Novogene 诺禾致源 | 提供领先的基因科技解决方案
Leading Edge Genomic Services & Solutions

诺禾致源扩增子报告 FAQ

诺禾致源宏基因组业务线

2021.03

目录

一、实验部分

1 样品制备

1.1 为什么提取的样品跑胶都看不见条带但是扩增是可以扩增出来的？

样品的浓度太低了所以才检测不出来条带。PCR 呈阳性，是因为 PCR 模板起始量需要很低。

1.2 样品浓度检测方法？

采用的是 Qubit，准确定量的检测 DNA 浓度。

1.3 样本降解的可能原因？

样本本身保存或取样位置导致降解 ；裂解时间过长、操作过慢导致降解；核酸保存方式不佳导致降解等。

1.4 DNA 样品纯化（RNA 消化）条件？

实验的消化条件为：取原液 3 μL +6 μL ddH2O+1 μL RNase A（原浓度 10 mg/mL），37℃，15 min。取 2 μL 进行电泳。

1.5 GC 过低或过高对 PCR 的影响？

DNA 的 AT 含量高，热力学稳定性低，序列的重复性。会导致聚合酶无法复制，或引入错误。高的 GC 含量，带来了高的热稳定性。在 PCR 扩增的过程中，高 GC 的模板难以完全变性，即使变性，当退火步骤温度降低时，这些链也可能形成很强的分子内氢键，从而使聚合酶难以发挥作用。

1.6 酒曲类、发酵类样本，真菌 ITS 选择哪对引物较好？

如果扩增子项目老师要扩增真菌，样本类型是酒曲类、或发酵类可能里面有酵母，那么建议老师做 ITS2。常规 ITS1 的话，酵母会大于 500 bp；如果做 ITS2 的话，电泳检测图可能会有双带出现， ITS2 片段大小在 430 bp 左右。

1.7 专家询问我们为什么用 2%的凝胶做 PCR 产物电泳？

实验用的琼脂糖凝胶浓度通常在 0.5～2%之间，低浓度的用来进行大片段核酸的电泳，高浓度的用来进行小片段分析。我们做的是 PCR 扩增产物，条带一般不超过 500bp，所以选用高浓度胶去进行电泳。

分离不同大小的 DNA 片段所用的最适凝胶浓度是不同的，数据见下表。

凝胶浓度（%）	线性 DNA 长度（bp）
0.5	1000～30000
0.7	800～12000
1.0	500～10000
1.2	400～7000
1.5	200～3000
2.0	50～2000

2 引物

2.1 各区域使用的引物序列如下：

类型	区域	引物名称	引物序列（5’→3’）
细菌 16S	V4	515F	GTGCCAGCMGCCGCGGTAA
		806R	GGACTACHVGGGTWTCTAAT
	V3+V4	341F	CCTAYGGGRBGCASCAG

		806R	GGACTACNNGGGTATCTAAT
	V4+V5	515F	GTGCCAGCMGCCGCGGTAA
		907R	CCGTCAATTCCTTTGAGTTT
	V5+V7	799F	AACMGGATTAGATACCCKG
		1193R	ACGTCATCCCCACCTTCC
古菌 16S	V4+V5	Arch519F	CAGCCGCCGCGGTAA
		Arch915R	GTGCTCCCCCGCCAATTCCT
	V8	1106F	TTWAGTCAGGCAACGAGC
		1378R	TGTGCAAGGAGCAGGGAC
真核生物 18S	V4	528F	GCGGTAATTCCAGCTCCAA
		706R	AATCCRAGAATTTCACCTCT
	V9	1380F	CCCTGCCHTTTGTACACAC
		1510R	CCTTCYGCAGGTTCACCTAC
真菌 ITS	ITS1-5F	ITS5-1737F	GGAAGTAAAAGTCGTAACAAGG
		ITS2-2043R	GCTGCGTTCTTCATCGATGC
	ITS1-1F	ITS1-1F-F	CTTGGTCATTTAGAGGAAGTAA
		ITS1-1F-R	GCTGCGTTCTTCATCGATGC
	ITS2	ITS3-2024F	GCATCGATGAAGAACGCAGC
		ITS4-2409R	TCCTCCGCTTATTGATATGC

2.2 引物设计中简并碱基的目的

简并碱基：是其所设计的引物序列某位置的核苷酸可以分别是两个或两个以上不同的碱基，结果所合成的引物是该位置上不同序列的混合物，目的是为了增加扩增的效率，因为研究标明不同样本该片段的碱基位点不是固定的保守序列。

3 阴性对照

3.1 为什么扩增子建库测序没有设置阴性对照

诺禾致源扩增子建库测序过程中，扩增环节都是有阴性对照的 PCR 反应孔，即使用 ddH2O 作为阴性试剂进行后续 PCR 操作，只有阴性对照没有条带的情况下才会进行后续实验，这是目前基本所有公司的正常操作（除少部分小公司目前该阴性操作仍不做）。我们的引物稀释等用到的试剂，均是用灭菌的 ddH2O

配制的，所以对照使用 ddH2O 作为阴性对照能说明情况，便没有在进行阴性样品建库，测序等后续工作。另外实验实验使用 ddH2O 等为阴性对照，即使有轻微的污染，数据产出率也会极低（也就是 CK 没有条带），reads 数目很少（一般不会超过 1000 条），按照目前的信息分析经验，这个数量级别的数据是不足以引起我们文章结果中在统计学水平的如此高丰度的差异，即结果中的高丰度差异，不应该是试剂污染所导致的差异。

二、生物信息学分析部分

1 交付结果

1.1 结果中图标注释数据是否清晰完整，并且可以进一步调整？

目前我们提供的分析报中的结果展示都会提供矢量图（pdf 格式或 svg 格式）形式，符合分辨率，并且可以借助额外的编辑器修改图片（颜色，字体，柱形，大小等）。以下网盘链接是一些修图软件及修图指引，可以根据需要下载：

链接：https://pan.baidu.com/s/1Wxtd0gVGw9u8qMjFf

U-axg 密码：ix8b

1.2 为什么用网页结题报告不用 word

使用网页版结题报告，主要是为了更直观、便捷地展示测序分析结果，如果用 PDF 或 word，在展示图片的时候会非常的麻烦，而且可能还会造成部分结果的丢失，网页版报告也是我们公司经过长时间的经验积累保留下的呈现方式。

2 拼接和质控

2.1 Flash 软件拼接问题

用 Flash（Mago， T.， et al， 2011）软件将有 overlap 的 reads 对进行拼接;拼接条件是什么？不满足条件的怎么处理？是将 3'端切除多少 bp 后继续拼接吗?

A:Flash 软件在拼接时主要有两个重要参数：重叠区域的最大错配率（0.1）和最小重叠区域（10bp），也即是说，我们在拼接时要保证不大于 0.1 的错配率和 PE reads 最小不低于 10 个碱基的重叠。考虑到 3'端序列质量存在系统性降低趋势，我们会根据片段长度在保证 PE reads 重叠区长度的基础上在 3'端对 PE reads 进行部分截取，这样有利于保证重叠区碱基的质量，提高拼接率。

2.2 抽平、均一化和标准化的区别？

抽平和均一化都指按某一规则（比如我们按最小序列数）对不同样品的序列进行随机抽取处理，使不同样品中的序列数一致；标准化一般是对大范围内波动的数据进行求 Z 值计算，将数据按比例缩放，使之落入一个小的特定区间，方便在图中展示。

2.3 effective%低的问题？

effective%与样品本身及扩增区域密切相关，但与后续分析相关性不大，我们后续分析均基于 effective tags 进行，effective tags 是去除了低质量序列及嵌合体的高质量序列，您的 effective tags 达到 3.5w 以上，且稀释曲线也趋于平缓，证明用于分析的数据质量及数量都是可靠的。

3 ASV 分析

3.1 细菌结果解读

3.1.1 用 NCBI 在线 blast 方法对我们注释的结果进行比对，发现可以注释到更细化的分类层级，如原来注释到属的，可以 blast 到种?

在物种注释过程中，我们用选取的 ASV 代表序列跟数据库中的已知序列比对，如果这个未知的序列跟谁的相似度最高，就认为这个位置的序列属于比对上的那个序列所代表的物种，这个就是物种注释的基本原理。但是，有很多的未知序列本身就属于未知物种，其可能与不同的两种或几种分类都有一定的相似性，这个时候就难以区分。这种情况下，我们用 RDP Classifier 注释，其原理类似于 LCA 算法，即最近公共祖先，其在得到两个同分类等级的两个不同注释结果时默认上一级共有分类单位为最后注释结果，这样就减少了错误分类。我们默认所选取的置信度阈值范围为（0.8~1）（是比较严格的要求），也即是说，如果某个 ASV 注释到某分类层级的置信度小于 0.8 ，算法会计算高一级层级分类的置信度，直到符合阈值条件，并把此部分信息作为注释结果。

Blast 方法是基于 Best hit 算法，这个算法的原则是选取最优比对结果，即会根据 e-value 值等选取最优注释结果进行展示，也就是说，在上一段的例子里，如果用 Blast 注释，就会在这些注释结果中选取一个打分最高的分类进行展示。

此外，在 NCBI 做 Blast 时采用的数据库是 NT 数据库，即核酸数据库，其中的信息比较多，我们在 16S 和 18S 注释中选用 Silva 数据库，都是高引用率的专用数据库，数据库的不同也会引起注释结果的差异。

3.1.2 NT 数据库和结题报告中描述的 Silva 数据库是什么关系?

Silva 数据库是一般细菌选择的数据库，目前一直在更新；NT 库为 NCBI 核酸库，包含较多信息，所得注释结果会很多，含有细菌之外的物种。

3.1.3 16S 结果中有 mitochondria（线粒体？）和 chloroplast（叶绿体？），这两个单词怎么解释?

我们的注释结果中出现 mitochondria、chloroplast，中文翻译为线粒体和叶绿体，但是在这里不要理解为细胞器更不要人为的翻译成线粒体和叶绿体，我们注释的数据库里命名都是其官方命名，是有其来源和依据的。

mitochondria 属于 Proteobacteria（变形菌门），Alphaproteobacteria（α 变形菌纲），Rickettsiales（立克次氏体目），它的确是一种原核微生物，另外文献：

Fitzpatrick D A， Creevey C J， Mcinerney J O. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales[J]. Molecular Biology & Evolution， 2006， 23（1）:74-85.中提到将 mitochondria 划为 Rickettsiales；

对于 chloroplast，在蓝细菌门比较常见的，文献：尹琦. 南太平洋环流区表层海水微生物群落结构研究[D]. 中国海洋大学，2012. 和文献：侯梅锋，何士龙，李栋，等. 连云港海底底泥及青海湖底泥细菌多样性研究[J]. 环境科学，2011，32（9）:2681-2688. 中都有提到蓝细菌门中的该物种，也有将其翻译为类叶绿体蓝细菌，但是不建议如此翻译，直接用其拉丁文命名即可。

3.2 ITS 注释结果解读

3.2.1 ITS 注释中 Un--s-和 IS--s-的理解

Un--s-表示可以比对到数据库中的某一参考序列，但该参考序列在该分类水平上尚无具体注释信息（在 Unite 数据库中表示为 unidentified，为方便展示，我们在注释结果中将其简写为 Un--s-）。

IS--s-，表示地位不明，即无法在该水平上进行区分（在 Unite 数据库中表示为 Incertae sedis，为方便展示，我们将其简写为 IS--s-）。

英文原文如下：

This specifies the hierarchical classification of the sequence. k = kingdom; p = phylum ; c = class ; o = order ; f = family ; g = genus ; and s = species. Missing information is indicated as "unidentified" item; “f__unidentified;” means that no family name for the sequence exists.

解释来源：https://unite.ut.ee/repository.php

3.2.2 为何会出现种水平有具体种名，而上级分类单元显示）Un--s-或 IS--s-

例如：

k__Fungi;p__Ascomycota;c__Un--s-Ascomycota sp;o__Un--s-Ascomycota sp;f__Un--s-Ascomycota sp;g__Un--s-Ascomycota sp;s__Ascomycota sp

这种情况表示在比对到的数据库的某一参考序列有具体的种水平注释信息，但是在上一层级的分类水平上无法区分（IS--s-）或所属上一层级没有定义好的注释名称（Un--s-），这种情况在微生物中较为常见，比如上述例子中的这一条参考序列，其种水平可以定义到 Ascomycota sp，然而其在科、目水平上看在数据库中没有确定能对应的注释名称，于是用 Un--s-表示，但是通过分子生物学手段可明确得知该序列分属 Ascomycota 门，于是得到完整注释信息如上。

3.2.3 种水平注释信息不是一个完整的物种名

例如：

k__Fungi;p__Ascomycota;c__Un--s-Ascomycota sp;o__Un--s-Ascomycota sp;f__Un--s-Ascomycota sp;g__Un--s-Ascomycota sp;s__Ascomycota sp

在我们的注释结果中，s__中表示的物种的最低分类信息，因此这个水平上的展示结果会包括一部分的不完全种名（如上例中的“Ascomycota sp”），还有一些尚未确认的分类信息（如“Unidentified basidiomycete”），这些都反映了当前的真菌分子生物学鉴定状态。

英文原文如下：

The “Species” column represents the lowest assignment available for that SH， it is not always a full species name. Partial species names （e.g.，“Candida sp.”）and other expressions of uncertainty （“Unidentified basidiomycete”） are not uncommon – this reflects the current state of molecular identification of fungi.

解释来源：http://www.mothur.org/wiki/UNITE_ITS_database

3.2.4 ITs 注释结果中可用信息很少，这是为什么

ITs 是真菌中的序列信息，注释分析方法是采用 Blast 与 Unite 数据进行比对分析。ITs 注释信息比较少，主要是 ITs 相对于原核 16s 的研究信息本身就太缺少，注释方法目前也没有针 ITS 的算法和软件。另外目前 ITS2 的注释信息普遍比 ITS1 要少。

3.3 古菌注释问题

3.3.1 注释结果中无奇古菌门（Thaumarchaeota）

原因：古菌注释中，若选用的数据库是 Greengene 数据库（ 13_8 版，最新版本），在这个数据库中，古菌信息有以下 4 种门（后面的数字是序列条数）：

p__Crenarchaeota 833
p__Euryarchaeota 1518
p__Nanoarchaeota 5
p__[Parvarchaeota] 89

即在目前最新的 Greengene 数据库中，就没有奇古菌门（Thaumarchaeota）的对应信息，因此也就无法注释到奇古菌门。

解决方法：目前已将 NCBI 古菌库加入流程，如果老师关注奇古菌等 Greengene 数据库未包含的信息，可换用 NCBI 古菌库为老师注释。

NCBI 古菌数据库主要包括以下几个门的数据（后面的数字是序列条数）：

p__Crenarchaeota 11558
p__Euryarchaeota 36513
p__Korarchaeota 219
p__Nanoarchaeota 166
p__Parvarchaeota 2
p__Thaumarchaeota 3561
p__Unclassified 13346

3.3.2 古菌注释结果中注释到很多细菌

首先，样本中古菌含量太低，由于古菌与细菌的同源性很高，所以在 PCR 时会扩增出那些丰度高的细菌序列，这是不能避免的；其实，这种情况也很正常，就像做植物内生菌一样，当样本中污染了植物 DNA 时，16S 的引物（目前来说特异性相当好的了）会扩增出叶绿体的序列（叶绿体序列与细菌序列同源性也很高），这个目前也不好避免，都是分析中进行过滤。因此，样本物种分布情况直接决定了前期 PCR 和后期数据分析的结果。

3.4 相对丰度水平中“相对”如何理解？

在某一分类水平上（门纲目科属种水平，或者 ASVs），样本中某个物种/ASV 对应的 tags 数目除以该样本总共聚类得到的 ASVs 对应的 tags 总数目，就是该物种/ASV 的相对丰度。在某分类水平上，同一个样本中所有物种的相对丰度加和应该为 1。

3.5 关于生物学重复偏离甚远的处理方式

生物学重复我们通常建议 5 个以上，至少 3 个。对于重复样品间存在较大差异的个别样品，建议：

1）从样品选取入手分析，生物学重复的样品，除了设定的分组条件外，可能还受到很多其他因素的影响，造成分析结果的差异；

2）对于显著离群的个别样品，推测可能为样品原因（如在采样、保藏、提取、扩增过程中出现问题等），建议剔除该样品进行分析。

3.6 维恩图以组为单位 ASV 的并集取法：

在计算每个组里的 ASV 数目的时候，对于组的 ASV 计数，采用的取并集方式（也就是说当该组的重复样品中只要有一个样品存在该 ASV，那么认为该组内存在该 ASV，若所有重复样品中都不存在该 ASV，即认为改组内不存在该 ASV）。ASV_num 是从 total tag 里面按照 97%的相似性聚类得到的 ASV 数目。

3.7 扩增子数据里宿主污染严重，剔除完宿主后，大部分只能注释到界水平的原因？

是因为剔除完宿主后所剩的序列不多，且剩余序列片段长度短，一条短序列会注释到多个物种，在使用 lca 算法进行注释时会得到上个层级的注释结果，结果就只能注释到界了。

3.8 相对丰度柱形图中 others 含义？

对应水平除去丰度前 10 之外的全部物种，包括丰度排名 11 及以后的物种以及该水平未注释到物种信息的物种。

4 样品复杂度分析

4.1 什么是 Alpha 和 Beta-diversity？

Alpha-diversity 主要关注局域均匀生境下的物种数目，因此也被称为生境内的多样性。在分析中，选取 Observed-species，Chao1，Shannon，Simpson，Good-coverage 几种不同的 Alpha 多样性指数，以表征样品中物种分布的多样性和均匀度，并直观展示测序深度和数据量情况。

参考网站：

Observed-species - observed_ASVs （http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.observed_ASVs.html?highlight=observed#skbio.diversity.alpha.observed_ASVs）

Coverage - the Good's coverage (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.goods_coverage.html#skbio.diversity.alpha.goods_coverage);

Chao - the Chao1 estimator （http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.chao1.html#skbio.diversity.alpha.chao1）；

ACE - the ACE estimator (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.ace.html#skbio.diversity.alpha.ace)；

Shannon - the Shannon index (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.shannon.html#skbio.diversity.alpha.shannon);

Simpson - the Simpson index (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.simpson.html#skbio.diversity.alpha.simpson);

PD_whole_tree - PD_whole_tree index(http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.faith_pd.html?highlight=pd#skbio.diversity.alpha.faith_pd)

Beta-diversity 指沿环境梯度不同生境群落之间物种组成的的相异性或物种沿环境梯度的更替速率也被称为生境间的多样性。我们主要采用的是（un）

weighted unifrac 即加权和非加权的 unifrac 算法，此外还有 Bray-curtis 算法等。unweighted unifrac 算法是没有考虑 ASV 丰度而计算的样品之间的距离矩阵，weighted unifrac 考虑到了 ASV 的丰度信息。Bray-curtis 算法加入了物种丰度信息，但未考虑到物种进化关系。

参考网站：

（un）weighted unifrac：http://en.wikipedia.org/wiki/UniFrac Bray-curtis：http://en.wikipedia.org/wiki/Bray-Curtis_dissimilarity

4.2 Alpha 多样性计算和稀释曲线绘制中的序列抽取原则

采用 Qiime 软件默认参数，设定最小抽取序列数为 10，最大抽取数为所有样品中最少的对应序列条数，步长为（最大抽取数-最小抽取数）/10，每一步重复取样 10 次。

详细信息可参考：http://qiime.org/scripts/alpha_diversity.html。

4.3 几种 Alpha 多样性指数的意义和区别？

Observed_species 指数：从样品中随机抽取一定测序量的数据，统计直观观测到的物种数目（也即是 ASVs 数目）。以此抽取的数据量与对应物种数来构建的曲线即为稀释曲线（Rarefaction Curve）。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 ASVs。

Goods_coverage 指数：是一个较为常用的测序深度指数，其在计算中加入了只有含一条序列的 ASV 数目和抽样中出现的总序列数目，因此能较为真实的反映样品的测序深度。

Chao1 指数：是生态学中广泛使用的 Alpha 多样性测度指数之一，用于估计群落样品中包含的物种总数，同时由于其计算中加入了丰度为 1 和 2 的物种信息，因此能很好的反映群落中低丰度物种的存在情况。

ACE 指数（Abundance Coverage-based Estimator）：是用来估计群落中 ASV 数目的指数，也是生态学中估计物种总数的常用指数之一，和 Chao1 算法不同，其计算中分别统计了只出现 1 次的物种数目、出现 10 次或以下的物种和出现 10 次以上的物种信息，并以此为基础评估未测出物种的多少。ACE 不仅考虑到了物种的丰度，同时也考虑到了物种在样品中出现的概率，因此是一个比较好的反正群落总体情况的指数。

Shannon 指数（Shannon's diversity index）：也叫香农-维纳（Shannon-Wiener）或香农-韦弗（Shannon-Weaver）指数，它的计算考虑到样品中的分类总数（Richness），和每个分类所占的比例（Abundance）。群落多样性越高，物种分布越均匀，Shannon 指数越大。

Simpson 指数（Simpson's Index）：通过计算随机取样的两个个体属于不同种的概率，来表征群落内物种分布的多样性和均匀度。simpson 指数说明可参考 http://www.countrysideinfo.co.uk/simpsons.htm

PD_whole_tree 指数（PD_whole_tree's Index）：PD 指数是基于进化距离计算得到的，反应的是群落内物种的亲缘关系，亲缘关系越复杂，进化距离越远，PD 指数越大。

shannon 和 simpson 指数都是用来描述群落多样性（物种丰富度和均匀度）的指标，而对于 simpson 指数拥有 3 种展示形式，即 Simpson's Index （D），Simpson's Index of Diversity （1 - D）和 Simpson's Reciprocal Index （1 / D），它们对于反映群落多样性的效果相近但是计算的结果形式不同；而我们用的是

Simpson's Index of Diversity （1 - D）指的是对群落里随机取样的两个个体属于不同种的概率，即 simpson 指数越大群落多样性越高（取到的的 2 个个体是不同种的概率越大），所以其计算结果和反映的趋势是与 shannon 指数同步的；另外还有一种 Simpson's Index （D）即对群落里随机取样的两个个体属于同一个种的概率，也就是与 shannon 指数相反的情况。

4.4 ASV 稀释曲线不饱和（未达平台期）是否意味数据量不够？

稀释曲线（Rarefaction Curve），是从样品中随机抽取一定测序量的数据，统计它们所代表物种数目（也即是 ASVs 数目），以数据量与物种数来构建的曲线。稀释性曲线图中，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 ASVs。但是，随着测序数据量的不断增加，在 QC 条件相对宽松的情况下，曲线的增长趋势可能会比较大，这主要因为测序存在错误，测序产品的丰度信息也不断增加，导致曲线会保持一个增加的趋势。

在增长相对不太平缓的情况下，我们可以参考 Shannon 曲线，它的计算考虑到样品中的分类总数（Richness），和每个分类所占的比例（Abundance）当曲线趋向平坦时，说明测序数据量足够大，可以反映样品中绝大多数的微生物信息。

5 多样品比较

5.1 PCA 分析的解读

在多元统计分析中，主成分分析（Principal components analysis，PCA）是一种分析、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。PCA 的数学定义是：一个正交化线性变换，把数据变换到一个新的坐标系统中，使得这一数据的任何投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，依次类推。在欧几里得空间给定一组点数，第一主成分对应于通过多维空间平均点的一条线，同时保证各个点到这条直线距离的平方和最小。去除掉第一主成分后，用同样的方法得到第二主成分。依此类推。

PCA 图作图的输入数据是样品微生物物种丰度，横坐标为第一主成分，即对样品间物种丰度差异贡献最大维度，纵坐标为第二主成分，即对样品间物种丰度差异贡献其次的维度。每个样品点表示的是物种丰度多维空间的样品点在第一第二主成分平面上的投影。

5.2 PCoA 分析解读

PCoA 是一种从复杂的多维变量数据中提取主要变量，并进行可视化的方法。与 PCA 的主要思想类似，PCoA 的目的也是找到一个矩阵中的主要的一些坐标系。

5.3 Weighted 和 Unweighted 绘制的 PCoA/UPGMA 图的区别与选择

两者的区别主要在于前者是在计算群落样品之间的距离时候会考虑到样品中 ASVs 的丰度信息，而后者不考虑相对丰度信息。如果研究的生物学问题与丰度信息密切相关，使用 Weighted 的结果可能更为恰当；如果研究的生物问题与丰度关系不密切，或者各组的区分与低丰度的 ASV 更为密切，使用 Unweighted 的结果可能更为合适。在不知道所研究的生物问题是否与丰度密切相关的情况下，看哪种方法的分类效果更加符合预期，则使用该方法。

5.4（Un）Weighted Unifrac 和 Weighted Unifrac 的分析原理

我们的 Unweighted Unifrac 和 Weighted Unifrac 是利用 QIIME 软件得到的，没有计算公式。我们构建 Unifrac 距离[9，10，11]过程如下：

对于 16s rDNA 可以利用 PyNAST 构建 ASVs 之间的系统发生关系，进一步计算 Unifrac 距离（Unweighted Unifrac）。Unifrac 距离是一种利用各样品中微生物序列间的进化信息计算样品间距离，两个以上的样品，则得到一个距离矩阵。然后，利用 ASVs 的丰度信息对 Unifrac 距离（Unweighted Unifrac）进一步构建 Weighted Unifrac 距离。

6. 统计分析

6.1 比较各组之间群落结构组成的差异的方法

若想要得到不同处理之间的群落结构组成的差异，可以用 MetaStat 分析提取出组与组之间的具有显著性差异的物种信息，还可以以组间的距离矩阵作为输入，进行 Anosim 分析，得到的结果为一个类似于 p-value 的值，并以此来判断组与组之间是否具有显著性差异。

6.2 metastat 组间显著性差异分析结果的解读

我们会提供门，纲，目，科，属，种水平的组间显著性差异分析结果。

说明：我们利用 Metastats 软件（http://metastats.cbcb.umd.edu/ ）对组间的物种丰度数据进行假设检验得到 p 值，通过对 p 值的校正，得到 q 值；最后根据 p 值或 q 值筛选具有显著性差异的物种[12]。

.p.mat 是对比的 2 组间共有的物种列表。

.test.xls 是假设检验得到各组在的各物种上的显著性差异情况。

.psig.xls 是选取的组间显著性差异的物种（P <0.05 ）。

表格说明：Taxo 是物种分类信息；Mean（G1）， Variance（G1），Std.err（G1） 分别是第一组的平均值，方差和标准差；Mean（G2）， Variance（G2），Std.err（G2） 分别是第二组的平均值，方差和标准差；P value 是假设检验的 p 值，Q value 是 p value 矫正的 q 值。

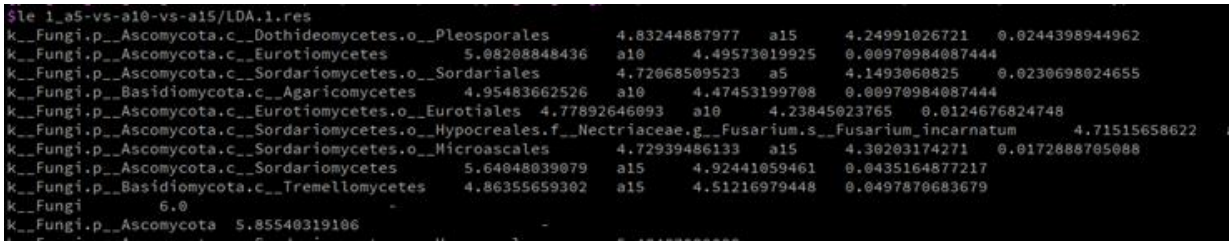
6.3 LefSE 解读问题

6.3.1 LefSE 软件参数设置：

lefse 软件默认的设置 LDA score 是 2， LDA score 的大小代表差异物种的影响大小，LDA score 大于 2 的都是可信的差异物种，值越大，代表差异物种的影响越大。文献中 LDA score 设置为 2，3，4 的都有，我们是以 LDA score 为 4 来做的，因为相对 2 来说更加严格，但是如果老师的样本组间差异并不是很大，找出的差异物种物种较少，达不到分析要求，可以降低 LDA score 来做，以期找到更多的差异物种。

```
-l float        set the threshold on the absolute value of the logarithmic
                LDA score (default 2.0)
```

6.3.2 lefse.res 文件的具体内容说明：



lefse 结果文件里有个 LDA.1.res 文件，这是 lefse 默认输出的格式，Output：输出.res 格式文件内容如下两行。

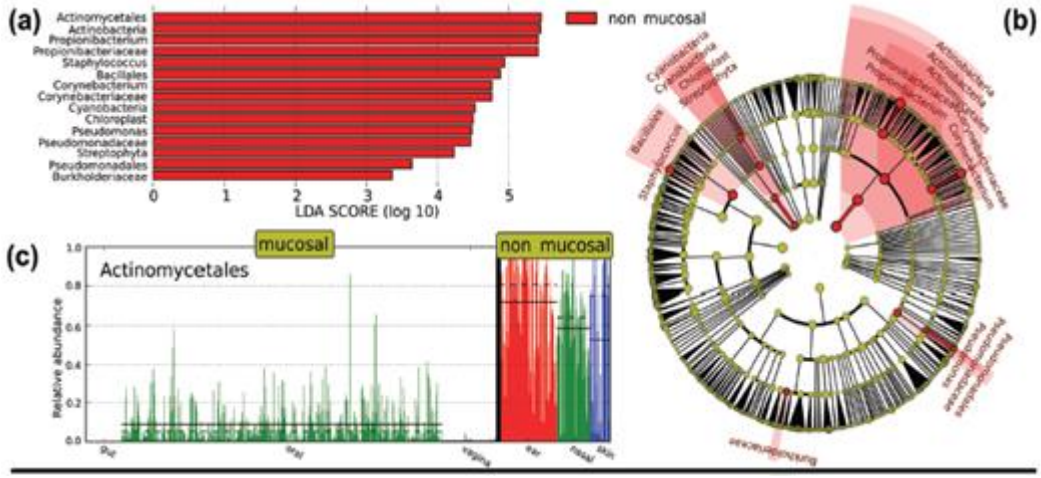
Bacteria.Firmicutes.Clostridia.Clostridiales.Ruminococcaceae 5.0923016841 Low_O2 4.74694106197 2.91304680962e-

07Bacteria.Tenericutes.Mollicutes.Mycoplasmatales.Mycoplasmataceae.Mycoplasma 2.55257491798

总共 5 列：第一列 biomarker 名称，第二列是平均丰度最大的 log10 的值，如果平均丰度小于 10 的按照 10 来计算，第三列是差异基因或物种富集的组名称，第四列是 LDA 值，第五列是 Kruskal-Wallis 秩和检验的 p 值，如果不是 biomarker 则用“-”表示。

6.3.3 LefSE 进化分支图和 LDA 柱形图中出现的分组颜色少于作图分组数目

例如：



图[16]中有 mucosal 和 non_mucosal 两个分组，但是进化分支图和 LDA 柱形图中都只有一组展示出来（红色的 non_mucosal 分组）

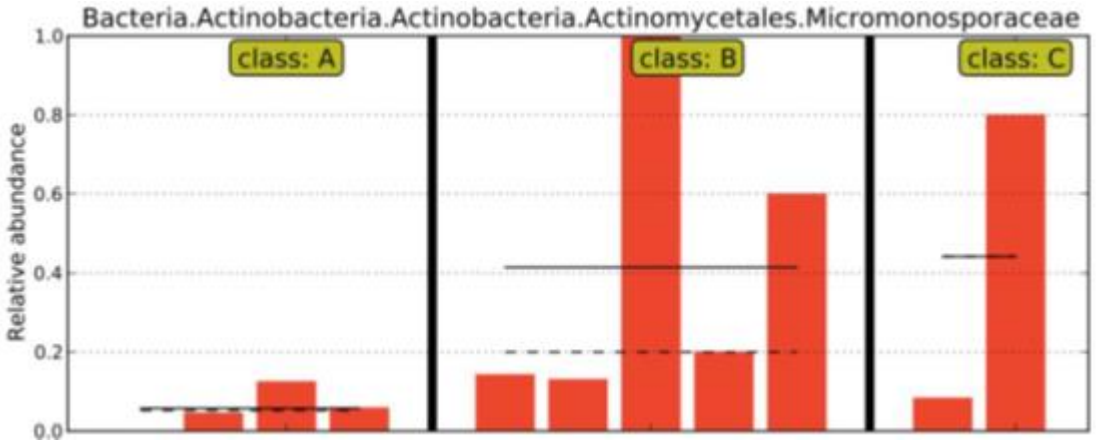
在 lefse 的展示结果中，所有展示物种都是在全部分组间差异显著的物种，这些物种的着色原则是：在哪个分组中富集（在那个分组中丰度最高）就展示这个分组的颜色，所以，如果有特定分组没有出现在展示图，原因就是在这些分组间所有差异显著的物种在这个分组中的丰度都比较低或不存在（也就是说这些物种没有在这个分组中富集）。

如上图的例子中，统计得到的差异物种共有 17 个（图 a），在这 17 个差异物种中，Actinomycetales 在分组 non_mucosal 中的丰度远远大于在 mucosal 分组

中的丰度（图 c），因此将其着色为红色（non_mucosal 分组代表颜色），其他的差异物种也同样，因此，在进化分支图（图 b）中就只显示了红色以表征这些差异物种的富集情况。

6.3.4 LefSE 分析中差异物种 在不同组各样品中的丰度比较图解读

例如：



组间不同样品丰度分布图中，将丰度最高的样品中的丰度设为 1，其他样品中该差异物种 的丰度为相对于丰度最高样品的相对值。

图中实线和虚线的含义：实线和虚线分别表示分组中各样品相对丰度的均值和中值。

如果有一个分组中无柱形，表示该分组中不存在此差异物种。

6.3.5 进化分枝图（图 b）中扇形顶部的分类名字是什么含义

图 b 中的分类名（如 Actinobacteria）代表统计学和生物学分组差异变化趋势的一致性。图中每个叶节点对应一个物种，每个叶节点其圆圈的直径与对应的物种丰度成正比，差异的着色以高丰度物种所在分组为标准。若下级分类层级的叶节点与其上面一层的祖先分支在不同分组间变化趋势是一样的，就将对应的上级分类名标注在图中，以得到一些有代表性的差异分类信息。

英文原文如下：

Taxonomic representation of statistically and biologically consistent differences between mucosal and non-mucosal body sites. Differences are represented in the color of the most abundant class（red indicating non-mucosal，yellow non-significant）.Each circle's diameter is proportional to the taxon's abundance. This representation，here employing the Ribosomal Database Project（RDP）taxonomy，simultaneously highlights high-level trends and specific genera - for example，multiple differentially abundant sibling taxa consistent with the variation of the parent clade.

6.3.6 LDA.tree 图从中可以得到什么信息?

LDA.tree 图中，由内至外辐射的圆圈代表了由门至属（或种）的分类级别。在不同分类级别上的每一个小圆圈代表该水平下的一个分类，小圆圈直径大小与相对丰度大小呈正比。着色原则：无显著差异的物种统一着色为黄色，差异物种 Biomarker 跟随组进行着色，红色节点表示在红色组别中起到重要作用的微生物类群，绿色节点表示在绿色组别中起到重要作用的微生物类群。

6.3.7 LDA 分支图的物种数和柱形图的物种数不一致的原因?

分支图只展示纲目科 3 个层级的物种，柱形图展示全部具有显著性差异的物种

6.4 LEfSe （LDA Effect Size） 和 metastat 的差别？biomaker 很少的情况如何解释？

Lefse 和 metastat 是两种不同的展现形式，其结果不同很大的原因是二者算法的不同，且 lefse 是多组放在一起分析，而 metastat 是两两组间的比较。

Lefse 算法：首先使用 non-parametric factorial Kruskal-Wallis （KW） sum-rank tes（t 非参数因子克鲁斯卡尔—沃利斯和秩验检）检测不同分组间丰度差异显著的物种，然后用成组的 Wilcoxon 秩和检验来进行组间差异性判断，最后用线

性判别分析（LDA）来实现降维和评估差异显著物种的影响大小（即为 LDA Score）。

metastat 的计算方法是首先对组间的物种丰度数据进行假设检验得到 p 值，通过对 p 值的校正，得到校正后的 q 值;最后根据 p 值或 q 值筛选具有显著性差异的物种。在门， 纲，目，科，属，种 6 个层级分别做组间物种差异显著性分析，得到不同层级， 两两比较的差异显著的物种。

所以说若想看两两组间的差异物种的时候，就可以参照 metastat 的结果，想多组放在一起看的话，可以参照 lefse。

Biomaker 很少一方面可能是由于我们的 Lefse score 值设置较为严格，默认为 4，Lefse 软件默认为 2。另一方面可能是由于组间物种没有较大的差异。

6.5 为什么要看 t-test 和 Metastat，哪种方法的结果更好呢？

t-test 要求样本正态分布，应用广泛；我们流程上 metastat 只做组间的检验，Metastat 可以不是正态分布，并且会根据样本情况自动调整统计方法，用的是 permuted t-statistics 和 fisher 检验，对于低频的物种，即：在两个组内各自的频数和（物种绝对丰度）小于组内样品数的用 fisher 精确检验。Metastat 方法保守（q 值筛选更严格），所以找出的结果相对少，我们建议物种多的情况用 metastat 较好，但是如果 Metastat 的结果并不符合研究，可以参照 t-test 的结果。

6.6 metastat 结果问题

6.6.1Metastats 分析中 family、genus、order 及 species 的 q<=0.05 是空的？是没做还是没有满足条件的？若 Metastats 分析 p<=0.05 筛选出数个 biomarker，但是 q<=0.05 没有，说明什么？

我们 metastats 的具体方法，首先对组间的物种丰度数据进行假设检验得到 p 值，通过对 p 值的校正，得到校正后的 q 值，最后根据 p 值或 q 值筛选具有显著性差异的物种，针对 q<=0.05 是空的情况，是我们根据条件没有筛选出符合条件的物种。

另外，由于即 P value 是假设检验的 p 值，Q value 是 p value 矫正的 q 值，严格意义上来说，q 值的筛选条件较 p 值更为严格，但是若 q 值筛选出的物种较少，可以按照 p 值来进行分析。

6.6.2 Metastats 中 taxa 的有的 mean 值是 0？为什么?

mean 值为 0，有两种原因：该物种丰度太低，我们在设定的某一数量级上不足以取到数值；或者该物种在该分类水平上并不存在。

6.6.3 Metastats 中 mean=0 是没有该分类水平的话，为什么还能得出 p<=0.05 的结论 ？

因为我们的 metastats 是在两组之间进行比较，其中一组的 mean 值为 0，另一组却存在该物种信息，所以也是可以计算 p 值的，关于高级分析 Metastat 中 P value 与 Q value 的计算方式如下[30，31]：

P-value：

用 Metastats 软件，对组间的物种丰度数据进行假设检验得到 p 值；

Q-value：

用 FDR 错误控制法（Benjamini and Hochberg False Discovery Rate）对 p-value 作多重假设检验校正，FDR 校正后的 p-value，即 q-value；

FDR 的计算公式如下：q-value（i）=p（i）*length（p）/rank（p）

6.7 在组间差异物种统计学分析部分中，使用了 T-test、Metastat、LEfSe 3 种不同统计方法，结果不尽相同。

由于三种统计分析方法使用的统计检验方法不同，结果不尽相同也属正常；其中 T-test 使用的 t 检验，Metastat 会根据样本情况自动调整统计方法（秩和检验或 fisher 检验），而 LEfSe 则使用了秩和检验和线性判别分析（LDA），因此 3 种统计分析方法筛选结果均是可信的，可以根据自己的研究背景选择最为符合的分析结果。

6.8 关于 t 检验分析涉及到的问题总结

6.8.1 t 检验是需要满足正态分布的，但咱们做的过程中没有检验数据是否满足正态分布。

t 检验只能用于两组数据间的比较，进行 t 检验的前提是假设数据呈正态分布，根据两组数据间的关系是否独立，t 检验又分成了独立样本 t 检验（不配对）和非独立样本 t 检验（配对），而我们流程里默认给出的是独立样本 t 检验的结果；也就是说我们给出的结果是选用 t 检验的统计方法，来寻找所关注变量（均值）在哪两组间具有显著差异；当然，我们在做 t 检验前，对数据可进行是否正态性的验证，若符合正态性分布，这些验证反过来也增强了我们对于所得结果的信心（参考书籍，R in action 第七章基本统计分析和第九章方差分析）。

6.8.2 Tukey 和 wilcox，两个分析方法的区别；多组的情况下哪个更有参考性?

这两种分析都是两两比较的方法，都是可以给出显著性检验，不同的是：tukey 是参数检验，数据要求符合正态分布；wilcox 是非参数检验，不要求数据是否符合正态分布，因此后者更为通用一些。结果文件中应该是基于这两种检验得

到的数据，后续针对两组数据的使用方法为：如果只有两组，选用 T-test 和 wilcox 检验，如果多于两组，选用的是 Tukey 检验和 agricolae 包的 wilcox 检验。

6.8.3 *psig.xls 和*qsig.xls 的差异?

一个是 p 值显著的的结果，一个是通过校正 p 值得到 q 值显著的结果

6.8.4 t 检验结果出图，是根据 p 值还是 q 值?

根据 p 值作图，但在根据 p 值显著的进行作图时，将组均值丰度都小于 0.001（默认值 0.001）的结果过滤给去掉了，剩余的进行作图，所有图中显示会比 *psig.xls 里面的结果有可能少；

6.8.5 p 值校正得到 q 值的方法?

方法为：Benjamini and Hochberg False Discovery Rate，这种方法对于 p 值的修正过程如下：

Here is how it works:

1） The p-values of each gene are ranked from the smallest to the largest.

2） The largest p-value remains as it is.

3） The second largest p-value is multiplied by the total number of genes in gene list divided by its rank. If less than 0.05， it is significant.

Corrected p-value = p-value*（n/n-1） < 0.05， if so， gene is significant.

4） The third p-value is multiplied as in step 3:

Corrected p-value = p-value*（n/n-2） < 0.05， if so， gene is significant.

And so on.

参考文献[1]： John D. Storey . A direct approach to false discovery rates. J. R. Statist. Soc. B （2002） 64， Part 3， pp. 479–498

参考文献[2]：Benjamini， Y. and Hochberg， Y. （1995） Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B， 57， 289–300.

6.8.6 结果图中，关于右边置信区间的解释

中心圆圈的着色和均值较大组的颜色一样，在竖向虚线左边表示均值差为负，在竖向虚线右边表示均值差为正，整张图形右边部分进行的分析是对两组数据均值差异的区间估计，默认显著水平为 α=0.05，一条横向点线代表均值差落在该区间范围内的概率为 95%，区间变化越大，说明均值差波动的可能性越大，某种程度上也说明组内数据重复性不好。

6.8.7 为什么 T test 那里显示物种在组间有差异（p 小于 0.05），但是却没有给图?

因为 T test 检验画图需要满足两个条件：p_value 值小于 0.05；物种在各组间的平均丰度大于 0.001。

如果两个物种在样本中 p 值都小于 0.05，但是它们在各组中的丰度均小于 0.001，此部分也是没有图的。

6.9 箱形图出现很多离散点的原因及解释？

这些离散点是异常值，程序脚本在绘制箱形图的时候会把异常值标识出来。【即出现 2 个接近的点：纵坐标相同，横坐标不同】；关于异常值的说明，可以参看下面的链接：https://www.zhihu.com/question/36172806

6.10 如果老师需要 beta.div 里面的 unweighted_unifrac 箱型图的这个图的最大值，最小值，中位数，异常值和边界值？

这些具体的数值是根据距离矩阵算出来的，是我们分析过程中的中间文件，所以在结果中没有提供给老师。如果需要这些数值的话，需要老师提供给我们

04.BetaDiversity-PCoA- unweighted_unifrac_dm.txt 文件和的分组信息，我们可以为您计算下这些数值

6.11 Anosim 分析中 R 值大于 0，p 值大于 0.05 如何解释？

anosim 分析是先得到 R 值，R 值大于 0 说明组间差异大，但是只有 R 值不足以说明两组之间存在差异，还需要有 p 值，p 值大于 0.05 说明 R 值大于 0 这个结果不具有可信度。

7 功能预测

7.1 picrust2 是否可以提供物种与功能对应关系的文件，与 meta 功能分析的区别？

Picrust 我们是将在线的 picrust 软件进行了本地化，流程都是官网的流程，picrust 是拿 clean tags 作为输入文件与 greengene 数据库比对。Picrust 官网已经把 greengene 数据库的物种与功能的对应关系对应好了，所以比对的时候可以直接得到功能信息。由于 picrust 软件没有物种和功能对应的列表，这个是整合在软件内部的，我们拿不到这个信息。

而 meta 是将 Unigenes 与 microNR 比对得到物种信息，unigenes 翻译成蛋白序列，kegg/eggnog/数据库里也是蛋白序列，两者比较得到功能的注释结果。通过 Unigenes 桥梁可以看到物种和功能的对应关系。

三、参考文献

[1] Bokulich， Nicholas A.， et al. Quality-filtering vastly improves diversity estimates from Illuminaamplicon sequencing. Nature methods 10.1 (2013): 57-59.

[2] Edgar， Robert C.，. et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27.16 (2011):

[3] Wang， Qiong，. et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 73.16 (2007):

[4] DeSantis， Todd Z.， et al. Greengenes，. a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 72.7 (2006):

[5] Caporaso， J. Gregory， et al. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26.2 (2010): 266-267.

[6] DeSantis， T. Z.， et al. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA ge4nes. Nucleic acids research 34.suppl 2 (2006): W394-W399.

[7] Price， M.N.， Dehal， P.S.， and Arkin， A.P. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution 26(2009):1641-1650.

[8] Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research32.5(2004): 1792-1797.

[9] Lozupone， Catherine，. and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology 71.12 (2005):

[10] Lozupone， Catherine， et al. UniFrac: an effective distance metric for microbial community comparison. The ISME journal 5.2 (2011): 169.

[11] Lozupone， Catherine A.， et al. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology 73.5 (2007): 1576-1585.

[12] White， James Robert， Niranjan Nagarajan， and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS computational biology 5.4 (2009): e1000352.

[13] 张金屯. 数量生态学. 科学出版社. 2004

[14] Jan Leps & Peter Smilauer. Multivariate analysis of ecological data using CANOCO. Cambridge University Press. 2003

[15] 赖江山， 米湘成. 基于 Vegan 软件包的生态学数据的排序分析. 中国生物多样性保护与研究进展 IX， 2005

[16] Nicola Segata， Jacques Izard， Levi Waldron， et al. Metagenomic biomarker discovery and explanation. Genome Biology 2011， 12:R60

[17] Fitzpatrick D A， Creevey C J， Mcinerney J O. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales[J]. Molecular Biology & Evolution， 2006， 23(1):74-85.

[18] 尹琦. 南太平洋环流区表层海水微生物群落结构研究[D]. 中国海洋大学，2012. 和文献：侯梅锋， 何士龙， 李栋，等.

[19] 连云港海底底泥及青海湖底泥细菌多样性研究[J]. 环境科学， 2011，. 32(9):

[20] Mcdonald D， Price M N， Goodrich J， et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.[J]. Isme Journal， 2012， 6(3):610-618.

[21] ITS1 versus ITS2 as DNA metabarcodes for fungi， 2013， Molecular ecology resources.

[22] Comparison of ITS1 and ITS2 rDNA in 454 sequencing of hyperdiverse fungal communities， 2013， Sciverse sciencedirect.

[23] Large-scale fungal diversity assessment in the Andean Yungas forests reveals strong community turnover among forest types along an altitudinal gradient， 2014， Molecular Ecology.

[24]Changes in fungal communities along a boreal forest soil fertility gradient， 2015， New Phyttologist.

[25]Relationship between soil fungal diversity and temperature in the maritime Antarctic， 2015， Nature climate change.

[26] Moon C， Baldridge M T， Wallace M A， et al. Vertically transmitted faecal IgA levels determine extra-chromosomal phenotypic variation[J]. Nature， 2015， 521(7550).

[27] Wang Z， Roberts A， Buffa J， et al. Non-lethal Inhibition of Gut Microbial Trimethylamine Production for the Treatment of Atherosclerosis[J]. Cell， 2015， 163:1585–1595.

[28] Barberán A， Ladau J， Leff J W， et al. Continental-scale distributions of dust-associated bacteria and fungi[J]. PNAS， 2015， 112(18).

[29] Korajkic A， Parfrey L W， Mcminn B R， et al. Changes in bacterial and eukaryotic communities during sewage decomposition in Mississippi river water.[J]. Water Research， 2015， 69:30-39.

[30] Audic， S. and J. M. Claverie (1997). The significance of digital gene expression profiles. Genome Res 7(10): 986-95.

[31] Benjamini， Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. 29: 1165-1188.