样本量和测序深度的Alpha多样性稀释曲线

刘永鑫、文涛

2020-06-22

### 样本量和测序深度的Alpha多样性稀释曲线

本节作者：刘永鑫，文涛

版本1.0.1，更新日期：2020年6月22日

本项目永久地址： <https://github.com/YongxinLiu/MicrobiomeStatPlot> ，本节目录 212RareCurve，包含R markdown(\*.Rmd)、Word(\*.docx)文档、测试数据和结果图表，欢迎广大同行帮忙审核校对、并提修改意见。提交反馈的三种方式：1. 公众号文章下方留言；2. 下载Word文档使用审阅模式修改和批注后，发送至微信(meta-genomics)或邮件([metagenome@126.com](mailto:metagenome@126.com))；3. 在Github中的Rmd文档直接修改并提交Issue。审稿人请在创作者登记表 <https://www.kdocs.cn/l/c7CGfv9Xc> 中记录个人信息、时间和贡献，以免专著发表时遗漏。

#### 基本概念

稀释曲线(Rarefaction Curve，也称稀疏曲线)：一般在微生物组研究中用于评估测序数据量或样本量的饱和情况。

本方法主要用于检测测序数据量是否充足。这里用到的方法是逐步扩大随机抽样的测序深度，如果样本测序深度增大但曲线不再有明显升高（准确来讲，曲线斜率平滑，变化较小）时，则认为测序数据量已充足，再增加测序数据量，样本的alpha多样性指标也不会有明显的变化，即样本alpha多样性指标达到稳定。

本方法也常用于评估样本量是否足够，是从样本中随机抽取一定数量的个体，统计出这些个体所代表的物种数目，并以个体数与物种数来构建曲线。它可以用来比较不同数量样本测序后的物种丰富度，也可以用来说明样本量大小是否合理。评估样本量是否足够，相关分析通常采用对原始样本进行随机抽样的方法，以抽到的不同取样量的样本与它们所代表的所有特征(如OTU)的数目构建稀释曲线。在样本稀释曲线图中，当曲线趋向平坦时，说明取样量充足且合理，更多的取样只会产生少量新的特征，反之则表明继续取样还可能产生较多新的特征，有必要进一步增加取样量。因此，通过绘制稀释性曲线，可以得出样品的取样量是否充足的结论。

其实就目前扩增子测序研究中普遍达到的测序深度而言，依据稀释曲线来判断样本测序数据量是否足够的问题已经不再是被重视的科学问题，目前在样本量是否充足和宏基因组测序基因集是否饱和方面有着越来越广泛的应用。alpha多样性的计算目前只能通过抽平来计算。但是一次抽平有概率（小概率）在一定程度上错误地评估alpha多样性结果。所以现在有一些研究者通过多次抽平计算alpha多样性，并通过求取均值的方式来叫矫正alpha多样性。稀释曲线是对单个alpha多样性结果的补充，可以从不同梯度全面地分析和展示结果。因此，基于不通测序深度或样本量水平上展示了alpha多样性，更加有利于对微生物群体多样性的综合评估。

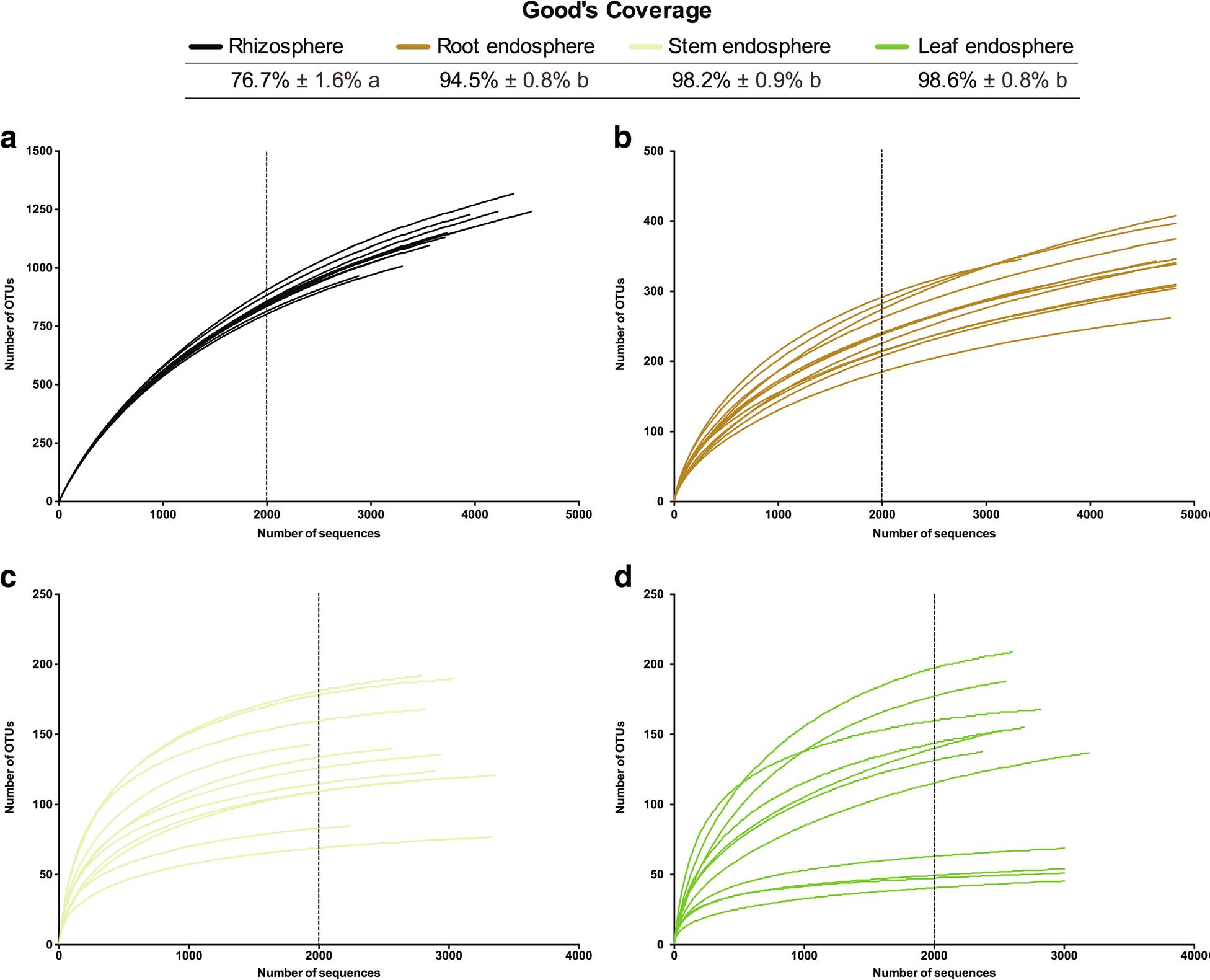
#### 文献解读

稀释曲线多用于测序数据量和样本量是否饱和的评估，在高通量测序初期(5前的文章)中应用较多，目前文章结果种类越来越多样，而且更多注重结果的新发现而不是评估，在扩增子文章中使用频率逐渐下降，但在宏基因组的文章中使用频率越来越多。

##### 例1. 各组中各样本的多样性随测序深度变化

本文是在Microbiome杂志上发表的杨树各部分微生物组的16S测序描述文章(Beckers et al., 2017)，图1采用稀释曲线描述各样本的测序深度与多样性的变化。这篇文章分析思路和内容都非常简单，文章发表3年引用过百次，详见 - [《Microbiome: 简单套路发高分文章–杨树微生物组》](http://mp.weixin.qq.com/s/Zod4Ca1tk3y5lIdI6RCiQQ)

注：关于Microbiome的图片格式和质量说明。Microbiome杂志的文章图片都是位图，不仅有的图片字看不清，而且无法被搜索引擎检索到。此图在文章主页中插入的图片质量非常差，是仅有37.5 KB的webp格式，点击查看原图(Full size )图片仍为webp，且仅为48.1 KB，图中文字比较模糊。再使用Adobe Reader打开PDF，复制到Word中，再另存为jpg/png，图片更清楚，分别为200/500 KB。



**图1. 不同取样部位(Compartment)中每株杨树测序数据绘制的覆盖率估算值稀释曲线**。A 根际土、B 根、C 茎、D 叶。 展示测序的饱合情况，同时展示不同生态位的差异(Y轴坐标不同，即Alpha多样性差别很大)，还有每颗树间也有较大的差别(图中的每条线代表来自一棵树的样品)

Average Good’s coverage estimates (%) and rarefaction curves of individual poplar trees per plant compartment (a rhizosphere soil, b root, c stem, d leaf). Good’s coverage estimates represent averages of 15 independent, clonally replicated poplar trees (rhizosphere soil and root samples) and 11 replicates (stem and leaf samples) (± standard deviation) and were calculated in mothur based on 10,000 iterations. Lowercase letters represent statistical differences at the 95% confidence interval (*P* < 0.05). Rarefaction curves were assembled showing the number of OTUs, defined at the 97% sequence similarity cut-off in mothur, relative to the number of total sequences.

结果：为了构建alpha稀释曲线（图1），我们从数据集中删除了“单体”（只有一个序列的OTU），因为这些“单例”可能是由于测序错误造成的。我们为每个单独的样品构建了稀释曲线，显示了观察到的OTU的数量，相对于已鉴定的细菌rRNA序列的总数（图1），该数量定义为以Mohur表示的97%序列相似性阈值下的序列数量。 正如预期的那样，内生细菌群落（图1b–d）的多样性远低于根际群落（图1a）。 此外，与根际样品相比，内生样品的稀释曲线形状变化程度更高。 评估每个样品的OTU丰富度的稀释曲线通常接近饱和度。 大多数根内生样品的饱和度范围为250–300 OTUs，而对于茎和叶样品这一范围只有50–150 OTUs左右。

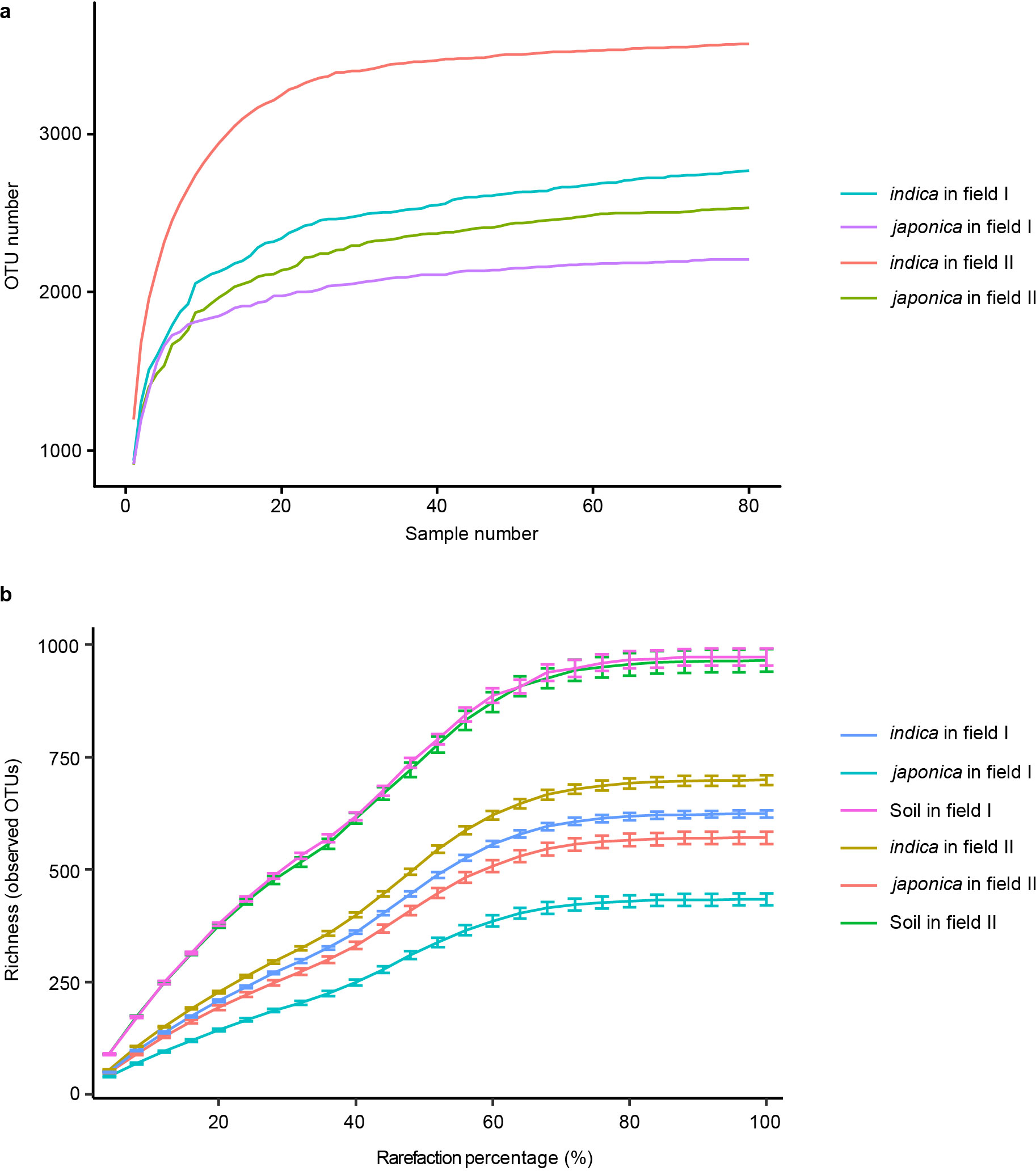
To construct alpha rarefaction curves (Fig. 1), we removed singletons (OTUs with only one sequence) from the dataset since these singletons could be due to sequencing artefacts. Rarefaction curves were constructed for each individual sample showing the number of observed OTUs, defined at a 97% sequence similarity cut-off in mothur, relative to the number of total identified bacterial rRNA sequences (Fig. 1). As expected, endophytic bacterial communities (Fig. 1b–d) were much less diverse than rhizospheric communities (Fig. 1a). Furthermore, the endophytic samples exhibited a higher degree of variation in the shape of their rarefaction curves as compared to the rhizospheric samples. Rarefaction curves evaluating the OTU richness per sample generally approached saturation. The majority of the root endophytic samples saturated around 250–300 OTUs and around 50–150 OTUs for the stem and leaf samples.

讨论：当比较根际土和根内样品时，我们观察到OTU稀释曲线的形状明显不同（图1）。根际土样品的稀释曲线间较为均匀（图1a），而内生样品的稀释曲线形状的变化要大得多，尤其是茎和叶样品（图1b-d）。内生OTU丰富度的高变异性可能是由杨树的根和植物地上部的散发和非均匀定植引起。Gottel等人将这种变异的一部分归因于无法对细菌内生菌群落进行足够深而均匀的测序，这是由于宿主16S rRNA基因（本研究检测到67,000条叶绿体和65,000条线粒体序列）的高度共扩增引起。但是，我们的数据显示出大致相同的模式，没有对非目标DNA进行共扩增，并且覆盖率估算值很高（图1）。因此，我们的数据表明内生菌落的变化是稀释曲线高度变化的主要原因。根际定植主要是由以下因素驱动的：（a）植物（根际沉积）沉积大量碳（例如，根系分泌物，根冠粘液等），以及（b）相对简单或不完善的化学作用将细菌（和其他微生物）吸引到根系分泌物中。

We observed remarkably dissimilar shapes of the OTU rarefaction curves when comparing rhizosphere soil and endosphere samples (Fig. 1). Rhizosphere soil samples displayed uniform rarefaction curves (Fig. 1a) whereas the variation in the shape of the rarefaction curves from the endophytic samples was much higher, especially for the stem and leaf samples (Fig. 1b–d). High variability of endophytic OTU richness, as depicted by the rarefaction curves, could possibly be caused by sporadic and non-uniform colonization of the roots and aerial plant compartments of Populus [36]. Gottel et al. attributed part of the variation to their inability to sequence the bacterial endophytic community deeply and uniformly enough because of the high co-amplification of organellar 16S rRNA (67,000 chloroplast and 65,000 mitochondrial sequences) [36]. However, our data exhibit roughly the same pattern without the co-amplification of non-target DNA (Table 1) and with high Good’s coverage estimates (Fig. 1). Therefore, our data suggest considerable variation in endophytic colonization as a major reason for the high variability in the rarefaction curves. Indeed, rhizosphere/rhizoplane colonization is primarily driven by (a) the deposition of large amounts of carbon (e.g., root exudates, mucilage by the root caps, etc.) by plants (rhizodeposition) and (b) the relatively simple or inelaborate chemo-attraction of the bacteria (and other microorganisms) to the root exudates.

##### 例2. 样品和百分比抽样稀释曲线

本文是我负责分析发表于Naute Biotechnology(简称NBT)的封面文章(Zhang et al., 2019)，介绍了水稻群体层面微生物组的研究并揭示宿主调控根系微生物参与氮利用的现象。详见[《NBT封面：水稻NRT1.1B基因调控根系微生物组参与氮利用》](https://mp.weixin.qq.com/s/s7Q1_MeodqJ0hjwDumeiXQ)。

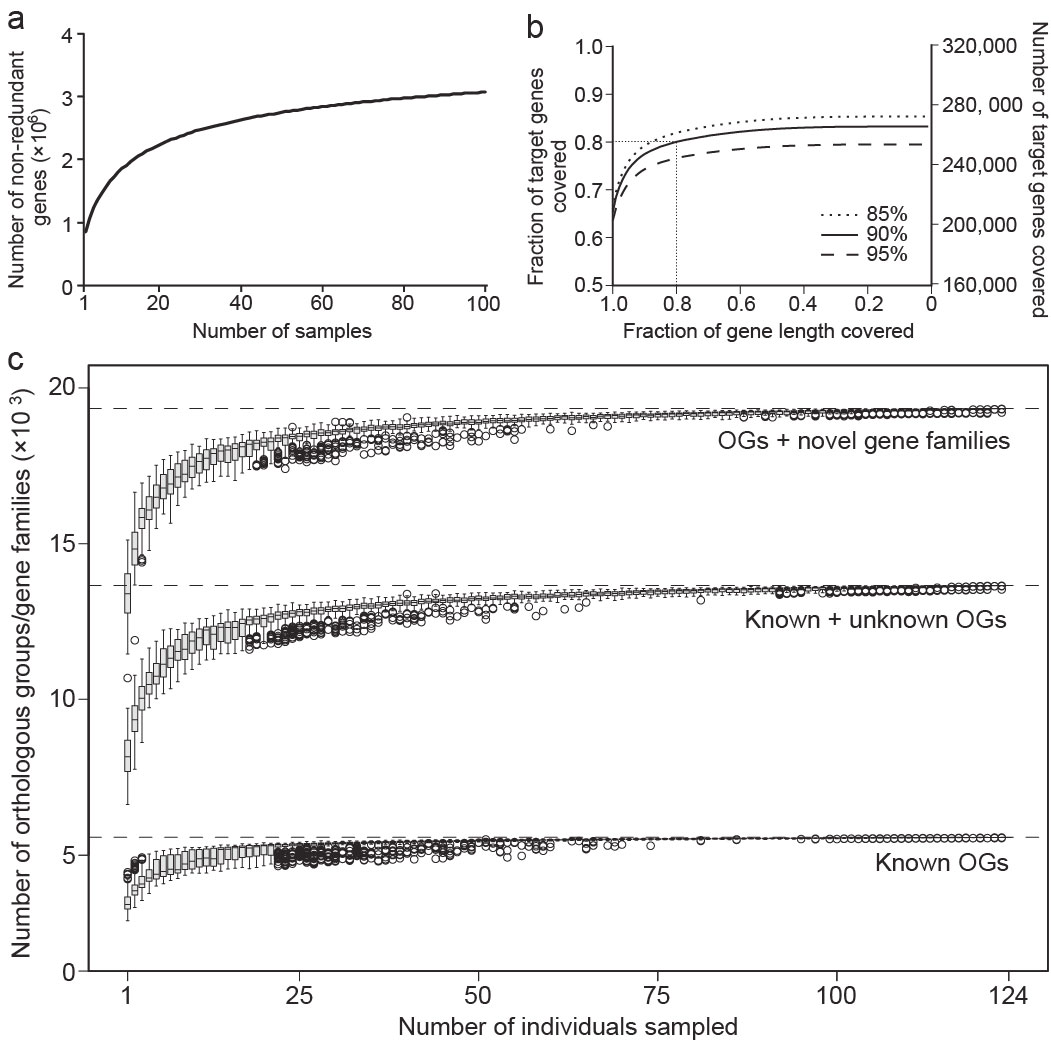


**附图1. 代表性的籼稻和粳稻品种在根细菌群成员中的覆盖度**。 （a）样本稀释曲线：随着样品数量的增加，根微生物群的细菌种类稀释曲线达到饱和阶段，这表明我们群体中的根微生物捕获了每个水稻亚种的大部分根细菌种类。 分别显示了两个位置的籼稻和粳稻品种。 （b）随着测序深度的增加，从籼稻和粳稻品种根系菌群中检测到的细菌OTU的稀释曲线达到饱和阶段。 每个误差线代表标准误差。该图中重复样本的数量如下：在地块I中，籼稻（*n* = 201），粳稻（*n* = 80），土壤（*n* = 12）； 在地块II中，籼稻（*n* = 201），粳稻（*n* = 81），土壤（*n* = 12）。

Supplementary Figure 1. Coverage of members in the root bacterial microbiota by the representative indica and japonica varieties. (a) Rarefaction curves of detected bacterial species of the root microbiota reach the saturation stage with increasing numbers ofsamples, indicating that the root microbiota in our population capture most root bacteria members from each rice subspecies. Indicaand japonica varieties in two locations are shown separately. (b) Rarefaction curves of detected bacterial OTUs of the root microbiotafrom indica and japonica varieties reach saturation stage with increasing sequencing depth. Each vertical bar represents standard error.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).

##### 例3. 样品和基因(簇)数量的稀释曲线或等差箱线图

本文是华大基因覃俊杰、李瑞强、王俊等负责分析发表于Naute的文章(Qin et al., 2010)，构建了人类肠道基因集1.0版本，虽然发表近10年，但是里程碑式的成果，目前被引用近8千次。详见：[《Nature：基于宏基因组测序构建人类肠道微生物组参考基因集》](https://mp.weixin.qq.com/s/KLMm62kD5uhpaE24b5P1Xg)。



**图2. 预测人体肠道微生物组中的开放阅读框(稀释曲线展示样本量与基因或基因家族数量的关系)**。a，测序样本量与非冗余基因数量的稀释曲线。基因积累曲线对应于Sobs值（观察到的基因数），该值是使用EstimateS 8.2.0对随机选择的100个样本（由于内存限制）计算得出。b，采用三种不同相似度计算来自89种常见肠道微生物物种的基因覆盖数量和比例的关系。c，基于已知直系同源基团（OG；底部），已知加未知直系同源基团（包括例如假定的、预测的、保守的假定功能；中间）和从宏基因组中恢复直系同源的基因，通过调查的样本数量捕获的功能同源簇和新基因家族（> 20个蛋白质）（上）。箱线表示第一和第三四分位数（分别为第25个和第75个百分位数）之间的四分位间距（IQR），内部的线表示中位数。轴须线分别表示距第一个和第三个四分位数的1.5倍IQR内的最小和最高值。圆圈表示轴须以外的异常值。

Figure 2: Predicted ORFs in the human gut microbiome. a, Number of unique genes as a function of the extent of sequencing. The gene accumulation curve corresponds to the Sobs (Mao Tau) values (number of observed genes), calculated using EstimateS21 (version 8.2.0) on randomly chosen 100 samples (due to memory limitation). b, Coverage of genes from 89 frequent gut microbial species (Supplementary Table 12). c, Number of functions captured by number of samples investigated, based on known (well characterized) orthologous groups (OGs; bottom), known plus unknown orthologous groups (including, for example, putative, predicted, conserved hypothetical functions; middle) and orthologous groups plus novel gene families (>20 proteins) recovered from the metagenome (top). Boxes denote the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the line inside denotes the median. Whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote outliers beyond the whiskers.

**结果**：

我们检查了在所有个体中发现的流行基因数量，至少两个读长的基因才被计算在内，绘制该基因数量与测序样本量累计分布曲线（图2a）。是由100个人类个体确定的（EvaluateS程序可以容纳的最高人数）基于指示的覆盖范围丰富度估计值，表明我们的目录涵盖了85.3％的流行基因。尽管这可能被低估了，但它仍然表明该基因集包含了该队列的绝大多数流行基因。

We examined the number of prevalent genes identified across all individuals as a function of the extent of sequencing, demanding at least two supporting reads for a gene call (Fig. 2a). The incidence-based coverage richness estimator (ICE), determined at 100 individuals (the highest number the EstimateS program could accommodate), indicates that our catalogue captures 85.3% of the prevalent genes. Although this is probably an underestimate, it nevertheless indicates that the catalogue contains an overwhelming majority of the prevalent genes of the cohort.

我们将330万个肠道ORF映射到人类肠道中89个常见微生物参考基因组的319,812个基因（目标基因）上。在90％的相似度阈值下，80％的靶基因至少有80％的长度被ORF覆盖（图2b）。这表明该基因组包括了大多数已知的人类肠道细菌基因。

We mapped the 3.3 million gut ORFs to the 319,812 genes (target genes) of the 89 frequent reference microbial genomes in the human gut. At a 90% identity threshold, 80% of the target genes had at least 80% of their length covered by a single gut ORF (Fig. 2b). This indicates that the gene set includes most of the known human gut bacterial genes.

为了研究流行基因集的功能组成，我们计算了n个个体（n = 2–124；见图2c）的任何组合中存在的直系同源基因簇和/或基因家族的总数。这种稀释性分析表明，“已知”功能（在eggNOG或KEGG中注释）迅速饱和（观察到5569个簇）：对50个个体的任何子集进行采样时，大多数能够被检测到。然而，四分之三的普遍肠道功能由未表征的直系同源基因簇和/或全新的基因家族组成（图2c）。当包括这些基因簇时，稀释曲线仅在最后阶段才开始趋于平稳，并达到更高的水平（检测到19,338个簇），这证实了大量个体的大量采样对于获得如此大量新颖或未知功能的基因是必须的。

To investigate the functional content of the prevalent gene set we computed the total number of orthologous groups and/or gene families present in any combination of n individuals (with n = 2–124; see Fig. 2c). This rarefaction analysis shows that the ‘known’ functions (annotated in eggNOG or KEGG) quickly saturate (a value of 5,569 groups was observed): when sampling any subset of 50 individuals, most have been detected. However, three-quarters of the prevalent gut functionalities consists of uncharacterized orthologous groups and/or completely novel gene families (Fig. 2c). When including these groups, the rarefaction curve only starts to plateau at the very end, at a much higher level (19,338 groups were detected), confirming that the extensive sampling of a large number of individuals was necessary to capture this considerable amount of novel/unknown functionality.

#### 绘图实战

测试数据和代码准备教程，详见- [211.Alpha多样性箱线图(样章，11图2视频)](https://mp.weixin.qq.com/s/cbeLDj4Q-Zsu7jjyn1iffQ)。 安装R包出现问题，可以下载预编译的R包，地址项目 <https://github.com/YongxinLiu/MicrobiomeStatPlot> - Data 目录 - BigDataDownlaodList.md 文档。

##### 安装和加载依赖R包

检查依赖关系是否安装，有则跳过，无则自动安装。

# 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")  
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")  
# library加载包，suppress不显示消息和警告信息  
suppressWarnings(suppressMessages(library(amplicon)))  
  
# Biconductor包安装，需要BiocManager  
if (!requireNamespace("BiocManager", quietly = TRUE))  
 install.packages("BiocManager")  
library("BiocManager")  
# 检测amplicon包是否安装，没有从源码安装  
p\_list = c("phyloseq", "microbiome")  
for(p in p\_list){  
 if (!requireNamespace(p, quietly = TRUE))  
 BiocManager::install(p)}

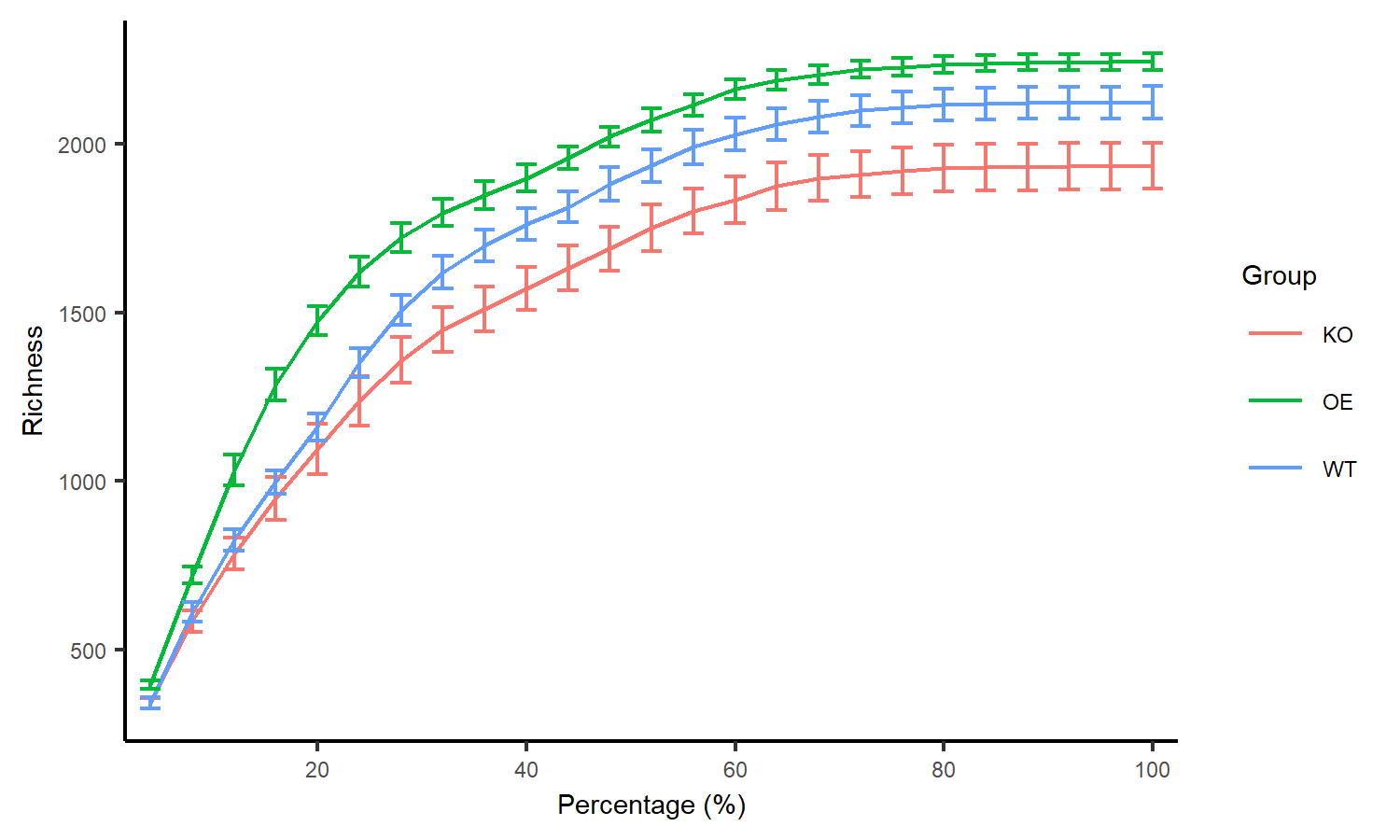
##### USEARCH结果绘制稀释曲线+标准误

USEARCH中的usearch -alpha\_div\_rare可以快速计算抽平后特征表的稀释曲线数据(详见：USEARCH流程)，我们配合amplicon包中的alpha\_rare\_curve函数，可以基于稀释曲线数据一行命令绘制稀释曲线(Rarefaction curve)+标准误(standard error)的图。

通过?alpha\_rare\_curve查看函数内容。本函数使用计算好的alpha稀释曲线表格仅仅用于对图形的绘制，按照分组展示不同处理的稀释曲线。

使用内置数据快速绘制，输入文件为对样本从1-100%重采样的丰富度(richness / Observed OTU)，样本元数据和分组列名

(p = alpha\_rare\_curve(alpha\_rare, 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("p1.rare\_curve.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p1.rare\_curve.png"), p, width=89\*1.5, height=56, units="mm")

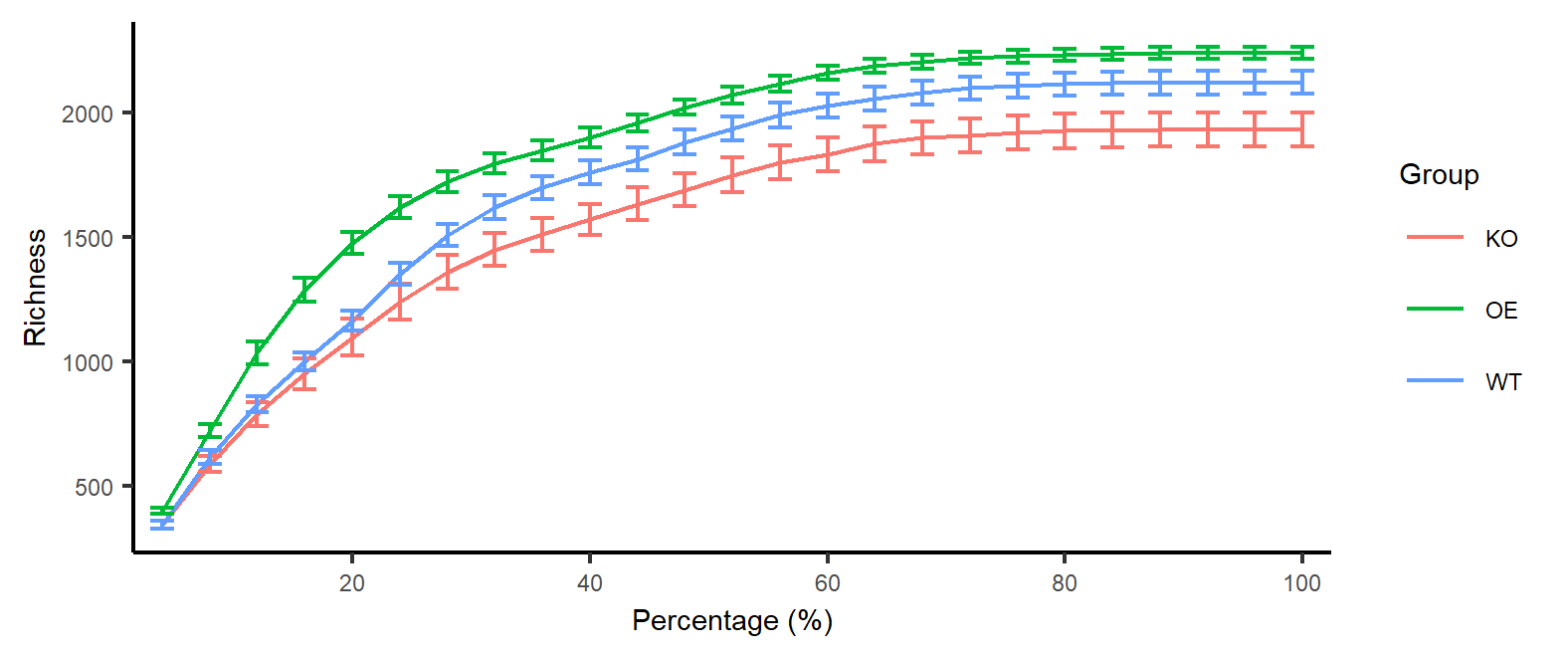


图1. 按分组绘制的稀释曲线+标准误。我们可以看到三组的丰富度存在明显区别。输出图片可以拉长宽度，或减少高度，以使图片尺寸更宽，可用于突出曲线平滑，测序量充足的效果。

我们更常用的方法，是从外部读取数据，查看输入数据格式，逐步绘图，最后保存图片。

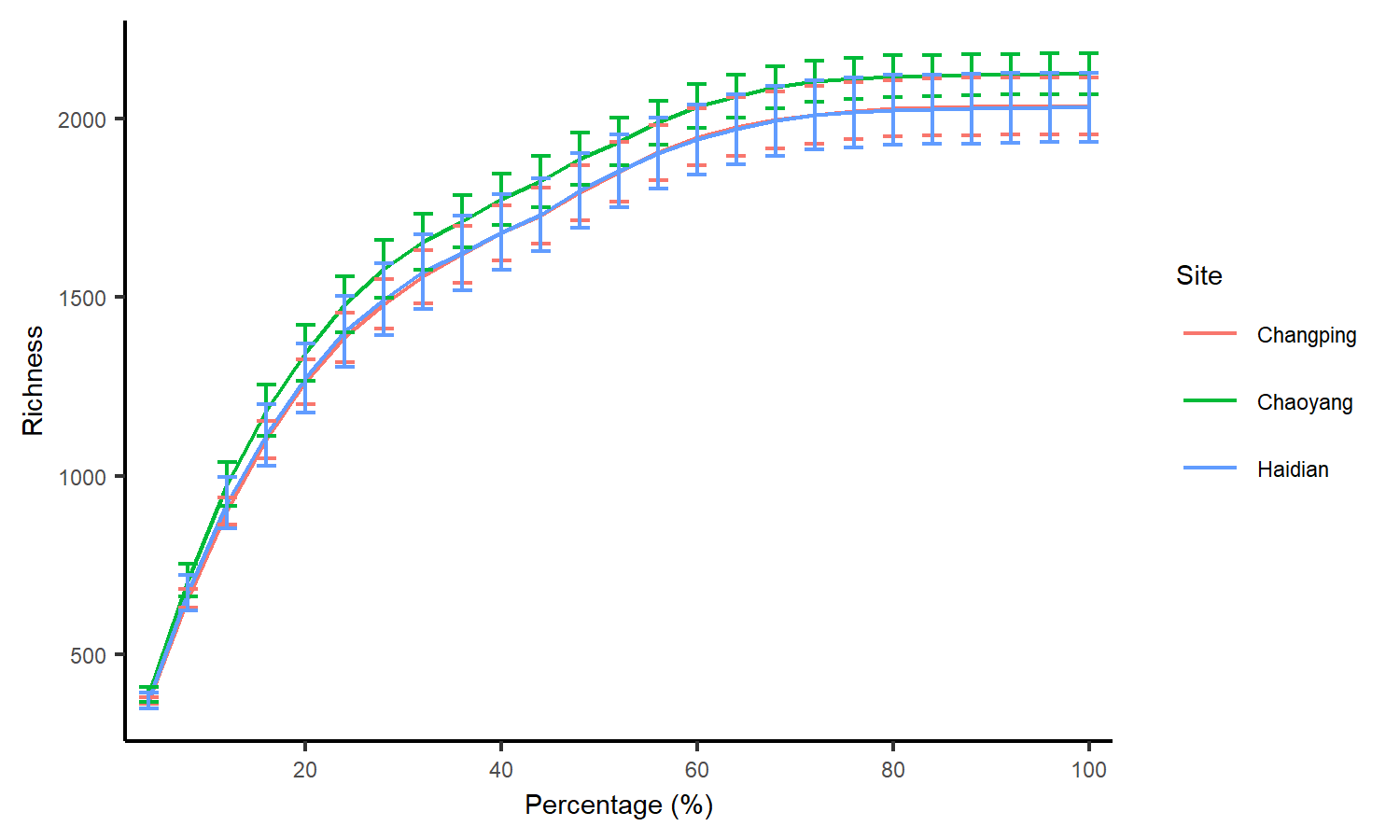
# 设置数据目录位置，可以为本地或网络；这里设为网络地址，方便大家直接运行  
dir="http://210.75.224.110/github/MicrobiomeStatPlot/Data/Science2019/"  
# 读取元数据，参数指定包括标题行(TRUE)，列名为1列，制表符分隔，无注释行，不转换为因子类型  
metadata <- read.table(paste0(dir, "metadata.txt"), header=T, row.names=1, sep="\t", comment.char="", stringsAsFactors = F)  
# 预览元数据前3行，前6列，注意分组列名  
metadata[1:3, 1:6]

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

# 读取usearch生成的稀释表  
alpha\_rare = read.table(paste0(dir, "alpha/alpha\_rare.txt"), row.names= 1, header=T, sep="\t", comment.char="", stringsAsFactors = F)  
# 预览稀释表前3行和9列  
alpha\_rare[1:3,1:9]

KO1 KO2 KO3 KO4 KO5 KO6 OE1 OE2 OE3  
1 104 113 108 107 106 114 107 99 93  
2 211 208 208 192 172 210 223 207 192  
3 318 297 274 260 245 288 351 303 286

# 绘制稀释曲线+标准误，本次选择地点"Site"分组   
(p = alpha\_rare\_curve(alpha\_rare, metadata, groupID = "Site"))



ggsave(paste0("p2.rare\_curve.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p2.rare\_curve.png"), p, width=89\*1.5, height=56, units="mm")

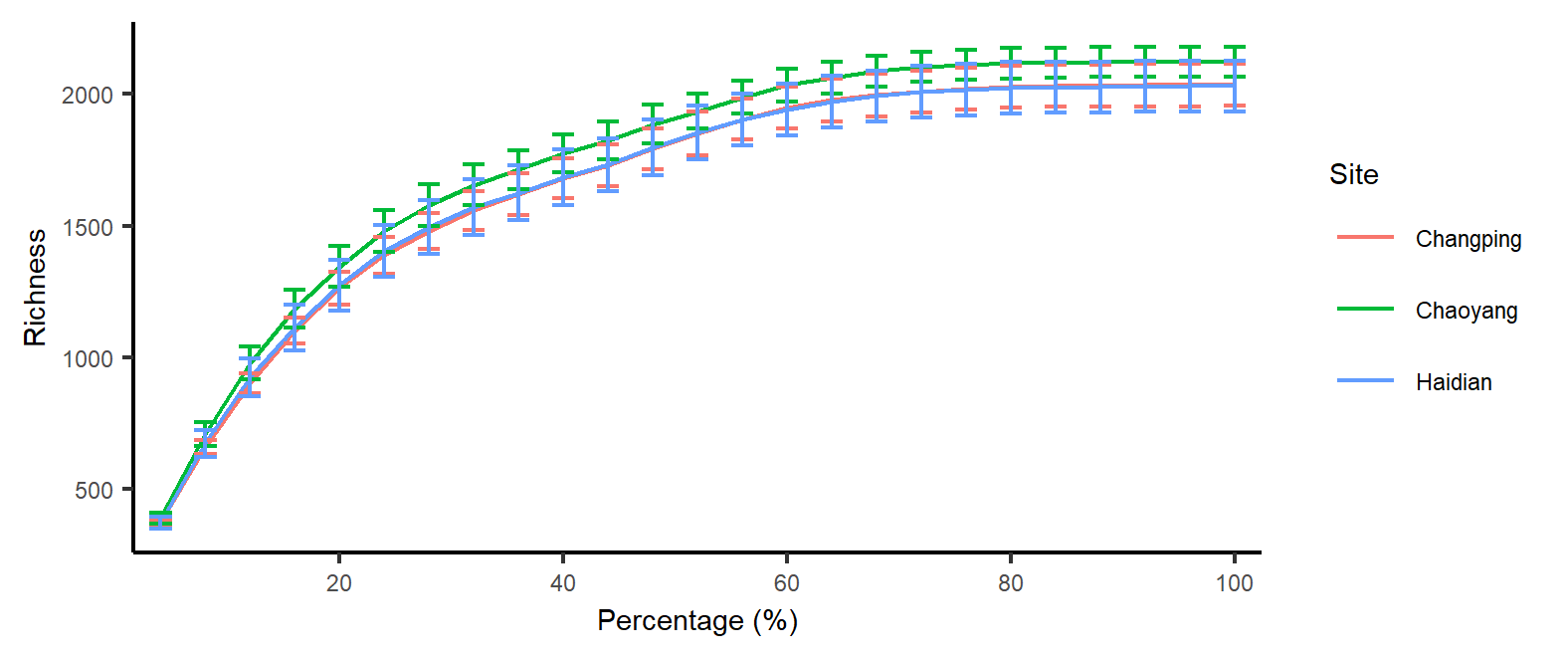


图2. 按地点(Site)分组绘制的稀释曲线+标准误。我们可以看到朝阳、昌平和海淀三组的丰富度没有明显区别。

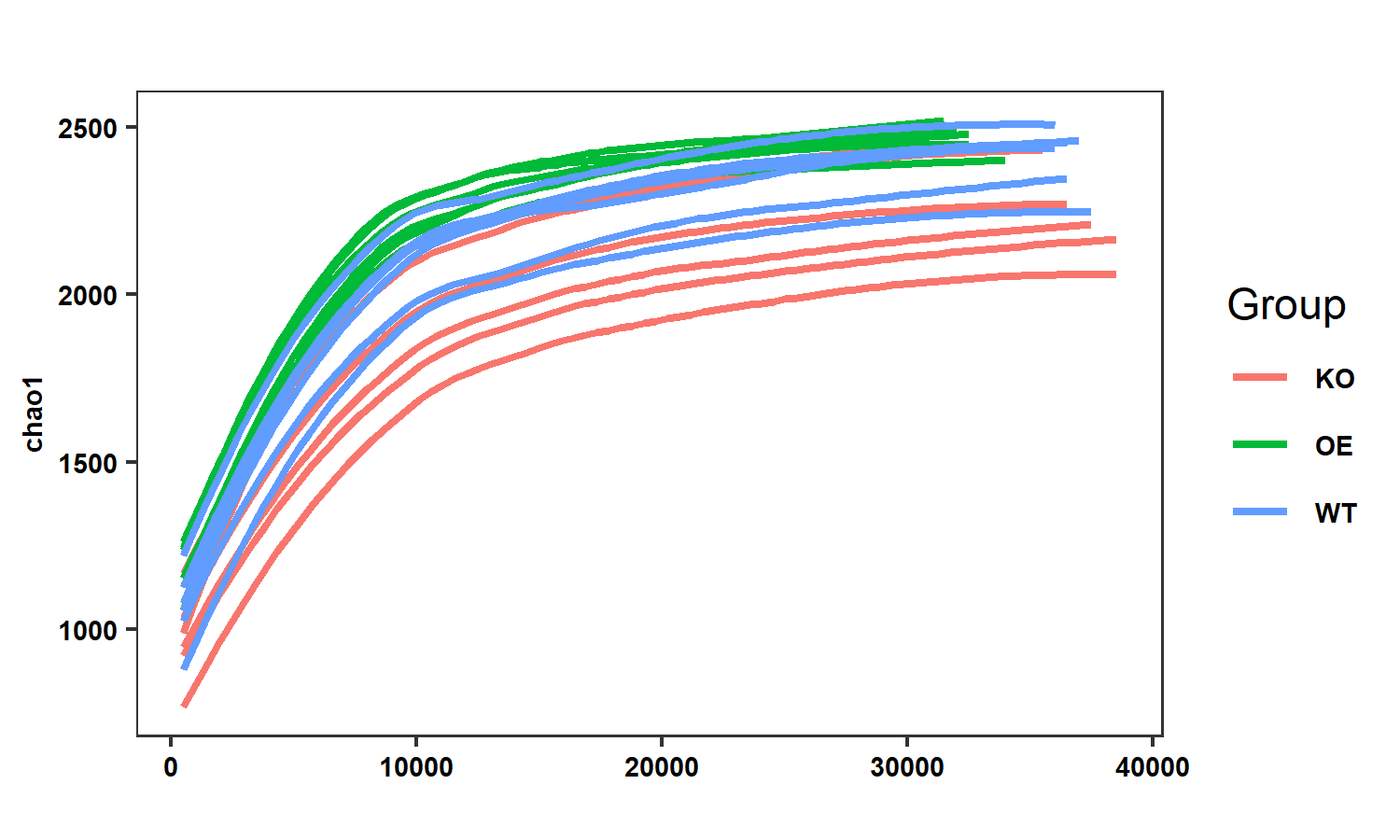
想要修改图片的细节，或进一步修改代码，可以直接运行函数名称(如alpha\_rare\_curve)，显示完整代码，进一步编辑修改。

##### 基于特征表绘制稀释曲线

我们更多的时候是只有特征表，如计算型(reads count)的OTU表。可以使用alpha\_rare\_all函数计算并绘制不同处理的稀释曲线，?alpha\_rare\_all查看函数功能。

* 计算alpha多样性部分，包含了phyloseq和microbiome包的全部alpha多样性指数，总共超过20种alpha多样性指数可供选择。
* 提供start参数可以指定合适的抽平数量
* 提供step参数用于控制抽平序列的间隔，默认100，意思是按照100条序列间隔多次抽平，直到达到最大序列数量。这里的最大序列数量为所有样本中序列数量最多的那一个，其他序列数目较少的样本抽平到自己的最大条数后便自动停止。为了缩短抽平时间，可以将这个参数设置大一些。

# 依赖phyloseq和microbiome包  
result = alpha\_rare\_all(otu = otutab, map = metadata, group = "Group", method = "chao1", start = 500, step = 500)  
# 结果返回列表，1为样本稀释曲线，2为数据表，3为按组均值的稀释曲线，4为组置信区间  
  
# 样本稀释曲线  
(p = result[[1]])



ggsave(paste0("p3.rare\_curve.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p3.rare\_curve.png"), p, width=89\*1.5, height=56, units="mm")

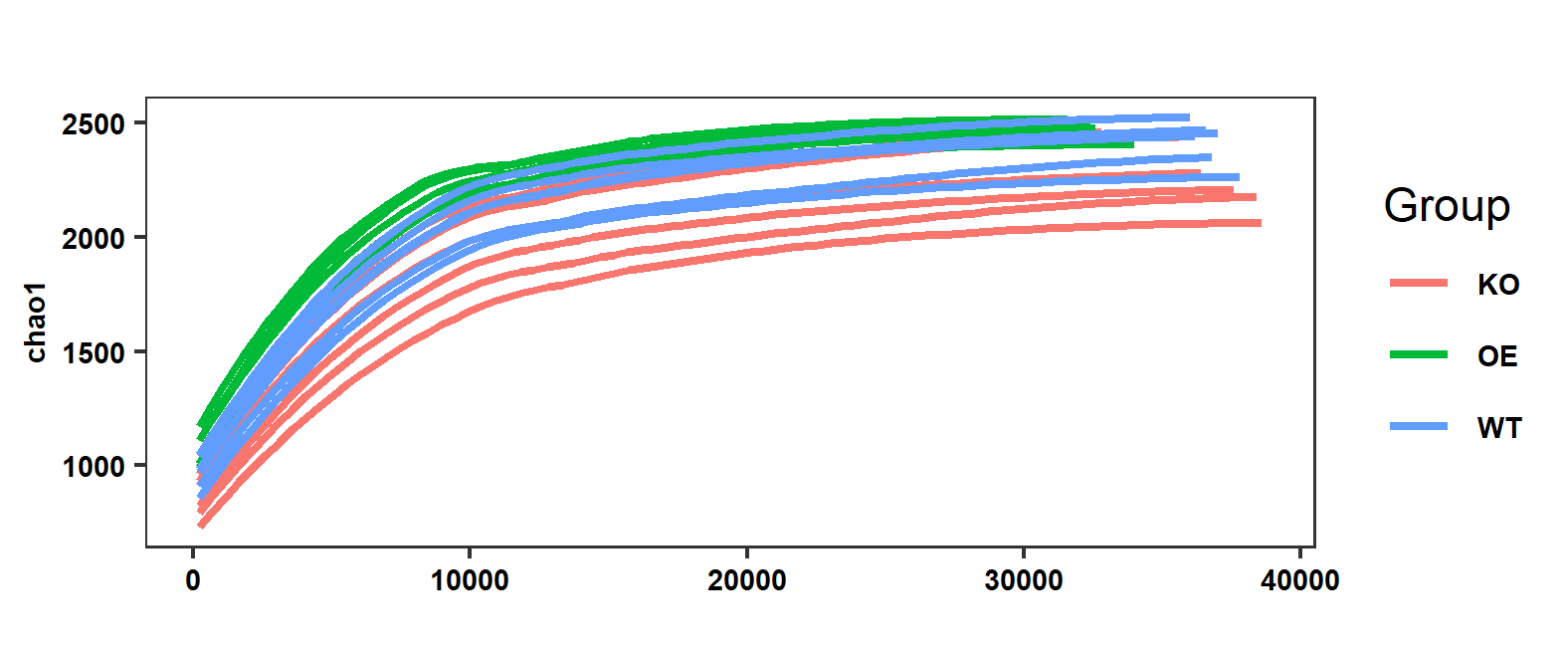


图3. 按样本绘制的稀释曲线，并按组着色。类似于例1 Microbiome的结果，但样本多时会出现曲线互相重叠，很难观察规律，使用较少。

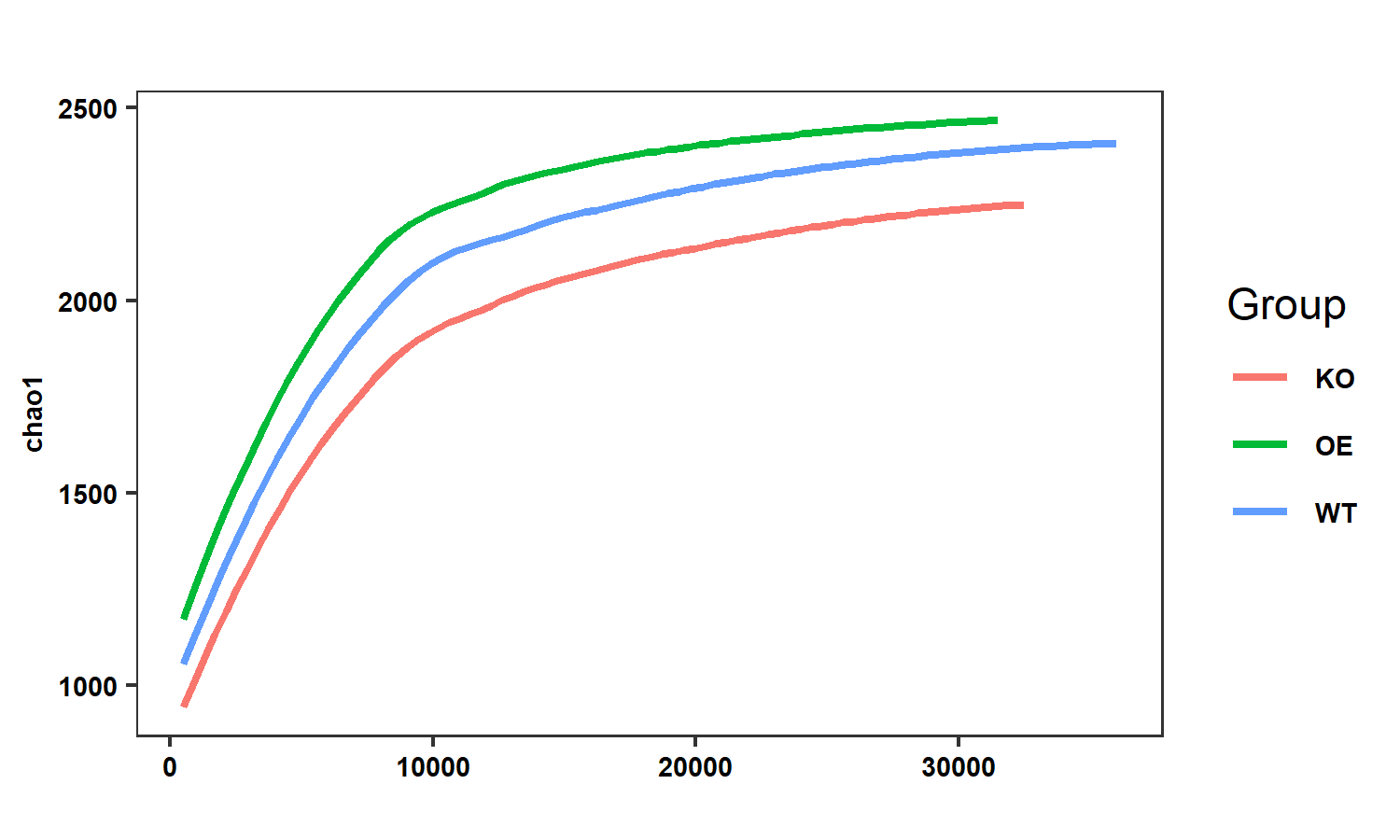
也可以导出原始数据，作为文章的附表，或使用其它工具进一步绘图。

# 预览数据前3行  
head(result[[2]], n=3)

ID i index Group  
1 KO1 500 845.4286 KO  
2 KO2 500 940.6667 KO  
3 KO3 500 570.0000 KO

write.table(result[[2]], file="t1.rare\_curve.txt", sep="\t", quote=F, row.names=F)

# 按组均值绘图  
(p = result[[3]])



ggsave(paste0("p4.rare\_curve\_group.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p4.rare\_curve\_group.png"), p, width=89\*1.5, height=56, units="mm")

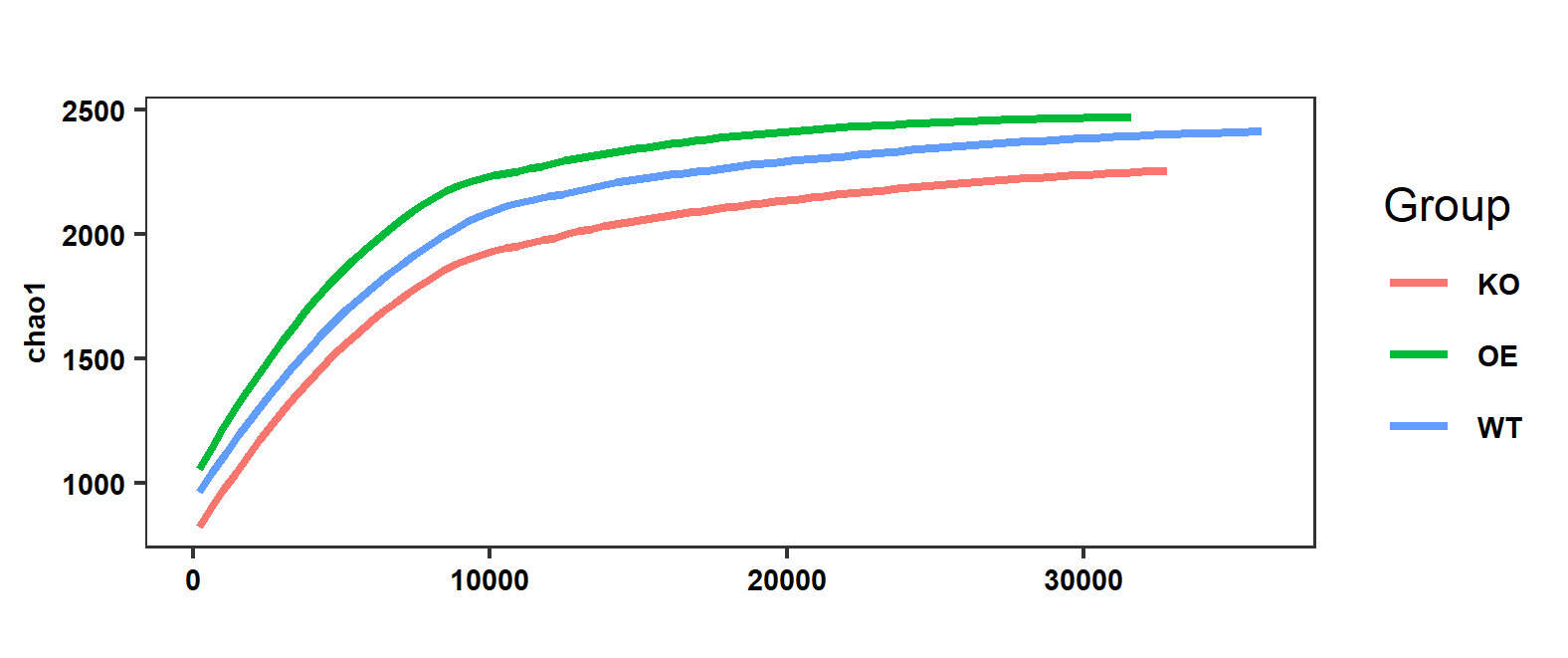
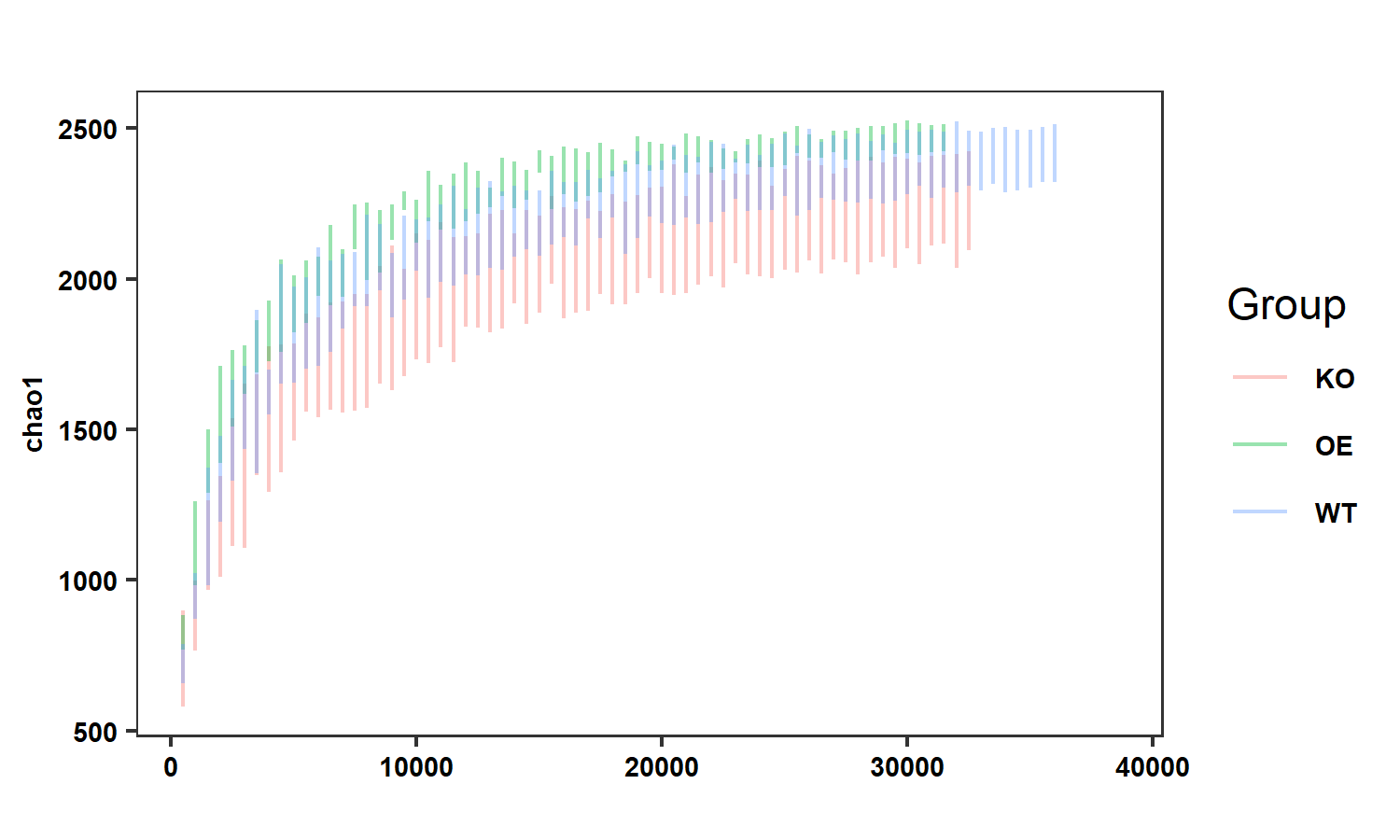


图4. 按样本分组绘制的稀释曲线，并按组着色。类似于图1，不同的是usearch是基于抽平的结果，各组线长度相同，而本函数可基于末抽平的特征表，绘制与实际测序量相同的结果。

# 按照分组绘制标准差稀释曲线  
(p = result[[4]])



ggsave(paste0("p5.rare\_curve\_group\_CI.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p5.rare\_curve\_group\_CI.png"), p, width=89\*1.5, height=56, units="mm")

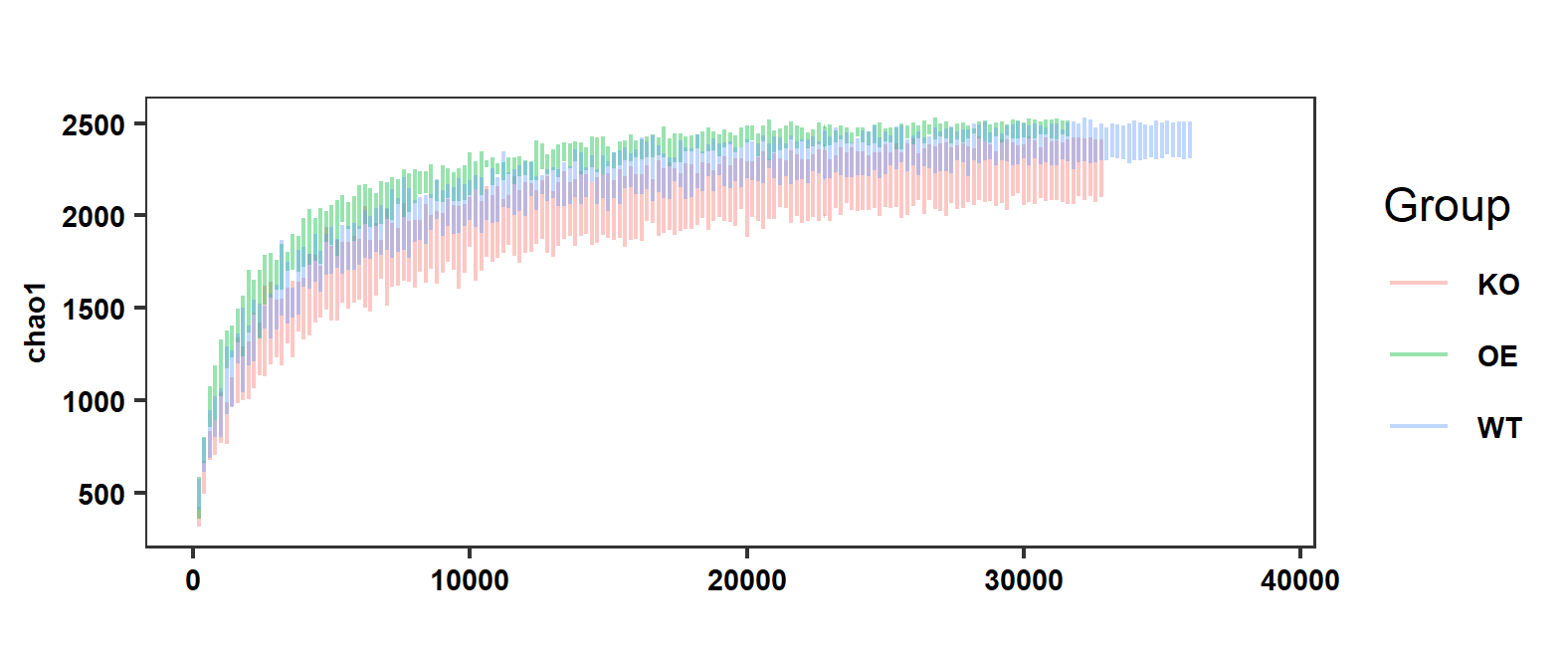
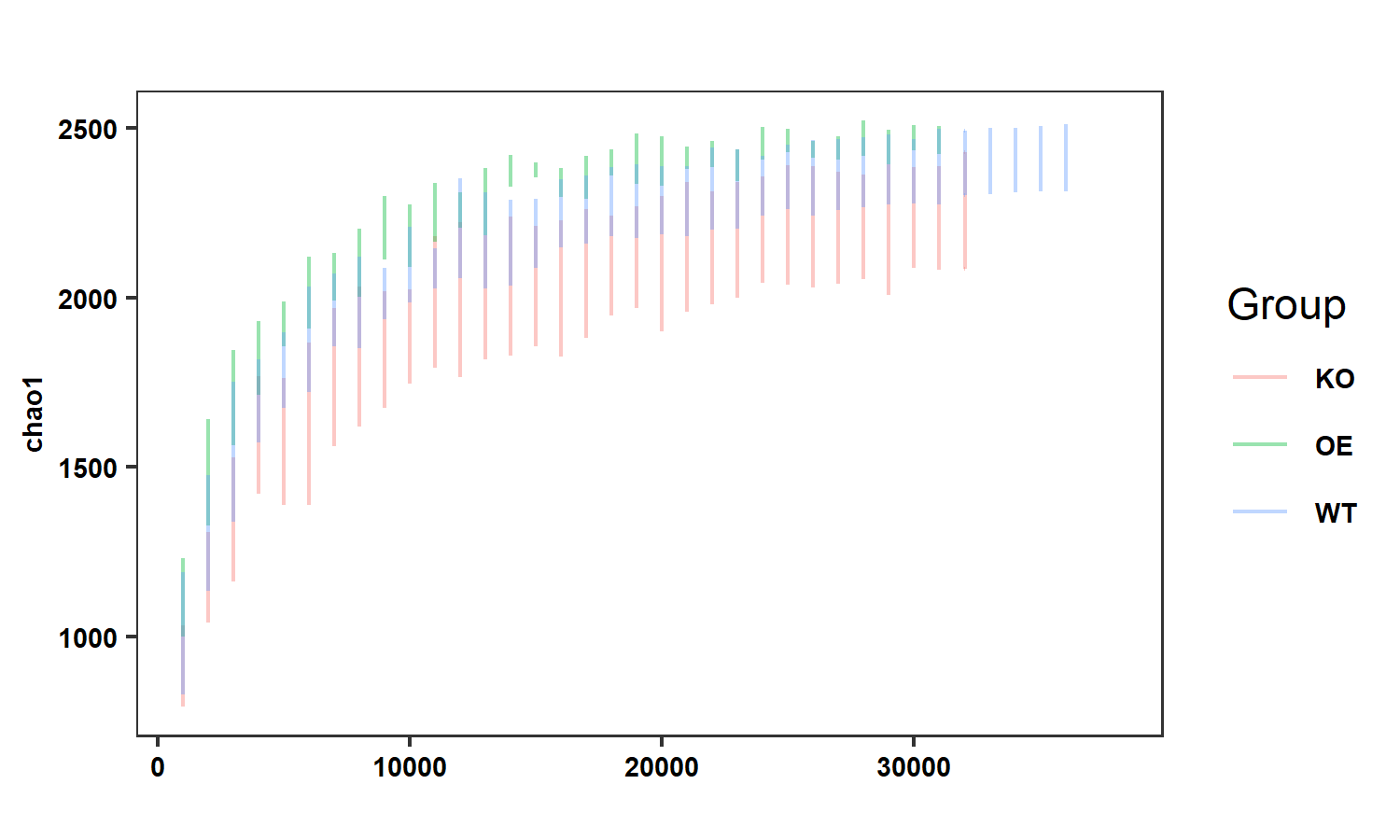


图5. 按样本分组+置信区间绘制的稀释曲线，并按组着色。

##### Phyloseq输入的稀释曲线

这里设置从1000条序列开始抽平，并按照1000条间隔进行逐步抽样，速度快很多，但是图形锯齿化化程度会更多一下。

library(phyloseq)  
# 构造phyloseq对象  
ps = phyloseq(otu\_table(otutab, taxa\_are\_rows=TRUE), sample\_data(metadata))  
# 输入为Phyloseq的绘图  
result = alpha\_rare\_all(ps = ps, group = "Group", method = "chao1", start = 1000, step = 1000)  
(p = result[[4]])



ggsave(paste0("p6.rare\_curve\_group\_CI.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p6.rare\_curve\_group\_CI.png"), p, width=89\*1.5, height=56, units="mm")

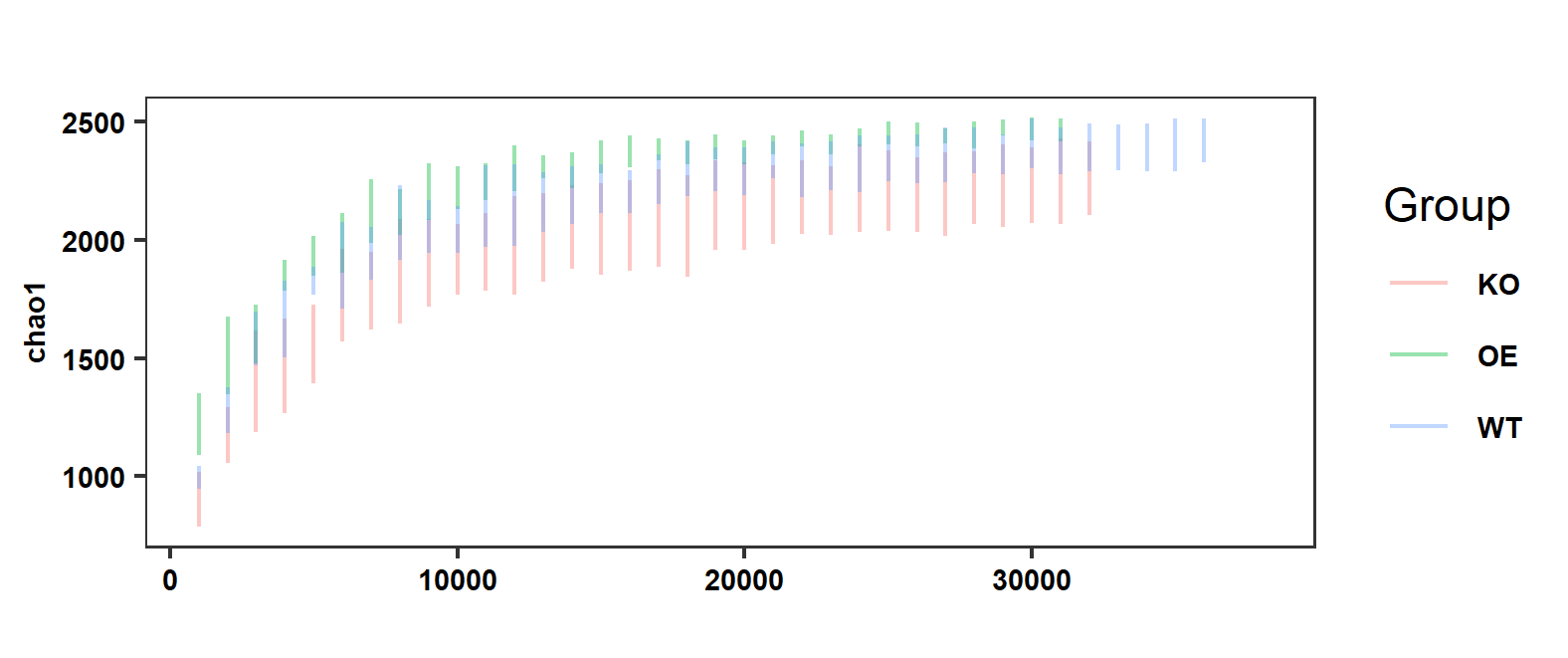


图6. 按样本分组+置信区间绘制的稀释曲线，并按组着色，步长为1000。

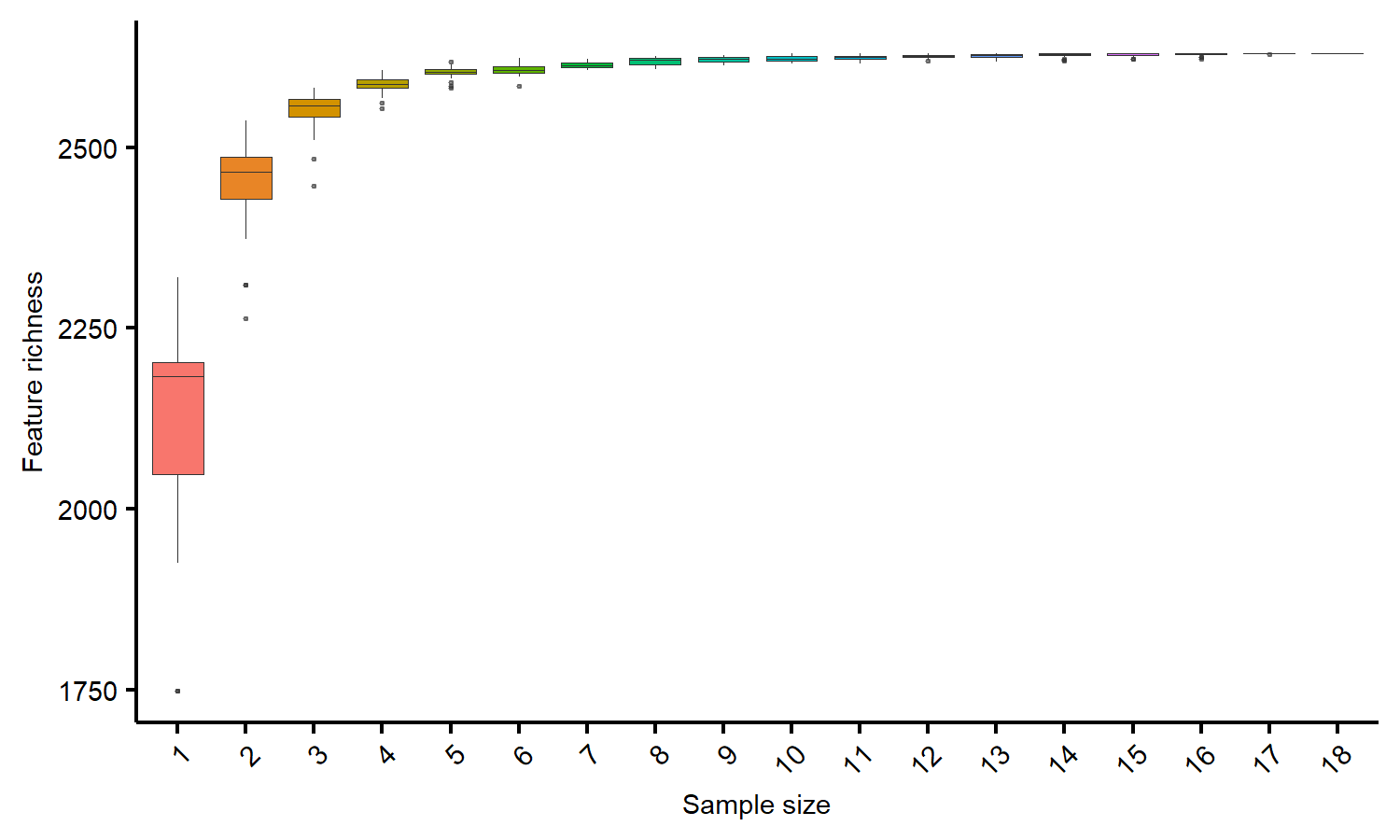
##### 样本箱线图稀释曲线

我们也经常要评估样本量是否达到物种、非冗余基因、基因家庭的饱和。这里编写了alpha\_sample\_rare函数可以基于reads counts值的特征表，直接绘制箱线图稀释曲线。详细帮助见?alpha\_sample\_rare

主要参数：

* otutab：特征表，推荐使用计数值的特征表(OTU/ASV/基因/KO)，也可以是抽平或标准化的。
* length：样本重采样的梯度数量，对应图中的箱体数量，默认为18；本版图推荐6-10，全版图推荐15-10；最大值<样本量，不然会有重复的箱体；
* rep: 每个样本梯度下的抽样次数，即对应每个箱体中的样本量，默认为30。提高会增加计算量。

# 默认值绘制样本箱线图稀释曲线  
library(amplicon)  
(p = alpha\_sample\_rare(otutab, length=18, rep=30, count\_cutoff=1))



ggsave(paste0("p7.sample\_rare.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p7.sample\_rare.png"), p, width=89\*1.5, height=56, units="mm")

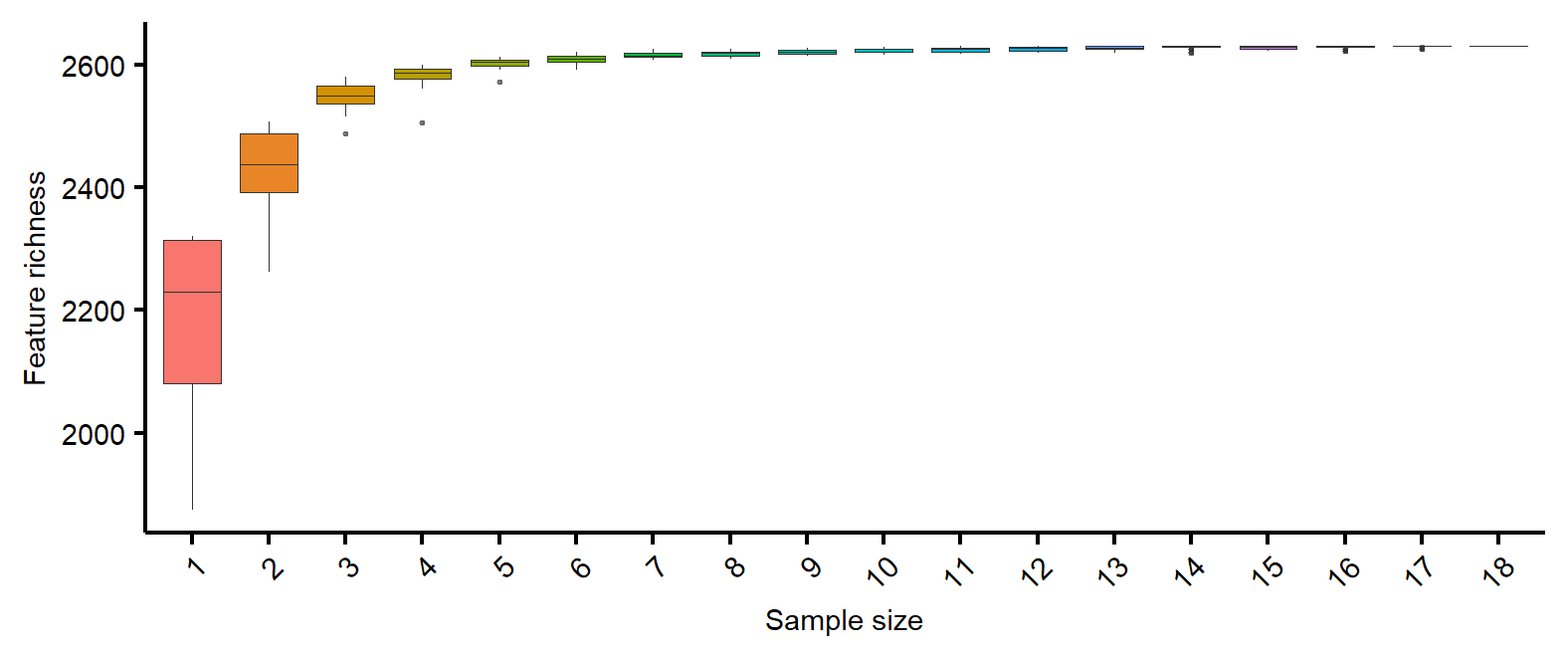
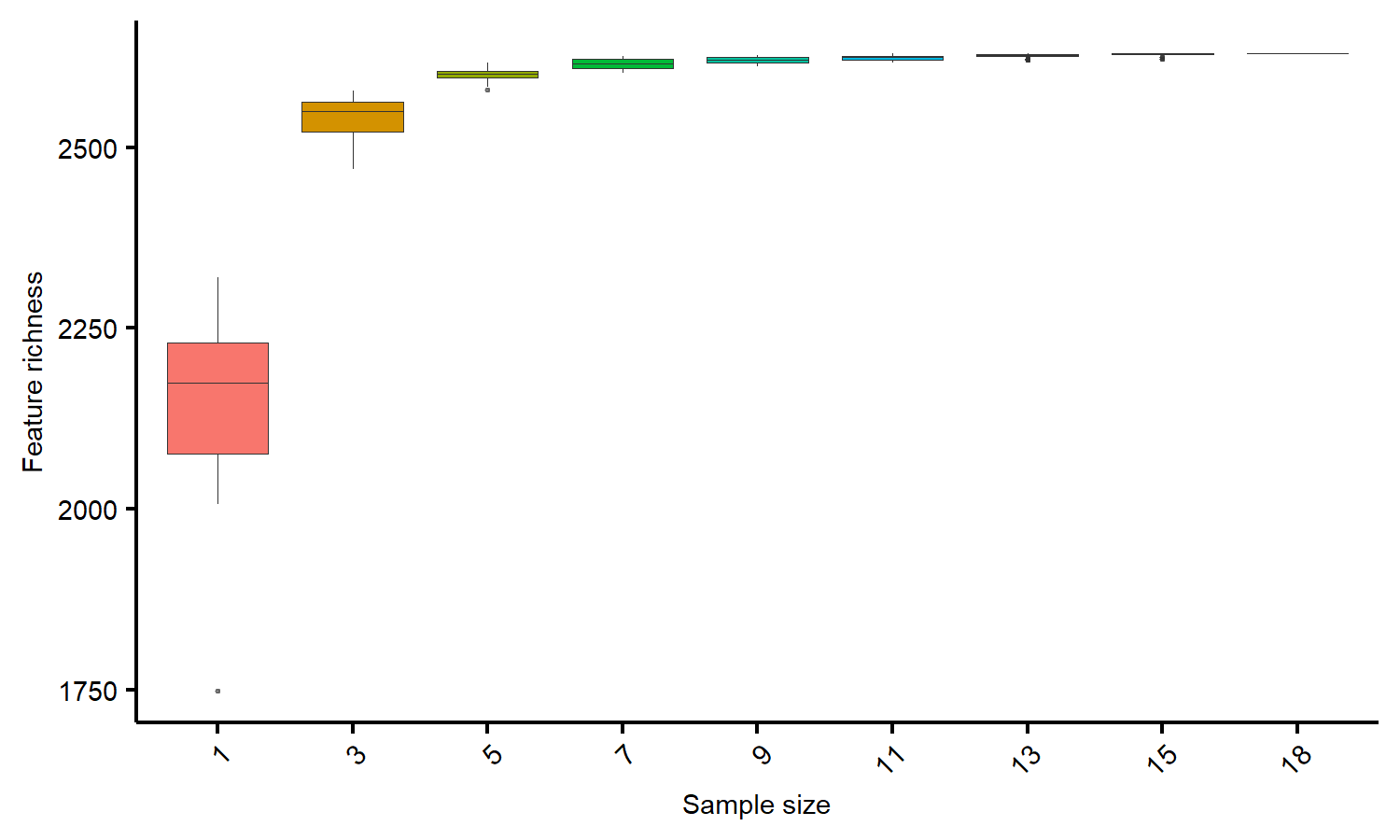


图7. 样本稀释梯度箱线图，从1-18个样本对应的丰富度值。可以看到在5个以上样本时多样性趋于稳定。

# 修改样本量箱体数量，length从默认18修改为9，用于不同趋势或图片布局  
(p = alpha\_sample\_rare(otutab, length=9))



# 箱体少时，可减少图片的宽度比例，如从1.5-2降低为1  
ggsave(paste0("p8.sample\_rare.pdf"), p, width=89\*1, height=56, units="mm")  
ggsave(paste0("p8.sample\_rare.png"), p, width=89\*1, height=56, units="mm")

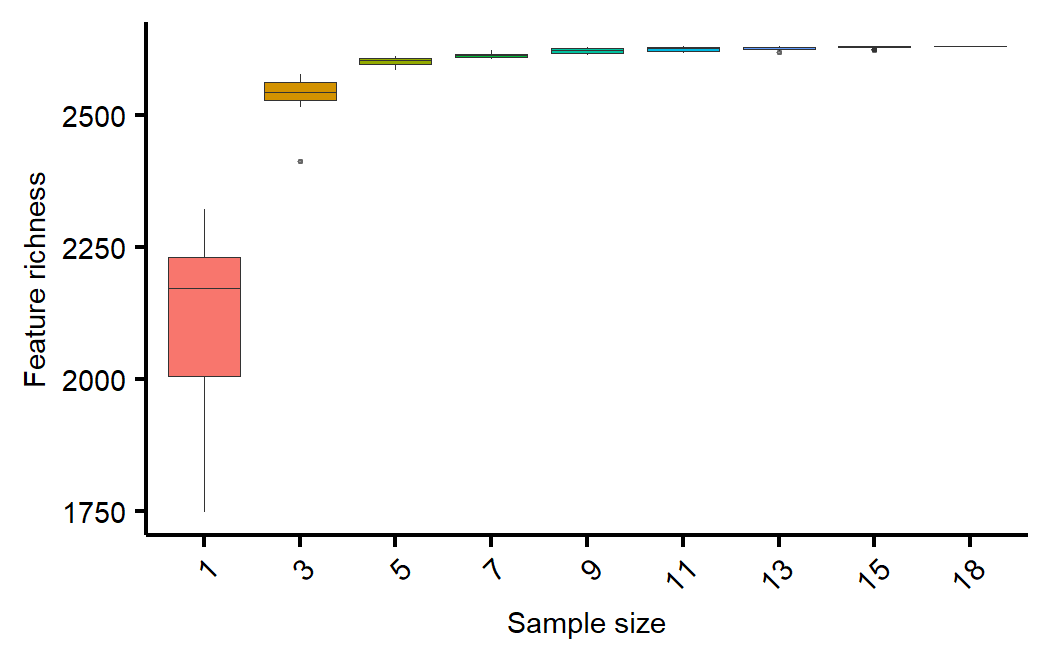
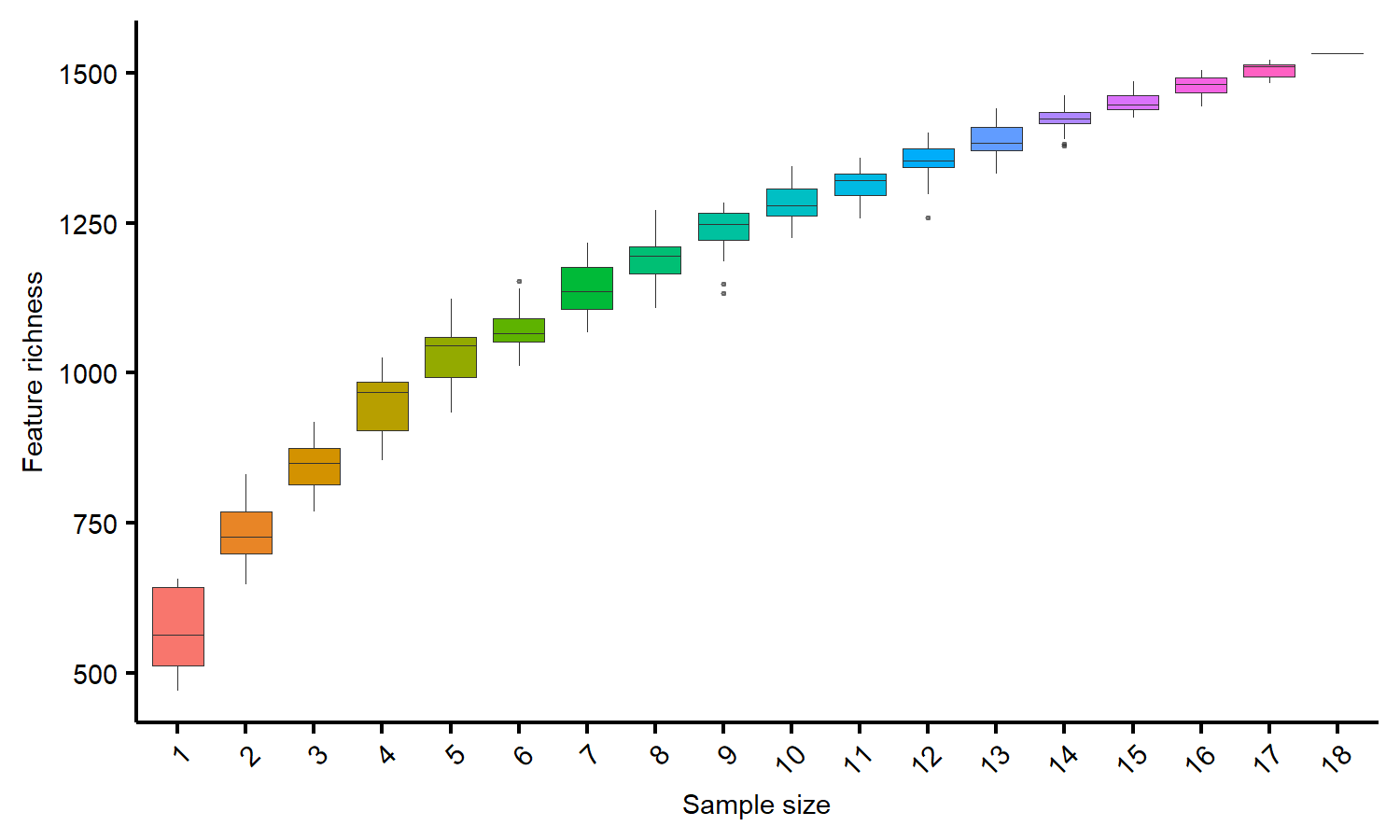


图8. 样本稀释梯度箱线图，从1-18个样本对应的丰富度值。只计算并展示9个梯度。

# 默认值绘制样本箱线图稀释曲线  
(p = alpha\_sample\_rare(otutab, count\_cutoff=9))



ggsave(paste0("p9.sample\_rare.pdf"), p, width=89\*1.5, height=56, units="mm")  
ggsave(paste0("p9.sample\_rare.png"), p, width=89\*1.5, height=56, units="mm")

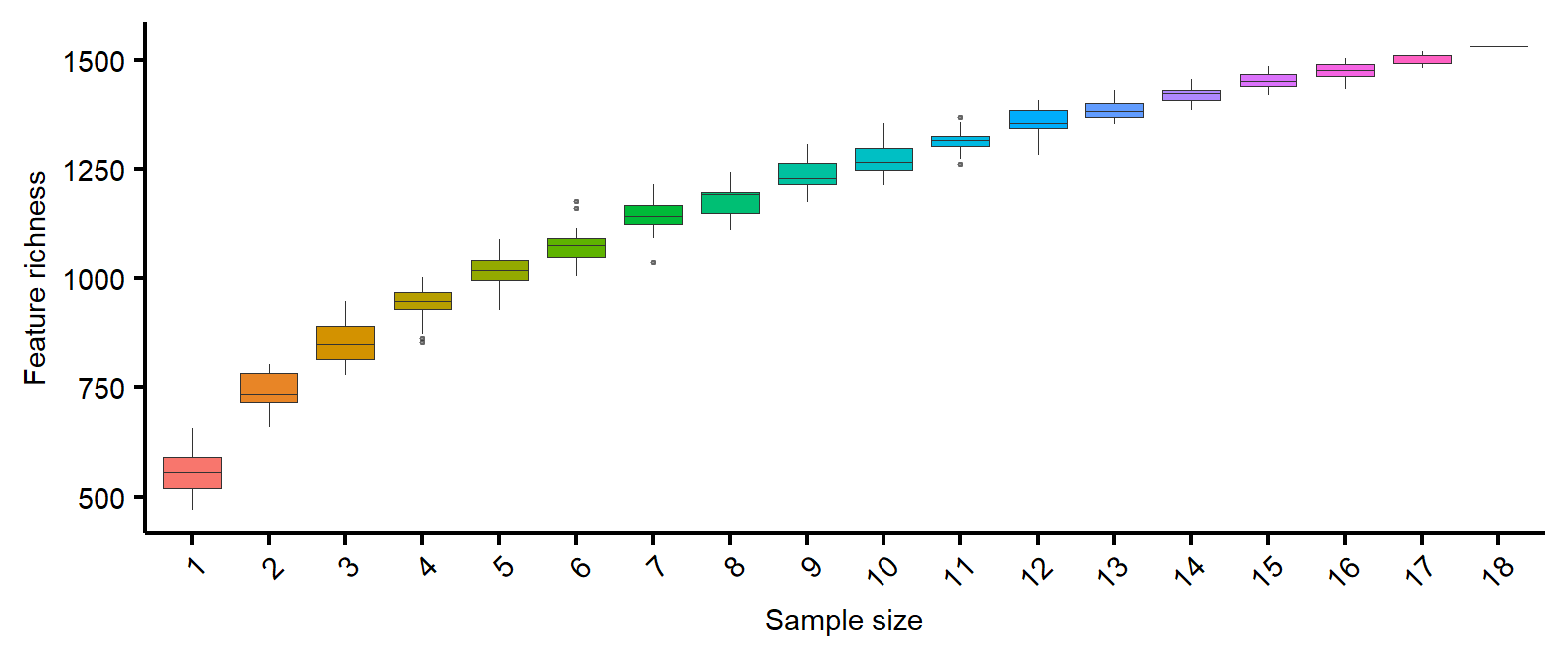


图9. 样本稀释梯度箱线图，从1-18个样本对应的丰富度值。阈值(count\_cutoff)从1修改为9，即9个读长才算可检测的特征，多样性增长的趋势变明显。因此阈值对多样性有极大的影响，可以适合不同场景表达不同的意义。如你有特别多的样品，如果count\_cutoff=1显示很少样本就达到饱和，则应该提高阈值，来突出本项目有足够多的样本才收集到如此高的多样性，即表达大样本量是非常有必要且有意义的。

此外，QIIME 2中都有相应绘制稀释曲线的方法，详见之前的教程：

* [QIIME 2. 《4人体各部位微生物组分析Moving Pictures》](https://mp.weixin.qq.com/s/c8ZQegtfNBHZRVjjn5Gyrw)中的图10

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)](https://doi.org/10.1007/s13238-020-00724-8)
* [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).](https://doi.org/10.1038/s41587-019-0104-4)

**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\_rare\_curve、alpha\_sample\_rare函数；文涛参与本文部分创作，编写了alpha\_rare\_all函数。

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

Bram Beckers, Michiel Op De Beeck, Nele Weyens, Wout Boerjan & Jaco Vangronsveld. (2017). Structural variability and niche differentiation in the rhizosphere and endosphere bacterial microbiome of field-grown poplar trees. Microbiome 5, 25, doi: <https://doi.org/10.1186/s40168-017-0241-2>

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>

Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, H. I. T. Consortium Meta, Maria Antolin, François Artiguenave, Hervé Blottiere, Natalia Borruel, Thomas Bruls, Francesc Casellas, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Miguel Forte, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Michiel Kleerebezem, Jan Knol, Michel Kristensen, Severine Layec, Karine Le Roux, Marion Leclerc, Emmanuelle Maguin, Raquel Melo Minardi, Raish Oozeer, Maria Rescigno, Nicolas Sanchez, Sebastian Tims, Toni Torrejon, Encarna Varela, Willem de Vos, Yohanan Winogradsky, Erwin Zoetendal, Peer Bork, S. Dusko Ehrlich & Jun Wang. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-65, doi: <https://doi.org/10.1038/nature08821>

责编：刘永鑫，中科院遗传发育所

版本1.0.0，提供USEARCH稀释结果、OTU表输入、QIIME2和样本稀释曲线多种方案 版本1.0.1，整合席娇的审稿意见，并全文修改