微生物 16S 扩增子测序分析
结题报告

项目名称：	16s rRNA 测序分析
检测类型：	16S 扩增子测序
客户姓名：	XXX
客户单位：	XXXXX
制定日期：	2023 年 10 月 30 日

目录

1 实验流程

实验室收到样品后，首先对接收的样本进行核对后保存于超低温冰箱，执行严格的样本质控流程，合格样本才会进入正式的实验流程。

合格的样品首先根据不同的样品来源，选用最合适的总 DNA 提取方法进行核酸提取并检测提取质量和测定核酸浓度。通常以能够反映菌群组成和多样性的细菌核糖体 RNA 序列为靶点，根据序列中的保守区域 16S 的 V3-V4 区设计相应引物，并添加样本特异性 Barcode 序列，进而对合格样品的 16S 的 V3-V4 区域进行 PCR 扩增。扩增完成后，利用磁珠纯化回收扩增产物并进行荧光定量，根据荧光定量结果，按照每个样本的测序量需求，对各样本按相应比例进行混合。采用 Illumina 公司的 TruSeq Nano DNA LT Library Prep Kit 制备测序文库，文库经质检合格后，使用二代高通量平台 NovaSeq6000 进行 PE250 测序。实验流程如下：

图 1 微生物 16S 扩增子实验流程图

2 项目分析方法及结果

2.1 测序数据基本统计

为了获得可供分析的数据，我们将下机后得到的原始数据导入 QIIME2[1]，使用 QIIME2 中的分析流程实施处理，包括去除 PE reads 接头，通过内置的 DADA2[2]进行数据质控、去噪、拼接和去嵌合体等一系列操作，最后按 99%的相似水平对序列进行聚类得到 ASV 特征表。

ASV 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）设置的同一标志。要了解一个样本测序结果中的菌属、菌种等数目信息，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 ASV。可根据不同的相似度水平，对所有序列进行 ASV 划分，通常在 99%的相似水平下的 ASV 进行生物信息统计分析。

基于 ASV 可以进行多种多样性指数分析，基于 ASV 聚类分析结果，可以对 ASV 表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，可以进行一系列群落结构和系统发育等深入的统计学和可视化分析。

每一步数据后的样本序列统计如下表：

表一 数据处理统计表

sample-id	input	filtered	denoised	merged	non-chimeric	percentage of data process
A1002	134360	127627	127417	126967	125921	93.72
A1020	107251	102091	101504	99395	93864	87.52
A1113	91389	86734	85532	80183	65516	71.69
A1233	140202	133134	132015	126651	94709	67.55
A1287	121773	115615	114833	111469	92212	75.72
A1316	109091	103208	100013	85115	57175	52.41
A1323	124199	118527	118239	117510	114364	92.08
A1679	101151	95983	94424	89414	76060	75.19
A1704	117512	111600	111114	109836	108092	91.98
A1735	128249	121596	120853	117726	113140	88.22
B1756	131312	124306	123581	120087	96953	73.83
B1782	113848	107935	107163	104150	91611	80.47
B1785	99260	93868	92941	87540	68153	68.66
B1787	110962	105103	103545	94718	64936	58.52
B1819	142835	136017	135444	133672	130526	91.38
B1841	100249	95315	93948	88010	72449	72.27
B1847	90554	85807	83170	71007	51336	56.69

B1855	74082	70468	69990	68262	63690	85.97
B1947	90266	85642	84428	79533	60431	66.95
B1953	122776	116205	114200	104009	64070	52.18

各列意义如下：

input：最开始输入的序列数；
filtered：去引物、质量过滤后的序列数；
denoised：去噪后的序列数；
merged：拼接成功的序列数；
non-chimeric：鉴定为非嵌合体的序列数；
percentage of data process：处理后序列数占原始序列数的百分比。

2.2 ASV 过滤

得到特征表以后，为了使用更加可靠的数据结果进行分析，ASV 特征表需要进行过滤操作。低表达水平的 ASV 会增加计算工作量，以及影响高丰度结果差异比较时 FDR 校正 Pvalue 而导致假阴性，通常需要选择一定的阈值进行筛选，这里过滤掉丰富度小于 100 以及在单样本中出现的 ASV；使用 QIIME2 自带的 classifier 对 RDP[3](v11.5)数据库训练 Naive Bayes[4] 分类器，在分析过程中，并非所有 ASV 代表序列都被收录到了参考数据库中，未能比对到参考数据库中的代表序列在所有分类水平都是 ‘Unclassified’，这些代表序列的 ASV 也进行去除。

最后使用 QIIME2 对序列进行抽平处理，使得分析的样本丰度处于相同深度下，即样本总物种丰度相同。这样使得分析的样本处于同一水平进行分析，大大减少误差。我们把最小的丰度值 38181 作为抽平的深度，使用抽平后的数据进行后续分析。

每一步过滤后的样本序列统计和抽平后部分样本的 ASV 如下：

表二　ASV 特征过滤统计表

sample-id	input	filter-feature	classified	rarefied	percentage of input rarefied
A1002	134360	123644	123644	38181	28.42
A1020	107251	93037	92114	38181	35.6
A1113	91389	61354	61354	38181	41.78
A1233	140202	91015	90997	38181	27.23
A1287	121773	90783	90610	38181	31.35
A1316	109091	50835	50835	38181	35
A1323	124199	102402	102402	38181	30.74
A1679	101151	73160	57638	38181	37.75
A1704	117512	106019	106019	38181	32.49
A1735	128249	110273	110273	38181	29.77
B1756	131312	94526	94526	38181	29.08
B1782	113848	90165	90165	38181	33.54

B1785	99260	65675	65650	38181	38.47
B1787	110962	63706	63706	38181	34.41
B1819	142835	127534	127531	38181	26.73
B1841	100249	71445	71445	38181	38.09
B1847	90554	48959	48674	38181	42.16
B1855	74082	63194	63191	38181	51.54
B1947	90266	58233	58233	38181	42.3
B1953	122776	62090	62090	38181	31.1

各列意义如下：

input：过滤前的特征序列数；

filter-feature：丰度过滤后的特征序列数；

classified：物种鉴定后的特征序列数；

rarefied：抽平后的特征序列数；

percentage of input rarefied：过滤后的特征序列数占过滤前特征序列数的百分比。

3 ASV 统计

3.1 物种注释

为了得到每个 ASV 对应的物种分类信息，使用训练好的分类器对 ASV 代表序列进行物种注释，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。得到 ASV 注释表，根据分类学分析结果，可以得知一个或多个样本在各分类水平上的分类学比对情况。在结果中，包含了两个信息：

1）样本中含有何种微生物；

2）各微生物的序列数，即各微生物的绝对丰度。

3.2 VENN 图

得到各个样品点对应的 ASV，根据不同样品或者分组类别的 ASV 组成， 通过 venn 图可以直观的展示 2-5 个样品或者组别的特有，共有或者差异 ASV 数量。绘图通过 R[5](3.2.5)语言的 VennDiagram[6]包做出。不同颜色填充的图形代表不同的样品或者组别，从 Venn 图又外往里看，单个椭圆形图形覆盖的区域表示样品点或者组别特有的 ASV，越往 Venn 图中心，图形与其他椭圆重叠区域表示与其他有重叠关系的样品点共有的 ASV，Venn 图椭圆都覆盖的区域表示所有样品点或者组别共有的 ASV 数。分组数大于 5 时不展示 Venn 图。

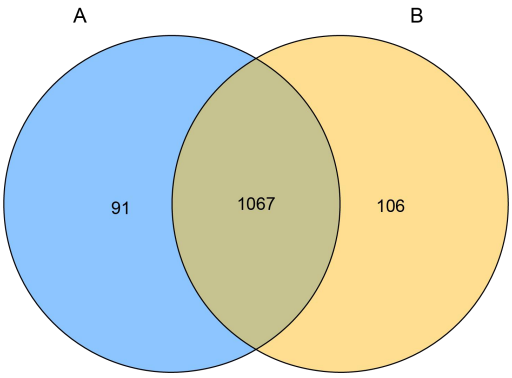


图 2 Venn 图（分组<=5）

3.3 物种柱状图

ASV 数据经过与数据库的比对，得到包括界、门、 纲、 目、 科、 属等几个分类等级的物种分类信息。依据不同分类等级信息和各个样品点或者组别内物种百分比绘制柱状图。物种丰度在所有样品低于 0.01%时，归类于 Others。

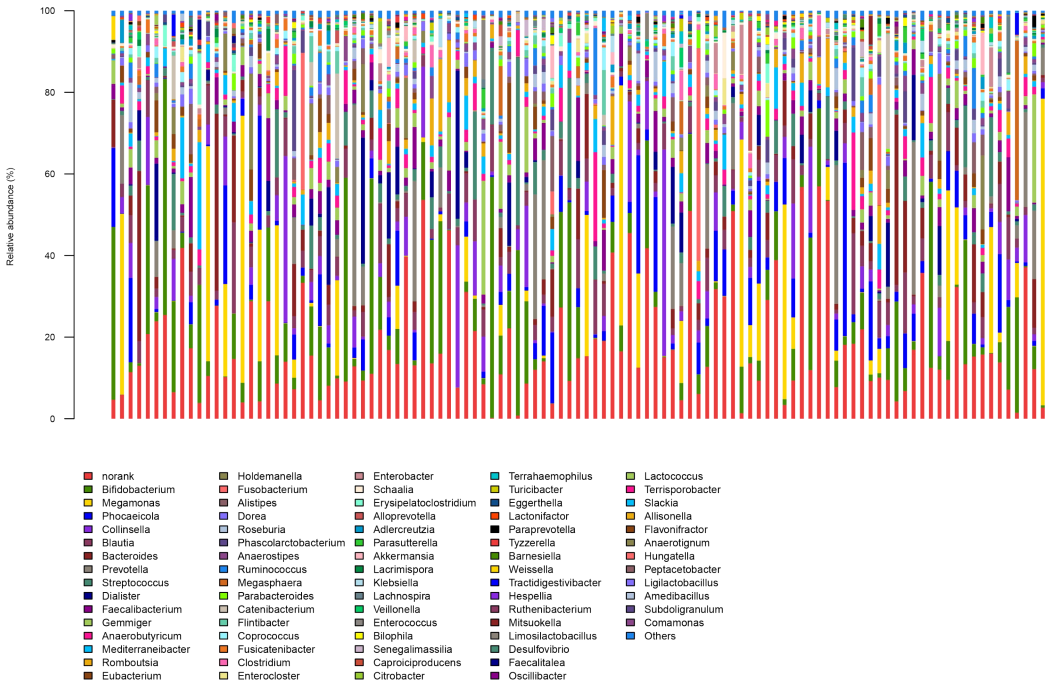


图 3 属水平柱状图

3.4 物种热图分析

热图可以用颜色变化来反映二维矩阵或表格中的数据信息，它可以直观地将数据值的大小以定义的颜色深浅表示出来，可将高丰度和低丰度的物种分块聚集，通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的相似性和差异性。热图分析包括门，纲，目，科，属分

类等级，为了更方便的分析每个物种在每个样品的相对丰度，故对丰度进行了 log10 转化。 热图通过 R（V3.2.5）语言中 ggplot2[7]包绘制.

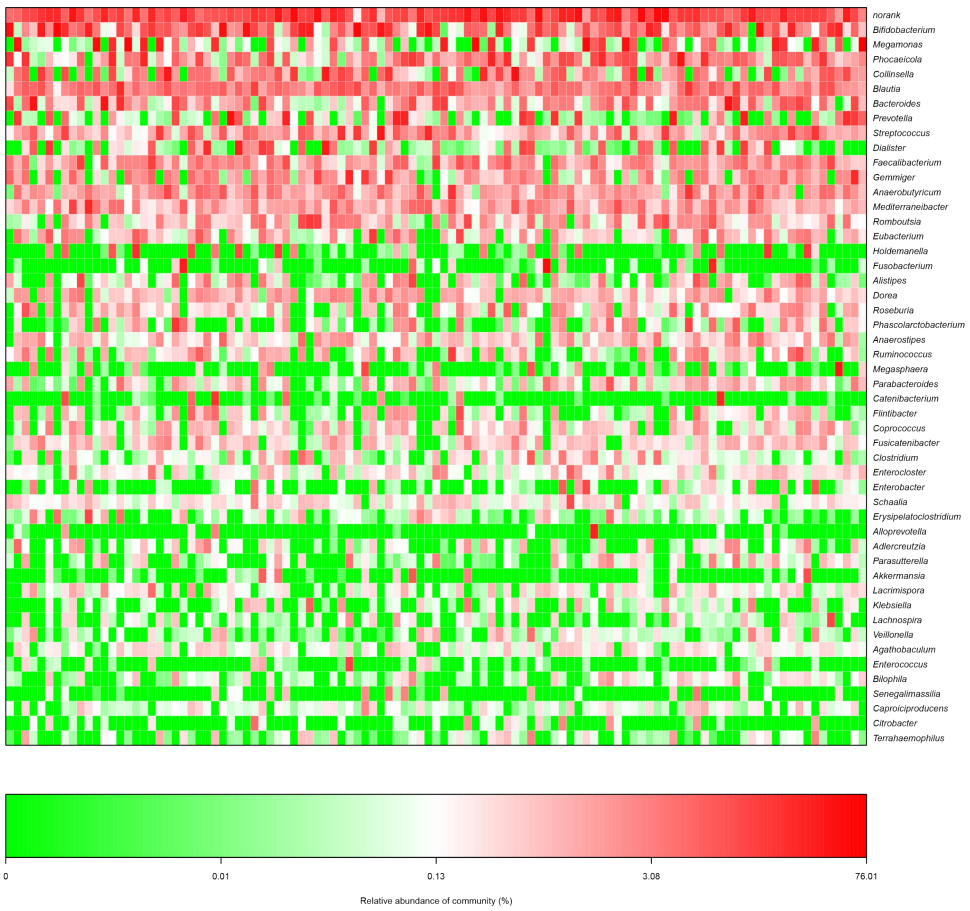


图 4　属水平热图

4　Alpha 多样性分析

4.1 Alpha 分析

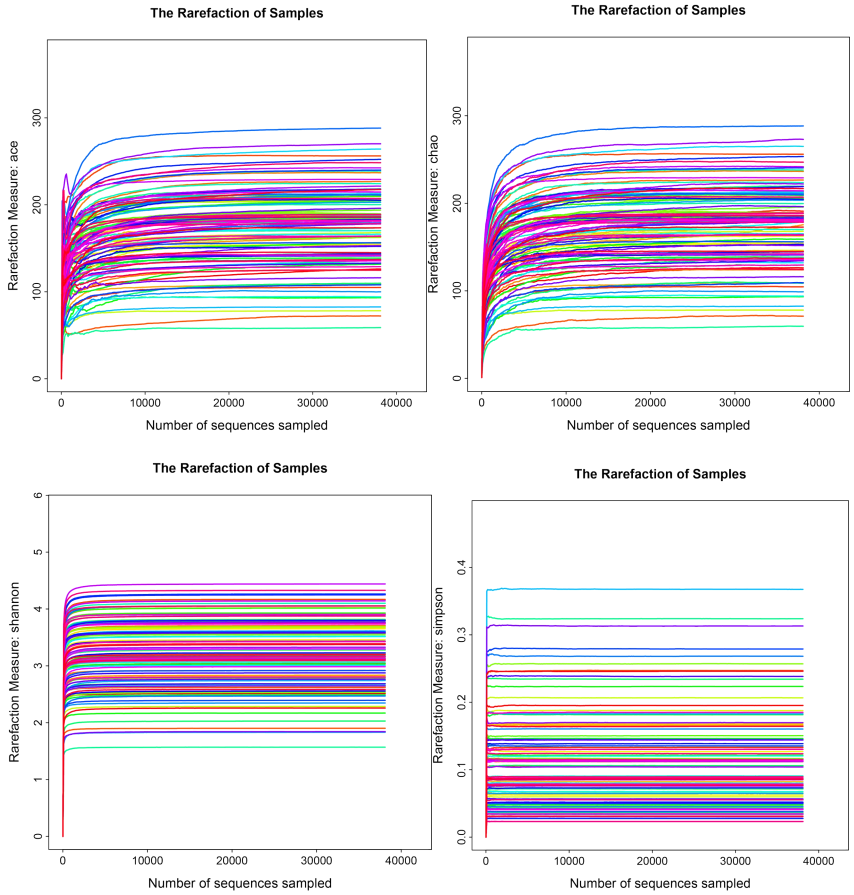
Alpha 多样性(https://en.wikipedia.org/wiki/Alpha_diversity)分析是研究特定环境内或者单个样品内的多样性，分析结果通过软件 Mothur[8]完成，数据通过 R 软件（v3.2.5）实现可视化。常用的度量指标包括：observed species 指数，Ace 指数，Chao 指数，Shannon 指数，Simpson 指数等。其中，observed species 指数，Ace 和 Chao 指数是估计群落中的 ASV 数目，根据 3 项指数所绘制的稀疏曲线，如果曲线平缓说明测序深度足够且更准确反映样品内的物种数量，反之，测序深度不够，还有物种未被检测到。Shannon 和 Simpson 指数用来计算菌群的多样性指数，通过丰富度指数和均匀度指数结合起来才能有效反映群落内物种多样性。 Shannon 指数与前面 3 项指数呈正相关，shannon 越大多样性越丰富，simpson 则相反，多样性越丰富数值越小。

表三　物种 Alpha 多样性统计

Group	Observe	ACE	Chao1	Shannon	Simpson
A	178.94 +/- 43.9189	180.9918 +/- 44.2859	180.9762 +/- 44.3428	3.2679 +/- 0.6353	0.8868 +/- 0.0799
B	167.7119 +/- 45.166	169.5947 +/- 45.0227	169.9744 +/- 45.2216	3.25 +/- 0.6028	0.8982 +/- 0.0666

4.2 稀疏性曲线图

稀释性曲线[9]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 ASV 的数目构建 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 ASV，反之则表明继续测序还可能产生较多新的 ASV。因此，通过作稀释性曲线，可得出样本的测序深度情况。其中，横坐标代表序列的数目，纵坐标为 Alpha 多样性分析指数，不同颜色代表不同的样品。



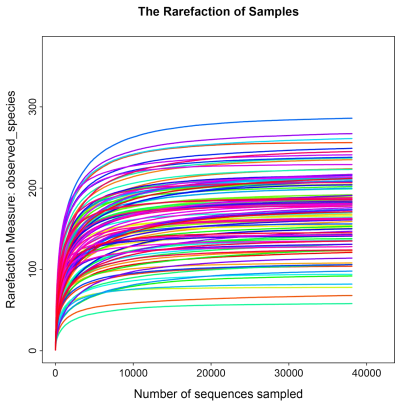
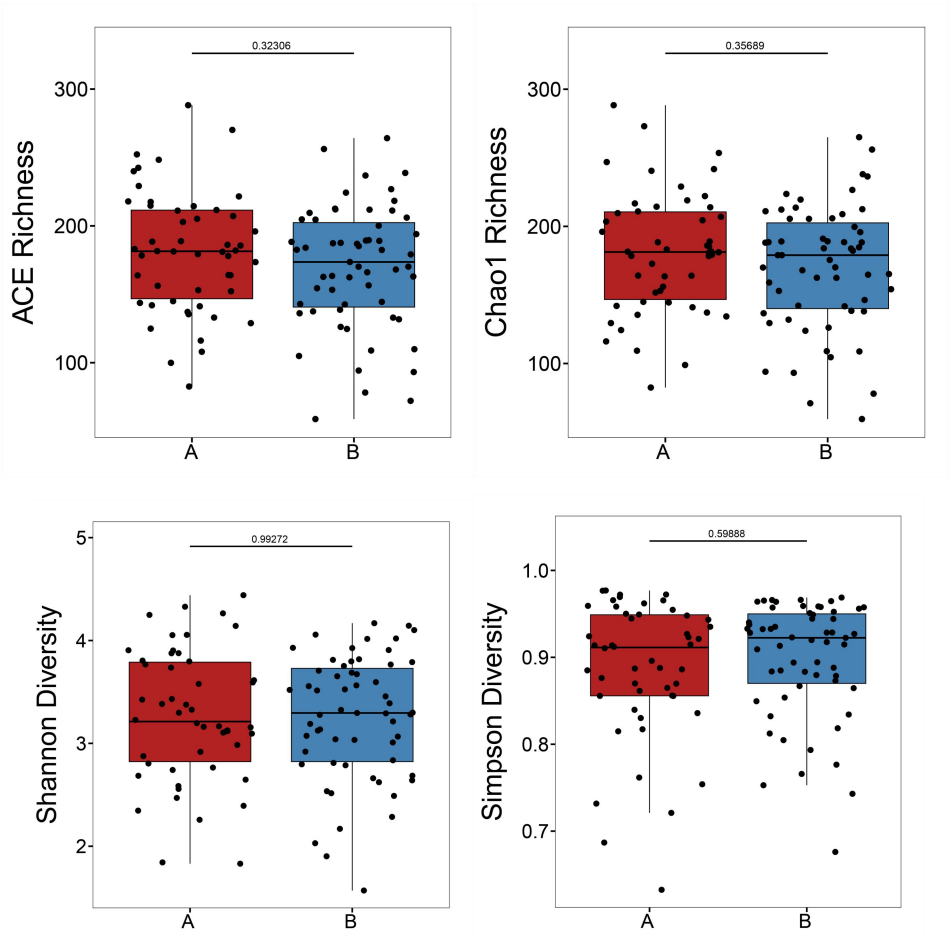


图 5　稀疏性曲线图

4.3 盒型图

Alpha 多样性盒形图，更加直观地展示不同分组中多样性差异。盒形图可以显示 5 个统计量（最小值，第一个四分位数，中位数，第三个中位数和最大值，及由下到上的 5 条线），异常值以“o”标出。如图所示：



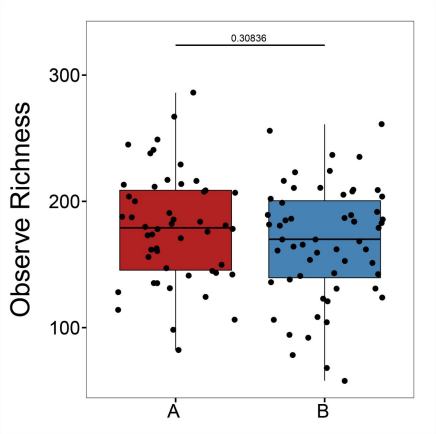


图 6 盒型图

5 Beta 多样性分析

5.1 Beta 分析

与 Alpha 多样性分析相比，Beta 多样性(https://en.wikipedia.org/wiki/Beta_diversity)则是分析不同区域间物种组成的差异。首先因为样品中不同测序深度，测出来的序列条数不一样，为了统一测序深度，以样品中序列数最少的那个样品的序列条数作为统一的标准来随机抽取序列进行分析。

5.2 样本聚类分析

根据样品中的 ASV 组成情况，展示多个样品之间的关系，样品越相似，相互之间越靠近，枝长越短。聚类分析通过 Qiime 软件进行，在 UniFrac 和加权 UniFrac 两种情况下采用迭代算法，聚类方法为 UPGMA，以样品中序列数的最小的 75%进行抽样分析，迭代 100 次，结果通过 R(v3.2.5)绘制出图形。

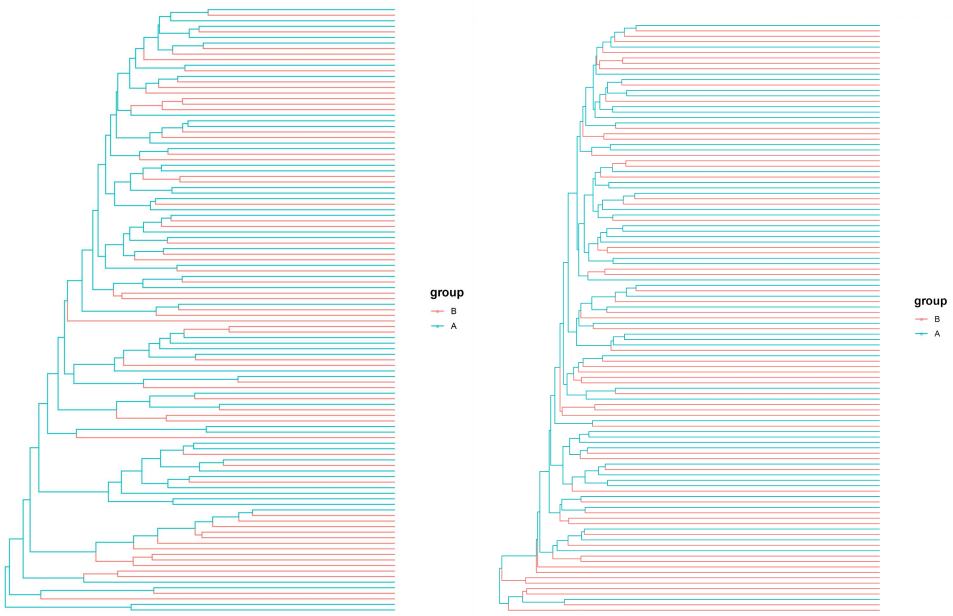


图 7　样本聚类分析图 加权(左)和未加权(右)

5.3 PCA 分析

PCA 分析(Principal Component Analysis)，即主成分分析，是一种分析和简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。通过分析不同样品 ASV 组成可以反映样品的差异和距离，PCA 运用方差分解，将多组数据的差异反映在二维坐标图上，坐标轴取能够最大反映方差值两个特征值。如果两个样品距离越近，则表示这两个样品的组成越相似。不同处理或不同环境间的样品可能表现出分散和聚集的分布情况，从而可以判断相同条件的样品组成是否具有相似性。图中，横坐标表示第一主成分，括号中的百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，括号中的百分比表示第二主成分对样品差异的贡献值；图中各点分别表示各个样品。不同颜色代表样品属于不同的分组。

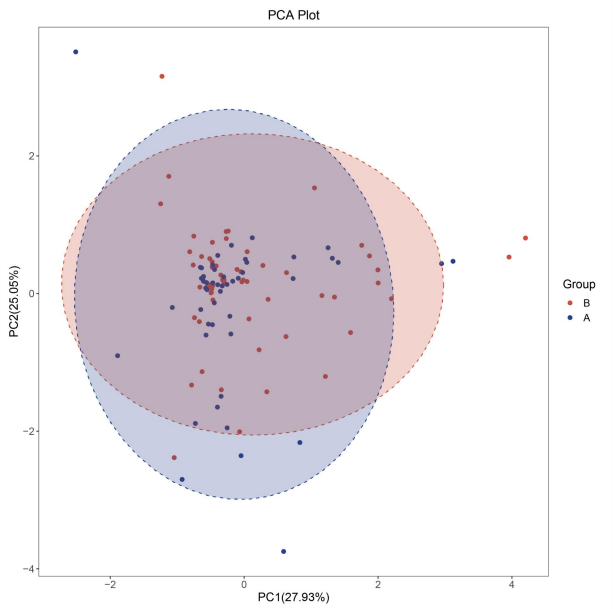


图 8　基于 ASV 丰度的 PCA 分析

5.4 PCoA 分析

为了进一步了解样品间物种多样性关系，通过主坐标分析 PCoA（Principal coordinates analysis）研究样品中各个群体及个体间的位置关系反映他们在遗传上的相似性，如果两个样品距离越近，则表示这两个样品的物种组成越相似。与非限制性排序分析 PCA 分析相比，PCA 分析基于原始的物种组成矩阵做排序分析，而 PCoA 分析基于物种组成计算得到的距离矩阵，计算距离矩阵利用样品中物种的进化信息。通过软件 QIIME2 完成 PCoA 分析，所有样品以样品中序列最小的样品 75%的序列数为标准进行抽样分析，采用迭代算法，在加权和不加权物种丰度的情况下，迭代 100 次最终得到结果。

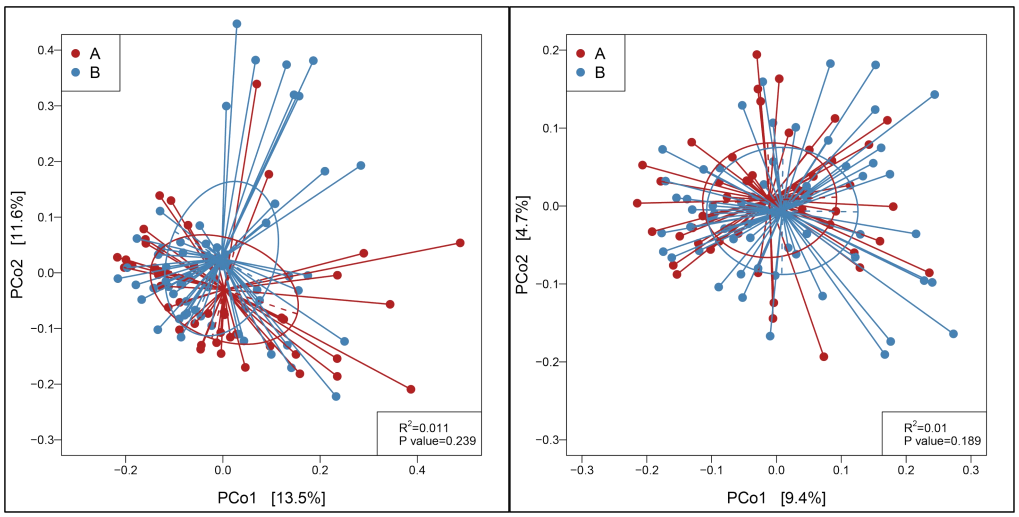


图 9　PCoA 分析图 加权(左)和未加权(右)

6 LEfSe 分析

LEfSe[10] (Linear discriminant analysis Effect Size) 是一种用于发现高维生物标识和揭示基因组特征的软件，用于区别两个或两个以上生物条件（或者是类群）。LEfSe 通过生物学统计差异使其具有强大的识别功能，然后评估这些差异是否符合预期的生物学行为。具体来说，首先使用 non-parametric factorial Kruskal-Wallis (KW) sum-rank test（非参数因子克鲁斯卡尔—沃利斯和秩验检）检测具有显著丰度差异特征，并找到与丰度有显著性差异的类群。最后，LEfSe 采用线性判别分析（LDA）来估算每个组分（物种）丰度对差异效果影响的大小。

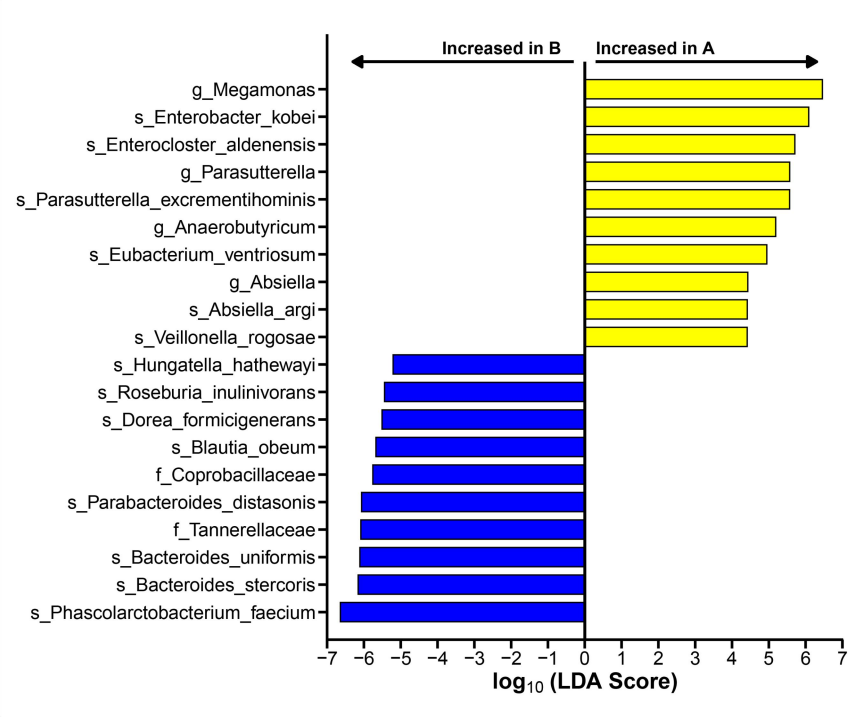


图 10 LEfSe 差异物种展示条形图

7 参考文献

1. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37:852–857. https://doi.org/10.1038/s41587-019-0209-9.
2. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul;13(7):581-3. doi: 10.1038/nmeth.3869. Epub 2016 May 23. PMID: 27214047; PMCID: PMC4927377.
3. Cole JR, Tiedje JM. History and impact of RDP: a legacy from Carl Woese to microbiology. RNA Biol. 2014;11(3):239-43. doi: 10.4161/rna.28306. Epub 2014 Feb 27. PMID: 24607969; PMCID: PMC4008557.
4. Zhang, H. (2004). The Optimality of Naive Bayes. The Florida AI Research Society.
5. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. Hanbo Chen (2022). VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.7.3. https://CRAN.R-project.org/package=VennDiagram.
7. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
8. Schloss PD et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75:7537–7541.
9. Andrew F. Siegel (2006). "Rarefaction Curves". In Kotz, Samuel; Read, Campbell B.; Balakrishnan, N; Vidakovic, Brani (eds.). Encyclopedia of Statistical Sciences. doi:10.1002/0471667196.ess2195.pub2. ISBN 9780471667193.
10. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011 Jun 24;12(6):R60. doi: 10.1186/gb-2011-12-6-r60. PMID: 21702898; PMCID: PMC3218848.