# 序

## 推荐序

蓝灿辉、赵方庆、王军、褚海燕、韦中等写推荐。

## 编者序

概述，历史背景，我的基础，主要动机。你将学到什么？

## 写在前面

2020 年 4 月底，宏基因组公众号编辑部启动了《微生物组数据分析和可视化实战》专项，该项目是对 2017 年宏基因组公众号成立之初，刘永鑫博士发布的《扩增子图表解读、分析流程和统计绘图》三部曲系列教程的更新和扩展。

2017 版扩增子分析学习三部曲(共 27 篇文章)：

- 扩增子图表解读-理解文章思路
- 扩增子分析流程-把握分析细节
- 扩增子统计绘图-冲击高分文章

上面的教程虽然收获了上万的读者，帮助了大量同行解决了入门难的问题。这本领域是目前世界科研的热点，发展极快。虽然过了不到三年的时间，但以列在的经验回看之前的教

程，是非常有必要进行更新为扩展的，以便让同行把握本领域最新的动态、技术和发展方各。

总结过去这三年，扩增子技术的发展已经从狂热到归于理性，分析技术和相关流程层出不穷。本领域出现了罕见的主流软件 mothur、QIIME 和 USEARCH 三足鼎立的局面，均轻松引用过万拉开了全民研究微生物组的新时代。2017 年是 QIIME 2 公测的开始，进一步助力 QIIME 成为本领域首个引用过 2 万的传奇软件，并成就作者 Rob Knight 教授以 20 万引用成为微生物组领域高引第一人。USEARCH 虽然 64 位版是商业软件，但 VSEARCH 继续填补这一空白，推动了易用性和跨平台分析的广泛使用。其次是大量 R 包的出现，使用包治百病的效率进一步扩大，如 DADA2 包的子现，使用 R 语言也可以实现扩增子数据全流程的分析。

# 我们的优势

**本次更新和扩展微生物组数据分析和可视化系列教程，我们具有以下优势**：

- 三年来每天坚持不断的学习和分享，持续分享**近两千篇本领域科研经验、分析方法和文献解读等，300 多万字**；
- 宏基因组公众号编辑部从我**一人独行发展到 50 多位投稿作者**，80%具博士学位，50%具有高级职称，而且**关注人数近 9 万人，同时形成了 10 个 500 人的专业同行微信交流群**；
- 参与本领域最广泛使用软件 **QIIME 第二版的开发**，并 2017-2020 **每年翻译并更新 QIIME 2 中文帮助文档**，目前 2020.2 版本帮助文档共包括 **32 节 10 余万字**；

- 团队主要成员负责分析的微生物组项目相继发表于 ***Science***、***Nature***、***Nature Biotechnology***、***Cell Host & Microbe***、***Microbiome***、***ISME Journal*** 等几十种主流期刊，得到国内外同行的认可；
- 团队受到数十家杂志邀请撰写方法学综述，目前已经在 ***Protein & Cell***、***Current Opinion in Microbiology***、***Chinese Medical Journal*** 和***遗传*** 杂志发表方法学综述，并受 ***JoVE*** 杂志邀请组织微生物组方法学视频专刊，成绩得到了国内外同行的认可和肯定。

基于以上基础 ，我们计划结合近 3 年发展的新方法、以及我们更新的知识体系，完成一套《微生物组数据分析和可视化的零基础教程》，**解读广大国内同行入门难的痛点，帮助研究生快速成长，帮助导师节约时间。让每位同行，一本书在手，实现数据分析的理解、开展、统计可视化、论文写作和投稿的全程指导。**

# 你能学到什么

**通过简单阅读本书，你能看懂，并可亲手绘制如下常用统计和可视化结果**：

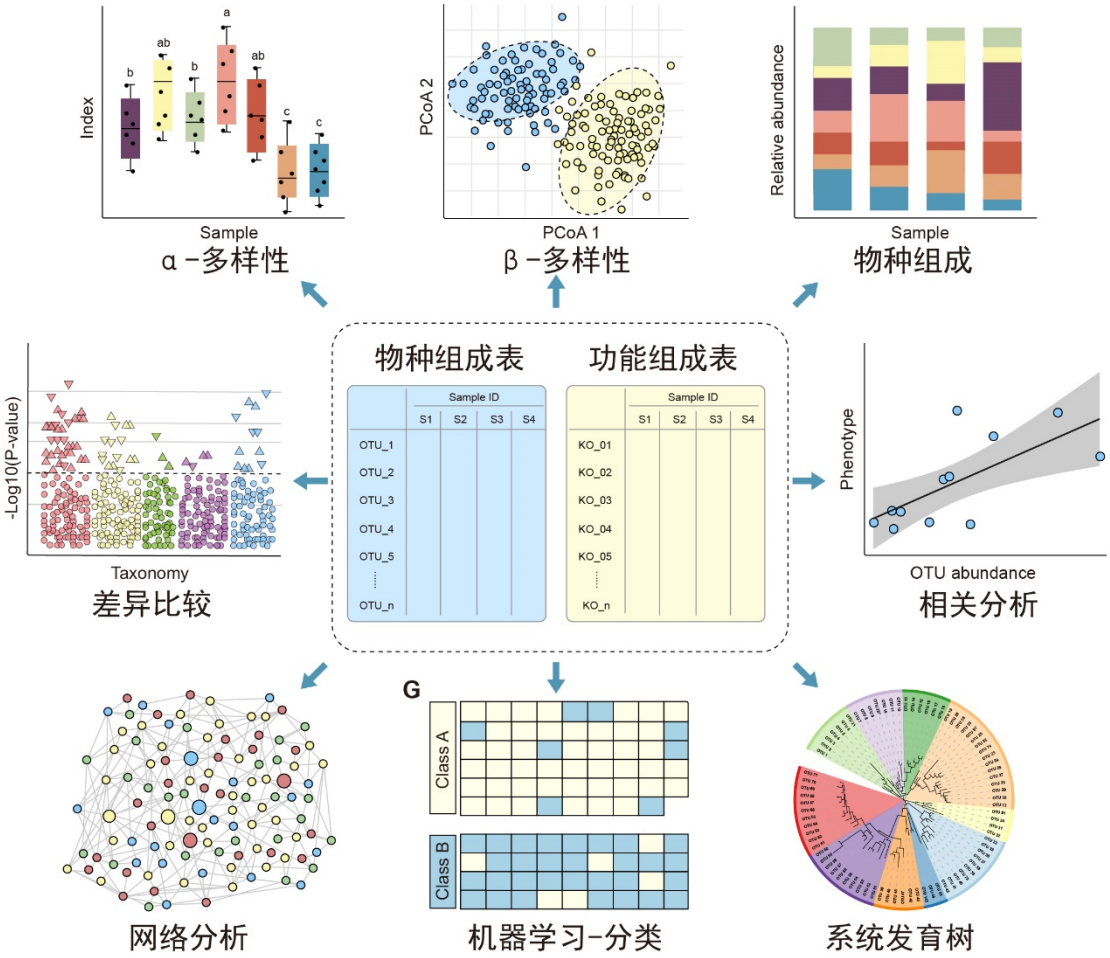



**图 1. 微生物组数据核心特征表及常用可视化方案**(Liu, et al. 2020)

本书的分析部分全程配合 github 更新，以确保随时相关软件的发展代码仍然可用。同时也会录取相关的视频教程，实现零基础自学的目标。

# 总结和展望

目前微生物组数据的可视化工具仍然处于发展的初级阶段，绝大多数的分析工作需要作者编写~~代码~~大量代码，这对于生物学家是极其困难的。斯坦福大学的 Susan Holmes 教授于 2012 年发布的是目前最主要的分析可视化包，该软件包于 2013 年正式发表于 PloS one，方便数据筛选，同时提供了常用 alpha、beta 多样性、物种组成可视化，帮助了近 4 千篇文章的发表。Phyloseq 的核心是将原始数据分析结果转化为 S4 类存储对象(提供了

封闭特征表、样本和特征元数据及进化树 4 类文件的封装格式)，结合 dplyr 进行数据框转换处理，stringr 进行字符串处理，ggplot 进行可视化处理即可高效完成基本统计出图。其次 2016 年以后逐渐开发一些 R 包，大大增强了扩增子下游分析。例如: microbiome 包专门为扩增子数据分析准备，丰富了微生物群落分析的内容；ggtree 增强了进化树可视化方案，可以使用简单的 ggplot 语法绘制美观的进化树；ggraph，tidygraph 增强网络可视化性能，可以使用 ggplot 语法轻易完成网络图可视化。这些 R 包出现，让新一代的基于 R 语言的扩增子分析变得简单和高效。随这扩增子测序的逐渐繁荣，其次在 19 年提出了基于 R 语言平台的新的算法 FEAST：用于预测微生物组来源，并发表在 Nature Methods 上，让微生物溯源分析更加高效快捷；随着人工智能的繁荣，让机器学习在微生物领域大放光彩，这些分析的实现离不开机器学习相关 R 包:randomforest，e1071，caret，pROC 等；随着 picrust 功能预测的开发，更适合环境微生物生态的功能预测 R 包 Tax4Fun2 可以使用扩增子数据更加准确的预测环境微生物群落功能的变化。这些包共同造就了 R 语言在扩增子数据后续分析得一完整生态，并在迅速完善发展。宏基因组团队近年来积累的代码汇编成的 EasyAmplicon 流程和 amplicon 包，提供了几十种扩增子常用分析和可视化方案，可更快速有效获得出版级结果，同时也为更高要求的读者提供代码框架，方便进一步修改。在未来，后续的数据分析变化也越来越多样，主要是以多个 R 包在内的新包出现和应用，更应用的窗口软件、网络服务器分析云平台也将快速发展，同时开源代码和保持可重复是重要的要求。

近些年我参与组织了数十场微生物组分析研讨会，学员主要来自中国大陆各高校和研究所，也有来自茅台、五粮液、安琪酵母、华为等大厂的科研人员，甚至**有海外华人不远万里从美国、欧洲、澳州、新西兰、新加坡等地飞来北京参加微生物组专题学习研讨会。这**

**也说明不只国内，国外也同样缺少该领域的入门教程。希望在不远的未来，发行此书的英文版，供海外同行学习。**

# 作者简介

**刘永鑫**，博士。2008 年毕业于东北农业大学微生物学专业，2014 年于中国科学院大学获生物信息学博士学位，2016 年中科院遗传发育所博士后出站留所任工程师。目前主要研究方向有微生物组数据分析、方法开发和科学传播。目前以第一作者(含共同)或微生物组数据分析负责人在 **Science**、**Nature Biotechnology**、**Cell Host & Microbe** 等杂志发表论文 20 余篇，引用千余次。作为中国唯一单位代表参与微生物组分析平台 **QIIME 2** 开发。受邀以第一作者和/或通讯作者(含共同)在 **Protein & Cell**、**Current Opinion in Microbiology**、**遗传** 等杂志发表微生物组研究方法综述。2017 年 7 月创办“宏基因组”公众号，目前分享本领域相关原创文章 1800 余篇，代表作品有《微生物组图表解读、分析流程和统计绘图》、《QIIME2 中文教程》等系列，关注人数 9 万+，累计阅读 1400 万+。

**文涛**，博士在读，2016 年就读于南京农业大学。荣拜资环院沈其荣教授课题组，研究方向为根际微生物生态，具体为植物介导下根际小分子代谢组同土壤微生物群落在防控土传病害方面的相互作用，关注宏基因组和代谢组。“微生信生物”公众号创始人，2019.1 加入“宏基因组”公众号任编辑，2019.12 起任副主编，发表《Microbiome：根系分泌物驱动土壤记忆抵御植物病原菌、《DADA2 中文教程 v1.8》和《R 语言绘制带聚类树的堆叠柱形图》等文章 20 余篇。

# Reference

- 如何优雅的提问
- 我的生物信息之路
- 微生物组入门必读+宏基因组实操课程
- Yong-Xin Liu, Yuan Qin, Tong Chen, Meiping Lu, Xubo Qian, Xiaoxuan Guo & Yang Bai. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 41, 1-16, doi:10.1007/s13238-020-00724-8 (2020).
- 扩增子图表解读-理解文章思路
- 扩增子分析流程-把握分析细节
- 扩增子统计绘图-冲击高分文章
- 引统计自谷歌学术：https://scholar.google.com/，时间截止 2020 年 5 月 27 日

# 微生物组分析

## 发展史

摸索，初步探索，建立方法，百花齐放。

## 分析前准备

## 生物信息

## 元数据和测序数据

## 分析的基本思路

## Shell 和 Linux

## R 统计与绘图

## 常用分析流程

---

盘点主流软件。高级阶段应该是各种方法步骤的自由组合，甚至是根据需要设计、开发方法。

## 认识特征表

---

特征表是上游大数据分析的终点，是里程碑式的成果，同时也是下游分析的起始。

# 特征表的分析和可视化

## Alpha 多样性

---

### Alpha 多样性概念和常用指数

物种多样性主要从三个层面进行衡量，分别是α多样性、β多样性和γ多样性。每个衡量尺度所呈现的多样性角度不同。Alpha 多样性也被称为生境内多样性（within-habitat diversity），是指一个特定区域或生态系统内的多样性。以医学领域为例，α多样性是指一个样本中物种的多少、丰度和均匀度（图 1）。我们用动物园来打个形象的比喻，α多样性是指这个动物园中动物的种类数、每种动物的只数和每种动物数量的平衡关系。β多样性又称生境间多样性（between-habitat diversity），是指生境群落之间物种组成的相异性或物种沿环境梯度更替的速率。同样以医学领域为例，它主要指样本间物种组成的相异性（图 1）。β多样性相当于 2 个动物园中动物种类的差异情况。γ多样性是指一个区域内总的多样性，由于其在微生物组研究中极少使用，此处不作介绍。


图 1. α多样性和β多样性示意图(Qian, 2020)。α多样性主要体现样本内物种多少、丰度和/或均匀度，而β多样性指样本间多样性异同。

α多样性的计算主要与 3 个因素有关：一是物种数目（richness），二是丰度（abundance），三是均匀度（evenness）。物种数目是指一个样本中物种存在的个数，与每个物种量的多寡无关。丰度是指每个物种的多寡，比如一个粪便样本中物种 A 出现 10 次，物种 B 出现 1000 次；如果将每个样本所有物种求百分比，这样每个样本的物种丰度合计数为 1，这种丰度叫相对丰度。均匀度主要考量物种之间的相对比例。α多样性常用的衡量指标有：

- 观察到的物种数（Observed OTU/ASV）：是指每个样本中能够观察到的 OTU 或 ASV 的数量，与每个 OTUs 或 ASVs 的多寡无关。如果把动物园比喻成一个样本，则“Observed OTUs”是指这个动物园中动物的种类数，与每种动物具体有几只无关。

- Chao1 指数：是物种数目的衡量标准之一，它考虑 3 个因素，一是物种数目，二是只有 1 条序列的物种数目，三是 2 条序列的物种数目，计算公式是：Schao1=Sobs+n1(n1-1)/2(n2+1)，其中 Schao1 为估计的 OTU 数，Sobs 为观测到的 OTU 数，n1 为只有 1 条序列的 OTU 数目，n2 为只有 2 条序列的 OTU 数目。Chao1 指数越大，表明某群落物种数目较多。注意，从公式可以看出，Chao1 指数受 1 条和 2 条序列的物种影响较大。

- 基于丰度的覆盖估计值（Abundance-based Coverage Estimator, ACE）：是用来估计群落中含有 OTU 数目的指数，同样由 Chao 提出(Chao and Yang, 1993)，是生态学中估计物种总数的常用指数之一。默认将序列量 10 以下的 OTU 都计算在内，从而估计群落中实际存在的物种数。ACE 指数越大，表明群落中物种数目越大。

- 香农指数（Shannon-Wiener index）：香农指数综合考虑了群落的物种数目和均匀度这两个因素。Shannon 指数值越高，表明群落的α多样性越高。注意，该指标对于丰度低的物种有较大的权重，即计算时受丰度低的物种影响较大，在解释香农指数时需要注意这点。

- 辛普森指数（Simpson index）：用来估算样品中微生物的多样性指数之一，由 Edward Hugh Simpson（1949）提出。Simpson 指数值越大，说明群落多样性

越低。辛普森指数在计算时将丰度高的物种设置了较大权重，所以高峰度物种较多时该指数值较大，这与香农指数有明显区别。

- Pielou’ s 均匀度指数（Pielou’ s Evenness Index）：这是最常用的均匀度指数，它其实就是香农指数与 Observed OTU/ASV 对数的比值。很显然，这个指标受 Observed OTU/ASV 影响很大，这是这个指标的主要缺点之一。由于香农指数和辛普森指数本身就包含了均匀度信息，实际研究工作中这 2 个指标更常用。

# 认识箱线图

箱形图（Box-plot）又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的统计图。因形状如箱子而得名。在宏基因组领域，常用于展示样品组中各样品 Alpha 多样性的分布。

下面两张图参考自斯坦福医学统计课程第一单元第三节，PPT32-33 页，中文翻译参考百度百科。直接上图。

- 第一种情况：最大或最小值没有超过 1.5 倍箱体范围；

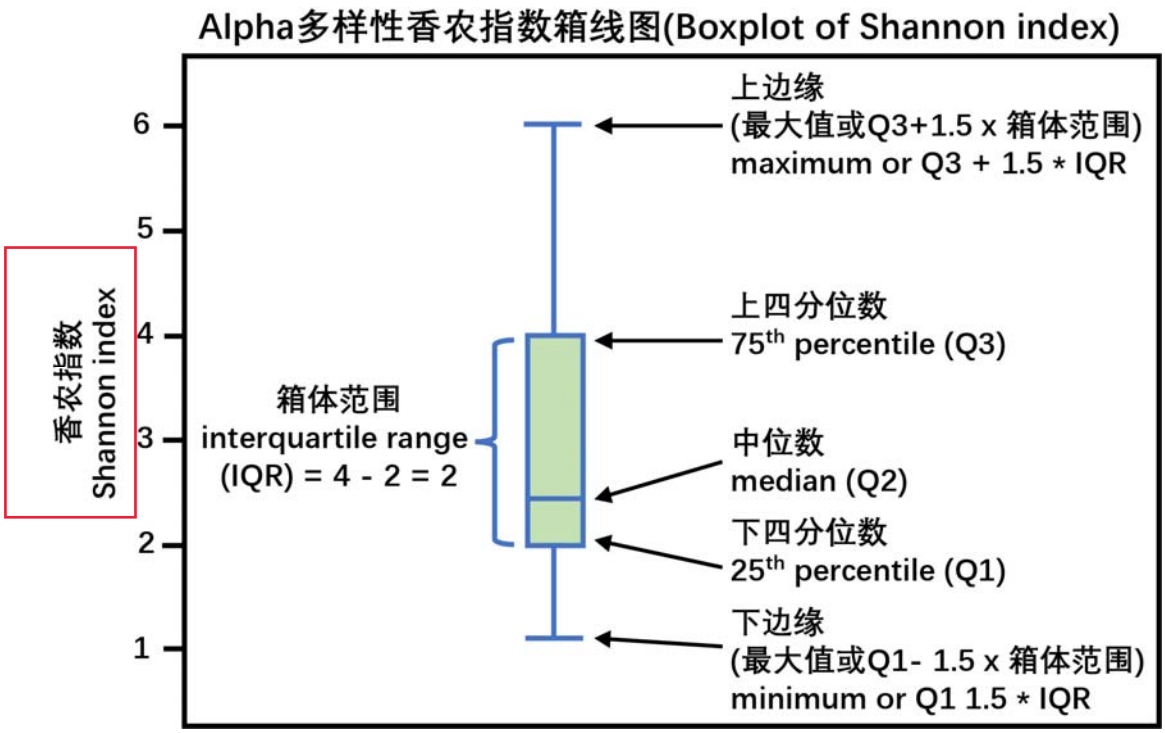



图 1. 以 Alpha 多样性最常用的香农指数(Shannon index)为例

- 第二种情况：最大或最小值超过 1.5 倍箱体范围，外位延长线外，即异常值 (outliers)：

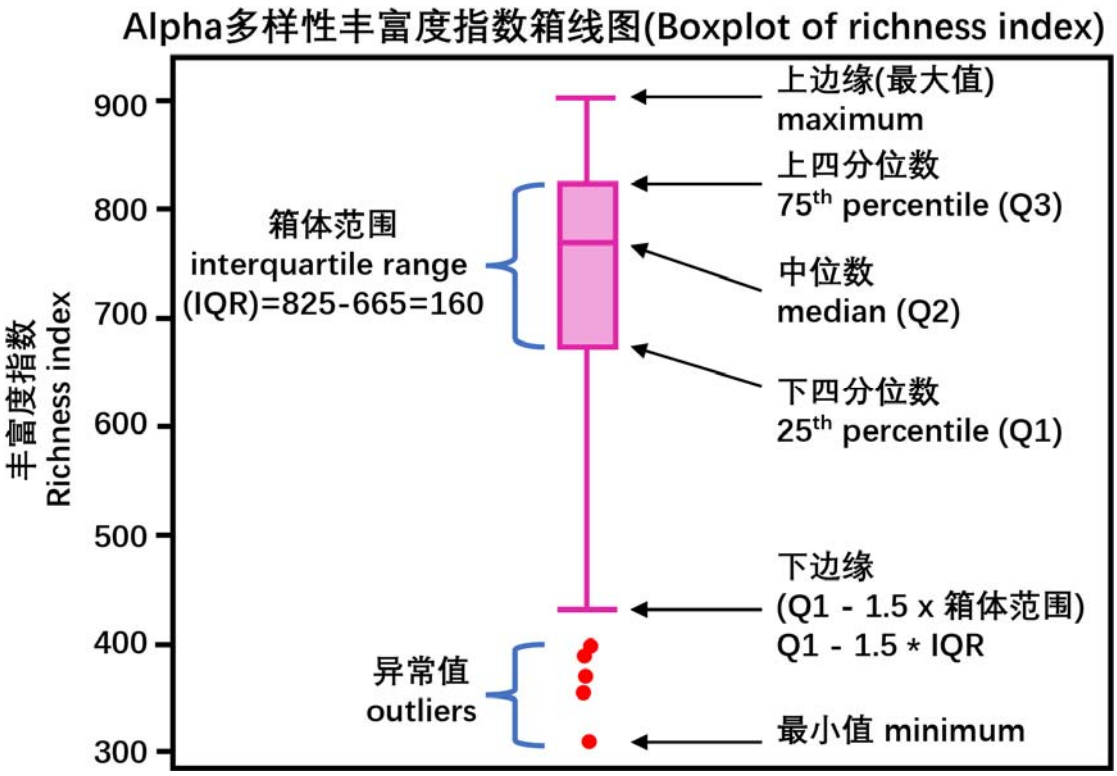



图 2. 以 Alpha 多样性丰富度指数(Richness index)为例

# 箱线图实例解读

1. Nature Biotechnology 2019 图 1e

此文章是中科院遗传发育所白洋团队于 2019 年发表于国际技术类顶级期刊 Naute Biotechnology(简称 NBT，$IF^{2018}$ = 31.864)的文章，介绍了水稻群体层面微生物组的研究并揭示了宿主机制调控根系微生物参与氮利用的现象。详细内容参见中文解读原文：- 《NBT 封面：水稻 NRT1.1B 基因调控根系微生物组参与氮利用》。

在这里，我们选择了文章中包含 alpha 多样性的描述性语句，方便大家更好的使用 alpha 多样性结果。然而我们也必须注意，alpha 作为一个经典和古老的指标，在如今的微生物

组学发展中已经不是核心指标了，所以~~刘老师~~这篇文章在讨论种并未对 alpha 多样性进行讨论，但是~~却~~对于下游微生物数据的分析方向有着重要的指导意义。

~~我们~~的学习思路是每个图提供图、图注和结果的描述，带领大家庖丁解牛式的学习每一种图的展示样式、图注内容和结果描述的中、英文表达方式，并批注一些注意事项、相关经验和套路的总结。~~使读者不仅~~看懂图，会写图注、写结果。最后提供参考代码绘制个性化的图表，让你独立完成科学论文中的每一部分，并能更好地传递科学发现。

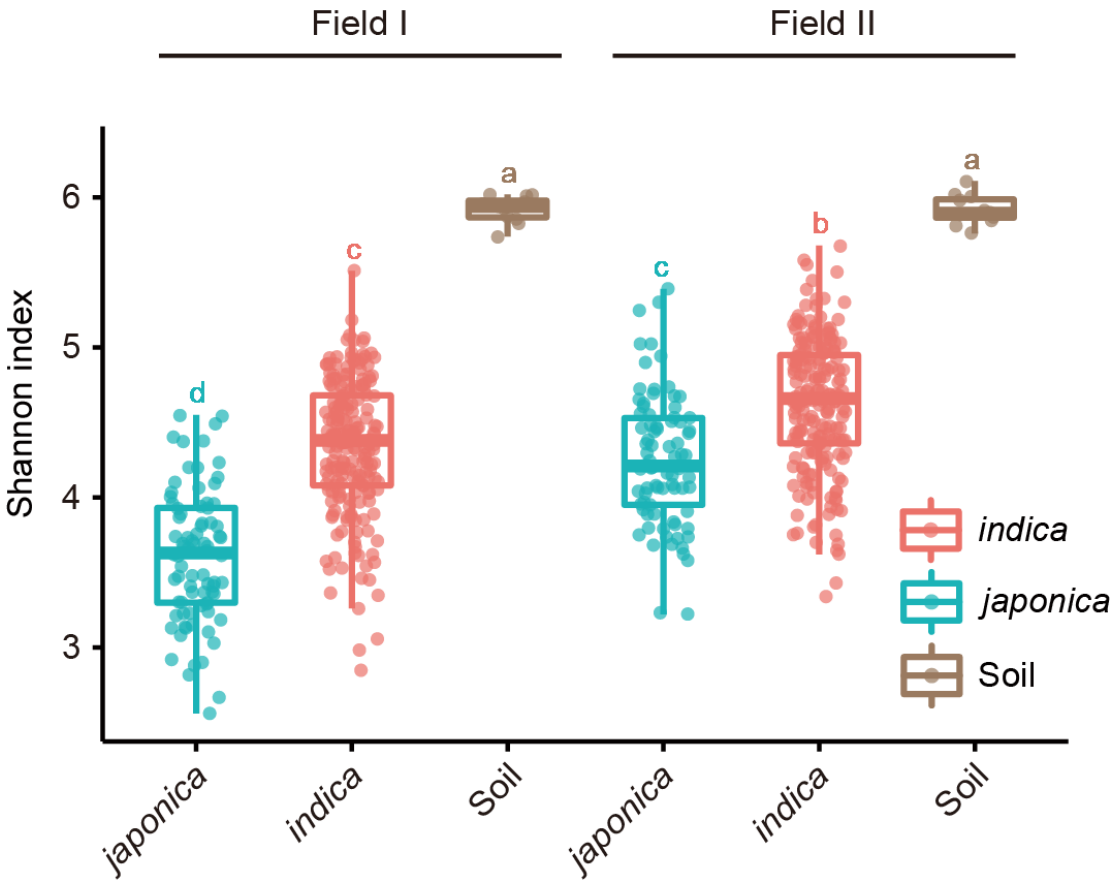



**图 3**. 箱线图展示籼粳稻和土壤的香农多样性指数。箱体上中下线分别为 75、50(中位数)和 25 分位数，轴须线最长不超过 1.5x 箱体范围。字母用于区分组间是否存在显著区别，不同字母表示组间存在显示差异($P < 0.05$，ANOVA, Tukey-HSD test)。图中的样本量如

下：地块 1：籼稻（*n* = 201），粳稻（*n* = 80），土壤（*n* = 12）；地块 2，籼稻（*n* = 201），粳稻（*n* = 81），土壤（*n* = 12）。

Shannon index of the microbiota of roots from indica, japonica and the corresponding bulk soils in two fields. The horizontal bars within boxes represent medians. The tops and bottoms of boxes represent the 75th and 25th percentiles, respectively. The upper and lower whiskers extend to data no more than 1.5× the interquartile range from the upper edge and lower edge of the box, respectively. The numbers of replicated samples in this figure are as follows: in field I, *indica* (*n* = 201), *japonica* (*n* = 80), soil (*n* = 12); in field II, *indica* (*n* = 201), *japonica* (*n* = 81), soil (*n* = 12).

图注描述注意事项：

1. 图表标题写法有两类，第一个类写做了什么(如本例)，另一类是写发现的结果(如 A 比 B 多)，以第一类使用较多，第二类更突出发现的规律但有时杂志不允许；
2. 箱线图要对箱体的上、中、下水平线，两端延长线的位置和意义进行描述。虽然是固定套路，但 Nature 系列杂志要求必须描述清楚；
3. 样本数量要在图注结果详细描述，每个实验组的样本量(*n* = xxx)，其中 n 要任何，等号两边要有空格；如果 *n* < 30，必须在箱线图中添加抖动图(jitter)展示每个样本点的分布位置。

**结果**：粳稻和籼稻的根系微生物的 alpha 多样性具有显著差异(图 1e 和附图 4)。两块地中粳稻根系微生物多样性显著高于粳稻(图 1e)，表明粳稻根系可以招募更多微生物种类。

Measurement of within-sample diversity (α-diversity) revealed a significant difference between indica and japonica varieties (Fig. 1e and Supplementary Fig. 4). The root microbiota of indica had higher diversity than those of japonica in both fields (Fig. 1e), indicating that indica roots recruited more bacterial species than japonica rice.

结果描述注意事项：

1. 一般提到显著(significant)就必须要描述准确的 P 值和统计方法，如($P$ = 0.03 或 $P$ < 0.05，ANOVA 和 Tukey HSD test 等方法)，但有时篇幅有限和感觉重复，只在方法部分定义，结果和图注中会省略，注意 $P$ 要斜体，<前后有空格。
2. 结果一般是图中信息的描述、比较和规律总结，有图时且已经发现了规律，写起来是非常容易的，要注意尽量陈述事实而不要过度引深或推断。

2. Gut Microbes 2020 图 1

这是南医大刘星吟团队发表在 Gut Microbes(简称 NBT，$IF^{2018}$ = 7.823)的文章，本研究揭示了肠道微生物谱的改变与孤独症谱系障碍的异常神经递质代谢活动相关。详细内容参见原文解读《Gut Microbes：南医大刘星吟团队揭示肠道微生物与孤独症相关》。

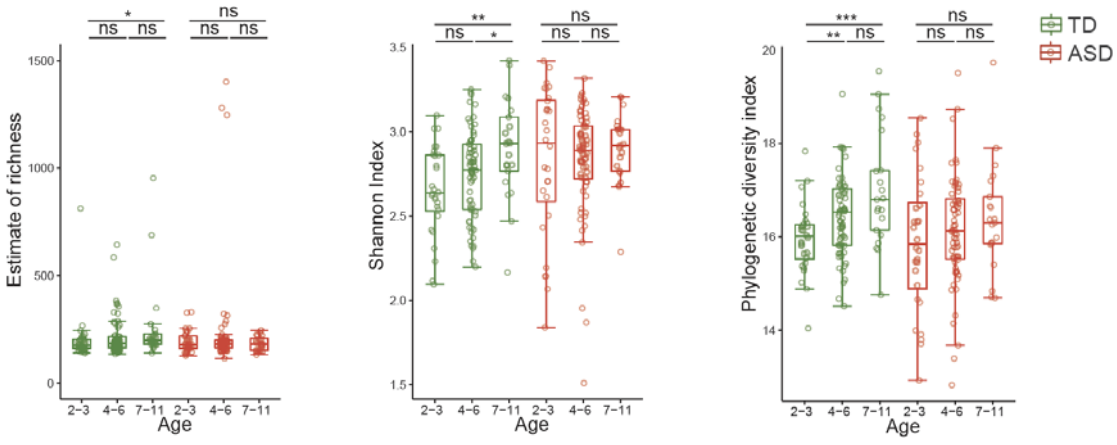


图 4. 基于年龄梯度的α多样性系数分析，包括丰富度估计量指数（H）、香农指数（I）和系统发育多样性指数（1J。

**附图注原文：**

(h), Shannon index (i), and phylogenetic diversity index (j).

**result**：我们进一步评估了与年龄有关的 细菌 alpha 多样性的变化。不同的 alpha 多样性指数仅反映了样本内多样性的一个方面。因此，我们使用三种方法来估算两组之间与年龄相关的 alpha 多样性变化。如图 1（h）所示，物种丰富度（breakaway estimates）显示出 TD 组的 7-11 岁年龄组与 2-3 岁的年龄组不同；然而，ASD 组没有显示出随着年龄增长而变化。香农指数解释存在物种的丰度和均匀性。如图 1（i）所示，4-6 岁年龄组的香农指数和 2-3 岁年龄组想比没有显着性 有所变化，但在 TD 组的 7 至 11 岁年龄组的 Shannon 指数显示出增加。系统发育多样性（PD）指数用于衡量两组之间的进化多样性差异程度。如图 1（j）所示，相比与 2-3 岁亚组，4-6 岁亚组的 PD 指数和 TD 组的 7-11 岁均增加。

We further assessed the age-related change of bacteria diversity. Different alpha diversity index reflects only one aspect of within-sample diversity; hence, we used three methods to estimate the age-related change in alpha diversity between the two groups. As shown in Figure 1(h), the richness of species (breakaway estimates) showed increased in 7–11 years age subgroup of TD group compared to 2–3 years age subgroup; however, the ASD group showed no change with age growth. Shannon index accounts for both abundance and evenness of species present. As shown in Figure 1(i), the Shannon index at the 4–6 years age subgroup showed no

significant change compared to the 2–3 years age subgroup, but the Shannon index at the subgroup of 7–11 years age in TD group showed increased compared to both 2–3 years and 4–6 years age subgroups, respectively. The phylogenetic diversity (PD) index was used to measure the degree of evolutionary divergence between two groups. As shown in Figure 1(j), the PD index of the subgroup of 4–6 years and 7–11 years in TD group was increased compared to 2–3 subgroups, respectively.

**result**：与 TD 组相比，C-ASD 组的物种丰富度和多样性显著降低。

Both species richness and diversity were significantly lower in the C-ASD group than in the TD group, as measured by breakaway and Shannon index analyses (Figure 3(b,c)).

同样是南医大刘星吟团队发表在 Gut Microbes 的文章。

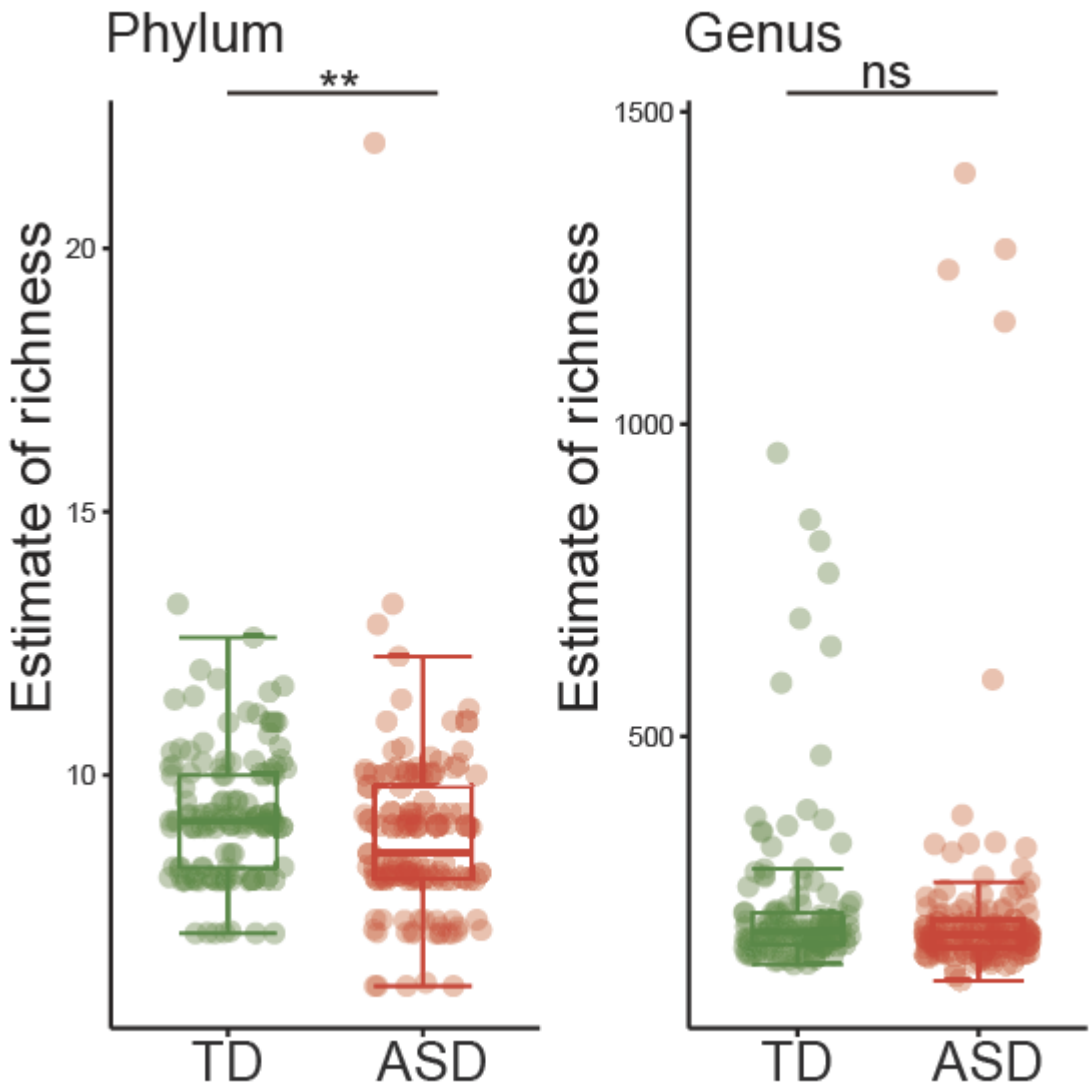



图 5. 在门 （b）和属 （c）水平评估微生物群落丰富度。

示例二和示例三均来自同一文章，所以接下来我一并流程本文中关于 alpha 多样性的摘要，讨论和结论部分内容，方便大家做一个宏观了解。

**Abstract** ： 孤独症组（ASD）的α多样性系数随着年龄的增加无显著的变化，而健康组（TD）的α多样性系数同年龄呈正相关，暗示孤独症患者肠道菌群发育处于相对停滞的状态。

We found that the α-diversity of ASD group showed no significant change with age, while the TD group showed increased α-diversity with age, which indicates that the compositional development of the gut microbiota in ASD varies at different ages in ways that are not consistent with TD group.

**Discussion** ： 另外：ASD 组幼儿的结果表明 alpha 多样性并没有年龄相关的变化。NC-ASD 组降低了 alpha 多样性，并改变了微生物组成。但是，C-ASD 组增加了 alpha 多样性，并进一步暗示便秘可能会增加肠道微生物异质性。

In addition, the α-diversity of ASD children showed no age-related change, while TD children showed increased α- diversity with age. NC-ASD showed decreased α - diversity and alternation of gut microbiota compared to TD.However, C-ASD showed increased α-diversity compared to NC-ASD, which further implicated that constipation might add heterogenous characteristics of gut microbiota in ASD.

**Conclusion**：ASD 组随着年龄的增加，其α多样性没有变化，而 TD 组α多样性随着年龄增长而增加，这提示 ASD 组的菌群发育可能处于相对停滞的状态。

Moreover, the α-diversity in the gut microbiota of ASD group showed no significant change with age; however, the TD group showed incre

**附图注原文：**

(b) The estimate of richness index analysis between two groups at the level of Phylum (b) and genus (c).

# 箱线图绘图实战

- 安装和加载 R 包：amplicon

从 github 上安装开发包；

```
# 基于 github 安装
# library(devtools)
# install_github("microbiota/amplicon")
```

如果 R 安装包存在问题。可以在宏基因组公众号后台回复"amplicon"，获得 R 3.6 版本的包合集下载链接。如 Windows 10 用户 R 包默认位置为 `文档 - R - win-library - 3.6` 目录，解压后替换 3.6 文件夹即可完成安装。

载入 R 包

```
suppressWarnings(suppressMessages(library(amplicon)))
```

- 数据读取和参数设定

这里有几个概念我们必须明确，由于我们的分析统计最终绝大多数在文章中表现为图片，所以需对文章中图形的布局和大小做一个简单的认识。

文章中纸图片尺寸为单栏 89 mm，双栏 183 mm，页面最宽为 247 mm，这里~~刘老师~~推荐比例 16：10，即半版 89 mm x 56 mm; 183 mm x 114 mm。

这里~~我还是要感叹感叹~~，绘图高手不一定是发文章高手，因为有数不清的细节需要学习注意，这就是其中之一。

```
# Data reading
metadata = read.table("./data/metadata.tsv", header=T, row.names=1, sep="\t",
comment.char="", stringsAsFactors = F)
head(metadata, n = 3)
# colnames of group ID in metadata
# 设置实验分组列名
group = "Group"
# Output figure width and height
```

```
# Letter 纸图片尺寸为单栏 89 mm，双栏 183 mm，页面最宽为 247 mm
# 推荐比例 16：10，即半版 89 mm x 56 mm; 183 mm x 114 mm
width = 89
height = 59

#--创建 alpha 文件夹用于结果保存

dir.create("./alpha")
# 手动指定分组列和顺序，默认为字母顺序
# metadata[[group]] = factor(metadata[[group]], levels = c("WT","KO","OE"))
# 按实验设计中分组出现顺序
# metadata[[group]] = factor(metadata[[group]], levels = unique(metadata[[group]]))
```

- 箱线图+统计 Boxplot+Statistics(alpha_boxplot)

使用?alpha_boxplot 查看函数功能和用法，本函数支持 14 种 alpha 多样性指标，并将结果调用方差分析进行统计检验，使用 TukeyHSD 函数进行多重比较。最后使用箱线图进行可视化。

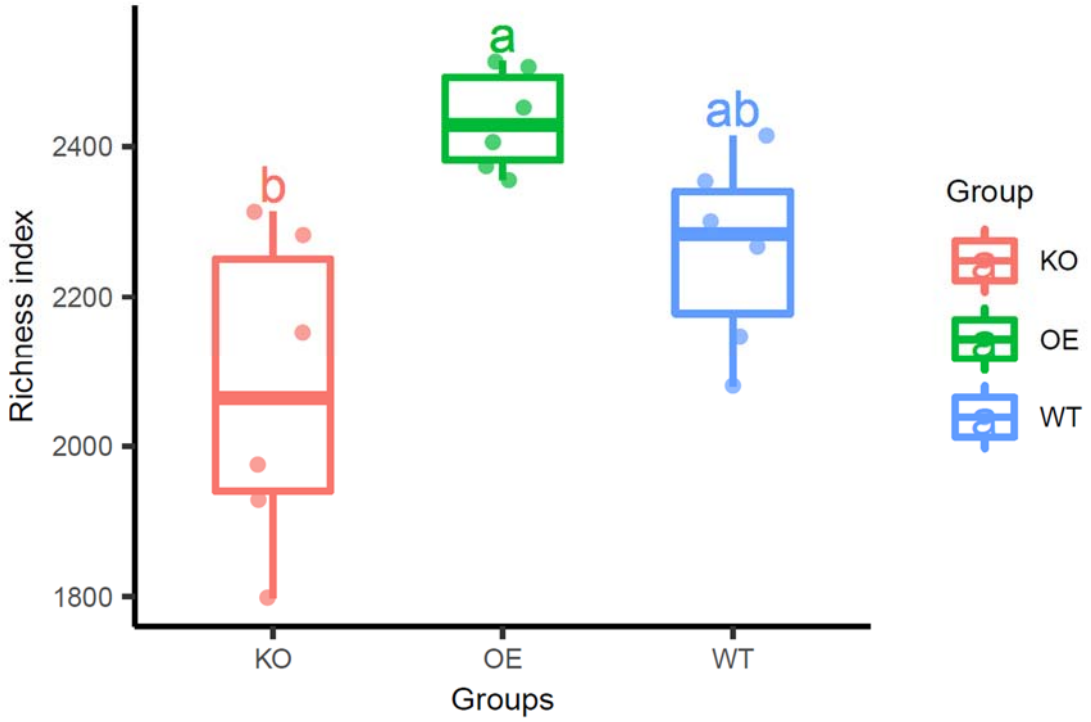
绘制主要分三步：

~~Plotting each figure mainly include three step:~~

1. 读取数据并预览格式 ~~Reading data and viewing format;~~
2. 参数调整和绘图 ~~Paramters adjustment and plotting;~~
3. 保存图片 ~~Saving figure~~.

```
# vegan.txt 中还有 6 种常用α多样性，alpha.txt 中有 11 种α多样性
alpha_div = read.table("./data/vegan.txt", header=T, row.names=1, sep="\t",
comment.char="")
head(alpha_div, n = 3)
# capitalize
library(Hmisc)
colnames(alpha_div) = capitalize(colnames(alpha_div))
```

```
colnames(alpha_div)
# 选择指数"Richness","Chao1","ACE","Shannon","Simpson","Invsimpson"
alpha_index = "Richness"

# Plotting alpha diversity Richness boxplot and stat
(p = alpha_boxplot(alpha_div, index = alpha_index, metadata, groupID = group))

# Saving figure
# 保存图片，大家可以修改图片名称和位置，长宽单位为毫米

ggsave(paste0("alpha/alpha_boxplot_",alpha_index,".pdf"), p, width = width, height
= height, units = "mm")

# 另存图片用于后期拼图
p1 = p

# 尝试探索不同的多样性各类
colnames(alpha_div)
alpha_boxplot(alpha_div, index = "Shannon", metadata, groupID = group)

# 尝试不同的分组方式
colnames(metadata)
alpha_boxplot(alpha_div, index = "Chao1", metadata, groupID = "Site")
```



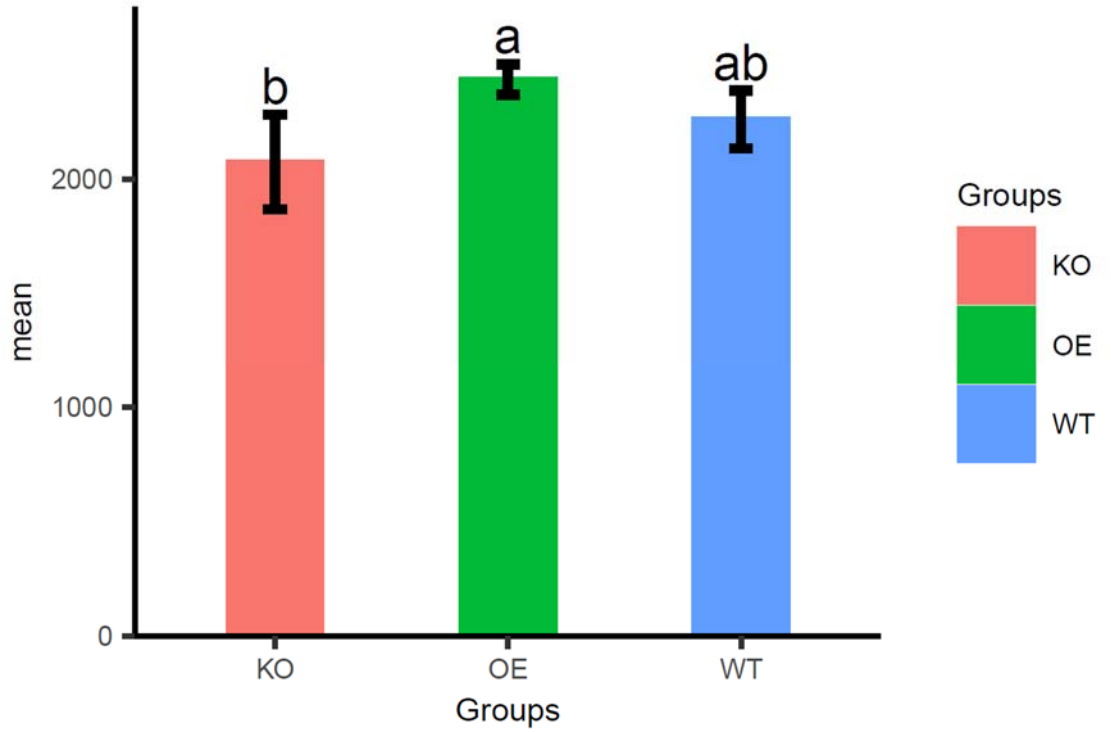

- 误差棒图+统计(alpha_barplot)

使用?alpha_barplot 查看函数功能和用法，本函数支持 14 种 alpha 多样性指标，并将结果调用方差分析进行统计检验，使用 TukeyHSD 函数进行多重比较。最后使用柱状图进行可视化。在主题功能上和上一个函数一样，只是修改了可视化的部分。

这给了我们另外的一种可视化 alpha 多样性的选择。其实柱状图及近年来在高水平文章种出现的次数越来越少了，尤其是单独作为文章正图或者一部分，因为从功能上来讲只能展示均值和方差，不如箱线图结合散点展示的多。最近在一篇 NBT 上看到将柱状图结合散点和误差线进行展示，解决的这种弊端。可能在接下来的使用中更多一些。

```
# vegan.txt 中还有 6 种常用α多样性，alpha.txt 中有 11 种α多样性
alpha_div = read.table("./data/vegan.txt", header=T, row.names=1, sep="\t",
comment.char="")
head(alpha_div, n = 3)
# capitalize
library(Hmisc)
colnames(alpha_div) = capitalize(colnames(alpha_div))
colnames(alpha_div)
# 选择指数"Richness","Chao1","ACE","Shannon","Simpson","Invsimpson"
alpha_index = "Richness"

# Plotting alpha diversity Richness boxplot and stat
(p = alpha_barplot(alpha_div, index = alpha_index, metadata, groupID = group))

# Saving figure
# 保存图片，大家可以修改图片名称和位置，长宽单位为毫米

ggsave(paste0("alpha/alpha_barplot_",alpha_index,".pdf"), p, width = width, height
= height, units = "mm")

# 另存图片用于后期拼图
p1 = p

# 尝试探索不同的多样性各类
colnames(alpha_div)
alpha_barplot(alpha_div, index = "Shannon", metadata, groupID = group)

# 尝试不同的分组方式
colnames(metadata)
```

```
alpha_barplot(alpha_div, index = "Chao1", metadata, groupID = "Site")
```




# Reference


1. Xu-Bo Qian, Tong Chen, Yi-Ping Xu, Lei Chen, Fu-Xiang Sun, Mei-Ping Lu & Yong-Xin Liu. A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. Chin. Med. J., doi:10.1097/CM9.0000000000000871 (2020).
2. Shannon, C.E. (1948). A mathematical theory of communication. The Bell System Technical Journal 27, 379-423.
3. Simpson, E.H. (1949). Measurement of Diversity. Nature 163, 688.
4. Chao, A., and Yang, M.C.K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. Biometrika 80, 193-201.

5. Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics 11, 265-270.
6. 刘尧博客：http://wap.sciencenet.cn/blog-3406804-1179809.html?mobile=1
7. 百度百科箱形图：https://baike.baidu.com/item/箱形图
8. Zhang, et.al. Transporter NTR1.1B contributes the association of root microbiota and nitrogen use in rice. 2019. Nature Biotechnology. doi: http://doi.org/10.1038/s41587-019-0104-4
9. Zhou Dan, Xuhua Mao, Qisha Liu, Mengchen Guo, Yaoyao Zhuang, Zhi Liu, Kun Chen, Junyu Chen, Rui Xu, Junming Tang, Lianhong Qin, Bing Gu, Kangjian Liu, Chuan Su, Faming Zhang, Yankai Xia, Zhibin Hu & Xingyin Liu. Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes, 1-22, doi:10.1080/19490976.2020.1747329 (2020).

## 稀释曲线

## 维恩图

维恩图的变形，如 UpsetView，网络图等。

# Beta 多样性

## 非限制性排序 PCoA/NMDS

1. 主成分分析 PCA
2. 主坐标分析 PCoA
3. 非度量多维尺度分析 NMDS
4. 对应分析 CA
5. 其他排序 pls-da，opls-da，t-sne

## 统计方法 PERMANOVA

1. PERMANOVA
2. ANOSIM
3. MRPP

## 限制性排序

1. 限制性主坐标分析 Constrinaed PCoA
2. 冗余分析 RDA
3. 典范对应分析 CCA
4. LDA

# 物种组成

## 堆叠柱状图

## 弦图

## 树图/气泡图

# 差异比较

## t 检验和秩和检验

## 匀二项分布和计数型差异分析 edgeR/DESeq2

1. 什么是物种数据的过度离散现象和负二项分布
2. 用 edgeR 包进行差异分析
3. DESeq2 包进行差异分析

## STAMP 与扩展柱状图

## LEfSe 和 Cladogram

## 其他常用差异分析方法

1. ANCOM 分析
2. ALDEx2 分析
3. songbird 和 DEICODE 介绍
4. limma

# 网络分析

## 网络基础知识

## 可视化入门

1. 按分类或模块着色网络
2. 网络属性
3. 全局属性
4. 节点属性

## 可视化进阶

1. 双网络比对
2. 多网络时间序列

# 机器学习：分类与回归

---

## 机器学习的常用算法

## 随机森林

1. 分类
2. 分类评估-ROC 曲线及 DCA 分析
3. 回归
4. 回归及效果评价

## Adaboost

## 深度学习

## 其他常用算法

1. 人工神经网络分类
2. 支持向量机分类
3. 逻辑回归（GLM）

# 树形图

---

## 进化树构建

1. 多序列比对
2. 建树 Fastree/RaxL
3. 宏基因组中建树 Phylophlan3
4. iTOL 美化进化树
5. ggtree 美化进化树

## 分类树构建

1. Graphlan 与 Cladogram
2. Krona

3. Metacoder

# 相关分析

## 特征与环境因子相关

## 特征间相关(同网络)

## 相关分析的可视化

# 统计学基础

## 正态性检验和方差齐性分析

## t 检验、方差分析、卡方检验使用注意事项

## 两组和多组秩和检验

## 多重比较的 P 值校正

## 物种数据标准化方法和注意事项

# 文章套路总结

## 扩增子

## 宏基因组

### 参考基因集

### 碳水化合物

### 抗生素抗性

## 扩增子+宏基因组

## 其他研究热点

- 人类：肠型、肥胖、二型糖尿病、IBD、早产、关联分析
- 动物：无菌小鼠、牛瘤胃、食性、宿主和微生物共进化
- 植物：根际、叶际、代谢物、氮利用、抗病
- 环境：抗生素耐药、抗生素挖掘、极端环境、生命之树

# 附录

# 实验设计

实验方案，样本元数据收集，样本名命名规则和示例。

# 测序平台和测序技术

# 数据备份与发布

NCBI，GSA，EBI

# 图片排版和美化

# 杂志点评

CNS，Microbiome，ISME

相关文章按杂志分类

# 机遇与挑战

目前的优缺点和不足，未来的发展方向。

# 三代测序

NBT 的 PacBio 和 ONT 文章简介

# 宏基因组精品文章分类

按专题分类