# Beta 多样性 PCoA/NMDS 排序分析

文涛 南京农业大学

2020/6/17

## Beta 多样性与 PCoA 和 NMDS 排序

作为高通量测序的代表之一，扩增子目早已成为表征微生物群落的主流手段，在后续的数据处理，生物信息学分析中最基础也是最重要的分析就是群落多样性分析（alpha 多样性和 beta 多样性）。今天我们来学习的就是群落 bate 多样性分析中最重要-非限制性排序。

### 基本概念

#### *β 多样性(Beta diversity)*

β 多样性又称生境间的多样性(between-habitat diversity)，是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，用于研究群落之间的种多度关系。Beta 多样性本身代表了一个复杂的问题，可以被视为物种更替（物种沿空间、时间或环境梯度的定向过程）或物种组成的差异（数据集内物种组成的异质性的非定向过程）。

群落的 Beta 多样性分析包括非约束排序（如 PCA、PCoA 等）、层次聚类等。但无论哪种形式的 Beta 多样性分析，均以群落相似（或距离）为基础。

非限制性排序和层次聚类并是不独立的，下面这张图表示的就是非约束排序和层次聚类的关系：

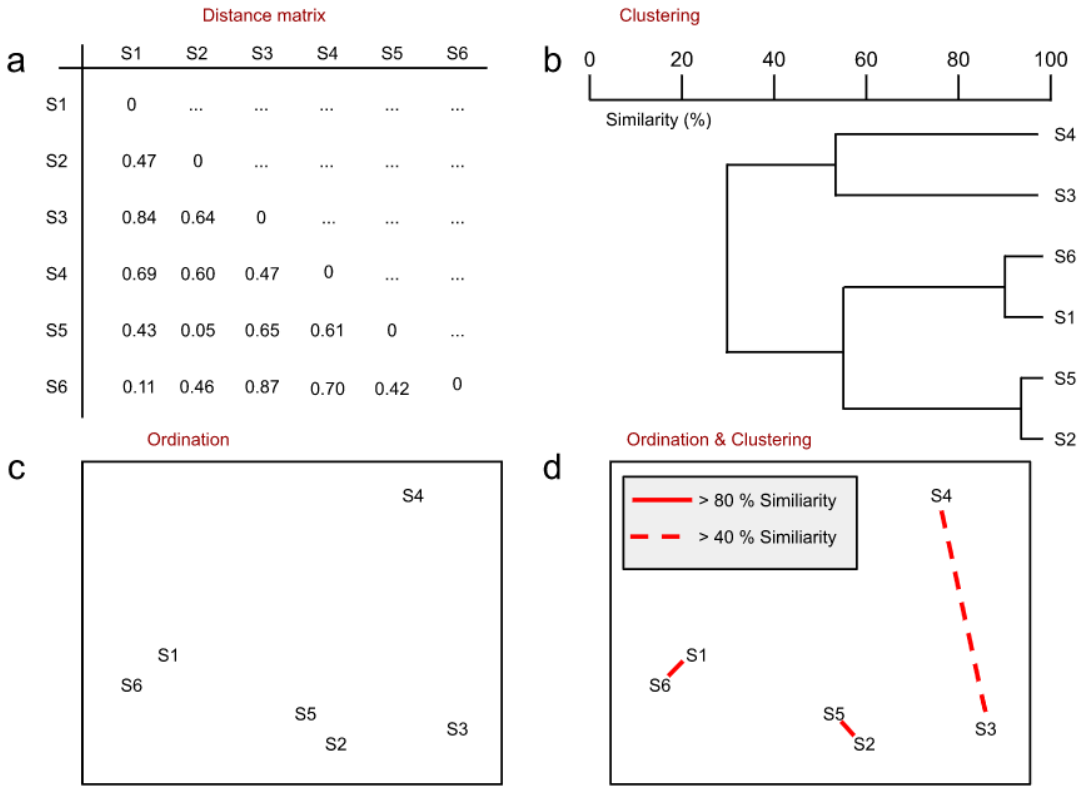



图 1. 图形展示 a 距离矩阵，b 聚类结果，c 排序结果，d 聚类结果叠加排序结果。

将分层聚类分析的结果与诸如非度量多维尺度分析(Non-metric multidimensional scaling,NMDS)产生的排序之类的结果结合，在图中，聚类和 NMDS 结果叠加在左下方面板中。在此示例中，结果展现出现一致性：聚集在一起的对象也彼此接近。

备注：图片来源（http://mb3is.megx.net/gustame/home）

*相似性和距离*

**生态相似性**（Ecological resemblance）以计算样方之间的群落组成相似程度或距离（相异程度）为基础，是处理多元生态数据的基本方法之一。在群落数据的分析中，常用其反映 Beta 多样性。

如在物种数据的分析中，对于两个群落，若它们共享相同的物种，并且所有物种的丰度也一致，那么这两个群落就具有最高的相似程度（或最低距离 0）。**生态学数据分析中的很多统计方法都以样方之间的相似性或距离为基础**，例如上述提到的 Beta 多样性分析中的聚类、排序等，即使对于 PCA 和 CA，实质上在计算时也分别基于欧几里得（euclidean）和卡方（chi-square）距离考虑的。

若两个对象在各属性上越近似，那么它们的相似性就越高。对于群落数据，这些属性一般就是物种组成，或者环境属性等。通常使用物种组成数据，依据相似性指数（similarity indices）判断群落相似性，范围由 0（两个群落不共享任何物种）到 1

（两个群落的物种类型和丰度完全一致）。所有相似性指数均可以转换为距离指数，转化公式大致就是“距离指数=1 – 相似性指数”的关系。

（1）可以转化为相似性指数的距离指数，例如定量数据的相异百分率（也称为 Bray-Curtis 距离）等。二者相互转换的公式通常表示为 D=1-S 或 S=1-D，其中 S 是相似性指数，D 为距离指数。

（2）无法转化为相似性指数的距离指数，例如欧几里得距离、卡方距离。

距离指数（distance indices）或称距离测度（distance measures），与相似性指数相反，距离数值越大表明群落间差异越大。存在多种距离类型，例如欧几里得（Euclidean）距离、Bray-Curtis 距离、UniFrac 距离等。对于物种组成数据，距离指数的最小值为 0（两个群落的物种类型和丰度完全一致），最大取值取决于距离类型和数据本身。

相似性或距离的衡量标准有很多种，Legendre 和 Legendre（1998）列出大约 30 种方法，并对生态相似性作了更详细的介绍，有兴趣可自行参阅 Legendre 和 Legendre（1998）“Numerical Ecology”第七章“Ecological resemblance”的内容。

*常见的相似性/距离指数*

**Jaccard**:

Jaccard 相似性指数（Jaccard similarity index）将两个样方共享的物种数量（a）除以两个样方中出现的所有物种的总和（a + b + c，其中 b 和 c 是仅在第一个和第二个样方中出现的物种数量）。计算公式如下：

$$J_{sim} = \frac{a}{a + b + c}$$

其中，y1j 和 y2j 分别是对象 1 和 2 中元素 j 的数值。若是群落物种数据，那么 y1j 和 y2j 即分别是样方 1 和 2 中物种 j 的丰度。p 是物种数（样方-物种矩阵中的物种数）。如下展示了仅包含两个物种的两个群落之间的欧几里得距离的计算过程。

但是在物种数据的分析中，欧几里得距离却表现得不很理想。主要原因在于它是一个对称的指数，即它将“双零”现象视作相同存在的方式处理，因此会缩小两个共享很少物种的群落之间的距离（实际上，它们差异很大）。可参考上文“对称指数和非对称指数”所述。并且，他还有对“物种丰度的差异”比对“物种是否存在”更加敏感的这么一个特点，也会影响我们对群落相似程度的判断。本文的末尾，详细展示了一例在物种数据处理中使用欧几里得距离可能会带来的问题。

如果仍要将欧几里得距离应用在物种数据的分析中，常见的解决方法是首先对原始物种数据执行预转化（常见的如弦转化、Hellinger 转化等），然后再使用转化后的数据计算欧几里得距离（即对应于下文提到的弦距离、Hellinger 距离，事实上，

它们仍然属于欧氏距离）。尽管弦距离、Hellinger 距离等然是对称指数的范畴，但是相较于使用原始物种丰度数据所得的欧几里得距离，弦距离、Hellinger 距离的优势体现在存在距离的“上限”，降低了欧几里得距离对“物种丰度”的敏感性，有效减少了“双零”问题导致的误差。

更多情况下，我们在处理物种数据时，会尽可能避开使用欧几里得距离这类的对称指数。例如，通常我们选择使用非对称的 Bray-curtis 距离等。除非特定需要，不得不使用欧氏距离的情况下，可再考虑先转化数据再求欧几里得距离的方法。

**Bray-curtis 距离**(Bray-curtis distance):

Bray-curtis 距离或称 Bray-curtis 相异度（Bray-curtis dissimilarity）、相异百分率（Percentage difference）。其计算公式如下：

$$D_{Bray} = \frac{\sum_{i=1}^{p} |y_{ij} - y_{ik}|}{\sum_{i=1}^{p} (y_{ij} + y_{ik})}$$

**欧几里得距离**(Euclidean distance):

欧几里得距离是多变量分析中经常使用的一种距离，如在线性排序方法 PCA、RDA，以及某些层次聚类算法中。欧几里得距离没有上限，最大值取决于数据。

欧几里得距离计算公式如下：

$$D_{Eucl} = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

其中 p 是物种数（样方-物种矩阵中的物种数），y1j 和 y2j 表示两个样方中对应的物种多度。

Bray-curtis 距离的取值范围范围由 0（两个群落的物种类型和丰度完全一致）到 1（两个群落不共享任何物种），因此它也可以直接通过“1 – 距离指数=相似性指数”转化为相似性指数（上文提到的“相似百分率”）。Bray-curtis 距离适用于群落物种数据分析的原因在于它是一个非对称指数，可有效忽略双零。

**Unifrac 距离**(Unifrac distance):

Unifrac 距离，它常用于微生物群落的研究中（例如，16S 扩增子测序）。上述距离的计算方法，仅考虑了物种的存在与否及其丰度，没有考虑物种之间的进化关系，距离 0 表示两个群落的物种组成结构完全一致。在 Unifrac 距离中，除了关注考虑了物种的存在与否及其丰度外，还将物种之间的进化关系考虑在内，距离 0 更侧重于表示两个群落的进化分类完全一致。

例如在 16S 扩增子测序中，根据 16S 序列组成构建 OTUs 进化树，OTUs 之间存在进化上的联系，因此不同 OTUs 之间的（系统发育）距离实际上有“远近”之分。将系统发育树和 OTUs 丰度数据一起添加至距离的计算中，计算 Unifrac 距离。而若使用上述提到的其它只基于 OTUs 丰度数据计算群落距离的方法，则忽略了 OTUs 之间的进化关系，认为 OTUs 间的关系平等。当然，并不是说 Unifrac 距离是最合适 16S 群落分析的，很多情况下它其实也并没有比只基于 OTUs 丰度数据计算群落距离的方法（如 Bray-curtis 距离）“更好”，总之具体问题具体分析吧，根据实际情况选择合适的距离测度。

Unifrac 距离分为非加权 Unifrac 距离（Unweighted unifrac distance）和加权 Unifrac 距离（Weighted unifrac distance）。两种的主要区别是否考虑了物种的丰度。非加权 Unifrac 距离只考虑了物种有无的变化，不关注物种丰度，若两个微生物群落间存在的物种种类完全一致，则距离为 0；加权 Unifrac 距离同时考虑物种有无和物种丰度的变化，若两个微生物群落间存在的物种种类及丰度完全一致，则距离为 0。

对于非加权和加权 Unifrac 距离的选择，看网上很多帖子给的经验性建议：在环境样本的检测中，由于影响因素复杂，群落间物种的组成差异更为剧烈，因此往往采用非加权方法进行分析。但如果要研究对照与实验处理组之间的关系，例如研究短期青霉素处理后，人肠道的菌落变化情况，由于处理后群落的组成一般不会发生大改变，但群落的丰度可能会发生大变化，因此更适合用加权方法去计算。

### *排序*

排序过程是将样品或物种排列在一定的空间，在一个低维空间中，使相似的样品或物种距离相近，相异的样品或物种距离较远。也就是说排序可以揭示微生物-环境间的生态关系，降低维数，减少坐标轴的数目，使排序轴能够反映一定的生态梯度。降维的过程就像投影，找到最好的角度使投影后的物种或者样品的位置关系尽量保持原始的位置关系。常见的方法有：PCA、PCoA、CA、DCA、NMDS、RDA、CCA 等等。

### *PCoA 排序*

PCoA(Principal Coordinate Analysis)即主坐标分析，可呈现研究数据相似性或差异性的可视化坐标，是一种非约束性的数据降维分析方法，可用来研究样本群落组成的相似性或相异性。它与 PCA 类似，通过一系列的特征值和特征向量进行排序后，选择主要排在前几位的特征值，找到距离矩阵中最主要的坐标，结果是数据矩阵的一个旋转，它没有改变样本点之间的相互位置关系，只是改变了坐标系统。两者的

区别为 PCA 是基于样本的相似系数矩阵(如欧式距离)来寻找主成分，而 PCoA 是基于更多种类的距离矩阵(如常用 Bray-Curtis、Unifrac 距离)来寻找主坐标。

PCoA 和 PCA 的不同之处是 PCA 是基于 OTU table 也就是基于欧式距离，而 PCoA 是基于两两样品之间的距离矩阵（可以是任何一种距离），如果 PCoA 也使用欧式距离矩阵的话，那么 PCA 和 PCoA 的分析结果是一样的。

另外，PCoA 是基于距离矩阵，它的排序的目的是将 N 个样品排列在一定的空间，使得样品间的空间差异与原始距离矩阵保持一致，这类排序方法也称作多维标定排序（Multi—dimensional scaling）。**如果排序依赖于相异系数的数值，就叫有度量多维标定法（metric multi—dimensional scaling）所以说 PCoA 分析也叫有度量多维标定法；如果排序仅仅决定于相异系数的大小顺序（秩次排序），则称为无度量多维标定法（Non—Metric Multi—Dimensional Scaling；NMDS）**。

### *NMDS 排序*

非度量多维尺度法是一种将多维空间的研究对象（样本或变量）简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法。适用于无法获得研究对象间精确的相似性或相异性数据，仅能得到他们之间等级关系数据的情形。其基本特征是将对象间的相似性或相异性数据看成点间距离的单调函数，在保持原始数据次序(秩)关系的基础上，用新的相同次序的数据列替换原始数据进行度量型多维尺度分析。换句话说，当资料不适合直接进行变量型多维尺度分析时，对其进行变量变换，再采用变量型多维尺度分析，对原始资料而言，就称之为非度量型多维尺度分析。其特点是根据样品中包含的物种信息，以点的形式反映在多维空间上，而对不同样品间的差异程度，则是通过点与点间的距离体现的，最终获得样品的空间定位点图。

NMDS 过程是迭代的，并且分几个步骤进行：

- 在多维空间中定义群落的原始位置；
- 指定降低维度的数量（通常为 2 个维度）；
- 二维构造样本的初始配置；
- 该初始配置下的距离相对于观察到的（测量的）距离进行回归；
- 根据回归确定应力(stress)或二维构造与预测值之间的差异；

如果应力较高，则按减小应力的方向重新定位 2 维中的点，然后重复进行直到应力低于某个阈值。**经验法则：应力<0.05 可很好地表示尺寸减小，<0.1 非常好，<0.2 还不错，而应力<0.3 有待提高。**

附加说明：**最终结果可能会因初始配置（通常是随机的）和迭代次数而有所不同，因此建议多次运行 NMDS 并尽可能减降低应力值**。

首先，NMDS 需要距离矩阵或相异矩阵。原始欧几里得距离并不是达到此目的的理想方法：它们对总丰度敏感，因此即使物种的标识不同，也可能将具有相似数量物

种的站点(site)视为相似物种。它们对物种的缺失也很敏感，因此可以将缺少相同物种数的站点视为相似物种。

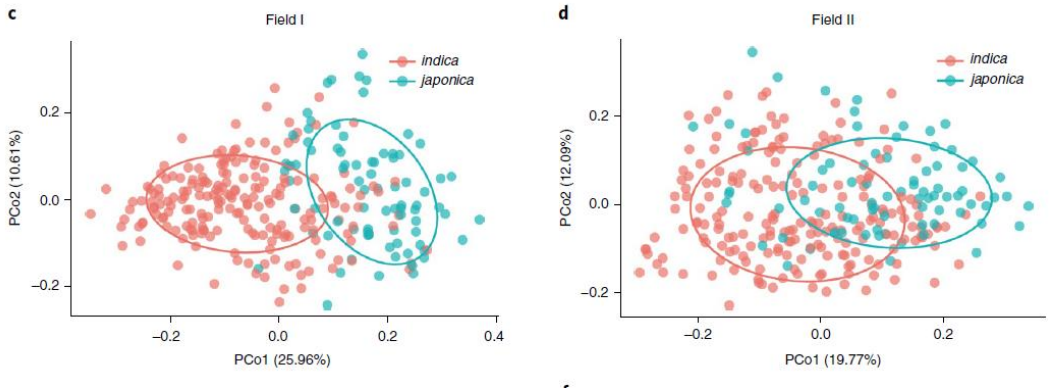
因此，生态学家使用 Bray-Curtis 相异性计算，该计算具有许多理想属性：

- 它不随单位的变化而变化
- 它不受添加/删除两个群落中不存在的物种的影响
- 它不受新增群落的影响
- 它可以识别总丰度的差异

## 实例分析

### *例 1.两地点重复的两组 PCoA*

本文于 2019 年 6 月 5 日发表在 Nature Biotechnology 杂志(37 卷第 6 期)，并当选为当期封面文章。点击查看中文解读




两图并列展示两组间明显的微生物组差异且在不同地点可重复。不同组采用着色配置信椭圆突出组间差异。

c. 基于 Bray-Cutis 距离的主坐标轴分析(PCoA)表明籼粳稻的根系微生物组在第一主轴分开($P < 0.001$，PERMANOVA 采用 Adonis 函数置换检验)。椭圆包括亚种 68%的数据。d. 基于 Bray-Cutis 距离的 PCoA 在地块 2 中结果表明籼粳稻根系微生物组也在第一主轴分开。

c, Unconstrained PCoA (for principal coordinates PCo1 and PCo2) with Bray–Curtis distance showing that the root microbiota of indica separate from those of japonica in field I in the first axis ($P < 0.001$, permutational multivariate analysis of variance (PERMANOVA) by Adonis). Ellipses cover 68% of the data for each rice subspecies. d, Unconstrained PCoA with Bray–Curtis distance showing that the root microbiota of indica separate from those of japonica in field II in the first axis (**P** < 0.001, PERMANOVA by Adonis).

**结果部分描述：**

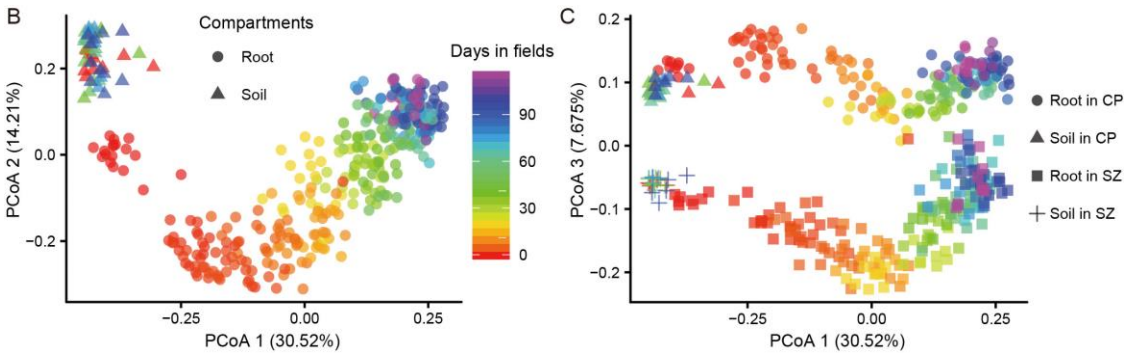
我们发现不同水稻亚种根系微生物组成存在差异。基于 Bray-Curtis 距离的非限制性主坐标轴分析(PCoA)表明籼粳稻在地块 1 的微生物组成明显形成两大类，且在

第一主轴分开(图 1c；附图 2)，表明水稻亚种分化是微生物组变异的最主要影响因素。同时也观察到了由于地块 2 的土壤不同，在地块 2 存在微生物组的变化(附图 3；附表 1)。但在两块地块中，籼粳稻显著分开保持一致(图 1d；附图 3)

We found that the composition of the root bacterial microbiotadiffered in rice subspecies. Unconstrained principal coordinateanalysis (PCoA) of Bray–Curtis distance revealed that theroot microbiota of indica and japonica in field I formed two distinctclusters, which separated along the first coordinate axis(Fig. 1c and Supplementary Fig. 2), indicating that the largest sourceof variation in the rice root microbiota was proximity to the subspeciationbetween indica and japonica. As expected, the root microbiotain field II differed from that in field I due to soil differences(Supplementary Fig. 3 and Supplementary Table 1), but the separationof root microbiota between indica and japonica varieties wasconsistent in the two locations (Fig. 1d and Supplementary Figs. 2and 3).

### *例 2. PCoA 时间序列*

本文是刘永鑫博士负责分析，于 2018 年发表在中国科学的一篇文章封面文章，详细描述了水稻田间全生育期根系微生物组的变化规律，发表 2 年被引 31 次。详见：手把手带你重现菌群封面文章图表。本文对图 1 中的 B/C 子图为例进行说明和点评。




田间水稻微生物组随生育时间变化。以水稻日本晴和 IR24 为材料,并分别种植于昌平和上庄两地，CP 代表北京昌平农场，SZ 代表北京海淀上庄。B-C. 主坐标轴分析(PCoA)展示水稻微生物组随时间变化，其中微生物群落结构主要在第 1/2 轴上随时间变化(B)，而不同土壤类型主要在第 3 轴上明显分开(C)

Figure 1 The rice root microbiota in fields shift over rice residence time in the field. B–C, Principle coordinate analysis showing that the rice microbiota shifts with rice residence time in the field and developmental stages in the first axis (B) and separated by geographical locations in the third axis (C).

**结果**：在所有样品的 Bray-Curtis 距离的主坐标分析（PCoA）中，土体土壤样品聚集在一起，并且水稻根样品在田间和发育阶段的第一个坐标轴上沿着水稻的生长时间从土壤移动（图 1B），表明 稻田在田间的停留时间和发育阶段是影响根系微生物组成的主要因素。 另外，尽管根微生物组在第三轴上被地理位置清楚地分开了，但水稻的生长时间和根系微生物组的动态变化在两个不同的地点显示出一致的趋势（图 1C）。

In Principle Coordinate Analysis (PCoA) of Bray-Curtis distance from all samples, bulk soilsamples clustered together, and rice root samples shifted farfrom the soil across rice residence time in the field

and developmentalstages in the first coordinate axis (Figure 1B),indicating that rice residence time in the field and developmentalstage are main factors influencing the root microbiotacomposition. Additionally, although the root microbiotawere clearly separated by geographic location in the thirdaxis, the rice residence time and development dependent shift of the root microbiota showed consistent trends in the two separate fields (Figure 1C).

- 总结

1. 图 1B/C 是基于 Bray-Curtis 距离进行的 PCoA 分析，采用散点图展示，并**按时间顺序填充彩虹色(比单色过渡明显，但对色盲人群不友好，有些杂志不接受)，按不同生态位和地点设置形状**，信息较丰富；一般人类颜色区分明显，把颜色赋予想要表达的第一变量，如本文的时间变量，形态分配给次要因素；
2. 图 1B 展示 PCo1/2 轴，组间最大差异为不同生态位与时间梯度上的变化，但不 **地点间是无法很好区分时，我们还需要继续探索其他主坐标轴**。本文在图 1C 展示 PCo1/3 轴，可进一步看到 1 轴的差异与时间变化一致，而 **3 轴可以很好分开不同地点**。

*例 3.NMDS 分析不同食物昆虫组肠道菌群*

本文由荷兰皇家科学院生态研究所的 S. Emilia Hannula 和中科院遗传发育所朱峰研究员于 2019 年 8 月发表于 Nature Communications (https://doi.org/10.1038/s41467-019-09284-w)。揭示了食叶昆虫微生物群落来源于土壤而不是取食植物。中文解读详见：Nature 子刊：植食昆虫微生物组来自土壤

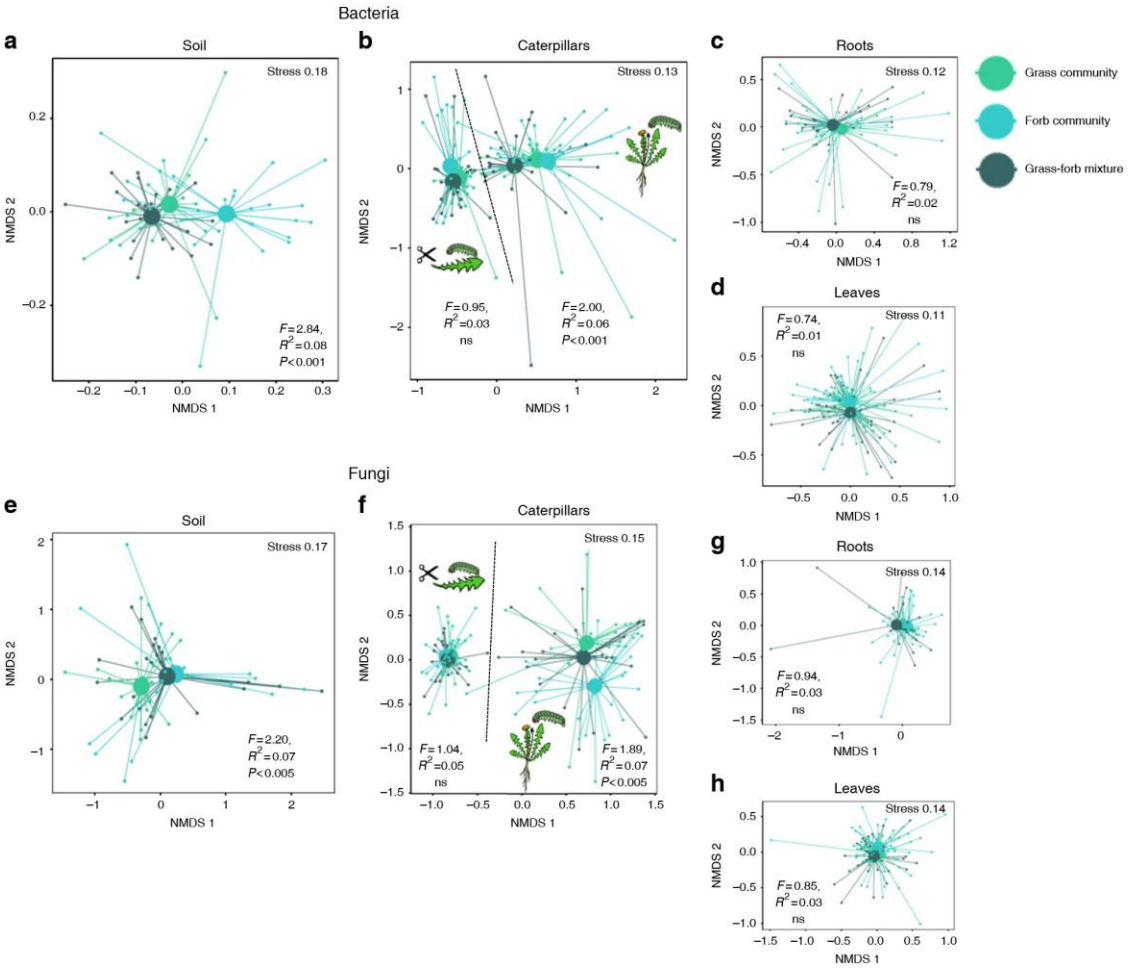



图 a-d 代表了植物群落对毛虫肠道、植物叶片、根系、土壤细菌群落的影响。图 e-h 代表了植物群落对毛虫肠道、植物叶片、根系、土壤真菌群落的影响。NMDS 分析基于 Bray-Curtis 相似性，二维应力值介于 0.11-0.18 之间。草地植被相关的群落使用亮绿色表示。 非禾本草本植物/阔叶草(forb)植被群落使用青绿色点表示。草地和阔叶草植被混合群落使用深绿色表示。每幅图中小点代表样品，大点代表每组样品的中心点。图中的标识为置换检验结果。a，e 代表土壤微生物群落。b，f 代表食用离体叶片和植株的毛虫肠道微生物。c，g 代表植物根系微生物群落，d，h 代表叶微生物群落。

Plant community identity effects on bacterial a–d and fungal (e–h) communities in caterpillars, leaves, roots, and soil. NMDS plots are presented based on Bray–Curtis similarity. The 2D stress value for each panel ranges between 0.11–0.18. Soils originating from grass communities are presented with light green symbols, soils from forb communities with turquoise symbols and soils from mixed grass and forb communities with dark green symbols. In each panel, smaller symbols depict individual samples, centroids are depicted with larger markers. Significance of the plant community treatment effect based on a PERMANOVA is also presented in each panel. a, e represent the composition of microbiomes in soils, b, f microbiomes in caterpillars both on intact plants and on detached leaves. c, g microbiomes in roots and d, h microbiomes in leaves.

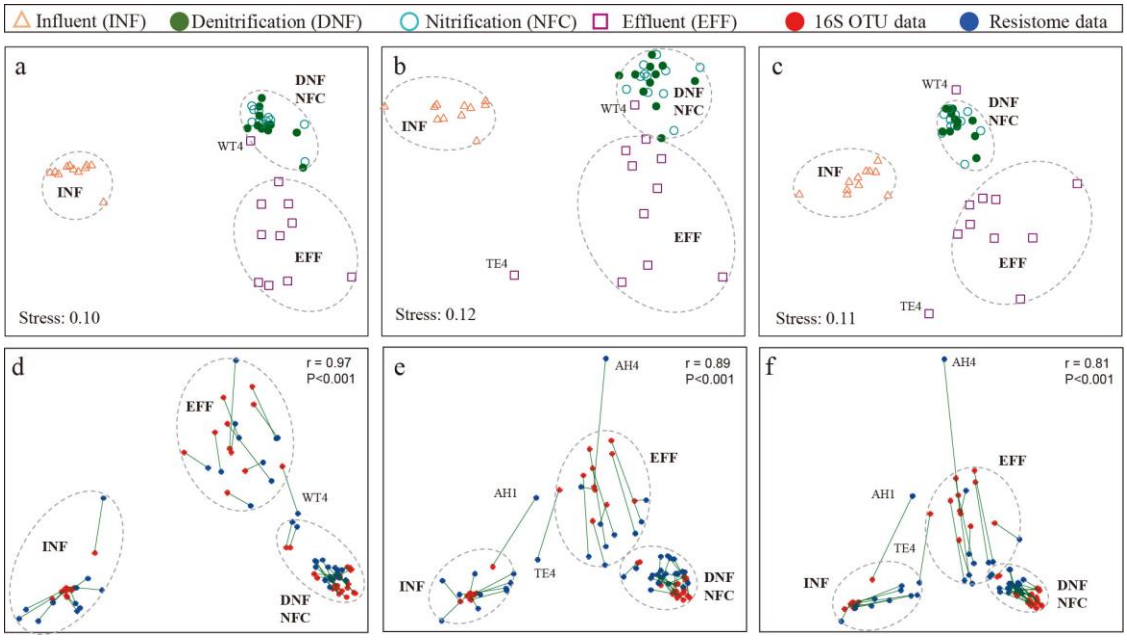
**结果：**

我们通过两个独立的平行试验，研究了田间植物群落对土壤中微生物群落组成、蒲公英和在这些植物上放养的毛虫的影响。植被群落改变了土壤细菌和真菌群落，但是令人惊讶的是并没有改变蒲公英根系和叶片微生物组成 **(图 3c, d, g, h)**。但是我们却检测到了不同植物群落对毛虫微生物群落的影响，但这只有在以完整植株为食的毛虫中检测到。

We investigated the legacy effects created by field-grown plant communities, on the composition of microbial communities in soils, dandelions grown in those soils, and caterpillars reared on these plants, in two parallel assays. The composition of the plant community (fast- and slow-growing grasses or forbs) that conditioned the soils that were used, influenced the fungal and bacterial community structure in these soils (Fig. 3a, e). Surprisingly, this did not alter the root- or leaf -associated microbiomes in the dandelion plants that were growing in these soils (Fig. 3c, d, g, h). However, we did detect these soil-derived plant community effects in caterpillar microbiomes, but only when the caterpillars were fed on intact plants (Fig. 3b, f), sug-gesting that, even though they are plant feeders, the caterpillars had been in direct contact with the soil.

## *例 4.NMDS 分析组间的功能基因类群*

瑞士 EAWAG 研究所，列西湖大学鞠峰教授于 2019 年发表于 The ISME Journal 的成果，发现污水厂抗性组受细菌组成和基因交换驱动且出水中抗性表达活跃 (https://doi.org/10.1038/s41396-018-0277-8)。全文解读详见：ISME：污水厂抗性组受细菌组成和基因交换驱动且出水中抗性表达活跃*。




污水厂不同处理部位抗性基因组的组成与细菌群落组成相关。a-c NMDS 分析描述了不同部位之间基于 ARG（a）、BRG（b）、MRG（c）组成的 Bary-curtis 距离。

Resistome composition correlates with bacterial community composition and phylogeny across wastewater treatment compart-ments. a–c Non-metric multidimensional scaling plots depict Bray-Curtis distances between treatment compartments based on relative abundance of antibiotic (a), biocide (b), and metal (c) resistance genes in the metagenomes.

**结果：**

细菌抗性组系统发育结构。为了测试在我们的数据集中是否存在这种情况，我们使用排序方法来跟踪抗性组（图 5）的结构变化。无论分析是基于看抗性组、杀菌剂和金属抗性基因的丰度指标（图 5a–c），样品始终分为三个主要类别。

Bacterial phylogeny structures soil resistomes. To test if this was the case in our dataset, we used ordination to follow structural variations in the resistomes (Fig. 5) both between and within treatment compartments. The samples consistently clustered into three main groups by treatment compartment with bioreactor samples closely clustered together, whether the analysis was based on abundance metrics of antibiotic, biocide, and metal resistance genes (Fig. 5a–c).

## PCoA/NMDS 实战

### *安装和载入 R 包*

关于更多本项目中示例文件的下载，R 包安装的内容，请参考之前的章节：

- 211.Alpha 多样性箱线图(样章，11 图 2 视频)

```
if (!requireNamespace("devtools", quietly=TRUE))
    install.packages("devtools")
library(devtools)
if (!requireNamespace("amplicon", quietly=TRUE))
    install_github("microbiota/amplicon")
suppressWarnings(suppressMessages(library(amplicon)))
```

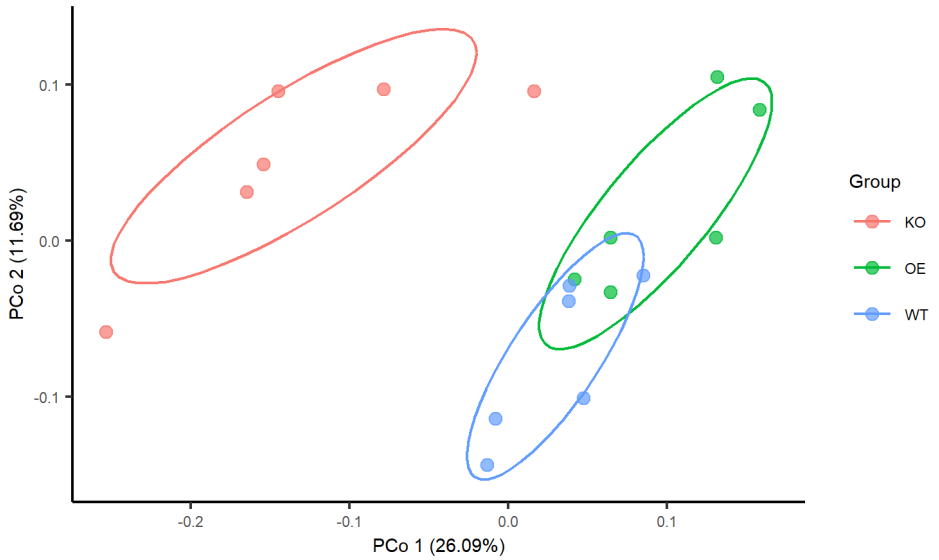
### *主坐标轴分析 PCoA*

主坐标轴分析(principal coordinate analysis, PCoA)

在 amplicon 包中有 beta_pcoa 函数可以快速绘制 PCoA 散点图，并按组着色和添加 68%的置信椭圆

本次绘制使用函数内置数据演示，查看函数帮助，打问题(?)+函数名，如?beta_pcoa

```
# 使用内置数据，输入距离矩阵、元数据和分组绘制 PCoA
(p=beta_pcoa(beta_bray_curtis, metadata, "Group"))
```




```
# 保存位图和矢量图，分别用于预览和出版
ggsave(paste0("p1.PCoA.bray.jpg"), p, width=89, height=56, units="mm")
ggsave(paste0("p1.PCoA.bray.pdf"), p, width=89, height=56, units="mm")
```

图 1. 散点图展示基于 Bray-Curtis 距离的 Beta 多样性 PCoA。点代表样本，颜色代表分组，并按每组添加 68%置信度的椭圆方便组间比较，图中展示主坐标分析的前两轴，解析率见坐标轴括号中。

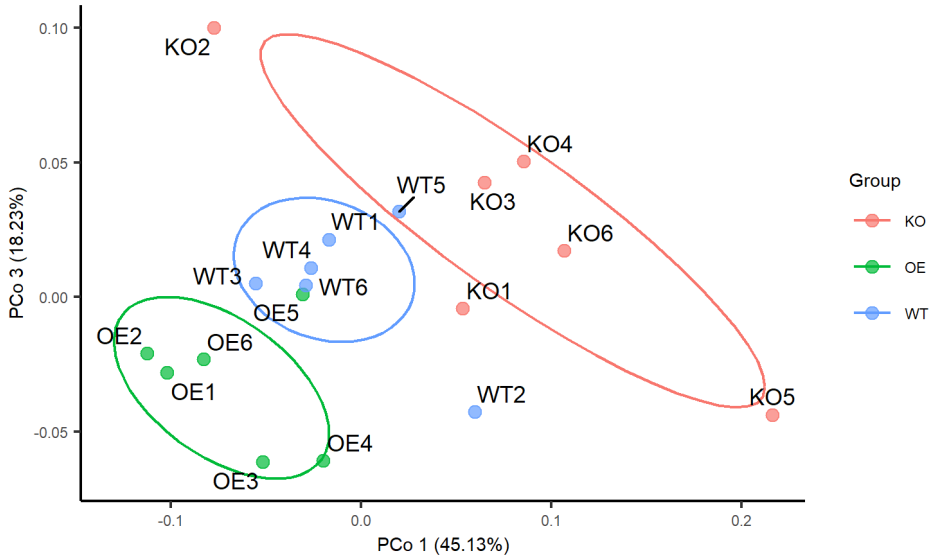
本次测试数据来自刘永鑫博士负责分析并于 2019 年发表于 Science 的文章(即上图展示的内置数据)，讨论了基因型对菌群的影响。详见宏基因组公众号详细解读-Science：拟南芥三萜化合物特异调控根系微生物组

我们再演示从文件读取距离矩阵和元数据，数据位于 Data/Science2019 目录，本次需要元数据(metadata.txt)和 Beta 多样性距离矩阵(alpha/unifrac.txt)两个输入文件(注：距离矩阵这里是由 USEARCH -beta_div 生成，将在扩增子流程部分详细介绍，也可由 vegan 包计算生成)。

```
# USEARCH 可选距离矩阵 bray_curtis、unifrac、unifrac_binary、jaccard、manhatten、euclidean
# 设置距离矩阵类似，本次使用 unifrac
distance_type="unifrac"
# 读取距离矩阵并预测前 3 行 3 列，再读取元数据
distance_mat=read.table(paste0("../Data/Science2019/beta/",distance_type,".txt"), header=T, row.names=1, sep="\t", comment.char="")
distance_mat[1:3, 1:3]
```

```
      KO1   KO2   KO3
KO1 0.000 0.178 0.159
KO2 0.178 0.000 0.201
KO3 0.159 0.201 0.000
```

```
metadata=read.table("../Data/Science2019/metadata.txt", header=T, row.names=1, sep="\t", comment.char="", stringsAsFactors=F)
# PCoA 散点图，按metadata 的Group 列着色，添加标签，PCo1/3
(p=beta_pcoa(distance_mat, metadata, groupID="Group", ellipse=T, label=T, PCo=13))
```




```
# 保存8：5 的半版图
ggsave(paste0("p2.PCoA.unifrac.jpg"), p, width=89, height=56, units="mm")
ggsave(paste0("p2.PCoA.unifrac.pdf"), p, width=89, height=56, units="mm")
```

图 2. 基于 Unifrac 距离的 PCoA。看到 PCo 1/2 的解析率比前面 Bray-Curtis 距离结果有提高，表明在 Unifrac 距离前两主轴一般可以解析更高比例的差异。由于 Unifrac 考虑进化距离，一般样本/组间差异会进一步缩小。

有时我们更想知识组间是否存在显著差异，使用?beta_pcoa_stat 查看函数帮助，使用距离矩阵指定分组，对全部组别两两差异使用 adonis 函数进行检测。

```
# 使用adonis 检测组件差异，注意是两两检测，并且将检测结果保存到当前路径下。
beta_pcoa_stat(distance_mat, metadata, "Group", "beta_pcoa_stat.txt")

  sampA sampB
1    KO    OE
  sampA sampB
1    KO    WT
  sampA sampB
1    OE    WT

# 结果文件默认见beta_pcoa_stat.txt
beta_pcoa_stat(dis_mat=distance_mat, metadata=metadata, groupID="Group",
 pairwise=F, pairwise_list="../Data/Science2019/compare.txt")
```

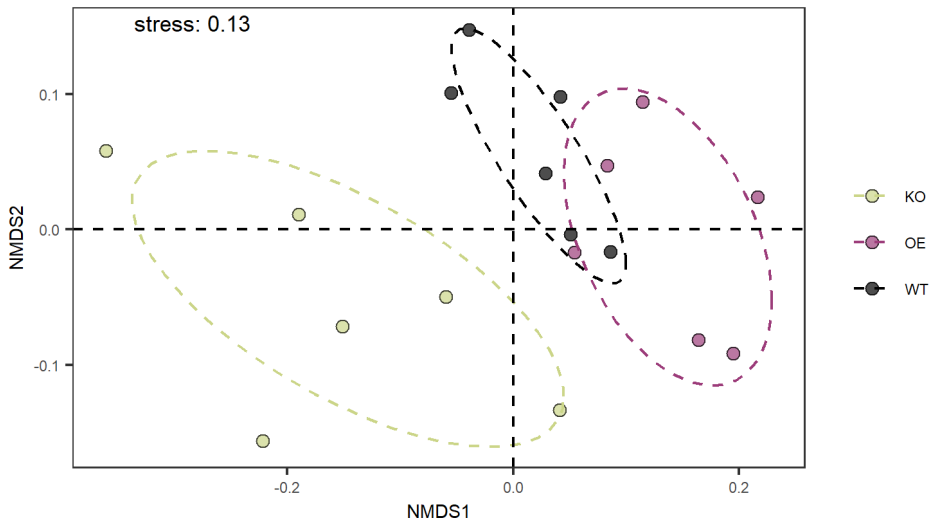
### *非度量多维尺度 NMDS*

我们将会用到 BetaDiv 函数，这个函数依赖 phyloseq 可以计算目前主流的降维排序方法，包括 DCA, CCA, RDA, NMDS, MDS, PCoA, PCA, LDA，t-sne，并且结合了群落差异分析，为大家带来相对全面的 beta 多样性分析。我们下面以 NMDS 为例演示函数的用法。?BetaDiv 显示帮助

```
# 安装Bioconductor 的R 包 phyloseq
if (!requireNamespace("BiocManager", quietly=TRUE))
    install.packages("BiocManager")
suppressWarnings(suppressMessages(library(BiocManager)))
if (!requireNamespace("phyloseq", quietly=TRUE))
    BiocManager::install("phyloseq")
library(phyloseq)

# 输入抽平标准化的特征表、元数据、分组列名、距离类型、降维和统计方法
result=BetaDiv(otu=otutab_rare, map=metadata, group="Group",
                 dist="bray", method="NMDS", Micromet="adonis")
```

```
Run 0 stress 0.1251324
Run 1 stress 0.1495259
Run 2 stress 0.1307573
Run 3 stress 0.144549
Run 4 stress 0.1356143
Run 5 stress 0.1568291
Run 6 stress 0.1504107
Run 7 stress 0.1639228
Run 8 stress 0.1317366
Run 9 stress 0.1262052
Run 10 stress 0.1251327
... Procrustes: rmse 0.0008382898  max resid 0.002266946
... Similar to previous best
Run 11 stress 0.1550039
Run 12 stress 0.1504272
Run 13 stress 0.1252244
... Procrustes: rmse 0.0308105  max resid 0.113413
Run 14 stress 0.1542445
Run 15 stress 0.1495225
Run 16 stress 0.1251328
... Procrustes: rmse 0.0008246685  max resid 0.002227507
... Similar to previous best
Run 17 stress 0.1452607
Run 18 stress 0.1251324
... New best solution
... Procrustes: rmse 5.105346e-05  max resid 0.0001411738
... Similar to previous best
Run 19 stress 0.237996
Run 20 stress 0.1252244
```

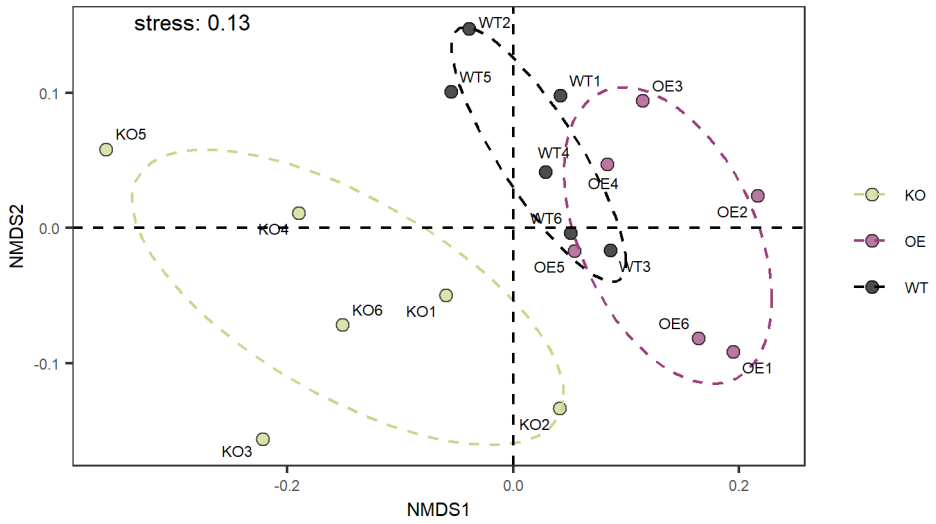
```
... Procrustes: rmse 0.03076905  max resid 0.1133602
*** Solution reached

# 返回结果列表：标准图，数据，标签图，成对比较结果，整体结果

#提取排序散点图(结果列表中的1)
(p=result[[1]])
```




```
ggsave(paste0("p3.NMDS.bray.jpg"), p, width=89, height=56, units="mm")
ggsave(paste0("p3.NMDS.bray.pdf"), p, width=89, height=56, units="mm")
```

图 3. NMDS 分析样本微生物群落结构，按组着色，stress 值显示于左上角。

```
# 提取出图坐标
plotdata=result[[2]]
plotdata[1:3,1:3]

               x           y Group
KO1 -0.05948976 -0.04984976    KO
KO2  0.04101286 -0.13349433    KO
KO3 -0.22212154 -0.15616441    KO

# 提取带标签排序散点图
(p=result[[3]])
```




```
ggsave(paste0("p4.NMDS.bray.label.jpg"), p, width=89, height=56, units="mm")
ggsave(paste0("p4.NMDS.bray.label.pdf"), p, width=89, height=56, units="mm")
```

图 4. NMDS 分析样本微生物群落结构，添加样本标签。

```
# 提取两两比较差异检测结果
(pair=result[[4]])
```

```
        ID           stat        p
1 KO_VS_OE adonis:R 0.258 p: 0.003
2 KO_VS_WT adonis:R 0.208 p: 0.004
3 OE_VS_WT adonis:R 0.169 p: 0.012
```

```
# 提取全部组整体差异检测结果
(Mtest=result[[5]])
```

```
[1] "adonis:R 0.269 p: 0.001"
```

*了解 PhyloSeq 对象*

输入数据除了支持特征表、元数据+分组；还支持 phyloseq 对象。

我们将特征表和元数据转换为 PhyloSeq 对象(简称 ps)

```
# 指定目标分组列为Group，作为默认分组
metadata$Group=metadata[["Group"]]
# 输入特征表和元数据为PhyloSeq 对象
ps=phyloseq(otu_table(as.matrix(otutab),taxa_are_rows=TRUE),
            sample_data(metadata),phy_tree(tree))
```

当然，除了常用的 adonis 置换检验，可选 anosim/MRPP 差异显著性检验方法。

```
result=BetaDiv(ps=ps, dist="bray", method ="NMDS", Micromet ="anosim")

Run 0 stress 0.1251141
Run 1 stress 0.1313194
Run 2 stress 0.1445213
Run 3 stress 0.1546294
Run 4 stress 0.1350167
Run 5 stress 0.1428286
Run 6 stress 0.125113
... New best solution
... Procrustes: rmse 0.0003985593  max resid 0.001238319
... Similar to previous best
Run 7 stress 0.1254256
... Procrustes: rmse 0.02702083  max resid 0.09492863
Run 8 stress 0.1350169
Run 9 stress 0.1313192
Run 10 stress 0.1348394
Run 11 stress 0.1554587
Run 12 stress 0.1300194
Run 13 stress 0.1441429
Run 14 stress 0.1254226
... Procrustes: rmse 0.02627008  max resid 0.09416715
Run 15 stress 0.1254256
... Procrustes: rmse 0.02705633  max resid 0.09500571
Run 16 stress 0.1313192
Run 17 stress 0.1300171
Run 18 stress 0.1501868
Run 19 stress 0.1300168
Run 20 stress 0.1300173
*** Solution reached

result[[5]]

[1] "ANOSIM.r 0.507 p: 0.001"
```

## 参考文献


Beta 多样性和生态相似性 http://blog.sciencenet.cn/blog-3406804-1195182.html

Xiao-Tao Jiang, Xin Peng, Guan-Hua Deng, Hua-Fang Sheng, Yu Wang, Hong-Wei Zhou & Nora Fung-Yee Tam. (2013). Illumina Sequencing of 16S rRNA Tag Revealed Spatial Variations of Bacterial Communities in a Mangrove Wetland. Microbial Ecology 66, 96-104, doi: https://doi.org/10.1007/s00248-013-0238-8

Jingying Zhang, Yong-Xin Liu, Na Zhang, Bin Hu, Tao Jin, Haoran Xu, Yuan Qin, Pengxu Yan, Xiaoning Zhang, Xiaoxuan Guo, Jing Hui, Shouyun Cao, Xin Wang, Chao Wang, Hui Wang, Baoyuan Qu, Guangyi Fan, Lixing Yuan, Ruben Garrido-Oter, Chengcai Chu & Yang Bai. (2019). NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nature Biotechnology 37, 676-684, doi: https://doi.org/10.1038/s41587-019-0104-4

Jingying Zhang, Na Zhang, Yong-Xin Liu, Xiaoning Zhang, Bin Hu, Yuan Qin, Haoran Xu, Hui Wang, Xiaoxuan Guo, Jingmei Qian, Wei Wang, Pengfan Zhang, Tao Jin, Chengcai Chu & Yang Bai. (2018). Root microbiota shift in rice correlates with resident time in the field and developmental stage. Science China Life Sciences 61, 613-621, doi: https://doi.org/10.1007/s11427-018-9284-4

S. Emilia Hannula, Feng Zhu, Robin Heinen & T. Martijn Bezemer. (2019). Foliar-feeding insects acquire microbiomes from the soil rather than the host plant. Nature Communications 10, 1254, doi: https://doi.org/10.1038/s41467-019-09284-w

Feng Ju, Karin Beck, Xiaole Yin, Andreas Maccagnan, Christa S. McArdell, Heinz P. Singer, David R. Johnson, Tong Zhang & Helmut Bürgmann. (2019). Wastewater treatment plant resistomes are shaped by bacterial composition, genetic exchange, and upregulated expression in the effluent microbiomes. The ISME Journal 13, 346-360, doi: https://doi.org/10.1038/s41396-018-0277-8

责任编辑：刘永鑫(Yong-Xin Liu)