及因生物
— TGENE —

16S分析报告

及因（上海）生物科技有限公司

Tgene Biotech (Shanghai) Co.,Ltd.

1 项目信息 😊
2 实验流程 🧪
3 分析流程 🖥️
4 OTU聚类及分类学分析
5 物种组成分析
6 Alpha多样性分析
7 Beta多样性分析 ⭐
8 物种差异分析 ⭐
9 功能预测分析
10 附录 🏷️

微生物多样性测序项目分析报告

(http://www.honsunbio.com/)

---

1 项目信息 😊

- 客户信息：张三
- 订单编号：
- 样本个数：18
- 报告日期：2024-01-09

2 实验流程 🧪

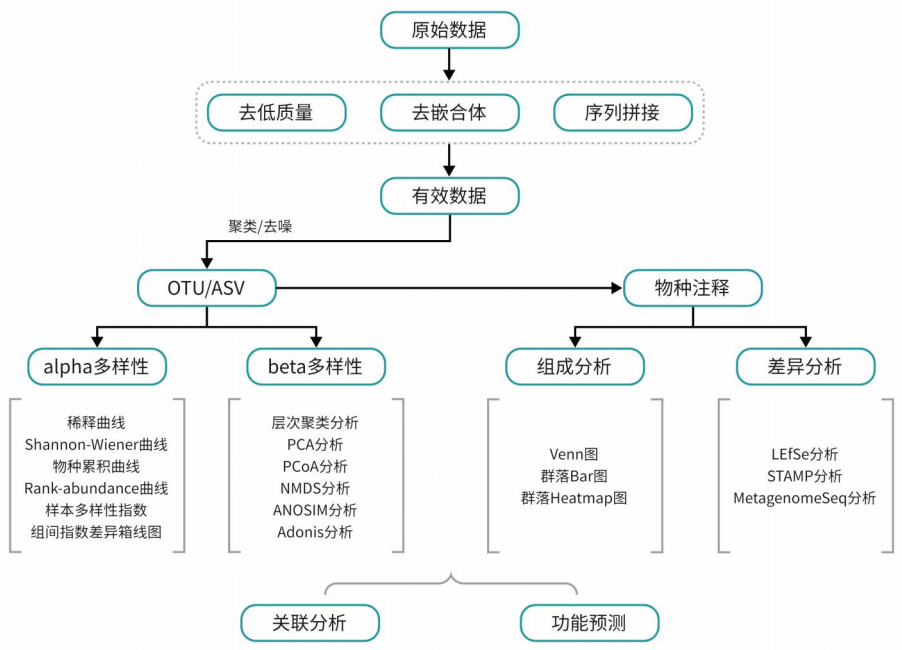
- 基因组DNA抽提
  - 完成基因组DNA抽提后，利用1%琼脂糖凝胶电泳检测抽提的基因组DNA。
- PCR扩增
  - 按指定测序区域，合成带有barcode的特异引物。为保证后续数据分析的准确性及可靠性，需满足两个条件，1) 尽可能使用低循环数扩增；2)保证每个样本扩增的循环数一致。随机选取具有代表性的样本进行预实验，确保在最低循环数中使绝大多数样本能够扩增出浓度合适的产物。

- PCR采用TransGen AP221-02：TransStart Fastpfu DNA Polymerase；
- PCR仪：ABI GeneAmp® 9700型；
- 全部样本按照正式实验条件进行，每个样本3个重复，将同一样本的PCR产物混合后用2%琼脂糖凝胶电泳检测，使用AxyPrepDNA凝胶回收试剂盒（AXYGEN公司）切胶回收PCR产物，Tris_HCl洗脱；2%琼脂糖电泳检测。

- 荧光定量
  - 参照电泳初步定量结果，将PCR产物用QuantiFluor™-ST蓝色荧光定量系统（Promega公司）进行检测定量，之后按照每个样本的测序量要求，进行相应比例的混合。
- Miseq文库构建
  - 1)连接“Y”字形接头；
  - 2)使用磁珠筛选去除接头自连片段；
  - 3)利用PCR扩增进行文库模板的富集；
  - 4)氢氧化钠变性，产生单链DNA片段。
  - 试剂：TruSeqTM DNA Sample Prep Kit
- Miseq测序
  - 1)DNA片段的一端与引物碱基互补，固定在芯片上；
  - 2)另一端随机与附近的另外一个引物互补，也被固定住，形成“桥(bridge)”；
  - 3)PCR扩增，产生DNA簇；
  - 4)DNA扩增子线性化成为单链；
  - 5)加入改造过的DNA聚合酶和带有4种不同荧光组合标记的dNTP，每次循环只合成一个碱基；
  - 6)用激光扫描反应板表面，读取每条模板序列第一轮反应所聚合上去的核苷酸种类；
  - 7)将“荧光基团”和“终止基团”化学切割，恢复3’端粘性，继续聚合第二个核苷酸；
  - 8)统计每轮收集到的荧光信号结果，获知模板DNA片段的序列。

3 分析流程 

测序得到的PE reads根据overlap关系进行拼接，同时对序列进行质控和过滤，区分样本后进行序列聚类或去噪分析和物种分类学分析。基于聚类或去噪分析结果，可以进行多种多样性指数分析，以及对测序深度进行检测；基于物种分类学信息，可以在各个分类水平上进行群落结构的统计。在上述分析的基础上，可以进行一系列更深入的数据挖掘分析并进行可视化。



4 OTU聚类及分类学分析

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，就需要对序列进行归类操作（cluster）。通过归类操作，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，通常在97%的相似水平下的OTU进行生物信息统计分析。

分析步骤如下：

- 对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/dereplication.html (http://drive5.com/usearch/manual/dereplication.html)）；
- 去除没有重复的单序列（http://drive5.com/usearch/manual/singletons.html (http://drive5.com/usearch/manual/singletons.html)）；
- 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。
- 将所有优化序列map至OTU代表序列，选出与OTU代表序列相似性在97%以上的序列，生成OTU表格。

为了得到每个分类单元对应的物种信息，采用RDP classifier贝叶斯算法对代表序列进行分类学注释，并分别在各个分类水平：domain（域），kingdom（界），phylum（门），class（纲），order（目），family（科），genus（属），species（种）统计各样本的群落组成。

物种参考数据库如下:

- 16S细菌和古菌：Silva（Release138 http://www.arb-silva.de (http://www.arb-si
- 18S真菌：Silva（Release138 http://www.arb-silva.de (http://www.arb-silva.de)）
- ITS真菌：Unite（Release7.0 http://unite.ut.ee/index.php (http://unite.ut.ee/index.php)）

结果目录及文件

result/Taxa (result/Taxa)

- rep_seqs.fasta 文件说明
- otu_table.xls 文件说明
- otu_taxa_table.xls 文件说明
- taxa_count.xls 文件说明
- tax_summary/ 文件说明
  - domain/kingdom/phylum/class/order/family/genus/species.xls
  - domain/kingdom/phylum/class/order/family/genus/species.relative.xls

结果预览

Show 10 entries Search:

Table. Taxonomy information of samples

Feature	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2	OE
OTU_1	756	433	780	782	928	637	266	179	
OTU_10	235	213	177	186	83	172	300	213	

OTU_100	16	3	14	13	5	12	16	7
OTU_101	17	15	9	17	14	24	23	17
OTU_102	15	15	11	7	2	5	21	28
OTU_103	13	17	21	26	24	12	8	10
OTU_104	0	6	0	1	0	6	31	50
OTU_105	3	1	0	196	1	3	9	46
OTU_106	13	16	6	29	12	6	10	15

OTU_107	11	10	10	19	13	19	11	7

Showing 1 to 10 of 300 entries

Previous 1 2 3 4 5 … 30 Next

注：第1列为不同分类水平的名称，第2列至最后一列为各样本在各分类中的丰度。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。

Show 10 entries Search:

Table. Taxonomy summary of samples

	phylum	class	order	family	ger
KO1	17	36	79	128	192	93
KO2	17	34	83	130	190	85
KO3	20	36	80	125	184	74
KO4	18	36	77	120	173	78
KO5	19	30	72	115	176	75
KO6	20	37	80	125	184	79
OE1	19	39	88	135	204	87
OE2	20	39	87	138	204	89
OE3	18	35	85	132	191	86
OE4	21	38	83	133	198	89

Showing 1 to 10 of 18 entries Previous 1 2 Next

注：第一列为样本名称，第2列至最后一列为各样本在各分类水平的分类单元数目。

5 物种组成分析

5.1 Venn图分析

Venn图也可以称为韦恩图、维恩图、文氏图等，用于展示不同样本中共有和独有的特征数目。每个形状的面积与其包含的元素数量成正比的Venn图称为面积比例（或按比例缩放）Venn图。在微生物数据分析过程中，经常需要对不同样本中共有或特有的微生物组特征进行统计。基于此需求，通常可以选择Venn图等进行可视化。

结果目录及文件

result/Venn (result/Venn)

- fig-*.pdf 文件说明
- venn-sets.xls 文件说明

结果预览

Venn/1/OTU | Venn/2/OTU

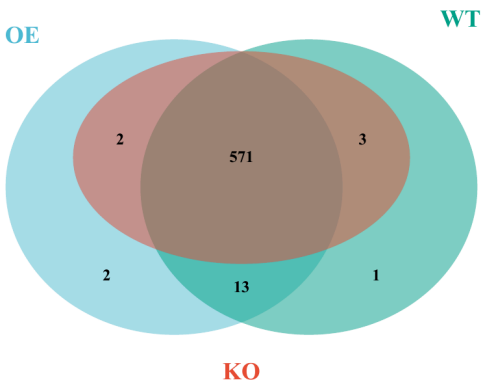


Fig. Venn plot

注：图中非重叠部分为每个组特有，重叠部分为各组共有，并将数字标示在相应范围内。

5.2 群落Bar图分析

使用统计学的分析方法，观测样本在不同分类水平上的组成结构。将多个样本的组成数据放在一起对比时，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。通常使用较直观的柱状图形式呈现。

结果目录及文件

result/Community (result/Community)

- *.bar.pdf 文件说明
- *.bar.xls 文件说明

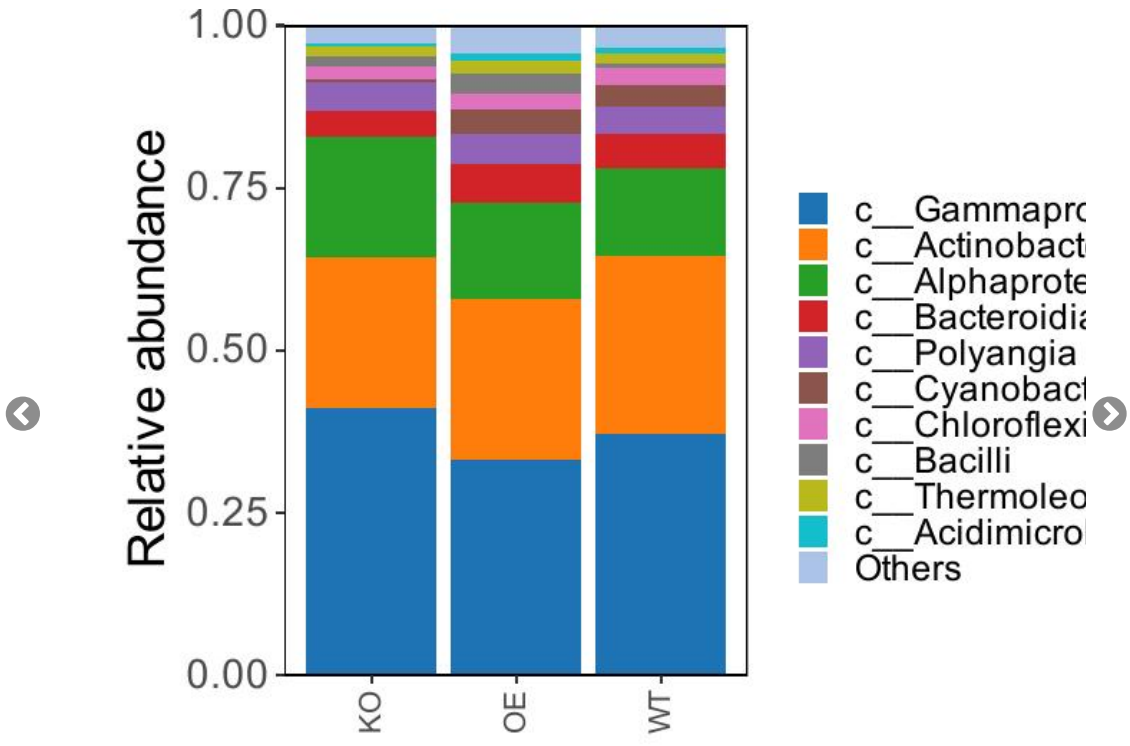
结果预览

Community/1/Class | Community/1/Species | Community/1/Family

Community/1/Phylum | Community/1/Genus | Community/1/Order

Community/2/Class | Community/2/Species | Community/2/Family

Community/2/Phylum | Community/2/Genus | Community/2/Order



class.xls.group.bar.pdf.png

Fig. Bar plot

注：为使视图效果最佳，作图时可将丰度低于1%的部分合并为other在图中显示。横坐标代表样本，纵坐标代表分类占比，颜色代表不同分类。

5.3 群落Heatmap图分析

热图（Heatmap）是一个以颜色深浅来显示数据大小的图片形式，是对数据分布情况进行分析的直观可视化方法，还可以对样本和变量进行聚类，发掘数据内部的潜在关系。生物学中热图经常用于展示多个基因在不同样本中的表达水平，然后可以通过聚类等方式查看不同组（如疾病组和对照组）特有的pattern。热图还可以用于展示其他物质的丰度比如微生物的相对丰度、代谢组不同物质的含量等等。当然，另一个热图的重要用处就是展现不同指标、不同样本等之间的相关性。

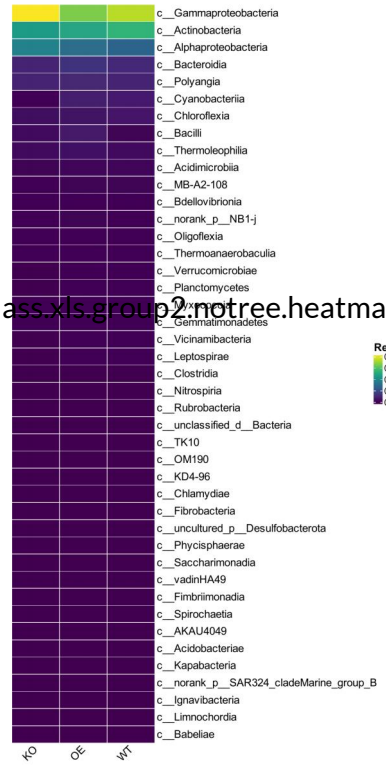
结果目录及文件

result/Heatmap (result/Heatmap)

- *.heatmap.pdf 文件说明
- *.top.xls 文件说明

结果预览

Heatmap/1/Class | Heatmap/1/Family | Heatmap/1/Phylum | Heatmap/1/Genus | Heatmap/1/Order | Heatmap/2/Class | Heatmap/2/Family | Heatmap/2/Phylum | Heatmap/2/Genus | Heatmap/2/Order



class.xls.group2.notree.heatmap.pdf.png

Fig

注：上图每一列代表一个样本，每一行代 的树状图是对物种和样本的聚类，通常聚类的结果把物种 tern相似的样本聚在一起。通常如果两组的差异较明显， 本聚类在一起。相反如果是差异较小的两组样本，就很可

6 Alpha多样性分析

群落生态学中研究微生物多样性，通过单样本的多样性分析（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。

- 计算菌群丰富度（Community richness）的指数有：
  - Chao - the Chao1 estimator (https://mothur.org/wiki/chao (https://mothur.org/wiki/chao))
  - Ace - the ACE estimator (https://mothur.org/wiki/ace/ (https://mothur.org/wiki/ace/))
- 计算菌群均一度（Community evenness）的指数有：
  - Pielou - the Pielou's Evenness index (https://en.wikipedia.org/wiki/Species_evenness (https://en.wikipedia.org/wiki/Species_evenness))
- 计算菌群多样性（Community diversity）的指数有：
  - Shannon - the Shannon index (https://mothur.org/wiki/shannon (https://mothur.org/wiki/shannon))
  - Simpson - the Simpson index (https://mothur.org/wiki/simpson (https://mothur.org/wiki/simpson))
- 计算系统发育多样性（Phylogenetic diversity）的指数有：
  - PD - the Faith's Phylogenetic Diversity (https://en.wikipedia.org/wiki/Phylogenetic_diversity (https://en.wikipedia.org/wiki/Phylogenetic_diversity))
- 测序深度指数有：

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

6.1 稀释曲线

稀释曲线是从样本中随机抽取一定数量的个体，统计这些个体所代表的物种数目，并以个体数与物种数来构建曲线。它可以用来比较测序数据量不同的样本中物种的丰富度，也可以用来说明样本的测序数据量是否合理。采用对序列进行随机抽样的方法，以抽到的序列数与它们所能代表的分类单元数目构建Rarefaction curve，当曲线趋向平坦时，说明测序数据量合理，更多的数据量只会产生少量新的分类单元，反之则表明继续测序还可能产生较多新的分类单元。因此，通过作稀释性曲线，可得出样本的测序深度情况。

结果目录及文件

result/RarefactionCurve (result/RarefactionCurve)

- fig-*.pdf 文件说明

结果预览

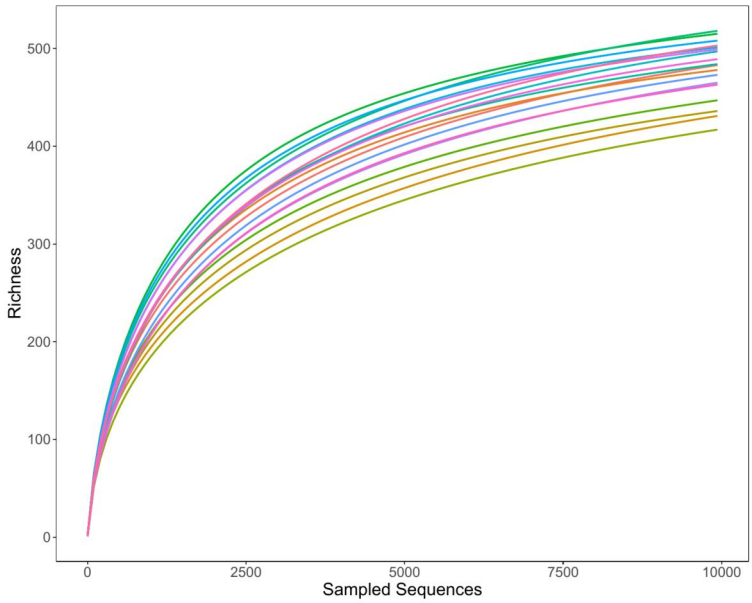


Fig. Rarefaction curves

注：横坐标表示随机抽取的测序数据量，纵坐标表示观测到的分类单元数量

6.2 Shannon-Wiener曲线

Shannon-Wiener是反映样本中微生物多样性的指数，利用各样本的测序量在不同测序深度时的微生物多样性指数构建曲线，以此反映各样本在不同测序数量时的微生物多样性。当曲线趋向平坦时，说明测序数据量足够大，可以反映样本中绝大多数的微生物信息。

结果目录及文件

result/ShannonCurve (result/ShannonCurve)

- fig-*.pdf 文件说明

结果预览

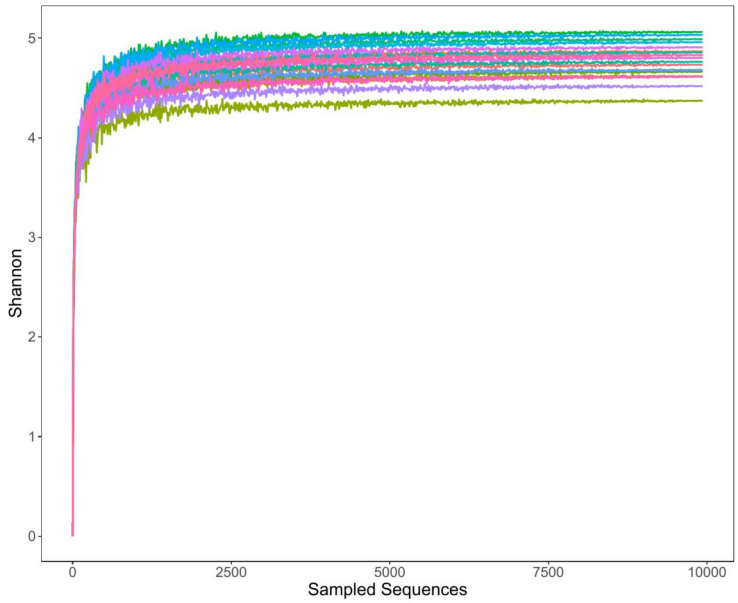


Fig. Shannon-Wiener curves

注：横坐标为测序深度，纵坐标为Shannon指数

6.3 Rank-abundance曲线

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本中，每一个分类单元所含的序列数，将分类单元按丰度（所含有的序列条数）由大到小等级排序，再以分类单元等级为横坐标，以每个分类单元中所含的序列数（也可用分类单元中序列数的相对百分含量）为纵坐标做图。Rank-Abundance曲线可用来解释多样性的两个方面，即物种丰度和物种均匀度。在水平方向，物种的丰度由曲线的宽度来反映，物种的丰度越高，曲线在横轴上的范围越大；曲线的形状（平滑程度）反映了样本中物种的均度，曲线越平缓，物种分布越均匀。

结果目录及文件

result/RankAbundance (result/RankAbundance)

- fig-*.pdf 文件说明

结果预览

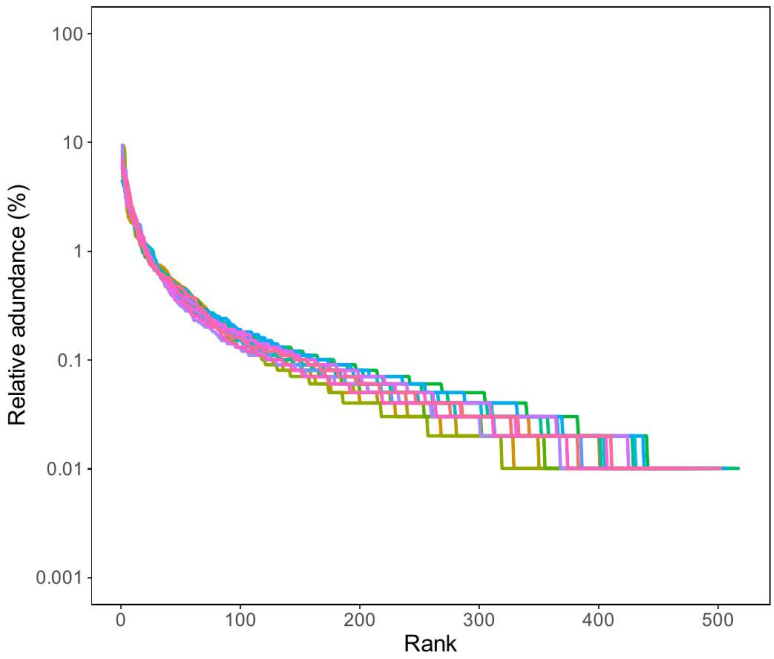


Fig. Rank-Abundance curves

注：横坐标为按丰度大小排列的分类单元等级，纵坐标为分类单元中序列数的相对百分含量

6.4 样本多样性指数

结果目录及文件

result/AlphaDiversity (result/AlphaDiversity)

- alpha_estimators.xls 文件说明

结果预览

Show 10 entries Search:

Table. Alpha diversity

Sample	Observed	Chao1	ACE	Shannon	Simpson		
KO1	483	571.596	553.932	4.73368	0.976615	0.989832	0.
KO2	478	528.05	525.75	4.86632	0.98269	0.992147	0.
KO3	431	515.726	515.622	4.61585	0.977592	0.989631	0.
KO4	436	489.443	499.191	4.61618	0.976716	0.991241	0.
KO5	417	494	497.28	4.37093	0.966518	0.990033	0.
KO6	447	526.222	514.088	4.66292	0.977414	0.990637	0.
OE1	515	562.034	554.896	5.06122	0.985991	0.992449	0.
OE2	518	573.625	576.536	4.98906	0.984097	0.990939	0.
OE3	484	550.725	529.994	4.76175	0.980271	0.991644	0.
OE4	497	555.603	559.676	4.85576	0.98283	0.990637	0.

Showing 1 to 10 of 18 entries Previous 1 2 Next

注：第1列是样本名称，之后各列的数值各个alpha多样性指数值。

6.5 组间指数差异箱线图

箱形图（Box-plot）又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的统计图。因形状如箱子而得名。在微生物组领域，常用于展示各样品分组中Alpha多样性的分布。同时可以结合假设检验统计各组间多样性指数是否存在显著性差异。

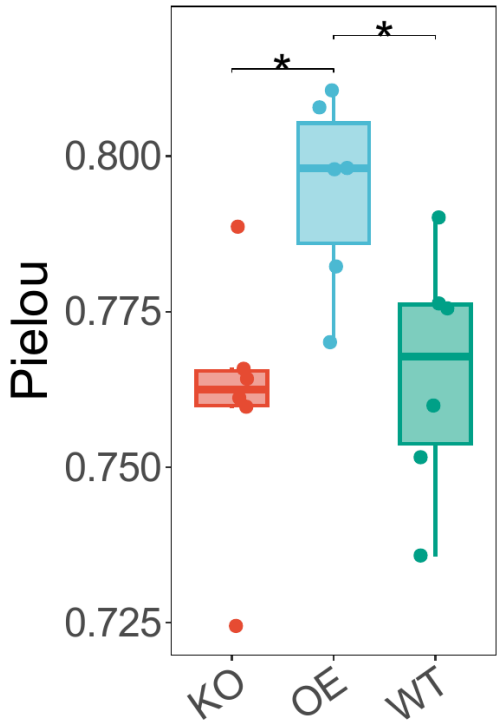
结果目录及文件

result/AlphaBoxplot (result/AlphaBoxplot)

- *-boxplot.pdf 文件说明
- *-test-result.xls 文件说明

结果预览

AlphaBoxplot/1/Alpha AlphaBoxplot/2/Alpha



Pielou-wilcoxon-boxplot.pdf.png

Fig. Boxplot of alpha index

注：横坐标表示样本分组，纵坐标表示多样性指数值。若有显著性差异，则组间会展示显著性标记，通常以“*”表示。

7 Beta多样性分析 ⭐

Beta多样性作为群落结构研究的根基，常用来比较不同生态系统之间的差异，反映生物种类因环境所造成异质性。通俗来讲，不同的处理（环境、健康状态等），会导致群落结构产生变化。指示Beta的方法有很多，在文章中常见的有距离指数（Jaccard、Bray-Curtis、Unifrac）；排序分析（PCA、PCoA、NMDS）；聚类分析（UPGMA）；差异分析（adonis、anosim）等几大类。通过这些分析，可以揭示分组是否符合预期，采样是否合理，还可对离群样本进行剔除等。

7.1 样本间距离矩阵

结果目录及文件

result/BetaDiversity (result/BetaDiversity)

- beta_distance_*.txt 文件说明

7.2 样本层次聚类分析

层次聚类的算法通过计算两个样本间的相似性，对所有样本中最为相似的两个进行组合，并反复迭代这一过程。简单的说，层次聚类的算法是通过计算每一个样本与所有其他样本之间的距离来确定它们之间的相似性，距离越小，相似度越高。并将距离最近的两个样本进行组合，生成聚类树。两个样本间相似性的度量通常是根据beta多样性距离矩阵，比如bray_curtis距离。

结果目录及文件

result/Hclust (result/Hclust)

- fig-hclust-tree.pdf 文件说明
- hclust-tree.tre 文件说明

结果预览

Hclust/1/OTU_Bray | Hclust/2/OTU_Bray

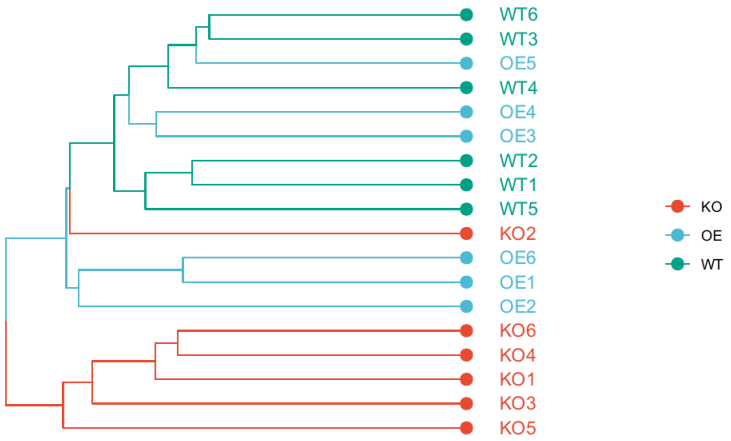


Fig. Cluster tree of samples

注：树枝长度代表样本间的距离，样本可按预先的分组以不同颜色区分。

7.3 PCA分析

PCA，即主成分分析（Principal Component Analysis），是一组变量通过正交变换转变成另一组变量，从而实现数据降维目的的一种分析方法。在一张PCA散点图中，数个样本的点聚在一起，那么就说明这几个样本之间的相似性非常高；反之，如果几个样本的点非常分散，则说明这几个样本之间的相似性比较低。例如几个组的样本对应的散点在组内呈现相互聚集的情况，说明组内的重复性

比较好，样本数据非常相似，而组间则有较好的区分度。有时候为了说明组内样本的相
会用一个椭圆将同一组内的样本对应的散点全部囊括起来。所以通过PCA分析，我们既可以直观地了解到每个样本的特征，又可以将样本进行聚类，看它们之间的相关性和差异性。

结果目录及文件

result/Pca (result/Pca)

- fig-pca-*.pdf 文件说明
- pca-importance.xls 文件说明
- pca-sites.xls 文件说明
- adonis-output.xls 文件说明

结果预览

Pca/1/OTU | Pca/2/OTU

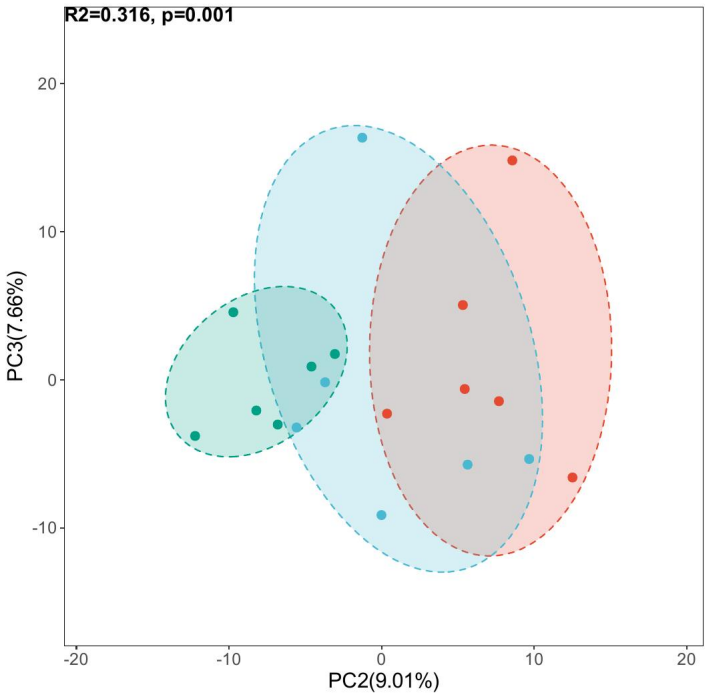


fig-pca-PC2PC3.pdf.png

Fig. PCA plot

注：PCA图形为散点图，图中的点代表样本，不同颜色/形状代表样本所属的分组信息。连线的距离体现每个样本之间的相似性，距离越短，相似性越大。同组样本距离远近说明了样本的重复性强弱，不同组样本的远近则反应了组间群落差异。通常情况下，来自不同环境的样本表现出各自聚集的现象。图中的横/纵轴分别代表了第一、第二主成分，横纵轴上所标注的百分比即该主成分对样品差异的贡献度，通常横轴百分比数值高于纵轴。

7.4 PCoA分析

PCoA，即主坐标分析（Principal Coordinates 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。通过PCoA分析可以观察个体或群体间的差异。与PCA分析的主要区别在于，PCA基于欧氏距离，PCoA基于除欧氏距离以外的其它距离，通过降维找出影响样本群落组成差异的潜在主成分。

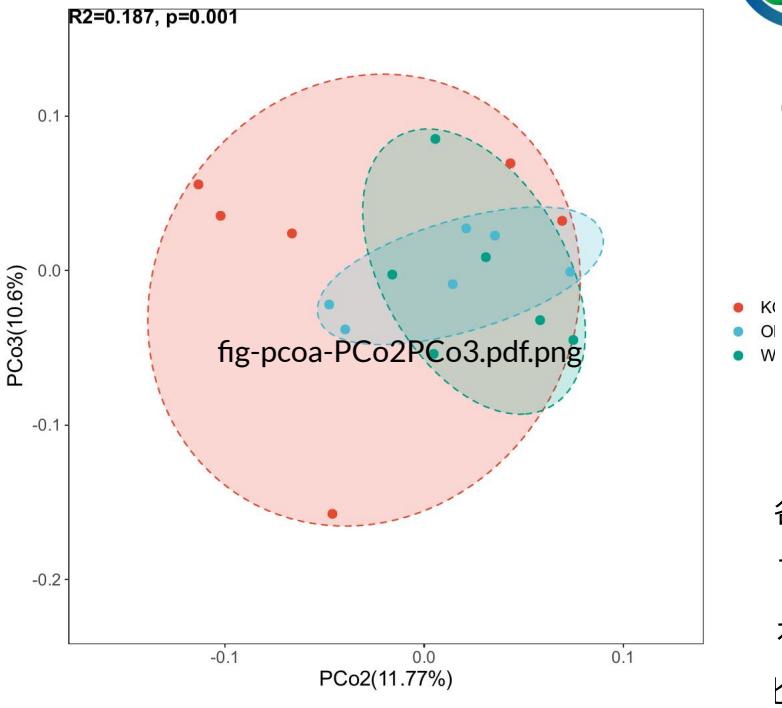
结果目录及文件

result/Pcoa (result/Pcoa)

- fig-pcoa-*.pdf 文件说明
- pcoa-importance.xls 文件说明
- pcoa-sites.xls 文件说明
- adonis-output.xls 文件说明

结果预览

Pcoa/1/OTU_Unweighted_unifrac | Pcoa/1/OTU_Bray | Pcoa/1/OTU_Weighted_unifrac | Pcoa/1/OTU_Jaccard | Pcoa/2/OTU_Unweighted_unifrac | Pcoa/2/OTU_Bray | Pcoa/2/OTU_Weighted_unifrac | Pcoa/2/OTU_Jaccard



注：PCoA图形为散点图，图中 各样本点连线的距离体现每个样本之间 了样本的重复性强弱，不同组样本的远 本表现出各自聚集的现象。图中的横/纵 比即该主成分对样品差异的贡献度，通常横轴百分比数值高于纵轴。

7.5 NMDS分析

NMDS，即非度量多维标度分析（Non-metric Multidimensional Scaling），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。其特点是根据样品中包含的物种信息，以点的形式反应在多维空间上，而对不同样品间的差异程度是通过点与点间的距离体现的，最终获得样品的空间定位图。NMDS优于PCA、PCoA的地方在于其在样本量过多的时候会更加准确。

结果目录及文件

result/Nmds (result/Nmds)

- fig-nmds.pdf 文件说明
- nmds-sites.xls 文件说明

结果预览

Nmds/1/OTU_Bray | Nmds/2/OTU_Bray

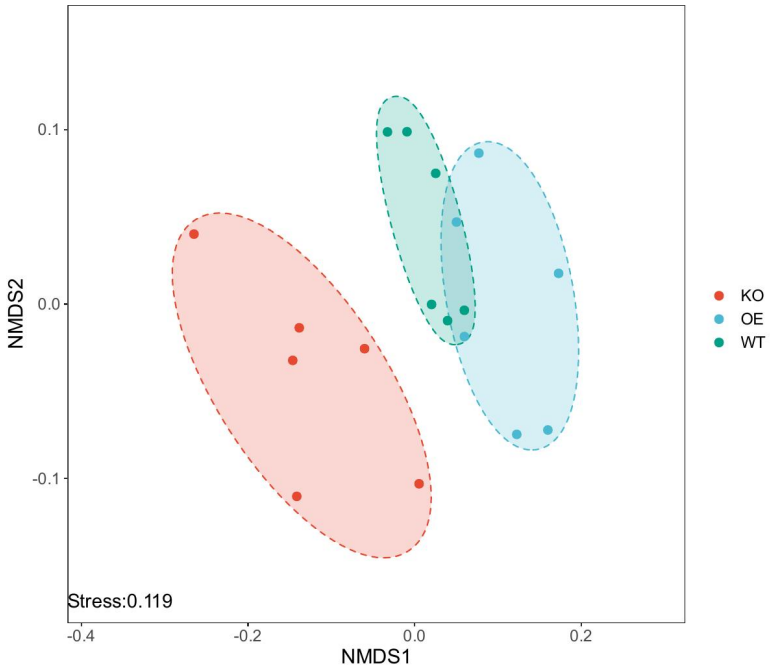


Fig. NMDS plot

注：NMDS图形中的点代表样本，不同颜色/形状代表样本所属的分组信息。各样本点连线的距离体现每个样本之间的相似性，距离越短，相似性越大。同组样本点距离远近说明了样本的重复性强弱，不同组样本的远近则反应了组间样本距离在秩次(数据排名)上的差异。样本相似性距离计算方式对结果有影响，选择输入不同相似性距离值的矩阵，得到的结果存在不同程度的差异。NMDS是距离值的秩次(数据排名)信息的评估，图形上样本信息仅反映样本间数据秩次信息的远近，而不反映真实的数值差异，横纵坐标轴并无权重意义，横轴不一定比纵轴更加重要。NMDS图形通常会给出该模型的stress值，用于判断该图形是否能准确反映数据排序的真实分布，stress值越接近0则降维效果越好，一般要求该值<0.1。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；当stress＜0.01认为结果极好。

7.6 ANOSIM分析

ANOSIM，即相似性分析（Analysis of similarities），是一种用于比较组间(两组或多组)的差异是否显著大于组内差异的非参数检验，通过判断群落结构差异，从而判断分组是否有意义。默认利用bray_curtis算法计算两两样品间的距离，然后将所有距离从小到大进行排序，最小的距离记为1，依次类推为2，3，4……基于排序后的关系排名（秩）计算R， R的值域范围是[-1, 1]。R<0表示组内差异大于组间差异，说明有可能实验设计存在缺陷，或者用于分析的数据存在问题。R=0表示组间没有差异，说明实验组和对照组之间没有差异。R>0表示组间差异大于组内差异，说明实验组和

对照组之间存在差异。 统计分析的可信度用P-value表示，常用的检验方法是Permutat
换检验)，将样本打乱随机分组，计算置换后的R值（记为Ri)，经过N次置换后（至少1000次)，Ri
大于原始R的概率即为p，p<0.05表示统计具有显著性。

结果目录及文件

result/Anosim (result/Anosim)

- fig-*.pdf 文件说明
- anosim-output.xls 文件说明

结果预览

Anosim/1/OTU_Bray | Anosim/2/OTU_Bray

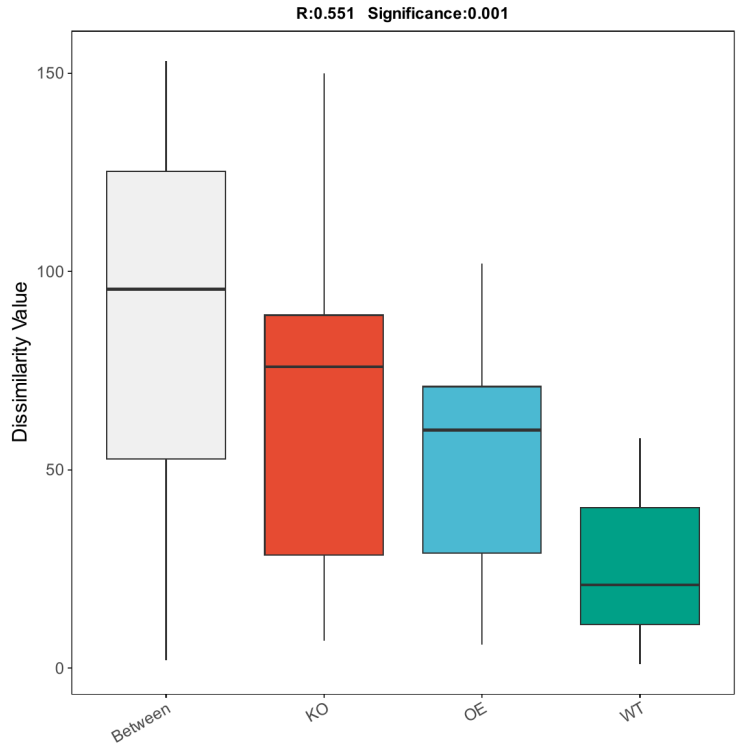


Fig. Boxplot of dissimilarity between groups

注：图上总共有N+1个箱子，N为分组数量，“Between”指代的是组间差异，其他箱子分别代表各组组内差异。R表示差异程度，一般介于（0，1）之间。R>0，说明组间存在差异（R>0.75：大差异；R>0.5：中等差异，R>0.25：小差异)；R=0或在0附近，表明组间没有差异；若R出现<0的情况，说明组内差异显著大于组间差异，这个时候表明分组或采样不合理，需要重做实验。Significance表示P值，数值小于0.05、0.01或在两者范围内，表明分组有显著性差异，进而反映出目标分组有意义。

8 物种差异分析 ⭐

8.1 LEfSe线性判别效应大小分析

LEfSe分析即LDA Effect Size分析，是一种用于发现和解释高维度数据生物标识(基因、通路和分类单元等)的分析工具，可以进行两个或多个分组的比较，它强调统计意义和生物相关性，能够在组与组之间寻找具有统计学差异的生物标识（Biomarker）。LEfSe分析主要有三大步骤：首先在多组样本中采用非参数因子Kruskal-Wallis秩和检验检测不同分组间丰度差异显著的生物学特征；再利用Wilcoxon秩和检验检查显著差异的生物学特征在分组亚组之间是否都趋同于同一分类（如果存在分组亚组时）；最后用线性判别分析（LDA）对数据进行降维和评估差异显著的生物学特征的影响力（即LDA score）。

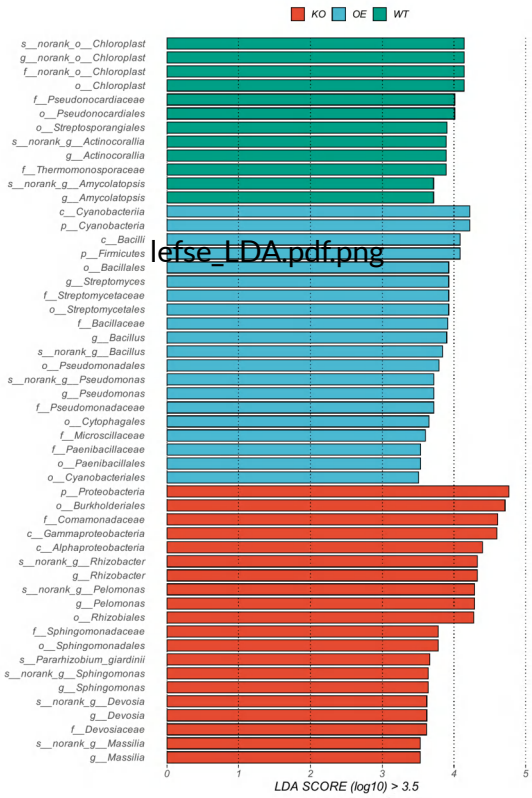
结果目录及文件

result/Lefse (result/Lefse)

- lefse_all.xls 文件说明
- lefse_LDA.xls 文件说明
- lefse_LDA.pdf 文件说明
- lefse_cladogram.pdf 文件说明

结果预览

Lefse/1/Taxa | Lefse/2/Taxa



lefse_LDA.pdf.png

注：条形图展示LDA score大于预设值的 Biomaker，默认预设值为2.0（看横坐标，只有LDA值的 颜色代表各自的组别，长短代表的是LDA score，即不同 式图中，图中由内至外辐射的圆圈代表了由门至属的分 分类级别上的每一个小圆圈代表该水平下的一个分类， 无显著差异的物种统一着色为黄色，差异显著的Biomar 在红色组别中起到重要作用的微生物类群，蓝色节点表 。未能在图中显示的Biomarker对应的物种名会展示在右侧，字母编号与图中对应。

8.2 Wilcoxon秩和检验

用于比较两组样本之间的特征丰度，通过此分析可获得显著性差异的特征。

结果目录及文件

result/Wilcoxon (result/Wilcoxon)

- *-full-out.xls 文件说明
- *-statistics-out.xls 文件说明
- *-filtered-statistics-out.xls 文件说明
- *-fig-stamp.pdf 文件说明

结果预览

Wilcoxon/2/Genus

本次分析无显著差异

Fig. Proportion difference analysis

注：左图所示为不同特征在各样本中的丰度占比，中间所示为95%置信度区间内差异效应，最右边为p值，p值＜0.05，表示差异显著。

8.3 Kruskal检验

用于比较两组以上样本之间的特征丰度，通过此分析可获得显著性差异的特征。

结果目录及文件

result/Kruskal (result/Kruskal)

- *-full-out.xls 文件说明
- *-statistics-out.xls 文件说明
- *-filtered-statistics-out.xls 文件说明
- *-fig-stamp.pdf 文件说明

结果预览

Kruskal/1/Genus

本次分析无显著差异

Fig. Proportion difference analysis

注：左图所示为不同特征在各样本中的丰度占比，中间所示为95%置信度区间内差异效应，最右边为p值，p值＜0.05，表示差异显著。

9 功能预测分析

9.1 PICRUSt2

PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) 是一个从标记基因（一般是16S rRNA）测序数据预测功能丰度的软件，支持基于多个基因家族数据库的预测，默认包括KEGG同源基因、KO直系同源物和EC酶分类编号。

结果目录及文件

result/Picrust2 (result/Picrust2)

- MetaCyc_pathways/KEGG_pathways 文件说明
  - path_abun_unstrat_descrip.tsv.gz 文件说明
  - path_abun_contrib_legacy.tsv.gz 文件说明
- KO_metagenome_out/EC_metagenome_out 文件说明
  - weighted_nsti.tsv.gz 文件说明
  - pred_metagenome_unstrat_descrip.tsv.gz 文件说明
  - pred_metagenome_contrib_legacy.tsv.gz 文件说明

9.2 功能单元PCoA分析

结果目录及文件

result/func_Pcoa (result/func_Pcoa)

- fig-pcoa-*.pdf 文件说明
- pcoa-importance.xls 文件说明
- pcoa-sites.xls 文件说明
- adonis-output.xls 文件说明

结果预览

func_Pcoa/1/EC_Bray | func_Pcoa/1/KO_Bray | func_Pcoa/1/KO_Jaccard | func_Pcoa/1/EC_Jaccard | func_Pcoa/2/EC_Bray | func_Pcoa/2/KO_Bray | func_Pcoa/2/KO_Jaccard | func_Pcoa/2/EC_Jaccard

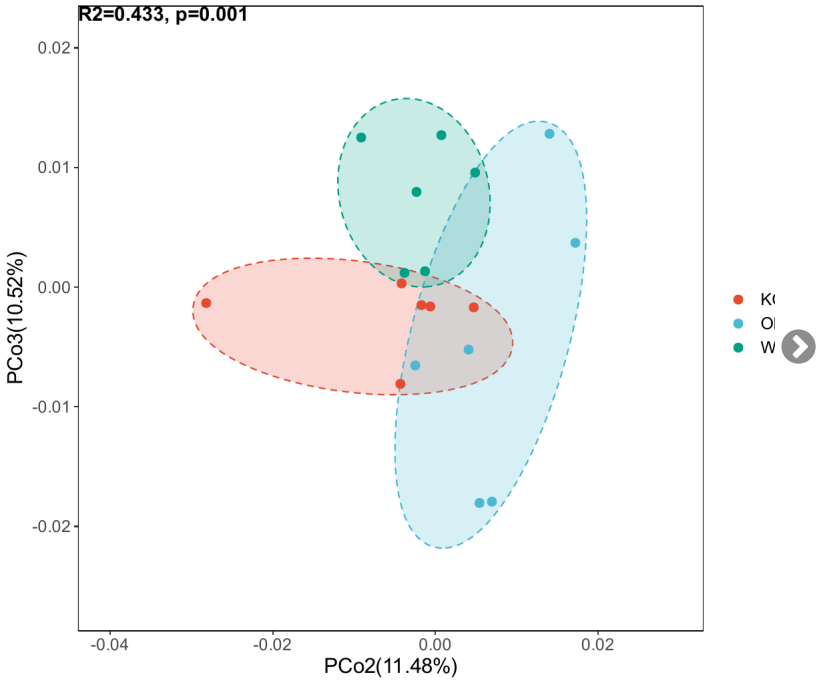


fig-pcoa-PCo2PCo3.pdf.png

Fig. PCoA plot

注：PCoA图形为散点图，图中的点代表样本，不同颜色/形状代表样本所属的分组信息，连线的距离体现每个样本之间的相似性，距离越短，相似性越大。同组样本距离远近说明了样本的重复性强弱，不同组样本的远近则反应了组间群落差异。通常情况下，来自不同环境的样本表现出各自聚集的现象。图中的横/纵轴分别代表了第一、第二主成分，横纵轴上所标注的百分比即该主成分对样品差异的贡献度，通常横轴百分比数值高于纵轴。

9.3 代谢通路LEfSe差异分析

结果目录及文件

result/func_Lefse (result/func_Lefse)

- lefse_all.xls 文件说明
- lefse_LDA.xls 文件说明
- lefse_LDA.pdf 文件说明
- lefse_cladogram.pdf 文件说明

结果预览

func_Lefse/1/KEGG_Path | func_Lefse/1/MetaCyc_Path | func_Lefse/2/KEGG_Path | func_Lefse/2/MetaCyc_Path

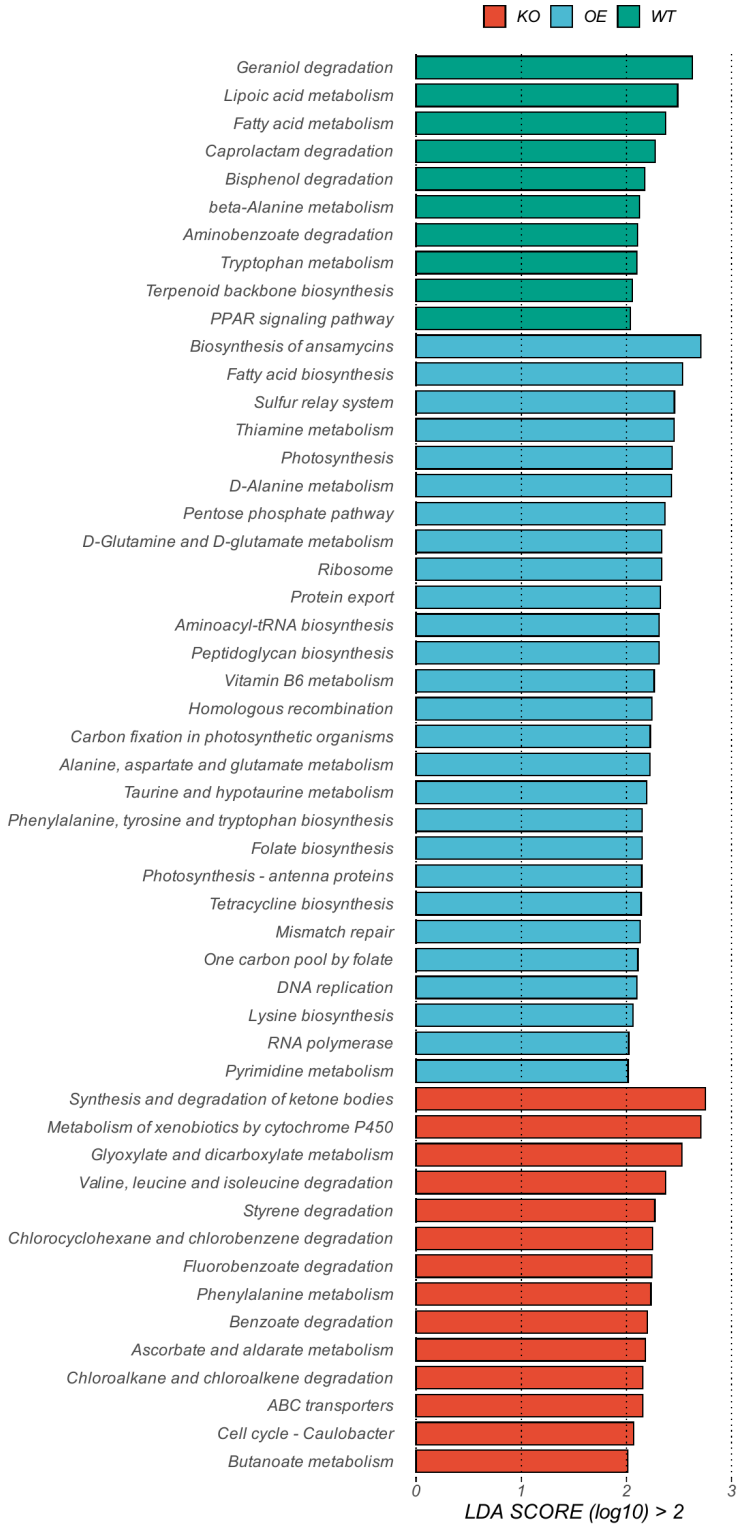


Fig. Lefse plot

注：条形图展示LDA score大于预设值的显著差异Biomaker，即具有统计学差异的Biomaker，默认预设值为2.0（看横坐标，只有LDA值的绝对值大于2才会显示在图中）；柱状图的颜色代表各自的组别，长短代表的是LDA score，即不同组间显著差异Biomarker的影响程度。

9.4 代谢通路MetagenomeSeq分析

结果目录及文件

result/func_MetagenomeSeq (result/func_MetagenomeSeq)

- metagenomeSeq_all.xls 文件说明
- metagenomeSeq_filtered.xls 文件说明
- fig_metagenomeSeq_bar.pdf 文件说明
- fig_metagenomeSeq_dot.pdf 文件说明

结果预览

func_MetagenomeSeq/1/KEGG_Path	func_MetagenomeSeq/1/MetaCyc_Path
func_MetagenomeSeq/2/KEGG_Path	func_MetagenomeSeq/2/MetaCyc_Path

本次分析无显著差异

Fig. MetagenomeSeq analysis

注：

条形图：横坐标表示log2(fold change)，纵坐标表示代谢通路名称，不同颜色表示显著性程度。

气泡图：横坐标表示log2(fold change)，纵坐标表示显著性水平，不同颜色表示通路富集的组别。

10 附录

10.1 分析软件列表

Show 10 entries

Search:

名称	版本	描述	链接
fastp	0.23.2	一种旨在为FastQ文件提供快速一体化预处理的工具，常用于原始数据质控	https://academic.oup.com/bioinformatics/article/34/17/i884/509
FLASh	1.2.11	一种非常快速和准确的软件工具，用于拼接二代测序实验中的成对末端读数	https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr507

Vsearch	2.22.1	一个开放源码和免费的多线程64位工具，用于处理和准备宏基因组学、基因组学和群体基因组学核酸序列数据	https://peerj.com/articles/2584/
QIIME2	2022.8		https://www.nature.com/articles/s41587-019-0209-9
RDP Classifier	2.13		https://journals.asm.org/doi/full/10.1128/AEM.00062-07

Python	3.9.1	一种广泛使用的解释型、高级编程、通用型编程语言，用于编写自定义脚本	https://www.python.org/
R	4.1.3	R 语言是为数学研究工作者设计的一种数学编程语言，主要用于统计分析、绘图、数据挖掘	https://www.r-project.org/
R包	-	数据清洗、绘图与统计	dplyr、reshape2、ggsci、RColorBrewer、vegan、ggplot2、VennDiagram、ggheatmap、ggtree

10.2 方法学描述

详见中英文方法学文档 (material/methods.pdf)

10.3 参考文献

1. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.
2. Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27(21), 2957-2963.
3. Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584.
4. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., … & Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology, 37(8), 852-857.
5. Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16), 5261-5267.