是生命科学领域内已通过实验数据阐明的最大的代谢参考数据库。目前，它包含了来自3009种不同生物体的2722个途径。MetaCyc包含了参与初级和次级代谢的各种通路以及相关代谢物，生物化学反应，酶和基因等信息，旨在通过存储具有代表性的实验验证的代谢通路，来对所有生命的代谢过程进行分类。

COG数据库是 NCBI 开发的用于同源蛋白注释的数据库。它将细菌、藻类和真核生物的21个完整基因组的编码蛋白，根据系统进化关系分类构建而成。通过将蛋白与数据库的比对，可以很好的预测蛋白质的功能。

分析软件：R语言。

分析步骤：使用归一化的pathway/group丰度表，依据选择的样本，计算第二层次通路/分类的平均丰度或全部数量。

代谢通路差异分析

在获得代谢通路的丰度数据后，我们可以尝试找出组间具有显著差异代谢通路。这里我们使用metagenomeSeq的方法(2.5.4中已有介绍)。

分析软件：R语言，metagenomeSeq包等

分析步骤：使用归一化的pathway/group丰度表，依据分组情况，按照metagenomeSeq的教程示例，调用fitFeatureModel函数使用zero-inated log-normal model对每个pathway/group的分布进行拟合，并使用该模型的拟合结果判别差异的显著性。

代谢通路的物种组成

获得了样本/分组的功能组成，特别是一些差异通路之后，我们还需要知道是哪些物种编码了这些具有这些功能潜能的基因。可使用分层的样本代谢通路丰度表进行通路的物种组成分析。需要注意的是，这里我们是将通路与物种进行关联，那么隐含的假设就是：这些代谢通路可由该物种独立实现全部或大部分的功能，如固氮，产甲烷等功能。如果一条通路被认为是由多种微生物协同完成的，那么此时就不建议在通路水平上进行物种组成的分析了，如硝化作用、反硝化作用等功能。

分析软件：R语言

分析步骤：依据选择的通路抓取分层的样本代谢通路丰度表中的相应数据，绘制物种组成柱状图。

2.8 附录

附表

Alpha多样性指数简介

指数名称	计算方法
Chao1丰富度估计指数(The Chao1 estimator)	由Chao首先提出，通过
Observed species指数(Observed species richness, Observed OTUs)	通过计算群落中不同的
Shannon多样性指数(Shannon diversity index, Shannon-Wiener diversity index H`)	综合考虑了群落的丰富
Simpson 多样性指数(The Simpson index)	通过计算群落中随机取
Faith’ s PD指数(Faith’ s Phylogenetic Diversity)	通过计算样本中的ASV/
Pielou’ s evenness指数(Pielou’ s Evenness index J`)	通过将香浓指数H`除以
Good’ s coverage指数(Good’ s nonparametric Coverage estimator)	是计算群落中非singlet

Beta多样指数简介

指数名称	计算方法
Jaccard distance	通过计算两个样本间非共有物种在所有物种中的比例，强调物种的有无，但不
Bray-Curtis distance	采用加权的计算方法，它通过计算两个样本间各物种丰度差值的绝对值之和与
unweighted UniFrac distance	在考虑物种的有无的同时考虑群落成员之间是否存在系统发育亲缘关系；它计
weighted UniFrac distance	在unweighted UniFrac的基础上，同时考虑物种丰度的差异；它通过给每个进

拓扑学指数简介

	拓扑学指数(Li, 2019)	描述
节点水平	Betweenness centrality	通过目标节点的节点间最短路径的数量。

	拓扑学指数(Li, 2019)	描述
	Closeness centrality	目标节点与所有其他节点的平均距离的倒数。
	Degree	目标节点的相连节点数。
	Transitivity	目标节点的相邻节点之间相连的概率，也称为聚类系数。
图形水平	Average nearest neighbor degree	目标网络中所有节点的相邻节点的平均度的均值
	Average path lenghth	网络中所有最短路径的长度之和。最短路径是指任意两个节
	Betweenness centrality	基于目标网络中所有节点水平值计算的图水平Betweenness
	Closeness centrality	基于目标网络中所有节点水平值计算的图水平Closeness ce
	Degree assortativity	基于节点Degree值计算的目标网络皮尔森相关系数
	Degree centralization	基于目标网络中所有节点Degree值计算的图水平Degree ce
	Density	目标网络中边数目与潜在最大边数目的比值
	Transitivity	所有节点的Transitivity的平均值
	Number of vertice	目标网络中的节点数目
	Number of edge	目标网络中的边的数目

参考文献

[1] Anderson, M.J., and Willis, T.J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology 84, 511-525.

[2] Anderson, M.J., Ellingsen, K.E. and McArdle, B.H. (2006) Multivariate dispersion as a measure of beta diversity. Ecology Letters 9, 683-693.

[3] Anderson, M.J., Walsh, D.C.I. (2013) Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? Ecological Monographs 83(4), 557-574.

[4] Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. Peerj 3.

[5] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[6] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological), 57(1), 289-300.

[7] Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. Philos Trans R Soc B-Biol Sci 360, 1935-1943.

[8] Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., and Caporaso, J.G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10, 57-U11.

[9] Bokulich, N. A. , Kaehler, B. D. , Ram, R. J. , Matthew, D. , Evan, B. , & Rob, K. , et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2’ s q2-feature-classifier plugin. Microbiome, 6(1), 90-.

[10] Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. Ecological Monographs, 27.

[11] Breiman, L. (2001). Random forests. Mach Learn 45, 5-32.

[12] Callahan, B. J., Mcmurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). Dada2: high-resolution sample inference from illumina amplicon data. Nature Methods, 13(7), 581-583.

[13] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7, 335-336.

[14] Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., & Fulcher, C. A., et al. (2008). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Research 40(Database issue), D742.

[15] Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics 11, 265-270.

[16] Chao, A., and Shen, T.J. (2004). Nonparametric prediction in species sampling. J Agric Biol Environ Stat 9, 253-269.

[17] Chao, A., and Yang, M.C.K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. Biometrika 80, 193-201.

[18] Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18, 117-143.

[19] Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37, D141-D145.

[20] Deng, Y., Jiang, Y. H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. BMC Bioinformatics 13(1), 113-.

[21] DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72, 5069-5072.

[22] Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

[23] Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194-2200.

[24] Franzosa, E.A., McIver, L.J., Rahnavard G., Thompson, L.R. Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso,J.G., Segata, N. & Huttenhower C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods 15, 962–968.

[25] Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. Biological Conservation 61, 1-10.

[26] Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. Nat Rev Microbiol 10, 538-550.

[27] Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C. & Langille, M.G.I. (2020). PICRUSt2 for prediction of metagenome functions. Nature Biotechnology 38, 685–688.

[28] Good, I. J. (1953). The population frequency of species and the estimation of the population parameters. Biometrics 40, 237-246.

[29] Hamilton, N. (2016). ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams.

[30] Heck, K.L., van Belle, G., and Simberloff, D. (1975). Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size. Ecology 56, 1459-1461.

[31] Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res 21, 1552-1560.

[32] Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudoise des Sciences Naturelles 44, 223-270.

[33] Katoh, & K. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Research, 30(14), 3059-3066.

[34] Kemp, P.F., and Aller, J.Y. (2004). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. FEMS Microbiol Ecol 47, 161-177.

[35] Koljalg, U., Nilsson, R.H., Abarenkov, K., Tedersoo, L., Taylor, A.F.S., Bahram, M., Bates, S.T., Bruns, T.D., Bengtsson-Palme, J., Callaghan, T.M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. Molecular Ecology 22, 5271-5277.

[36] Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Thurber, R.L.V., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology 31, 814-+.

[37] Legendre, P. & Legendre, L. (1998). Numerical ecology (second edition).

[38] Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R., et al. (2008a). Evolution of mammals and their gut microbes. Science 320, 1647-1651.

[39] Ley, R.E., Lozupone, C.A., Hamady, M., Knight, R., and Gordon, J.I. (2008b). Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol 6, 776-788.

[40] Li H.Y., (2019). The Biodiversity of Paddy Soil Microbial Community and its Correlations with the Chemodiversity of Dissolved Organic Matter across Typical Regions in China. Ph.D. Dissertation. Zhejiang: Zhejiang University.

[41] Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News 2, 18-22.

[42] Stilianos L., and Michael D., (2018). Efficient comparative phylogenetics on large trees. Bioinformatics, 34, Issue 6(34), 1053–1055.

[43] Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology, 71(12), 8228-8235.

[44] Lozupone, C.A., Hamady, M., Kelley, S.T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73, 1576-1585.

[45] Magoc, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957-2963.

[46] Mahadevan, S., Shah, S.L., Marrie, T.J., and Slupsky, C.M. (2008). Analysis of metabolomic data using support vector machines. Analytical Chemistry 80(19), 7562-7570.

[47] McArdle, B.H., and Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82, 290-297.

[48] Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., Gonzalez, A., Fontana, L., Henrissat, B., Knight, R., and Gordon, J.I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332, 970-974.

[49] Nilsson, R. H. , Ryberg, M. , Kristiansson, E. , Abarenkov, K. , Larsson, K. H. , & Urmas Kõljalg. (2006). Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLOS ONE, 1.

[50] Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). Interactive metagenomic visualization in a Web browser. Bmc Bioinformatics 12.

[51] Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. Journal of Theoretical Biology 13, 131-144.

[52] Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26, 1641-1650.

[53] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Gloeckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, D590-D596.

[54] Ramette, A. (2007). Multivariate analyses in microbial ecology. FEMS Microbiol Ecol 62, 142-160

[55] Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). Vsearch: a versatile open source tool for metagenomics. Peerj, 4(10).

[56] Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75, 7537-7541.

[57] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biol 12.

[58] Shannon, C.E. (1948a). A mathematical theory of communication. The Bell System Technical Journal 27, 379-423.

[59] Shannon, C.E. (1948b). A mathematical theory of communication. The Bell System Technical Journal 27, 623-656.

[60] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13, 2498-2504.

[61] Simpson, E.H. (1949). Measurement of Diversity. Nature 163, 688.

[62] Warton, D.I., Wright, S.T., and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution 3, 89-101.

[63] White, J.R., Nagarajan, N., and Pop, M. (2009). Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol 5.

[64] Whittaker, R. H. (1960). Vegetation of the siskiyou mountains, oregon and california. Ecological Monographs, 30(4), 279-338.

[65] Whittaker, R. H. (1972). Evolution and measurement of species diversity. Taxon, 21(2-3), 213-251.

[66] Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature 486, 222-227.

[67] Ye, Y., and Doak, T.G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. Plos Computational Biology 5(8), e1000465.

[68] Zgadzaj, R., Garridooter, R., Jensen, D. B., Koprivova, A., Schulzelefert, P., & Radutoiu, S. (2016). Root nodule symbiosis in lotus japonicus drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. Proc Natl Acad Sci USA, 113(49).