宏基因组测序专题手册

2017 年 5 月

编辑：施冰斌 葛长利

目录

宏基因组测序中的热门问题

1、微生物组测序设置生物学重复的意义是什么，一般情况下，设置多少个重复合适？

生物学重复是指样本重复，比如 3 只小鼠，同时做同一个处理，就是三个生物学重复。生物学重复对于测序的实验设计以及后续信息分析都非常重要，设置生物学重复主要有以下几个作用：

1）能够消除组内误差：生物学重复可以测量变异程度

2）增强结果的可靠性：测序的样本越多，越能够降低背景差异，从而增加结果可信度

3）检测离群样本：异常样本的存在，会严重影响测序结果的准确性，通过后续信息分析可以发现异常样本，将其排除。

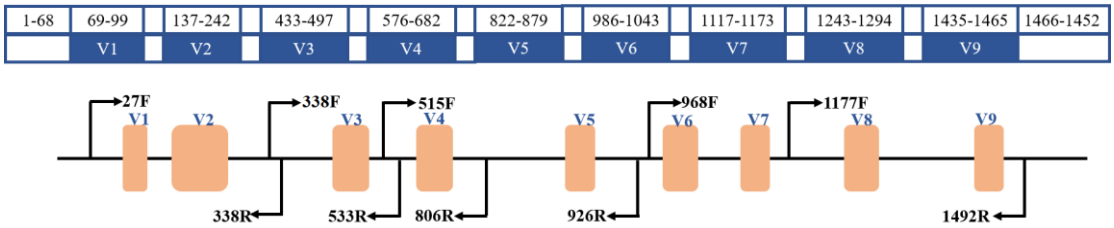
一般情况下，自然环境中（比如土壤，根系，植物等）及模式动物（比如大鼠，小鼠等）建议每组至少 5 个生物学重复，一般推荐 10 个生物学重复；若是人类肠道，粪便等样品，由于个体之间差别较大（比如环境，饮食，遗传条件，健康状态等影响），建议加大取样量，每组不少于 30 个生物学重复（取样少，可能会导致组内差异大于组间差异则项目无意义）。值得注意的是，如果将多个样本混在一起建库测序，多个样本则变成了一个样本，仍是无生物学重复的。

一般建议老师多做生物学重复，若是遇到个体之间差别较大的情况可以选择剔除个别样本，避免后期样本数量不够，后期补送重复性不好的情况发生。

2、16S 测序和宏基因组测序的主要区别是什么？

微生物测序研究常用手段包括 16S 等扩增子测序和宏基因组测序，这两者技术手段的主要区别如下：

1）测序原理不同：16S rDNA 基因存在于所有细菌的基因组中，具有高度的保守性。该序列包含 9 个高变区和 10 个保守区（下图），通过对某一段高变区序列（V4 区或 V3-V4 区）进行 PCR 扩增测序。宏基因组测序将微生物基因组 DNA 随机打断成 500bp 的小片段，然后在片段两端加入通用引物进行 PCR 扩增测序。



图：细菌 16S 区域

2）研究目的不同：16S 测序主要研究群落的物种组成，物种间的进化关系及群落的多样性。宏基因组测序在 16S 测序分析的基础上还可以进行基因和功能层面深入研究（GO、Pathway 等）。

3）物种鉴定深度不同：16S 测序序列大部分注释不到种水平，而宏基因组测序则能鉴定到种水平甚至菌株水平。对于 16S 测序，任何一个高变区或几个高变区，尽管有很高的

特异性，但是某些物种（尤其是分类水平较低的种水平）在这些高变区可能非常相近，能够区分它们的特异性片段可能不在扩增区域内。宏基因组测序对微生物基因组随机进行打断，并通过组装的方式，将小片段拼接成较长的序列。因此，在物种鉴定过程中，有较高的优势。

通常情况下，在微生态研究中，建议同时结合宏基因组测序和 16S 测序两种技术手段，更高效更准确地研究微生物群落组成结构，多样性以及功能情况。

3、16S 测序物种注释常用的数据库有哪些？这些数据库的特点（或优缺点）是什么？

微生态研究，是研究微生物之间以及微生物与环境之间相互关系，其主要研究对象包括真菌，细菌，古菌和病毒等。对于获得的大量 16S rRNA 测序序列，得到可靠的物种分类结果与全面的数据库是密不可分的。16S 测序物种常用的数据库有 RDP，SILVA ，Greengenes，NT-16S（NT 库提取整理 16S 序列数据库）等，这些数据库的详细信息见下表：

表：16S 注释常用数据库一览表

名称	版本/日期	16S 序列数目	链接
RDP（冗余）	2016.9.30	3,356,809	http://rdp.cme.msu.edu/
SILVA （非冗余）	2016.9.29	552,377	https://www.arb-silva.de/
Greengenes（冗余）	2013.5	1,242,330	http://greengenes.secondgenome.com/downloads/database/13_5
NT-16S（冗余）	2016.10.29	20,271,041	ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt.gz

注：第一列为数据库名称；第二列为数据库最新版本或更新日期；第三列为数据库 16S 序列数目；第四列为数据库网址。

RDP 数据库是目前较常用的比对、注释的参考数据库，版本更新比较快，16S 序列信息较全。

SILVA 数据库由于更新比较及时，因此也是目前最常选用的参考数据库之一。

Greengenes 数据库相对于 RDP，SILVA 数据库，长时间未更新，目前较长用于嵌合体去除的参考数据库，另外 16S 功能预测-PICRUST 软件是基于 Greengenes 的 gg_13_5 研发的，因此想做 PICRUST 分析也必须依托于 Greengenes 数据库进行比对。

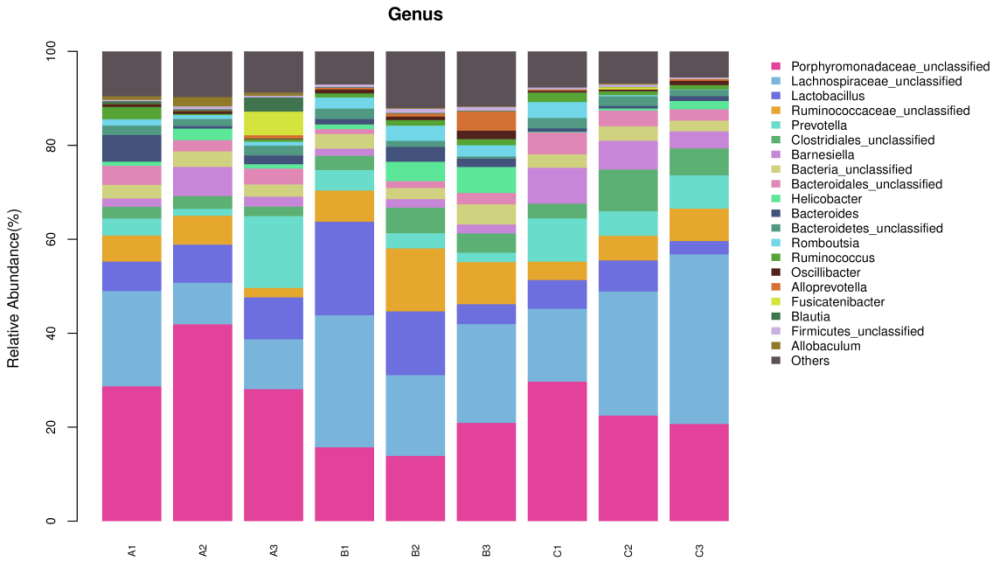
NT-16S 数据库是基于 NT 库提取整理的 16S 数据库，该数据库是联川特色数据库，由于 RDP 数据库只能注释到属水平，基于 RDP 注释结果，进一步对序列进行注释，获得种水平的注释结果。

4、16S 物种分类学注释中，常得到 UNCLASSIFIED，NO_RANK，OTHERS 等意义？

虽然理论上所有的微生物序列都应当能在种甚至菌株水平得到鉴定，但由于微生物种类繁多，目前上述常用数据库还很难包罗万象。加上二代测序读长的限制，往往只能选择 16S rRNA 1-3 个高变区（目前常用的扩增片段 V3-V4 区/V4 区）作为扩增片段进行测序，分类的精确性受到一定的限制（某些物种高变区也可能十分相近，能够区分它们的特异性序列片段有可能不在扩增区域内，因此在鉴别的过程中受到测序长度的限制），因此，在实际分析过程中，并非所有 OTU 代表序列都能获得属或种水平的分类学信息（即在对应的分类学水平尚且属于“Unclassified”）。另外，也总是有可能遇到某些较为新奇、尚未被充分研究的微生物（在较高分类水平，比如门水平鉴定为“Unclassified”），此为正常现象。

No_Rank 表示在某个分类水平上没有明确的分类信息或分类名称（与数据库有关）。

Others 一般在作图时自行定义的，比如做热图或柱状图分析时，一般选择丰度较高的物种（比如丰度前 20 的物种）进行展示，剩下的所有物种以 Others 定义，且 Others 的丰度为剩下所有物种丰度的求和。

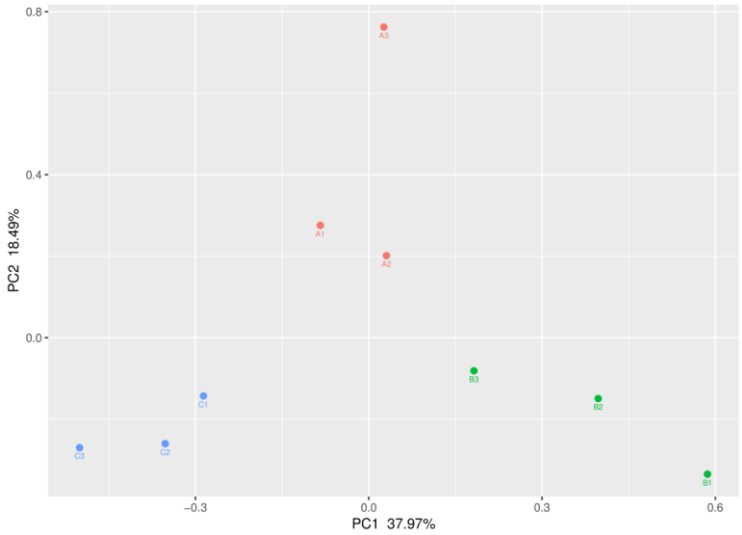


图：物种组成柱状图示例图

5、PCA、PCOA、NMDS 三种排序分析的生物学意义及区别是什么？

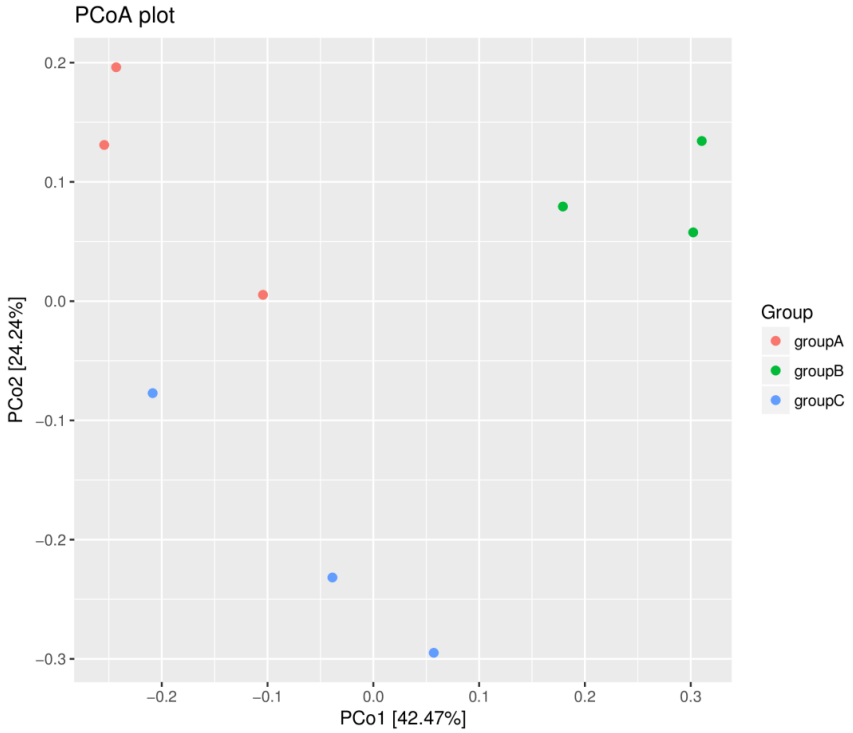
首先，PCA，PCoA，NMDS 均属于排序（Ordination analysis）分析，排序的过程就是在一个可视化的低维空间(通常是二维)重新排列这些样方,使得样方之间的距离最大程度地反映出平面散点图内样方之间的关系信息。这三种分析图都是用于比较样本或样本组间差异的。通过样本点中的距离，体现样本间的差异程度，样本间的距离越近表示样本组成相似性越高，差异越小。

主成分分析（PCA，Principal Component Analysis）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。



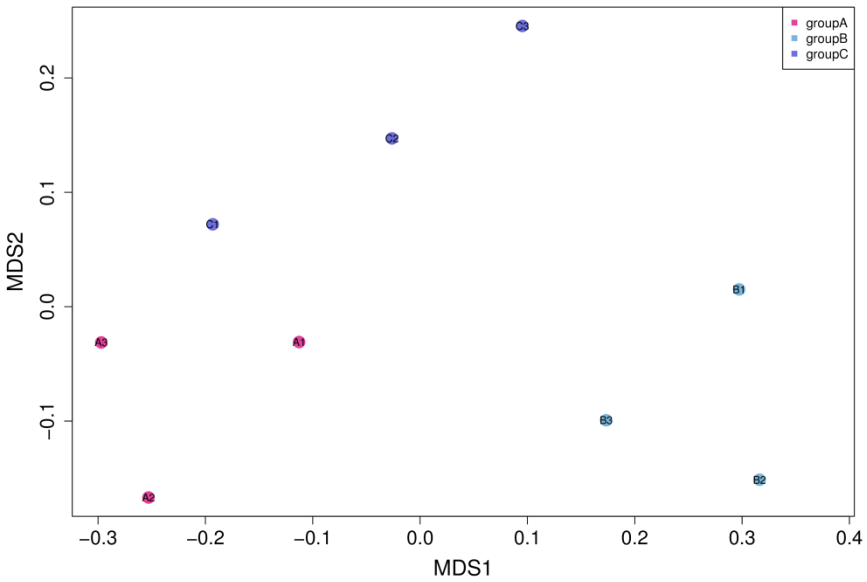
图：PCA 分析示例图

与主坐标分析（PCoA，Principal Coordinates Analysis）最大的区别在于，PCA 分析是基于原始的物种组成矩阵所做的排序分析，而 PCoA 分析则是基于物种组成矩阵计算得到的距离矩阵所做的排序分析。



图：PCoA 分析示例图

非度量多维尺度分析(NMDS，Nonmetric Multidimensional Scaling)，与 PCoA 分析一样也是基于距离矩阵所做的排序分析。区别在于 NMDS 不再基于距离矩阵数值，该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。



图：NMDS 分析示例图

当然，在 16S 测序分析中，这三者的分析结果可以相互参考，尤其是在去除偏离样本时，需要结合多个分析结果进行判定。

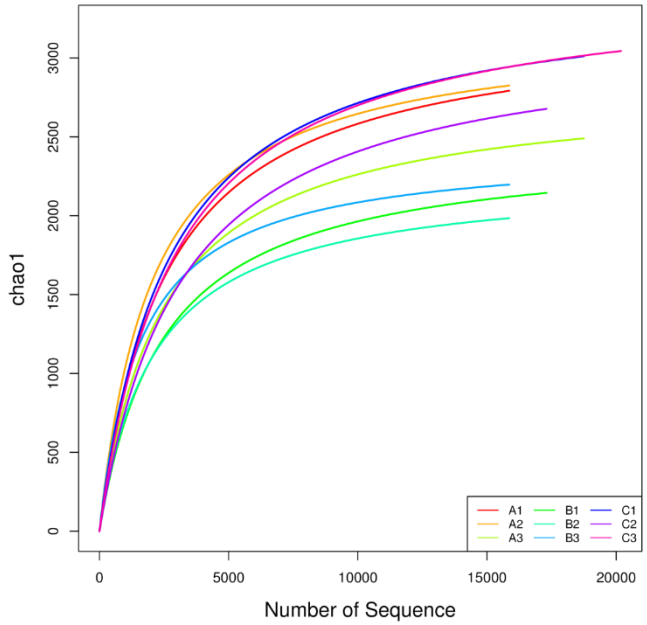
6、UNIFRAC 距离（WEIGHTED 和 UNWEIGHTED 区别）？

UniFrac 距离用于衡量不同微生物群落的距离，它利用各样品微生物序列间的系统进化信息来比较样品间的物种群落差异。UniFrac 包括 unweighted unifrac（非加权）和 weighted unifrac（加权）两种计算方法。其中非加权算法只考虑序列是否在样本中出现（即只考虑物种的有无），而不考虑序列的丰度，如果两个样本所包含的物种完全相同，那么不管每个物种的丰度是否有差别或者差别的大小，非加权计算的值为 0，相反如果两个样本的物种完全不一样，即他们是完全独立的两个进化过程，那么非加权计算的值为 1。但在某些情况下，研究者感兴趣的恰恰是群落中物种丰度的变化，比如研究人体肠道菌群在抗生素干扰下的变化情况，这时非加权算法就不能很好的解决问题了。weighted unifrac（加权）算法在非加权的基础上，将序列的丰度纳入考虑，因此它能够区分物种丰度的差别。

UniFrac 距离通过比较不同样品中群落的系统进化关系远近，从而反映样品间群落结构的相似性。Unifrac 分析得到的距离矩阵可用于多种分析方法，比如上述提到的 PCoA，NMDS 分析，除此之外，还可以进行相似性分析（Anosim，Analysis of similarities），多样品相似度树分析（如非加权组平均法 UPGMA 构建进化树）等。

7、测序数据中，有的样品测序量多，有的样品测序量少，这种情况怎么办？

由于存在人工混样操作误差，以及建库时 PCR 扩增效率和偏好性不同，一般每个样本测到的 reads 数目会存在一定的差别。从 alpha 稀释曲线来判断样本测序数据量是否足够，从下图来看，虽然每个样本的测序量不一样，但基本上在 reads 数目 10000 左右时，所有样本的 chao1 指数达到平台期，说明测序量已经足够。所以，尽管每个样本的测序数据量不完全一样，只要测序量能覆盖到样本中所有的 OTU 即可。此外，由于测序数据量的不同，不同比较组之间的物种差异分析采用的是归一化的相对丰度进行计算的，因此也不受测序量不同的影响。



图：alpha-chao1 指数稀释曲线

8、基于高通量测序的微生物多样性检测技术优势以及原理是什么？

常规的宏基因组学研究方法包括基因克隆文库、变性梯度凝胶电泳 DGGE/TGGE 等，但这些方法的通病是信息量太小，不能充分反映复杂的环境微生物多样性和分布。基因克隆文库构建和检测的工作量大，且自然界中 99%的微生物在实验室都没有办法纯化培养，从培养基上挑取克隆菌株，摇菌转化测序，效率低下。DGGE 法曾经广泛应用于检测微生物群落结构的多态性，但是需要标准菌株，且受到凝胶电泳特性的局限，无法检测到稀有菌群的种类，因此其重复性和分辨率都不甚理想。

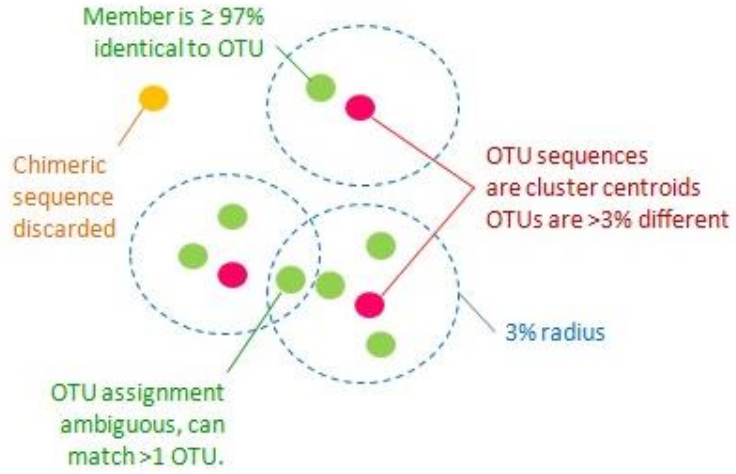
二代高通量测序无需构建质粒克隆文库，这避免了文库构建过程中利用宿主菌对样品进行克隆而引起的系统偏差，可以直接对环境样品中的基因组片段进行测序，简化了基本操作，提高了测序效率，它能够对一个群落中微生物的多样性作更加深入和全面的描述，且具有通量高，重复性好，精确度高的优点，因而在微生物生态学研究中逐渐占据了优势。

技术原理：16S rRNA 普遍存在于原核细胞中，且含量较高、拷贝数较多（占细菌 RNA 总量的 80%以上），便于获取模板，功能同源性高，遗传信息量适中，适于作为细菌多样性分析的标准。16S rRNA 序列长约 1542bp，其分子大小适中，且同时包含高保守区和高特

异性。通常我们利用保守区域设计引物来扩增 rRNA 基因的单个或多个可变区，然后测序分析微生物多样性。

9、OTU 是什么？作用是什么？

OTU（operational taxonomy unit）操作分类单元，在微生态研究中，为了便于分析，人为给定某一分类单元（品系，种，属等）设置的同一标志。简单的说，OTU 类似于种，是根据高变区序列比对得出的，一般将相似度大于 97%的序列聚为一类，称为同一个 OTU。



图：OTU 聚类原理

通常情况下，将来自同一环境的所有样品序列进行合并，将序列之间相似性大于 97%（相当于物种分类种水平之间的序列差异）归类为一个 OTU，理论上来说，每个 OTU 对应于一个不同的 16S rDNA 序列，也就是每个 OTU 对应于一个不同的物种。通过 OTU 聚类分析，不仅可以简化数据结构还可以得到样品中的微生物多样性以及不同微生物的丰度。

10、ALPHA 多样性和 BETA 多样性在微生物检测中的意义？

对群落多样性、生态多样性的研究主要应用多样性指数来进行比较，常用的指数包括 alpha 多样性指数和 beta 多样性指数。

alpha 多样性是指一个特定环境或生态系统内的多样性，主要关注局域均匀生境下的物种数目，因此又称为生境内的多样性。alpha 多样性常用的指数有四种：Observed species，Chao1，Shannon，Simpson。其中 Observed species 指数是指样本中实际包含的 OTU 数目，chao1 指数是指估算样本中 OTU 的数目，这两个指数均反映样本中 OTU（物种）数目的高低，这两个指数越高表明样本中的物种丰富度越高。Shanon 指数是用来描述 OTU 出现的紊乱和不确定性，不确定性越高，多样性指数越高，Simpson 指数是指从样品中随机取两条序列，这两条序列属于不同的 OTU 的概率，如果样本中只有一个 OTU，Simpson 指数为 0，多样性最低。Shannon 和 Simpson 指数不仅反映了样本中物种的数量（即丰富度）又反映了样本中各物种的丰度分布情况（即均匀度）。总之，样品中 OTU 数目越多，OTU 丰度分布越均匀，多样性指数越高。

beta 多样性指沿环境梯度不同生境群落之间物种组成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，主要是衡量群落之间的差别。beta 多样性的意义在于，①它可以反映生境变化的程度或指示生境被物种分割的程度；②beta 多样性的高低可以用来比较不同生境的多样性。

总的来说，alpha 多样性主要关注某一个群落中的物种多样性，而 beta 多样性主要关注不同群落之间的物种多样性差别。

11、16S 测序与其他项目关联原理及注意要点

16S rDNA 测序主要用于研究某一特定环境中微生物结构组成，及其在不同处理条件下微生物种类及丰度差异。虽然根据 16S rDNA 物种分析结果，可以进行功能预测（PICRUST 软件），但是 16S 功能预测的结果只能做到 KEGG pathway 水平，无法通过 map 图看到每种代谢内部基因调控的酶的变化，另外受到数据库（Greengenes 长时间未更新）影响，功能预测的结果只能作为参考。如果想对基因层面对微生物代谢功能更深入的研究，就需要结合其他研究手段（宏基因组，宏转录组等）研究复杂微生物群落变化，基因功能层面差异及挖掘潜在的新基因。

16S 测序相比宏基因组测序来说，价格便宜。利用 16S 技术对大量样本进行测序分析，首先发现并阐明不同条件下样本的物种组成结构以及差异情况，然后挑选个别有代表性的样本（注意：样本挑选的时候，挑选同一条件下样本组成相似，且与其他条件下样本组成差别较大的样本，为了差异分析具有统计学意义，建议每组至少挑取 3 个样本。），进行宏基因组测序，从而更加深入的阐明群落的功能。

宏转录组，兴起于宏基因组之后，从整体水平上研究某一特定环境，特定时期群体生命全部基因组转录情况以及转录调控规律，它以生态环境中的全部 RNA 为研究对象。与宏基因组学相比，宏转录组学能从转录水平研究复杂微生物群落变化，能更好地挖掘潜在的新基因。联合 16S，宏基因组，宏转录组三者分析，可以在群落水平，DNA 和 mRNA 水平上同时进行研究，实现多重角度互相验证，为全面解析微生物环境群落中的重要物种和基因信息提供了一种有效的手段。

除了与宏基因组和宏转录组联合分析外，16S 和代谢组学联合分析也是近年来研究的热点。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的影响是很困难的。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。我们采用 WGCNA 算法将代谢物进行功能模块聚类，然后根据聚类模块的特征值与 OTU 丰度表进行 Spearman 相关性系数分析（注意：该联合分析需要代谢物样本和 16S 样本完全一致，建议先做 16S 测序，去除偏离点，保留的样本再做代谢组测定分析），进而发现微生物（某几类门）与代谢物模块间的关系（如下图）。

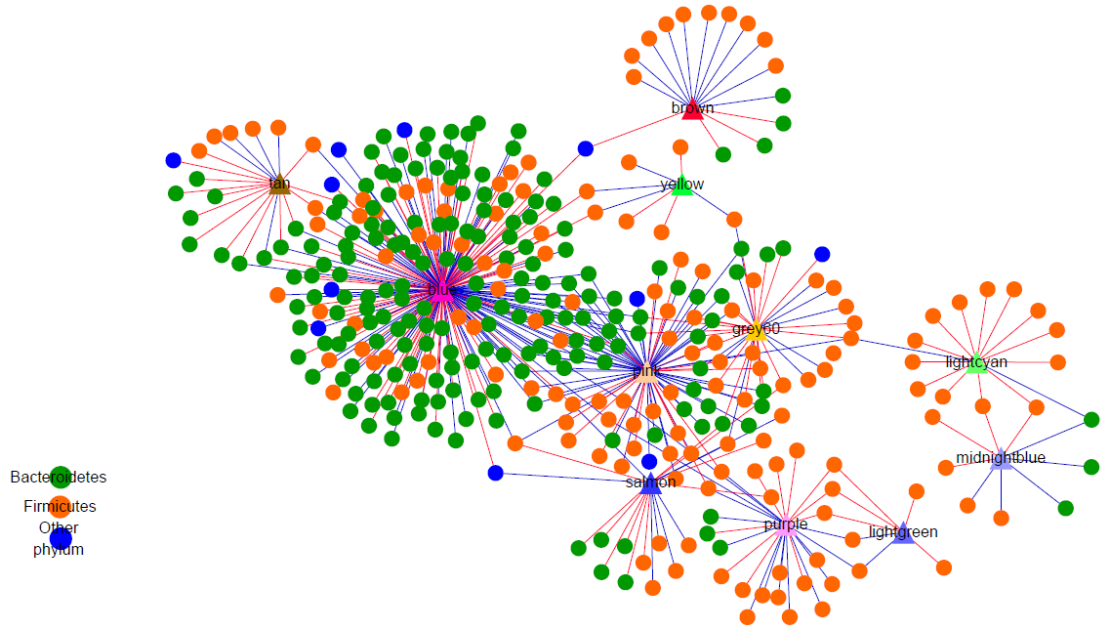


图:16S 和代谢组学联合分析结果

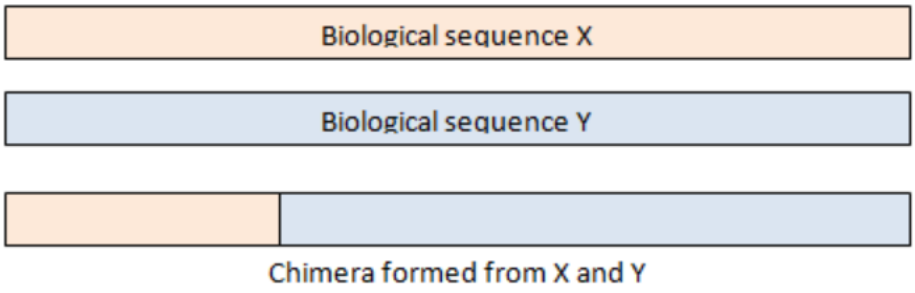
12、物种差异分析原理

根据是否有生物学重复，采用不同的统计检验方法来进行物种差异统计。对于无生物学重复样品之间的差异比较，采用 Fisher's exact test 方法，该算法用来判断两个变量之间是否存在非随机相关性的一种统计学检验方法。如果存在生物学重复，采用 Kruskal-Wallis test 方法，该算法适合三组或更多数据的非参数检验（参数统计是总体分布类型已知，用样本值来对总体参数进行估计或者做出假设检验的统计方法，非参数统计不考虑总体分布类型，对总体参数不做比较，比较的是总体分布的位置是否相同的统计方法），用来检验总体函数分布的一致性原假设和其替代假设。

根据统计检验得到的 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，如果 P 值很小，说明原假设情况发生的概率很小，如果出现了，根据小概率原理，有理由拒绝原假设，P 值越小，拒绝原假设的理由越充分）来判断该物种在不同组之间是否存在显著性差异，一般以 P<0.05（一个事件如果发生的概率很小的话，那么它在一次试验中是几乎不可能发生的，但在多次重复试验中几乎是必然发生的，数学上称为小概率原理，统计学上一般认为等于或小于 0.05 或 0.01 的概率为小概率）认为存在显著性差异。通过差异检验分析，可找到影响不同分组的重要菌属，对后续的深入研究或结果验证有指导意义。

13、什么是嵌合体，嵌合体是否需要去除

嵌合体，遗传学上用以形容不同遗传性状的嵌合或混杂表现的个体。嵌合体序列是由来自两条或者多条模板链的序列组成，见下图示例图：



图：嵌合体示例图

PCR 反应中，在延伸阶段，由于不完全延伸，就会导致嵌合体序列的出现（以上图为例）。在扩增序列 X 的过程中，序列延伸阶段，只产生了部分 X 序列延伸阶段就结束了，在下一轮 PCR 的反应中，这部分序列 Y 的引物接着延伸，扩增就会形成 X 和 Y 的嵌合体序列。在 PCR 过程中，大概有 1%的几率会出现嵌合体序列，嵌合体在正常生物体中是不存在的，所以在 16S 扩增子测序的分析中，需要去除嵌合体序列。采用 Vsearch 软件对嵌合体序列进行过滤。

14、什么是 SINGLETON OTU，分析过程中是否需要去除？

利用 Vsearch 对 unique 序列在大于 97%相似性条件下进行聚类，获得 OTU 代表序列及其对应的丰度。其中 OTU 丰度（这里的 OTU 丰度是指 OTU 在所有样本中表达量加和）为 1 的 OTU 即是 Singleton OTU，该 OTU 的序列与样本中其他序列之间的相似性低（即相似度小于 97%）。Singleton OTU 极有可能由于测序错误导致的，假阳性较高，建议在分析时去除。

除了 Singleton OTU 外，极低频率出现的 OTU 一般都认为是背景噪音，对分析结果无帮助，因此在实际分析中我们默认将丰度值低于所有样本测序总量 0.001%（十万分之一）的 OTU 进行过滤。但如果老师想研究一些极低丰度的物种，请提前告知分析人员，帮你保留低丰度的 OTU(由于 Singleton OTU 产生与测序错误有极大关系 ,这部分结果假阳性较高，建议将这部分 OTU 去除过滤）。

15、16S 多样性测序引物该怎么选择？

传统方法中最常用的引物是 27F 和 1492R，几乎能扩增出完整的 16SrRNA 基因全长序列，但是由于二代高通量测序读长的限制，该引物明显不适用于高通量测序平台。考虑到读长的限制，只能对 16SrDNA 某一段或某两段可变区进行测序。一般而言，环境微生物组学常用的，也是认可度比较高的测序区域，V3-V4，V4-V5，或者单个 V4 区。

常用的 V3-V4 区通用引物是 338F/806R，具体序列如下：

338F-ACTCCTACGGGAGGCAGCAG

806R-GGACTACHVGGGTWTCTAAT

宏基因组测序热门文献解读

微生物与女性健康

16S 和宏基因组揭示女性生殖道微生物群落多样性

Item：Cervicovaginal Bacteria Are a Major Modulator of Host Inflammatory Responses in the Female Genital Tract

Journal：Immunity，2015

IF：21.561

研究背景：

乳酸杆菌被认为对维持生殖器官健康至关重要，但是目前对生殖道微生物群落如何影响宿主免疫功能和调节疾病易感性知之甚少。通过分析无炎症年轻妇女的生殖道菌群，发现健康女性生殖道菌群相对简单，主要由 Lactobacillus 组成，这些微生物会产生细菌素、乳酸和过氧化氢，抑制一些致病菌和真菌的生长。细菌性阴道病（BV）主要是阴道菌群紊乱导致的一种炎症性疾病。其中菌群的优势地位被 Gardnerella 和 Mobiluncus 取代。另外，BV 会使 C. trachomatis、N. gonorrhoeae、T. vaginalis 和 HIV 发生的概率增加 1.5~2 倍。由此推断，宫颈阴道微生物群体影响生殖器炎症的发生，从而可能影响女性的生殖健康，包括她获得艾滋病毒的风险。

研究方法：

样本选取：18-23 岁的南非女性 146 名进行粘膜取样，每两周进行 HIV RNA 检测，每三个月进行盆腔盥洗及外周血检验；最终有 94 个研究对象具有完整的阴道拭样、灌洗样和细胞表型数据；

实验方法：16S 测序分析：Illumina Miseq， 300bp，16S rRNA 基因 V4 区；
宏基因组测序分析：12 个样品，Illumina Miseq，PE250bp。

研究结果：

1. 分析女性生殖道菌群构成

对 94 个健康女性的生殖道进行无菌棉签擦拭，将棉签擦拭样品进行 16S rRNA 基因 V4 区进行测序分析。分析发现，Lactobacillus 的在这些受试者中丰度较低，仅占 37%（图 1A），根据优势菌群 PCoA 分析，可以分为不同阴道型（Cervicotypes，CTs）（图 1B）。CT1 以 Lactobacillus crispatus 为主，CT2 以 L. iners 为主，CT3 以 Gardnerella 为主，其他的归为 CT4 型（含 Prevotella）。与已有报道的发达国家白人妇女 Lactobacillus 占 90%和黑人妇女占 62%不一致。通过研究发现，本研究中的 Lactobacillus 优势菌中，77%为 CT2 型（L. iners），L. iners 在生物学上与其他乳酸杆菌不同，因为它具有独特的可以与不同群落成员共同生存的适应性，并且具有更大的致病潜力。

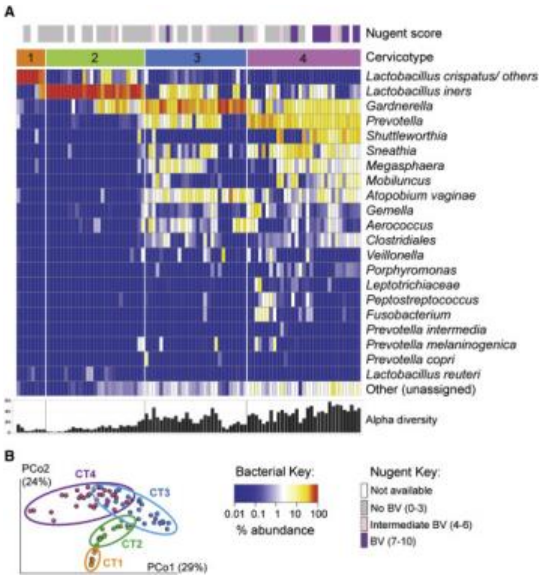
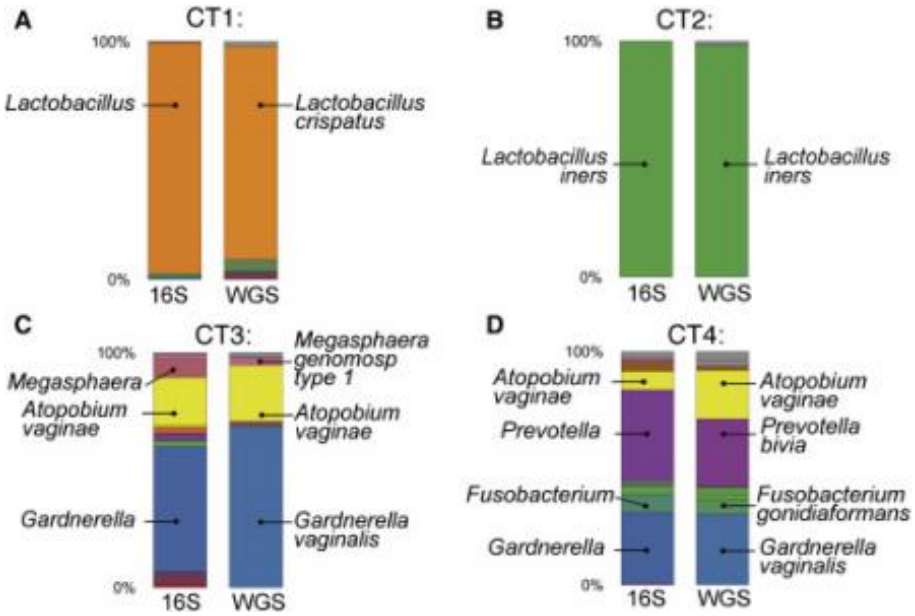


图 1：16S rRNA 序列分析微生物群落结构

2. 比较 16S 测序与宏基因组测序数据结果

通过对其中的 12 个受试者样本进行宏基因组测序，并将其结果与 16S 测序结果进行比较，证明宏基因组测序数据对微生物识别度更高。宏基因组测序分析发现 CT1 型的两名受试者均具有 Lactobacillus crispatus 优势物种，因此我们将分析视角放大至 CT1 型的每个受试者上，通过 Oligotyping 分类方法检测 16S rRNA 基因内微量核苷酸变异模式，发现几乎所有的受试者都具有相同的序列，进一步验证了 CT1 主要由 L. crispatus 组成。 因此，我们在 16S rRNA 测序和宏基因组测序获得的分类学分类之间表现出密切的一致性，并提高了细菌群落的物种水平分辨率。



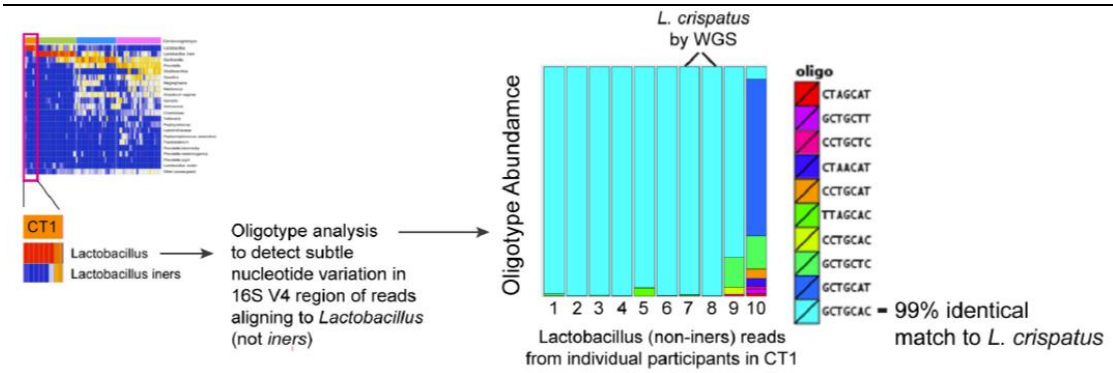


图 2：16S 和宏基因组测序数据比较

3. 微生物群落结构与 STI、激素避孕药使用及性行为无关

4. 通过对阴道灌洗液中 17 种细胞因子的检测发现，生殖道免疫活性与 STI（FGT 女性生殖系统疾病）无关

生殖器免疫激活已被认为是 FGT 疾病的重要危险因素，如产科并发症和艾滋病毒的获取等。因此，通过测量已完成至少一次粘膜采样的所有参与者的宫颈阴道灌洗液（CVL）液中 17 种可溶性细胞因子的浓度来评估生殖道免疫活性。通过询问有最高水平的细胞因子炎症的女性患有活动性 STI，发现患有沙眼衣原体的妇女比没有 STI 的妇女有更高的炎症，但是与以前的研究一致，没有观察到淋病奈瑟氏球菌的趋势。在发炎程度最高四分之一的 35 名妇女中，其中 75%没有可检测到的 STI（图 3B）。细胞因子水平也与性频率，避孕套使用，激素避孕用法或 STI 症状无关。因此，大多数妇女的生殖器官炎症升高的原因仍然无法解释。

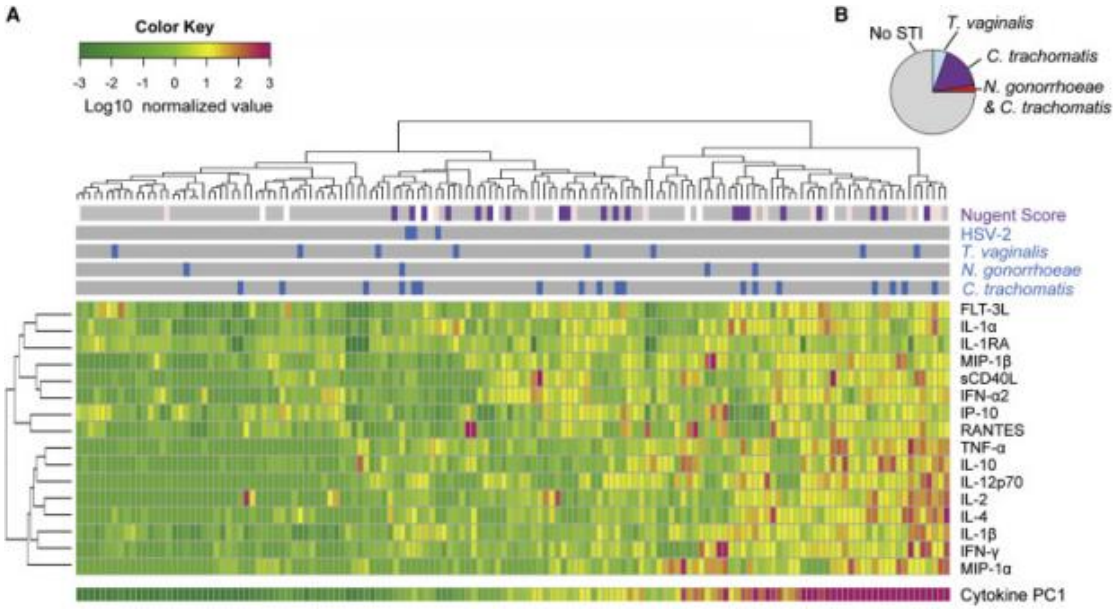


图 3：正常女性生殖道炎症与 STI 无关

5. 细菌多样性与促炎因子呈正相关

为了评估微生物群落是否可能与非 STI 相关的生殖器炎症有关，研究比较了促炎因子与不同 CTs 之间的关系，由于 L. crispatus 的有益作用，假设乳酸乳杆菌属优势 CT1 具有最低的炎症，发现高度多样化的细菌群落的 CT4 以及 CT3 与多种促炎细胞因子的存在密切相关，这表明特定的生殖细菌可诱导强大的局部免疫应答。

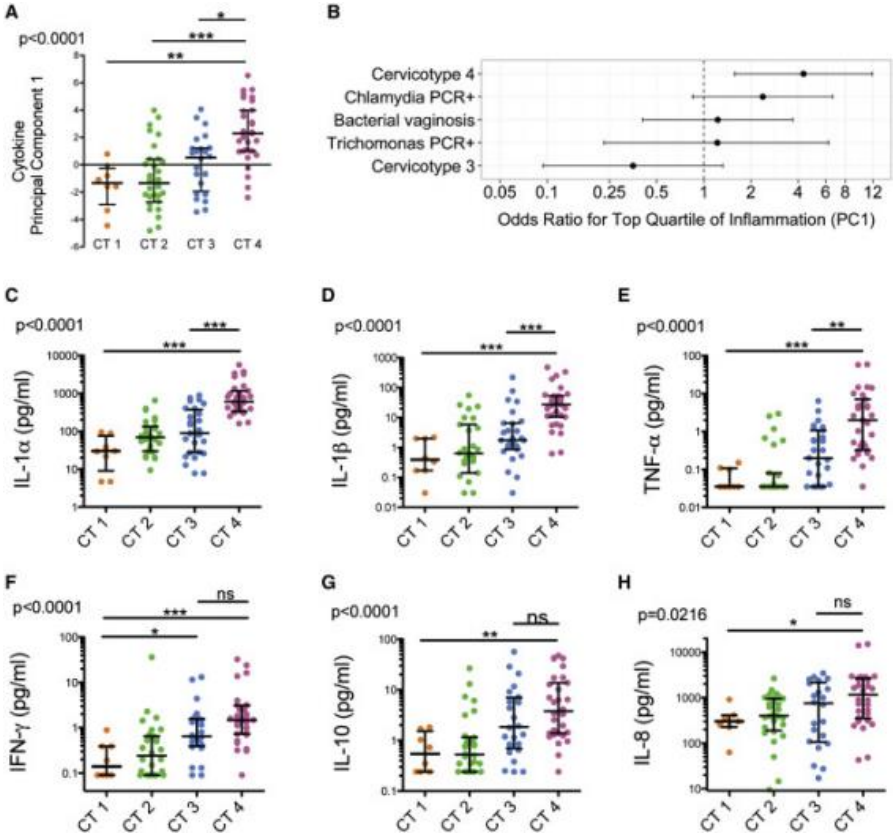


图 4 细菌多样性与促炎因子的关系

6. 时间梯度上阴道菌群与促炎因子之间的关系

阴道菌群与验证之间具有显著性关系性，研究对不同时间这一关系的稳定性进行研究。研究发现，阴道微生物菌群处于动态变化中，微生物变化差异不大时，炎症因子差异不大；当微生物发生显著变化时，炎症因子也发生显著变化，这说明生殖道菌群与炎症变化显著相关。

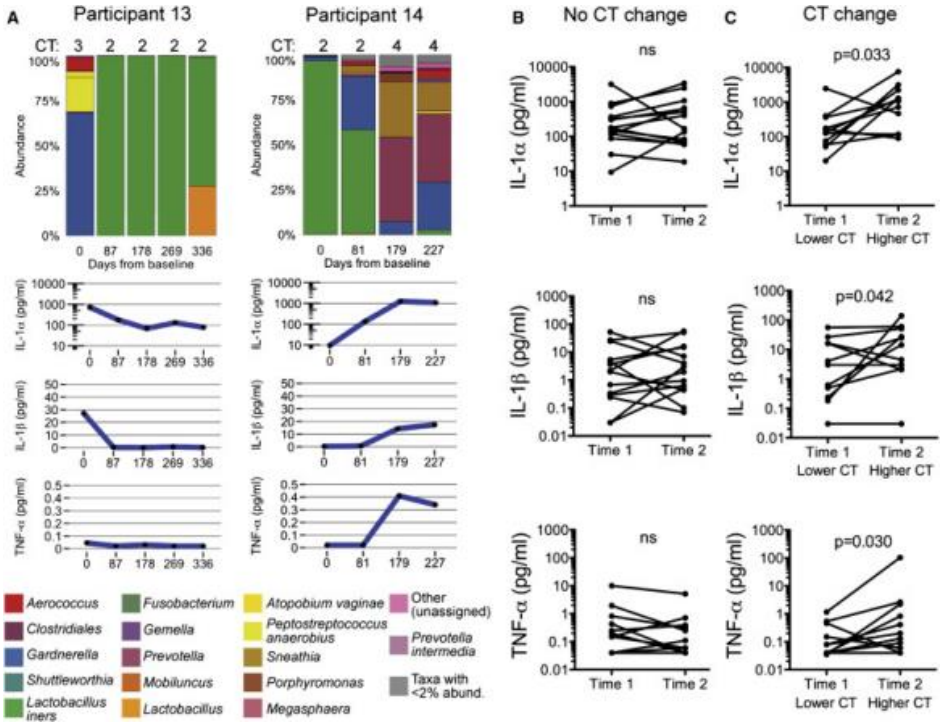


图 5 时间梯度上阴道菌群与促炎因子之间的关系

7. 对引起促炎因子分泌的微生物进行研究发现：Fusobacterium、 Aerococcus、 Sneathia、 Gemella、 Mobiluncus 和 Prevotella 与促炎因子的分泌具有显著相关性。

A

	Prevalence (n=94)	Scaled β coefficient	p value
Fusobacterium	6.4%	5.34	**0.0015**
Aerococcus	26.6%	5.29	**0.0012**
Sneathia	40.4%	3.96	**0.0168**
Gemella	16.0%	3.94	**0.0198**
Mobiluncus	19.1%	3.63	**0.0275**
Prevotella	55.3%	3.62	**0.0348**
Other (unassigned)	47.9%	2.11	0.2034
Shuttleworthia	21.3%	2.05	0.2141
Clostridiales	18.1%	0.76	0.6400
Mycoplasma	4.3%	-0.51	0.7570
Lactobacillus iners	62.8%	0.20	0.9068
Leptotrichiaceae	2.1%	-0.18	0.9120

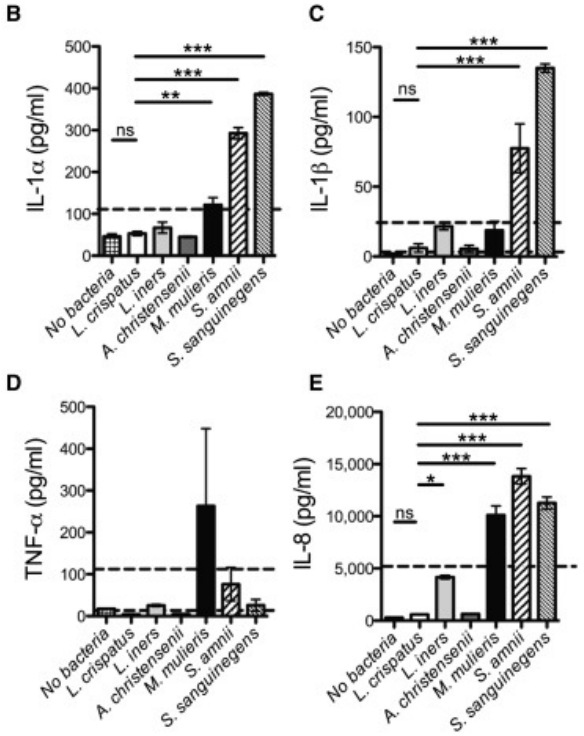


图 6 体外促炎因子分泌相关微生物

4. 体内检测实验表明，阴道菌群多样性高的妇女阴道菌群产物会激活抗原呈递细胞
8. 妇女阴道促炎因子的升高会增加 HIV 目标细胞的水平，增加了患 HIV 的风险。

研究亮点：

本研究实验设计严谨，从结合 16S 测序与宏基因组测序结果以及其他生理生化方面分析揭示了阴道菌群与炎症之间的相关性，研究思路清晰，层层递进，研究结果环环相扣，从各个方面说明了阴道微生物与阴道炎症之间的显著性关性。另外，文章对 16S rRNA 测序结果与宏基因组测序结果进行比较，结果表明宏基因组测序在微生物物种识别上的显著优势。

口腔微生物与儿童龋齿

通过口腔微生物群的时空变化预测幼儿龋齿的发生

Item：Prediction of early childhood caries via spatial-temporal variations of oral microbiota

Journal：Cell host & microbe，2015

IF：12.552

研究背景：

早期儿童龋齿（ECC，Early childhood caries），是儿童最常见的口腔疾病，影响着全世界近一半的儿童。ECC 导致牙釉质和牙本质的持续脱矿质，并且感染可以从受影响的牙齿扩散到周围的软组织，甚至导致肿胀和炎症。龋齿一旦发生，将对牙齿产生不可逆转的损害，而且新长出的牙齿也有高患病的风险，甚至导致整个生命当中的牙齿脱落。因此，对 ECC 的预防性干预具有特殊的临床意义。然而，对 ECC 的预测特别是对新疾病发病来说是很困难的。

研究方法：

1）2011 年 6 月至 2013 年 6 月期间，从同一幼儿园招收了 50 名学龄前儿童（平均 4 岁）作为研究对象，这些儿童有着相似的生活环境。在整个研究期间，这些儿童保持着相似的饮食习惯，口腔卫生及牙齿检查等。

2）2 年分四个不同时期共采集 284 份样品（分别来自唾液和牙菌斑两个生态位），根据纵向临床研究状况，将所有 50 名受试者分为三种类型（H2H, H2C, C2C，其中 H 表示健康，C 表示龋齿，H2H 为未患龋齿-17 位儿童（研究始末均健康，共 94 份样品），H2C 轻微龋齿-21 位儿童（研究起始未患龋齿，研究过程中患轻微龋齿，共 120 份样品），C2C 严重龋齿-12 为儿童（研究起始到结束均为严重龋齿，共 70 份样品）。根据 dmfs 判断 ECC 严重程度，在给定的时间内，dmfs 为零的微生物被指定为“健康”（“H”）；否则为“龋齿”（“C”），由“低龋齿”（1%dmfs <6）和“严重龋齿”（dmfs R6）组成。）的宿主型。

3）作为后续模型验证，额外的 40 个儿童（46~50 个月，其中包含 20 个健康的儿童，20 个患有严重的龋齿）用于分析唾液微生物。

4）分别对牙垢牙菌斑和唾液中提取 DNA，Roche 454(GS FLX Titanium System)进行 16S rDNA（V1-V3 高变区）测序。

5）采用 MOTHUR，QIIME 及 R 对 16S 序列进行分析，并使用 PICRUST 进行功能预测。口腔微生物群体年龄计算，使用 R（3.1.1）randomForest（随机森林）包计算健康群体（H2H 组）所有种水平的物种相对丰度分布适合相应的年龄（月）。使用 randomForest 包中的“rfcv”功能，确定了前 25 个重要 age-discriminatory 的物种。然后对随机森林模型进行训练，以确定来自“ConfidentH（相对健康）”和“Caries（严重龋齿）”组的样本疾病状况，并使用 ROC 曲线面积进行评估。这种模型称为“MiC”（龋齿的微生物指标）。为了构建和优化 MiC，首先测试了不同物种分类水平，口腔龛位和宿主成长对 MiC AUC 的影响。

研究结果：

1、基于健康儿童口腔微生物丰度进行随机森林分析

通过对 17 名健康儿童的牙菌斑和唾液样本的物种相对丰度进行随机森林回归，确定了牙菌斑（图 1A）和唾液（图 1B）中 age-discriminatory 的细菌分类群。值得注意的是，不论

是来自牙菌斑还是来自唾液的这 25 个物种主要是 Streptococcus，Neisseria，Fusobacterium，Capnocytophaga，Prevotella 以及 Porphyromonas，且这些物种在健康儿童口腔中的较为丰富。根据来自健康儿童口腔微生物的随机森林模型来定义 C2C 和 H2C 组中微生物学年龄。与 H2H 组比较，C2C 和 H2C 组儿童牙菌斑和唾液的微生物更成熟。在 C2C 组中，微生物菌群的成熟度明显高于病情发展期间，尤其是当龋齿特别严重（dmfs>10）的时候。在 H2C 组中，对于整体微生物结构，ECC 起始对细菌多样性的影响比年龄更重要，说明在口腔微生物在健康-龋齿的转变过程中的改变与宿主年龄增长引起的菌群正常变化是不同的。因此，在建立 ECC 发生模型时，需要考虑年龄因素。此外，H2C 组患病儿童的微生物成熟度略低于 H2H 组。总之，ECC 患者中口腔微生物群体的正常发育受到扰乱。

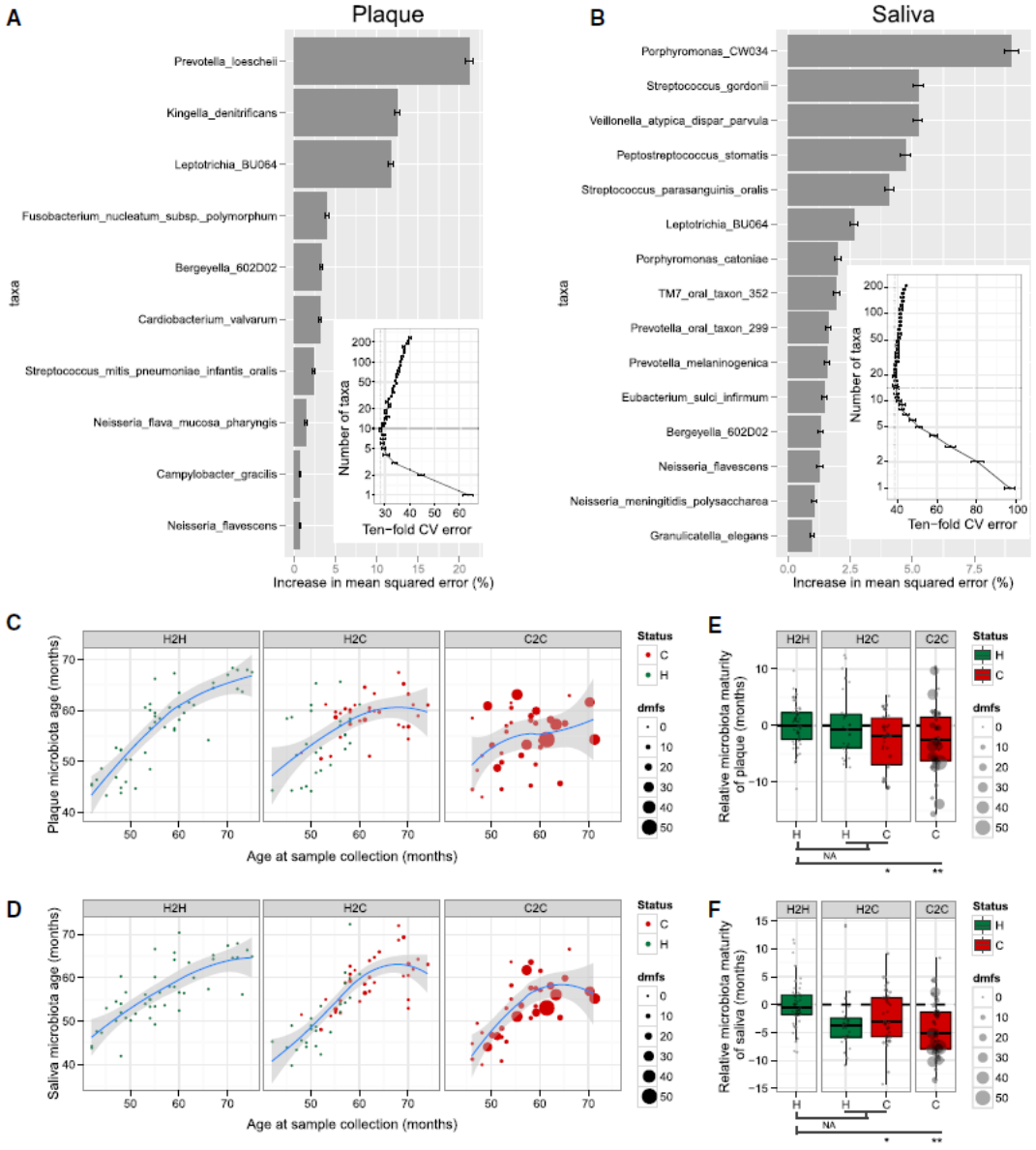


图 1：健康儿童口腔微生物群落的细菌分类标记

2、幼儿龋齿发病开始和发病过程中微生物区系的变化具有明显的不同。

牙菌斑和唾液微生物 alpha 多样性（Shannon 指数）在龋齿发作（H2C 组）期间明显发生改变，而且与龋齿严重性发展进程的 dmfs 没有相关性。在 C2C 组，牙菌斑微生物 beta 多样性与龋齿严重性发展进程没有相关性。在牙龈炎研究中，牙龈炎退化和进展期间追踪临床症状和微生物群的时间变化，Shannon 指数随着 MGI 的变化而显著变化。

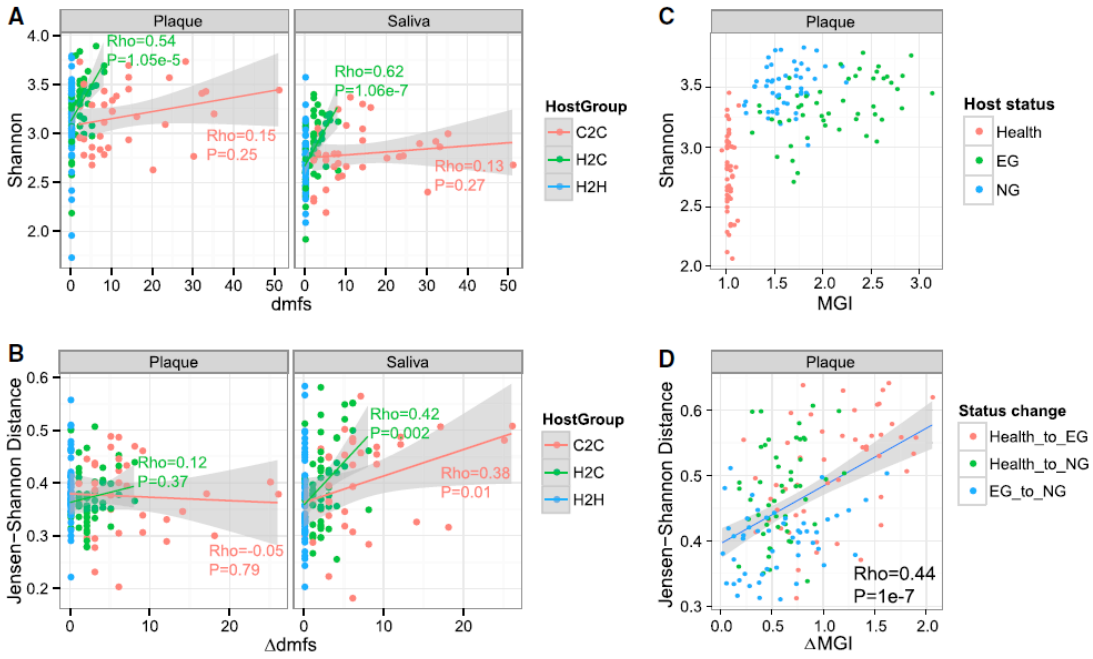


图 2：细菌多样性与 ECC 和牙龈炎等临床症状的相关性

3、构建了一个预测模型（MiC）、龋齿的微生物指标，能够对健康儿童龋齿的发生的做出诊断和预测。

基于随机森林模型在牙菌斑，唾液以及两者共同的物种特征，按照 ROC 曲线的面积进行评估。从牙菌斑和唾液中，收集了与 ECC 状态相关的 20 个关键微生物（具体物种见图 3C）。此外，Prevotella 在 H2H 组中与年龄不存在相关性，表明它们可以检测 ECC 发生并且不受年龄的影响。

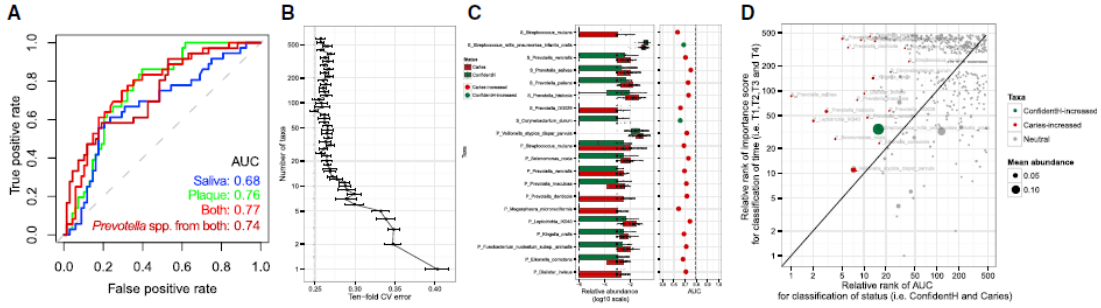


图 3：基于口腔微生物进行疾病分类

4、预测模型评估

尽管在明显的临床症状之前，患者没被诊断为龋齿，但可以通过该模型来预测相对健康的微生物群受试者是否会发展为龋齿（即，估计未来患病的风险）。将 MiC 模型应用于 42 个相对健康的样品（21 例牙菌斑样品，21 例唾液样品），在牙菌斑中，17 例被正确预测为龋齿，与随后的采样时间点表现的临床症状一致，而只有 3 例样本被归为健康状态。而且相比唾液样本而言，对牙菌斑样本的预测结果更准确。

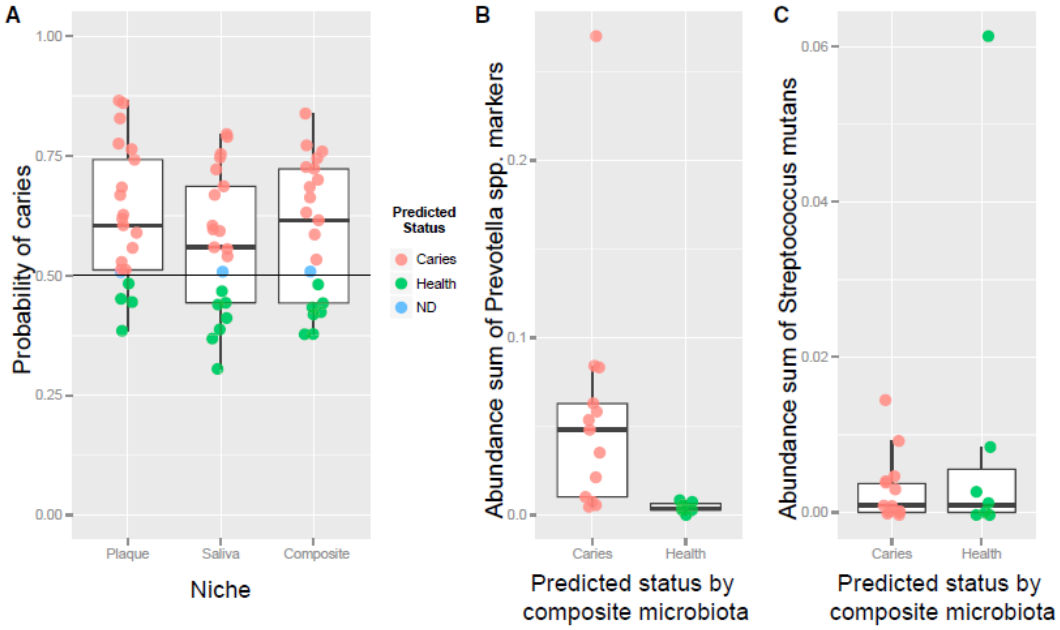


图 4：基于口腔微生物对相对健康儿童进行分类

研究亮点：

1、对 50 个 4 岁左右儿童口腔微生物跟踪研究了两年（获得了时间变化结果）

2、ECC 扰乱年龄相关的微生物群落

3、微生物群的变化先于 ECC 的临床症状表现

4、建立预测龋齿微生物指标，可预测 ECC 的发生

肠道微生物与儿童营养不良

儿童营养不良与肠道菌群紊乱的关系

Item：Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children

Journal：Science，2016

IF：33.61

研究背景：

儿童营养不良是一种及其严重的儿童健康问题，并不能总是通过改善营养来补救，而后期极易造成发育障碍、神经系统发育异常、免疫系统紊乱等问题，且通过治疗干预无法有效缓解疾病。已有研究表明人类肠道微生物可以有效地移植到无菌小鼠中以重述其相关表型。利用这个模型，发现健康儿童的微生物群体减轻了营养不良儿童的微生物所造成的对生长的有害影响。多个研究已经证实一些具体的有益微生物可能被利用来解决营养不良综合征。本文通过构建 0-3 岁儿童肠道菌群发育模型，证明营养不良是造成儿童肠道细菌紊乱的主要原因。

研究方法：

样本选取：

实验一：从马拉维南部地区的 5 个农村地区收集 317 对双胞胎和 3 对三胞胎（年龄在 0-36 个月）的粪便样品，选择其中 27 对双胞胎和 2 对三胞胎的 220 个粪便样品进行 16S 测序。

实验二：收集 19 个婴儿（其中 9 个婴儿 6 个月大，另外 10 个婴儿 18 个月大）的粪便，利用粪菌移植技术，将其分别移植到 5 周大雄性无菌小鼠体内，每个粪便移植给 5 个小鼠。移植前 3 天，所有的小鼠统一饲喂无菌饮食。移植成功后，连续观察 4-5 周，收集小鼠粪便，进行 16S 测序。

实验三：为了探究这些生长速度随年龄变化而变化的菌群能否治疗婴儿的生长畸形，将接受健康儿童肠道菌群移植和营养不良儿童肠道菌群移植的小鼠放在一起饲养，统一饲喂食物，连续观察 3 个周，每周称 2 次体重，收集一次粪便进行 16S 测序。

实验四：将带有营养不良儿童肠道菌群的小鼠单独饲养，并向肠道中添加包括长双歧杆菌在内的 5 种微生物，连续观察 21 天，并定期检测粪便菌群组成。

实验方法：Illumina MiSeq，16S rRNA V4 区，PE250

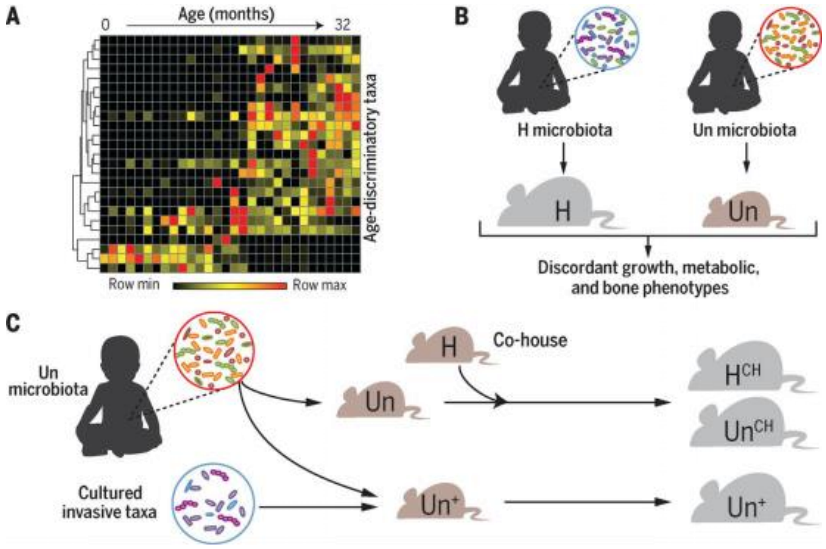


图 1 未发育成熟的肠道菌群与营养不良有关的实验设计思路

研究结果：

1. RF 模型构建

利用 0-36 个月大的 27 对健康双胞胎的粪便样品，构建肠道菌群的 RF 模型，如图 2 所示，发现肠道菌群成熟度与婴儿年龄呈正相关。分析结果表明 MAZ 得分（年龄相关微生物 Z 得分）或许可以用来预测儿童身体健康状态，但还需进一步地研究和分析。

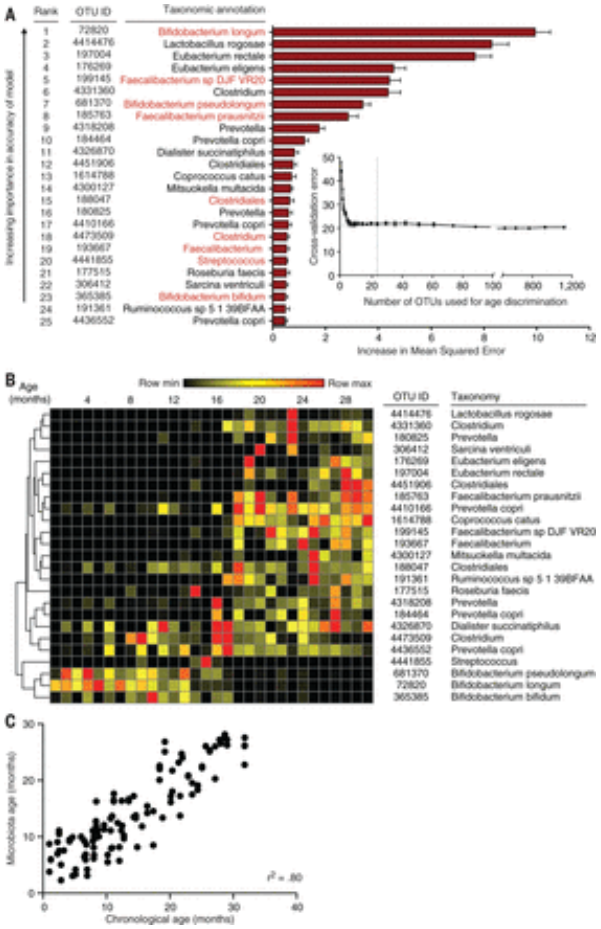


图 2 肠道微生物群的 RF 模型

2. 粪菌移植实验

粪菌移植实验设计如图 3A 所示，对接受粪菌移植后的小鼠粪便进行 16S 测序，结果显示接受健康儿童粪菌移植的小鼠明显比接受营养不良儿童粪菌移植的小鼠要胖，然而所有小鼠在食物吸收方面却没有显著性差异（图 3B-C）。接受 6 个月大婴儿粪菌移植的小鼠的生长率要比接受 18 个月大儿童粪菌移植的小鼠快，无论是移植健康状态还是营养不良状态下的粪菌。13 个 OTUs 的相对丰度与体重增加显著相关，7 个 OTUs 的相对丰度则与瘦体重增加显著相关。此外，研究还发现长双歧杆菌和柔嫩梭菌群的生长速率有很大差异，其中，长双歧杆菌在婴儿发育到 5 个月大时就已表现出最高的相对丰度，而柔嫩梭菌群最高相对丰度则出现在第 19 个月。并且接受 6 个月大婴儿粪菌移植的小鼠有更高的骨密度、BV/TV 值，更低的骨小梁间隙。

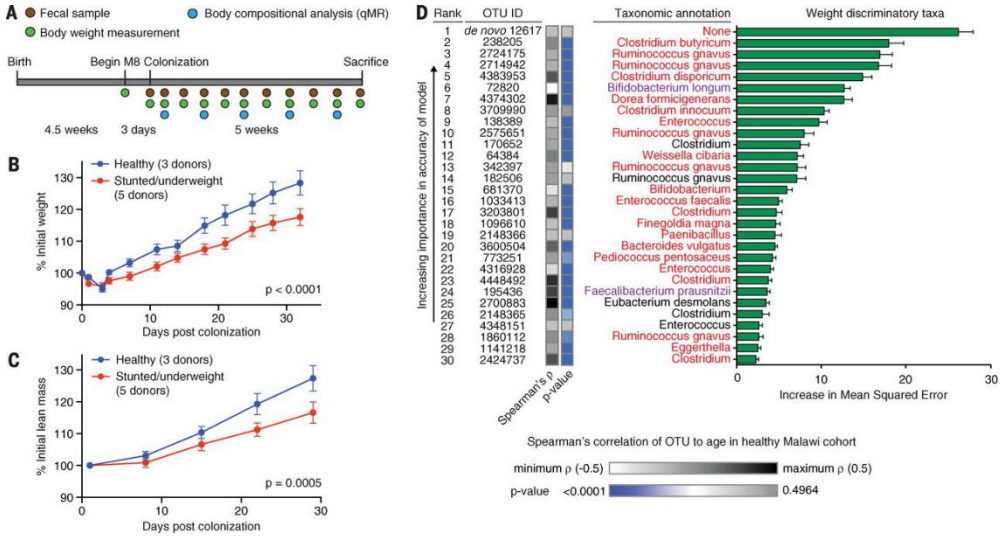


图 3 粪菌移植实验揭示肠道菌群与年龄之间的关系

3. 修复受损的生长表型

为了确定年龄和生长相关细菌种类是否能够修复与来自发育迟缓和体重不足的供体的微生物群的异常。研究从两个 6 个月大的供体中选出了微生物群的生物体受体，这些供体在 19 微生物群的初始筛选中表现出最不一致的生长表型。研究发现将分别带有健康儿童菌群和不健康儿童菌群的小鼠共室饲养后，二者瘦体重的增加量显著高于带有不健康儿童菌群小鼠单独饲养，但是与带有健康儿童菌群小鼠的瘦体重增加量之间则无显著差异（图 4B）。该结果表明不同个体之间，健康菌群可以慢慢融入不健康菌群（图 4C）。其中，以长双歧杆菌和活泼瘤胃球菌为代表的 9 种微生物可从健康菌群转移至不健康菌群中，而只有肠球菌和粘液真杆菌 2 种微生物能够从不健康菌群转移至健康菌群中。

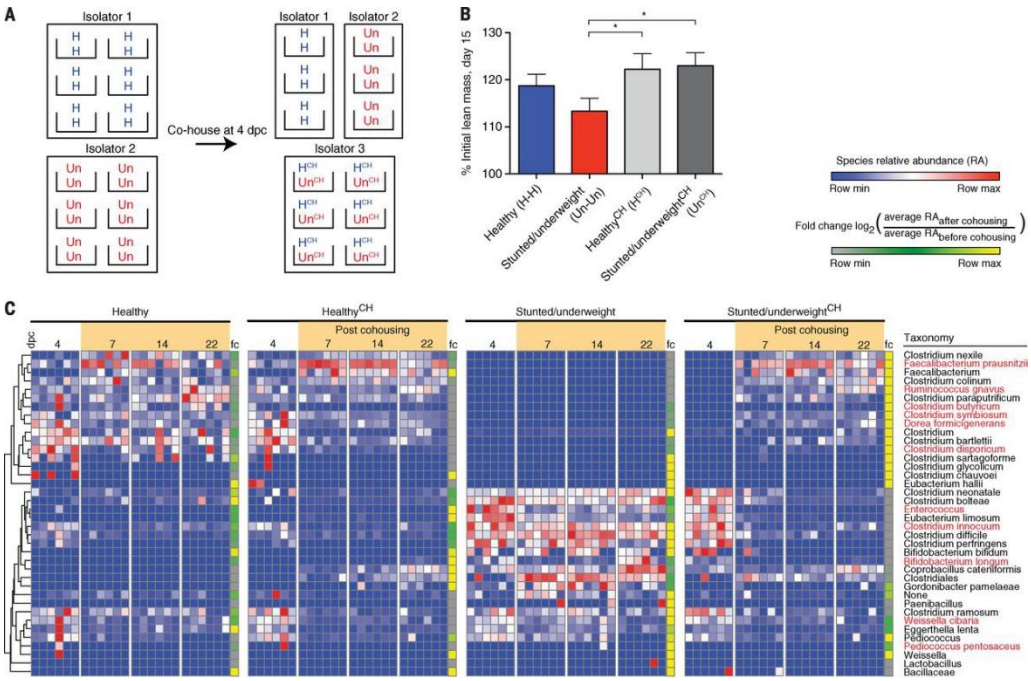


图 4 接受不同来源粪菌移植的小鼠共室饲养及 16S 测序结果

5. 为进一步了解哪些 OTUs 能够调控生长表型，选择 2 组均带有不健康儿童肠道菌群的小鼠，其中一组小鼠额外添加长双歧杆菌、柔嫩梭菌群、系结梭菌、共生梭菌和 Dorea formicigenerans 五种细菌，另外一组作为对照，不作任何处理（图 5A），连续观察 3 周。结果发现，添加长双歧杆菌等 5 种菌群的一组小鼠的体重和瘦体重增加量显著高于对照（图 5B）。物种组成柱状图显示只有长双歧杆菌和共生梭菌成功定植在受体小鼠中，且小鼠肠道原有菌群并没发生显著性改变（图 5C）。此外，研究还发现，这两种菌能够影响小鼠的代谢型（图 5D）。对照组小鼠盲肠中酰肉碱很多、肝脏中酰肉碱很少，而实验组小鼠则正好相反，表明上述两种细菌能够调控宿主代谢，避免氨基酸氧化。

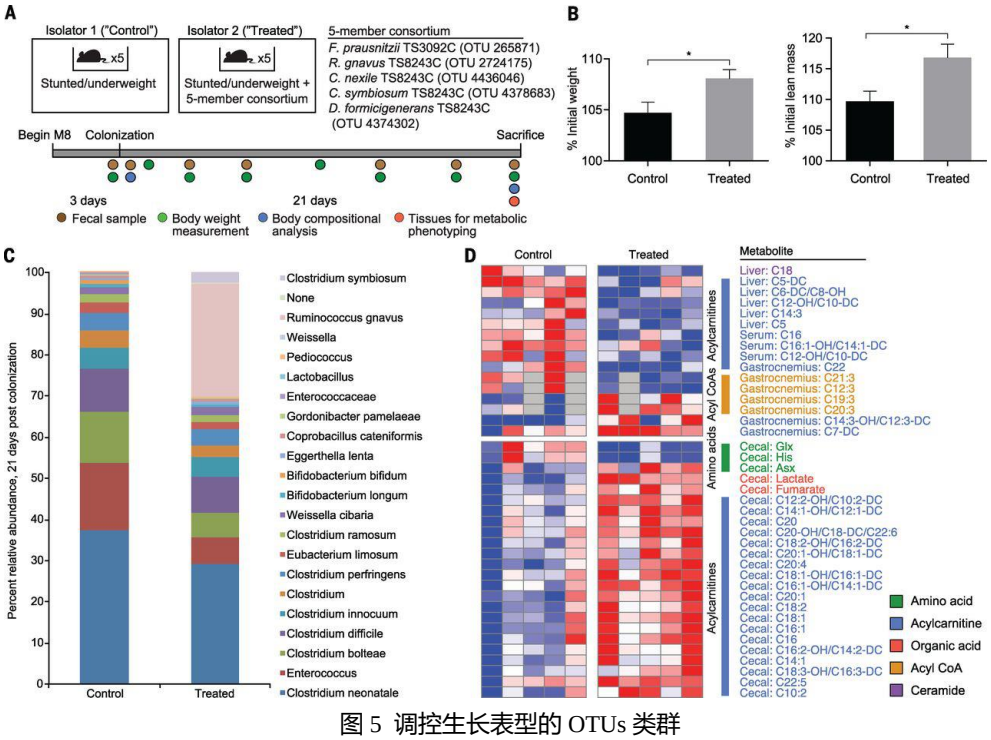


图 5 调控生长表型的 OTUs 类群

本研究不仅证明人体肠道菌群的成熟度与机体营养状态有关，而且证明二者之间存在因果关系，即由于饮食不当等原因，机体肠道菌群逐渐紊乱，而后导致营养不良。在停止纯母乳喂养后提供的某些适当的补充食品可能具有促进生长-歧视性肠道分类群的定植能力，其比例适合年龄，并且可以在系统试验中对其临床价值进行测试。

研究亮点：

1.实验设计十分完善，层层递进，关系紧密。

2.将 16S 测序与人体指标巧妙结合解释常见的科学问题。

3.证实了肠道菌群紊乱与营养不良之间关系。

肠道微生物与疾病

肠道微生物研究发现黑色素瘤治疗会引起病人患有结肠炎

Item ： Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis

Journal：NATURE COMMUNICATIONS，2016

IF：11.329

研究背景：

肠道微生物群的组成影响炎症性疾病的发展。然而，将炎性疾病与微生物群的特定微生物成员结合是具有挑战性的，因为临床可检测的炎症及其治疗可改变微生物群的组成。Ipilimumab 是一种阻断共同抑制分子细胞毒性 T 淋巴细胞相关抗原-4（CTLA-4）的单克隆抗体，是一种能够有效治疗转移性黑色素瘤的免疫调节疗法。在接受 ipilimumab 治疗的患者中，部分病人会引发结肠炎，另一部分病人则不会。

研究方法：

1）本研究招募患有转移性黑色素瘤 34 名成年人，这些受试者之前没有结肠炎和肠切除病史， 并且在两个月以内没有接受抗生素治疗。如果结肠炎发病后收集的第一个粪便样本，如果病人没有接受 ipilimumab 治疗或者结肠状态位置的患者被排除。除了三名患者外，ipilimumab 治疗以 3mgkg^{-1} 的剂量每 3 周 4 次。作为临床试验的一部分，10#和 15#患者在 ipilimumab 治疗之前和期间接受了 vemurafenib。13#患者是盲症临床试验的一部分，以 3 或 10mgkg^{-1} 的剂量接受 ipilimumab。对于 30 例患者，在首次给予 ipilimumab 之前，收集每个参与者粪便样品，对于发生肠胃道炎症的 2 例患者和 2 例保持无炎症的患者，在 CTLA-4 阻断开始后，但在结肠炎发作前进行样本收集。

2）患者根据以下情况判定结肠炎评分：无腹泻（0 分），1 级腹泻（1 分），2 级腹泻（2 分），2 级腹泻和 2 级结肠炎（均为 3 级），3 级腹泻或 3 级结肠炎（1 例，4 分）。其中 C-F 病人（无结肠炎 n=24），PtC 病人（发展为结肠炎患者 n=10）。

3）采用 Illumina MiSeq 平台分别做 16S 和宏基因组测序

研究结果：

1、结肠炎患者（PtC）和非结膜炎患者（C-F）微生物组成不同

16S 测序分析结果发现，C-F 和 PtC 患者具有类似复杂的肠道微生物群落，共有 578 个 OTU，其中 239 个为两组共有的 OTU。且这些共有的 OTU 分别占 C-F 和 PtC 患者组的总 OTU 丰度的 79%和 83%。虽然两组共有的 OTU 大部分来自 Firmicutes（厚壁菌门），但 C-F 患者 Bacteroidaceae（拟杆菌科）占有较高的比例。

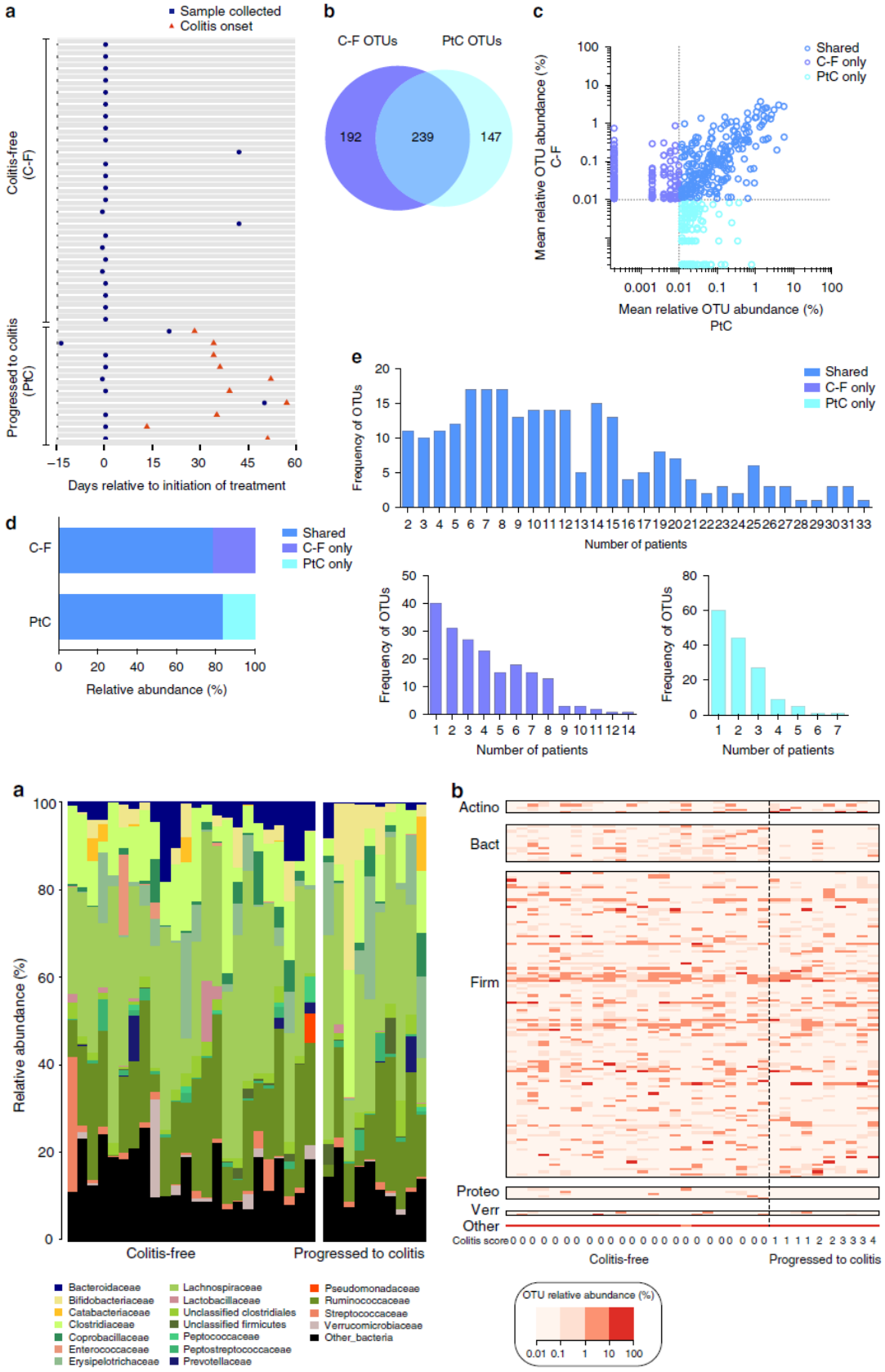


图 1：C-F 和 PtC OTU 和肠道微生物组成情况

2、结肠炎患者中拟杆菌门丰度升高

进一步探讨肠道微生物群落特异性细菌与 CTLA-4 阻断性结肠炎发生之间的关系，按照炎症程度对病人进行了分类，并在不同分类学上的分类物种的相对丰度进行了 Spearman 相关性检验。结果发现，在 C-F 患者样品中，门分类水平的分类群更为普遍。虽然肠道微生物群的组成可以在个体的寿命周期内变化，但是患者的年龄和物种丰度没有关联。

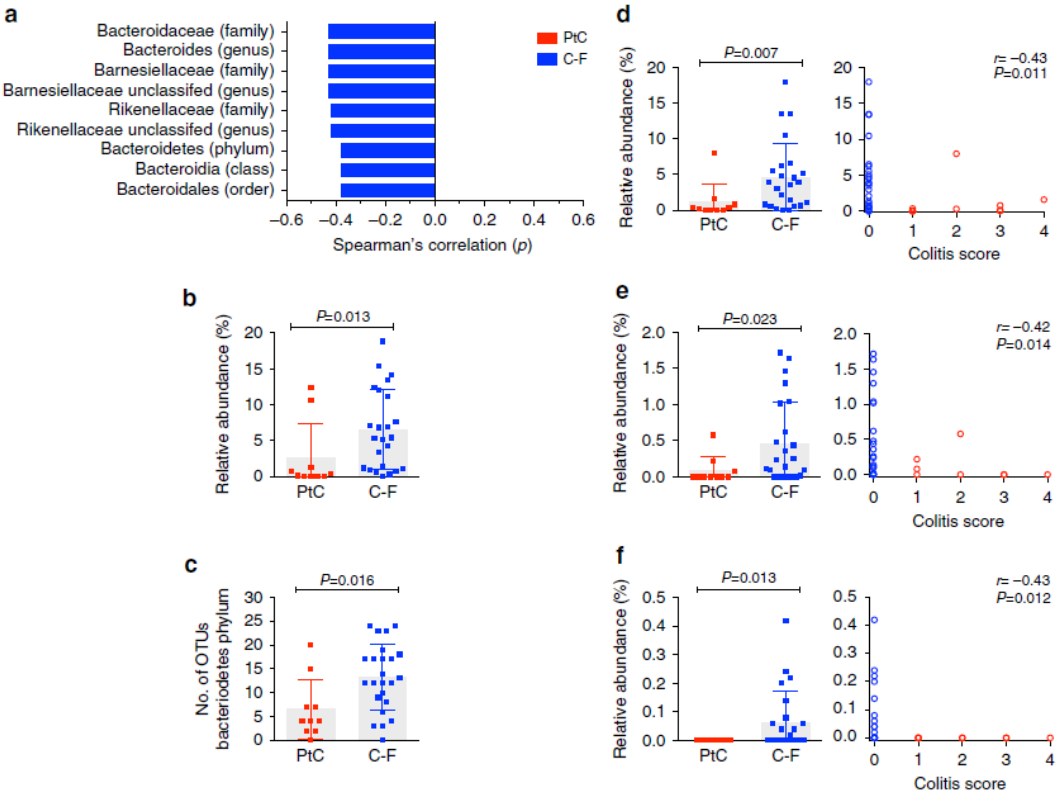


图 2：结肠炎患者中拟杆菌门升高

3、与保护相关的特定微生物功能模块

为了评估可能在免疫介导性结肠炎发展中发挥作用的遗传途径，对 10 例 PtC 和 12 例 C-F 患者粪便样品进行宏基因组测序，结果发现患者肠道微生物菌落功能在 C-F 和 PtC 患者之间大致相似。使用 LEfSe 分析发现，亚精胺/腐胺多胺转运系统 （spermidine/putrescine polyamine transport system）以及参与维生素 B 生物合成(biosynthesis of B vitamins (riboflavin (B2), pantothenate (B5) and thiamine (B1)）等抗结肠炎相关的功能模块在 C-F 患者中更丰富。

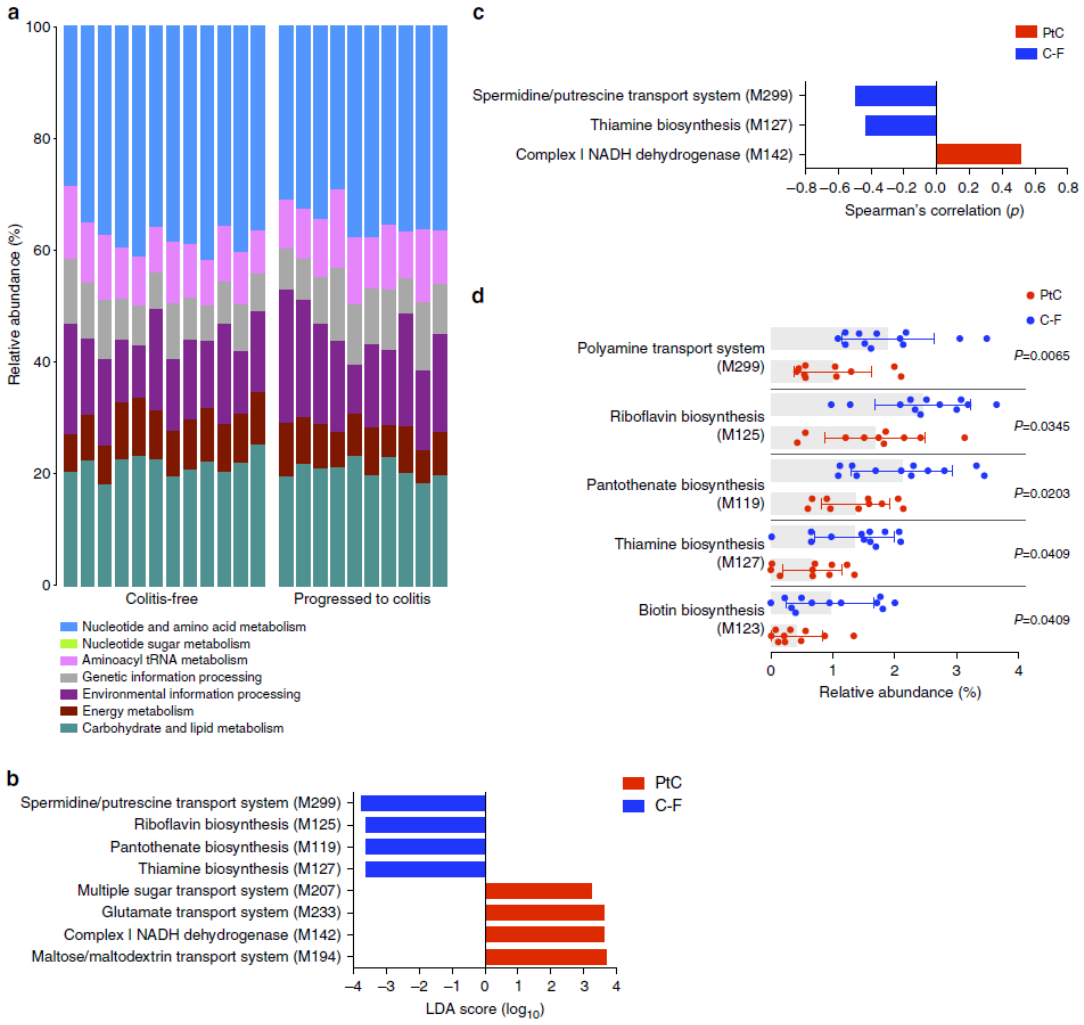


图 3：细菌模块参与与抗结肠炎相关的聚胺转运和维生素 B 合成等功能模块

4、相关通路作为 biomarker 可用于结肠炎预测

通过递归分区机器学习算法，基于 102 个功能模块形成分类树，仅使用多胺转运系统相对丰度，就成功将样本从 PtC 或 C-F 患者进行分类。多胺转运系统相对丰度为 41%的被认为属于 C-F，在 PtC 患者中，7 个样本被正确识别，3 个样本被错误分类为 C-F，预测的灵敏度 70%，特异性为 100%（下图 a）。同时使用与 C-F 患者相关的 4 个功能模块进行概率回归分析的一次性交叉验证，正确预测了 12 名 C-F 患者中的 10 例，该模型的灵敏度为 70%，特异性为 83%。总的来说，这些分析模型识别可能赋予对结肠炎抗性的细菌途径，并且可作为发生 CTLA-4 阻断性结肠炎的高风险患者的生物标志物。

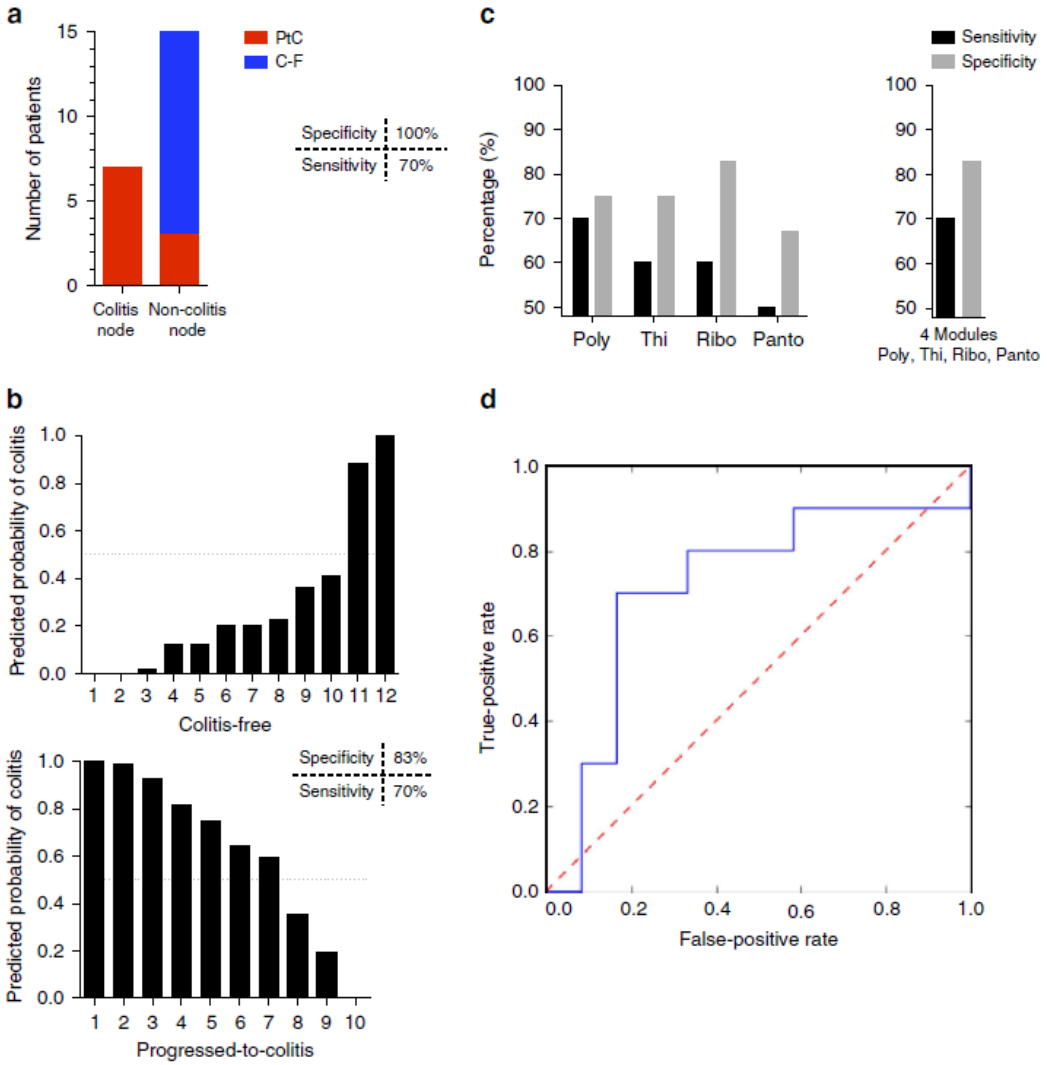


图 4：细菌模块预测结肠炎患者准确率

研究亮点：

1）同时使用 16S 和宏基因组方法对 ipilimumab 治疗过程中产生结肠炎和非结肠炎患者的粪便进行测序分析，且实验设计部分十分严谨

2）结合 16S 和宏基因组分析结果找到可预测 ipilimumab 治疗过程中患结肠炎风险的生物标志物

微生物与眼部健康

定义正常人类眼结膜微生物群落的“core microbiome”

Item：Defining the normal "core microbiome" of conjunctival microbial communities

Journal：Clinical Microbiology and Infection，2016

IF：5.768

研究背景：

眼部细菌感染是很常见的，但运用传统培养与分子生物学方法鉴定结膜微生物群落具有明显的局限性，而宏基因组研究可以弥补这些方法的缺陷。通过对结膜拭子样品中微生物的分析获得可能代表正常结膜微生物群落的核心属。正常成人人结膜中微生物群的组成和多样性用于调查多种微生物群在与眼表相关的疾病中发挥的潜在作用。

研究方法：

样本选取：选取 6 个月内没有任何系统性疾病，眼表疾病，葡萄膜炎，青光眼，视网膜疾病或眼外伤/移植的病史，或接受滴眼剂给药（抗生素，皮质类固醇和非淄体抗炎药）的正常受试者，局部麻醉后用一次性无菌干燥棉签在上、下睑结膜，阜和穹窿结膜收集样本。在收集每个样本后，将局部麻醉剂滴入一次性无菌干燥棉签中，作为空白对照，保存于-80℃。从 2014 年 10 月至 12 月，对 31 名正常成年人（16 名男性和 15 名女性）共收集 31 份 DNA 样本。

实验方法：16S 测序分析：Illumina Miseq，300bp，16S rRNA 基因 V3-V4 区；

研究结果：

1. 结膜样本的细菌多样性分析

通过将高质量的序列以 97%相似度聚类成 OTU 后，对这些序列进行物种鉴定，发现每个样本平均 452 个 OTU，通过稀释曲线分析评估微生物群落多样性。

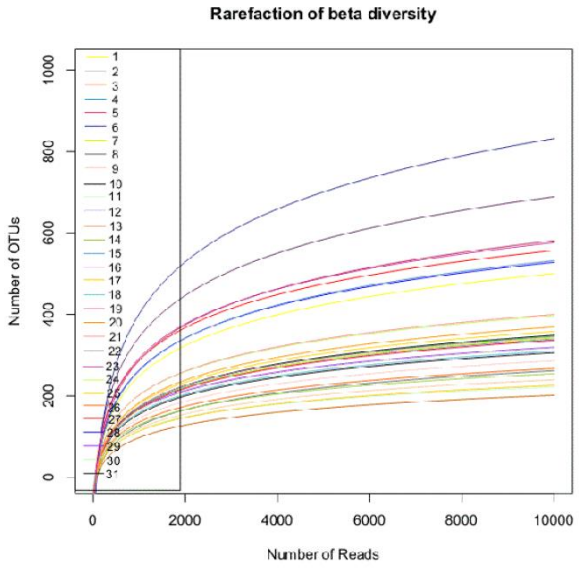


图 1 稀释曲线分析

2. 物种分类注释分析

利用 RDP 分类工具在门和属水平进行注释分类，共分为 25 个门，大多数物种属于十个重要的门（如图 2），变形菌门占比最大，其次是放线菌门、厚壁菌门、拟杆菌门、异常球菌-栖热菌门、梭菌门、蓝藻门、酸杆菌门、Candidatus Saccharibacteria 和螺旋体门，共计占所有测序读数的 99.99%（图 3）。其中三类门（Proteobacteria，Actinobacteria，Firmicutes）占 95.89%，前 5 名占 98.88%。单个变异似乎仅影响代表该微生物群落的 DNA 序列的相对丰度，但不影响其组成。两个普通门（Proteobacteria 和 Actinobacteria）在每个受试者中的相对丰度没有显著变化，只有 5 个受试者在 Proteobacteria 和 Actinobacteria 之间的 5 倍以上的 DNA 序列的相对丰度差异不大。

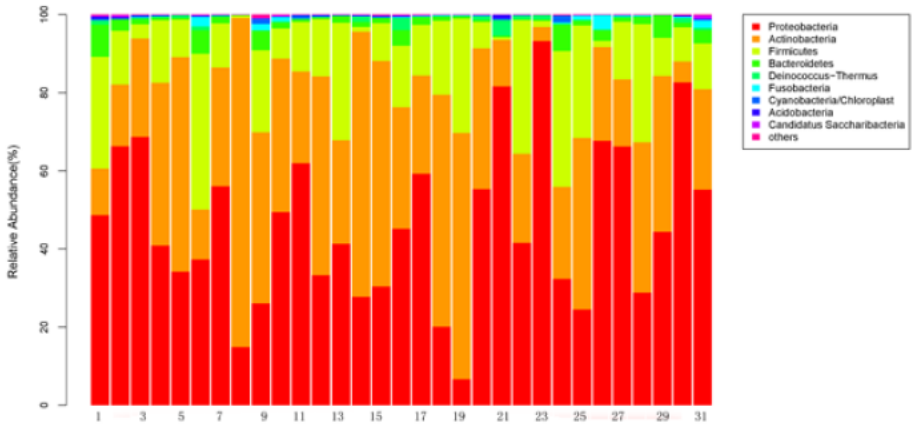


图 2 受试者样本物种组成及丰度

在属层级，共 526 个属中 24 个属在所有样本中均出现，而棒状杆菌属、假单胞菌属、葡萄球菌属、不动杆菌属、链球菌属、Millisia、Anaerococcus、Finegoldia、Simonsiella 和 Veillonella 十个属被鉴定为 core microbiome（定义相对丰度> 1%的检测到的属）。

与以前的研究相比，16S rDNA 对于有培养要求和生长特性的微生物具有明显的优势，其重点是使用革兰氏涂片和生化特性的表型检验。使用传统的微生物培养技术，<80%的结膜拭子将产生可培养的微生物。然而，已知可培养的微生物群仅代表可以定居于人的微生物的一部分。在我们以前的研究中，我们使用血板培养在 135 例正常结膜样品中发现 14 个细菌属，总阳性率为 66.7%（90/135）。革兰阳性菌占 58.5%（79/135），包括表皮葡萄球菌（31.85%，43/135），葡萄球菌（5.2%，7/135）和粪肠球菌（5.2%，7/135），而革兰阴性菌的比例为 8.1%（11/135）。随着发现原核生物 16S 核糖体 RNA 序列的差异可用于构建原核系统发育，以及随后对细菌和古菌的描述，这种不可培养技术通常用于细菌鉴定，因此“宏基因组“的环境可以进行评估，以概述微生物的多样性。还有核心微生物群和可变微生物群落的概念，表明某些微生物类型可能总是存在，但其他微生物类型是短暂的，并且可能依赖于诸如环境，生活方式和生理差异等因素。

研究亮点：

本研究运用高通量测序描述了正常成年人的结膜微生物群的组成成分和多样性，为调查微生物多样性在眼部表面疾病中所起的作用搭建了框架。

植物内生微生物

杨树根际与各部位组织微生物的组成多样性

Item：Structural variability and niche differentiation in the rhizosphere and endosphere bacterial microbiome of field-grown poplar trees

Journal：Microbiome，2017

IF：9.000

研究背景：

植物中的微生物群体是植物健康和生产力的关键决定因素之一，他们具有利用低丰度营养物，抑制植物病原体以及抵抗生物或非生物胁迫因子等多种功能。而对不同植物微环境中特别是地下和地上部分中微生物之间的结构组成的有力的了解仍然不够了解。本研究讨论了不同生态植物生态位的细菌群落的微生物群落分化和结构稳定性的假设。

研究方法：

样本选取：选择体外繁殖的克隆杨树（Populus tremula × Populus alba），收集 15 棵的根际土壤和根系组织样本，11 棵树的茎、叶样本，样本类型包括根际土壤，根，茎和叶。根际土壤被严格定义为粘附在根部的土壤颗粒。对于茎叶样品，收集了 11 棵杨树个体直接连接到中央树干的分支。为了标准化和最大化茎样品的再现性，将几个具有树皮的小茎“核”（5 -7 芯，每个 1 厘米）从基部到分支顶部的每个分支收集，以表示茎隔室。对于叶子样品，收集来自采样分枝的所有叶子以表示叶室。

实验方法：16S rRNA 967F-1391R，测序平台为 Roche 454 测序仪

研究结果：

1. 通过 16S rRNA 测序分析杨树根、茎、叶和根际土样本微生物组成，不同样本类型中 OTUs 差异 Venn 图（图 1）分析，表明 4 组不同样品既有共有 OTUs，也有特有 OTUs。但根际土中特有的 OTUs 比例更高，表明其微生物具有更高的多样性。

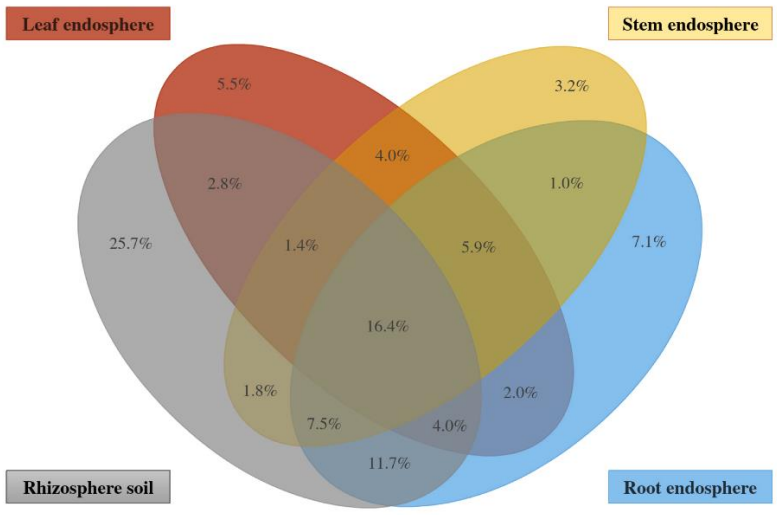


图 1 各组样本 Venn 图分析

2. α 多样性分析

从已有数据中去除 singleton OTUs，这些 OTU 为假阳性的概率很大。基于 OTU 丰度，计算 Simpson 多样性指数和 Pielou 均匀度（图 2）分析每个样本中的微生物多样性。为了控制差异，在计算多样性指数之前，将每个样本稀少到 2000 个序列。 OTU 丰富度高度依次为根际土壤，根系样本和茎样。叶片样品的 OTU 丰度指数与茎样品相当。对于多样性和均匀度估计，发现杨树各部位组织内生菌多样性远低于根际土微生物多样性（图 2,3），但内生菌样品个体间富含的菌群差异程度却要显著高于根际土（图 3）。

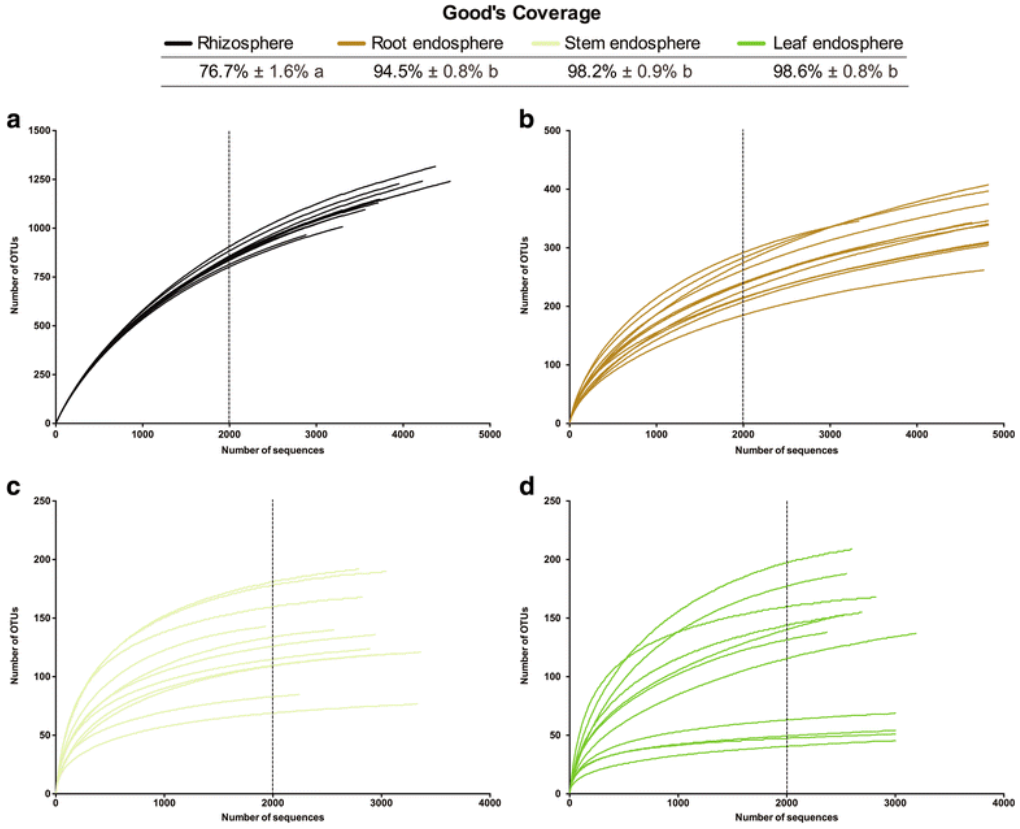


图 2 各个组织样本的稀释曲线

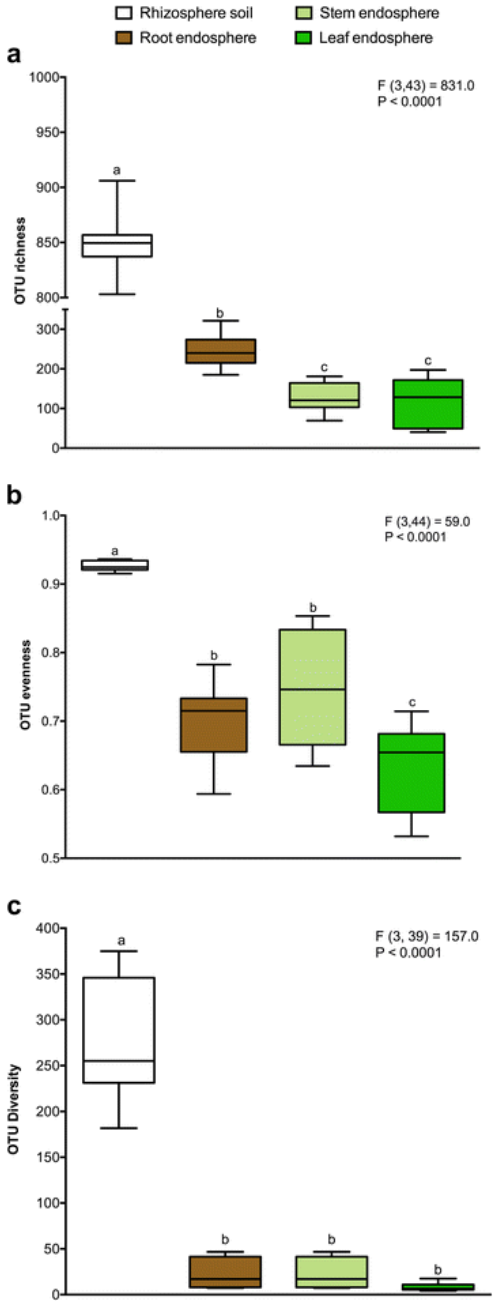


图 3 各组样品 α 多样性分析结果

2. β 多样性分析

基于 Bray-Curtis 距离，进行了 β 多样性分析。从 PCA 结果可以看出，杨树茎、叶中所含微生物显著区别于根部样本和根际土样本，而根部样本与根际土样本中的菌群组成也显著不同。尽管茎、叶内生菌组成相似（相似度在 40-60%之间），但也并非完全重合（图 4a）。聚类分析也支持了这一结果（图 4b）。

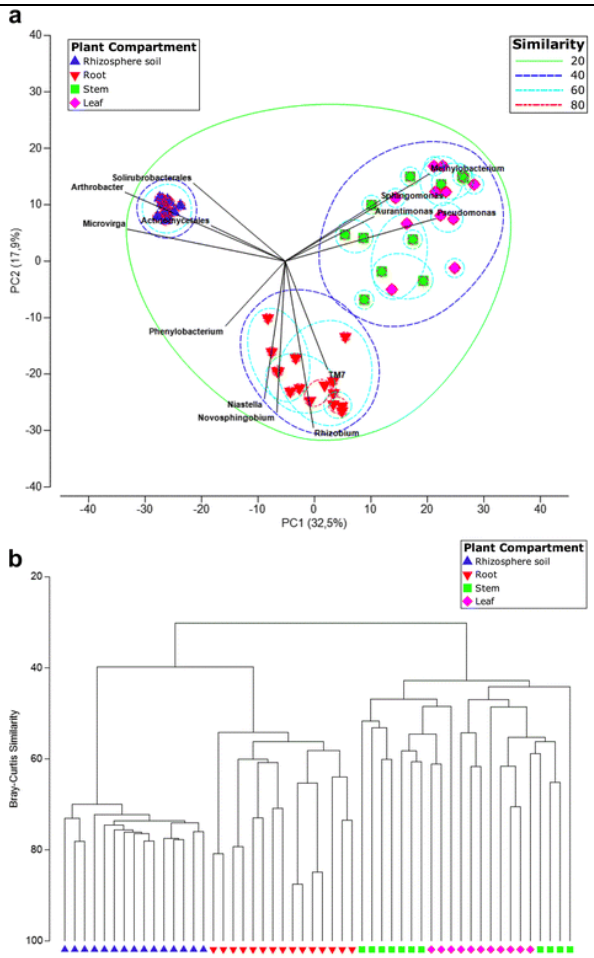


图 4　基于 OTU 水平的 PCA 分析和聚类分析

3. 物种注释分析

从门水平样品物种组成柱状图中不难看出，变形菌门、放线菌门、拟杆菌门和厚壁菌门均是各组样品中的优势菌群，但在各组中所占比例各有不同，且存在相对丰度的组间差异。根际土中放线菌门的相对丰度要显著高于其他三组，β 变形菌、拟杆菌门的相对丰度显著高于茎、叶样品。但茎叶样品中 γ 变形菌的相对丰度要显著高于根部样品和根际土样品(图 5)。

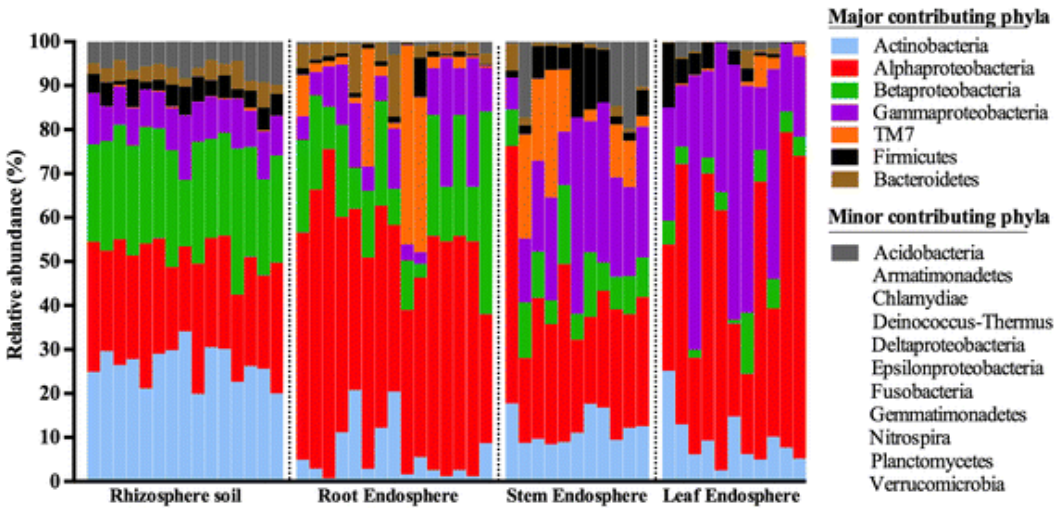


图 5 各样品门水平物种组成柱状图

4. 物种分类树

根据分析得到的 OTU 物种丰度表，将植物各个生态位中丰度最高的前 10 种微生物定义为 core microbiome，并进行分类树状图的构建（图 6）。其中，根际土中 top10 优势物种分别是放线菌目、伯克氏菌目、节细菌、Chitinophagaceae、芽孢杆菌目、微小病毒等；根部样品中丰度最高的物种依次是根瘤菌、贪嗜菌、新鞘脂菌和 Niastella 等。TM7 在茎叶组织中显著富集，而根瘤菌在茎叶组织中却显著降低。并且相比于根际土壤和根系样本，假单胞菌、甲基杆菌和鞘氨醇单胞菌在茎叶组织中富集。在叶片样本中还发现 Moraxellaceae，Aurantimonas 和 Sphingomonadales 的显著富集。

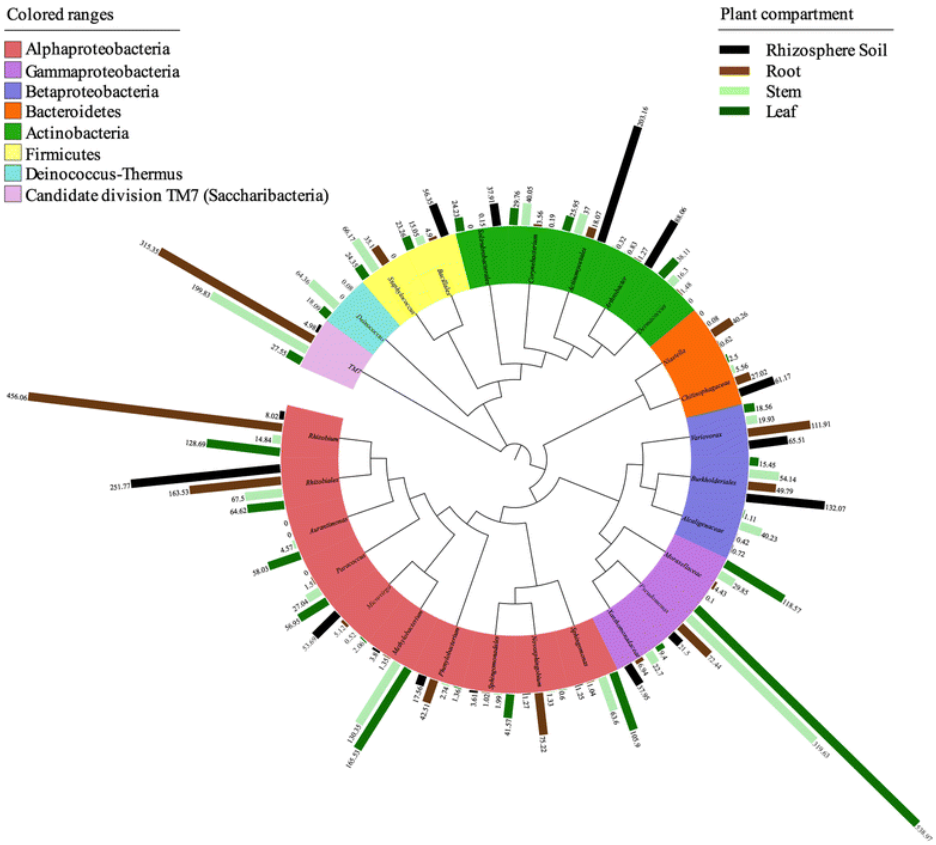


图 6 与植物生态位有关的优势菌群

5. 物种指示分析

为了进一步确定哪些 OTU 对杨树不同生态位之间的群落分化负责，我们使用物种指标分析来发现 OTU 和植物不同生态位之间的显着关联。在所有 OTU 上进行指标分析，不仅是核心 OTU，以发现核心 OTU 分析可能漏掉的影响。物种指标分析显示根际土壤中有 94 个指标 OTU，根内皮中有 18 个，茎内部有 5 个，而在叶片内皮样本中有 9 个。然而，当筛选平均相对丰度> 1%的 OTU，发现根际土壤（节细杆菌，硝螺菌，Norcardioides，Hyphomicrobiaceae，分枝杆菌和 Microvirga 的 6 个指标 OTU），根系样本中的 2 个（Novosphingobium 和 Niastella），茎样品中的 2 个（Alcaligenaceae 和 Amnibacterium）和 2 个叶样品（Sphingomonadales 和 Aurantimonas）。

表 1 指示物种分析

OTU (Genus or higher)	Plant compartment	Indicator value	*P*	Relative abundance (%)
Arthrobacter	Rhizosphere soil	0.978	0.0015**	4.403
Nitrospira	Rhizosphere soil	0.977	0.0024**	1.04
Nocardioides	Rhizosphere soil	0.97	0.0028**	1.117
Hyphomicrobiaceae	Rhizosphere soil	0.962	0.0036**	1.521
Mycobacterium	Rhizosphere soil	0.911	0.0068**	1.559
Microvirga	Rhizosphere soil	0.874	0.0119*	2.684
Novosphingobium	Root	0.981	0.0230*	3.761
Niastella	Root	0.96	0.0234*	2.013
Alcaligenaceae	Stem	0.886	0.0286*	2.205
Amnibacterium	Stem	0.83	0.0290*	1.104
Sphingomonadales	Leaf	0.937	0.0266*	2.079
Aurantimonas	Leaf	0.904	0.0270*	2.9

本研究证明野生生长杨树（P. tremula x P. alba）中根际微生物的结构变异性远低于其他部位微生物群。根际细菌群落的形成似乎是一个更加稳定和可控的过程，而内生根，茎和叶是变化较大的。此外，我们的数据不仅证实了根际土壤的微生物群落分类，而且还清楚地显示了茎和叶中的微生物群落的微调和适应。每个植物生态位代表细菌群落的独特组成，包括对特定宿主基因型效应（克隆，基因修饰基因型等）的分析的未来研究可以更好地了解细菌群落对宿主植物特定变化的可塑性或反应性。最后，我们确定了与杨树不同生态位相关的核心细菌微生物群。这可以为更多详细（隔离）研究确定的丰富的 OTU 提供依据，并进一步了解杨树的复杂宿主-微生物相互作用。

研究亮点：

（1）从独特的研究视角出法，关注植物内生菌的组成变化；

（2）比较植物各个部位不同微生物组成，证实植物内生菌多样性较低。

样本处理方式与微生物

样品处理及保存方式对微生物多样性的影响

Item：Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room?

Journal：Microbiome， 2016

IF：9

研究背景：

经常在进行微生物组多样性研究的过程中，由于条件限制（野外取样等）不可能取样后马上进行提取及测序分析，那么从取样到最后的分析，这一过程中的各种状况对最终的研究结果会有多大影响呢？本研究分析了不同样品保存及处理条件对微生物群落的影响，其相关结论对今后实验研究具有重要指导意义。

研究方法：

样本选取：来自两个不同的群体的共 103 个样品：出生 2-11 周的 8 个早产婴儿粪便样品（微生物组成相对简单）；出生 13-19 个月的 6 个幼儿粪便样品（较早产婴儿微生物组成比较复杂）。

实验设计：主要比较 5 种条件下微生物的差异：①提取：不同人员、不同试剂盒；②不同取样重量：20、50、100、200mg；③不同冻存时间；④室温保存；⑤邮寄。实验条件统计表格如下：

Table 1 Summary of study experiments

Experiment	Subject population	Faecal sample source	DNA extraction protocol	Sequencing region	Sequencing platform
Variation between extractions	Term infants	Infant 14	2	Vv4	Illumina Miseq
Effects of varied weight	Premature infants	Infants 1, 2, 3 and 4	2	V3–V5	Roche 454 GS FLX
Effects of long term freezing	Premature infants	Infants 5, 6, 7 and 8	1	V4	Illumina Miseq
Effects of room temperature storage	Premature and term infants	Infants 5, 6, 7, 8, 9, 10, 11, 12 and 13	2	V4	Illumina Miseq
Effects of mailing samples	Term infants	Infants 9, 10, 11, 12 and 13	2	V4	Illumina Miseq

实验方法：16S V3-V5 区测序采用 454 FLX 平台，对 16S V4 区测序采用 Illumina Miseq 平台。

研究结果：

1. 不同提取方式对微生物的影响

将单个样本分成十个等分的试样并分为两批处理。经不同人在不同时期采用不同批次试剂盒提取后，对不同处理的样本测序分析发现不同组之间 α 多样性并无显著差异，PCoA 分析也并未发现显著差异，因此得出结论不同提取方式对微生物多样性无显著差异。

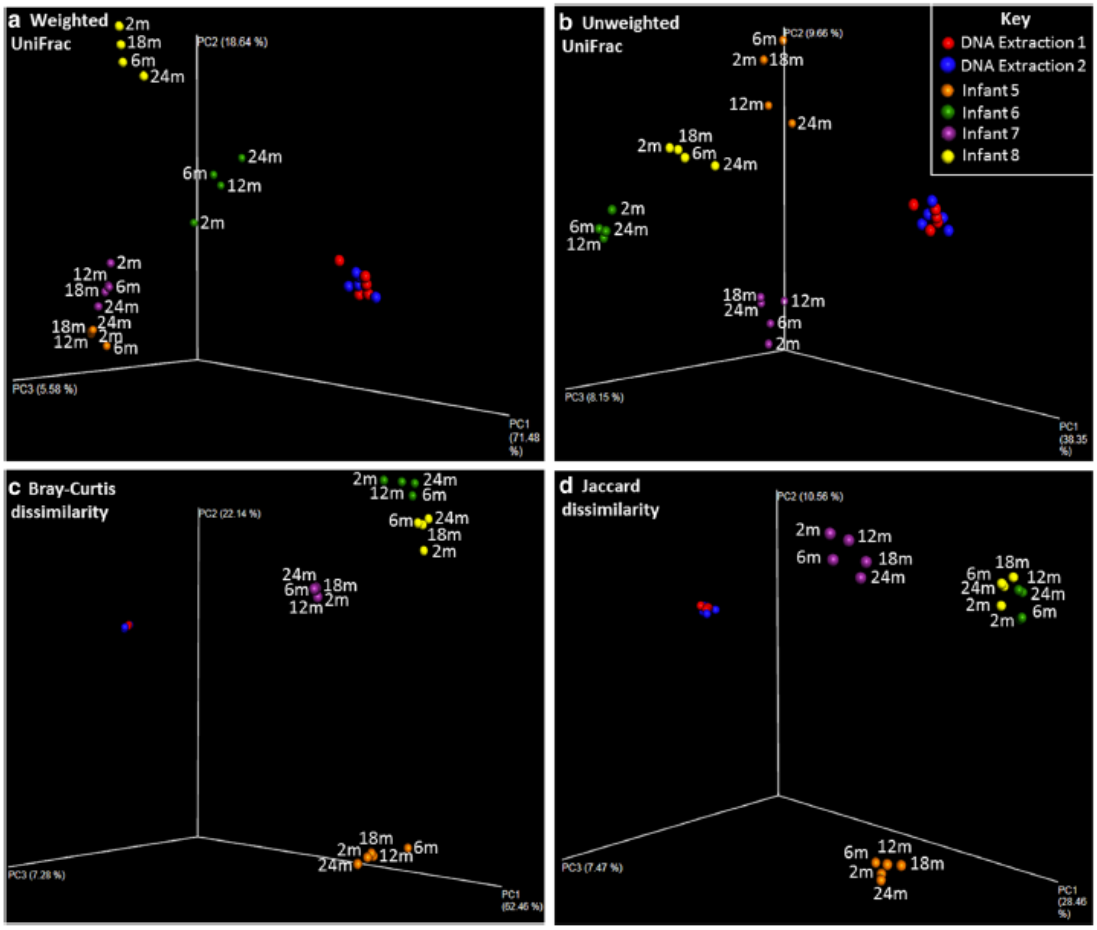


图 1　PCoA 分析

上图中 a. Weighted UniFrac 距离, b. Unweighted Unifrac 距离, c. Bray-Curtis dissimilarity 距离, d. Jaccard dissimilarity 距离；其中红点（DNA extraction 1 为在-80℃冻存 24 天）和蓝点（DNA extraction 2 为在-80℃冻存 115 天）表示同一个样品分为 10 份，经不同人采用不同批次试剂盒提取后检测微生物多样性，无显著差异；其他点表示保存不同时间后微生物多样性。

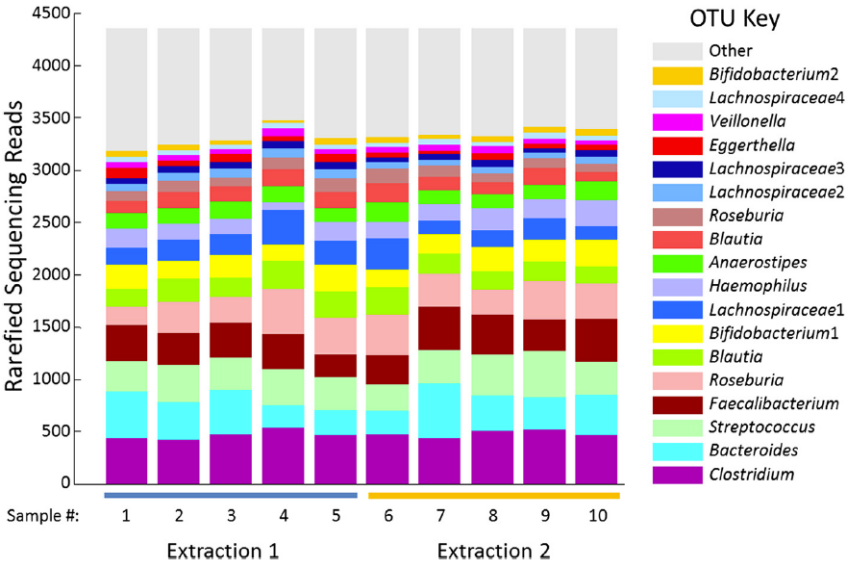


图 2 不同提取方法对微生物的影响

2. 不同量样品提取对微生物的影响

在研究过程中经常会遇到珍贵样本，总量很少，这样的样本进行测序分析会对结果有影响吗？这是很值得研究的问题。将样本分为 25mg、50mg、100mg、200mg 分别进行测序分析。一般线性模型分析（general linear models，GLMs）表明：alpha 多样性与样品质量无显著相关性（图 3）；同时，beta 多样性指数与样品质量也无显著相关性，但是与 50mg 样品相比，200mg 样品的 weighted UniFrac 指数较高。Weighted UniFrac distances 和 Bray-Curtis dissimilarity 结果显示：同一来源的不同质量样品之间更相似（图 4）。

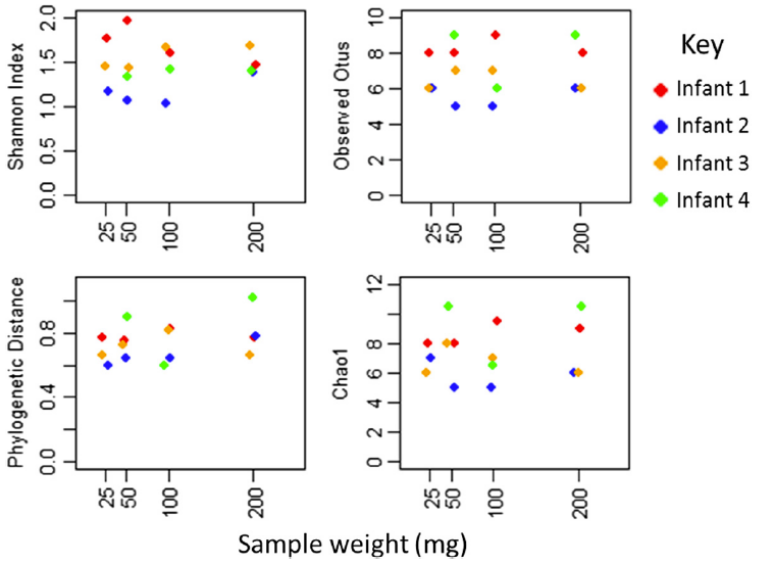


图 3 alpha 多样性与样品质量之间的关系

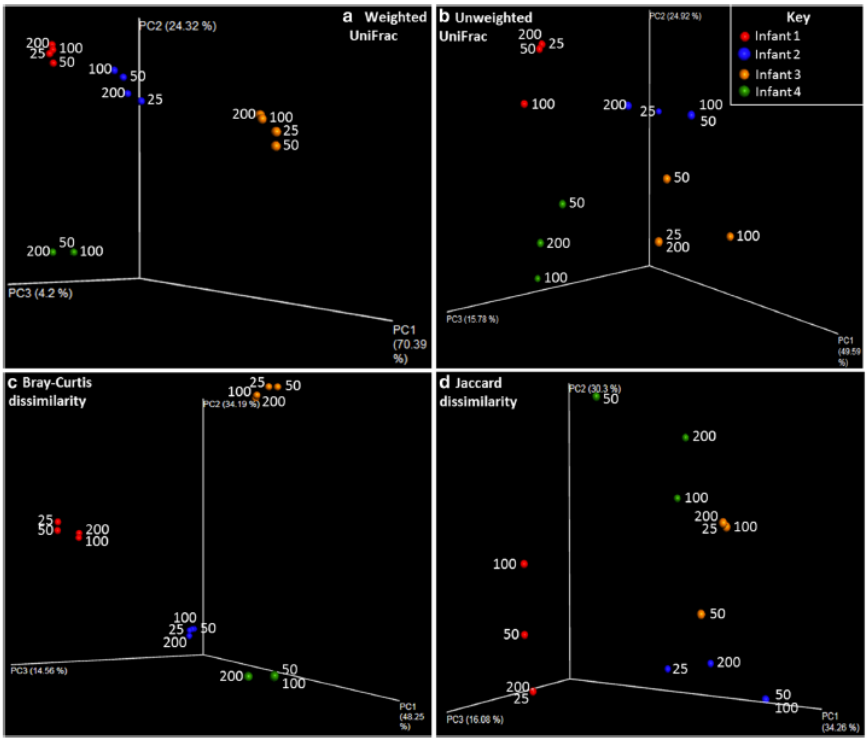


图 4 四个样品不同质量 beta 多样性分析

3. 样本长期冻存对微生物多样性的影响

当前样本保存的黄金标准为-80°C下保存，但是尚未有研究其对微生物群落的潜在长期影响。考虑到某些实验需要多年收集样本，本研究对冻存时间对微生物多样性的影响进行分析。结果发现四个多样性指数与冻存时间之间存在相关性，并且只有 observed OTUs 随着冻存时间的增加而减少（p=0.018）。GLMs 分析结果表明 beta 多样性与长时间冻存之间无显著相关性。四个样品（早产婴儿 5、6、7、8）中，有三个样品在冻存 24 个月后 weight UniFrac distances 偏高，说明这些样品中微生物丰度发生变化。通过 GLMs 分析，共发现 8 种微生物（Lactobacillus1、Lactobacillus2、Bacilli、Enterobacteriaceae、Gammaproteobacteria1、Enterococcus2、Staphylococcus1 和 Staphylococcus2）发生显著变化。图 6 结果表明，冻存 2 个月的样品中 OTUs 与其他月的样品 OTUs 相比无显著差异。

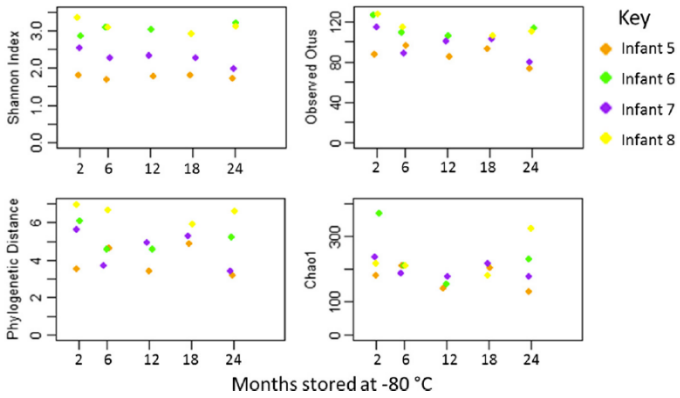


图 5 alpha 多样性与冻存时间之间的关系

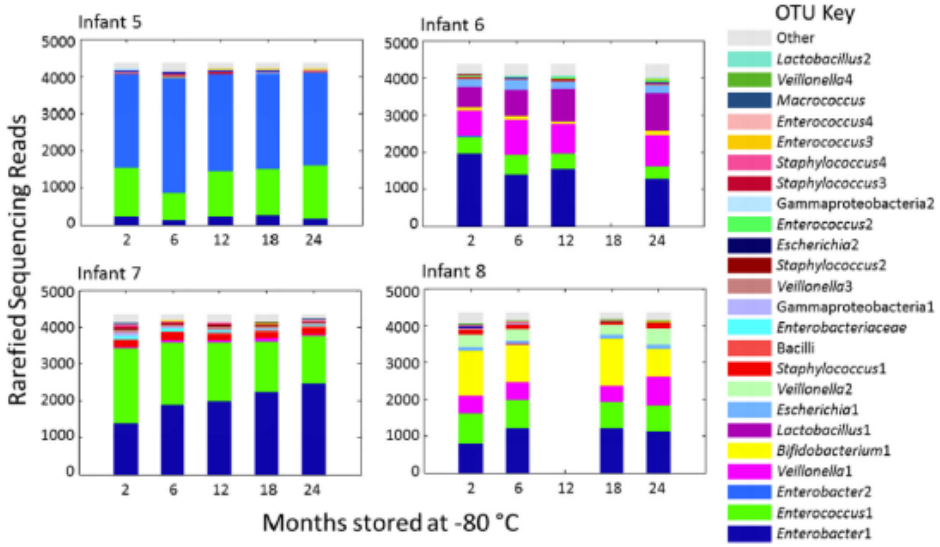


图 6 冻存时间与样品中微生物群落的关系

4. 室温保存对微生物群落结构的影响

在 alpha 多样性指数中，只有 Shannon index 随时间显著降低（图 7）。在 beta 多样性指数中（图 8），对于群落结构简单的样品，从第 4h 开始，随着时间延长，unweighted UniFrac distances 增加（p=0.038）；对复杂群落结构的样品，所有 beta 多样性指数都随着时间延长而增加。

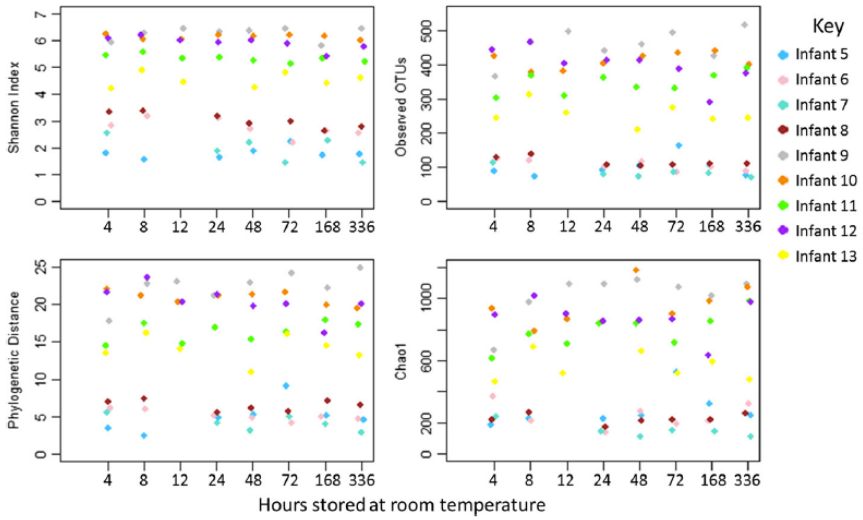


图 7 室温保存时间与 alpha 多样性的关系

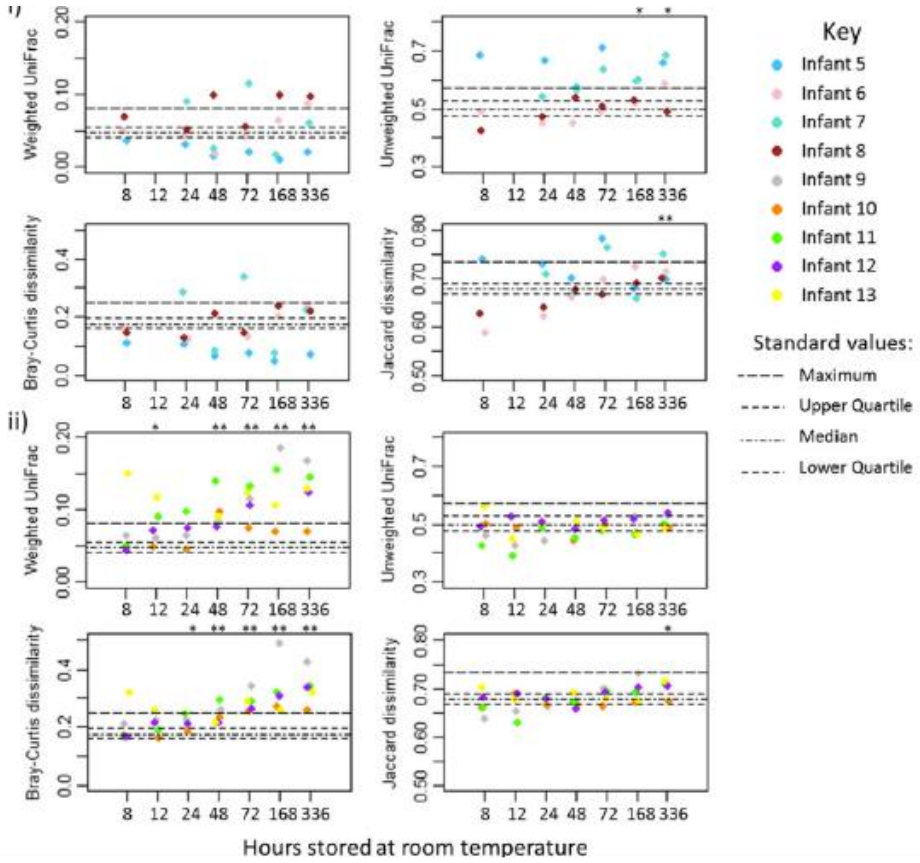


图 8 室温保存时间与 beta 多样性的关系

随着室温保存时间的延长，在门水平上，复杂微生物群落样品中厚壁菌门显著降低（p=0.004），变形菌门显著升高（p=0.018），而放线菌门微生物丰度在两种不同样品中都降低（p=0.013 和 0.033），校正后，只有厚壁菌门微生物丰度存在显著差异（图 9）。下图中 i)为简单微生物群落样品（早产婴儿）；ii 为复杂微生物群落样品（足月生产幼儿）。

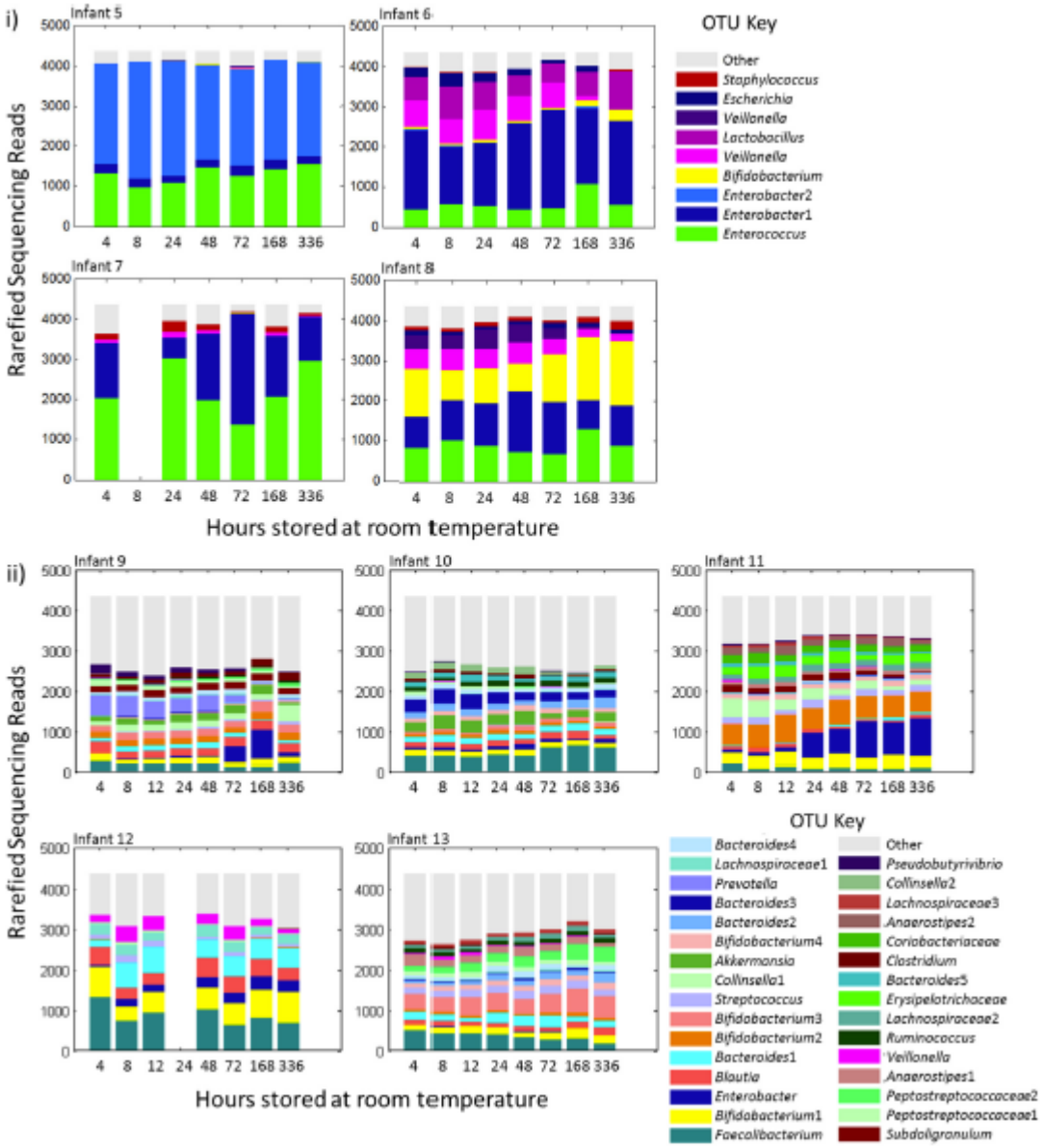


图 9 室温保存时间微生物群落结构变化的关系

4. 邮寄对微生物群落的影响

邮寄转移样本是常用的运输方式，虽然上述实验已经证明微生物在室温下相对稳定长达 2 天，但是邮寄系统本身也存在不稳定因素。通过对三组样品进行比较：①邮寄样品；②室温下保存 4h、然后-80℃保存的样品；③一直室温下保存样品。结果发现邮寄样品和一直室温下保存样品 alpha 多样性无显著差异，各 OTUs 及门水平丰度变化差异不显著；室温下 4h、然后-80℃保存的样品与其他两组样品 beta 多样性无显著差异。

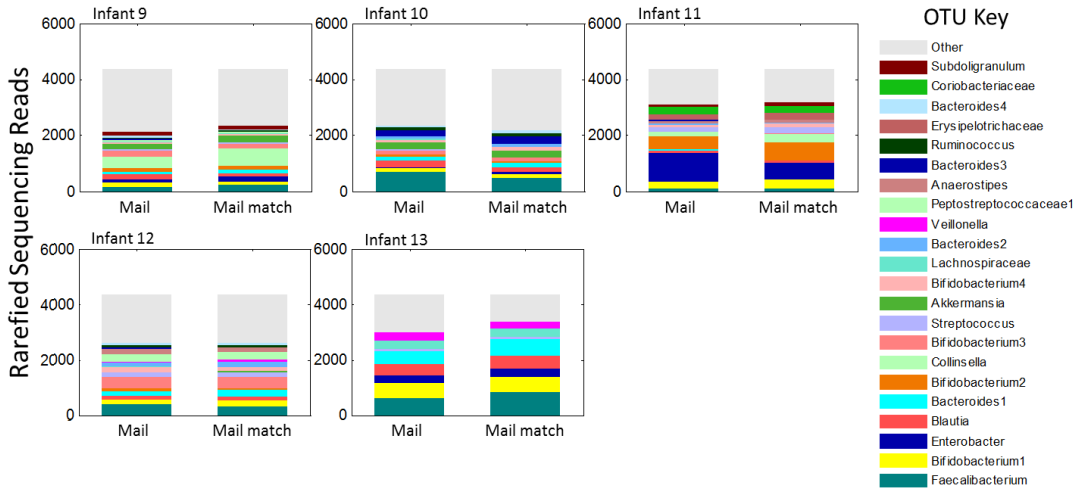


图 10 样本邮寄对微生物多样性的影响

通过以上研究得出以下结论：1. -80℃保存样品 2 年内对微生物群落结构的影响较小，只对其中部分微生物如 Lactobacillus 和 bacilli 等影响显著；2. 样品室温下放置 2 天就会对微生物群落产生显著影响，所以应在两天内处理或者放在-80℃保存；运输时要充分考虑到运输时间；3. DNA 提取人员与提取试剂批号对微生物群落结构的影响不显著；4. 对于微生物多样性较低的样品，减少样品提取用量对 alpha 多样性和一些 beta 多样性的影响不显著；但是低样品质量会影响 weight beta 多样性，同时对 OTUs 丰度影响显著。

研究亮点：

通过 16S 测序，系统性的研究多种因素（提取方式、保存时间、保存方法、运输、样品量）对微生物多样性的影响，对科研工作者如何处理及保存样品具有重要指导意义。

宏基因组测序特别专题

专题一 16S 测序样本要求及取样建议

在 16S测序项目中，高质量的DNA是得到完美结果的第一步。即使实验设计及样本处理都十分严谨，但是在取样时稍一不注意，样本遭到破坏，那么测序数据必然不会完全符合预期了。因此，取样及提取是大家都十分关心的问题，经过我们实验团队多年的经验，我们对各种不同的微生物样本取样及DNA提取均做了总结汇总，便于大家参考。

粪便样本的收集及提取:

1.收集样本：一般对于人的粪便，要求用无菌粪便收集器或其他灭菌器皿收集粪便样本，要注意即刻进行样本标记并低温保存（考虑到患者样本可能无法取样后直接分装标记，可以先在低温条件下保存，尽快进行分装标记）；对于小鼠粪便，分别收取采样小鼠的粪便，立即进行收取，尽量不要暴露在空气中太长时间，避免污染和降解；

2.分装保存：由于有些样本取样不易，因此我们建议一次收集后分装保存，对于人的粪便，我们建议使用无菌牙签挑取内部样本，尽量排除空气污染及防止样本降解，使用灭菌离心管分别称取 0.2g左右的样本，每个样本分装 10 管左右备用，标记好后即刻放入-80℃保存。当然由于小鼠个体较小，可能粪便样本不足 0.2g时，我们可以适当将多个生物学重复样本进行混合，进行保存。

3.建议：一般来说，每次实验取用一管，用过后将剩余样本丢弃；并且若遇上收集困难或者十分珍贵的样本，请老师注意备份，避免再次收集耗费更大的人力物力，并且无法保证实验条件完全一致，导致后续分析无法达到预期。

DNA提取：建议使用 E.Z.N.A.®Stool DNA Kit 或者使用联川试剂盒

土壤样本的收集及提取：

1.样本收集：取样区域一般是根据实验设计进行选择，建议选取具有代表性的土壤；采样时使用的所有工具、采集袋或其他物品均要使用已灭菌的；若野外没有合适的条件进行灭菌处理，可以使用采集的土壤对取样工具进行擦拭，尽量去除干扰；采集时一般选 取 5-10 cm处土壤，去除杂质，将一定量的土壤进行分装标记，每袋样品约 5-10 g， 密封后立即低温保存；

2.建议：土壤取样时时常会碰到杂质含量较高，取样时需要对杂质进行过滤，避免石块等杂质戳破采集袋等情况的发生，也会给后续提取造成干扰；

DNA提取：建议使用 E.Z.N.A.®Soil DNA Kit 或者使用联川试剂盒

肠道内容物的收集及提取

1. 样本收集：一般取样时实验对象已死亡（动物），需要使用无菌的解剖刀，尽量在无菌状态下取出整个肠道，将实验所需的肠段的内容物切下，将内容物取出放置在无菌离心管中，建议分装保存，便于后续实验需求，低温保存运输；若是无法立即抽提，建议将样本先置于液氮中冷冻 4 小时以上，保证冷冻充分再转移到-80℃保存；

2. 建议：由于肠道内容物来源不一，一般来说保存单个样本取样量如下：大型动物，如牛、羊 500 mg-1000 mg/管；中型动物，如大鼠、家兔 200-1000 mg/管；小型动物，如小鼠 200-500 mg/管；对于肠道内容物较少动物，如鱼、 虾 ≥100 mg/管。为保证实验顺利进行，在采样允许的情况下，尽量多采集样品。

DNA提取：建议使用 E.Z.N.A.®Stool DNA Kit 或者使用联川试剂盒

肠道粘膜微生物的收集及提取

1.样本收集：实验操作在无菌条件下进行，将整个肠道移出体外，将肠系膜的对面纵向切开，使肠腔外露；使用无菌器械将肠道内容物清理干净后，用无菌生理盐水清洗肠腔，为了避免扰破坏乱肠粘膜 ,需将看得见的肠道内容物轻轻去除 ;用无菌载玻片轻刮肠道粘膜(避免穿透基底膜）。所有样品取完后立即放入液氮冷冻，并放在-80°C冰箱保存。

2.建议：由于肠道黏膜微生物收集要求较多，而采集到的黏膜微生物比较少，建议重复上述步骤多次，但必须在无菌环境下操作，保证足够取样及保存。

DNA提取：建议使用OMEGA E.Z.N.A.® Stool DNA Kit或者使用联川试剂盒

生殖道微生物的收集及提取

1. 样本收集：取样对象需 48h内无性行为，不得进行私处清洗、上药等改变菌群结构的行为， 30 天内不能用抗生素和抗真菌类药物，以免导致菌群结构变化；取样时用无菌棉签擦拭，充分取样，用无菌剪刀剪下棉签头部，放入装有Amies培养基的无菌离心管中，立即-80°C保存；

2. 建议：生殖道微生物收取不易，样本珍贵，需进行多管备份保存；提取DNA时需先将棉签低温解冻，漩涡处理 5min以重悬细胞。
DNA提取：建议使用联川试剂盒

口腔微生物的收集及提取

1. 样本收集：口腔微生物收集可以有两种方法供选择：

一、（**推荐**）根据研究目的选定时间段对 24h内不刷牙的受试者进行取样 ;取样前漱口，用无菌金属环轻轻刮取牙齿表面 10-20 次；将其置于装有 1ml盐溶液的 1.5ml无菌离心管，盐溶液经过UV照射以避免DNA污染；将样本混合完全后-80°C保存；

二、根据研究目的选定时间段对 24 h内不刷牙的受试者进行取样，使用特定唾液取样器；受试者向取样器中吐 2-3 次唾液（一次约 1 ml）；将配套的DNA稳定剂加到唾液中；密封后，做好样品标记，-80°C保存。由于唾液中含有抑菌成分且宿主DNA过多，不推荐使用此方法。
2. 建议：不同研究目的受试者情况不同，建议取样时进行多管备份保存，以备后续实验需求。

DNA提取：建议使用联川试剂盒

皮肤微生物表面的收集及提取

1. 样本收集：取样前 24h不能洗澡，不能使用润肤乳及抗菌活性的肥皂等；取样前 7 天不能使用抗菌成分的沐浴用品；取样时采用无菌棉签或无菌手术刀片轻轻刮取皮肤表面取样，区域大小约为 4 cm^2；若是取指甲等其他部位样本，需将指甲样品剪碎后用蛋白酶K过夜 55°C处理，其他样品用yeast lysis buffer and lysozyme处理 1h（37°C）；-80°C保存。

2. 建议：由于皮肤微生物采集较困难，菌量可能不多，因此建议取样时进行多管备份保存，以备后续实验需求。

DNA提取：建议使用联川试剂盒

血液微生物的收集及提取

1. 样本收集：用EDTA抗凝管抽取 1-2ml全血；颠倒抗凝管 8-10 次，充分混匀EDTA 和全血，保证抗凝效果；-20℃保存。

2. 建议：由于血液样本十分特殊，提取的DNA宿主会占多数，且样本中本身细菌含量较低，无法保证扩增建库成功，因此需要根据实际情况及经验调整提取方案。但实际情况，无论是何种方法提取，成功率均不高。

DNA提取：建议使用联川试剂盒或其他试剂盒

植物内生菌微生物的收集及提取

1. 样本收集：根据研究目的选取对应的新鲜植物组织；依次用 70%无水乙醇浸泡 40s，再用 2.5%次氯酸钠浸泡 10 min，每 100 ml 2.5%次氯酸钠加一滴吐温 80； 无菌水清洗 2-3 次，-80℃保存。

2. 建议：由于植物内生菌微生物样本十分特殊，提取的DNA宿主会占多数，若宿主DNA去除失败，则无法保证扩增建库成功，因此需要根据实际情况及经验调整提取方案。

DNA提取：

建议参考文献：内生细菌提取： Illumina-based analysis of endophytic bacterial diversity and space-time dynamics in sugar beet on the north slope of Tianshan mountain；

内生真菌提取：Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis

空气微生物的收集及提取

1. 样本收集 抽取针对实验目的空气微颗粒，用不同孔径大小的无菌滤膜筛选目的颗粒；（将含颗粒过滤膜装入无菌铝箔内，密封在-80℃长期保存）；截取适量大小过滤膜，置于含 1×PBS缓冲液的 50 ml离心管；在 4℃条件下，加速度离心 2 h；温和漩涡处理，0.2 μm的Supor 200 PES Membrane Disc Filter 过滤纸过滤重悬液。

2. 建议：由于样本特殊，微生物含量较少，建议多管备份保存。

DNA提取：建议使用OMEGA E.Z.N.A.® Water DNA Kit

水体微生物的收集及提取

1. 样本收集：根据实验设计确定取样的深度和范围，由于水体样本的特殊性，取样体积需大于 5 L，取样工具需灭菌处理，采样后进行滤膜过滤，选择合适孔径的滤膜。对于澄清水体或者略浑浊水体，选用 0.22 μm或者 0.45 μm的滤膜；对于浑浊水体，建议先用大孔径的滤膜过滤一遍，再用小孔径的滤膜过滤；水体泥样的采集提供大于 5 g样本。

2. 建议：由于水体样本的特殊性，样本直接保存困难，因此建议一次取样足够，一般提取一次需要 2 张滤膜，分装保存足够量的滤膜，便于后续实验的使用。

DNA提取：建议使用 E.Z.N.A.®Water DNA Kit

活性污泥及海洋沉积物微生物的收集及提取

1. 样本收集：活性污泥样本取自活性污泥装置中，一般取样量大于 20 ml，将其置于无菌管中，建议立即提取后-80℃保存；海洋沉积物样本根据研究目的进行取样，一般建议取

1kg以上的沉积物装入无菌的塑料袋中，低温运输并尽快进行提取后-80°C保存。

2. 建议：活性污泥样本等一般建议及时提取，样本体积较大，长时间-80°C保存一方面储存不便，另一方面冻融提取也比较困难，加大了实验工作量及难度。建议之间提取后保存DNA即可。

DNA提取：建议使用E.Z.N.A.®Soil DNA Kit

物体表面微生物的收集及提取

1. 样本收集：根据物体形状大小采用不同的方式进行收集，对于形状较小的物体，直接将其放入无菌的容器中，加入PBS进行冲洗，振荡使微生物脱落到PBS中，利用PBS提取DNA或者经滤膜过滤之后进行DNA提取；形状较大物体：采用无菌棉签沾取无菌水擦拭物体表面取样，或用无菌手术刀片轻轻刮取物体表面取样，均使用无菌离心管收集样品，-80°C保存。

2. 建议：物体表面微生物根据实际情况不同存在差异，若微生物较少可能会对提取存在干扰，建议多取几次，保证足够的量。

DNA提取：建议使用联川试剂盒

样品送样建议

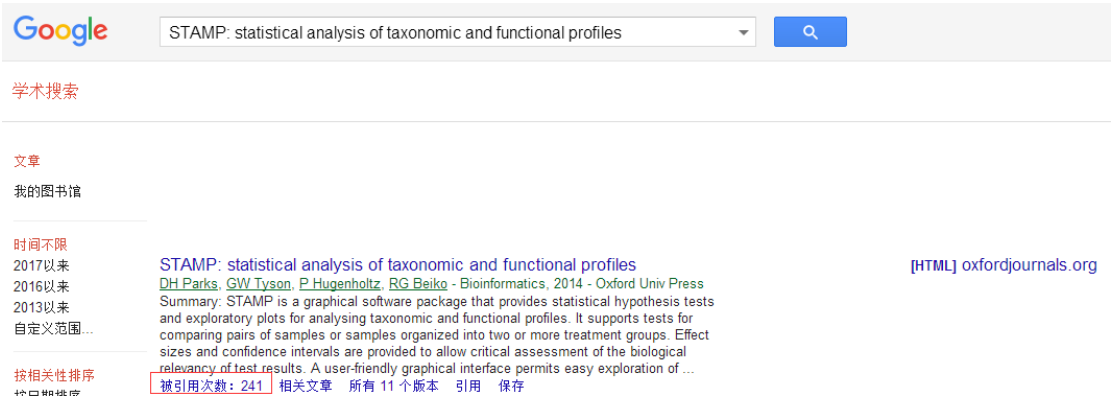
宏基因组项目：c ≥10 ng/μl；m ≥1 μg；浓度以Qubit质检结果为准，建议送 2 次建库的量，即 2 μg。OD 260/280 在 1.8-2.0 之间；OD260/230 在 1.8-2.0 之间。如样品难以取得，最低总量可低至 100 ng DNA，浓度 ≥1 ng/μl。

扩增子项目：c ≥2.5 ng/μl （Qubit）； m ≥100 ng；建议送 2 次建库的量，即 200 ng。OD260/280 在 1.8-2.0 之间；OD260/230 在 1.8-2.0 之间。

专题二 组间差异分析神器-STAMP

在微生态研究中，当我们做完 16S 或宏基因组测序等多样性测序后，想找出不同处理组之间差异物种或差异基因，一般常用的组间差异分析 metastats（只能用于两组之间的差异比较），LEfSe，秩和检验等。对于不会编程的我，如何根据自己的数据特征选择不同的差异统计方法并获得相应的差异分析结果呢？在这里给大家介绍一款简单实用的组间差异分析软件-STAMP，而且该软件分析获得的图片可直接用于文章的发表。

STAMP 来源于 2014 年的文章（下图），在短短的时间内受到众多科研者的青睐，目前其被引用 240+次，非常值得肯定。该软件的强大之处，不仅能够对两组甚至多组样本及两两样本之间的 KEGG、COG、基因及任何分类水平的物种等进行显著性差异分析，同时带有 10 多种可选择的差异检验方法以及图形展示形式（柱状图，散点图，热图，pca 图等）。最重要的是，每种图形基本上都能直接用于发表文章，而且该软件的操作简单易学。



STAMP 软件下载网址：http://kiwi.cs.dal.ca/Software/STAMP （可支持 windows，linux 及 OS X 多种操作系统，根据需要自行下载安装。该软件的安装也十分简单，按照默认的参数选择安装即可）

言归正传，在使用 STAMP 软件之前，首先需要准备文件（丰度表文件和分组信息文件），文件的格式如下（以 OTU 丰度表为例）。

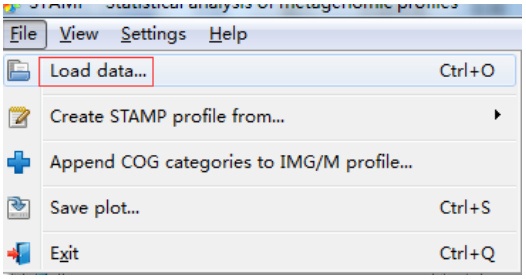
1）丰度文件，每列之间用 tab 键隔开的 txt 文件（可在 excel 表格内编辑，然后保存为 txt 格式文件，需要注意的是该文件一定要包含表头）

#OTU ID	A1	A2	A3	B1	B2	B3	C1	C2	C3
OTU1	0	0	1	0	0	0	103	0	29
OTU2	70	15	59	8	21	78	278	36	63
OTU3	6	4	5	7	8	1	5	14	19
OTU4	0	0	0	4	3	3	0	0	0
OTU5	7	25	27	66	37	9	10	22	6
OTU6	313	646	603	1292	856	266	438	260	168
OTU7	15	62	15	37	61	63	17	54	7
OTU8	26	18	50	142	77	65	21	20	10
OTU9	2	1	2	46	95	54	18	0	0
OTU10	2	0	2	0	0	1	10	0	2
OTU11	324	65	418	42	64	47	222	310	171
OTU12	0	0	0	3	3	0	0	1	0
OTU13	52	28	73	279	149	43	226	287	150
OTU15	10	33	24	30	67	90	0	0	1
OTU16	0	2	0	0	1	3	0	0	0
OTU17	53	14	88	0	0	0	0	0	0
OTU18	0	0	2	6	5	12	0	1	0
OTU19	0	0	0	0	3	8	0	11	0
OTU20	0	0	1	0	0	0	1	0	0
OTU21	0	0	0	1	1	5	0	0	0
OTU22	23	4	15	5	8	17	84	20	23
OTU23	1	0	0	2	6	2	3	0	1
OTU24	0	0	1	0	1	3	0	1	0
OTU25	0	0	0	1	1	0	0	0	0
OTU26	5	0	4	3	0	0	0	0	0

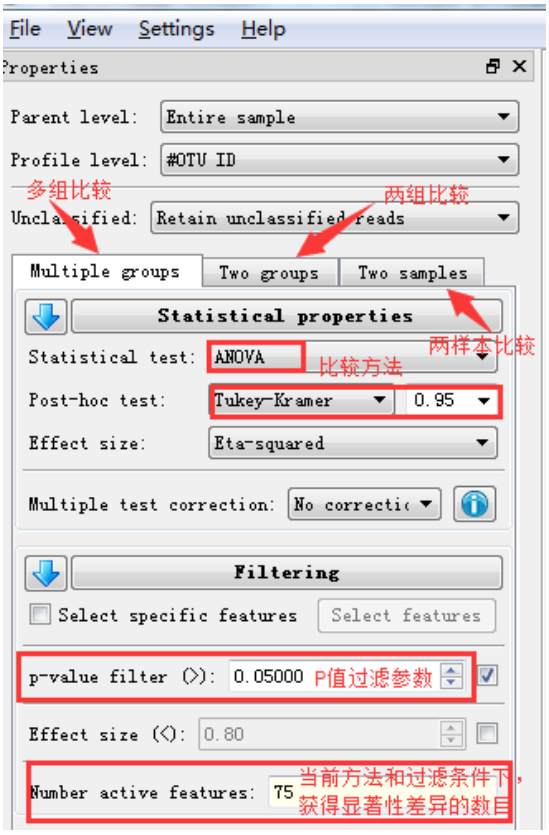
2）分组信息文件（格式同丰度表格式，该文件也需要加入表头，否则会默认第一行为表头，导致样本缺失。）

#SampleID	Group
A1	A
A2	A
A3	A
B1	B
B2	B
B3	B
C1	C
C2	C
C3	C

1、文件导入(方法 File-load data，选择文件导入，注意文件存放的路径中不能包含中文字符)



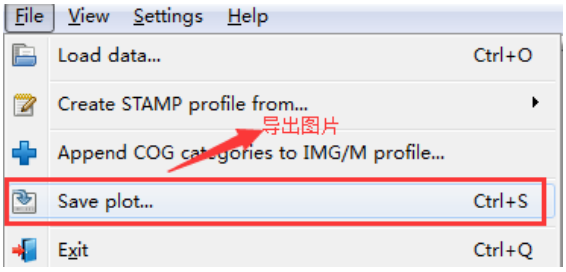
2、文件导入成功后，就可以设置参数，绘制专属的图片了。具体的参数设置见下图：



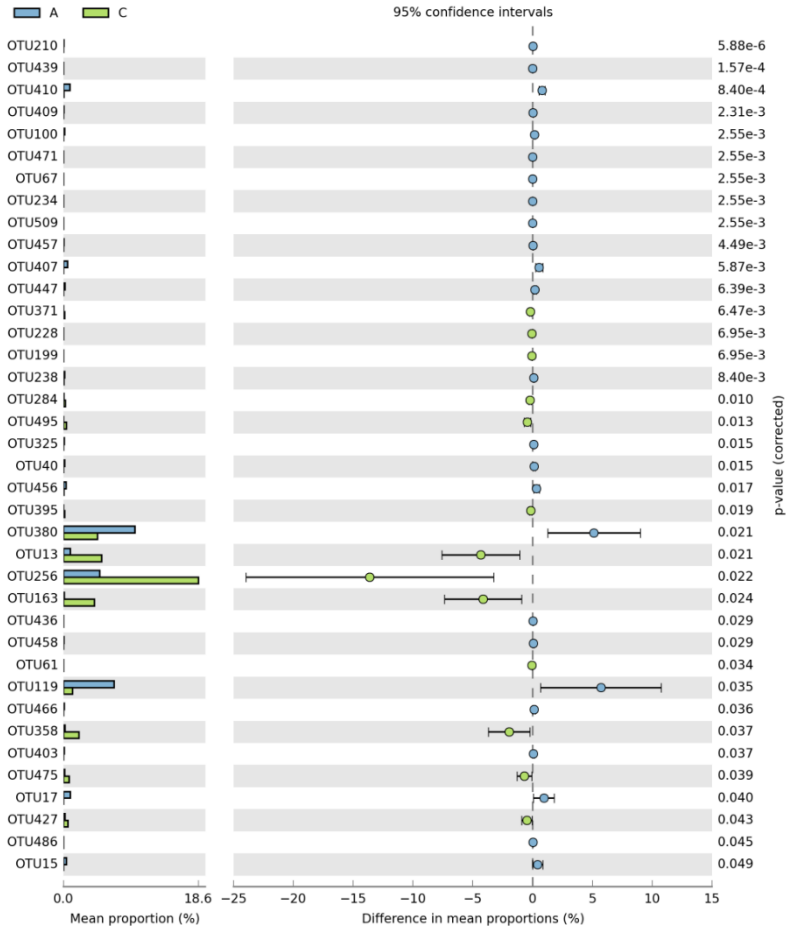
该软件默认打开界面 Multiple groups（多组比较），根据实际需要的比较方案进行选择，比如想进行两组之间的比较，首先点击 Two groups，然后选择需要比较两组的组名以及统计方法和过滤条件，即可进行显著性差异统计分析。其中多组分析统计学方法包括 ANOVA 和 Kruskal-Wallis H-test。

两组之间比较统计学方法包括 t-test(equal variance)，Welch's t-test 和 White's non-parametric t-test。为了确保统计学意义和结果的准确度，需要选择合适的检验方法。t-test 检验可以在最少样本数为 4 的时候保持较高的准确度和精确度，当两个分组之间具有相同的方差时，用 t-test 更为准确，当方差不同时，Welch's t-test 更为准确。White's non-parametric t-test 算法计算时间较长，当样本数目少于 8 的时候，可以使用该检验方法，当样本数目过多时，不宜使用该检验方法。

3、作图类型和图形导出



4、结果示例



5、STAMP 软件使用注意事项

1）STAMP 作图原始数据来源？

STAMP 软件可以直接使用 QIIME 的 biom 文件和 PICRUST 的 KEGG 和 ko 文件，groupfile 需要老师根据自己实验设计进行样本分组。

值得注意的是，这些结果文件存放路径中不能存在中文字符，否则在数据无法导入到软件。比如文件路径为：E:\16S\stamp_data\OTU_table.txt（不存在中文字符是可以的），如果文件路径为 E:\stamp_分析\OTU_table.txt(存在中文字符报错）或者文件存在中文字符也是不行的，比如 E:\stamp_data\OTU_丰度表.txt

2） Unclassiffied 选项中，remain Unclassiffied reads、remove Unclassiffied reads、和 use only for calculating frequency profiles 方法的区别？

当输入的丰度表文件和分组文件样本不一致时，对该参数进行设置。其中 remain Unclassiffied reads 和 use only for calculating frequency profiles 方法会保留所有的数据，而 remove Unclassiffied reads 仅仅保留有确定分组信息的数据。注意的是，分组文件的样本一定在丰度表文件中存在，否则会报错，反之，是可以的。

3）当打开了一个分析文件后，如果再重新打开新的文件会显示错误？

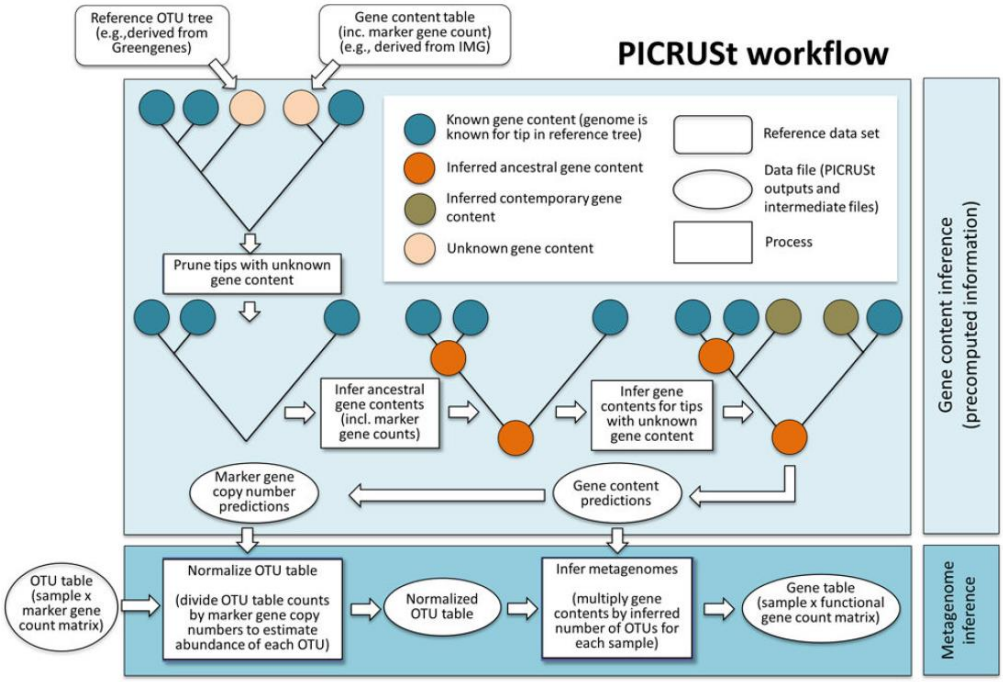
主要原因是目前版本的 STAMP 存在一些小的 bug，一次分析只能使用一个数据文件，如果要打开新的数据文件，需要关闭软件后重新打开。如果分析完成，一定要记得保存哦，不然需要再重新分析一次。

专题三 PICRUST 分析介绍

PICRUSt(phylogenetic investigation of communities by reconstruction of unobserved states) 通过隐性状态重建进行群落系统进化的研究，该软件基于 16S rDNA 和参考序列数据库，预测宏基因组功能组成。

PICRUSt 的原理基于已测细菌基因组的 16S rRNA 全长序列，推断它们的共同祖先的基因功能谱，对 Greengenes 数据库中其它未测物种的基因功能谱进行推断，构建古菌和细菌域全谱系的基因功能预测谱，最后，将测序得到的菌群组成"映射"到数据库中，对菌群代谢功能进行预测（详见下图）。

为了能够通过 16S 测序数据来准确的预测出功能构成，首先需要对原始 16S 测序数据的种属数量进行标准化，因为不同的种属菌包含的 16S 拷贝数不相同。然后将 16S 的种属构成信息通过构建好的已测序基因组的种属功能基因构成映射获得预测的功能结果。



图：PICRUSt 工作流程

该分析的缺点是古菌和细菌域全谱系的基因功能预测谱是基于 Greengenes 数据库进行构建的，Greengenes 版本为 gg_13_5，已经长时间未更新，因此很多古菌和细菌并未包含在内。此外，该预测结果只能预测到 KEGG 某个 pathway 水平，但不能从基因层面预测研究。如果在做宏基因组研究之前，想先看下关注的 pathway 是否存在或是否在不同分组中存在显著性差异，此时可以先做下功能预测，基于功能预测结果进行后续宏基因组实验设计。

PICRUSt 在线分析网址：http://huttenhower.sph.harvard.edu/galaxy/

具体使用步骤如下：

1、输入数据格式

# Constructed from biom fil		
#OTU ID	staggered	even
290426	1	0
265267	10	0
248563	0	1
316653	0	6
61102	2	2
465087	1	12
591907	152	215
456375	8	48
9670	294	46
140061	0	1
368134	1805	493
152176	0	4
582866	1	3
255326	1	5
201726	1	0
138395	10	9
165676	2391	789
551871	14	24

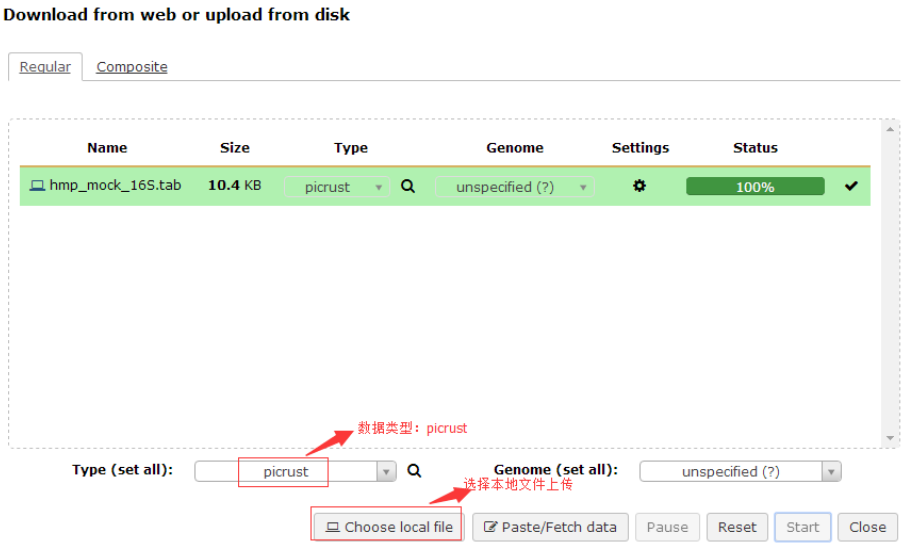
注意：该ID号为Greengene物种ID号

```
228054  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
844608  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
178780  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
198479  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
187280  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
179180  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
175058  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
176884  k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae; o__Synechococcales; f__Synechococcaceae; g__Synechococcus; s__
228057  k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__Pelagibacteraceae; g__; s__
```

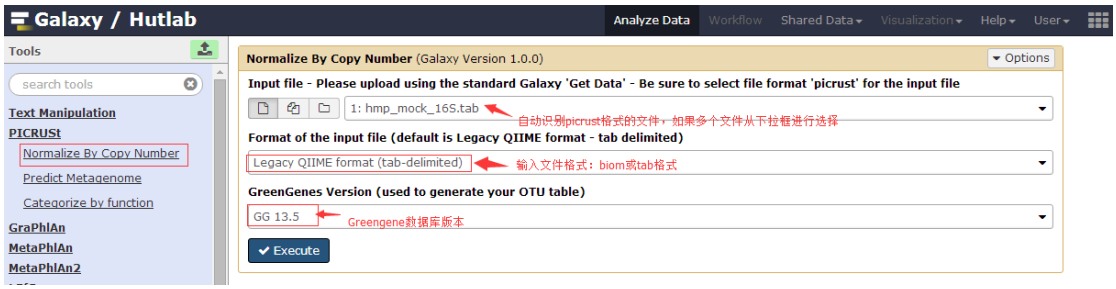
注：数据输入格式第一列为 Greengene 物种 ID 号，第二列-第 n 列为不同样本的表达量。如果 OTU 注释所用的数据库不是 Greengene ,需要根据 OTU 物种注释信息 ,转换为 Greengene ID。Greengene 对应的物种信息下载网址如下：

http://greengenes.secondgenome.com/downloads/database/13_5 （由于 Greengene 长时间未更新，用其他数据库注释得到的物种可能在 Greengene 物种列表中不存在）

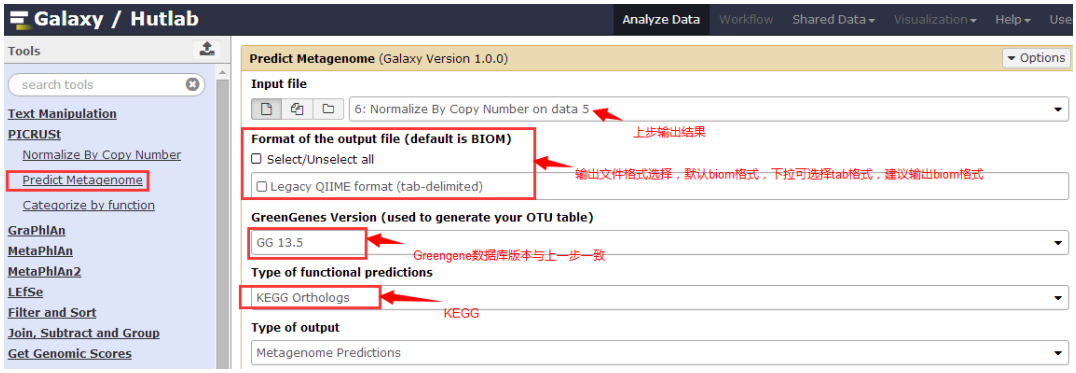
2、将数据上传至网页（需要注意的是数据类型需要选择：picrust）



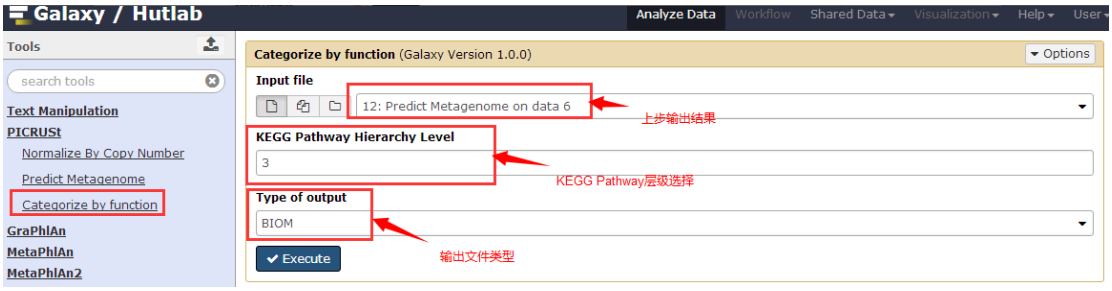
3、Normalize By Copy Number



4、Predict Metagenome（在预测的分类中，按需选择“KEGG Orthologs”、“COG”或者“Rfam”）



5、Categorize by function



其中“KEGG Pathway Hierarchy Level“有三种层级（具体的层级含义见下图）可选，输出文件格式建议使用更常见的格式“Legacy QIIME format(tab)”。

KEGG Pathway 网址：http://www.kegg.jp/kegg/pathway.html

2. Genetic Information Processing 第一层级

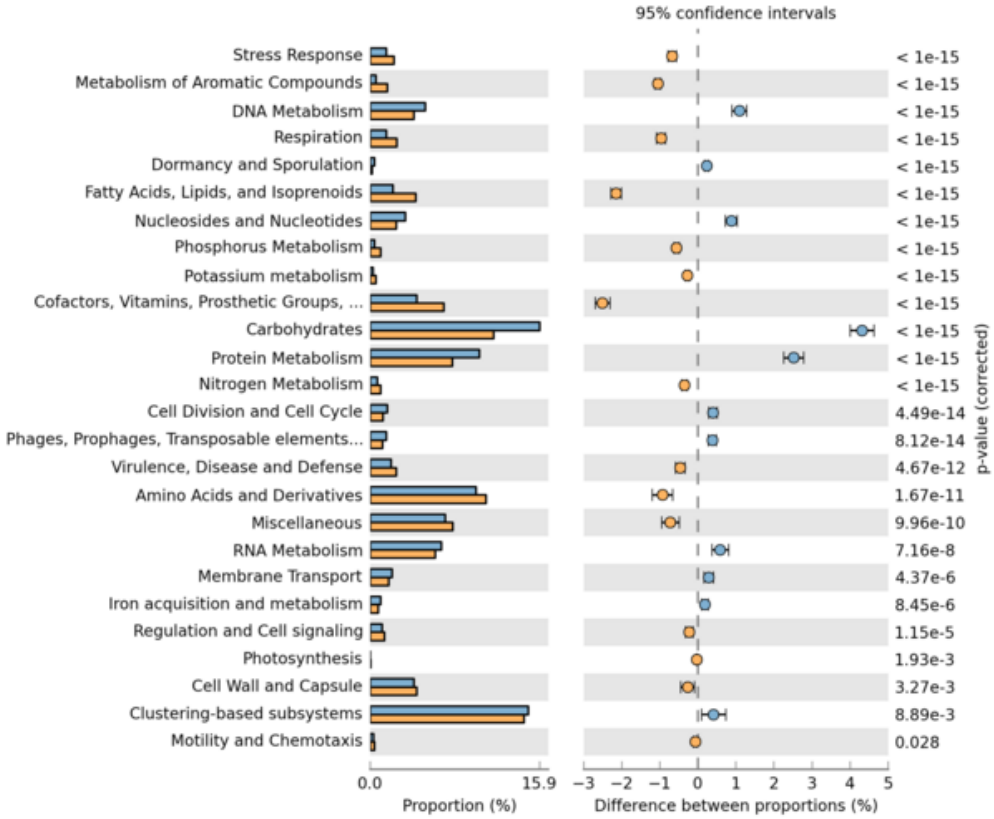
2.1 Transcription 第二层级
- RNA polymerase — Transcription factors
- Basal transcription factors — Transcription machinery
- Spliceosome — Spliceosome

2.2 Translation 第三层级
- Ribosome — Ribosome
- Aminoacyl-tRNA biosynthesis — Ribosomal proteins
- RNA transport — Ribosome biogenesis
- mRNA surveillance pathway — Transfer RNA biogenesis
- Ribosome biogenesis in eukaryotes — Translation factors

2.3 Folding, sorting and degradation
- Protein export — Chaperones and folding catalysts
- Protein processing in endoplasmic reticulum — SNAREs
- SNARE interactions in vesicular transport — Ubiquitin system
- Ubiquitin mediated proteolysis — Proteasome
- Sulfur relay system
- Proteasome
- RNA degradation

2.4 Replication and repair
- DNA replication — DNA replication proteins
- Base excision repair — Chromosome
- Nucleotide excision repair — DNA repair and recombination proteins
- Mismatch repair
- Homologous recombination
- Non-homologous end-joining
- Fanconi anemia pathway

6、基于预测结果，统计不同分组间显著性差异的 KEGG Pathway（该分析可直接使用 Windows 本地软件 STAMP（下文会对该软件进行详细使用说明）实现）



至此，16S 在线功能预测分析就完成了，有没有觉得很简单呢？如果 16S 功能预测的结果不能满足您对微生物功能层面的研究，建议选择合适的样本进行宏基因组或宏转录组测序分析，会获得意想不到的收获。

专题四 MEGA 绘制系统发育树详解

通常我们会有绘制系统发育树的需求，在得到一条或几条感兴趣的序列后，在 NCBI 上 blast 得到相关的一些序列，想要拿这些做一张序列之间的关系图，那么请跟着我们的提示一步步操作吧，简单方便，自己就能做出美观的图片。

第一步：下载软件并安装

该软件为开源软件，可在百度搜索并下载；或者之间点击下方链接进行下载即可；链接：http://pan.baidu.com/s/1qXQrb9i 密码：9i3x；

双击即可进行安装，无特殊要求。

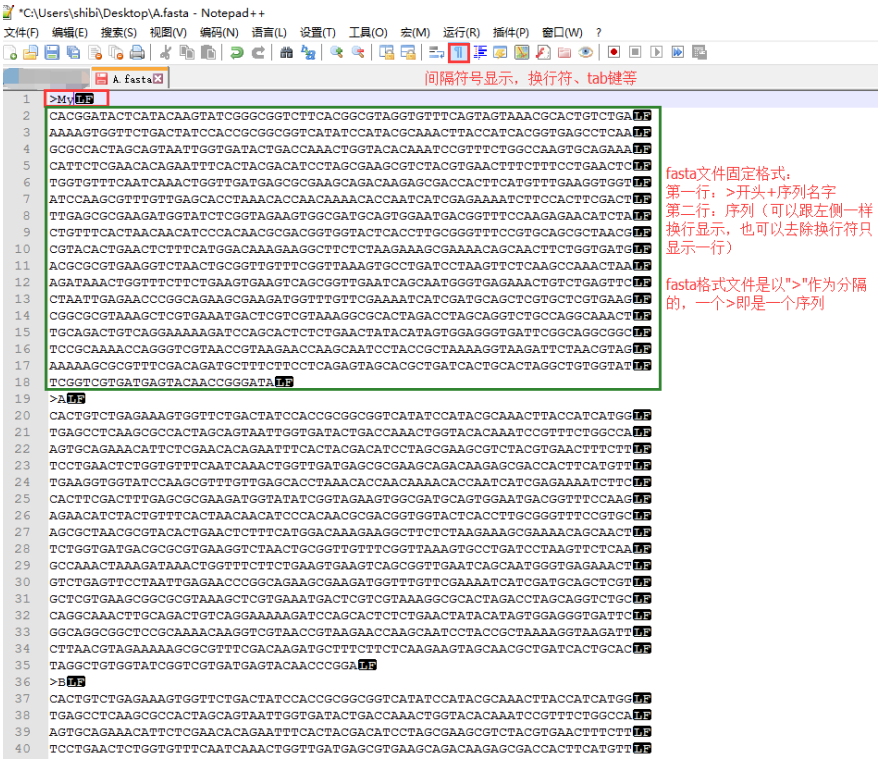
第二步：下载 notepad++软件并安装

该软件可打开比较大的 txt 文档，并且有良好的编辑功能及可视化功能，便于序列文件的调整，建议客户安装；

可直接在 360 软件管家中搜索并下载安装；或者直接百度搜索并下载安装。

第三步：整理需要作树的序列

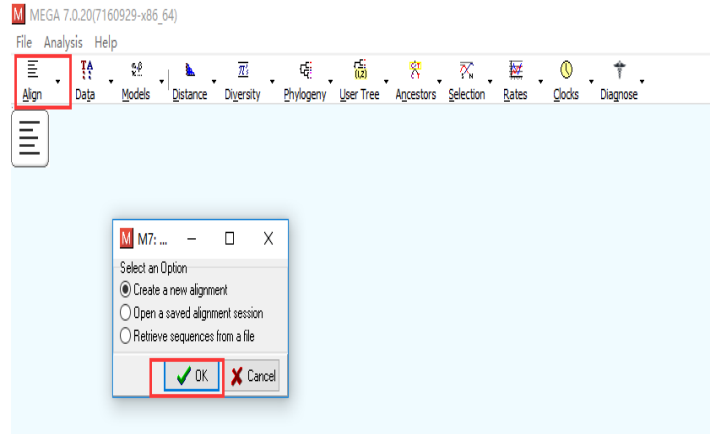
将下载好的序列整理成 fasta 格式（必须是该格式），并通过 notepad 打开，格式如下：

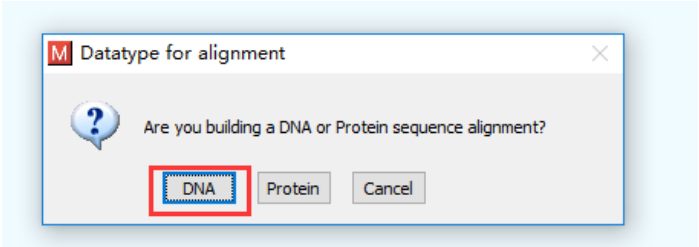


将整理好的文件保存为“文件名.fasta”。

第四步：将序列导入 MEGA 软件进行 align

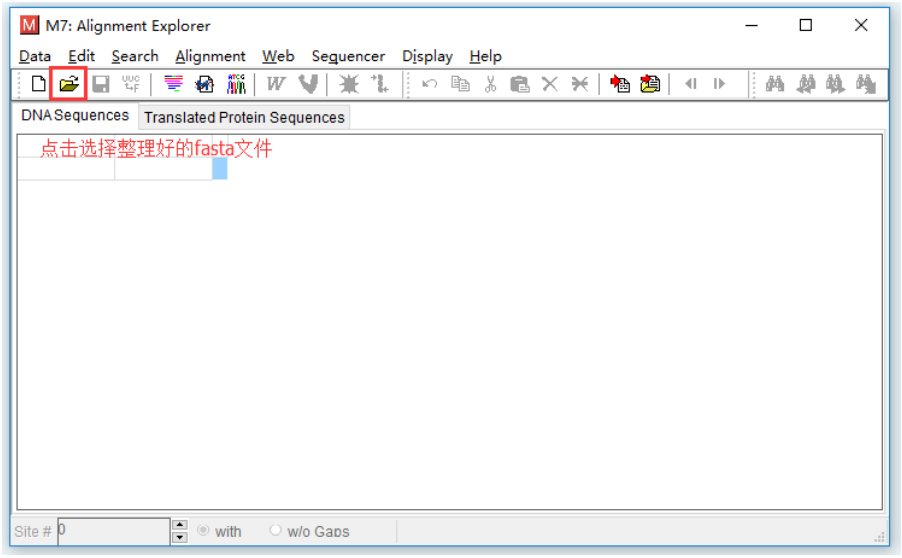
双击打开 mega 软件，点击左上角 Align 按钮，点击 edit\bulid alignment，点击 ok。

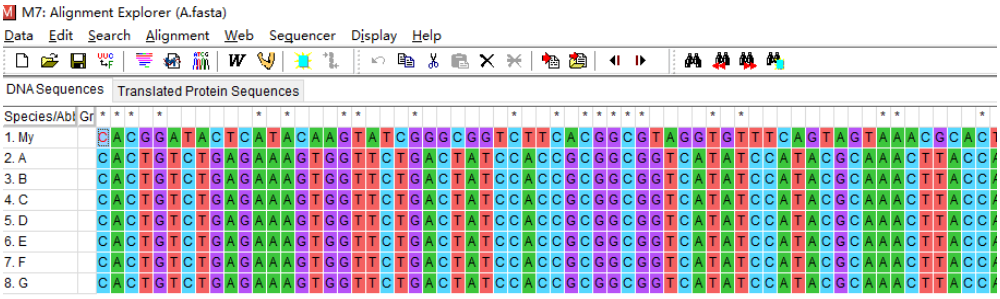




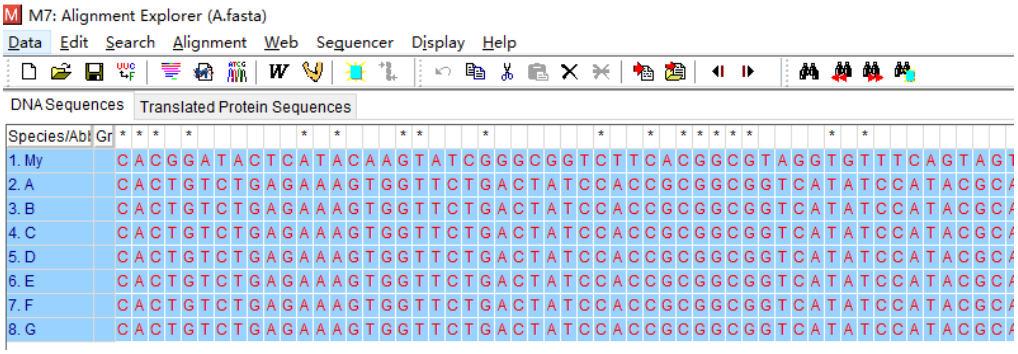
我们比对的是 DNA 序列，勾选 DNA;

（1）导入整理好的 fasta 文件

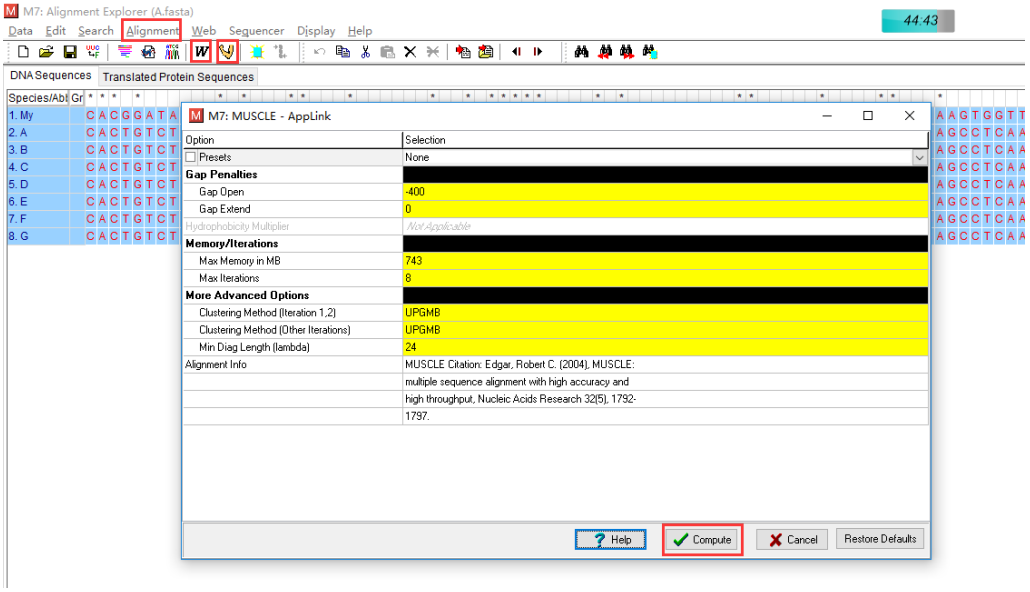




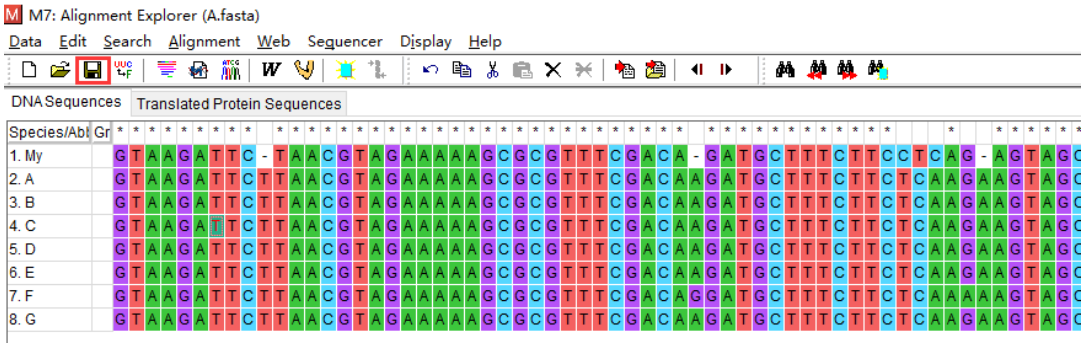
（2）将导入序列全部选中，点击第一条序列，按住 shift 键，点击最后一条序列；



（3）开始比对：可以之间选中快捷键，或者点击 Aligment 后选中相应的比对方式；

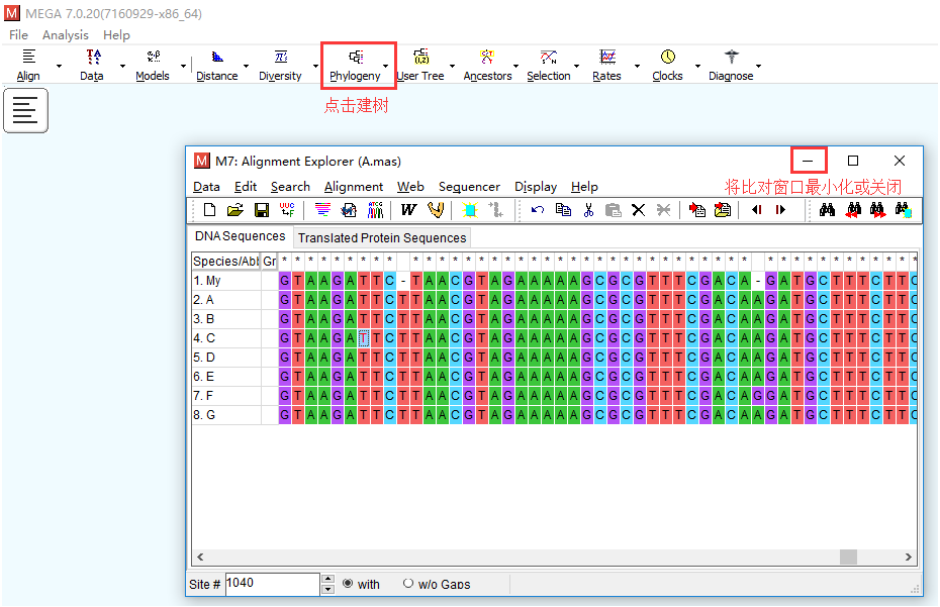


（4）保存比对结果，保存为 mas 文件；



（5）开始建树

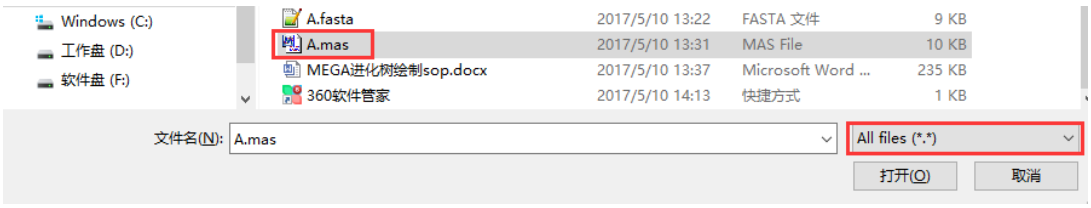
点击 Phylogeny 按钮，点击 Construct/Test Maximum Likehood Tree；

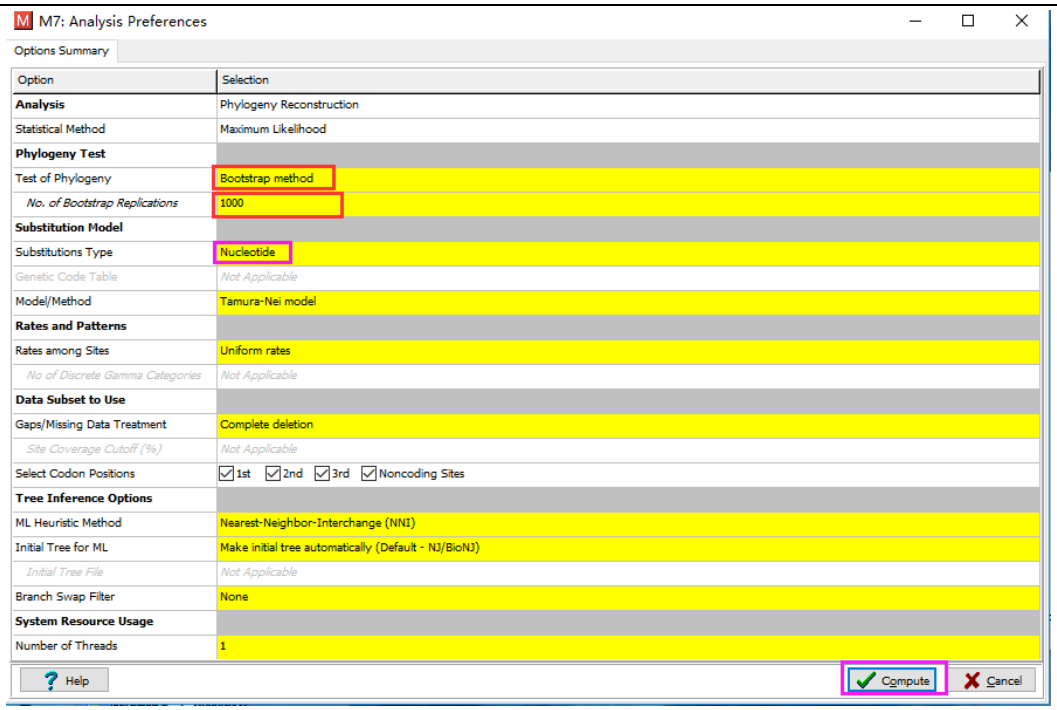


（6）导入上述比对好的序列，选择 Bootstrap，并设置值为 1000；解释如下：

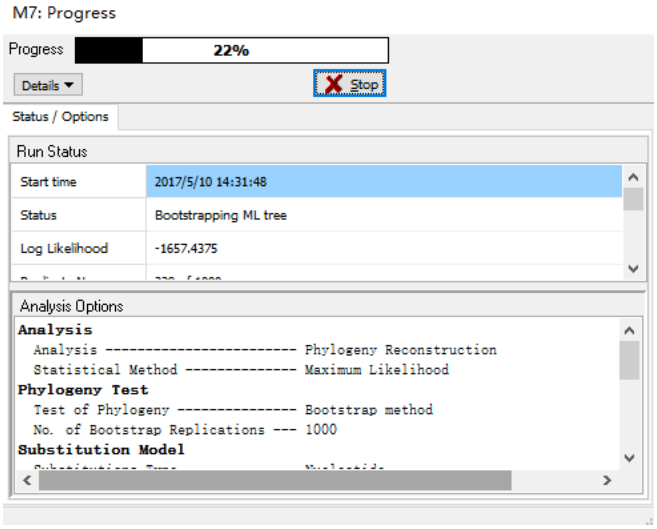
Bootstrap 值即自展值，可用来检验所计算的进化树分支可信度。Bootstrap 几乎是构建系统进化树一个必须的选项。一般 Bootstrap 的值>70%，则认为构建的进化树较为可靠。如果 Bootstrap 的值太低，则有可能进化树的拓扑结构有错误，进化树是不可靠的。

Bootstrap 值是指根据所选的统计计算模型，设定初始值 1000 次，就是把序列的位点都重排，重排后的序列再用相同的办法构树，如此让模型计算并绘制 1000 株系统发育树，这是命令阶段产生的。如果原来树的分枝在重排后构的树中也出现了，就给这个分枝打上 1 分，如果没出现就给 0 分，这样给进化树打分后，每个分枝就都得出分值。系统发育树中每个节点上的数字则代表在命令阶段要求的 1000 次进化树分析中，有多少次。重排的序列有很多组合 ,值越小说明分枝的可信度越低 ,最好根据数据的情况选用不同的构树方法和模型。

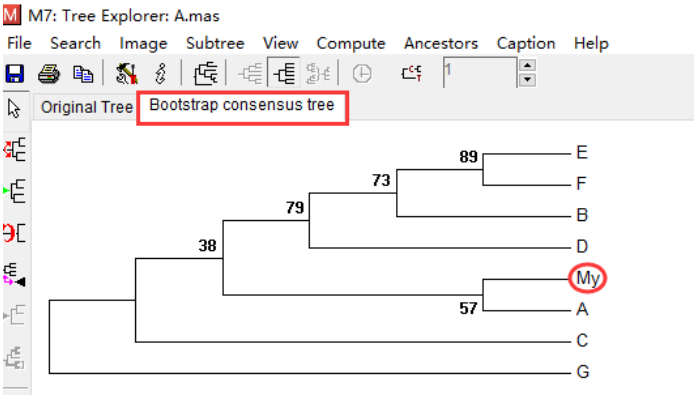




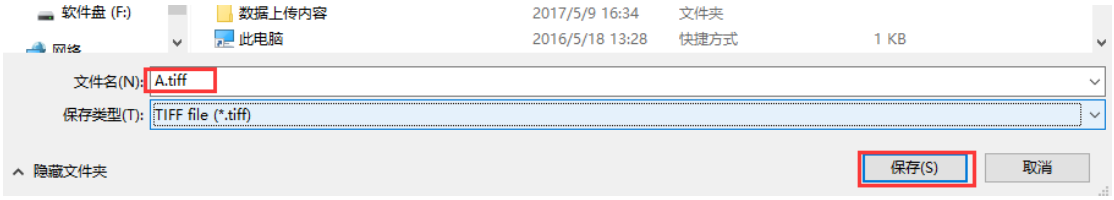
（7）建树过程需要等待，一般来说序列长度越长，导入序列数量越多，建树需要的时间越长；



（8）选择对应方法的图片，检查结果，并导出图片；



点击 Image 按钮，可以保存为 tiff、png、pdf 等图片。



宏基因组测序实用教程分享

LEfSe 在线分析教程

LEfSe（LDA Effect Size）分析，可以用于两个或多个分组之间的比较，从而找到组间有显著性差异的物种（即 biomarker），分析步骤主要分为三步：

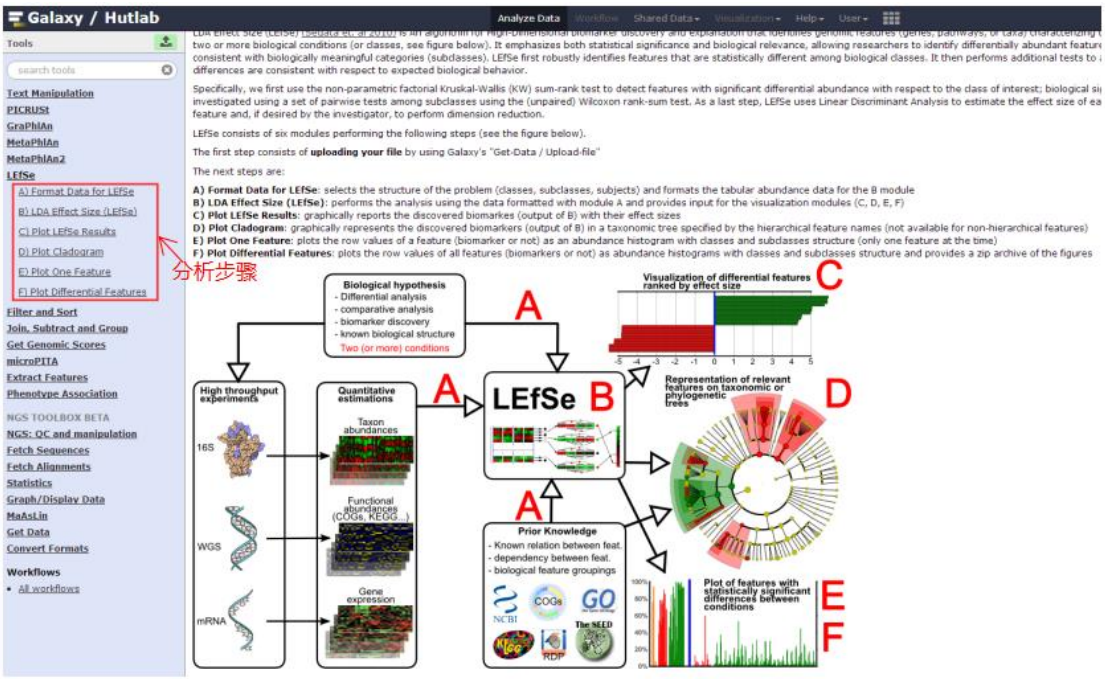
Step1：利用 Kruskal-Wallis 秩和检验检测所有的特征物种，通过检测不同组间的物种丰度差异，获得显著性差异物种。

Step2：再利用 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：最后用线性判别分析（LDA），得到最终的差异物种（即 biomarker）。

LEfSe 在线分析网址：http://huttenhower.sph.harvard.edu/galaxy/

网页界面如下：



1、数据格式

一般微生物物种分析结果中都能得到 OTU 在不同分类水平的物种丰度结果，将数据格式转化为下图所示的网站标准输入格式，即可进行后续的分析。

Bodysite 样品分组行名

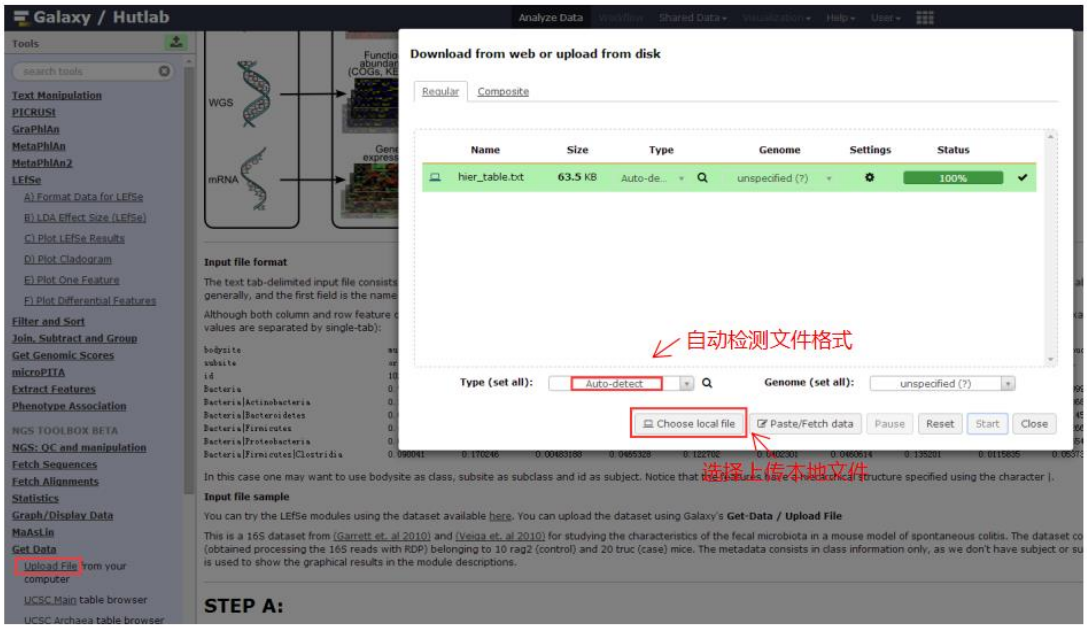
Subsite 亚组行名（一般无该分组）

Id 样品所在行行名

bodysite	mucosal	mucosal	mucosal	mucosal	mucosal	non_mucosal	non_mucosal	non_mucosal
subsite	oral	oral	oral	oral	gut	skin	nasal	skin
id	1023	1023	1672	1876	1672	159005010	1023	1023
Bacteria	0.99999	0.99999	0.999993	0.999989	0.999997	0.999927	0.999977	0.999987
Bacteria|Actinobacteria	0.311037	0.000864363	0.00446132	0.0312045	0.000773642	0.359354	0.761108	0.603002
Bacteria|Bacteroidetes	0.0689602	0.804293	0.00983343	0.0303561	0.859838	0.0195298	0.0212741	0.145729
Bacteria|Firmicutes	0.494223	0.173411	0.715345	0.813046	0.124552	0.177961	0.189178	0.188964
Bacteria|Proteobacteria	0.0914284	0.0180378	0.265864	0.109549	0.00941215	0.430869	0.0225884	0.0532684
Bacteria|Firmicutes|Clostridia	0.090041	0.170246	0.00483188	0.0465328	0.122702	0.0402301	0.0460614	0.135201

2、上传数据

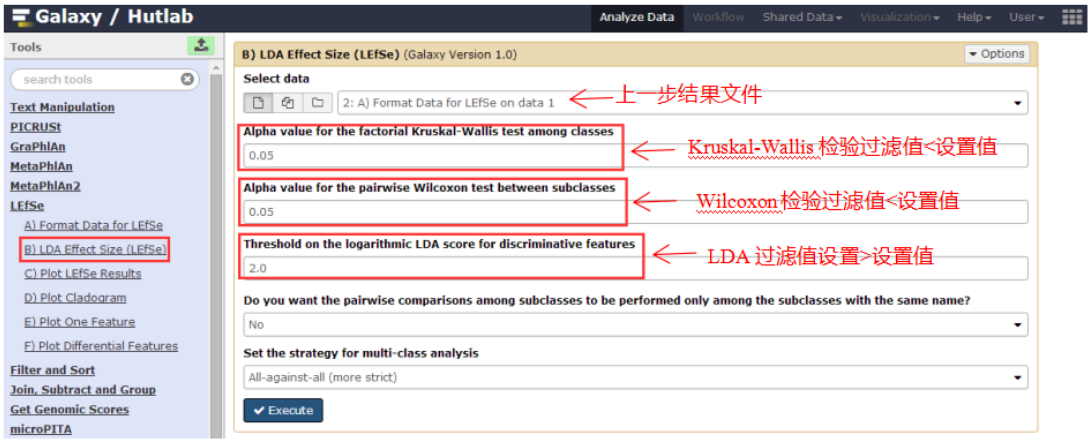
将标准格式的文件上传至网站：（Get Data->Upload File），具体设置如下图：



3、LEfSe 分析-步骤 A）Format Data for LEfSe 格式转化

输出文件格式（该格式文件只是中间文件，具体意义不需要详细理解）如下：

4、LEfSe 分析-步骤 B） LDA Effect Size (LEfSe) 差异计算与统计



输出结果：

```
d__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Lactobacillaceae.g__Lactobacillus.s__uncultured_Lactobacillus_sp     2.96913507904                   -
d__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria.o__Campylobacterales.f__Helicobacteraceae.g__Helicobacter.s__Helicobacter_ganmani          1.90754877721                   -
d__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Prevotellaceae.g__Prevotella.s__Prevotella_unclassified 4.78542282878                   -
d__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Desulfovibrionales.f__Desulfovibrionaceae.g__Desulfovibrio.s__uncultured_Desulfovibrio_sp_       1.90627010197                   -
d__Bacteria.p__CandidatusSaccharibacteria.c__Saccharibacteria_genera_incertae_sedis.o__Saccharibacteria_genera_incertae_sedis.f__Saccharibacteria_genera_incertae_sedis 3.40003712696    groupC
3.09273315396    0.0273237224473
d__Bacteria.p__Lentisphaerae     2.10308505856                   -
d__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae.g__Parasporobacterium.s__Parasporobacterium_unclassified     2.62164049461                   -
d__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae.g__Hespellia.s__uncultured_Hespellia_sp_     2.66322086159                   -
d__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Alteromonadales.f__Shewanellaceae       1.60651878154                   -
d__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Peptostreptococcaceae.g__Filifactor 3.3604965062                   -
d__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae.g__Acetatifactor.s__Acetatifactor_unclassified      3.81513410947                   -
d__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Porphyromonadaceae.g__Odoribacter.s__Odoribacter_unclassified   2.69693290971    groupB  2.73424451532    0.0339856065141
d__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Coriobacteriales.f__Coriobacteriaceae.g__Olsenella    1.77976382916                   -
```

各列意义如下：

第一列：物种信息

第二列：平均丰度最大的 log10 的值，如果平均丰度小于 10 按照 10 来计算

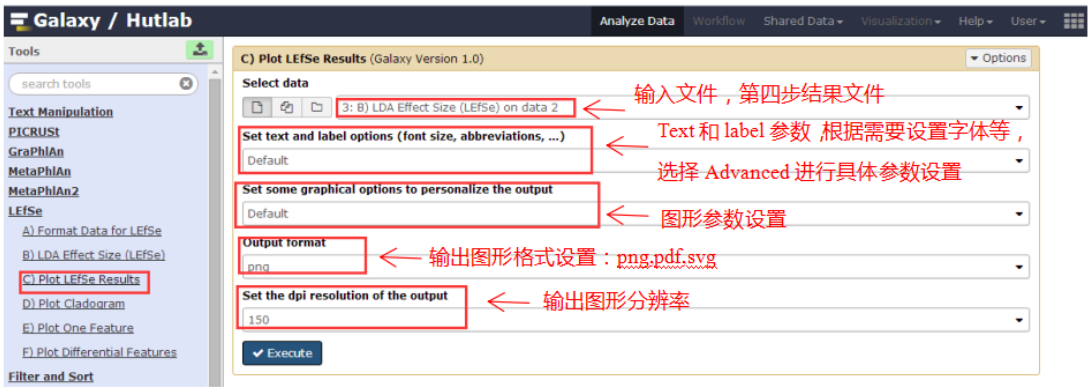
第三列：差异物种富集的组名称

第四列：LDA 值

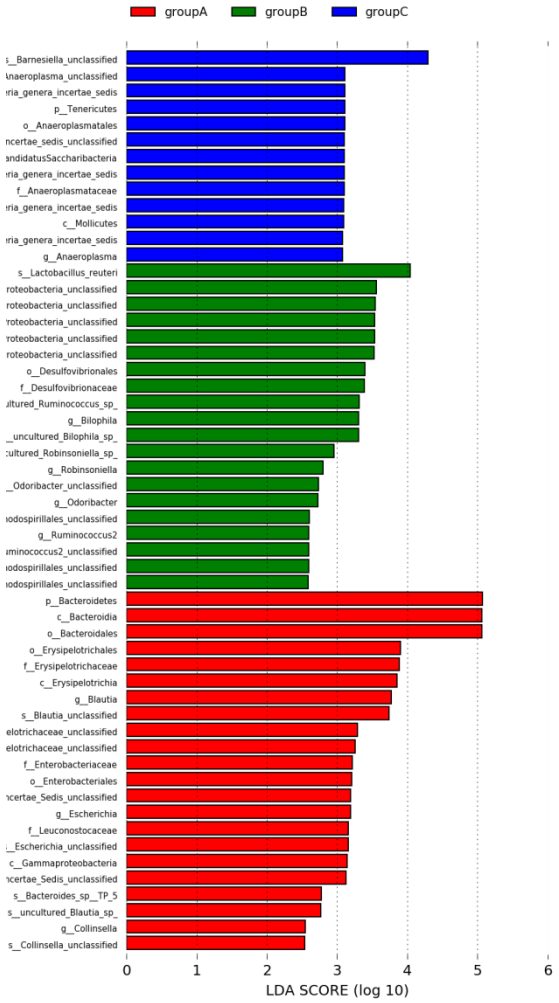
第五列：Kruskal-Wallis 秩和检验的 P 值，如果不是 biomarker，则用‘-’表示

5、LEfSe 分析 C-F 步骤均为作图

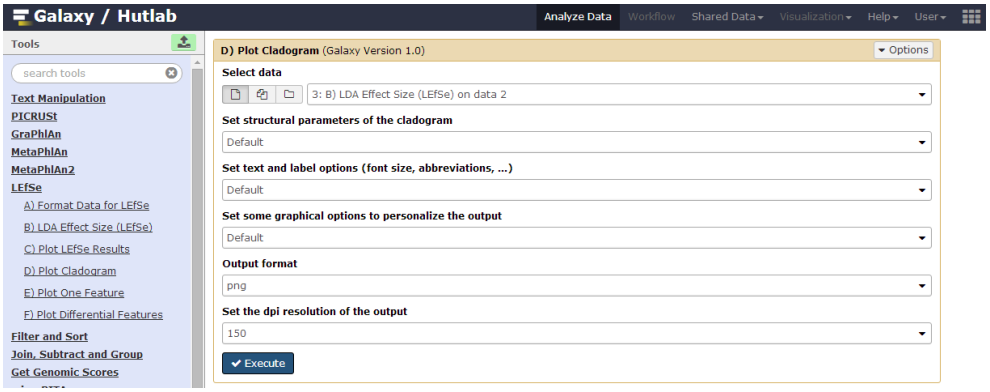
步骤 C）Plot LEfSe Results



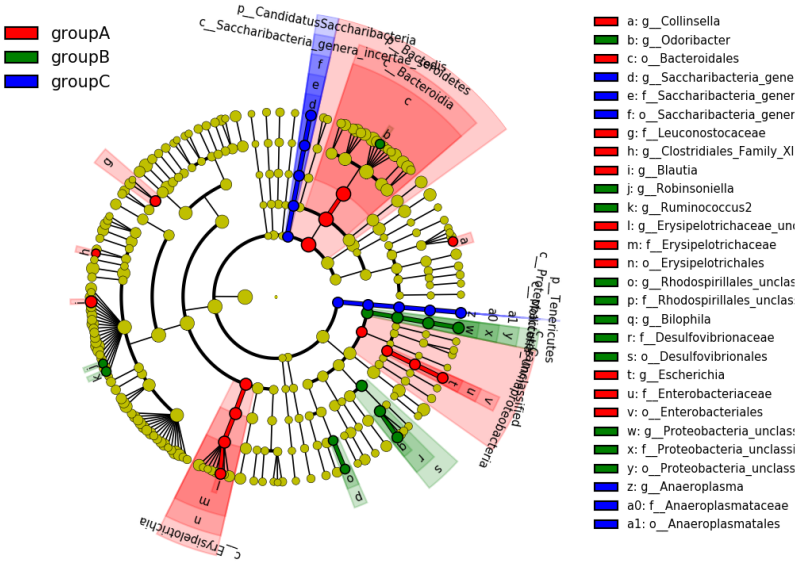
输出图形结果如下（如果差异过多或过少，可重新设置第四步参数并运算，然后作图）：



步骤 D）Plot Cladogram（参数设置与步骤 C 类似）

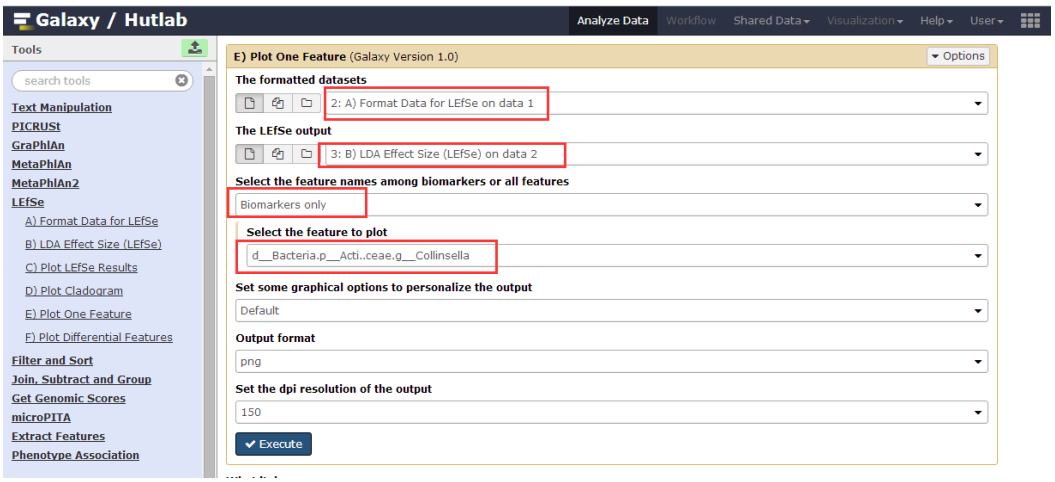


输出结果：

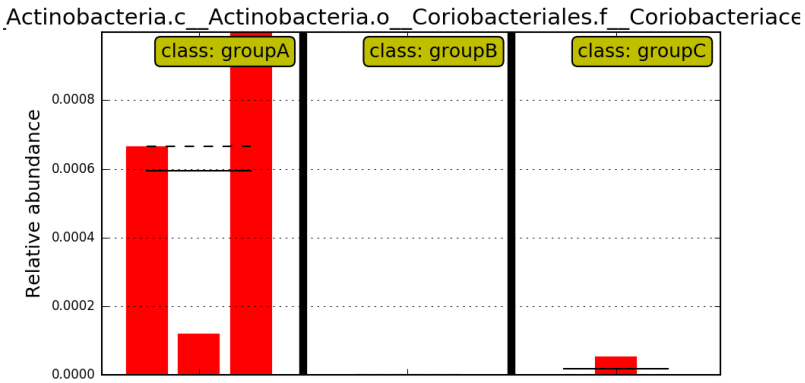


图注：不同圆圈表示不同分类层级，从内至外，依次为门-纲-目-科-属。每个节点表示一种物种，黄色表示该物种在三组中无显著性差异，其他颜色，以 groupA 中的红色为例，如果节点颜色为红色则表示该物种在比较组中有显著性差异 ,且该物种在 groupA 中的丰度更高，其他颜色以此类推。具体差异的门和纲直接在左侧图中标出，其他差异以字母表示，具体代表的物种在右侧标出。

步骤 E）Plot One Feature （画其中一个物种在不同组样品中的柱状图）

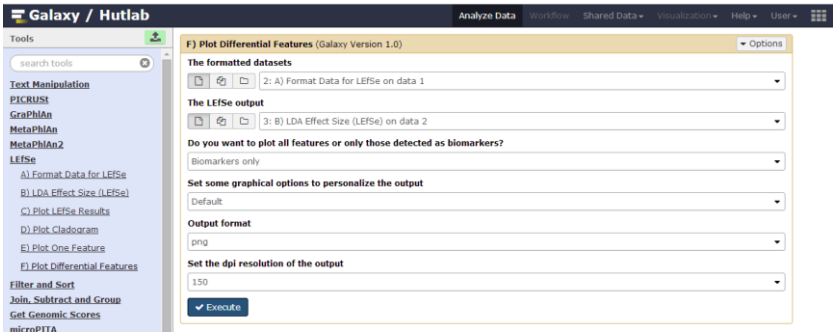


结果展示：



图注：图中不同分组用黑实线隔开，每组柱状图中的实线表示该组样品表达量的平均值，虚线代表该组样品表达量的中位值。

步骤 F ）Plot Differential Features (绘制差异物种或基因柱状图，输入设置与 E 基本类似)



步骤 E 和 F 的区别在于，E 每次只能做一张图，F 可以绘制 biomarker（或所有的物种）柱状图。一般建议跳过步骤 E，直接做步骤 F，步骤 F 的结果一般以压缩包的形式，下载到本地解压后即可查看每个 biomarker 在不同样品中的表达柱状图。

6、结果下载

物种柱状图&饼图绘制教程

在16S分析结果中物种柱状图占有十分重要的作用，不仅可以直观的看到各个样本的物种的组成及比例，还可以看到样本间的物种变化情况；在通常情况下我们是可以使用最常见的excel表格进行绘制的，这样方便老师有个性化需求时可以及时的得到结果，无需长时间的等待及沟通了。

物种柱状图绘制

1、数据来源

我们测序完成后会给到老师完整的结果（summary文件），其中6_taxonomy_community/2_Phylum/1_abundance_stats/all/Phylum_count.xls为例，即是门水平所有样本及物种的丰度信息；其他表格处理雷同；

Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
Firmicutes	4720	3600	4030	6770	6360	5230	4170	5490	5840
Bacteroidetes	4710	5480	5470	2710	2710	3510	5470	4020	3690
Proteobacteria	217	539	173	192	622	811	38.5	86.8	198
Bacteria_unclassified	291	330	260	311	238	430	284	308	228
Actinobacteria	51.5	44.6	32	17.8	72	12.1	4.16	15.2	3.24
Tenericutes	0.61	0.60	39.4	0	0	0	3.12	59.1	19.4
Candidatus Saccharibacteria	0	0.60	1.5	3.9	1.83	9.64	32.8	25.5	17.1
Cyanobacteria	7.87	1.19	0	1.11	0.61	1.21	0	0	5.55
Lentisphaerae	1.21	0.60	2	0	0	0	0	0	0
Fusobacteria	1.21	0	0	0.56	0	0.60	0.52	0	0.46
Candidatus_Saccharibacteria	0	0	0	0	0	0	0	1.08	0
Deferribacteres	0	0.60	0	0	0	0.60	0	0	0

2、数据处理：相对丰度计算及删选

（1）各物种在各个样本中的相对丰度计算

第一步：求和

Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
p__Proteobacteria	32195	23749	18439	14533	24064	23793	26339	23456	16579
p__Actinobacteria	1753	5975	7442	670	772	1600	76	85	257
p__Firmicutes	1912	4000	1176	509	279	650	238	190	6979
p__Deinococcus-Thermus	12	47	5	0	18	13	832	149	184
p__Bacteroidetes	30	301	46	28	20	82	158	124	221
p__Fusobacteria	7.00	373.00	5	16	31	2	0	6	62
p__Cyanobacteria	45	113.00	5	17	5	4	27	0	2
p__Candidatus Saccharibacteria	3	77	0	2	5.00	0	1	0	2
p__Spirochaetes	0	21.00	1	0	0	0	0	0	2
sample_sum	35957	34656	27119	15775	25194	26144	27671	24010	24288

第二步：计算相对丰度，每个物种在各个样本中的表达量/该物种序列总和

Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
p__Proteobacteria	32195	23749	18439	14533	24064	23793	26339	23456	16579
p__Actinobacteria	1753	5975	7442	670	772	1600	76	85	257
p__Firmicutes	1912	4000	1176	509	279	650	238	190	6979
p__Deinococcus-Thermus	12	47	5	0	18	13	832	149	184
p__Bacteroidetes	30	301	46	28	20	82	158	124	221
p__Fusobacteria	7.00	373.00	5	16	31	2	0	6	62
p__Cyanobacteria	45	113.00	5	17	5	4	27	0	2
p__Candidatus Saccharibacteria	3	77	0	2	5.00	0	1	0	2
p__Spirochaetes	0	21.00	1	0	0	0	0	0	2
sample_sum	35957	34656	27119	15775	25194	26144	27671	24010	24288

每个样本中各个物种的表达量除以该物种序列总和

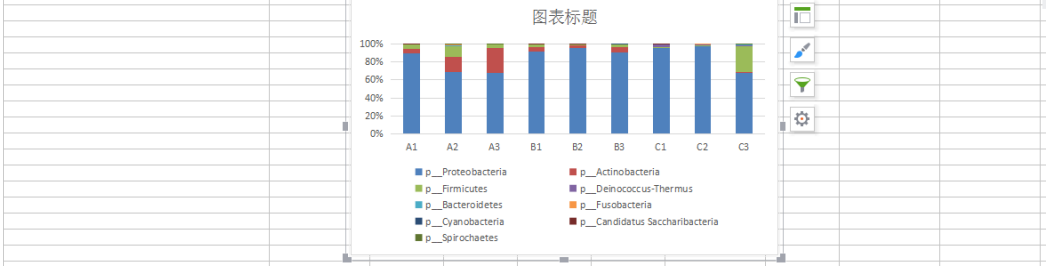
Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
p__Proteobacteria	0.895375031	0.685278	0.679929	0.921268	0.955148	0.910074969	0.951863	0.976926281	0.682600461
p__Actinobacteria	0.048752677	0.172409	0.27442	0.042472	0.030642	0.06119951	0.002747	0.003540192	0.010581357
p__Firmicutes	0.053174625	0.11542	0.043364	0.032266	0.011074	0.024862301	0.008601	0.007913369	0.287343544
p__Deinococcus-Thermus	0.000333732	0.001356	0.000184	0	0.000714	0.000497246	0.030068	0.006205748	0.007575758
p__Bacteroidetes	0.00083433	0.008685	0.001696	0.001775	0.000794	0.003136475	0.00571	0.005164515	0.009099144
p__Fusobacteria	0.000194677	0.010763	0.000184	0.001014	0.00123	7.64994E-05	0	0.000249896	0.002552701
p__Cyanobacteria	0.001251495	0.003261	0.000184	0.001078	0.000198	0.000152999	0.000976	0	8.23452E-05
p__Candidatus Saccharibacteria	8.3433E-05	0.002222	0	0.000127	0.000198	0	3.61E-05	0	8.23452E-05
p__Spirochaetes	0	0.000606	3.69E-05	0	0	0	0	0	8.23452E-05

若是其他水平（科属种）时，物种可能很多，此时若是将所有物种均在图上展示结果不太美观，一般来说我们会选取丰度前 20 的物种进行绘图展示，我们提供的结果已经是按照丰度从高到低进行排序了，筛选可以直接选取前 20 行，将其余物种变为others，将其余物种的各个样本中丰度进行相加即可。

3、作图

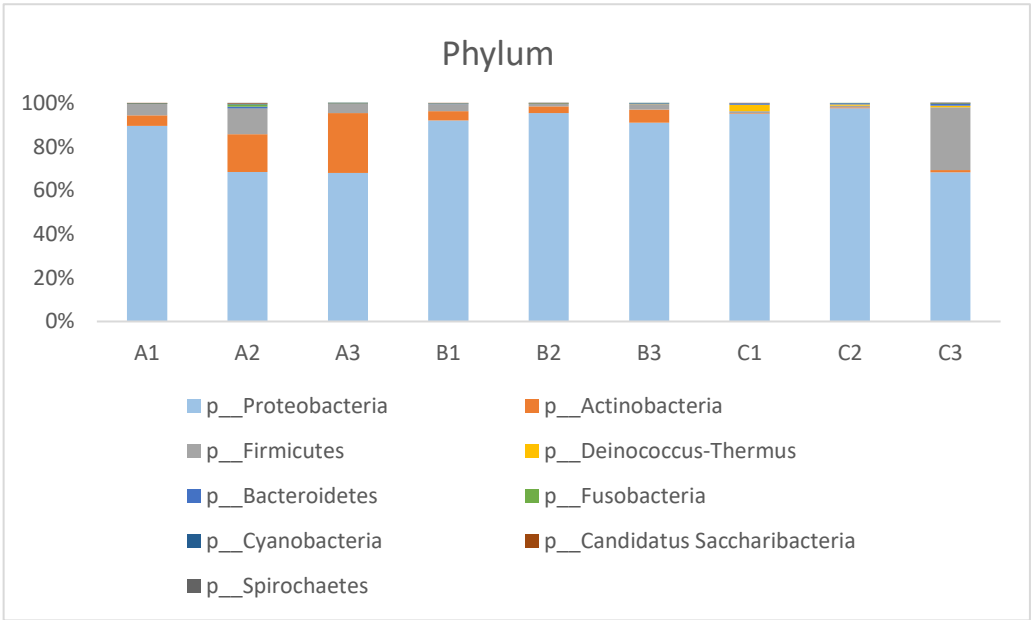
选中作图数据区域，选择excel中插入功能，选择柱状图；

Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
p__Proteobacteria	0.895375031	0.685278	0.679929	0.921268	0.955148	0.910074969	0.951863	0.976926281	0.682600461
p__Actinobacteria	0.048752677	0.172409	0.27442	0.042472	0.030642	0.06119951	0.002747	0.003540192	0.010581357
p__Firmicutes	0.053174625	0.11542	0.043364	0.032266	0.011074	0.024862301	0.008601	0.007913369	0.287343544
p__Deinococcus-Thermus	0.000333732	0.001356	0.000184	0	0.000714	0.000497246	0.030068	0.006205748	0.007575758
p__Bacteroidetes	0.00083433	0.008685	0.001696	0.001775	0.000794	0.003136475	0.00571	0.005164515	0.009099144
p__Fusobacteria	0.000194677	0.010763	0.000184	0.001014	0.00123	7.64994E-05	0	0.000249896	0.002552701
p__Cyanobacteria	0.001251495	0.003261	0.000184	0.001078	0.000198	0.000152999	0.000976	0	8.23452E-05
p__Candidatus Saccharibacteria	8.3433E-05	0.002222	0	0.000127	0.000198	0	3.61E-05	0	8.23452E-05
p__Spirochaetes	0	0.000606	3.69E-05	0	0	0	0	0	8.23452E-05



4、图形调整及保存

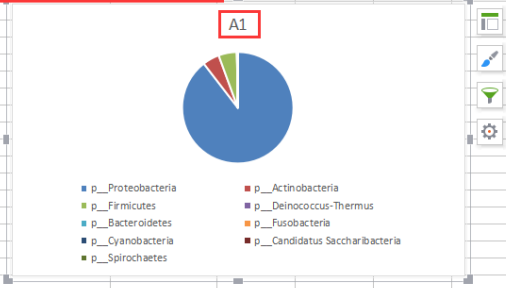
可以改变颜色，标题等等，看老师自己的需求进行调整；



单样本饼图绘制

与柱状图类似，使用上述筛选过的数据，选择如下图红框中数据，插入图表选择饼

Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
p__Proteobacteria	0.895375031	0.685278	0.679929	0.921268	0.955148	0.910074969	0.951863	0.976926281	0.682600461
p__Actinobacteria	0.048752677	0.172409	0.27442	0.042472	0.030642	0.06119951	0.002747	0.003540192	0.010581357
p__Firmicutes	0.053174625	0.11542	0.043364	0.032266	0.011074	0.024862301	0.008601	0.007913369	0.287343544
p__Deinococcus-Thermus	0.000333732	0.001356	0.000184	0	0.000714	0.000497246	0.030068	0.006205748	0.007575758
p__Bacteroidetes	0.00083433	0.008685	0.001696	0.001775	0.000794	0.003136475	0.00571	0.005164515	0.009099144
p__Fusobacteria	0.000194677	0.010763	0.000184	0.001014	0.00123	7.64994E-05	0	0.000249896	0.002552701
p__Cyanobacteria	0.001251495	0.003261	0.000184	0.001078	0.000198	0.000152999	0.000976	0	8.23452E-05
p__Candidatus Saccharibacteria	8.3433E-05	0.002222	0	0.000127	0.000198	0	3.61E-05	0	8.23452E-05
p__Spirochaetes	0	0.000606	3.69E-05	0	0	0	0	0	8.23452E-05



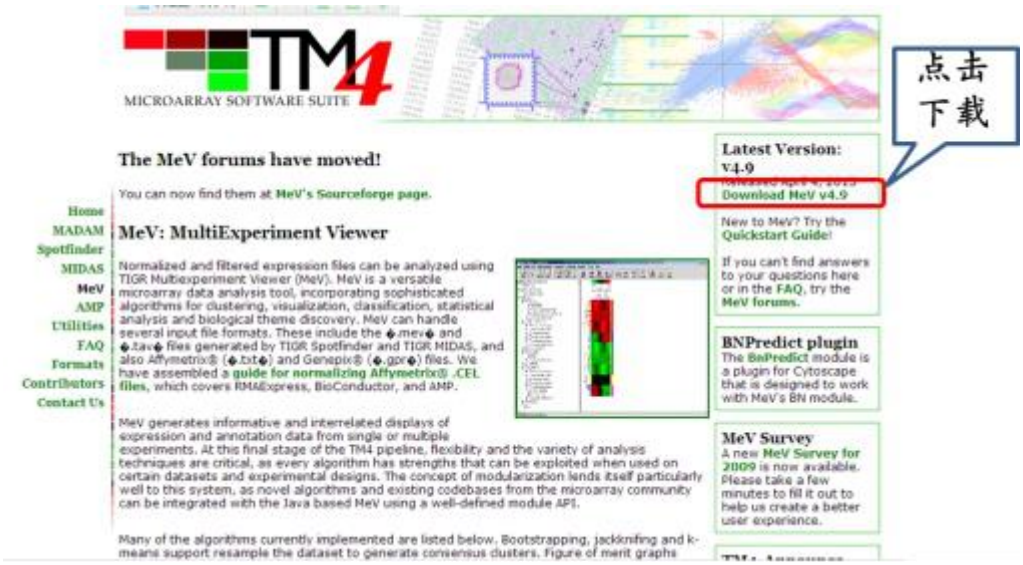
图，即可绘制出单样本饼图，其他样本以此类推。

对上述图形进行编辑修改后保存即可。

TMEV 绘制热图教程

一、软件下载

下载地址：http://www.tm4.org/mev.html



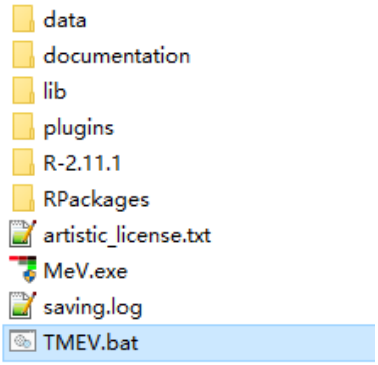
二、安装和启动

在启动前，电脑操作系统需要安装好java运行环境

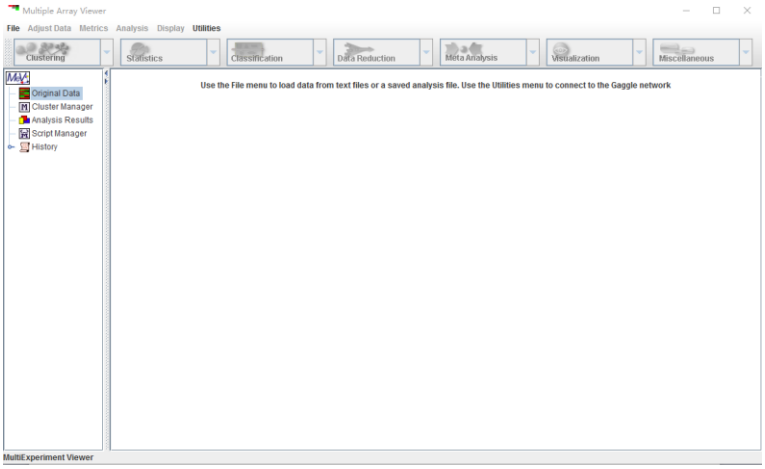
下载网址：http://www.java.com/zh_CN/

下载完成后，双击该文件选择安装即可

解压下载的压缩包，双击TMEV.bat文件启动TMEV软件



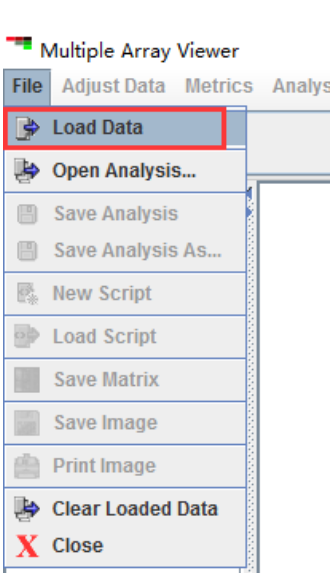
出现以下界面时，启动完成

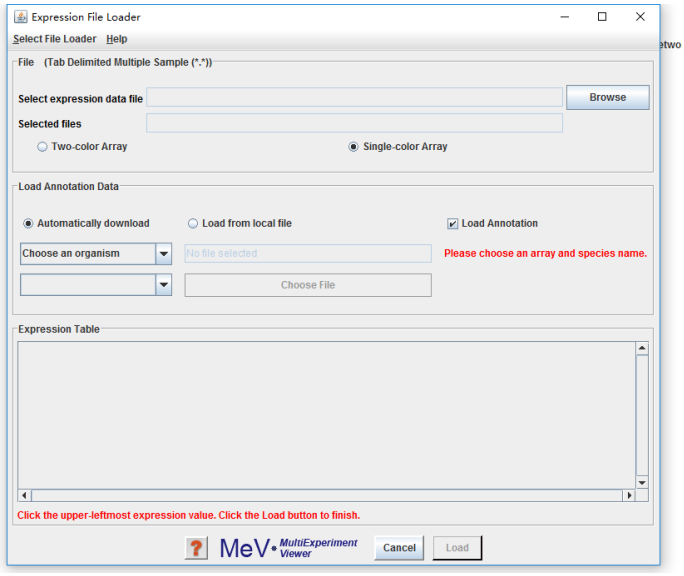


三、Heatmap图绘制

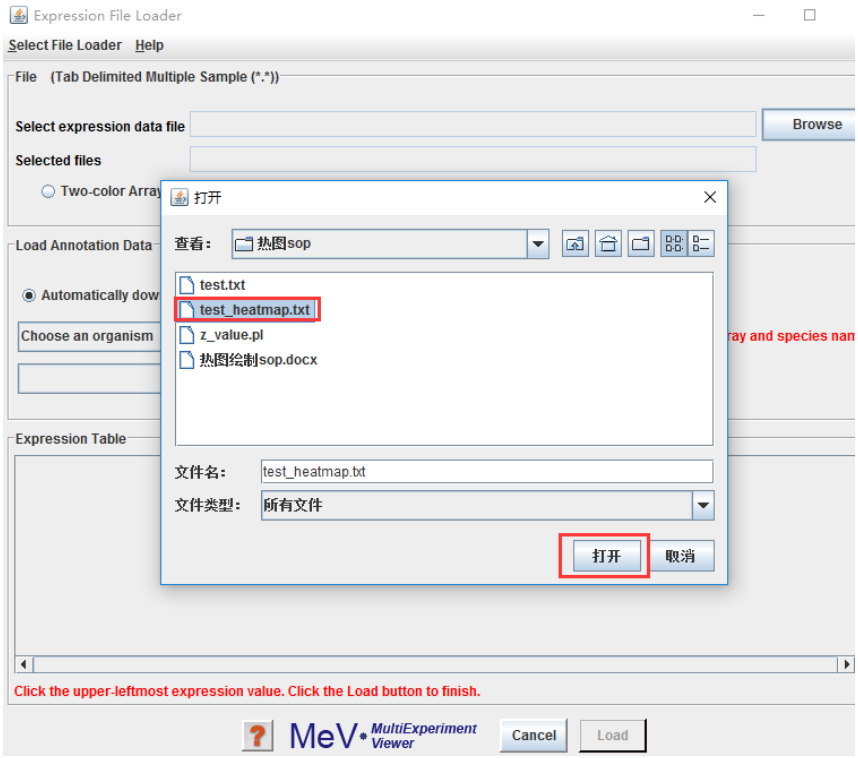
（1）导入数据

File→Load Data→Browse





选择输入文件，单击按钮“打开”

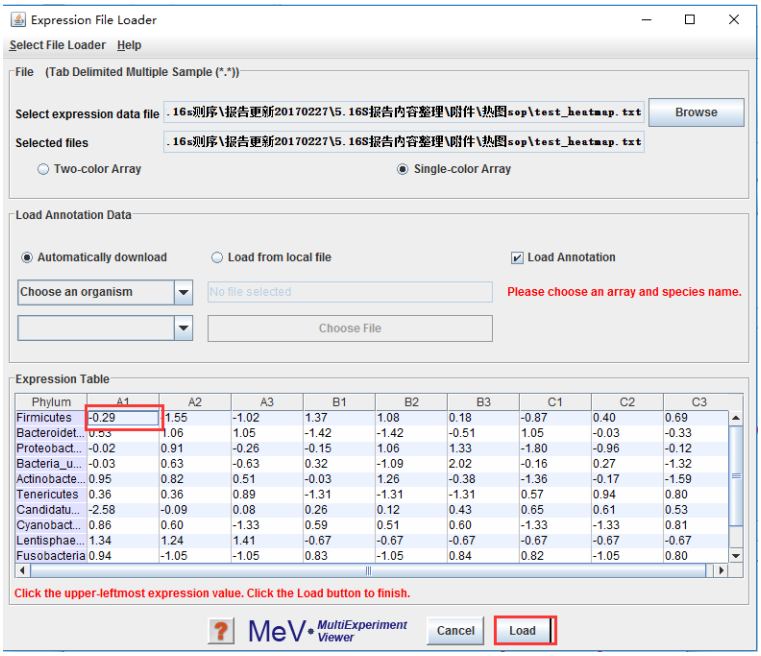


这里需要注意的是，数据必须是txt格式的文本文档，其中的数据格式如下：

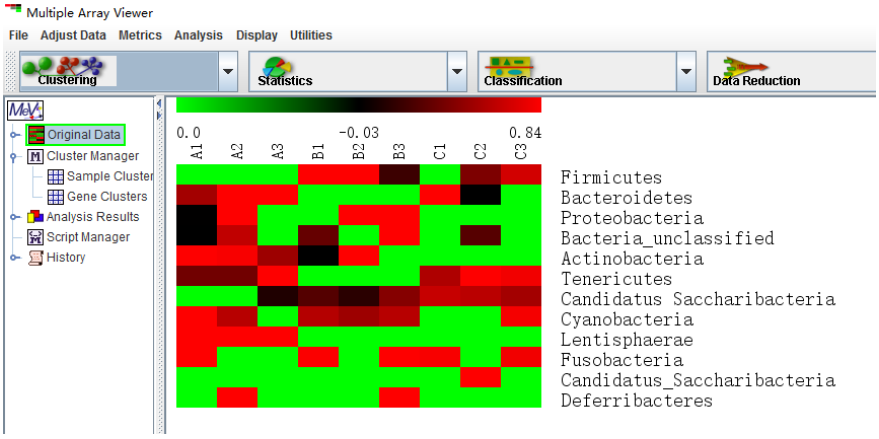
Phylum 物种分类	A1	A2	A3	B1	B2	B3	C1	C2	C3	样本名称
Firmicutes	-0.29	-1.55	-1.02	1.37	1.08	0.18	-0.87	0.4	0.69	
Bacteroidetes	0.53	1.06	1.05	-1.42	-1.42	-0.51	1.05	-0.03	-0.33	
Proteobacteria	-0.02	0.91	-0.26	-0.15	1.06	1.33	-1.8	-0.96	-0.12	
Bacteria_unclassified	-0.03	0.63	-0.63	0.32	-1.09	2.02	-0.16	0.27	-1.32	
Actinobacteria	0.95	0.82	0.51	-0.03	1.26	-0.38	-1.36	-0.17	-1.59	
Tenericutes	0.36	0.36	0.89	-1.31	-1.31	-1.31	0.57	0.94	0.8	
Candidatus Saccharibacteria	-2.58	-0.09	0.08	0.26	0.12	0.43	0.65	0.61	0.53	
Cyanobacteria	0.86	0.6	-1.33	0.59	0.51	0.6	-1.33	-1.33	0.81	
Lentisphaerae	1.34	1.24	1.41	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67	
Fusobacteria	0.94	-1.05	-1.05	0.83	-1.05	0.84	0.82	-1.05	0.8	
Candidatus_Saccharibacteria	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33	2.67	-0.33	
Deferribacteres	-0.5	1.76	-0.5	-0.5	-0.5	1.76	-0.5	-0.5	-0.5	

Z-value值是通过每个物种在样本中的表达量转换而成，具体方法请参看附录。

选择之前准备好的文件导入数据，然后点击“Load”即可。这里需要特别说明一下的是，鼠标需要放到导入文件中的第一个Zvalue值上面，代表数据由此开始进行图片的绘制。

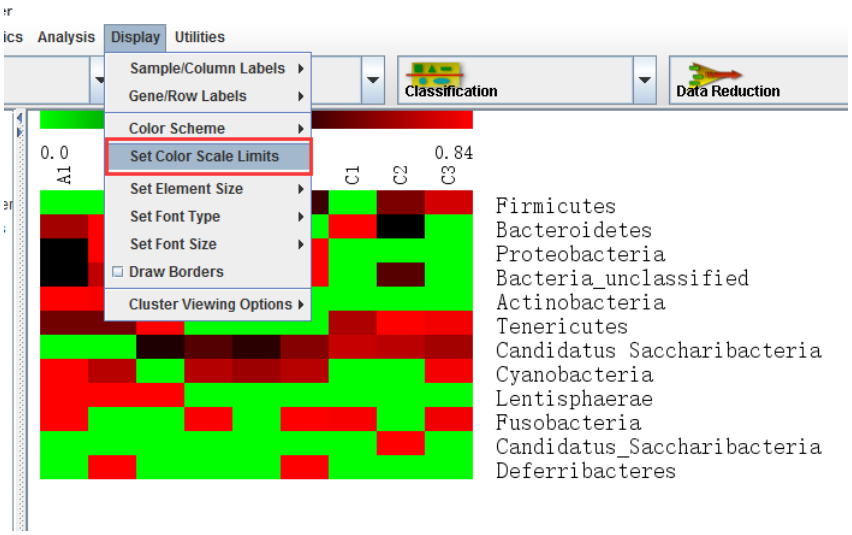


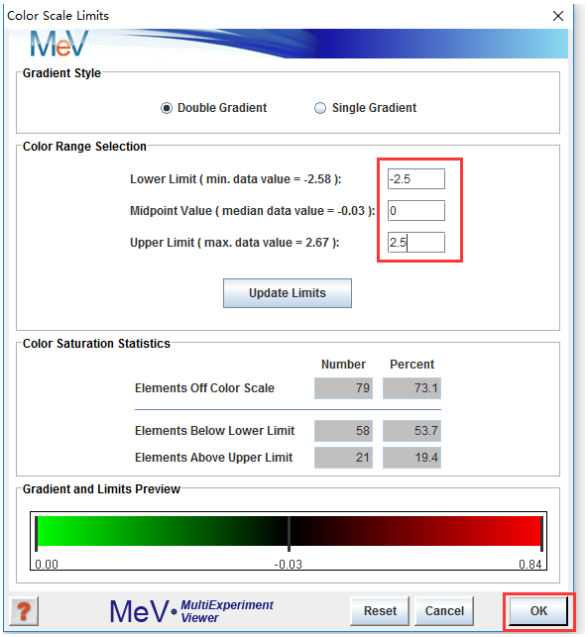
数据导入完成后出现如下图片：



色度范围调整：Display→Set Color Scale Limits

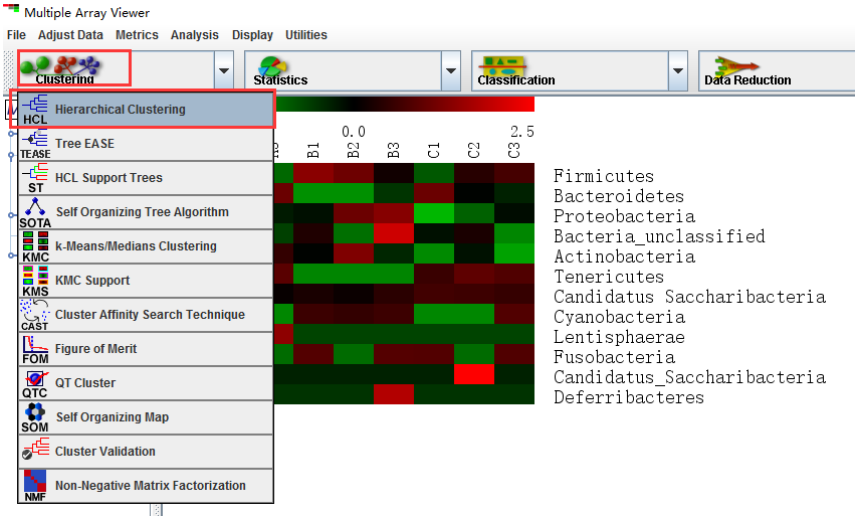
一般会根据最大值和最小值来设置，将 Low Limit 与 Upper Limit数值设置对称，然后再填写位于最大值和最小值之间的中值即可。



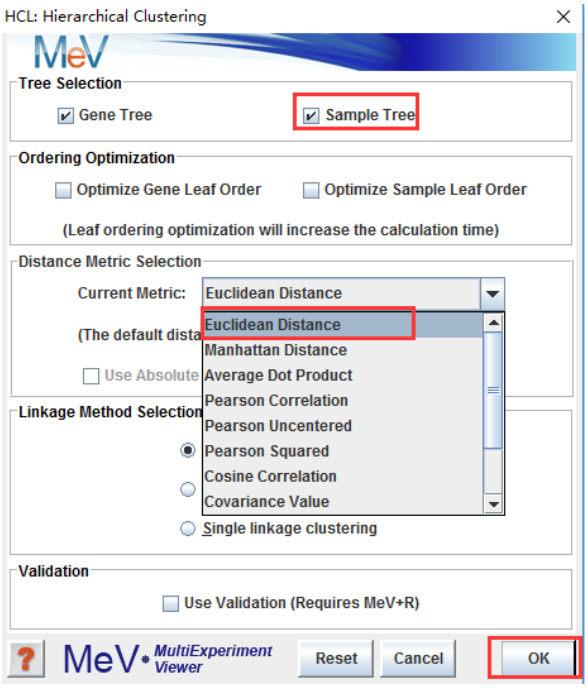


（2）聚类分析

Clustering→HCL (Hierarchical Clustering)

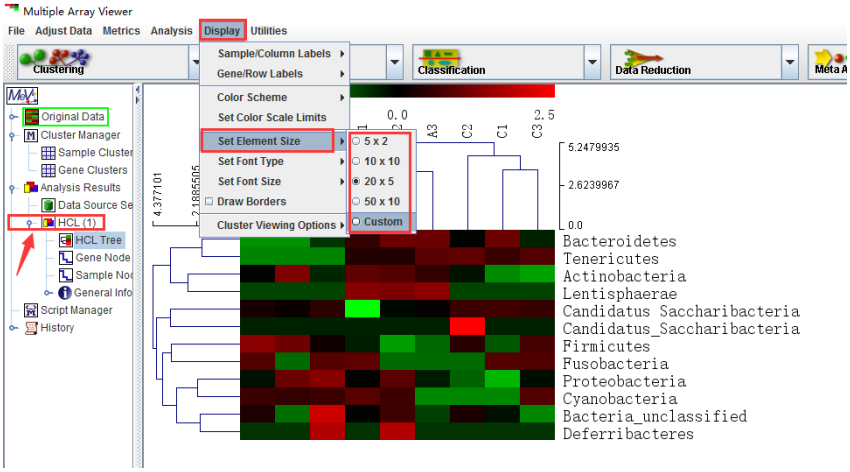


参数设置：Sample Tree用来选择是否需要进行样本聚类，选择Euclidean Distance（欧氏距离），点击”OK”



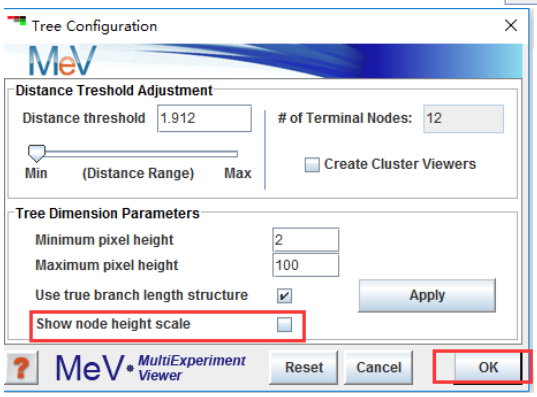
（3）图形大小调整：Display→Set Color Element Size

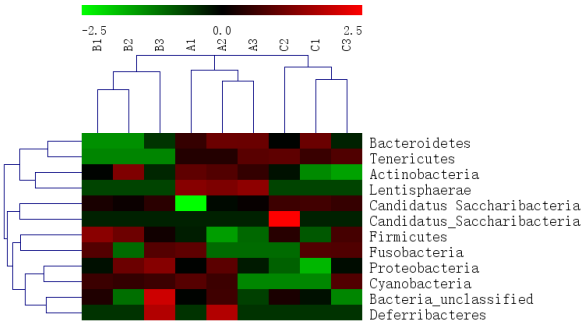
可以选择软件中的尺寸，也可以选择Custom自己选择宽度和高度



去除两端标尺

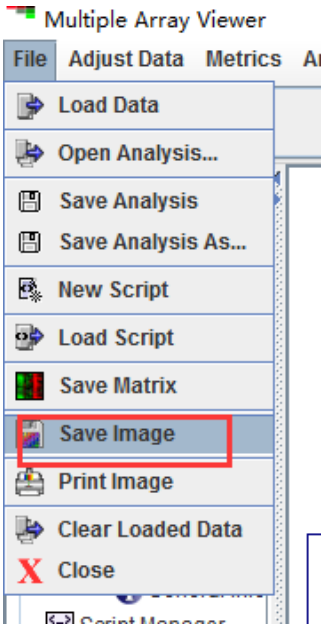
点击鼠标右击，分别点击Gene Tree 和Sample Tree，去除height scale选项，点确定，得到去除标尺后的最终结果。

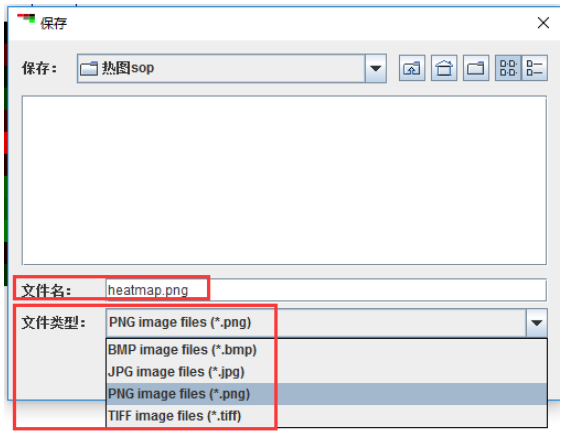




（4）. 图片保存：File→Save image

选择自己需要保存的路径及图片格式





Z 值计算（以门水平为例）：

数据来源：/mnt/data1/Customers/16S_ITS/16S_demo_V3V4/Output/summary/6_taxonomy_community/2_Phylum/1_abundance_stats/all/Phylum_abund_all.xlsx

计算公式：

$$Z_{samplei} = \frac{log_2(SignalSamplei) - Mean\ (log_2(Signal)\ of\ all\ Samples)}{Standard\ deviation\ (log_2(Signal)\ of\ all\ Samples)}$$

计算过程：

对每个样本的表达量进行 log2 的对数计算

对每个物种的所有 log2 后的值（每一行数据）进行平均值（mean）和标准差（stdev）的计算

如下图所示：

Phylum	A1	A2	A3	B1	B2	B3	C1	C2	C3
Firmicutes	4720	3600	4030	6770	6360	5230	4170	5490	5840
Bacteroidetes	4710	5480	5470	2710	2710	3510	5470	4020	3690
Proteobacteria	217	539	173	192	622	811	38.5	86.8	198
Bacteria_unclassified	291	330	260	311	238	430	284	308	228
Actinobacteria	51.5	44.6	32	17.8	72	12.1	4.16	15.2	3.24
Tenericutes	0.61	0.6	39.4	0	0	0	3.12	59.1	19.4
Candidatus Saccharibacteria	0	0.6	1.5	3.9	1.83	9.64	32.8	25.5	17.1
Cyanobacteria	7.87	1.19	0	1.11	0.61	1.21	0	0	5.55
Lentisphaerae	1.21	0.6	2	0	0	0	0	0	0
Fusobacteria	1.21	0	0	0.56	0	0.6	0.52	0	0.46
Candidatus_Saccharibacteria	0	0	0	0	0	0	0	1.08	0
Deferribacteres	0	0.6	0	0	0	0.6	0	0	0

每个值取log2的对数，原始值为0，仍取0

Phylum	logA1	logA2	logA3	logB1	logB2	logB3	logC1	logC2	logC3
Firmicutes	12.20	11.81	11.98	12.72	12.63	12.35	12.03	12.42	12.51
Bacteroidetes	12.20	12.42	12.42	11.40	11.40	11.78	12.42	11.97	11.85
Proteobacteria	7.76	9.07	7.43	7.58	9.28	9.66	5.27	6.44	7.63
Bacteria_unclassified	8.18	8.37	8.02	8.28	7.89	8.75	8.15	8.27	7.83
Actinobacteria	5.69	5.48	5.00	4.15	6.17	3.60	2.06	3.93	1.70
Tenericutes	-0.71	-0.74	5.30	0.00	0.00	0.00	1.64	5.89	4.28
Candidatus Saccharibacteria	0.00	-0.74	0.58	1.96	0.87	3.27	5.04	4.67	4.10
Cyanobacteria	2.98	0.25	0.00	0.15	-0.71	0.28	0.00	0.00	2.47
Lentisphaerae	0.28	-0.74	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Fusobacteria	0.28	0.00	0.00	-0.84	0.00	-0.74	-0.94	0.00	-1.12
Candidatus_Saccharibacteria	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00
Deferribacteres	0.00	-0.74	0.00	0.00	0.00	-0.74	0.00	0.00	0.00

代入上述公式进行计算

Phylum	logA1	logA2	logA3	logB1	logB2	logB3	logC1	logC2	logC3	Mean	stdev
Firmicutes	12.20	11.81	11.98	12.72	12.63	12.35	12.03	12.42	12.51	12.30	0.31
Bacteroidetes	12.20	12.42	12.42	11.40	11.40	11.78	12.42	11.97	11.85	11.98	0.41
Proteobacteria	7.76	9.07	7.43	7.58	9.28	9.66	5.27	6.44	7.63	7.79	1.40
Bacteria_unclassified	8.18	8.37	8.02	8.28	7.89	8.75	8.15	8.27	7.83	8.19	0.27
Actinobacteria	5.69	5.48	5.00	4.15	6.17	3.60	2.06	3.93	1.70	4.20	1.57
Tenericutes	-0.71	-0.74	5.30	0.00	0.00	0.00	1.64	5.89	4.28	1.74	2.68
Candidatus Saccharibacteria	0.00	-0.74	0.58	1.96	0.87	3.27	5.04	4.67	4.10	2.20	2.14
Cyanobacteria	2.98	0.25	0.00	0.15	-0.71	0.28	0.00	0.00	2.47	0.60	1.24
Lentisphaerae	0.28	-0.74	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.45
Fusobacteria	0.28	0.00	0.00	-0.84	0.00	-0.74	-0.94	0.00	-1.12	-0.37	0.53
Candidatus_Saccharibacteria	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.01	0.04
Deferribacteres	0.00	-0.74	0.00	0.00	0.00	-0.74	0.00	0.00	0.00	-0.16	0.32

Phylum	Z(A1)	Z(A2)	Z(A3)	Z(B1)	Z(B2)	Z(B3)	Z(C1)	Z(C2)	Z(C3)
Firmicutes	-0.29	-1.55	-1.02	1.37	1.08	0.18	-0.87	0.40	0.69
Bacteroidetes	0.53	1.06	1.05	-1.42	-1.42	-0.51	1.05	-0.03	-0.33
Proteobacteria	-0.02	0.91	-0.26	-0.15	1.06	1.33	-1.80	-0.96	-0.12
Bacteria_unclassified	-0.03	0.63	-0.63	0.32	-1.09	2.02	-0.16	0.27	-1.32
Actinobacteria	0.95	0.82	0.51	-0.03	1.26	-0.38	-1.36	-0.17	-1.59
Tenericutes	-0.91	-0.92	1.33	-0.65	-0.65	-0.65	-0.04	1.55	0.95
Candidatus Saccharibacteria	-1.02	-1.37	-0.75	-0.11	-0.62	0.50	1.32	1.16	0.89
Cyanobacteria	1.91	-0.28	-0.48	-0.36	-1.06	-0.26	-0.48	-0.48	1.50
Lentisphaerae	0.48	-1.79	2.11	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13
Fusobacteria	1.23	0.71	0.71	-0.88	0.71	-0.69	-1.08	0.71	-1.42
Candidatus_Saccharibacteria	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33	2.67	-0.33
Deferribacteres	0.50	-1.76	0.50	0.50	0.50	-1.76	0.50	0.50	0.50

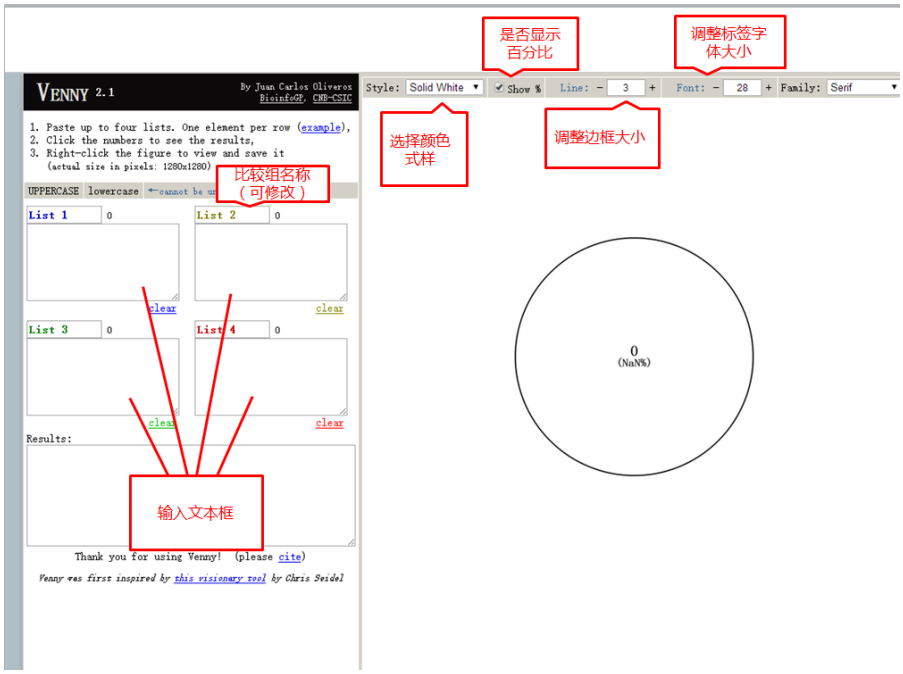
代入公式计算

将最终的结果黏贴到一个新的 txt 文件中，用于作图。

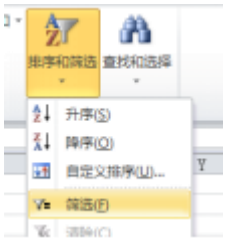
在线 Venn 图绘制

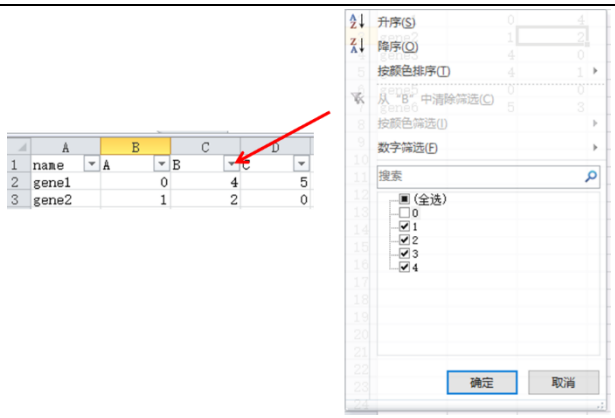
Venn 图绘制网址：http://bioinfogp.cnb.csic.es/tools/venny/index.html

页面介绍：

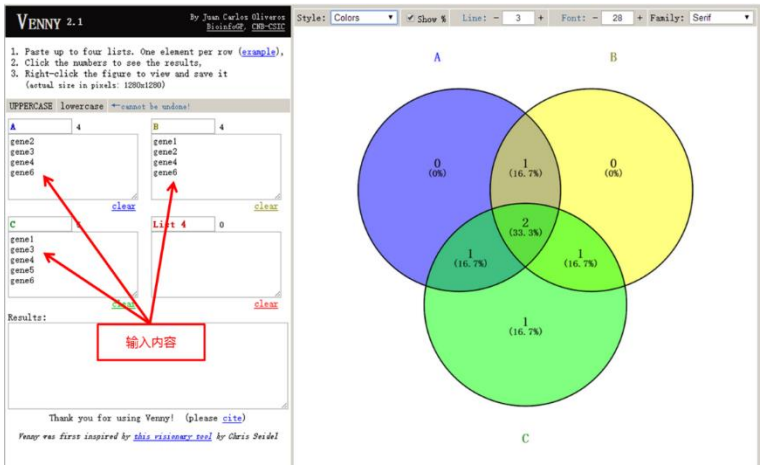


第一步：准备输入内容。

使用 Excel 表格的筛选功能（界面右上角）  ，点击样本名称栏里出现的下拉三角，可选择筛选出表达量不为 0 的 gene（或其他名称）。



每个样本筛选出的 gene 名称（其他名称）复制到不同的文本框内。如下图所示：

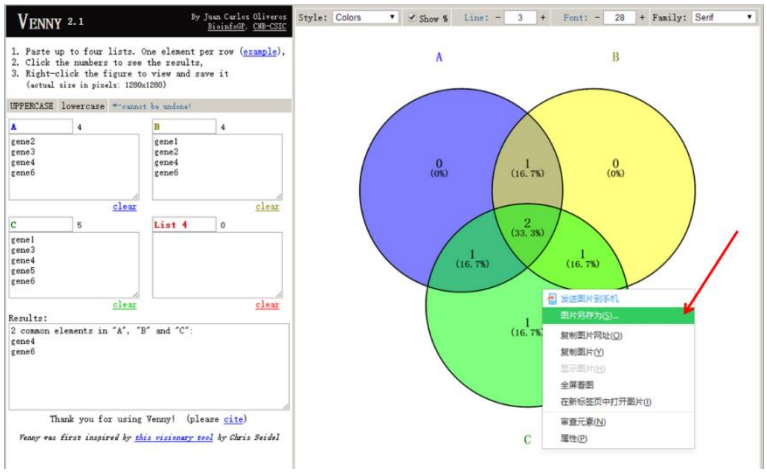


第二步：

调整图片。通过 venn 图上方的选项，调整颜色式样，是否显示百分比等，将图片修改到满意的样子为止。

第三步：

保存图片。将鼠标停留在 venn 图上，右击鼠标，选择“图片另存为”，保存到本地。一张 venn 图就制作完成了。



Tips：1.网页版制作的 venn 图，最多只有 4 个比较组；

2.输入文本时要注意，一行只能输入一个名称

3.鼠标点击 venn 图中的数字 ,网页左边“Results”文本框内 ,会告诉你这个数字的意义。