扩增子报告 FAQ

目录

1 数据质控

1.1 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 PEreads 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。

1.2 数据过滤问题和判断标准

使用 QIIME(Caporaso, J.G., et al, 2010)软件对拼接数据进行过滤，过滤掉含 N 较多或含低质量碱基较多的序列; 含低质量碱基较多的序列。

Qiime 的质控主要有两个：a)、Tags 截取：将 Raw Tags 从连续低质量值(默认质量阈值为<=19)碱基数达到设定长度(默认长度值为 3)的第一个低质量碱基位点截断；b)、Tags 长度过滤 :Tags 经过截取后得到的 Tags 数据集 ,进一步过滤掉其中连续高质量碱基长度小于 Tags 长度 75%的 Tags[1]。

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

1.3 过滤掉拼接序列中的嵌合体序列。嵌合体序列的判断标准是什么？

首先，将质控得到的 Tags 与数据库(16S: Gold database，ITS: Unite database，http://drive5.com/uchime/uchime_download.html，18S 没有对应的数据库信息，因此不做这一步检测)进行比对(UCHIME Algorithm，http://www.drive5.com/usearch/manual/uchime_algo.html) 检测嵌合体序列(http://www.drive5.com/usearch/manual/chimera_formation.html)并去除[2]。

其次， Uparse 软件(Uparse v7.0.1，http://drive5.com/uparse/)在进行 OTUs 聚类时也会对嵌合体形成的假性 OTUs 进行判断和去除。

1.4 数据质控软件 Mothur 和 Qiime 的区别

目前文献中使用 mothur 和 qiime 进行数据质控的均比较多 ,由于 Mothur 无人维护，已经很久不更新了，而 QIIME 是有专人维护，而且不断有更新的，所以我们使用的为 qiime 软件。

2 物种注释

2.1 16s，ITs 和 18s 分析的物种注释时使用的方法、软件及版本、与什么数据库比对？

流程中 16s 和 18S 注释默认采用的是 Silva 数据库（也可以根据老师的需要进行选择）。二者注释方法都为 mothur 方法 ,mothur 为类 LCA 算法 ,即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。在注释过程中，16S 设定阈值为 0.8~1，18S 设定阈值为 0.6~1，即置信度高于设定阈值的注释结果才能完整输出。

ITs 的物种注释是使用的 Unite 数据库（unite+INSD），注释方法是采用 Blast 方法，选取最优比对结果，即在设定条件下从注释结果中选取一个打分最高的分类进行展示。

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

2.2 16S 物种注释部分， other 和横线的区别

例如：

Root;p__Actinobacteria;c__Actinobacteria;o__Coriobacteriales;Other;Other

Root;p__Actinobacteria;c__Actinobacteria;o__Coriobacteriales;f__;g__

other，表示分类时，程序无法根据规则判断应该分为哪一类；f__;g__表示，程序把未知序列比对到了某一参考序列，但是，参考序列本身没有鉴定到 f__;g__等级（比如一个特征序列能区分到目，但在该目下一些不同的科所对应的微生物上，都是一致的，那么该序列就没办法在科水平上进行区分）。

英文原文如下：

The "Other" assignments are due to ambiguity when the RDP classifier tries to assign below the order level in this case (can't decide between distinct taxa). The f__;g__ means that it did match a reference sequence well, but that reference sequence is poorly defined (not named at family level or lower). Additionally, UNITE has some species names that have no lower levels identified so you will get p__unidentified or o__unidentified and a know species.

解释来源：

https://groups.google.com/forum/#!searchin/qiime-forum/other$20_$20difference$20taxonomy/qiime-forum/uJ89GGM6C8E/d3d9XAyxS_kJ

2.3 ITS 注释结果解读

1） No blast hit 的含义

ITS 物种注释中的“No blast hit”表示依据比对规则（序列最小长度、e-value 值等设定），无法在数据库中比对到参考序列。也即是说 QIIME 无法用 BLAST 方法对此 OTU 作出分类注释。

英文原文如下：

This means that when BLAST was run to perform taxonomy assignment, there was no hit in the database that met the minimum length, e-value, and percent id requirements. In other words, QIIME isn't able to make a taxonomic assignment using BLAST for the OTU.

解释来源：https://groups.google.com/forum/#!topic/qiime-forum/SF0V8qP-oAU

2）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）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。

4)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

2.4 维恩图以组为单位 OTU 的并集取法:

在计算每个组里的 OTU 数目的时候,对于组的 OTU 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 OTU,那么认为该组内存在该 OTU,若所有重复样品中都不存在该 OTU,即认为改组内不存在该 OTU)

2.5 OTU 代表序列如何选择?

把 effective tags 去冗余,按丰度排列去掉 singleton 之后就按 97%相似聚类,代表序列就是频数最高的那一条,还有种情况,如果有 otuc 的代表序列 c 和 otua 的代表序列 a 和 otub 的代表序列 b 都是 97%相似,就把 abc 合并成 1 个 otu 然后把 c 作为代表序列。

3 物种多样性

3.1 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_otus (http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.observed_otus.html?highlight=observed#skbio.diversity.alpha.observed_otus)

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

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

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

PD_whole_tree-the PD_whole_tree Index(http://scikit-bio.org/docs/latest/generated/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 OTU 丰度而计算的样品之间的距离矩阵，weighted unifrac 考虑到了 OTU 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

3.2 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 OTUs 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 OTUs。

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 OTU 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 OTU 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。

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。

3.3 OTU 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 OTUs 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 OTUs。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。

3.4 Rank-abundance 曲线的理解

Rank-abundance 曲线是分析多样性的一种方式。构建方法是：先统计每个样品中每个 OTU 的相对丰度（OTUs 包含的序列数除以总的序列数目），将 OTUs 按丰度由大到小进行排序，得到每个 OTU 的等级（从大到小排序的序号），再以 OTU 等级为横坐标，以 OTU 的相对丰度为纵坐标做图。Rank-abundance 曲线可用来解释物种丰度和物种均匀度，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（平缓程度）反映了样品中物种的均度，曲线越平缓，物种分布越均匀，如果曲线迅速下降，表明少数优势物种，占据了样品中微生物数量的很大比例。

4 多样品比较

4.1 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。

4.2 PCoA 作图的输入文件都是哪些？

输入文件：sorted_otu_table.biom（biom 格式的 OTU table）、all.mf（分组文件）、

rep_set.tre（进化树文件）、beta_params.txt（参数文件，可以不设置），得到 unifrac 距离文件（un）weighted_unifrac_dm.txt 作为输入用 R 画图得到 PCoA 图。

4.3 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。

我们 16S 信息分析流程，PCoA 的输入文件为 beta diversity 距离矩阵，即两两样品 OTU 丰度组成的差异值矩阵。经过 PCoA 分析，生成如下的文件：

pc vector number	1	2	3	4	5	6
ZC.1	-0.0414476533655	0.0110218603597	-0.0373923021828	-0.102295507962	0.0329808773807	-2.71896307177e-09
ZC.2	0.0577765424331	-0.0701363785111	-0.000528805251401	-0.0245372229418	-0.078500940836	-2.71896307177e-09
ZC.3	0.203636172398	-0.045633826204	-0.00659817612732	0.0336002979523	0.0423029450154	-2.71896307177e-09
ZC.4	-0.151303196579	-0.048087668294	-0.0680891635468	0.0635001561793	0.00545733542362	-2.71896307177e-09
ZC.5	0.0230864113891	0.178938860715	-0.00127244208763	0.0232612017527	-0.0182152334487	-2.71896307177e-09
ZC.6	-0.0917482762758	-0.0261028480661	0.113880889196	0.00647107501948	0.0159750164649	-2.71896307177e-09
eigvals	0.0783671134193	0.0421359374837	0.0190486100512	0.0168106544351	0.00965645353248	4.4356561114e-17
% variation explained	47.203767338	25.3802252344	11.4737690051	10.1257553855	5.81648303695	2.6717799079e-14

其中，行为样品，列为主坐标轴，最下方为特征值和每个坐标可以解释的差异的百分比。

4.4 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 OTUs 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 OTU 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老师的预期，则使用该方法。

4.5 (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 OTUs 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 OTUs 的丰度信息对 Unifrac 距离(Unweighted Unifrac)进一步构建 Weighted Unifrac 距离。

4.6 PCoA 结果中的 discrete vs. continuous 两组数据有什么区别

Q :使用 QIIME 绘制的 PCoA 时候，会有 discrete vs. continuous 两组数据，他们有什么区别？

A：discrete vs. continuous 是指对样品(组)进行着色的方法。例如，如果是针对于 PH，时间，温度等这样的连续型因子时候，continuous 便会根据分组从 red 到 blue 之间按照颜色梯度进行着色；反之，如果是研究的不同群，离散型因子，这建议用 discrete。总之，两者的区别是绘图时候的着色方法不一致，PCoA 的空间分布是一致的。

4.7 UPGMA Weighted Unifrac 是什么意思？为什么用它做指标？

两张图都是 UniFrac 软件聚出来 UPGMA 图，通常研究目的只关心物种是否存在时，选用 unweighted 方法；如果聚类是需要考虑物种丰度，则选用 weighted 方法。

UniFrac 的详细说明见网站：http://bmf.colorado.edu/unifrac/help.psp，

它是指的距离矩阵计算使用的数据类型，是微生物群落研究中广泛应用的一种聚类距离。建议老师可以查阅一下这方面的文献：《宏基因组数据分析中的统计方法研究》。

5 统计分析

5.1 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。

5.2 LefSE 解读问题

1）LefSE 软件参数设置：

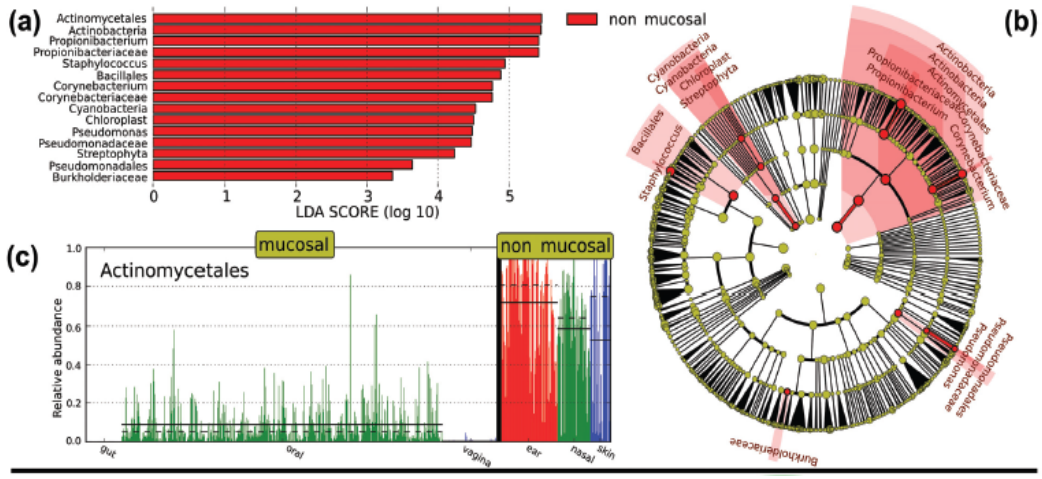
lefse 软件默认的设置 LDA score 是 2，

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

，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，达不到您的分析要求，可以降低 LDA score 来做，以期找到更多的差异物种。

2）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）

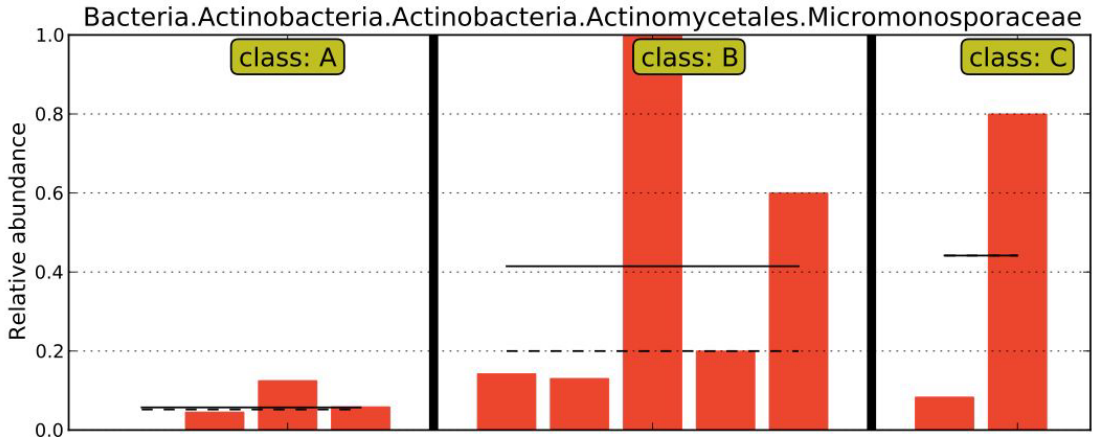
在 lefse 的展示结果中，所有展示物种都是在全部分组间差异显著的物种，这些物种的

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。

3）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。

4) 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.

5）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）.res 格式文件解读：
该文件共计 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。

5.3 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。

5.4 在 Anosim 和 MRPP 分析中，两种分析返回来的所有两两比较 P 值都是 0.1，但从 PCoA 图上能明显看出是很明显分开的两组。此种现象如何解释？

1，首先，PCoA是基于unifrac距离的（OTU间的进化距离）。而ANOSIM和MRPP是基于BC距离的，与其对应的是NMDS降维图。

2，其次，对于ANOSIM和MRPP，要求样本量要大（10个以上，样本量越多越好）。没有达到显著水平可能是以下两个原因导致的：a 即使降维图分的比较开，但是确实没有达到显著的程度；b 对于每组只有三个样品这样的情况，由于样本量太少，其内部算法，很难得到差异显著的情况（$p<0.05$）。

5.5 LEfSe 分析分析结果显示两组间有显著差异的物种；但是 Anosim 和 MRPP 组间差异分析的 R-value 和 A 值均小于 0（说明组内差异大于组间差异）

lefse和anosim、mrpp分析是两种完全不同的计算方法，其分析目的也不尽相同：

lefse主要是统计组间的差异显著物种，并将其展示在图片中；而anosim和mrpp分析主要

是看组间和组内差异哪个更大，这不代表组间没有差异；换句话说，anosim和mrpp主要看我们的组内样品的平行情况，这个是要算距离的，可能组内的某一个或两个样本的平行性较差，进而影响了组内的距离情况，而lefse分析是有计算均值的，这样就降低了个别离群值的对总体的影响；

5.6 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的问题

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，老师可以根据自己的研究背景选择最为符合的分析结果。