Alpha 多样性箱线图

刘永鑫，文涛，钱旭波

2020/6/17

《微生物组数据分析与可视化实战》专著

- 众筹编写《微生物组数据分析与可视化实战》——成为宏基因组学百科全书的创始人(目录)
- 编者序：初衷、计划、要求、优势、目标和展望

本文为样章，211 代表目录中的第二章，第一节的第一部分，部分引用格式文字为格式说明。

每节的基本逻辑：背景知识 —— 实例解读 —— 实战代码，三步走

1. 背景知识：基本概念讲解，扫清阅读障碍，成为专业人士；本文主题是 Alpha 多样性箱线图，重点讲解 alpha 多样性概述和常用指数、箱线图的概念和基本元素，同时注意尽量图文并茂，本文中大部分图为作者手绘，引用他人文章原图，要规范引用并获得作者授权。
2. 实例解读：通过近 3 年发表的高质量文章中 2 ~ 5 个矢量图、图注、以及相关的结果、讨论、方法的中英文描述，方便同行读懂相关文章，同时在撰写中、英文章中提供可参考的模板；同时要总结每个图的优缺点和注意事项，让同行获得更高层级的理解和经验，在自己的研究中可以做的更好；本节选取了两篇近一年内发表的高水平文章中的 3 组矢量图为例，提取文中相关英文描述段落，并配中文翻译，最后附上作者的总结和点评的改进建议。
3. 实战代码：本质上是分析的教程，参考 Nature Protocols 分析类文章的思路，必须有软件环境要求(系统和软件版本)和安装方法、可重复分析的代码(推荐 Rmarkdown 可一键运行，尽量打包成函数，避免代码冗余)、提供测试数据；最好还有软件安装和操作的视频(推荐但不必须)，参考 JoVE 杂志的视频格式 JoVE 微生物组专刊征稿，写方法拍视频教程发 SCI，视频录制推荐使用 截图、录屏全能王 FastStoneCapture。

特征表的分析、可视化和解读

Alpha 多样性分析

Alpha 多样性与箱线图

本节作者：刘永鑫，文涛，钱旭波

版本 1.0.2，更新日期：2020 年 6 月 17 日

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 et al., 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：

- 观测的特征数（Observed OTU/ASV）：是指每个样本中能够观察到的 OTUs 或 ASVs 的数量，与每个 OTU 或 ASV 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：Chao1 = Sobs + n1(n1-1)/2(n2+1)，其中 Chao1 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，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 定义为稀有并单独计算，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。计算公式详见 http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html

- 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。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，所以高丰度物种较多时该指数值较大，这与香农指数有明显区别。
- Pielou 的均匀度指数（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 个指标很常用。

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；

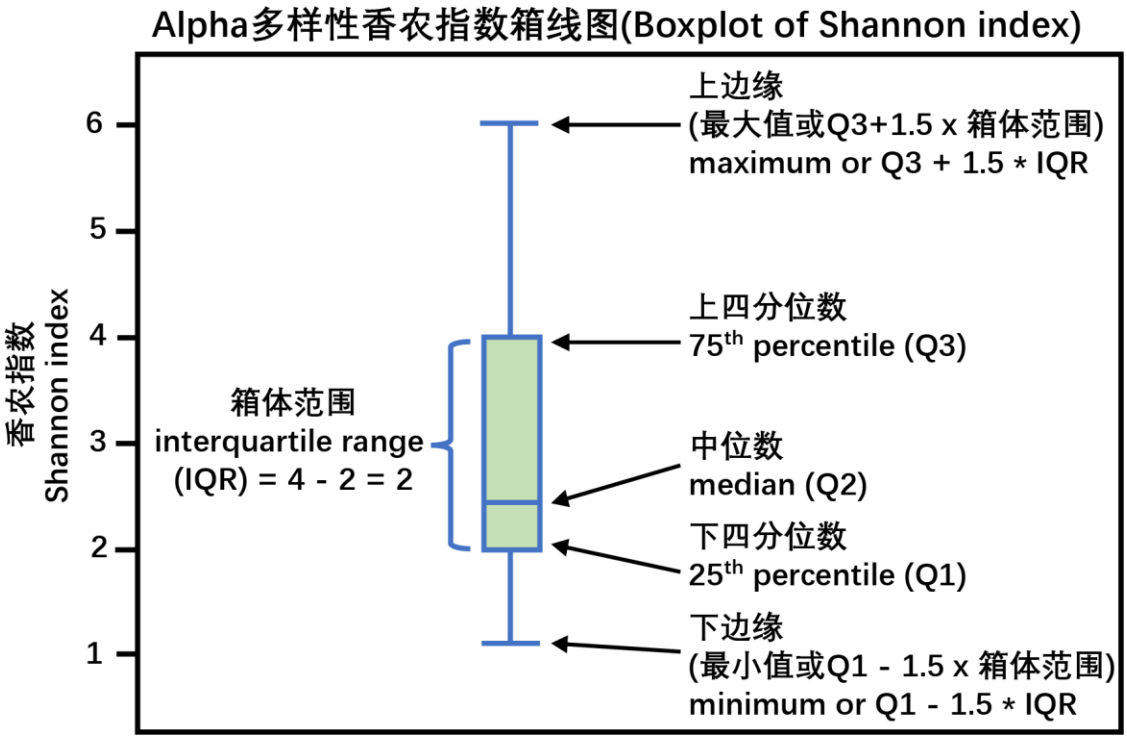


图 2. 以 Alpha 多样性最常用的香农指数(Shannon index)为例

- 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)：

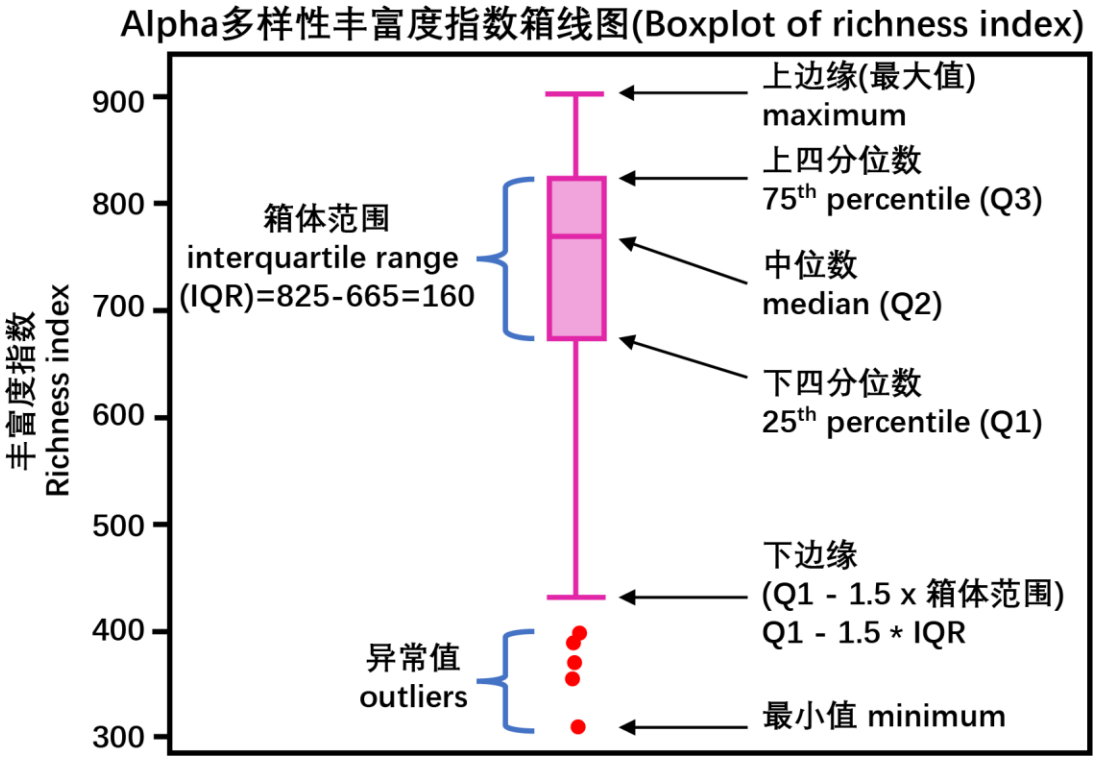


图 3. 以 Alpha 多样性丰富度指数(Richness index)为例

实例解读

例 1. 两地点组间香农指数比较

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，IF2018 = 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调控根系微生物参与氮利用的现象。详细内容参见作者的文章解读：- 《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，我们选择了文章中图 1e 为例进行讲解，并包含 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，所以刘永鑫负责分析的这篇文章在讨论中并未对 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。

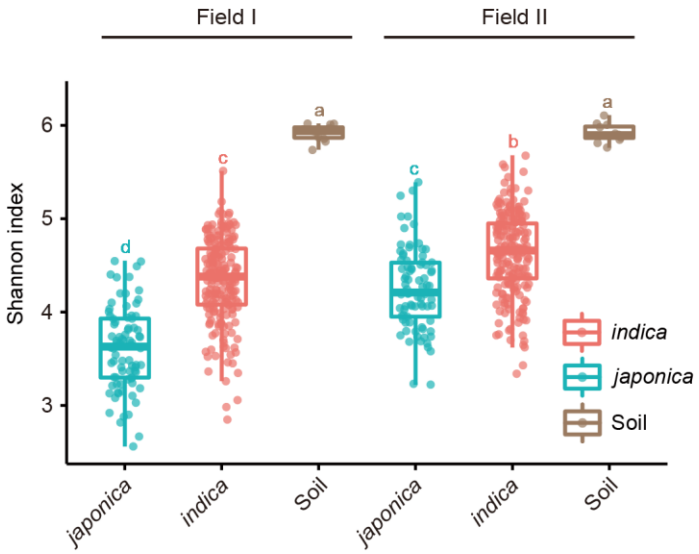


图 4. 箱线图展示籼粳稻和土壤的香农多样性指数(Zhang et al., 2019)。箱体上中下线分别为 75、50(中位数)和 25 分位数，轴须线最长不超过 1.5x 箱体范围。字母用于区分组间是否存在显著区别，不同字母表示组间存在显示差异($P < 0.05$，ANOVA, Tukey-HSD test)。图中的样本量如下：地块 1：籼稻 ($n = 201$), 粳稻 ($n = 80$), 土壤 ($n = 12$); 地块 2, 籼稻 ($n = 201$), 粳稻 ($n = 81$), 土壤 ($n = 12$)。

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 要斜体，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。
4. 本文由于整体结果过多，只使用了香农指数。如果文章结果不是特别多，可以使用多种指数同时展示，使结果充实，同时突出结果的重复性、稳定性更好，详见下面的例子。

结果：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。

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。

例 2. 时间梯度多指数比较

这是南医大刘星吟团队发表在 Gut Microbes(简称 NBT，IF2018 = 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》。这里以文中的图 1 中的 h-j 为例

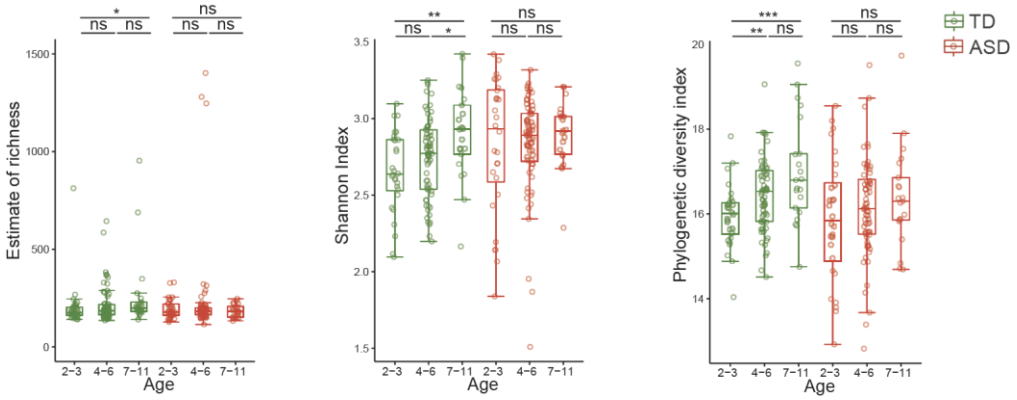


图 5. 基于健康(typically developing, TD)组和自闭症(Autism Spectrum Disorder, ASD)组年龄梯度的 α 多样性系数分析(Dan et al., 2020)，包括丰富度估计量指数(h)、香农指数(i)和系统发育多样性指数(j)。

Alpha diversity indices of genus for TD and ASD according to age from 2 to 11, Estimate of richness (h), Shannon index (i), and phylogenetic diversity index (j)

摘要： 自闭症组的 α 多样性系数随着年龄的增加无显著的变化，而健康组的 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.

结果：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，但在健康组的 7 至 11 岁年龄组的香农指数显示出增加。系统发育多样性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.

讨论：另外，自闭症组幼儿的结果表明 alpha 多样性并没有年龄相关的变化。非便秘自闭症(non-constipated ASD, NC-ASD)组降低了 alpha 多样性，并改变了微生物组成。但是，便秘自闭症(constipated ASD, 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.

结论：自闭症组随着年龄的增加，其 α 多样性没有变化，而对照组 α 多样性随着年龄增长而增加。确定了与便秘自闭症(C-ASD)中代谢物改变相关的变化物种。

Moreover, the α-diversity in the gut microbiota of ASD group showed no significant change with age; however, the TD group showed increased diversity. The changed species associated with metabolite alteration in C-ASD were identified.

- **总结**

1. 时间序列可以与多样性进行相关分析，将在后面的相关分析章节详细讲到；
2. 时间序列的数据有时无法得到较好的相关性或规律，可尝试随机森林回归分析，也可尝试**将时间进行人为分组，转化为通常 2 ~ 7 组的分类型变量**，将复杂的时间序列分析**转化为最常用的组间比较问题**，最常用分 3 组(本文示例，因为 2 组只有一种比较方案，但 3 组有 3 种比较方案，更多组过于复杂)；
3. 本图将样本点和统计结果标注在图上，是非常规范的作图方案；同时推荐将各分组的样本量情况写在图注中，参考上图 NBT 文中的示例；

例 3. 组间物种分类级展示

同样是南京医科大学刘星吟团队 2020 年发表在 Gut Microbes 杂志上文章中图 1 的 b 和 c。

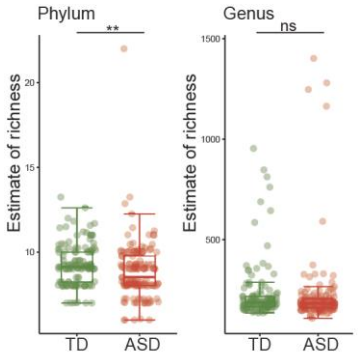


图 6. 在门（b）和属（c）水平评估微生物群落丰富度(Dan et al., 2020)。

附图注原文： > (b) The estimate of richness index analysis between two groups at the level of Phylum (b) and genus (c).

结果: 与健康组相比，便秘自闭症组的物种丰富度和多样性显著降低。

The richness of species(breakaway estimates) at the phylum level wassignificantly lower in the ASD group than that inthe TD group (Figure 1(b)). However, there wereno significant differences in richness between theASD group and TD group at the genus level(Figure 1(c)).

- **总结**

1. 通常 Alpha 多样性在 ASV 和 OTU 层面分析，测序结果的聚类和去噪也可以比较好的反映多样性的情况；
2. 物种注释常用 7 级分类法：界、门、纲、目、科、属和种；扩增子测序读长短，通常只在属水平较准确，所以将 OTU/ASV 归类为属水平再分析多样性也是不错的选择。同时门水平，因为数量有限，人类可读性更好，在分析中也比较常用。本文在门、属水平展示多样性，是比较典型的应用，同时对 OTU/ASV 层面进一步的证明和补充。

绘图实战

测试数据和代码准备

数据和代码下载

为了保证数据和代码的可用性和安全性，我们将结果同步保存于 Github、个人服务器和百度云三处（狡兔三窟，这对数据分析人员极其重要）。

通常以下三种方式通过任选其一，可获得原始数据和代码，但三种都会可极大提高数据获取的成功率：

- 方法 1. 项目 Github 地址：https://github.com/YongxinLiu/MicrobiomeStatPlot ，可以点击 Clone or download —— Download ZIP 下载整个项目，链接为 https://github.com/YongxinLiu/MicrobiomeStatPlot/archive/master.zip ，但不推荐全部下载，因为项目每天更新，而且很多文件你可能用不到，随时项目

发展体积会非常大。 可选单个文件下载，如本节的目录为 `211AlphaBoxplot`；点击文件夹中具体的文件，非文本文件会出现 `Download` 按扭下载；文本文件直接可预览、编辑和复制保存，也可在 `Raw` 按扭点右键另存文件。注：Github 也可能经常不可用，可以换个时间尝试，有些地区、或有些时期根本可能无法访问 Github，请尝试另外两种方法。

- 方法 2. 公网下载：个人备份服务器，http://210.75.224.110/github/MicrobiomeStatPlot/ 再添加在 github 中对应的具体目录及文件名，可作为备选的下载方式，如本文的 markdown 笔记下载地址：http://210.75.224.110/github/MicrobiomeStatPlot/211AlphaBoxplot/211.Alpha 多样性箱线图.md 。此方式也可能会因为服务器故障或网络安全原因无法访问，请尝试另外两种方法。

- 方法 3. 百度云下载，大文件压缩包如 R 包合集、数据库等分享，链接详见 https://github.com/YongxinLiu/MicrobiomeStatPlot/blob/master/Data/BigDataDownlaodList.md 页面。文件夹文件过多时，基本无法分享，但可以发送好友，大家可以加我百度帐号 woodcorpse 为好友，定期在百度云群里分享项目的文件。

注：以上三种下载方式，一种不行马上尝试另一种，再不行尝试第三种，如果确定三种方式均无效，请阅读本文留言区补充信息寻找备用链接(微信推文无法修改，但可以留言补充)。如果是操作没经验，请观看下面 3 种方式的操作视频。

视频教程：Windows 下 Github、公网和百度云的 3 种下载方式。

视频链接：https://v.qq.com/x/page/u0978k38dl5.html

软件和数据库安装

本教程需要在 R 语言环境下运行，推荐在 RStudio 界面中学习更高效。以 Windows 10 环境，R 3.6.3 和 RStudio 1.2.5019 为例进行过测试。理论上 Mac、Linux 系统，以及 R 或 RStudio 的更新版本是兼容的，但并没有广泛测试，有问题欢迎自行解决并在群在与同行分享经验。2020 年 6 月 7 日发布前测试，软件和代码全为最新，也可轻松安装，顺利使用。

- R 语言：https://www.r-project.org/ 下载最新版：Downad CRAN - China Tsinghua - Download R for Windows/Mac —— base —— Download，本次为 R 4.0.1+；安装时注意语言选择英文，减少出现乱码；
- RStudio，R/Shell 语言运行界面: https://www.rstudio.com/products/rstudio/download/#download ，下载最新版，如本次为 1.3.959+
- Rtools40，源码包编译工具: https://cran.r-project.org/bin/windows/Rtools/

安装和加载 R 包：amplicon

视频教程：Alpha 多样性箱线图代码运行和讲解

视频链接：https://v.qq.com/x/page/b0978gl6by2.html

检查依赖关系是否安装，有则跳过，无则自动安装。

```
# 基于github 安装包，需要devtools，检测是否存在，不存在则安装
if (!requireNamespace("devtools", quietly = TRUE))
    install.packages("devtools")
# 注：提示安装源码包的最新版时，推荐选否，加速安装预编译的稳定版。原码包编译时间长且容易出错
# 第一次运行，会自动在C:\Users\User\Documents\R\win-library\4.0 目录中安装75 个包
# 加载github 包安装工具
library(devtools)

# 检测amplicon 包是否安装，没有从源码安装
if (!requireNamespace("amplicon", quietly = TRUE))
    install_github("microbiota/amplicon")
# 提示升级，选择3 None 不升级；升级会容易出现报错
# library 加载包，suppress 不显示消息和警告信息
suppressWarnings(suppressMessages(library(amplicon)))
```

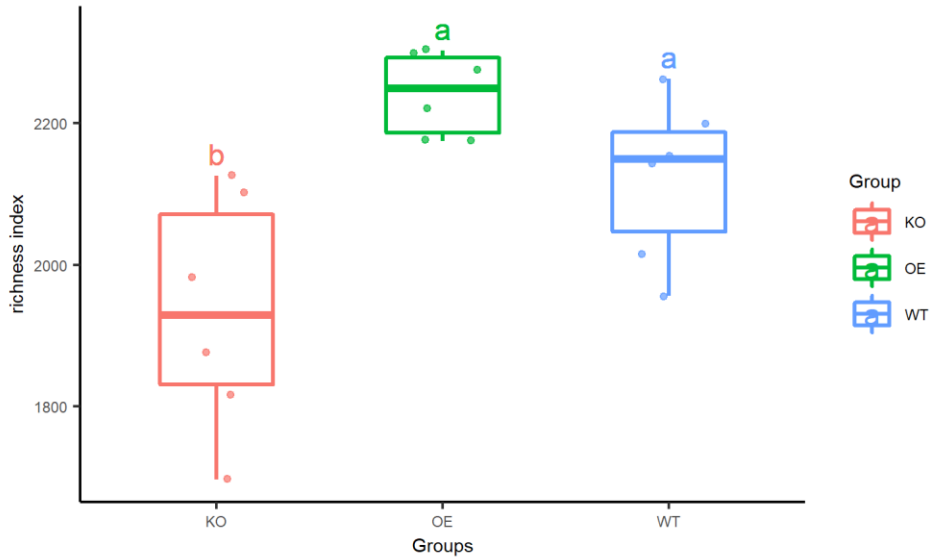
本项目更新较快，建议使用中存在问题，运行 install_github 行安装最新版。

alpha_boxplot 函数

在 amplicon 包中有 alpha_boxplot 函数可以一行命令快速绘制箱线图+统计标记的图

本次绘制使用函数内置数据，进行快速演示；查找命令使用，可打问题(?)+命令名。

```
# 查看函数帮助
?alpha_boxplot
# 使用内置数据，绘制以Group 分组下的丰富度指数
(p = alpha_boxplot(alpha_div, metadata, "richness", "Group"))
```




```
ggsave(paste0("7.png"), p, width=89, height=56, units="mm")
```

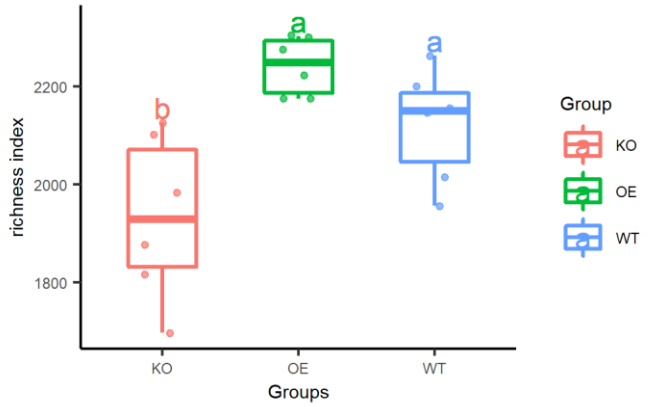


图 7. 箱线图展示 Alpha 多样性丰富度(richness)指数在三组内在分布和组间统计。不同字母代表组间存在显著差异(Adjust $P < 0.05$，ANOVA，Tukey HSD test)。

从图中可以看到 KO(基因敲除，knock-out)组与 OE(过表达，over-expression)和 WT(野生型，wild-type)组存在显著差异(字母不同)，即基因的有无可对微生物群落的丰富度引起显著变化。同时观察到丰富度在三组间 KO < WT < OE，根据背景知识这三组的基因表达量也是逐渐升高的，因此可以推断该基因的表达可能在促进菌群丰富度中起重要作用。而且很多研究报导疾病组一般多样性较低，因为可以考虑进一步研究该基因在维持宿主健康中的作用。

绘制精讲

绘制主要分三步：

1. 读取数据并预览格式;
2. 参数调整和绘图;

3. 保存图片。

本次测试数据来自我们负责分析的一篇 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 Science：拟南芥三萜化合物特异调控根系微生物组

原文实验较复杂，这是只截取了 3 个实验组各 6 个样品的结果用于演示。数据位于 Data/Science2019 目录，本次需要元数据(metadata.txt)和 Alpha 多样性指数(alpha/vegan.txt)两个输入文件。

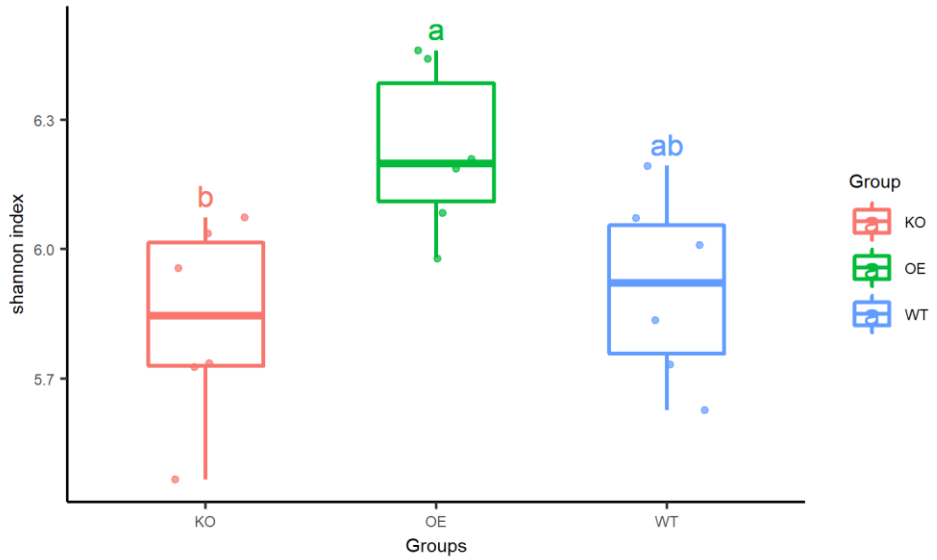
```
# 读取元数据，参数指定包括标题行(TRUE)，列名为1 列，制表符分隔，无注释行，不转换为因子类型
metadata = read.table("../Data/Science2019/metadata.txt", header=T, row.names=1, sep="\t", comment.char="", stringsAsFactors = F)
# 预览元数据前3 行，注意分组列名
head(metadata, n = 3)
```

```
    Group      Date      Site        CRA       CRR BarcodeSequence
KO1    KO 2017/6/30  Chaoyang CRA002352  CRR117575      ACGCTCGACA
KO2    KO 2017/6/30  Chaoyang CRA002352  CRR117576      ATCAGACACG
KO3    KO  2017/7/2 Changping  CRA002352 CRR117577      ATATCGCGAG
    LinkerPrimerSequence      ReversePrimer Batch              Species
KO1  AACMGGATTAGATACCCKG ACGTCATCCCCACCTTCC     1 Arabidopsis thaliana
KO2  AACMGGATTAGATACCCKG ACGTCATCCCCACCTTCC     1 Arabidopsis thaliana
KO3  AACMGGATTAGATACCCKG ACGTCATCCCCACCTTCC     1 Arabidopsis thaliana
    Sequencing  Platform           Description
KO1   Novagene HiSeq2500 Knock-out replicate 1
KO2   Novagene HiSeq2500 Knock-out replicate 2
KO3   Novagene HiSeq2500 Knock-out replicate 3
```

```
# 读取vegan 计算6 种alpha 多样性指数，计算方法见"分析流程 - 扩增子"部分
alpha_div = read.table("../Data/Science2019/alpha/vegan.txt", header=T, row.names=1, sep="\t", comment.char="")
# 预览多样性指数前3 行，注释各指数列名
head(alpha_div, n = 3)
```

```
    richness    chao1      ACE  shannon   simpson invsimpson
KO1     2125 2432.412 2443.642 6.036140 0.9897723   97.77403
KO2     2089 2429.512 2419.509 6.074251 0.9913350  115.40716
KO3     1756 2060.862 2072.541 5.726450 0.9882047   84.77983
```

```
# 绘制各组香农指数分布，外层()可对保存的图形同时预览
(p = alpha_boxplot(alpha_div, index = "shannon", metadata, groupID = "Group"))
```




```
# 保存图片，指定图片为pdf格式方便后期修改，图片宽89毫米，高56毫米
ggsave(paste0("alpha_boxplot_shannon.pdf"), p, width=89, height=56, units="mm")
ggsave(paste0("8.png"), p, width=89, height=56, units="mm")
```

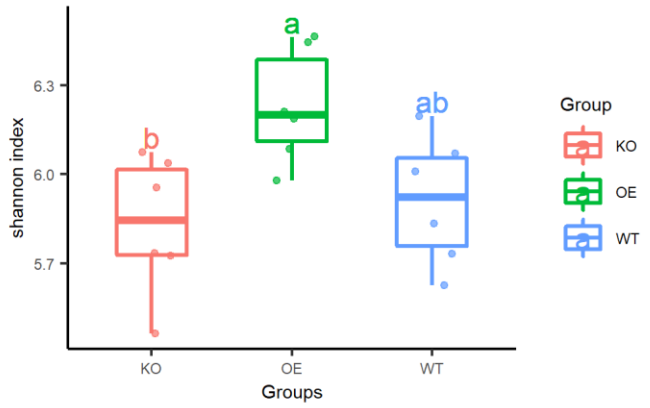


图 8. 箱线图展示 Alpha 多样性香农(shannon)指数在三组内在分布和组间统计。

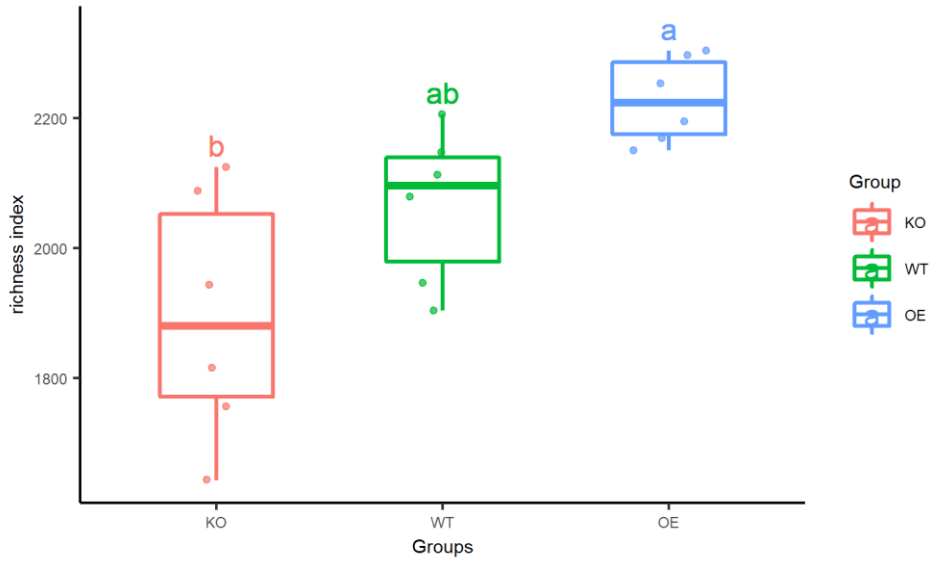
我们看到与丰富度相似，但又不完全相同的结果。在 Shannon 指数角度，只有 KO 和 WT 组存在显著差异。

常用技巧

修改分组顺序

我们经常要按照一定的逻辑指定分组顺序。如图 7 中发列多样性分布存在一定规律，我们想按多样性由小至大顺序手动重排分组。在 R 语言中，可以通过设置 level 来指定顺序

```
metadata$Group = factor(metadata$Group, levels = c("KO","WT","OE"))
(p = alpha_boxplot(alpha_div, metadata, "richness", "Group"))
```




```
ggsave(paste0("9.png"), p, width=89, height=56, units="mm")
```

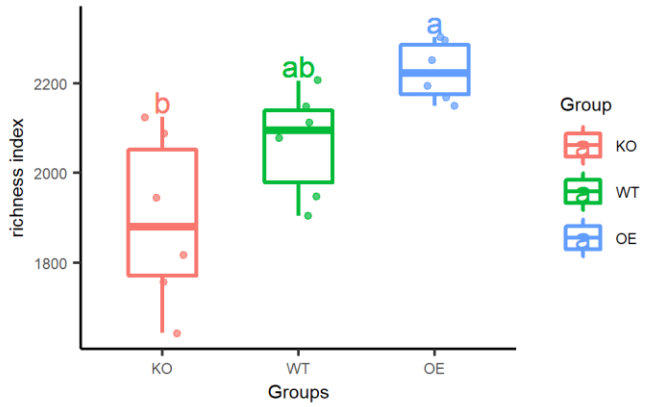
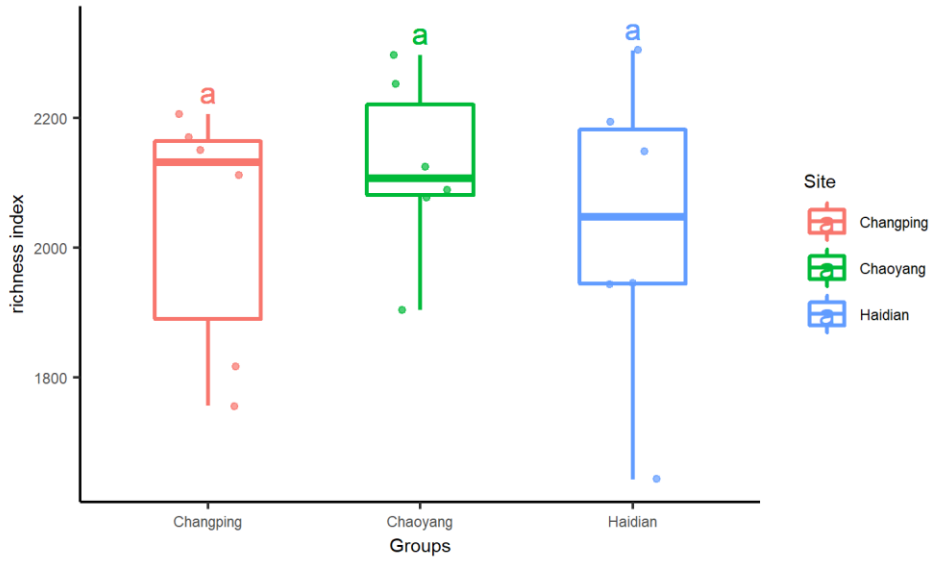


图 9. 手动指定分组顺序，表达观察到的规律。

讨论干扰因素是否显著影响多样性

实验中还可能涉及多时间、多地点、不同测序批次、序列标签、DNA 提取方法不同等多种混淆因子。如果有分组内存在以上因素，判断它们是否引起群落多样性的变化是至关重要的。本研究中每个批次还涉及多个实验地点，以判断不同实验地点或批次是否会结果有影响。我们将分组列指定为地点(Site)

```
(p = alpha_boxplot(alpha_div, metadata, "richness", "Site"))
```




```
ggsave(paste0("10.png"), p, width=89, height=56, units="mm")
```

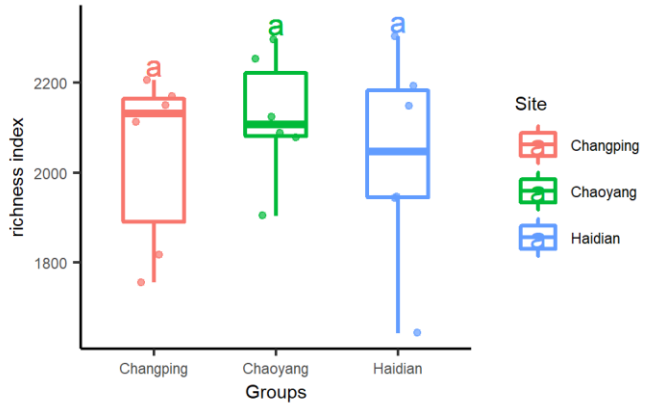


图 10. 讨论不同地点是否对 Alpha 多样性存在影响，图中显示无显著影响。

查看函数原代码进一步修改

只输入函数名称，不加后面的括号和任何参数，即显示函数的全部代码。

你可以复制输出的代码，在文档中修改更加个性化的分析结果。

```
alpha_boxplot

function(alpha_div, metadata, index = "richness", groupID = "Group") {
  # 依赖关系检测与安装
  p_list = c("ggplot2", "dplyr", "multcompView") # "agricolae"
  for(p in p_list){
    if (!requireNamespace(p)){
    install.packages(p)}
    suppressPackageStartupMessages(library(p, character.only = TRUE, qu
ietly = TRUE, warn.conflicts = FALSE))
```

```
  }

  # 测试默认参数
  # library(amplicon)
  # index = "richness"
  # groupID = "Group"
  # metadata = subset(metadata, Group %in% c("KO","OE"))

  # 交叉筛选
  idx = rownames(metadata) %in% rownames(alpha_div)
  metadata = metadata[idx,]
  alpha_div = alpha_div[rownames(metadata),]

  # 提取样品组信息,默认为 group 可指定
  sampFile = as.data.frame(metadata[, groupID],row.names = row.names(me
tadata))
  # colnames(sampFile)[1] = "group"

  # 合并 alpha_div 和 metadata
  df = cbind(alpha_div[rownames(sampFile),index], sampFile)
  colnames(df) = c(index,"group")

  # 统计各种显著性
  model = aov(df[[index]] ~ group, data=df)
  # 计算 Tukey 显著性差异检验
  Tukey_HSD = TukeyHSD(model, ordered = TRUE, conf.level = 0.95)
  # 提取比较结果
  Tukey_HSD_table = as.data.frame(Tukey_HSD$group)

  # 保存统计结果
  # 保存一个制表符，解决存在行名时，列名无法对齐的问题
  write.table(paste(date(), "\nGroup\t", groupID, "\n\t", sep=""), file
=paste("alpha_boxplot_TukeyHSD.txt",sep=""),append = T, quote = F, eol
= "", row.names = F, col.names = F)
  # 保存统计结果，有 warning 正常
  suppressWarnings(write.table(Tukey_HSD_table, file=paste("alpha_boxpl
ot_TukeyHSD.txt",sep=""), append = T, quote = F, sep="\t", eol = "\n",
na = "NA", dec = ".", row.names = T, col.names = T))

  # 函数：将 Tukey 检验结果 P 值转换为显著字母分组
  # 输入文件为图基检验结果和分组
  generate_label_df = function(TUKEY, variable){
    # library(multcompView)
    # 转换 P 值为字母分组
    ## 提取图基检验中分组子表的第 4 列 P adjust 值
    Tukey.levels = TUKEY[[variable]][,4]
    ## multcompLetters 函数将两两 p 值转换为字母，data.frame 并生成列名为 Let
```

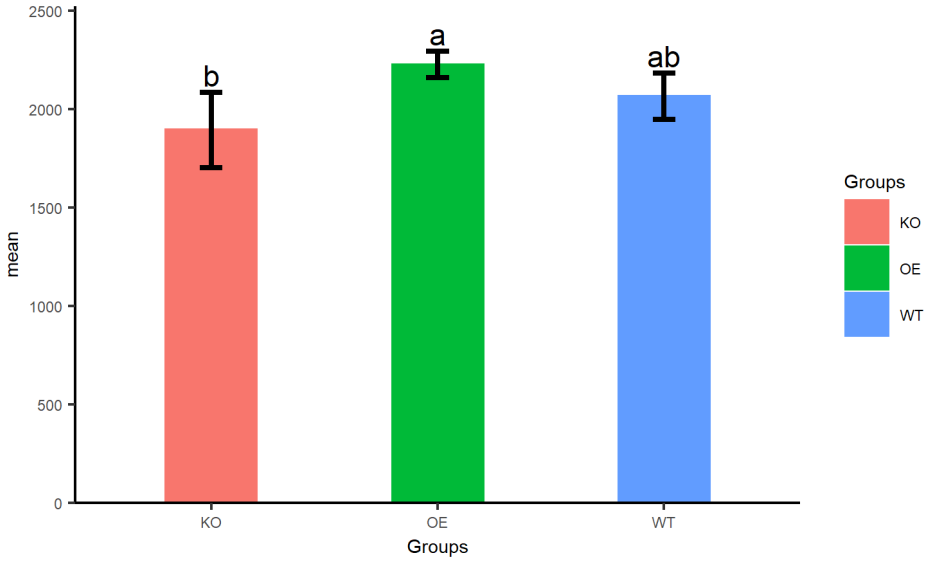
```
ters 的数据框
    Tukey.labels = data.frame(multcompLetters(Tukey.levels)['Letters'])
    # 按分组名字母顺序
    ## 提取字母分组行名为 group 组名
    Tukey.labels$group = rownames(Tukey.labels)
    # 按组名的字母顺序排列，默认的 Levels
    Tukey.labels=Tukey.labels[order(Tukey.labels$group), ]
    return(Tukey.labels)
  }

  # 当只有两组时，用 LSD 标注字母
  if (length(unique(df$group)) == 2){
    # LSD 检验，添加差异组字母
    library(agricolae)
    out = LSD.test(model, "group", p.adj="none")
    stat = out$groups
    # 分组结果添入 Index
    df$stat=stat[as.character(df$group),]$groups
    # 当大于两组时，用 multcompView 标注字母
  }else{
    # library(multcompView)
    LABELS = generate_label_df(Tukey_HSD , "group")
    df$stat=LABELS[as.character(df$group),]$Letters
  }

  # 设置分组位置为各组 y 最大值+高的 5%
  max=max(df[,c(index)])
  min=min(df[,index])
  x = df[,c("group",index)]
  y = x %>% group_by(group) %>% summarise_(Max=paste('max(',index,')',s
ep=""))
  y=as.data.frame(y)
  rownames(y)=y$group
  df$y=y[as.character(df$group),]$Max + (max-min)*0.05

  # 绘图 plotting
  p = ggplot(df, aes(x=group, y=df[[index]], color=group)) +
    geom_boxplot(alpha=1, outlier.shape = NA, outlier.size=0, size=0.7,
 width=0.5, fill="transparent") +
    labs(x="Groups", y=paste(index, "index"), color=groupID) + theme_cl
assic() +
    geom_text(data=df, aes(x=group, y=y, color=group, label=stat)) +
    geom_jitter(position=position_jitter(0.17), size=1, alpha=0.7)+
    theme(text=element_text(family="sans", size=7))
  p
}
<bytecode: 0x0000000015956ce0>
<environment: namespace:amplicon>
```

柱状图加误差棒

大多数情况下还是推荐箱线图的，但有时觉得箱线图用的太多，审美疲劳。或是数据分布过散，规律不明显时，也可以尝试使用历史更加悠久的柱状图+误差棒

```
(p = alpha_barplot(alpha_div, index = "richness", metadata, groupID = "
Group"))
```

```
[1] "Statistic table is in alpha_boxplot_TukeyHSD.txt"
```




```
ggsave(paste0("10.png"), p, width=89, height=56, units="mm")
```

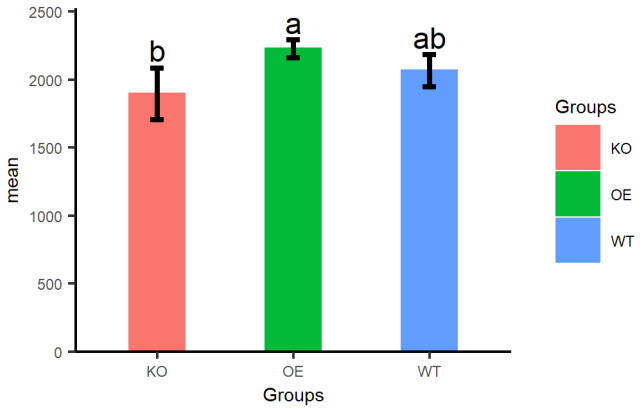


图 11. 误差柱状图展示 Alpha 多样性。

如果你使用本教程的代码，请引用:

- Yong-Xin Liu, Yuan Qin, Tong Chen, et. al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 41, 1-16, doi:10.1007/s13238-020-00724-8 (2020)

- Jingying Zhang, Yong-Xin Liu, et. al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nature Biotechnology 37, 676-684, doi:10.1038/s41587-019-0104-4 (2019).

声明：由于个人时间和知识有限，文中定有很多不足之处，欢迎大家留言批评指正。

作者贡献：刘永鑫负责本文的主体框架和大部分写作，编写了 alpha_boxplot 函数；文涛参与本文部分创作，绘制了典型箱线图的两张模式图，编写了 alpha_barplot 函数；钱旭波编写了 Alpha 多样性的基本概念和常用指数，绘制了模式图，并修改了本文。

致谢：感谢西北农林科技大学的席娇对本文的校对，并提出宝贵修改意见。

参考文献

Xu-Bo Qian, Tong Chen, Yi-Ping Xu, Lei Chen, Fu-Xiang Sun, Mei-Ping Lu & Yong-Xin Liu. (2020). A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. Chinese Medical Journal, doi: https://doi.org/10.1097/CM9.0000000000000871

Python3 中 skbio.diversity.alpha 包提供 34 种 alpha 多样性指数的说明和计算方法 http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html

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

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. (2020). Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes, 1-22, doi: https://doi.org/10.1080/19490976.2020.1747329