

16S rDNA 测序分析总结报告

客户: Demo

客户单位: Demo

项目编码: Demo

项目经理: Demo

您值得信赖的基因组学、转录组学与代谢组学技术合作伙伴

一、项目简介

16S rDNA 是一种对特定环境（或者特定生境）样品中所有的细菌进行高通量测序，以研究环境样品中微生物群体的组成，解读微生物群体的多样性、丰富度及群体结构，探究微生物与环境或宿主之间的关系的技术。传统的微生物研究依赖于实验室培养，16S 扩增子等高通量测序的兴起填补了对无法在传统实验室中培养的微生物的研究空白，扩展了对微生物资源的利用空间，为研究微生物相互作用提供了有效工具。16S rDNA 位于原核细胞核糖体小亚基上，指的是基因组中编码核糖体 16S rDNA 分子对应的 DNA 序列，也就是 16S rDNA 的编码基因。该基因全长约 1542bp，由 9 个可变区和 10 个保守区组成（可变区为 V1 到 V9），其中保守区反映了生物物种间的亲缘关系，而可变区则表明物种间的差异，且变异程度与细菌的系统发育密切相关，被认为是最适于细菌系统发育和分类鉴定的指标。

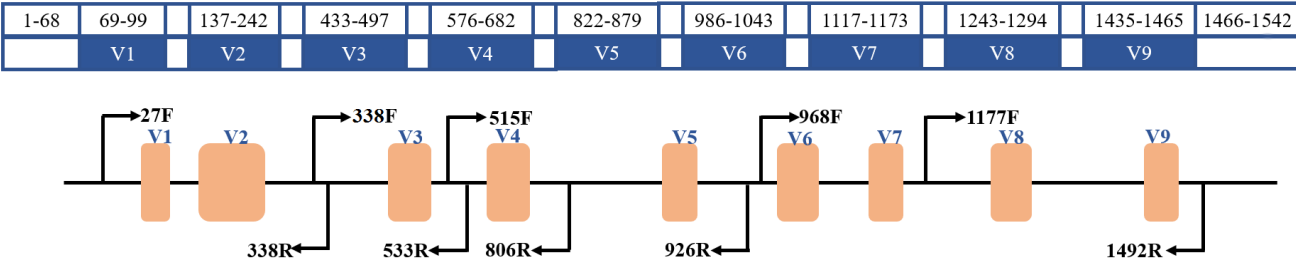


图 1 细菌 16S 区域

越来越多的研究表明，微生物与人类健康密不可分，如生长（如营养不良）、发育（如免疫系统发育成熟）、代谢疾病（如肥胖）以及很多癌症（结肠癌，直肠癌，乳腺癌，肝癌等）的发生都与肠道菌群紊乱有关。此外，微生物在农业生产和环境变换中也扮演者十分重要的作用。根系微生物与植物之间的互作关系、植物内生菌对植物生长的影响、空气（雾霾）水体（海洋，冰川，污水等）中微生物等也都是近年来研究的热点。二代测序技术的横空出世，无需培养、快速产出、数据庞大等各种优势使得微生物研究进入万千科学学者的视野。针对不同项目的特点，我们整理出不同项目的优势与能解决的问题，以促进后续科学研究有的放矢。常见的问题，如为什么要设置生物学重复，为什么 16S 结果中注释到叶绿体，我们的 FAQ 中都有相应说明，也欢迎各位老师补充。



图 2 微生物高通量测序项目

二、工作流程

2.1 实验流程



图 3 16S rDNA 实验流程

2.2 信息分析流程

上机测序完成后，我们得到原始的下机数据，利用 overlap 将双端数据进行拼接，并进行质控、嵌合体过滤，获得高质量的 cleandata。DADA2 (Divisive Amplicon Denoising Algorithm) [1]不再以序列相似度进行聚类，而是通过“去重复” (Dereplication, 相当于以 100%相似度聚类) 等步骤，进而获得单碱基精度的代表序列，大大提升了数据精确度与物种分辨率。DADA2 的核心是去噪，然后使用 ASVs (Amplicon Sequence Variants)的概念构建类 OTU (Operational Taxonomic Units) [3]表，获得最终的 feature 特征表以及特征序列，进一步进行多样性分析、物种分类注释和差异分析等。

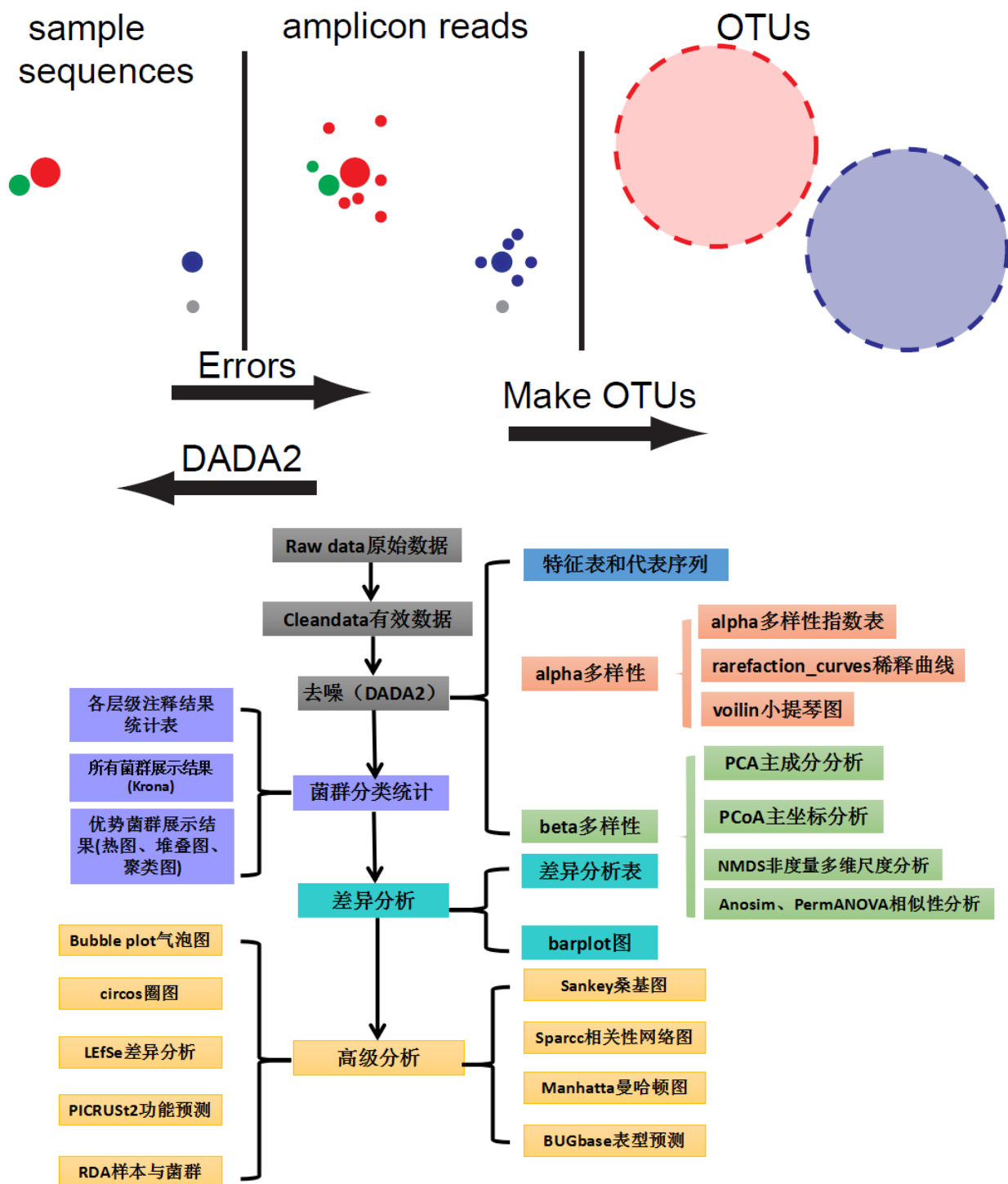


图 4 16S rDNA 信息分析流程

三、标准分析

在多样性论文中，物种分类结果会是讨论的重点。根据不同的实验设计和目的，我们总结出如下思路：

- 1.测序做为高效的筛选手段，筛选出与疾病（医口）或农艺性状（农口）相关的菌群。不仅可以快速筛选到已知的一些菌群，还可能找到新的未被报道的但有研究价值的新菌群；
- 2.从结果中筛选一个或几个值得重点关注的物种，重点关注门和属丰度上的变化，由此找出可以作为疾病biomarker 的菌群；
- 3.结合不同层级物种的统计学差异的菌群以及功能预测结果，进一步使用宏基因组测序来探讨菌群的基因功能和通路；
- 4.菌群代谢物是影响宿主的重要途径之一，结合菌群代谢组数据，进行多组学关联研究，有助于深入探讨菌群与宿主的互作机制。

以下是两个简单的思路，供您在整理数据和讨论结果时参考。如您是其他的实验设计，欢迎与售后支持人员交流结果讨论思路。

- 1.研究某一群体（如健康人肠道菌群，或高产植物的根际土壤）的微生物构成，找出这类群体的微生物共性，从而找到微生物对群体的影响（需要较多的样品数量）；
- 2.研究不同处理样品（比如疾病与健康，治疗组与安慰剂组，处理组与对照组）之间的微生物构成差异等。通过功能预测来推测该群体的菌群可能产生的功能，结合临床指标以及样本分组信息，将表型与微生物的组成联系起来。

3.1 有效数据统计

对原始下机数据进行双端拼接、质量控制、嵌合体过滤后，进行高质量数据统计，结果如下表：

表 1 有效数据统计表：

Sampl e	Raw_Ta gs	Raw_Bas es	Valid_Ta gs	Valid_Bas es	Valid %	Q20 %	Q30 %	GC %
feces_ 1	87124	43.56M	75925	31.93M	87.15	97.27	92.37	54.3 3
feces_ 2	86726	43.36M	78543	32.97M	90.56	97.39	92.61	54.7 9
feces_ 3	80186	40.09M	70228	29.39M	87.58	97.37	92.56	54.8 4
soil_1	87314	43.66M	81978	34.13M	93.89	97.13	91.68	55.9 6
soil_2	84437	42.22M	67883	28.40M	80.39	96.74	91.01	56.7 5

soil_3	80853	40.43M	62694	26.25M	77.54	97.76	93.24	55.21
water_1	72371	36.19M	58143	23.90M	80.34	97.35	92.19	51.13
water_2	86400	43.20M	75337	30.93M	87.20	96.91	91.24	50.78
water_3	84489	42.24M	71706	29.44M	84.87	97.17	91.67	50.89

结果路径: summary/2_clean_data/Demo_clean_data_stat.xlsx

各列意义如下:

Term	意义
Sample	测序样品名称
Raw_Reads	测序原始数据，以四行为一个单位，统计每个文件的测序序列个数
Raw_Bases	测序序列的个数乘以测序序列的长度，并以单位 M 表示
Valid_Tags	预处理后，以四行为一个单位，统计每个文件的拼接后测序序列个数
Valid_Bases	预处理后，测序序列的个数乘以测序序列的长度，并以单位 M 表示
Valid%	有效数据 (Valid) 与原始数据 (Raw) 比值，百分比表示
Q20%	有效数据中数据质量 $\geq$ Q20 的数据比例
Q30%	有效数据中数据质量 $\geq$ Q30 的数据比例
GC%	有效数据中数据 GC 含量

3.2 feature 特征表

QIIME2[2]相对于 QIIME，最大的改变是 DADA2 不再以序列相似程度进行聚类，而是通过过滤、去重 (Dereplication)、嵌合体过滤等方式修正扩增子的测序错误，确定更多的真实变异。扩增子测序本身就具有内在的限制，但是聚类 OTU 的方式进一步限制了它的发展。OTU 不是物种，它们不应该成为错误的一部分。经过去除背景噪音获得的特征表和特征序列，不仅极大提升了数据精确度与物种分辨率，还保证了结果可信程度。

表 2 各样品 feature 特征序列统计结果

feature ID	feces_ 1	feces_ 2	feces_ 3	soil_ 1	soil_ 2	soil_ 3	water_ 1	water_ 2	water_ 3
3ed7861da51934a4d1d1744ae76a2184	0	0	0	0	0	0	5058	7354	7520
4b38473099dc5f18b73be535e282921a	6564	379	4902	4	7	4	11	3	0
dde7915c2b470e6ab3e4787a2ac6b02a	0	0	0	2	0	4	2480	4304	3998
3bc875054ff1ca03c4b2419490818f87	0	0	0	0	0	0	2703	3360	3332
29d41f7724da8f39aa8dcdf8104add95	2450	2965	824	0	0	0	0	0	0
a5eef8793ddf5fb4b141fbd69429e7c5	2271	2882	1050	0	0	0	0	0	0
eeded886b5c875624baa05062416ebb	2011	2579	1366	0	0	0	0	0	0
8d4c73fc9277a090add2539a3c8a904c	996	1803	2513	0	0	0	3	0	0
68d91efd23bd572a2596634c5e1a2a90	1792	3308	173	0	0	0	0	0	0
8c031f2fdca43c7f6b666fc988445cd9	0	0	0	0	0	0	1398	1663	1621
b1c049aeefb630fe2937b155523c6e97	2860	1321	318	0	0	0	0	0	0
e60a1b18a29616f8d18b108040983e13	587	1215	2105	0	0	0	0	0	0
8fd82a31143f2ea7ef77a11eb026fd3f	0	0	0	0	0	0	1144	1427	1311
41f45c0ed64d083d2cc3bd5750efe9a2	1777	1427	582	0	0	0	0	0	0
cc8323d12abb9681d32a25b77a3b3cad	0	0	0	0	0	0	1015	1385	1297
5556388c13ce3c501de4589d7779702c	975	1210	694	0	0	0	0	0	0
b7563644fdabeddb89b31cac9c326f9f7	860	727	1054	0	0	0	0	0	0
2fddc807ed63b947ab8e68046a34a66e	0	0	0	554	427	1645	0	0	0
c28dac20d59b30aa554e98ddb7bb563	0	0	0	0	0	0	678	900	873
1									
0efcbcd171f35162777500721a618080c	838	1110	449	0	0	0	0	0	0
f8cbbeb5d541e3b903b0c0adf44774d4e	0	0	0	0	0	0	620	954	802
5e2f801c6bdd4ee934e6ee7234729007	1268	59	1020	0	0	0	0	0	0
6f1c70b6d9e037711c028629ca233185	1072	918	352	0	0	0	0	0	0
9f8e4954244a27164352f22367601959	0	0	0	0	0	0	698	819	818
9a5f1c980c1b870738c3a0616f8a7db5	467	834	907	0	0	0	0	0	0
310b4272773237dd3675b9790703c43	0	0	0	0	0	0	393	994	795
3									
6ca1326d225d72bba74fb6f0087ea590	535	836	691	0	0	0	0	0	0
8ca69dfd6a88cffad1d15d675d460149	0	0	0	0	0	0	559	727	742
46189723f5bf9a5f3c4371645544bd7e	742	936	339	0	0	0	0	0	0
6f602a19d82c8cd75af44fd809e32065	343	765	893	0	0	0	0	0	0
af70d3a4d7eb9610c2c307b40c8f9db2	222	89	1430	0	0	0	0	0	0
a13d09db91b720773b7cb6f13d3c81ab	0	0	0	0	0	0	474	635	544
aba8812f8d81ca8e10a2279eba3ccf4d	562	537	355	0	0	0	0	0	0
bf23854f0fabe70d01993762ae47da8b	0	0	0	0	0	0	372	543	529
3105a7cca2e029c29f652342b678313d	0	0	0	0	0	0	354	541	539
66c4212afbea18ed6c91448beba1e994	0	0	0	0	0	0	387	502	509
60dfd10c78be620d9e829923a0e44e91	0	0	0	0	0	0	380	529	489
e37d0259a2948afb6587241e964e6f78	0	0	0	0	0	0	379	456	470
8de013940c0887e172a8b68fb364ce63	385	459	451	0	0	0	0	0	0
3c82a03e8d39950d08d5f6b097b093cc	0	0	1295	0	0	0	0	0	0
9cc3f7f95ddfcf67153bf6969602a86b	0	0	0	0	0	0	365	482	441
32fb2369f2d7ae7e71342a4c8056d2d5	0	0	0	0	0	0	341	463	482
3ca53903eabce6ddb11f1a8b8e99c40b	258	312	640	0	0	0	0	0	0
26272d708f014f96d0512781f9413b35	0	0	0	0	0	0	331	449	408
f8234599ec7499a4553e8e715d8545d1	0	0	0	0	0	0	400	385	391

16399452af664116fe40843f7a70c85b	0	0	0	0	0	0	348	443	361
dc22e0f49c71e85bc6fb960fe883e0c7	0	0	0	0	0	0	248	483	415
830780e721a40691388d1ee15ceb7e52	0	0	0	0	0	0	325	410	404
7bf24eb4ec59c2f0b37fd3767168d147	420	717	0	0	0	0	0	0	0
ee595d4475e4eb0c50864be4dc94ce07	0	0	0	0	0	0	294	437	382
5f9bfe4d1b66c123b5d73aa327165c8f	718	300	56	0	0	0	0	0	0
e944b393b81184b288837f7a2e607bb1	0	0	0	0	0	0	297	390	376
18a82ec19b3331139173d5bf1d92331d	0	0	0	0	0	0	295	393	366
b4e3207d0c753480069badae058f0a67	0	0	0	0	0	16	215	364	400
52e5699c94de102da70ae27e4ce69996	517	385	69	0	0	0	0	0	0
4c7be0eabea9acb536d8e31b67c18e47	0	0	0	964	0	0	0	0	0
9ec1572ee1c2db0cba4f9bb058965331	0	0	0	0	0	0	232	376	316
be39e2de2e8aa8c43dd515a0776124a3	269	199	425	0	0	0	0	0	0
fe4e2deae37b20114c6d833dbff2f853	0	0	0	0	0	0	248	331	309
315bf6a49451981bd05eaf638a25ca7e	576	305	0	0	0	0	0	0	0
f3420aee4b202d1c6ca84c0fb4e6374a	0	0	0	0	0	0	252	324	293
2d8382265a1dcf43e7c06f540a92a0bc	286	410	172	0	0	0	0	0	0
aaccb74c20d80b419ea8b9cc5500284e	0	0	0	0	0	0	235	339	289
3acccb8e948cf60ad3474cd1c6e543d	0	0	0	0	0	0	249	297	312
f64cafb46d13d8e5825677d185fa1d7e	413	306	125	0	0	0	0	0	0
cc34f964109779dab9888d26986d1ae9	0	0	0	0	0	0	193	372	268
6958da4c979b96fb9628451ae32ce775	249	312	258	0	0	0	0	0	0
160c6c0d21f3ef288f0003faee24788b	0	0	0	0	0	0	212	313	293
9991030520fcf8ad16fda244f5d0c039	127	109	580	0	0	0	0	0	0
d1936a86defac30e61a0595fab1b5db3	22	172	607	0	0	0	0	0	0
e4fe27ac6bc30f5f0eeafb355861817a	306	341	148	0	0	0	0	0	0
ac6498acef3939cdf03a543d84834ca	0	0	0	20	24	726	0	0	0
b24dd27fac72fc74c728c27a6d408d98	0	0	0	0	0	0	207	285	276
7ba440b3aadf4ee1889a8ec8cfaa3f11	123	157	471	0	0	0	0	0	0
0897f5701faf86dea626fb525535add0	0	0	0	0	0	0	213	302	234
f268dd98cd2227df4f1d12b92d0a0f39	0	0	0	0	0	0	193	304	251
1ff70fbddf3ccd7c16a5fe7124baa271	374	307	63	0	0	0	0	0	0
eaad5e4d6325fd747fd5b0f44f6c4fce	251	355	131	0	0	0	0	0	0
8714640e9a5bd16f33cd985711805a34	158	311	240	0	0	0	0	0	0
924d57158f0c6892d37c545c6b609258	0	0	0	0	0	0	194	272	232
cbb49c2b90607a7f1161913c34d719de	321	375	0	0	0	0	0	0	0
b254f77ebd40f591ba241aa2f1bfd68a	435	257	0	0	0	0	0	0	0
5412be598dec00d512d42afa7bb9df4c	211	316	140	0	0	0	0	0	0
9ad864a1e8e5cfb2e0a1dbb80ffbc4b8	0	4	658	0	0	0	0	0	0
d5540963a6dcf53eaf027bc65b68caad	195	319	146	0	0	0	0	0	0
df8ca2692141028441f58603fb15dcd5	199	336	124	0	0	0	0	0	0
a5937cf062b47e56932e37f277e46454	342	312	0	0	0	0	0	0	0
66e6438009f2599c59265100ebae9779	0	0	0	0	0	0	154	241	245
3ad08ba70fdfe94d378145e6c738662d	0	0	0	242	269	126	0	0	0
408cda804ac9932e6fef80d5a5734756	0	0	0	0	0	0	197	207	210
8de2669d020df753ac5ff28d3a8faf98	261	255	92	0	0	0	0	0	0
becd24ecd39467d799512d817e4e6733	0	0	0	18	0	586	0	0	0
8c160916bba9a860adec0e3b1185330e	195	404	0	0	0	0	0	0	0
9547178ab221d3d96bd0f30b8b482c20	249	192	155	0	0	0	0	0	0
8e77f795bc7a02aa80804ccc32ecf48b	91	198	303	0	0	0	0	0	0

8f17ca772f951a71f996cc1efe210c23	263	176	132	0	0	0	0	0	0
57444bc83e1015847af7f763b0c75292	149	299	116	0	0	0	0	0	0
6e7078fe87c9fd8aa5bb04a2ded8887d	117	159	281	0	0	0	0	0	0
1beb251bd26ef27682e55493ada004e7	0	0	0	0	0	0	188	209	159

结果说明：其中 feature ID 表示 QIIME2 降噪后特征序列的 ID，它们是无序且无意义的字符串。其中的数值表示每个样本对应的该特征序列的 feature 数量。由于不同样本的原始测序量并非完全一致，因此不是简单来看数值的大小来比较丰度的高低，而是需要通过相对丰度的计算来比较不同特征值之间的差异。

结果路径：summary/3_feature_profiling/1_feature_tables

3.3 feature 分布 Venn 图

根据获得的特征值丰度表，计算各样品/分组共有 feature 的数量，并通过 Venn 图直观地呈现各样品/分组共有和特有的 feature 数目。我们对每个 feature 在各个分组中的表达丰度进行统计；若三个组中 feature 表达均为非零，则该 feature 为三个组共有的。同理，若存在 feature 只存在于一个样品/分组，则该 feature 为该样品/分组特有。

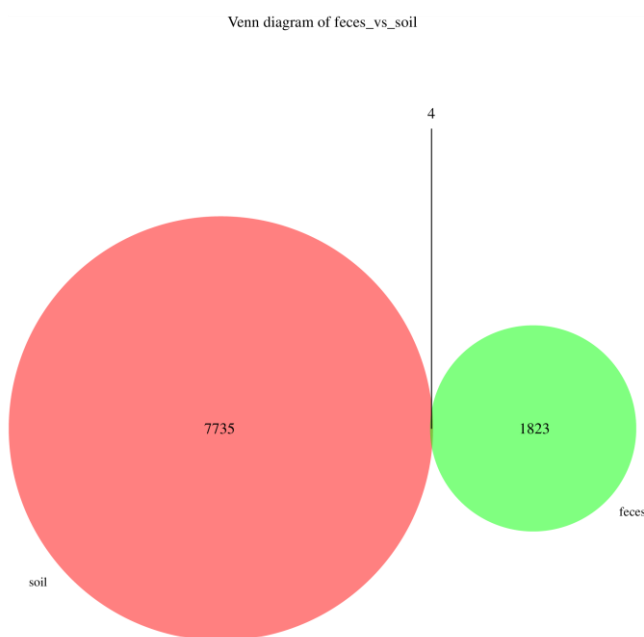


图 5 [Venn 图](#)

结果说明：我们对于样本/分组数 ≤ 5 的情况绘制 Venn 图。随着样本/分组数的增加，Venn 图的复杂度也急剧增大，当样本/分组数=5 时，考虑到 5 组样本间所有可能的共有和独有关系，这些关系可能在 Venn 图上形成超过 30 块分区，进一步增加样品/分组将使得图片非常拥挤，复杂度急剧增加，而可读性大大降低。如果样品/分组 ≥ 6 ，我们以花瓣图形式进行展示。它能反映所有样本/分组的共有和特有的特征

值数目，但不能反应两两分组间共有的特征值数目。图中每个圈代表一个样本/分组，圈和圈重叠部数目个数。

结果路径：summary/3_feature_profiling/3_venn_diagrams

3. 4 Alpha 多样性分析

Alpha 多样性是指一个特定环境或生态系统内的多样性，主要用来反映物种丰富度和均匀度以及测序深度。Alpha 多样性主要通过 Chao1、Observed species、Goods_coverage、shannon、Simpson 等指数来反映丰富度和均匀性。Chao1,Observed_species 指数主要反映样本的物种丰富度信息；Goods_coverage 反映样本的低丰度 feature 覆盖情况；Simpson/Shannon 主要综合体现物种的丰富度 (Richness) 和均匀度(Eveness)。其中Chao1和observed_species主要是估计群落中包含物种的数目；Goods_coverage 是指微生物覆盖率，其数值越高，则样本中新物种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；Shannon 指数来源于信息熵，Shannon 指数越大，表示不确定性大。不确定性越大，表示这个群落中未知的因素越多，也就是多样性高；Simpson 数值范围在 0-1 之间，当群里只有一种物种的时候，此时 Simpson 值最小为 0，同时也是我们直观理解的多样性最小。当物种种类无限多（丰富度最高），并且每个物种数目都一致（均匀度最高）的时候，Simpson 值为最大为 1。

稀释曲线 (rarefaction curve) 通过模拟重新取样的过程，观察其中物种变化的趋势，估计环境中的物种丰富程度。通过绘制稀释曲线，统计 feature 的丰富程度，对比不同样品的稀释曲线就可以直观显示样品间物种多样性的差异。稀释曲线可直接反映测序数据量的合理性，并间接反映样品中物种的丰富程度，当曲线趋向平坦时，说明测序数据量渐进合理，更多的数据量只会产生少量新的物种/特征值。

chaol.rare.png

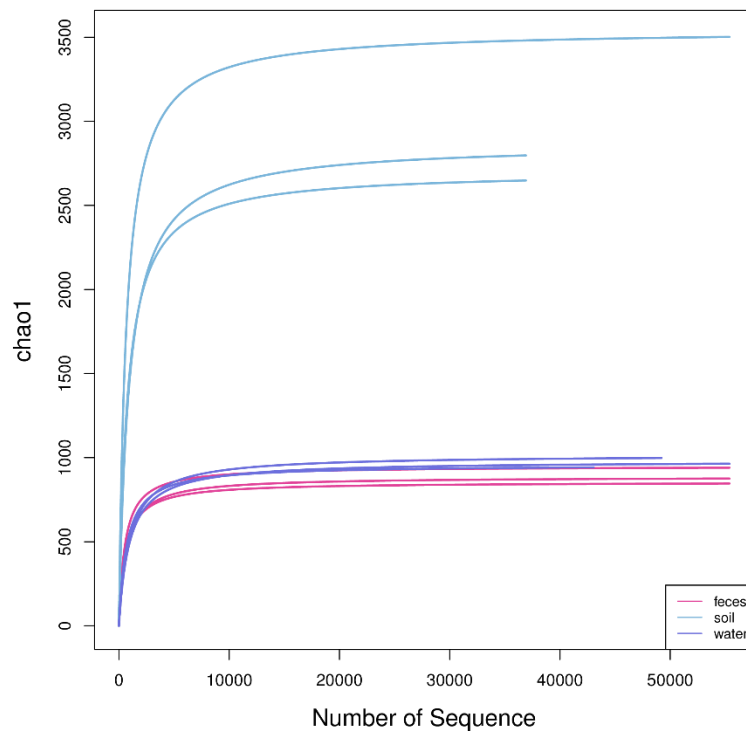


图 6 稀释曲线结果示意图

结果说明：横坐标代表随机抽取的序列数量；纵坐标代表观测到的 feature 数量。通过该图，可以比较在相同的测序深度下，不同样本中 feature 数目的多少，从而在一定程度上衡量每个样本的丰富度和多样性高低。从稀释曲线中我们也可以估算测序数据量是否足够，是一个参考性指标。若曲线急剧上升，表明测序数据量并未达到最佳；若曲线趋于平缓，则证明测序数据量已饱和。

如果您的结果中包含生物学重复，且组内重复数目 ≥ 5 ，则我们在结果中会提供小提琴图，其中左上角的 p 值是图中所有分组使用秩和检验所得 p 值，其中两两分组差异显著性会以*表示差异显著，**表示差异极显著，ns 表示表示无显著差异，对应的 p 值详见：[summary/4_alpha_diversity/4_violin_plot/alpha_diversity_diff.xlsx](#)。

结果路径：[summary/4_alpha_diversity](#)

3.5 Beta 多样性分析

Beta 多样性是指不同环境群落之间的物种差异性。Beta 多样性与 alpha 多样性一起构成了总体多样性或一定环境群落的生物异质性。Beta 多样性分析通常由计算环境样品间的距离矩阵开始，该矩阵包含任意两个样品间的距离，主要通过主成分分析（Principal component analysis, PCA）、主坐标分析（Principal coordinates analysis, PCoA）、聚类分析（Clustering analysis, UPGMA）、非多维尺度分析（Multidimensional scaling, NMDS）、相似性分析（Analysis of similarities, ANOSIM）、多元方差分析（PermANOVA, 又称 Adonis）等方法，观察样本之间的差异。其中标准分析中 PCoA 以及 NMDS 中 p 值是 ANOSIM 差异分析所得 p 值，如后期确认 PermANOVA 分析 p 值结果更适合您的结果，后期进行图片编辑即可。

3.5.1 PCA 分析:

主成分分析 (PCA, Principal Component Analysis), 简单来说就是在数据信息丢失最少的原则下, 对多特征的数据进行最佳综合简化, 即是对高维特征空间进行降维处理。它通过寻找矩阵的线性无关向量实现数据降维, 线性无关的向量就称为主成分, 第一主成分解释了数据最大部分的方差或波动性。PCA 能够提取出最大程度反映样品间差异的两个坐标轴, 从而将多维数据的差异反映在二维坐标图上, 进而揭示复杂数据背景下的简单规律。基于 Feature 丰度表, 我们使用 R 软件 vegan 包进行了 PCA 分析, 如果样品的物种组成越相似, 则它们在 PCA 图中的距离越接近。我们可以通过 PCA 分析直观看出生物学重复的组内相似性及组间差异是否明显。

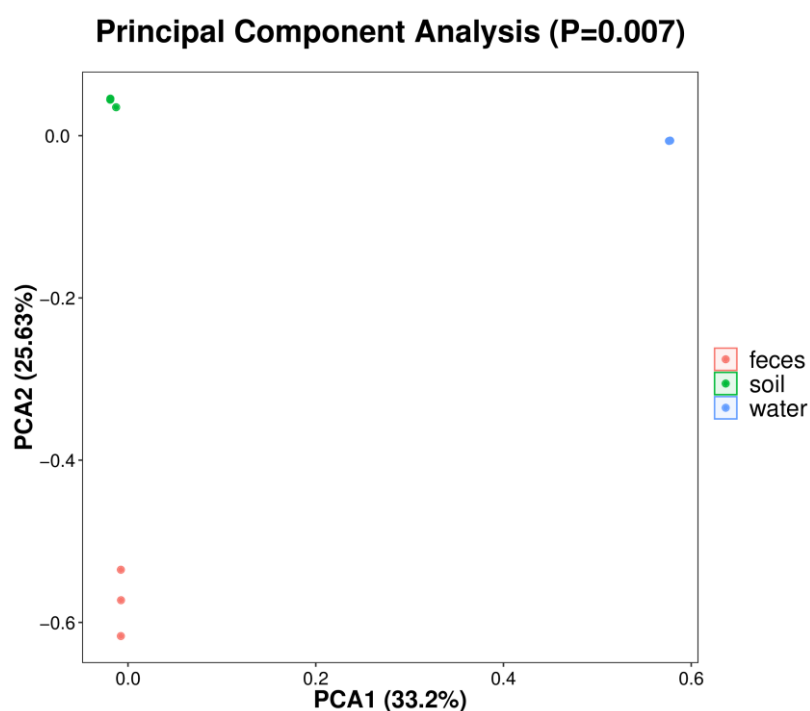


图 7 样品 PCA 分析展示

结果说明：不同颜色代表不同分组，同一个分组按照 95%置信区间以圈图(每组生物学重复数目 ≥ 5)形式展示。两个样品间距离越近，则说明样品之间的微生物组成结构越相似，差异性越小；反之，则说明样品之间的微生物组成结构差异越大，差异性越大。通过该分析可观察个体或群体间的差异。当生物学重复样本距离更近，不同分組组间距离更加远，说明处理有效或者说明差异显著，也预示着后期寻找差异菌群结果更趋近于真实情况(背景噪音小)；当所有样本均难以分开，说明该处理对菌群的分布影响甚微；当生物学重复样本距离大于不同分組样本之间的距离，则需要做相应排查，确认取样、寄送或后期处理中样本是否弄混。

结果路径：summary/5_beta_diversity/2_PCA_result/PCA.png

3.5.2 PCoA 分析：

基于上述计算获得的两种不同距离进行主坐标分析，通过 PCoA 分析可以观察单个样品或样品组之间的差异，不同样本间的距离代表其物种组成的差异情况，并判断是否与项目实验分組情况一致。

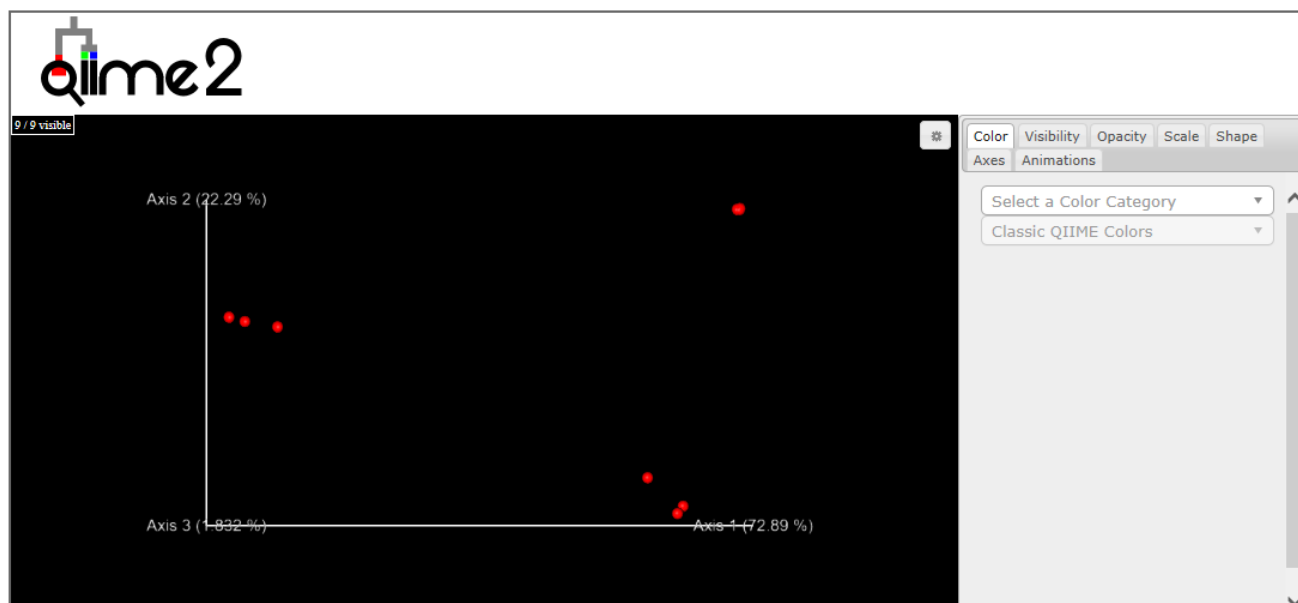


图 8-1 [PCoA 分析 3D 示意图](#)

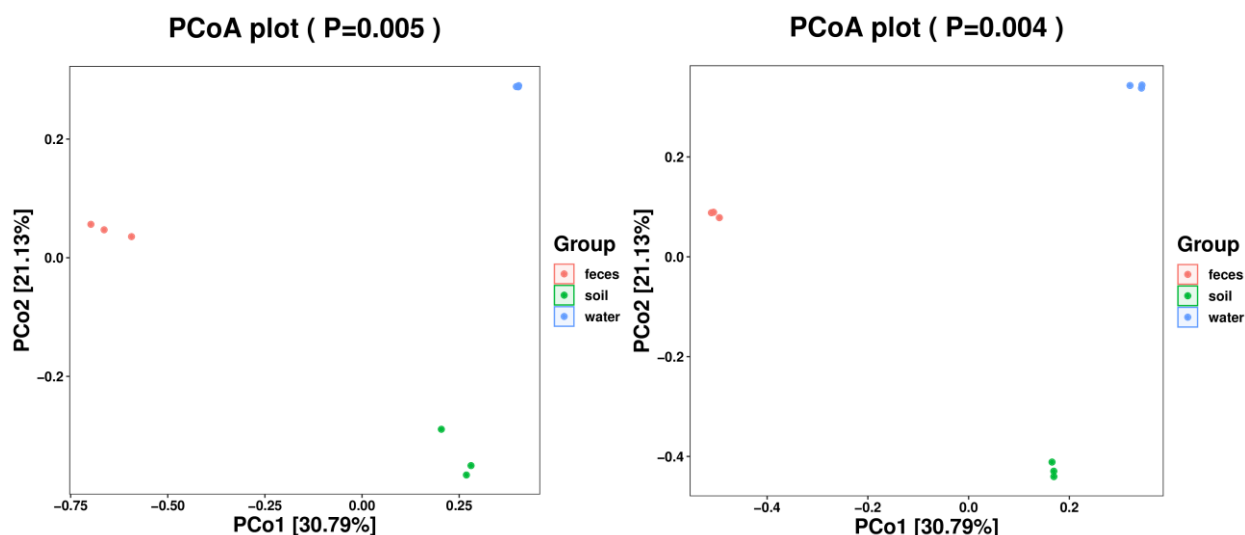


图 8-2 PCoA 分析 2D 示意图

(左图为 Weighted_UniFrac 结果，右图为 Unweighted_UniFrac 结果)

结果说明：和 PCA 一样，结果中不同颜色代表不同分组，样品距离越近，说明样品之间的微生物组成结构越相似，差异性越小。微生物结构组成相似的样品聚在一起，通过该分析可观察个体或群体间的差异。当样本数量庞大时，出现同一组内样本离散在群体之外，说明这个样本可能组间相似性不高，作为生物学重复可能会影响整个分组的结果。两者结果都能说明通过降维，选择主要的前几位特征值来进行组别的区分。和 PCA 不同的是，PCoA 使用的文件是距离矩阵。

结果路径：summary/5_beta_diversity/2_PCoA_result

3.5.3 样品聚类

为了适应不同的环境样品类型，基于各样品的 feature 分析结果，我们将使用最常用的 unweighted unifrac 和 weighted unifrac[4,5]两个指标来衡量两个样品间的相异系数，其值越小，表示这两个样品在物种多样性方面存在的差异越小。同时计算得到的样品距离矩阵，并利用 UPGMA (Unweighted Pair Group Method with Arithmetic Mean) 方法对样品进行聚类，生成聚类树数据文件，该文件可以导入聚类树查看软件如 Figtree。

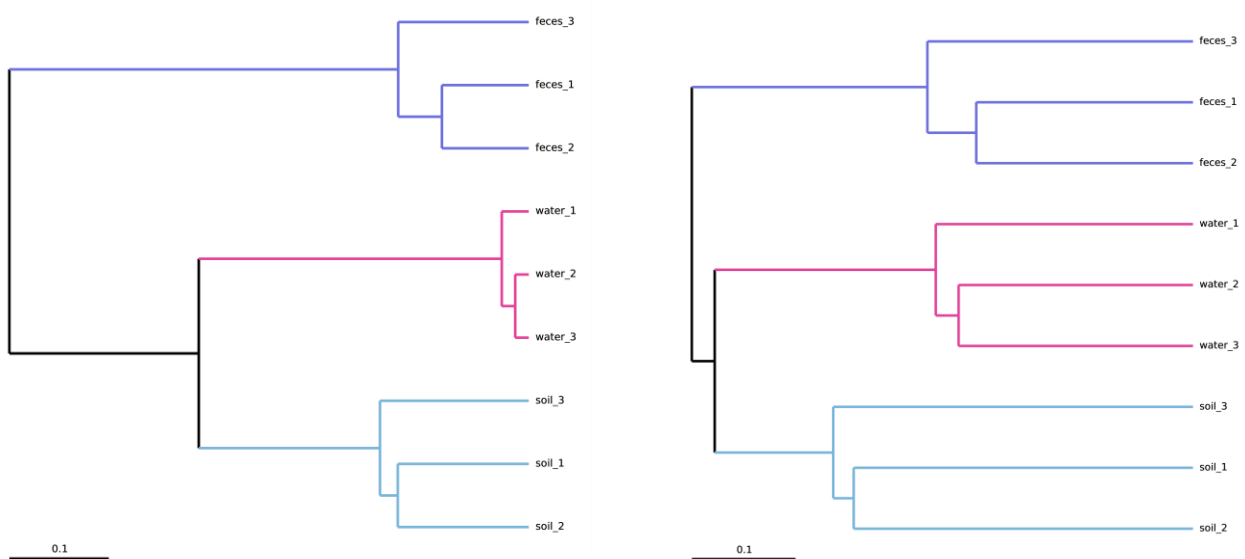


图 9 UPGMA 层次聚类分析

(左图为 Weighted_UniFrac 结果，右图为 Unweighted_UniFrac 结果)

结果说明：结果图中树枝不同颜色代表不同的分组。聚类树展现了样本间的相似度，样本间的分枝长度越短，两样本越相似。同时这样的可视化结果可以直观显示不同的环境样本中微生物进化的差异程度。本结果在论文中展示的情况根据结果是否与实验设计一致，聚类图结果清晰、直观，可有效展示良好的分组及样本重复性。当出现同一组样本离散在群体之外，说明这个样本可能组间相似性不高，作为生物学重复可能会影响整个组，因此，建议可以结合 PCA 图与样本柱状图结果，对该离群样本进行分析，判断是否需要删除该样本。

结果路径：summary/5_beta_diversity/3_upgma_cluster

3.5.4 NMDS 分析图

NMDS(Nonmetric Multidimensional Scaling)非度量多维尺度分析，与 PCoA 分析一样，也是一种基于样品（任何类型距离）距离矩阵的排序方法。与 PCoA 不同的是，NMDS 不再基于距离矩阵数值，而是根据距离排位顺序进行降维计算。NMDS 排序依赖于相异系数的大小顺序，并不依赖于准确的相异性系数值。如果排序的目的不是在于最大程度保留样品之间的实际距离，只是反映样品之间的顺序关系，这种情况下，非度量多维尺度分析是一种不错的解决方案。NMDS 分析目标是保持样品排序关系不变，因此，如果两个样品距离较近，说明这两个样品物种组成越相似。

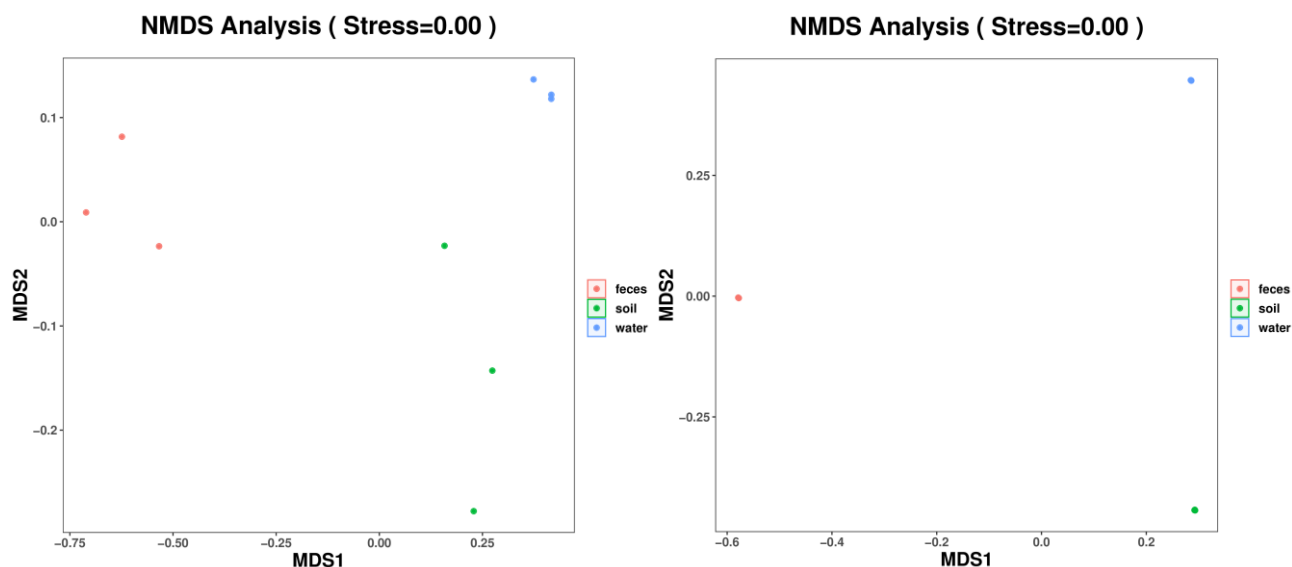


图 10 NMDS 分析图

(左图为 Weighted_UniFrac 结果, 右图为 Unweighted_UniFrac 结果)

结果说明：图中的点表示样品，不同颜色样品属于不同的分组，点与点之间的距离表示样品之间差异程度。检验NMDS分析结果的优劣用胁强系数（stress）来衡量，通常认为当 $\text{stress} < 0.2$ 时，可用NMDS的二维点图表示，其图形有一定的解释意义；当 $\text{stress} < 0.1$ 时，可以认为是一个好的排序；当 $\text{stress} < 0.05$ 时，具有很好的代表性。

结果路径：summary/5_beta_diversity/5_NMDS_result

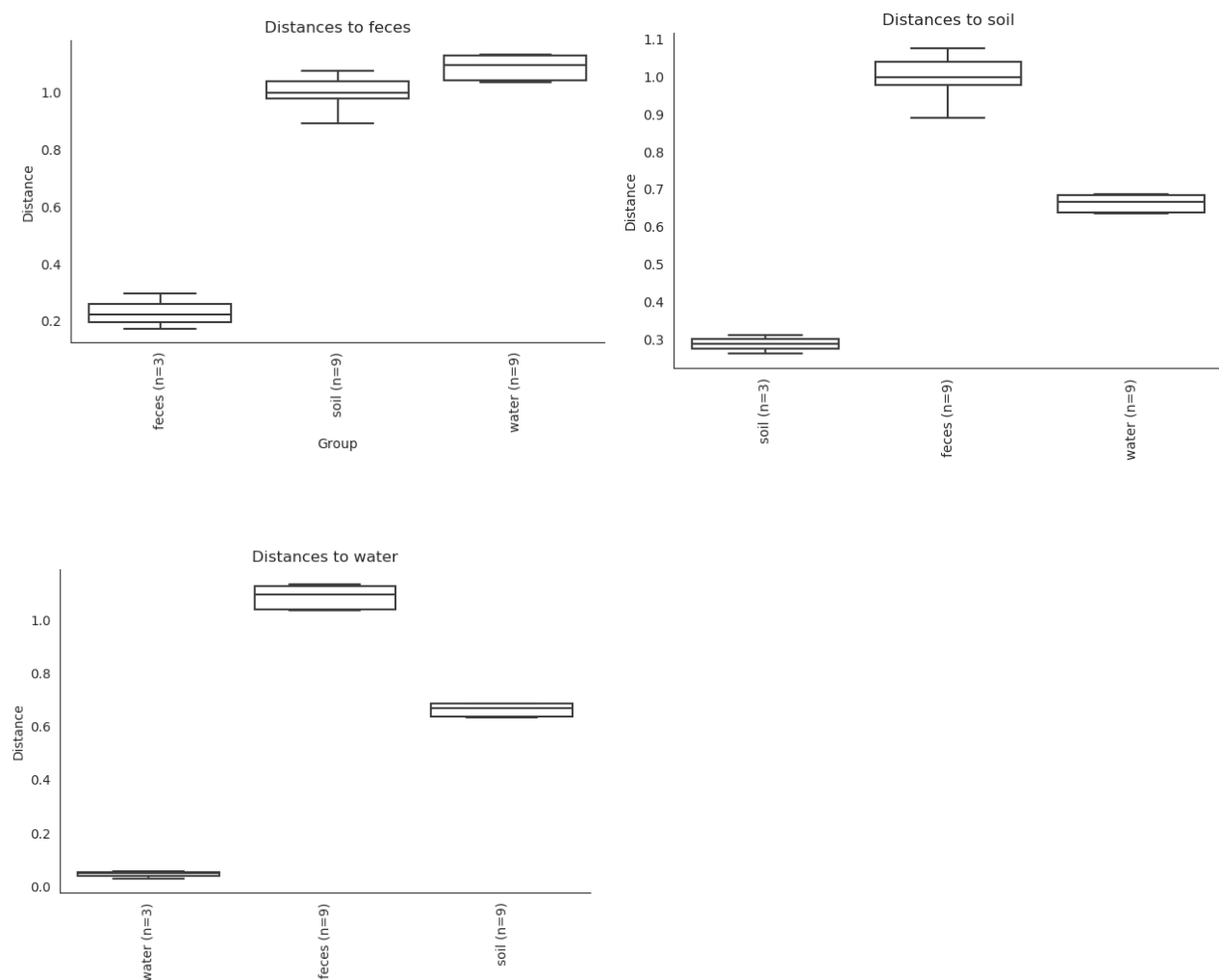
3.5.5 相似性分析

Anosim 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。首先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，并计算 R 和 P 值。Adonis 又称置换多元方差分析或非参数多元方差分析。它是通过距离矩阵，分析不同分组因素对样品差异的解释度，并使用置换检验对其统计学意义进行显著性分析。

	ANOSIM results
method name	ANOSIM
test statistic name	R
sample size	9
number of groups	3
test statistic	1
p-value	0.005
number of permutations	999

Group significance plots

[Download raw data as TSV](#)



结果说明：Statistic 即为 R 值，范围为-1 到+1。R 值越接近 1 表示组间差异越大，R 值越小则表示组间和组内没有明显差异，并且当 P 值小于 0.05 时说明检验的可信度高。

结果路径：summary/5_beta_diversity/5_anosim_result

表 3 adonis 多元分析

Df	SumOfSqs	R2	F	Pr(>F)
----	----------	----	---	--------

Group	2	2.54	0.95	53.92	0.01
Residual	6	0.14	0.05		
Total	8	2.68	1		

结果说明：Df:自由度；SumsOfSqs:总方差，又称离差平方和；Mean Sqs:表示均方（差），即 SumsOfSqs/Df；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；Pr:P 值，小于 0.05 说明本次检验的可信度高。

结果路径：summary/5_beta_diversity/6_adonis_result

3.6 物种分析

feature 特征序列是我们物种分类的原始文件，为了更加准确的分析物种组成，我们使用 SILVA (Release 132, <https://www.arb-silva.de/documentation/release-132/>) 以及 NT-16S 数据库做物种分类及后续分析，以保证注释结果完整准确。根据 feature 注释结果和各样品 feature 特征表，获得界、门、纲、目、科、属、种水平物种丰度表，并针对不同水平物种丰度表进行不同样品（组）物种组成及差异分析，参数阈值：置信度>0.7。

物种注释完成后，会统计产生 feature 对应的注释表格，对应各个物种信息。一般在论文中我们会对这部分结果使用大量的讨论，根据不同的实验设计和目的，从中筛选一个或几个重点关注的物种（可以从七个水平上分别查看，其中建议老师可以重点关注门和属物种的变化，这两个水平的研究是最多的），结合 feature 丰度信息表分析物种表达水平、聚类分析等。当样品量较大时（比如超过 100 个样品），导致样品柱状图十分拥挤，降低了图形的美观和可读性，建议按分组（分组物种丰度为组内样品平均相对丰度）图形进行展示。

表 4 feature 表达量注释表结果

feature ID	Domain	Phylum	Class	Order	Family	Genus	Species
3ed7861da51934a4d1d1744ae76a2184	d_Bacteria	p_Actinobacteria	c_Acidimicrobiia	o_Actinomarinales	f_Actinomarinaceae	g_Candidatus_Actinomarina	s_Candidatus_Actinomarina_unclassified
4b38473099dc5f18b73be535e282921a	d_Bacteria	p_Firmicutes	c_Bacilli	o_Lactobacillales	f_Lactobacillaceae	g_Lactobacillus	s_Lactobacillus_unclassified
dde7915c2b470e6ab3e4787a2ac6b02a	d_Bacteria	p_Proteobacteria	c_Alphaproteobacteria	o_SAR11_clade	f_Clade_I	g_Clade_Ia	s_Clade_Ia_unclassified
3bc875054ff1ca03c4b2419490818f87	d_Bacteria	p_Proteobacteria	c_Alphaproteobacteria	o_Rhodobacterales	f_Rhodobacteraceae	g_Planktomarina	s_Planktomarina_unclassified
29d41f7724da8f39aa8dcdf8104add95	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Muribaculaceae	g_Muribaculaceae_unclassified	s_Muribaculaceae_unclassified
a5eef8793ddf5fb4b141fbd69429e7c5	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Muribaculaceae	g_Muribaculaceae_unclassified	s_Muribaculaceae_unclassified
eededa886b5c875624baa05062416ebb	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Muribaculaceae	g_Muribaculaceae_unclassified	s_Muribaculaceae_unclassified
8d4c73fc9277a090add2539a3c8a904c	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Muribaculaceae	g_Muribaculaceae_unclassified	s_Muribaculaceae_unclassified
68d91efd23bd572a2596634c5e1a2a90	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Bacteroidaceae	g_Bacteroides	s_Bacteroides_unclassified
8c031f2fdca43c7f6b666fc988445cd9	d_Bacteria	p_Proteobacteria	c_Gammaproteobacteria	o_Oceanospirillales	f_Nitrincolaceae	g_Nitrincolaceae_unclassified	s_Nitrincolaceae_unclassified
b1c049aeefb630fe2937b155523c6e97	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Prevotellaceae	g_Prevotellaceae_UCG-001	s_Prevotellaceae_UCG-001_unclassified
e60a1b18a29616f8d18b108040983e13	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Muribaculaceae	g_Muribaculaceae_unclassified	s_Muribaculaceae_unclassified
8fd82a31143f2ea7ef77a11eb026fd3f	d_unclassified	p_unclassified	c_unclassified	o_unclassified	f_unclassified	g_unclassified	s_unclassified
41f45c0ed64d083d2cc3bd5750efe9a2	d_Bacteria	p_Bacteroidetes	c_Bacteroidia	o_Bacteroidales	f_Prevotellaceae	g_Alloprevotella	s_Alloprevotella_unclassified
cc8323d12abb9681d32a25b77a3b3cad	d_Bacteria	p_Proteobacteria	c_Alphaproteobacteria	o_Rhodobacterales	f_Rhodobacteraceae		

结果说明：该表代表每个特征值对应的界、门、纲、目、科、属、种不同层级的物种注释结果。（此处展示的为部分结果内容）

结果路径：summary/6_taxonomy_community/feature_taxonomy_split.xlsx

3.6.1 柱状堆叠图(stacked bar chart)

根据物种丰度表和物种注释表，我们默认选取丰度最高的 30 个物种分类，将每个样本/分组的相对丰度进行不同形式的展示。该柱状图以堆叠柱状图形式展现,便于更直观地进行样品丰度的比较。在各个层级中，我们可以直观的看到优势菌种的表达情况，及在各个不同处理中的变化趋势。当然，若是有关关注稀有菌群的表达情况时，我们可以展现所有物种分类。一般我们建议老师重点关注门、属水平的分类，代表性强且分类层级清晰便于比较，易于论文发表。

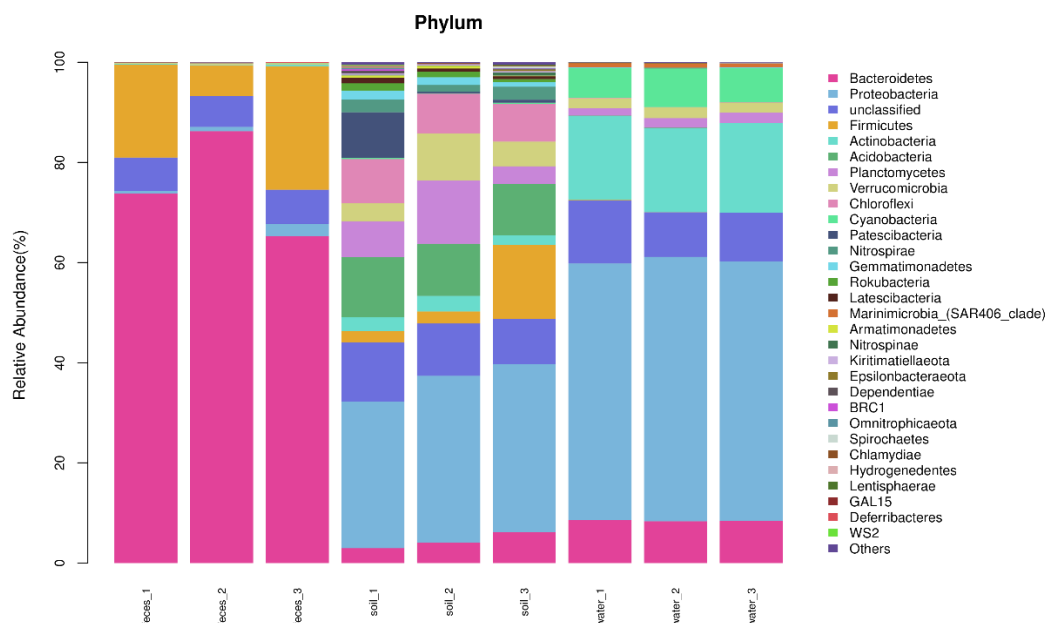


图 10 样品丰度柱状图

结果说明：图中横轴为样品名称，纵轴代表某分类的相对丰度；不同颜色对应同一层次不同物种，通过柱状图可以明了每个样本高表达物种组成，同时也可以观察组内物种组成及表达和组间物种表达趋势。根据样品物种丰度表，默认选取丰度最高的 30 个物种分类或功能分类（随着物种分类水平越低，鉴定的物种越精确，同时物种数目也随之增加，以属水平为例，可能会注释到几百个属。且有很大一部分属丰度很低，做物种组成柱状图分析时，这些属所占的比例非常低，在图中很难分辨这些丰度较低的物种，选择丰度前 20 的优势物种，将剩余的物种归为一类，一方面图形增加图形的辨识度，另一方面优势物种基本就能反映样品的物种组成结构），将其余的物种设置为 Others，进行相对丰度计算（即计算每个分类的百分比，不同样品间的序列数目不完全一致，直接用物种序列数表示样品丰度，并作为样品之间表达柱状图的相互比较与展示显然是不合理的，因此归一化处理是十分必要的），得到相对丰度文件，并以此绘制样品相对丰度堆叠柱状图，便于直观进行样品间物种丰度的比较。

结果路径：summary/6_taxonomy_community/*/All/*_abund_top30_stacked_bar.png

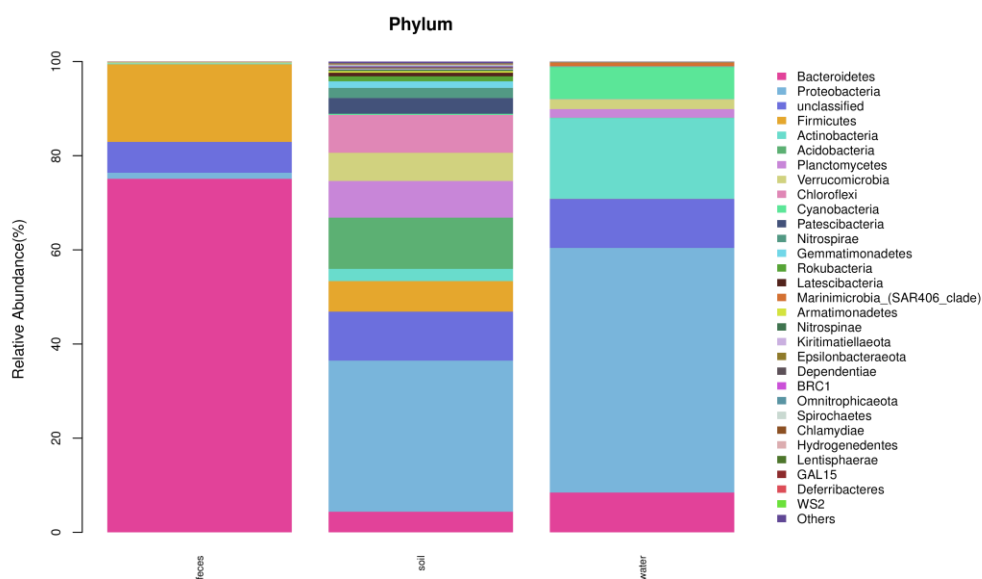


图 11 分组柱状图分析图

结果说明：图中不同颜色表示不同物种，柱子的最低的色块表示丰度在所有样本中丰度和的最高微生物，以此类推。不同柱子代表不同分组，每个分组的丰度是该组内所有生物学重复的平均值。

结果路径：summary/6_taxonomy_community/*/Group/*_abund_Group_stacked_bar.png

3.6.2 热图(heatmap)

根据样品相对丰度表，将各分类水平相对丰度最高 30 个的群落组成数据根据分类单元的丰度分布或样本间的相似程度加以聚类，根据聚类结果对分类单元和样本分别排序，并通过热图加以呈现。通过聚类，可以将高丰度和低丰度的分类单元加以区分，并以颜色梯度及相似程度来反映多个样品在各分类水平上组成的相似性和差异性。

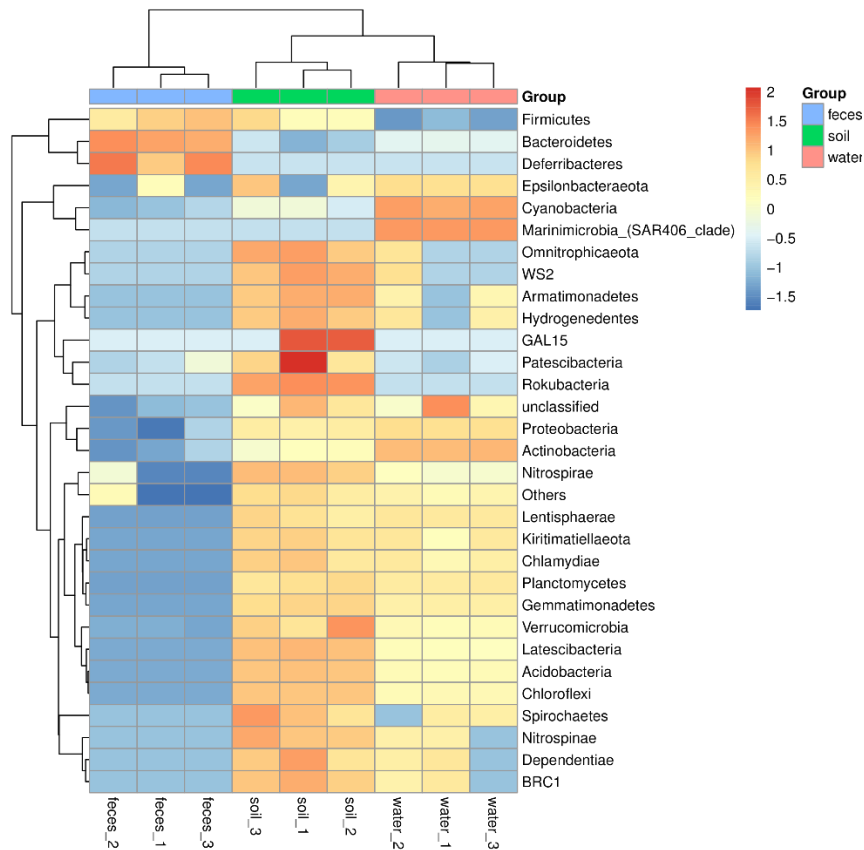


图 12 样品热图分析

结果说明：图中用蓝色到红色的渐变色反映丰度由低到高的变化，越趋近于蓝色，丰度越低，越趋近于红色，丰度越高。该热图经过 Z 值转化，将同一个菌的表达丰度进行归一化，因此该热图仅能进行横向比较，此时纵向比较已无意义。

结果路径：summary/6_taxonomy_community/*/All/*_abund_top30_heatmap.png

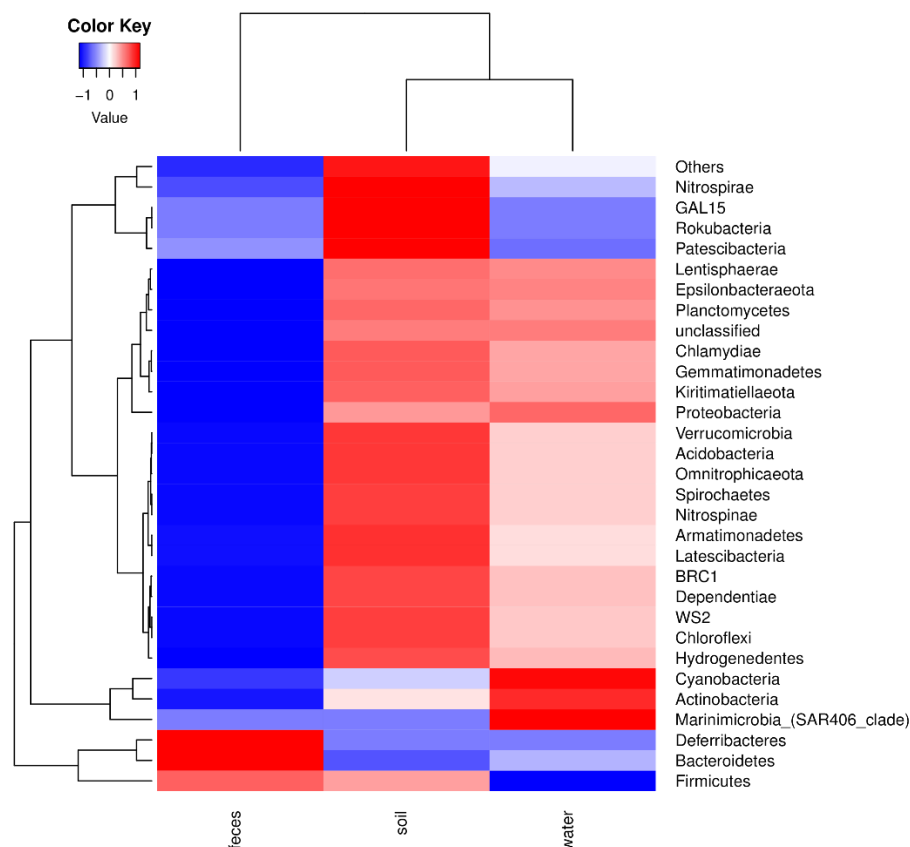


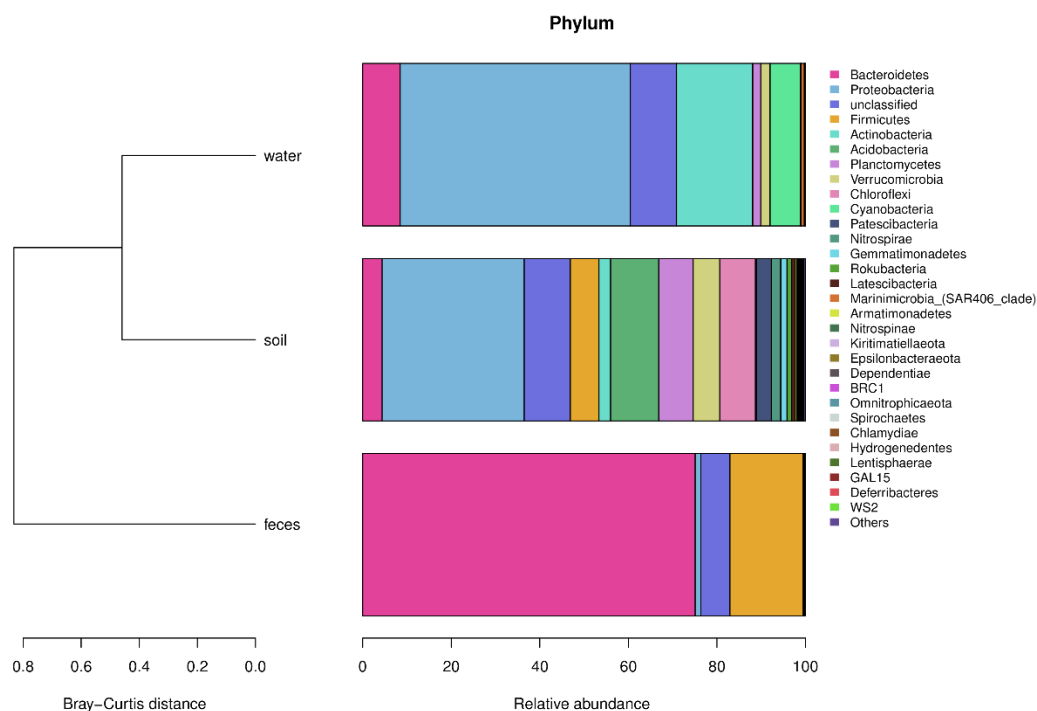
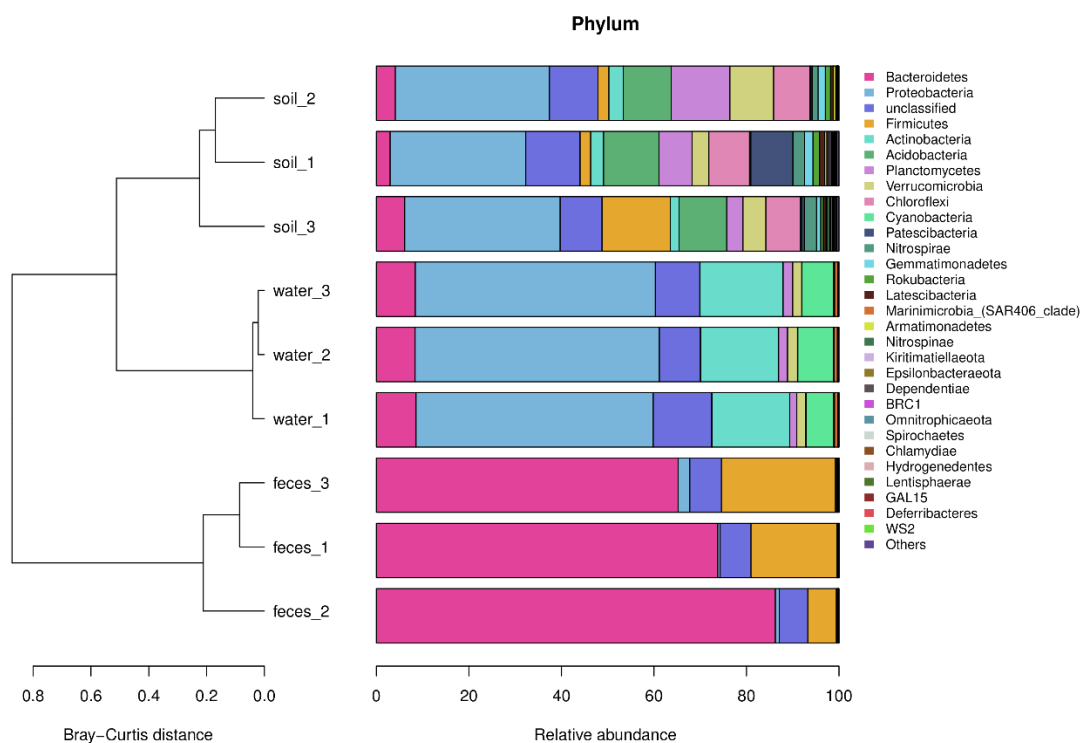
图 13 样品热图分析

结果说明：图中同一生物学重复组内取均值进行热图绘制，蓝色表示丰度越低，红色表示丰度越高。

结果路径：summary/6_taxonomy_community/*/All/*_abund_Group_heatmap.png

3.6.3 聚类图(cluster)

若该项目有生物学重复，我们将在结果中提供每组生物学重复均值的物种丰度柱状图、热图以及堆叠图。物种组成柱状图虽然可以直观看不同样品或分组的物种丰度情况，但是具体样品之间的相似性远近，是无法直接观察到的。为了更细致的研究不同样品间的差异和相似性，在柱状图的基础上，还可以根据样品物种组成距离对样品进行聚类分析。该分析使用 R 语言 vegan 包实现，采用 Bray-Curtis 距离（系统聚类法中使用最普遍的一个距离指标，主要用于描述样品间的相近程度，距离的大小是进行样品分类的主要依据）进行样品聚类。



结果说明：左侧是 Bray-Curtis 距离聚类树结构，样品聚类越近，分支越短，代表样品物种组成越相似。右侧的是各样品在门水平上的物种相对丰度分布图，所占比例越大表示丰度越高。

结果路径: summary/6_taxonomy_community/*/All/*_Phylum_abund_Group_cluster.png

3.6.4 物种注释结果 KRONA 展示

为了更直观展示各样品在不同分类水平（界门纲目科属种）的丰度情况，我们利用 Krona[6]软件对单样品物种注释结果及丰度进行动态可视化展示。该网页文件对样品物种分类学组成进行交互展示，扇形的大小反映了不同分类水平的物种相对丰度高低，并可以查看具体的丰度数值。该分析对于样品整体物种注释以及丰度情况进行动态可视化展示，客户可以根据自身的需要，选择查看不同分类水平的样品物种丰度表达情况，如可在 search 中输入老师重点关注的物种，即可在图中进行突出显示，方便查看在各个样本中该物种的表达情况。

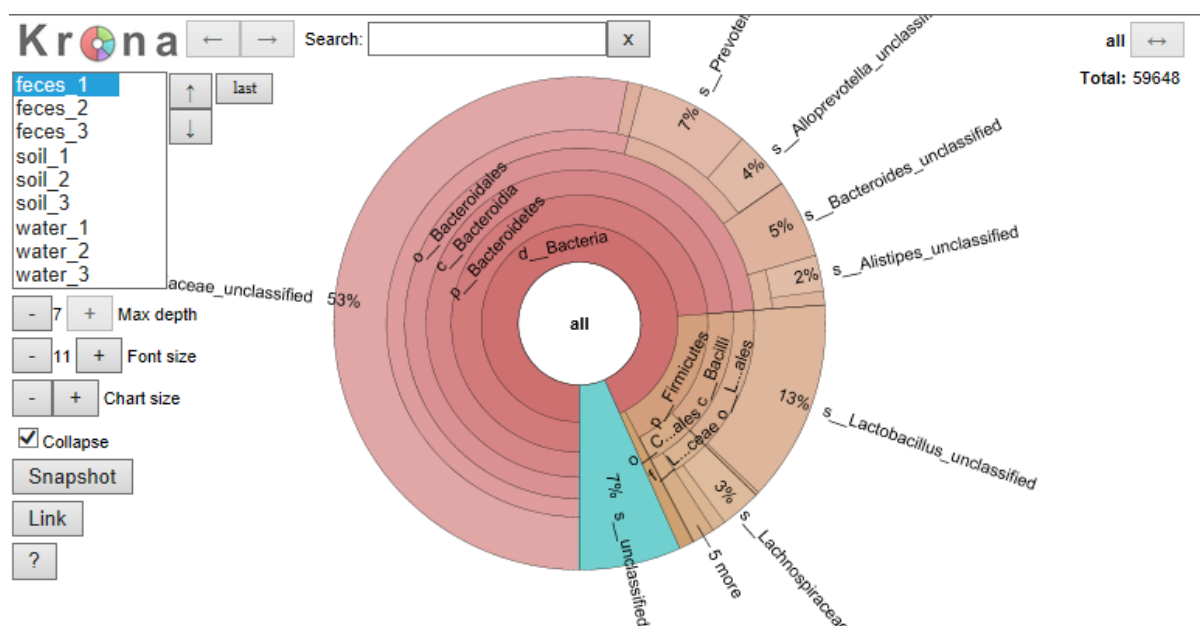


图 15 Krona 物种注释结果

结果说明：使用 Krona 软件动态可视化展示单样品在不同分类水平注释结果，通过调节不同的参数来调整展示的图片，具体的参数调整如下图所示。如果项目比较关注某一个物种，可通过在 Search 栏中输入关注物种的名称，可快速定位到关注物种在样品中的表达情况。通过点击左侧栏中的样品名，来展示该样品的物种注释表达情况。此外，还可以调整图中字体大小以及图片大小等。如果想对当前图片进行保存，点击 Snapshot 进行截图。

结果路径：summary/6_taxonomy_community/krona.html

3.6.5 物种进化树

在微生态研究中，往往需要研究特定生态系统（环境）中微生物间的系统进化关系。序列进化关系基本原理是从一条序列转化另一条序列所需要的变换越多，这两条序列之间的相关性就越小，从共同祖先开始发生分歧的时间就越早，两者之间的进化距离就越大。构建物种进化树过程是十分简单的，根据 features 特征序列的多序列比对结果构建 feature 特征序列进化树。

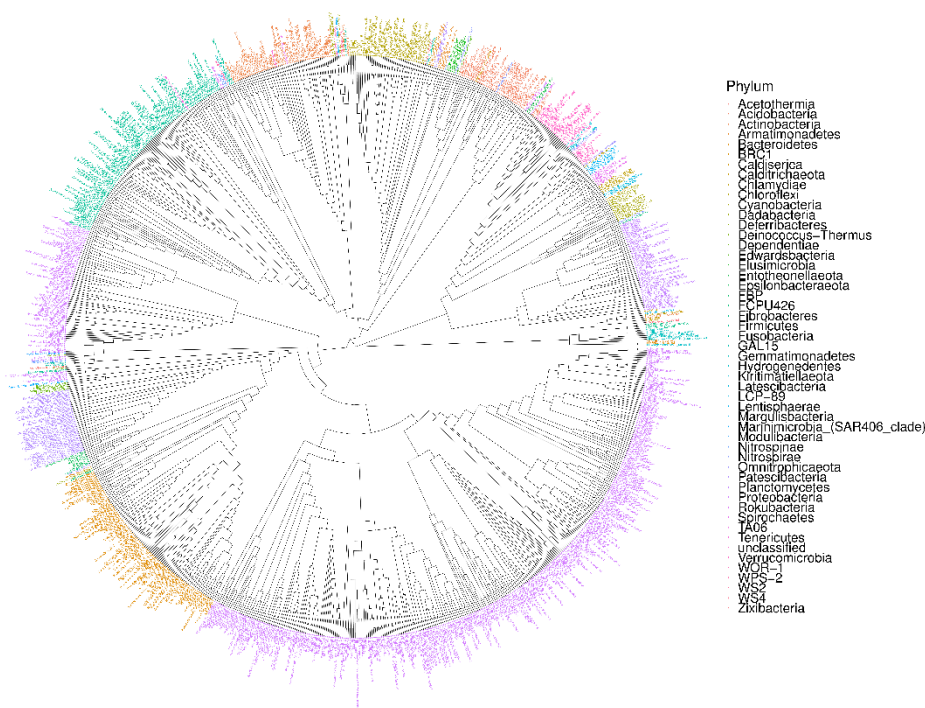


图 16 物种进化树

结果说明：图中不同分支代表不同属水平分类，不同属但是相同颜色，表示它们属于同一个门，两个物种距离越近表明他们之间的进化关系越近，可以直观的发现重点研究物种的发生关系及其生物学意义。

结果路径: summary/6_taxonomy_community/genus_phylotree/genus_phylotree.png

3.7 显著性差异分析

16S rDNA 测序分析，通过测序序列分析研究特定环境中微生物群体的构成情况。根据文献调研，16S rDNA 测序分析主要有以下研究目的：研究某一群体（如健康人肠道菌群）的微生物构成，找出这类群体

的微生物共性，从而找到微生物对群体的影响（这类研究通常样品量非常大）；研究不同处理样品（比如疾病与健康，不同肥料土壤）之间的微生物构成差异等。根据样品物种丰度表，采用 Fisher's exact test（适用于无生物学重复样品差异比较）、Mann-Whitney U test（适用于有生物学重复两组样品差异比较）、Kruskal-Wallis test（适用于有生物学重复样品的多组之间比较）进行物种差异检验。

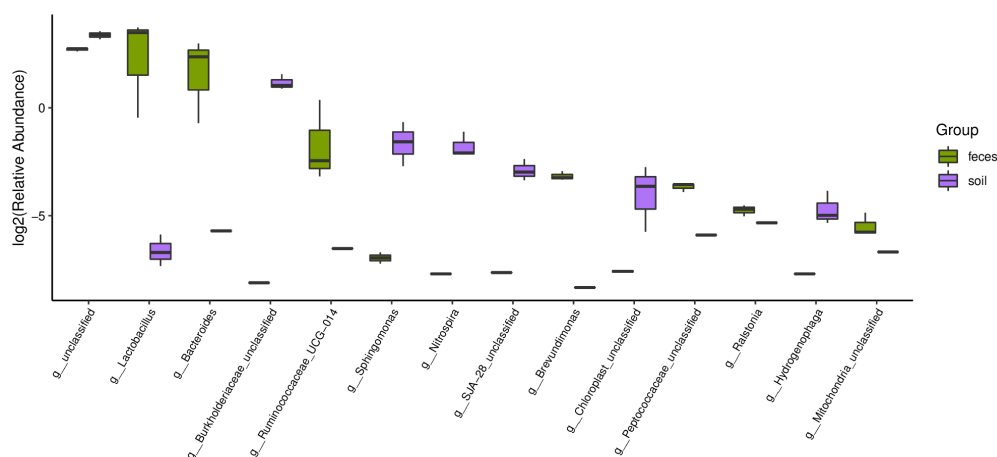


图 17 barplot 差异分析

结果说明：barplot 中展示的均是每个层级差异显著的微生物。根据上述统计检验得到的 P 值（在原假设正确的情况下，出现当前情况或者更加极端情况的概率）来判断该物种在不同组之间是否存在显著性差异，我们以 $P < 0.05$ 定义为显著差异物种。在七个层级以及不同分组进行差异分析，找到影响不同分组的重要菌。

结果路径：summary/7_diff_abundance/*/*/*_diff_boxplot.png

四、高级分析

针对不同的研究目的，我们特此更新九大高级分析。对于每个图形，对应文件夹中均含有相应的输入文件，我们根据整理好的输入文件，进行不同形式的展示。

4.1 气泡图 Bubble plot

气泡图（Bubble Plot）利用气泡的大小和颜色变化，直观反映物种注释与丰度二维矩阵中的数据信息。我们的云工具中也包含该模块，不仅可以展示属水平对应的差异，还能展示其它水平以及展示不同比较组的信息，并可以个性化修改颜色、字体等信息，按照示例文件填写提交即可。

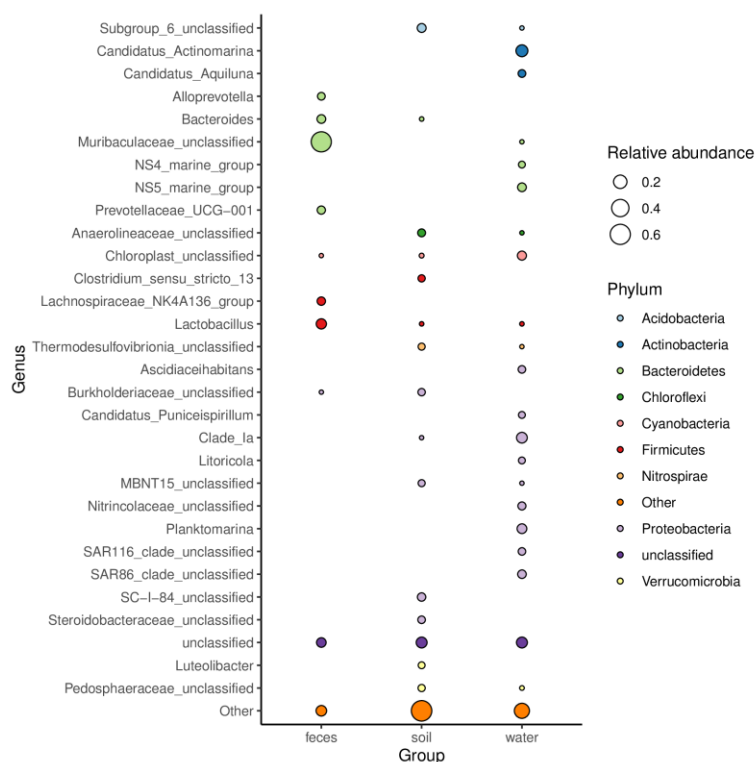


图 1 气泡图 (Bubble Plot)

结果说明：图展示了不同的样本分组中属水平的物种注释信息及相对丰度（圈的大小），同时又展示了物种对应的门的物种注释信息（圈的颜色）。

结果路径：summary/8_advanced_analyse/1_Bubble_plot/bubble_plot.png

4.2 圈图 Circos

优势菌群在不同的项目中有不同的定义，在本项目中，考虑到不同图形展示的美观程度，我们展示丰度 top5 的菌群的分布状况。您可以结合 7 个层级中 1-2 个层级的优势菌群进行展示，以满足您发表文章的需求。Circos 圈图展示样本与物种的对应关系，反映每个样本组的优势物种组成比例以及各优势物种在不同样本组之间的分布比例。以门水平结果为例：

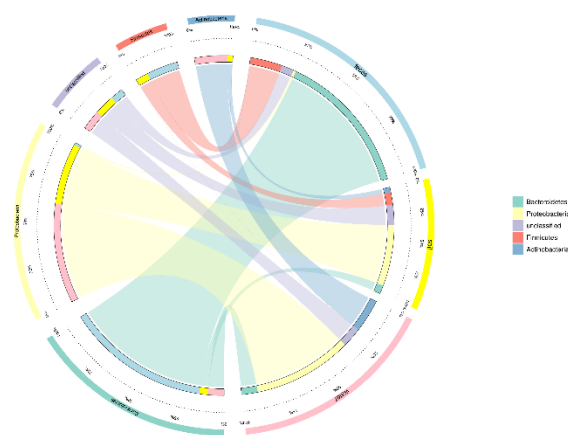


图 2 圈图 (Circos)

结果说明：左半部分是丰度 top5 的门以及其对应的丰度信息，右半部分是按照分组展示的分组信息，所占宽度越宽，表示丰度越高，相反所占宽度越短，表示丰度越低。

结果路径：summary/8_advanced_analyse/2_Circos_plot/*/*_Circos_top5.png

4.3 差异分析 LEfSe

LEfSe (LDA Effect Size) 分析主要目的两组或多组之间的比较，找到不同组间在丰度上有显著性差异的物种 (biomarker)，目前已获得了广泛的应用。LEfSe 分析绘图更加直观展现组间所有层级的差异物种，也是目前用的使用频率更高的差异物种的展示图。分析步骤主要分为三步：

Step1：利用 Kruskal-Wallis 秩和检验检测所有的特征物种，通过检测不同组间的物种丰度差异，获得显著性差异物种。

Step2：利用 Wilcoxon 秩和检验检测上步获得的显著性差异物种的所有亚种是否都趋同于同一分类级别。

Step3：使用线性判别分析 (LDA)，得到最终的差异物种 (即 biomarker)。

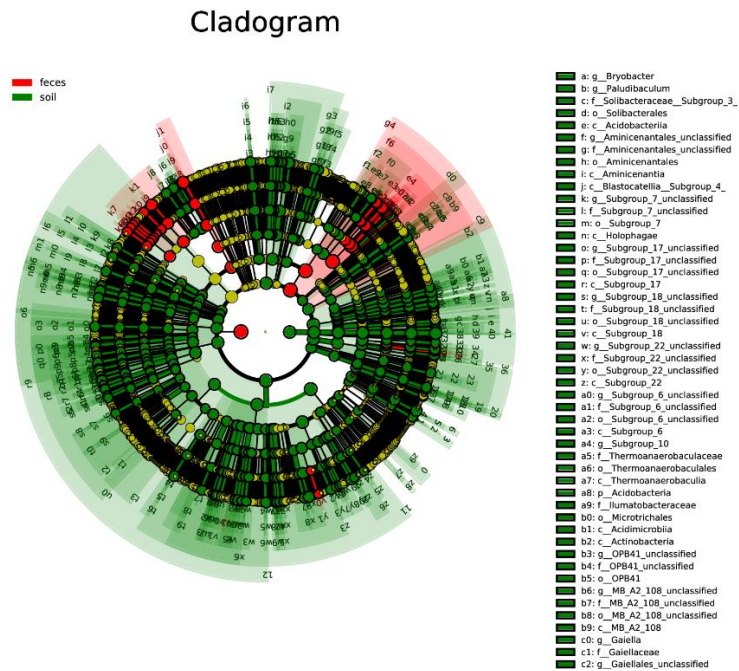


图 3 LEfSe 差异分析

结果说明：图中不同圆圈层从内至外辐射分别代表界门纲目科属种七个分类级别，各个节点代表该水平下的一个物种分类，该物种丰度越高节点越大。其中节点颜色为黄色的表示该物种在比较组中无显著性差异，如果节点颜色为红色则表示该物种在比较组中有显著性差异，且该物种在红色分组中的丰度更高，其他颜色以此类推。显著性差异的门在图中直接标出，其他水平差异的物种节点分别用字母进行标识，具体代表的物种。该结果表明两组或多组比较中，各个组中在不同层级显著差异的物种，根据实验设计及关注点，结合丰度统计等结果分析这些微生物在不同处理中的重要作用。

结果路径：summary/8_advanced_analyse/3_LEfSe_result/*_0.05_3.0

4.4 功能预测 PICRUST2

PICRUST[6]将菌群与功能建立了“映射”，从推出至今已近 6 年，引用 2500 余次，是菌群功能预测的不二之选。最近推出的 PICRUST2(<https://github.com/picrust/picrust2>)相比 V1 版本，具有以下两大优势：1.任何 OTU/ASV 直接预测功能；2.数据库扩大 10 倍。PICRUST2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States)是一款基于标记基因序列来预测功能丰度的软件。“功能”通常指的是基因家族，如 KEGG 同源基因和酶分类号，但可以预测任何一个任意的特性。同样，预测通常基于 16S rDNA 基因测序数据，但也可以使用其他标记基因。

基于 PICRUST2 功能预测结果，您可以得到包括来自以下数据集的基因家族：COG, EC, KO, PFAM, TIGRFAM 等数据库功能注释结果，可以用官方推荐的 STAMP 差异分析或者 LEfSe 进行后续分析。STAMP 以及 LEfSe 均是差异分析的利器，是微生物多样性结果中重要展示的分析方法。

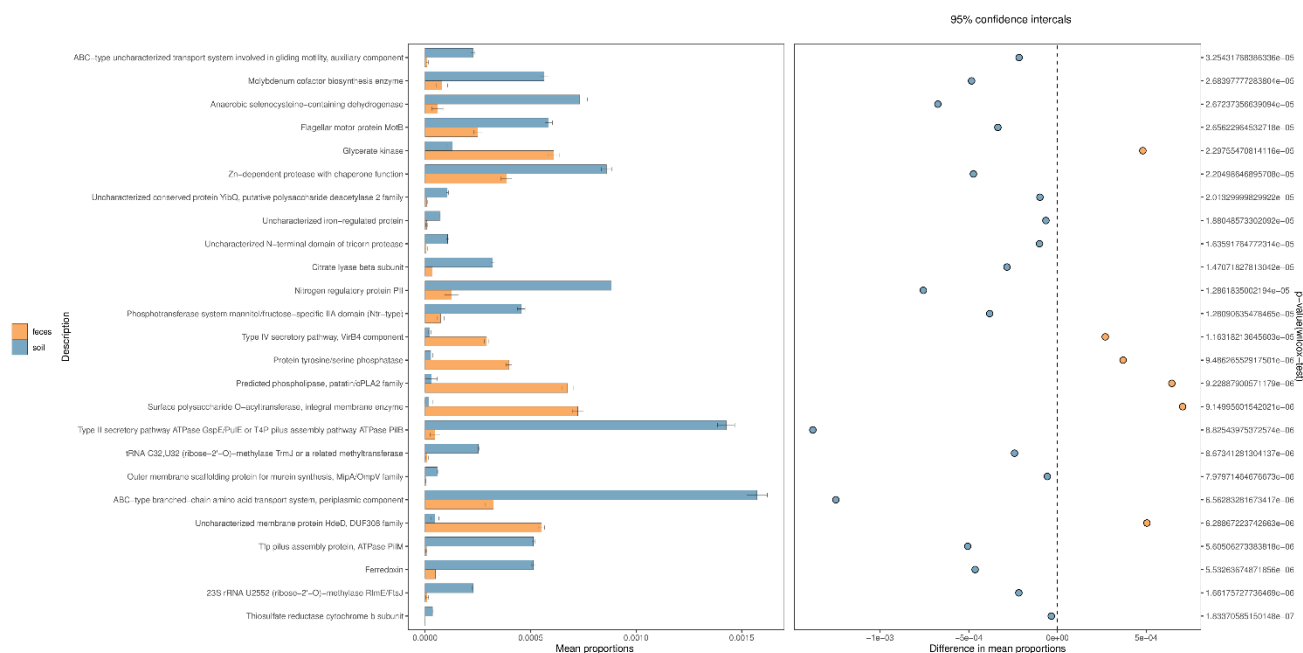


图 4 功能预测 STAMP 差异分析

结果说明：图中是根据 t-test 差异检验，p 值<0.05 的阈值进行的图片展示。结果中展示的功能数据库中的丰度数据统计学上差异显著的功能（置信区间 95%），可以初步推测您的菌群和以上图中相关功能有关。

结果路径：summary/8_advanced_analyse/4_PICRUSt2_result/*

4.5 RDA 分析

RDA 分析其实就是约束化的主成分分析（PCA），它的优点就是考虑了环境因子（如土壤研究中的 PH 值，酸碱度，疾病研究中临床理化因子等）对样本的影响，可同时反映样本，环境因子和物种三者或两两之间的关系。以门水平结果为例：

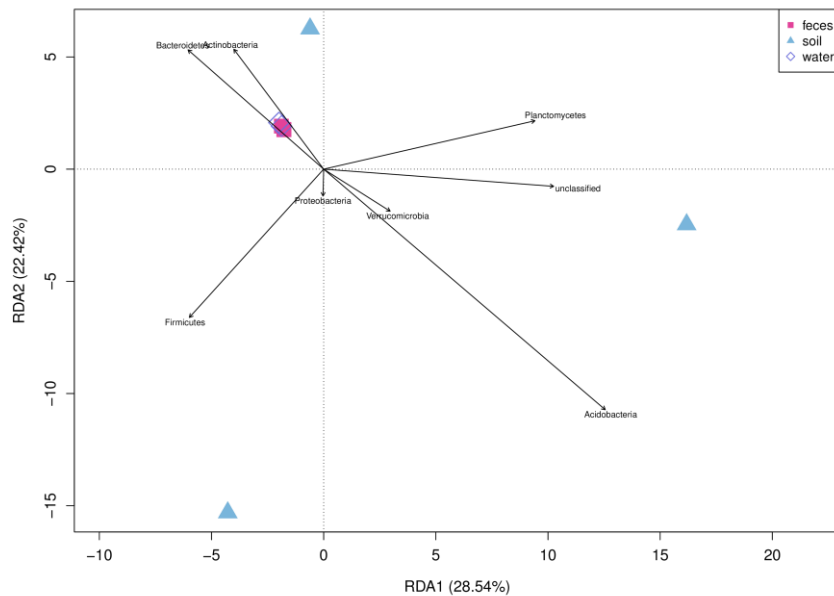


图 5 RDA 分析

结果说明：我们默认选取丰度 top10 的菌群与所有样本进行 RDA 分析，如果后期能提供相关临床指标或者环境指标（需要提供的信息与测序样本——对应）。图中每个点代表一个样本，两点之间的距离越接近，说明两个样本的群落结构相似度越高。箭头分别代表不同的影响因素，影响因素之间（因素与样本之间）的夹角为锐角时表示两个因素正相关，钝角时为负相关，射线越长，表明该因素的作用越大。样本在箭头上投影点的位置近似代表该因素在对应样本中的数值大小。

结果路径：summary/8_advanced_analyse/5_RDA_result/*/*_RDA_top10.png

4.6 桑基图 Sankey_plot

桑基图（Sankey plots）用来展示数据的“流动”变化，分支的宽度表示流量的大小。

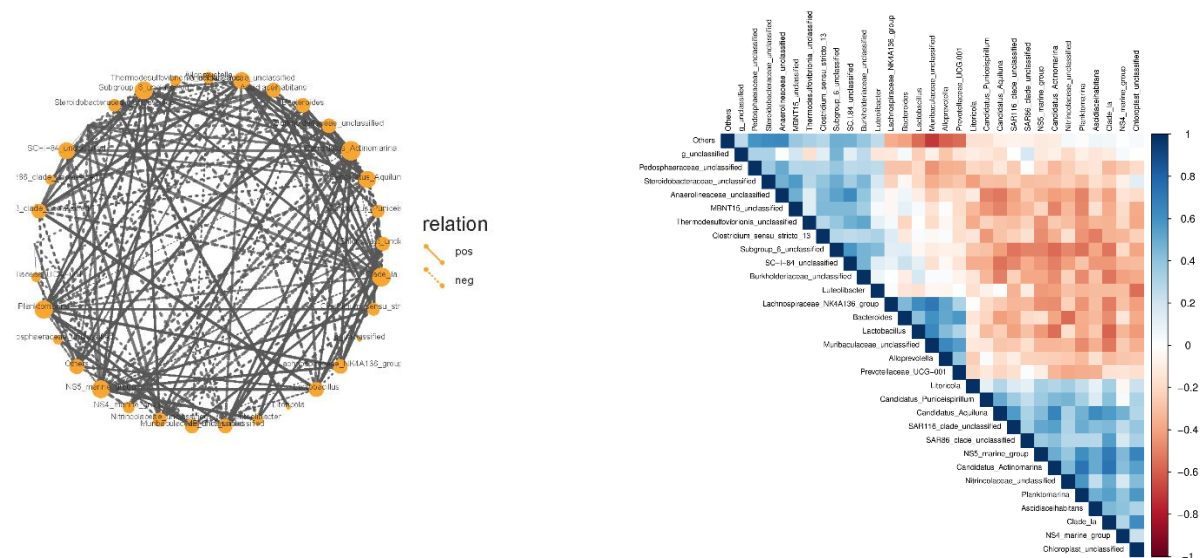


图 7 SparCC 网络图/热图

结果说明：结果说明：网络图中不同的节点代表不同的优势属，节点之间的连接表明两个属之间存在相关性，我们默认展示相关性系数 $|\rho| > 0.4$ 的关系对，线条粗细表示相关性的强弱，线条越粗，表示相关性越强；线条越细，表示相关性越弱。实线表明正相关，虚线表明负相关。节点的大小表示菌群的丰度，丰度越大，节点越大；丰度越低，节点越小。通过节点连接的数量，可以找出与菌群中其它成员的关联较多的物种，进而探索这些物种彼此相关性的生物学意义。热图的也能展示关系对的相关性，热图中通过聚类以及 p 值展示，*表示差异显著，**表示差异极显著，ns 表示表示无显著差异。

结果路径：

[summary/8_advanced_analyse/7_Sparcc_result/corrheatmap.png/network.png](#)

4.8 曼哈顿图 Manhattan

曼哈顿图本质上是一个散点图，用于显示大量非零大范围波动数值，最早应用于全基因组关联分析 (GWAS) 研究展示高度相关位点。它得名源于样式与曼哈顿天际线相似。使用差异 feature 结合分类学结果，采用 `Manhattan_plot` 展示有非常好的效果，倍受推崇。大数据中，曼哈顿图优点在于既能展示数据全貌，又能快速找到目标基因或 feature，同时可知目标的具体位置和分类、显著程度等信息。

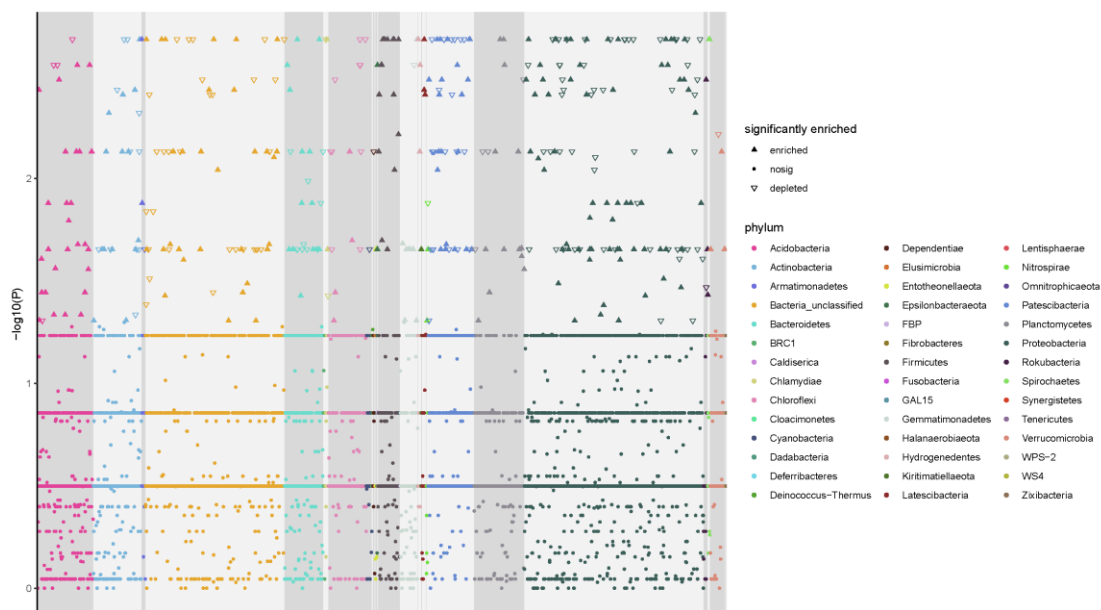


图 8 manhattan 图

结果说明：图中我们关注差异显著的 feature 以及它们对应的门，也可以初步看出该项目中哪个门差异菌群更多。其中 enriched 表示显著上调，depleted 表示显著下调，nosig 表示无显著性差异。

结果路径：summary/8_advanced_analyse/8_Manhatta_plot/*_manhattan.png

4.9 细菌表型预测

不同宿主体位的微生物群落涉及到人类健康的许多方面，包括代谢、免疫反应、疾病状态，甚至行为。随着微生物群落领域从检测疾病相关性转向识别疾病机制，了解微生物群落内微生物的功能能力至关重要。当试图设计微生物靶向治疗时，了解临床相关的微生物组表型也很重要。例如，更好地了解胃肠道中革兰氏阴性菌和革兰氏阳性菌的比例，将有助于更有针对性的抗生素治疗方法，而确定供体粪便样本的厌氧菌和病原体含量，有助于预测粪便微生物组移植的有效性和安全性。Rob Knight 实验室提供了一款神器——BugBase[7,8]，它是一款预测微生物组样品表型的工具，主要包括 9 个潜在的表型，包括好氧 (Aerobic)、厌氧 (Anaerobic)、含移动元素 (Contains Mobile Elements)、兼性厌氧 (Facultatively Anaerobic)、生物膜形成 (Forms Biofilms)、革兰氏阴性 (Gram Negative)、革兰氏阳性 (Gram Positive)、潜在致病性 (Potentially Pathogenic) 和耐应激性 (Stress Tolerant)。可以预测大量比较信息，非常适合医学领域的研究。

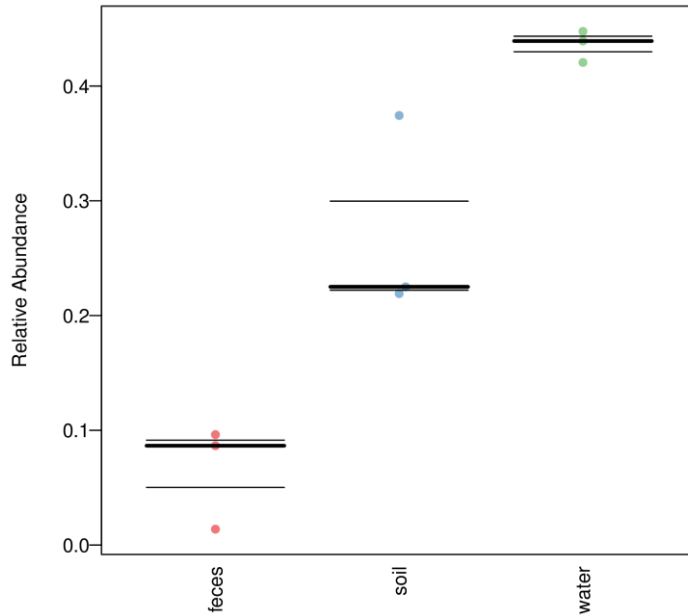


图 9 细菌表型预测 (BUGbase)

结果说明：通过预测得到的不同分组的不同表型使用散点图，对应不同比较组之间的 P 值详见：
summary/8_advanced_analyse/9_BUGbase_result/*_stats.xlsx

结果路径：summary/8_advanced_analyse/9_BUGbase_result

五、数据深入挖掘

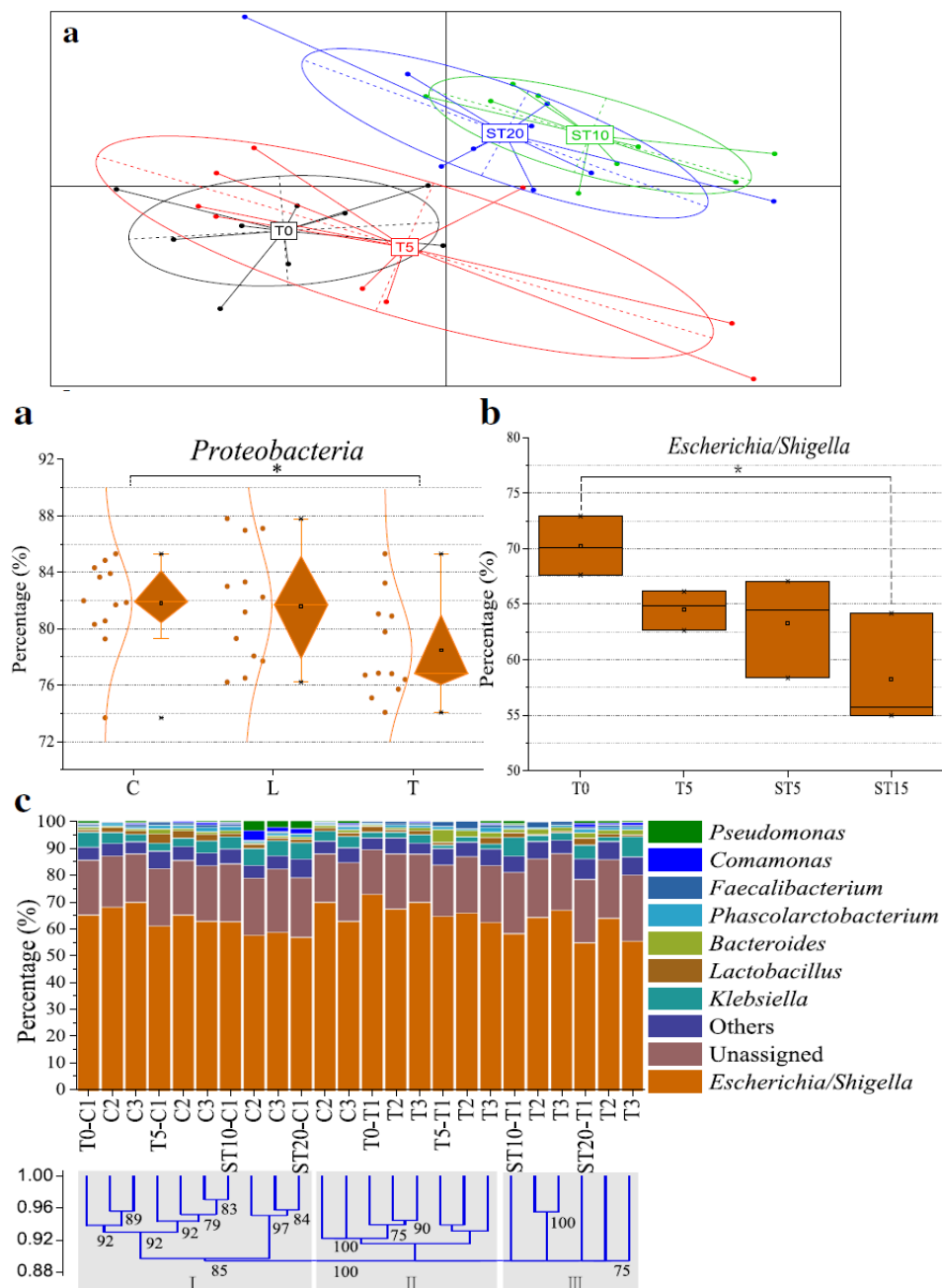
5.1 与宏基因组的关联分析

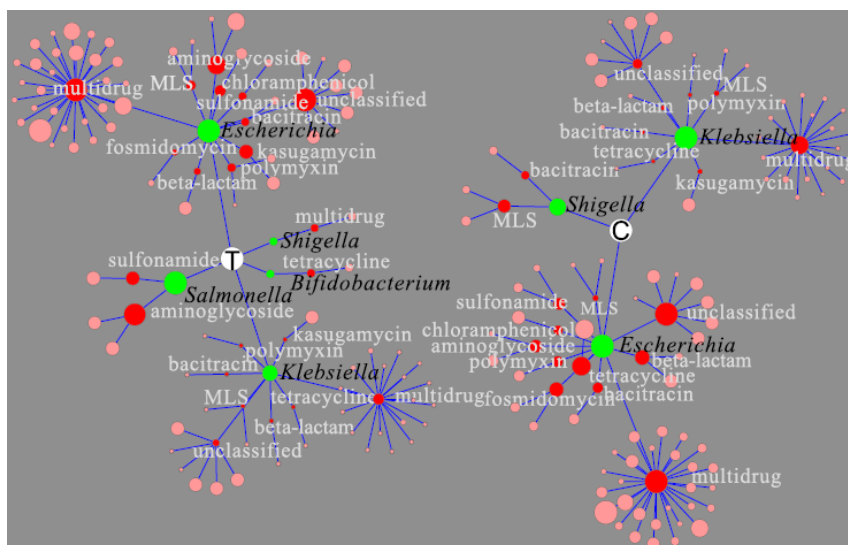
16S rDNA 测序用于研究样本的微生物结构组成，及其在不同条件下微生物种类及丰度差异，宏基因组可以进行基因水平的分析，从而研究微生物代谢功能。因此我们需要对 16S 鉴定到的微生物进行进一步分析，寻找不同条件下差异显著的微生物基因表达差异及微生物背后的运作机制。

抗生素 (antibiotics) 是由微生物 (包括细菌、真菌、放线菌属) 或高等动植物在生活过程中所产生的具有抗病原体或其它活性的一类次级代谢产物，能干扰其他生活细胞发育功能的化学物质。抗生素的普遍使用有力的抑制了普通细菌，客观上减少了微生物世界的竞争，因而促进了耐药性细菌的增长。但滥用抗生素会导致肝肾功能损害、菌群失调、抗生素相关性腹泻，以及细菌的耐药性增加，甚至是延误病情导致死亡等严重不良反应的发生。因此 Xiong 等[9]研究了抗生素对家禽的影响，进而发现抗生素介导的肉鸡肠道微生物群变化决定了抗生素耐药基因的发生。

该研究通过宏基因组结合 16S rDNA 两种测序的结果，说明抗生素对肉鸡肠道微生物的影响。首先通过宏基因组结果说明不同样品中 ARG 亚型在氯四环素给药过程中的变化(图 1)，其次通过 16S 扩增子测序展示不同细菌群落结构的变化 (图 2)，最后再经过网络图发现对照组携带 ARGs 的细菌宿主与 T5 治疗剂

量组比较的变化 (图 3), 文章得到的主要结果是: ①金霉素筛选压力对耐药基因的总量影响不显著, 只引起以四环素耐药为代表的特定耐药基因的富集; ②使用治疗剂量的金霉素降低了多重耐药基因的主要宿主菌——埃希氏杆菌属的数量, 导致粪便菌群中多重耐药基因和变形菌门的减少; ③携带金霉素耐药基因的双歧杆菌得以在金霉素筛选压力下增殖; ④金霉素导致氨基糖苷类耐药基因的主要宿主菌群由埃希氏杆菌属转变为克雷伯氏菌属; ⑤金霉素的作用具有剂量效应, 改变粪便菌群和耐药菌群结构。



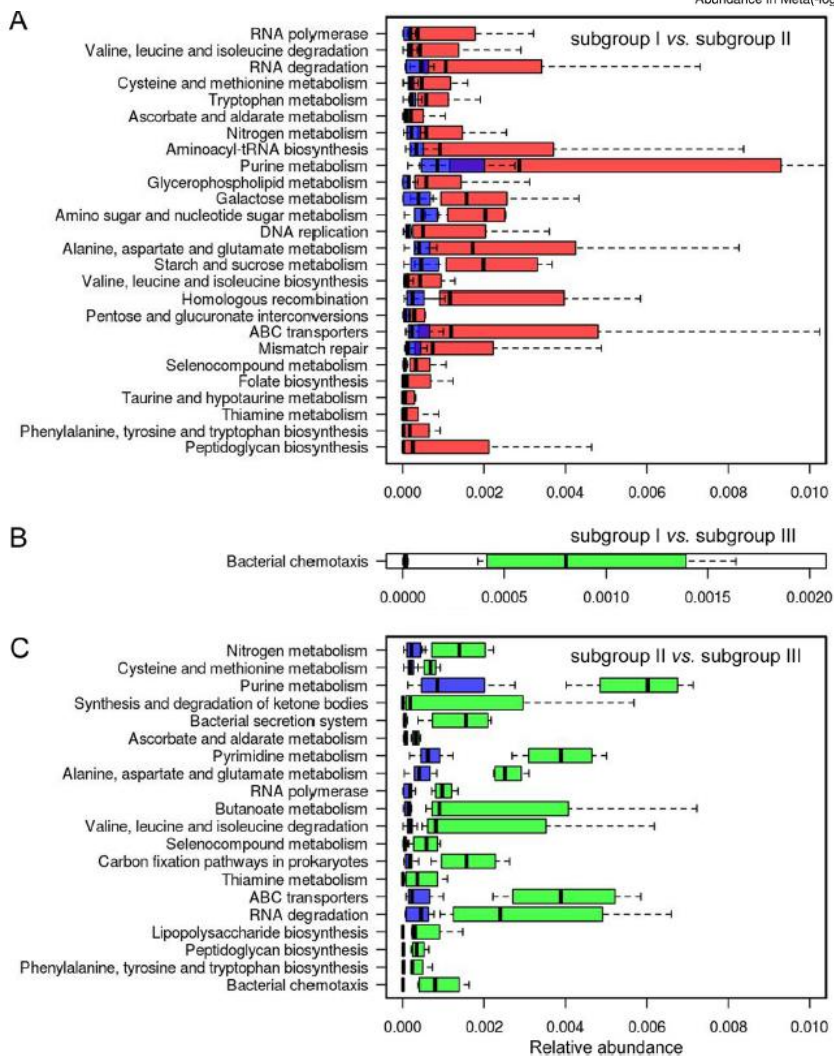
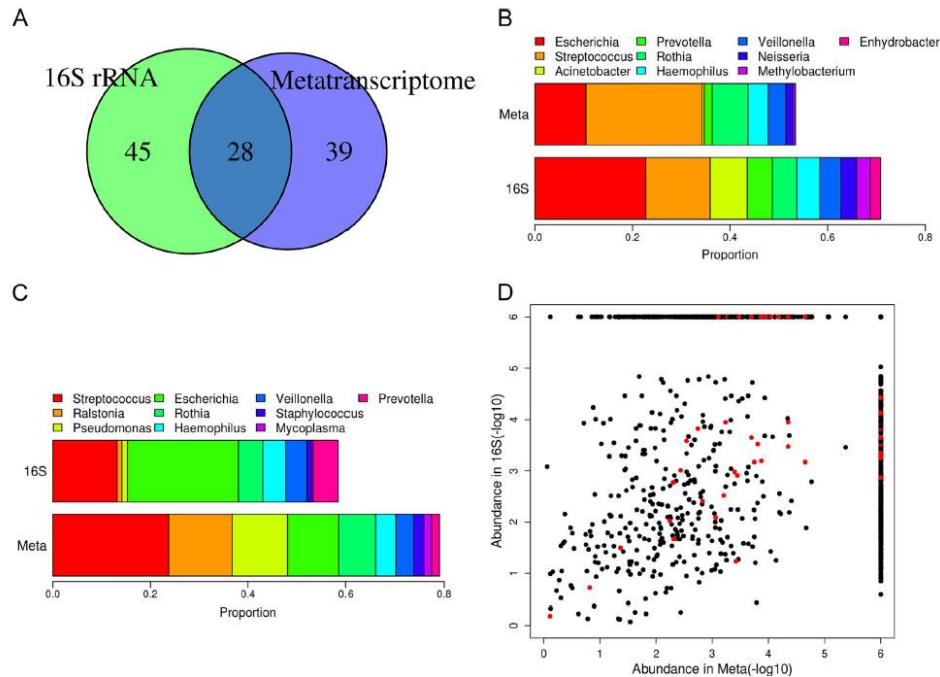


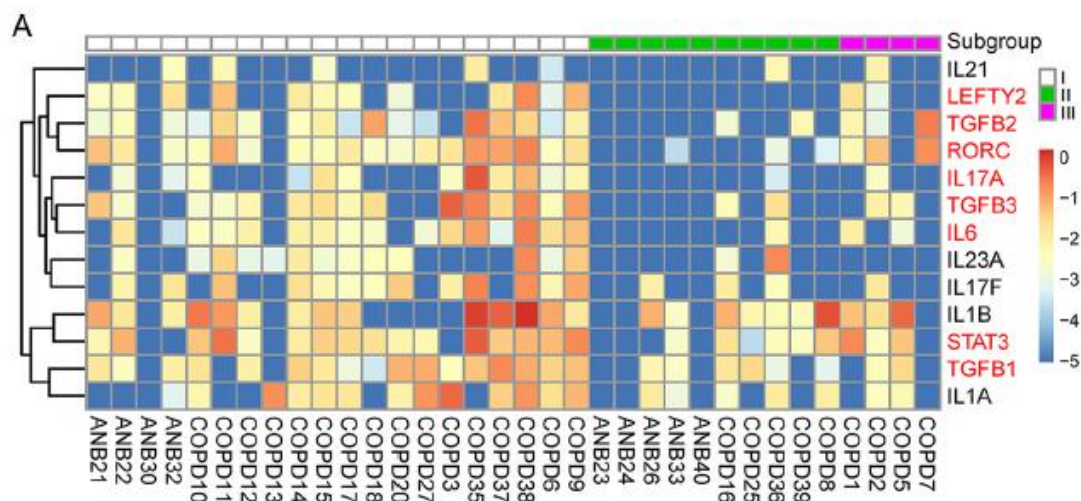
5.2 与宏转录组的关联分析

虽然宏基因组分析能够研究 DNA 水平预测基因表达变化和潜在的代谢功能，但是在特定的时空下并非所有的基因都会表达。为了研究某一特定时空下微生物群落在不同环境因子作用下的变化，比如包括高丰度菌种的组成及活性基因的表达，则需要通过宏转录组研究分析微生物群落的 mRNA 表达情况。

为了研究肺微生物群的活性及其与宿主免疫应答和临床特征的关系，Ren 等[10]对 25 例慢性阻塞性肺疾病(COPD)患者和 9 例未发现肺部疾病的对照患者的支气管肺泡灌洗液(BALF)。科学家们使用了宏转录组测序以及 16S rDNA 两种测序方式，分别对物种分类、生理指标以及肺活菌群分层结构和宿主菌相互作用进行研究。首先确认两种测序方式物种注释的异同（图 1），然后确认微生物与临床指标的关系，再说明不同样品微生物基因 KEGG 通路的富集（图 2），再展示富集出来的 KEGG 通路人体内 Th17 细胞分化展示与 Th17 细胞分化相关的 13 个关键基因的表达模式（图 3）。

主要得到的结果：①肺部的活性菌群结构复杂，菌群-宿主互作现象显著；②链球菌、罗氏菌和假单胞菌的转录活性与细菌生物量、人辅助 T 细胞-17 (Th17) 免疫应答、COPD 病情恶化率显著相关；③上述三种菌转录活性较高人群，其肺部细菌载量较高，Th17 免疫反应较强；④淋巴细胞的比例与假单胞菌转录活性正相关；⑤与假单胞菌转录活跃或细菌载量低的病人相比，链球菌或罗氏菌转录活跃的 COPD 患者病情恶化率降低。该研究结果支持了肺活菌群分层结构和宿主菌相互作用的观点，由此推测人群中存在着多种肺微生物，它们的存在和活动可能影响或反映肺健康的不同方面。

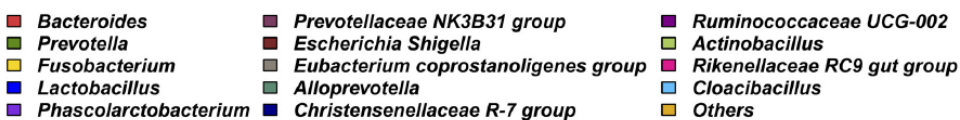
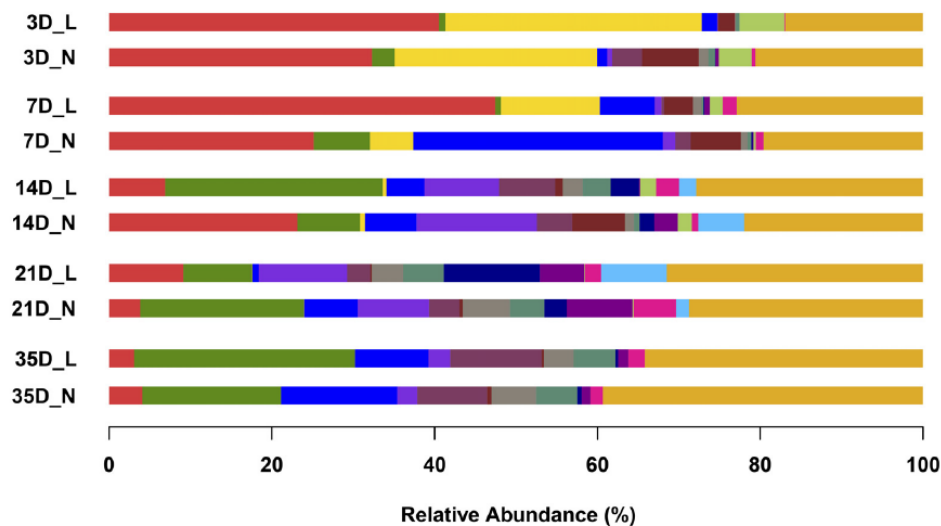




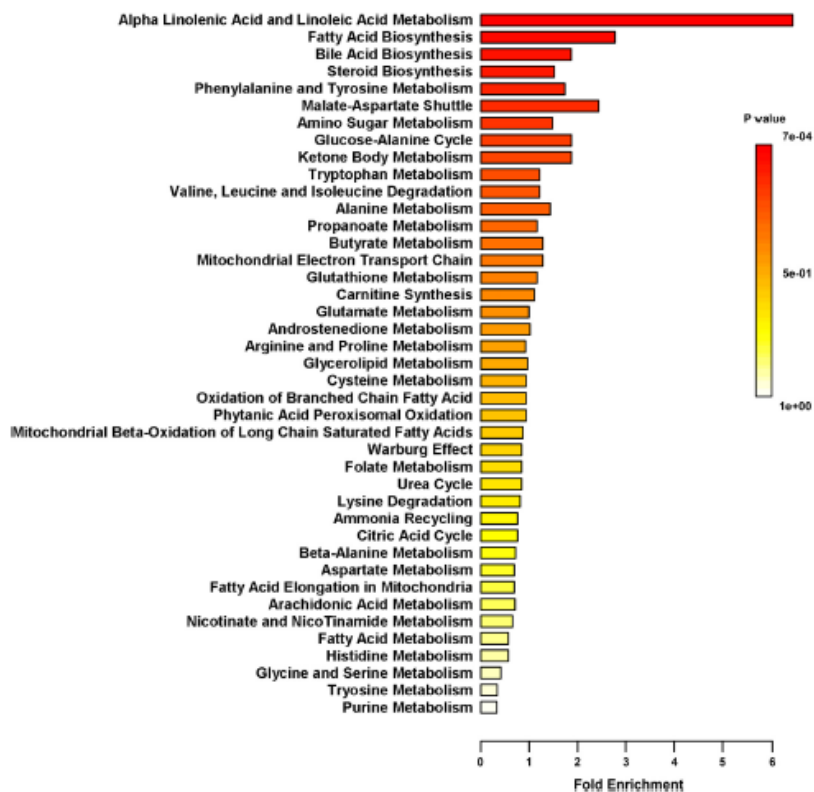
5.3 与代谢组的关联分析

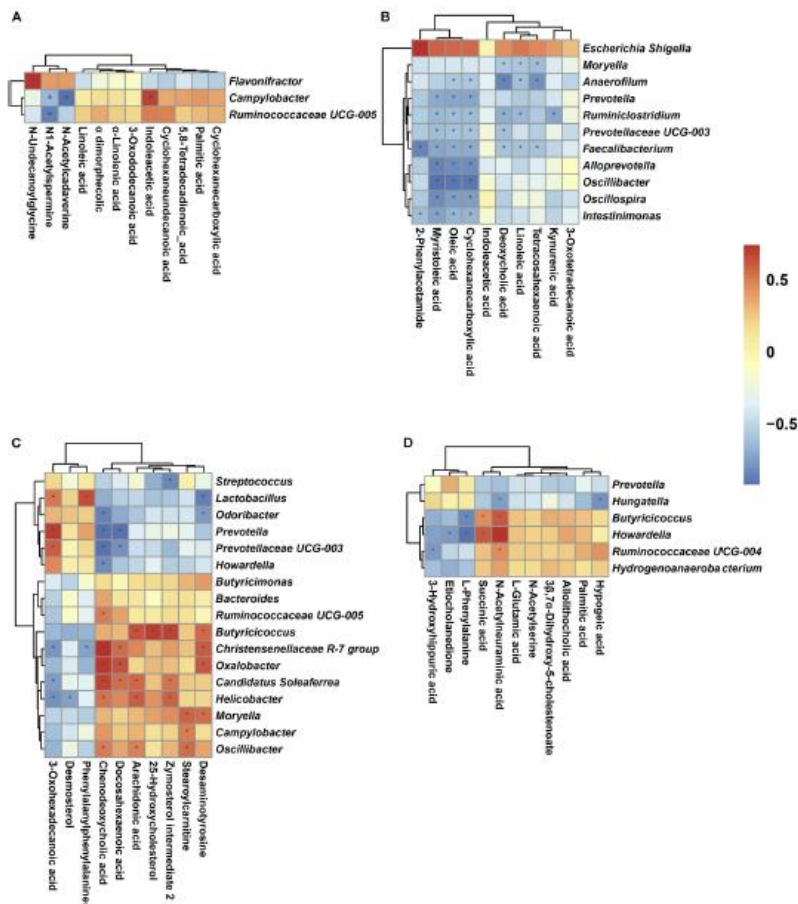
近年来，越来越多的研究发现，人类疾病与微生物组成改变有很重要的关系。但是由于微生物群落的复杂性，研究微生物如何通过功能改变对宿主产生影响是很困难的。而通过测定微生物代谢物发生的改变来研究菌群功能状态是一种有效的方法。因此，通过 16S 和代谢组学联合分析，可深入研究微生物对人体健康的作用。

Li 等[11]发现低出生体重(LBW)仔猪在出生后生长衰竭、死亡和后期代谢紊乱的高风险。早期微生物暴露是调节宿主健康和代谢的潜在有效干预策略。然而，LBW 仔猪肠道微生物群的发育是否与正常仔猪不同，及其与代谢产物谱的关系尚未得到很好的阐明。因此作者采用 16SrDNA 基因测序和代谢组学方法，研究 LBW 与正常仔猪早期(第 3 天(D3)、第 7 天(D7)、第 14 天(D14)、第 21 天(断奶前)、第 35 天(出生后))粪便微生物区系及代谢产物的差异。与正常仔猪相比，LBW 仔猪粪杆菌、乳酸菌、链球菌和普雷沃特氏菌等含量较低，而弯曲杆菌、拟杆菌等丰度明显较高。代谢组学分析结果表明，LBW 对脂肪酸代谢(如亚油酸、 α -亚麻酸、花生四烯酸)、氨基酸代谢(如缬氨酸、苯丙氨酸、谷氨酸)、胆汁酸生物合成(如糖胆酸、25-羟基胆固醇、去氧胆酸)涉及的粪便代谢物有显著影响。Spearman 相关分析显示，众多菌群与代谢物之间呈显著负相关。总的来说，LBW 仔猪与正常出生体重仔猪(NBW)相比具有不同的肠道细菌群落结构。因此猜想 LBW 仔猪独特的代谢状态可能与肠道微生物群的变化有关。



Metabolite Sets Enrichment Overview





参考文献

- [1]Callahan Benjamin J,McMurdie Paul J,Rosen Michael J et al. DADA2: High-resolution sample inference from Illumina amplicon data.[J] .Nat. Methods, 2016, 13: 581-3.
- [2]Bolyen Evan,Rideout Jai Ram,Dillon Matthew R et al. Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.[J] .Nat. Biotechnol., 2019, 37: 1091.
- [3]Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philosophical transactions of the Royal Society of London Series B, Biological sciences 360: 1935-1943. doi: 10.1098/rstb.2005.1725.
- [4]Rognes T, Flouri T, Nichols B, Quince C, Mahe F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584. doi: 10.7717/peerj.2584
- [5]Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA,

McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature methods* 7: 335-336. doi: 10.1038/nmeth.f303

[6]Langille Morgan G I,Zaneveld Jesse,Caporaso J Gregory et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.[J] .*Nat. Biotechnol.*, 2013, 31: 814-21.

[7]Tonya Ward, Jake Larson, Jeremy Meulemans, Ben Hillmann, Joshua Lynch, Dimitri Sidiropoulos, John Spear, Greg Caporaso, Ran Blekhman, Rob Knight, Ryan Fink, Dan Knights.BugBase Predicts Organism Level Microbiome Phenotypes,bioRxiv 133462

[8]Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 9: 386. doi: 10.1186/1471-2105-9-386

[9]Xiong Wenguang,Wang Yulin,Sun Yongxue et al. Antibiotic-mediated changes in the fecal microbiome of broiler chickens define the incidence of antibiotic resistance genes.[J] .*Microbiome*, 2018, 6: 34.

[10]Ren Lili,Zhang Rongbao,Rao Jian et al. Transcriptionally Active Lung Microbiome and Its Association with Bacterial Biomass and Host Inflammatory Status.[J] .*mSystems*, 2018, 3

[11]Li Na,Huang Shimeng,Jiang Lili et al. Differences in the Gut Microbiota Establishment and Metabolome Characteristics Between Low- and Normal-Birth-Weight Piglets During Early-Life.[J] .*Front Microbiol*, 2018, 9: 1798.