三元图

吴一磊

2020/6/6

Table of Contents

### 三元图在微生物组差异比较中的应用

本节作者：吴一磊，中科院微生物所

版本1.0.5，更新日期：2020年6月24日

本项目永久地址： <https://github.com/YongxinLiu/MicrobiomeStatPlot> ，本节目录 246TernaryPlot，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。

#### 背景知识

三元图(Ternary plot)1是描述三个变量之和为常数的质心图，其核心原理是：

1. 等边三角形内任意一点到三角形三边的距离之和等于其中一边上的高(常数)；
2. 过等边三角形内任意一点分别向三条边作平行线，按顺时针方向或逆时针方向读取平行线在各边所截取之三条线段，三条线段之和等于该等边三角形任一边之长(常数）。

在微生物组组领域，我们主要利用的是基于相位网格的三元图，其特点是每一种成分在相应等边三角形的顶点的比例为100%，在其对面的线为0%，将零点线和顶点按比例划分，用以估计各成分的含量。如图1：y对面的底线代表落在底线上的所有点在y中占比为0，随着平行线逐渐靠近顶点，落在相应平行线上的点在y中的占比越来越高。

if (!requireNamespace("ggtern", quietly=T))  
 install.packages("ggtern")  
library(ggtern)  
data <- data.frame(value=c(10, 30, 50, 70, 100),  
 x=c(90, 70, 50, 30 ,10),  
 y= c(10, 30, 50, 70, 90),  
 z=rep(0, 5))  
lables <- data.frame(value=c(10, 30, 50, 70, 90),  
 x=rep(0.5, 5),  
 y=c(0.2, 0.35, 0.5, 0.65, 0.8))  
p=ggtern(data,aes(x,y,z))  
p=p + geom\_crosshair\_tern() + geom\_mask() +  
 geom\_text\_viewport(x=lables$x, y=lables$y, label=lables$value, color='red') +   
 theme\_bw()  
p

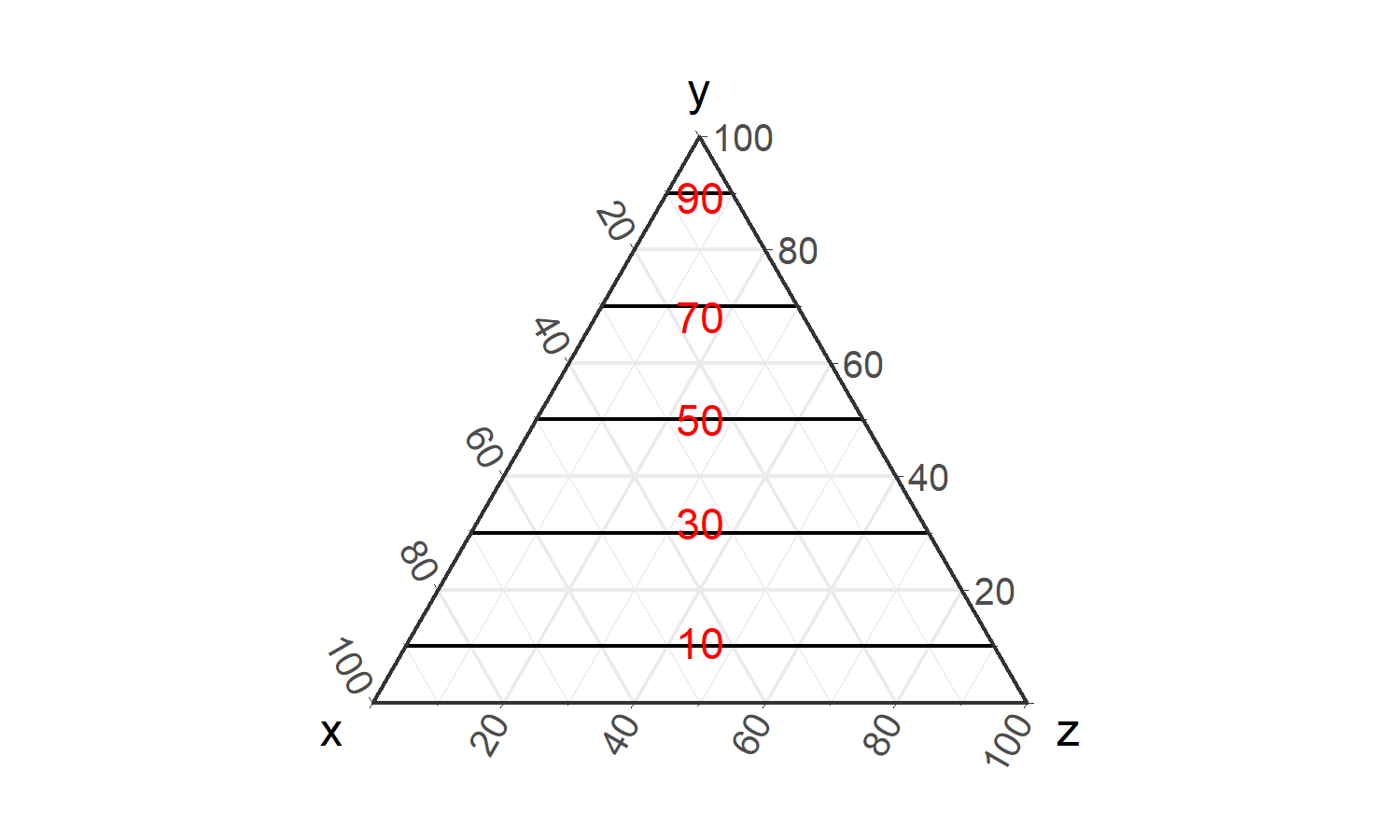


图1. 在y组中的占比逐渐增长

# 保存位图和矢量图，分别用于预览和排版  
ggsave(paste0("b1.TenaryPlot.png"), p, width=89, height=89, units="mm")  
ggsave(paste0("b1.TenaryPlot.pdf"), p, width=89, height=89, units="mm")

在解读三元图时，我们可以通过点的位置快速获得其在三个分组中的相对比例信息：目标点越靠近一角的顶点，说明他在相应的分组中比例越高；反之其相对比例就越低。根据图1，经过点的平行线在等边三角形两边的截距代表该点在对应顶点分组的占比，因此可以得出该点的在三个分组中的占比情况。如图2：过绿色点分别向三条边做平行线，然后按逆时针方向依次读取平行线在三条边的截距约为(0.33, 0.33, 0.33)，因此该点在x, y, z 三个组分中的占比为0.33, 0.33, 0.33；依此原则可得出，红点占比为0.1，0.2，0.7；蓝点占比为0.2，0.7，0.1；该结果跟我们的作图代码是相符的。

data <- data.frame( x=c(0.33, 1, 2),  
 y=c(0.33, 2, 7),  
 z=c(0.33, 7, 1))  
p = ggtern(data, aes(x=x, y=y, z=z)) +   
 geom\_point(size=5, alpha=0.7, color=c("green", "red", 'navy')) +   
 geom\_mask() +  
 annotate(geom ='text',  
 x =c(0.33, 1, 2),  
 y =c(0.33, 2, 7),  
 z =c(0.33, 7, 1),  
 vjust=c(1.5, 1.5, 1.5),  
 angle=c(0, 30, 60),  
 label=c("(0.33, 0.33, 0.33)", "(0.1, 0.2, 0.7)", "(0.2, 0.7, 0.1)" ),  
 color=c("green", "red", 'navy')) +   
 theme\_rgbw() + geom\_crosshair\_tern(size=1)   
p

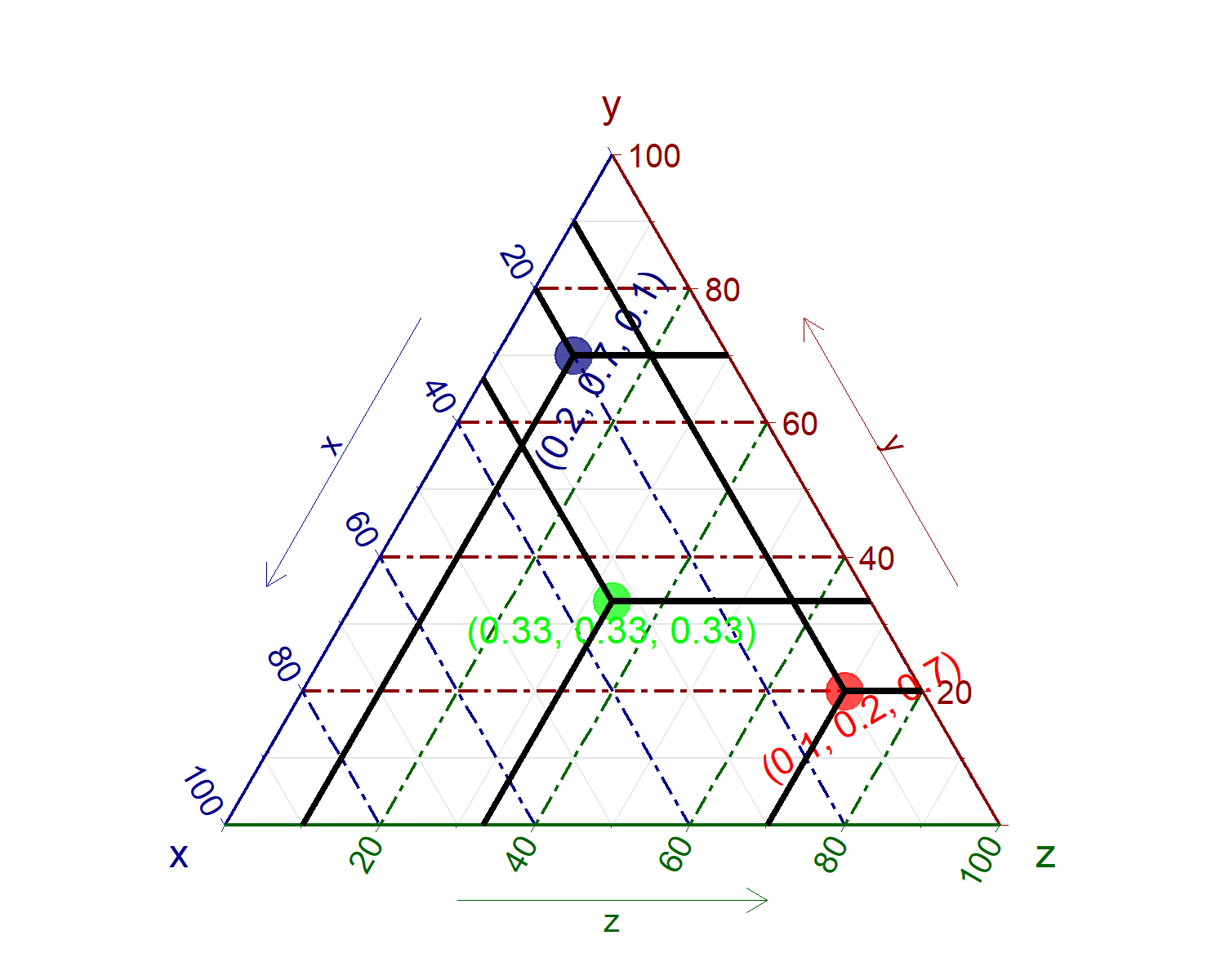


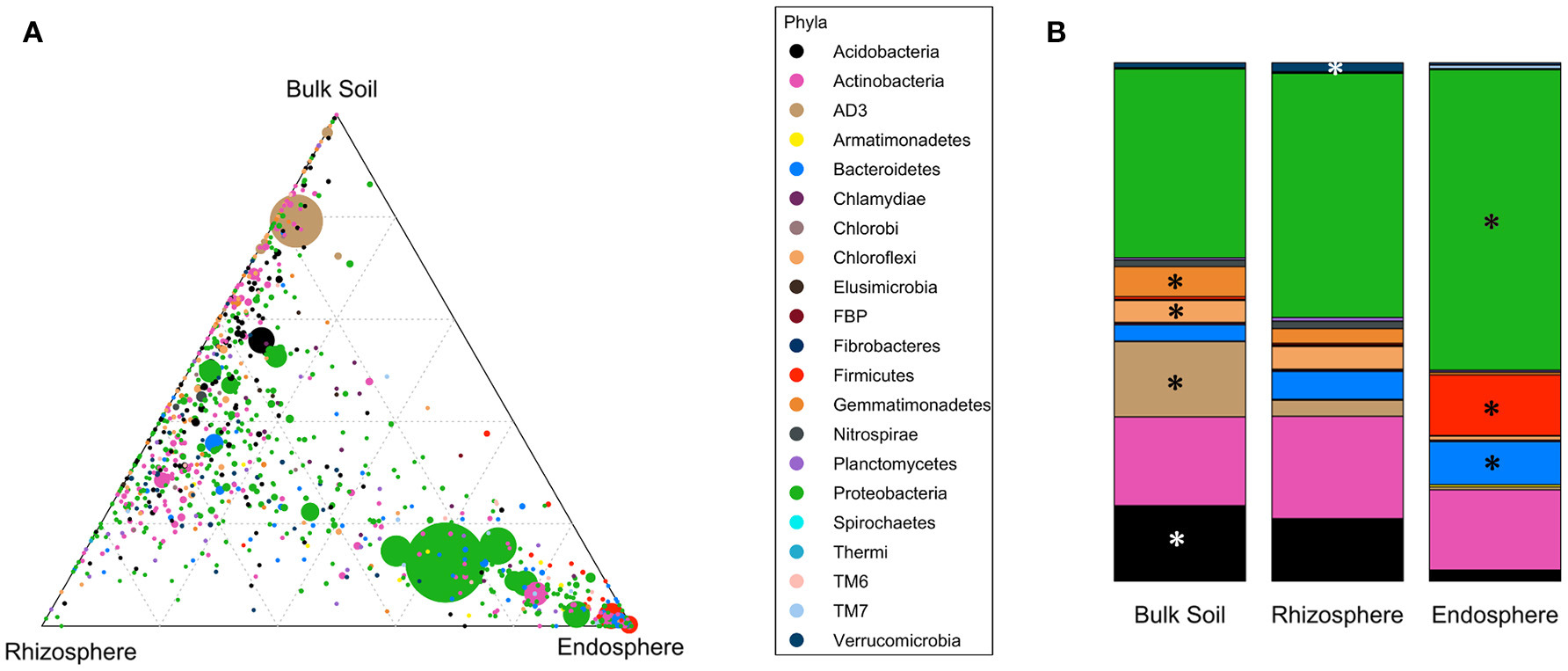
图2. 各点在x, y, z中的占比

在微生物多样性实际分析中，由于OUTs数目繁多，我们并不需要获得如此精确的占比情况。一般来说，三元图不同的点代表不同的OTUs（或其他分类水平），点的大小代表该OTUs的平均丰度(一般需要log2或log10等转换)。不仅如此，还可以对OTUs在各微环境中的丰度数据进行统计检验后，得出各OTU分别在哪种微环境中显著富集，此时三元图不仅表现出OTUs或者物种在微环境中的相对比例，还包含显著性统计结果，它打破了火山图或韦恩图两两比较的结果，总共展示了6次两两比较的结果，即每个组的富集情况是相对于其他两组的。

#### 实例解读

##### 例1. 三个部位的分布差异比较

本例2017年发表于自Frontiers in Microbiology文章的图4，描述了植物从三个北高寒带气候区的共同核心类群中组装特定物种的细菌群落2。

 图3. 三元图展示OTUs的空间特异性(Kumar, et al., 2017)

**图片描述：**

三元图展示不同区域(土体土、根际土、根内)的样本的群落结构在门水平上的差异情况,每个圆点代表一个OTU，OTU的大小、颜色和位置分别代表其相对丰度、菌门水平名称和分组情况。

**Distribution of OTUs and phyla across different compartments.** (A) Ternary plot of all OTUs plotted based on the compartment (Bulk soil, Rhizosphere soil, Endosphere) specificity. Each circle represents one OTU. The size, color and position of each OTU represents it relative abundance, bacterial phyla and affiliation of the OTU with different compartments, respectively.

**图注描述注意事项：**

1. 总述图表展示的信息：三元图展示OTUs在三个不同部分的空间特异性；
2. 详细描述图片中各元素代表什么：每个圆点代表一个OTU，OTU的大小、颜色和位置分别代表其相对丰度、菌门水平名称和分组情况。

**文中的图片解读：**

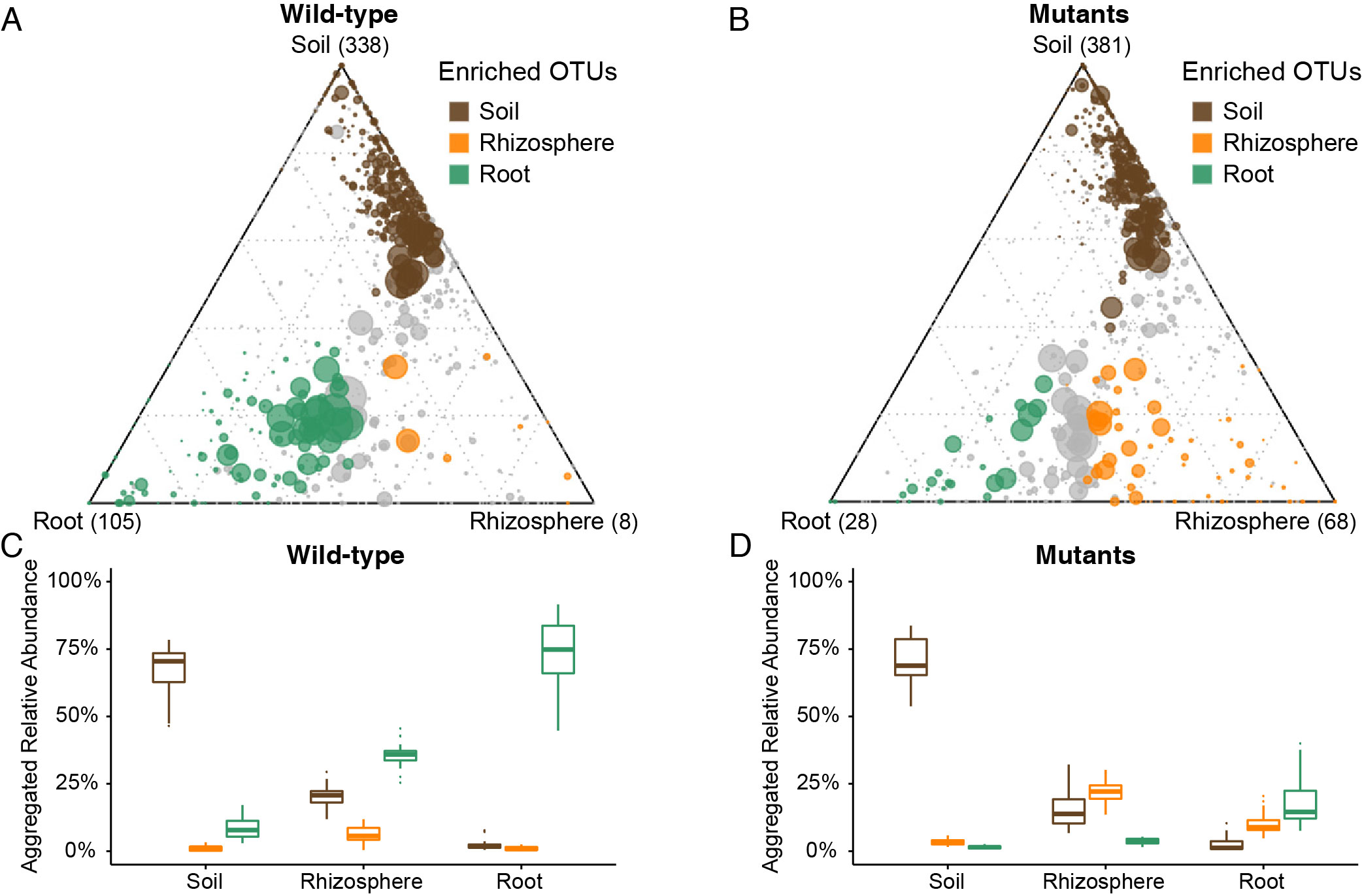
在门水平的不同部位中，细菌群落结构明显不同。 这些差异主要是由根内来源的序列数据集中相对较丰富的厚壁菌门驱动的，而在土体土和根际土中它们的丰度却很低。变形杆菌和拟杆菌的相对丰度从土体土到根际土再到根内逐渐增加，同时候选门AD3、芽单胞菌门和绿弯菌门的相对减少，它们共同构成根内群落的<4％。

**主要描述了比较突出的情况，如丰度较高、规律性变化或明显聚集在某区域的门。**

Bacterial community structures were clearly different in the different compartments at the phylum level. These differences were mainly driven by strong relative enrichment of Firmicutes in the endosphere-derived sequence data sets, compared to their very low abundances in the bulk and rhizosphere soils. The relative abundances of Proteobacteria and Bacteroidetes increased progressively from bulk to rhizosphere soil to the endosphere, with a concomitant decrease in those of candidate division AD3, Gemmatimonadetes and Chloroflexi, which collectively constituted <4% of endosphere communities.

##### 例2. 不同区域的分布差异

本例选自2016年发表于PNAS杂志上一文的图43，介绍了百脉根结瘤突变体中根际微生物组的研究。

 图4. 三元图展示不同取样部位中特异富集的OTUs(Zgadzaj et al., 2016)

**图片描述：**

1. 最上方的“Wild-type”和“Mutants”指明材料类型，分别为野生型和突变体；
2. 三个顶点分别为三个取样部分，并在括号中指出显著富集的OTU数量；
3. 点的大小代表三组样品的平均相对丰度；
4. 通过颜色指示显著富集情况：土壤=土色， 根际土=橘黄色，根系=绿色，灰色=两两比较中未全部显著富集的（均未富集，只相对其中一个分组富集）。

Ternary plots depicting compartment RA of all OTUs (>5 ‰) for WT SampleID (A; WT; n=73) and mutant SampleID (B; nfr5-2, nfr5-3, nin-2, and lhk1-1; n=118) across three soil batches (CAS8–CAS10). Each point corresponds to an OTU. Its position represents its RA with respect to each compartment, and its size represents the average across all three compartments. Colored circles represent OTUs enriched in one compartment compared with the others (green in root, orange in rhizosphere, and brown in root SampleID). Aggregated RAs of each group of enriched OTUs (root-, rhizosphere- and soil-enriched OTUs) in each compartment for the WT SampleID (C; WT; n=73) and mutant SampleID (D; nfr5-2, nfr5-3, nin-2, lhk1-1; n=118) are shown. In each compartment, the difference from 100% RA is explained by OTUs that are not significantly enriched in a specific compartment.

**总结：**

该图在例1的基础上添加了组间显著性差异比较的结果，包含了6次两两比较和三次韦恩图比较的结果，信息高度概括。同时，作者还进一步结合相对丰度的箱线图突出组间差异出；此外，通过与B图的横向比较相同的色系，明显的分布差异，两类材料微生物组的不同一目了然。

#### 绘图实战

由于前期数据处理是三元图主要的难点，所以在这里将数据处理和可视化分开，使用时便于检查异常、调整分析细节。

##### 环境设置

本教程需要在R语言环境下运行，推荐在RStudio界面中学习。目前测试版本为：Windows 10，R 4.0.x和 RStudio 1.3.x。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，有问题尝试搜索自行解决、欢迎分享经验。

**按需求安装，没必要每次都运行该安装代码，一般运行一次即可。**

p\_list = c("tidyverse", "ggtern", "BiocManager")  
for(p in p\_list){  
 if (!requireNamespace(p, quietly = TRUE))  
 install.packages(p)}  
if (!requireNamespace("edgeR", quietly=T))  
 BiocManager::install("edgeR")

##### 数据处理

OTU表数据源来自MicrobiomeStatPlot项目中Data/Science2019目录中的otutab.txt(<https://github.com/YongxinLiu/MicrobiomeStatPlot/blob/master/Data/Science2019/otutab.txt>)。

函数data\_clean参数介绍：

* 输入：特征表和样本元数据
* 输出：各组均值，用于三元图中的位置
* otu: 特征表，如OTU/ASV表
* design: 实验设计信息，**样品信息的列名设置为SampleID，分组信息列名设置为Group**
* type: otu\_table 类型，绝对丰度absolute/相对丰度relative；如果是绝对丰度，一定要注明。
* threshold: 相对丰度阈值，希望丢弃(获取)的相对丰度大小
* times: 点的倍数变化，主要跟可视化有关，如果结果图中的legend全是小数，可以通过倍数变化将其转变为正数，也可以在可视化时自己调整，不影响其传递的丰度相对大小信息。

library(tidyverse)  
# 数据处理函数  
data\_clean <- function(otu, design, type=c("relative", "absolute"), threshold=0.001, times=100){  
   
 # 函数测试数据  
 # library(amplicon)  
 # otu=otutab  
 # metadata$SampleID=rownames(metadata)  
 # design=metadata[,c("SampleID","Group")]  
 # type="absolute"  
 # threshold=0.0005  
 # times=100  
   
 # 绝对丰度转相对丰度1  
 if (type == "absolute"){  
 otu\_relative <- apply(otu, 2, function(x){x/sum(x)})  
 }else {otu\_relative <- otu}  
  
 # 至少有一个样本大于阈值即保留  
 idx <- rowSums(otu\_relative > threshold) >= 1  
 otu\_threshold <- as.data.frame(otu\_relative[idx, ])  
 otu\_threshold$OTUs <- row.names(otu\_threshold)  
   
 #转换宽表格为长表格  
 otu\_longer <- pivot\_longer(data=otu\_threshold,   
 cols=-OTUs,  
 names\_to="SampleID",   
 values\_to="value")  
   
 # 按"SampleID"对应添加元数据中的分组Group  
 merge\_data <- merge(otu\_longer, design, by ="SampleID")  
 # 去除样本列  
 # otu <- subset(merge\_data, select=-SampleID)  
 # 元数据不只有样本列，直接筛选OTUs、Group和value更稳健  
 otu <- subset(merge\_data, select=c("Group","OTUs","value"))  
 # 按OTUs和Group求均值  
 otu\_mean <- otu %>% group\_by(OTUs, Group) %>%   
 summarise(value=mean(value))  
 # 转换回宽表格  
 otu\_tern <- otu\_mean %>%  
 group\_by(Group, OTUs) %>%  
 mutate(index=row\_number()) %>%  
 pivot\_wider(names\_from=Group,values\_from=value) %>%  
 select(-index)  
 # 此处用group\_by可以直接合并均值，不用长、宽转换两次  
   
 # 调整点大小均值，可缩放 size of points  
 otu\_tern$size <- (apply(otu\_tern[2:4], 1, mean))\*times   
 return(otu\_tern)  
}

# 读取输入文件  
otutab <- read.delim("otutab.txt", header=T, row.names=1)  
design <- read.delim("metadata.txt", header=T, row.names=NULL)  
# 只提取元数据中的样本名和分组列，要求名称为SampleID和Group  
design = design[,c("SampleID","Group")]  
  
# 计算三元图输入数据：各组相对丰度均值  
otu\_tern <- data\_clean(otutab, design, type="absolute", threshold=0.001, times=100)  
head(otu\_tern,n=3)

# A tibble: 3 x 5  
# Groups: OTUs [3]  
 OTUs KO OE WT size  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 ASV\_10 0.0173 0.00800 0.00980 1.17   
2 ASV\_100 0.00229 0.00162 0.00115 0.169  
3 ASV\_101 0.00144 0.00222 0.00147 0.171

##### ggtern可视化

ggtern 是Nicholas Hamilton开发的，用于创建三元图的ggplot2的扩展包，详细参数和用法见?ggtern或[官方说明文档](http://www.ggtern.com/)。

library(ggtern)  
# x/y/x指定三个组名，显示顺序为左、中、右  
p <- ggtern(data=otu\_tern, aes(x=KO, y=OE, z=WT)) +   
 geom\_point(aes(size=size), alpha=0.8, show.legend=T) +  
 scale\_size(range=c(0, 6)) + geom\_mask() +   
 guides(colour="none") + theme\_bw() +  
 theme(axis.text=element\_blank(), axis.ticks=element\_blank())  
p

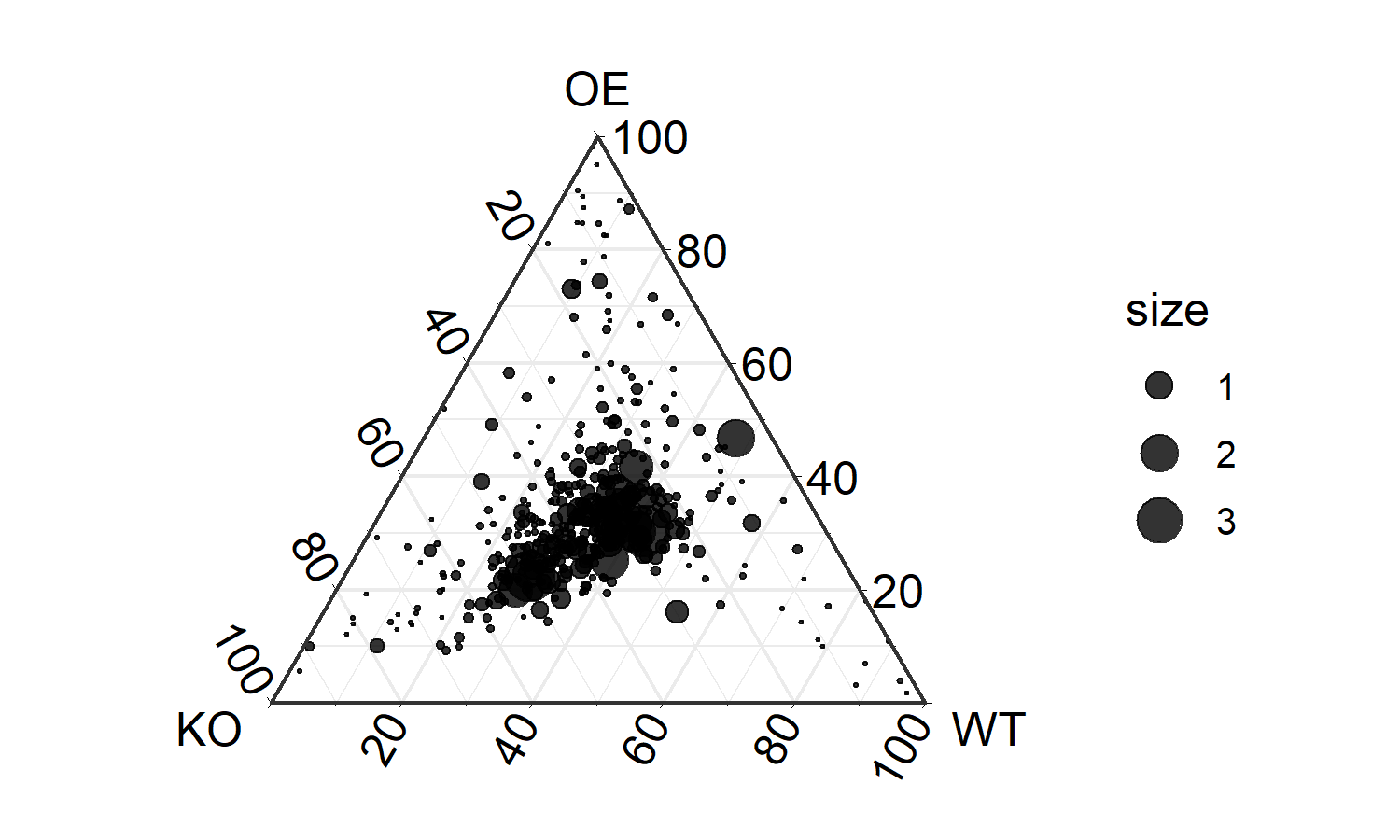


图5. 三元图展示OTUs的实验分组特异性

**图片描述：**

三元图展示不同处理组(KO, OE, WT)的样本的群落结构在OTUs上的丰度差异情况,每个圆点代表一个OTU，OTU的大小和位置分别代表其相对丰度和分组情况。

**图注描述注意事项：**

由于该图中并未全部展示所有OTUs，所以在描述的时候，最好对数据筛选的参数进行描述：

1. 图中只展示了相对丰度>0.1%的OTUs；
2. 点大小对应的平均相对丰度的倍数变化；

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)组存在丰度差异，即基因的有无可对微生物群落的丰富度引起变化。

##### 高丰度特征着色

图5能够展示的信息和按top丰度着色后的可视化方案是一致的，但叙述时可以像图3一样着重介绍突出情况，比如丰度较高的OTUs有那些，富集情况很特殊的有哪些。

函数top\_OTUs参数介绍：ggtern输入文件筛选丰度前N(10)的OTUs

* data: ggtern输入文件
* rank: 希望进行着色的丰度排名，推荐10个。如果超过10个，则需要自己制定配色方案。

top\_OTUs <- function(data, rank=10){  
 # 按丰度size降序排列  
 data\_order <- data[order(data$size, decreasing=T), ]  
 # if (missing(rank))  
 # rank=10  
 # 提取前N行  
 top <- data\_order[1:rank, ]  
 # 其余部分  
 otu\_el <- data\_order[-(1:rank), ]  
 # 返回top和其它的列表  
 return(list(top, otu\_el))  
}

plot\_data <- top\_OTUs(otu\_tern, rank=10)  
  
# 配色方案  
platte <- c('#a6cee3', '#1f78b4', '#b2df8a', '#33a02c', '#fb9a99',  
 '#e31a1c', '#fdbf6f', '#ff7f00', '#cab2d6', '#6a3d9a')  
  
p <- ggtern(data=otu\_tern,   
 aes(x=KO, y=OE, z=WT)) +   
 geom\_mask() +   
 geom\_point(data=plot\_data[[2]], aes(size=size), color="grey",  
 alpha=0.8, show.legend=F) +  
 geom\_point(data=plot\_data[[1]], aes(size=size, color=OTUs),   
 show.legend=T) +  
 scale\_colour\_manual(values=platte) +  
 scale\_size(range=c(0, 6)) +  
 # legend  
 guides(size="none") +  
 theme\_bw() +  
 theme(axis.text=element\_blank(),   
 axis.ticks=element\_blank())  
p

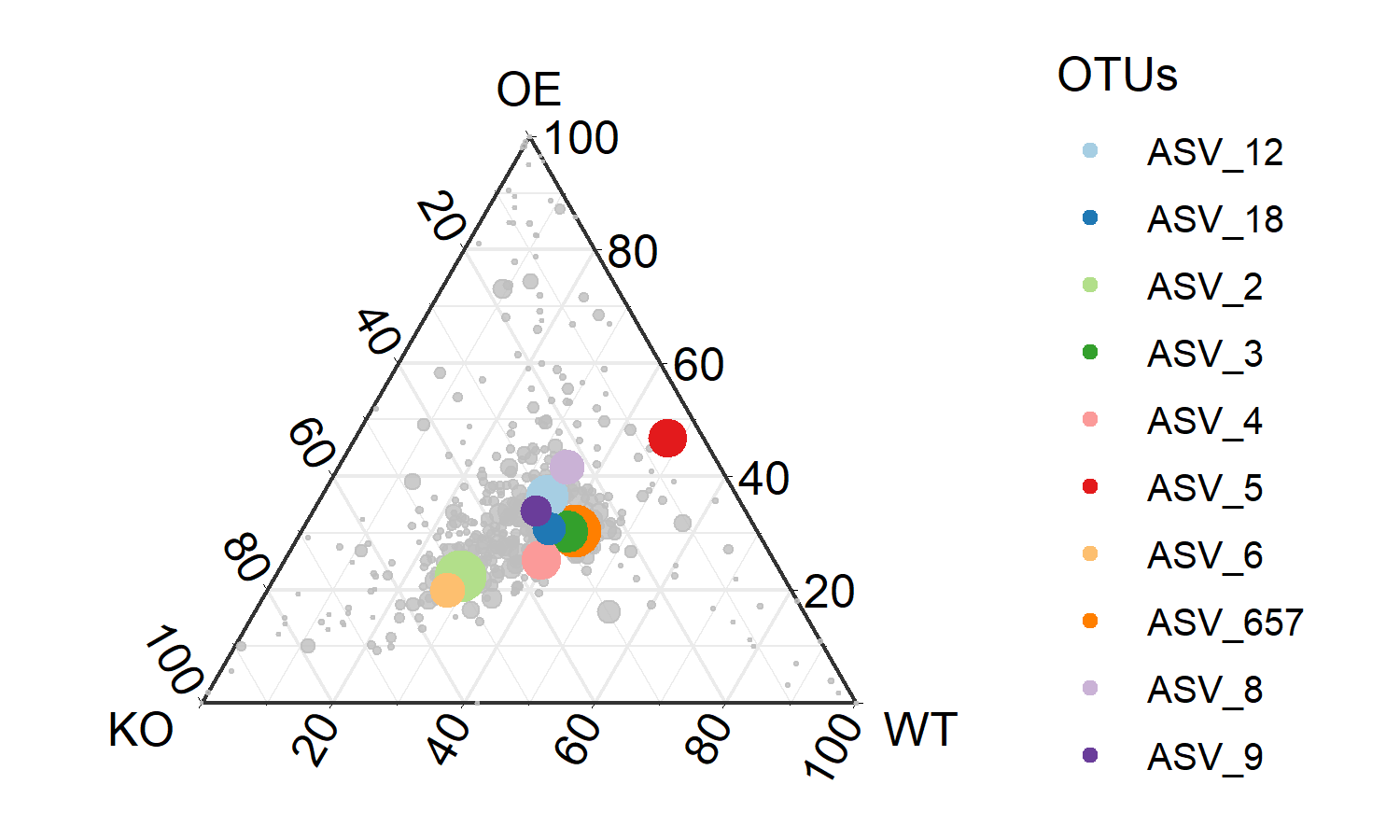


图6. 三元图展示OTUs的实验分组特异性

**图片描述：**

三元图展示不同处理组(KO, OE, WT)的样本的群落结构在OTUs上的丰度差异情况，每个圆点代表一个OTU，OTU的大小、颜色和位置分别代表其相对丰度、OUTs名称和分组情况。 值得注意的是，红色原点所代表的ASV\_5在KO组中丰度几乎为0，但在WT和OE均有相对较高的丰度。具体可以再结合物种注释进一步描述。

##### 富集显著性分析

除了上面直接展示OTUs在不同分组中的丰度富集情况，还可以通过显著性富集分析，获得在不同分组中显著富集的OTUs，最后进行可视化（例2）。

library(edgeR)  
# 计算3组比较各组特有的特征  
enrich\_data <- function(otu, design, p.value=0.05, adjust.method="fdr"){  
   
 # 函数测试数据  
 # library(amplicon)  
 # otu=otutab  
 # metadata$SampleID=rownames(metadata)  
 # design=metadata[,c("SampleID","Group")]  
 # p.value=0.05  
 # adjust.method="fdr"  
  
 dge\_list <- DGEList(counts=otu, group=design$Group)  
 # Remove the lower abundance/(cpm, rpkm)  
 keep <- rowSums(dge\_list$counts) >= 0  
 dge\_keep <- dge\_list[keep, ,keep.lib.sizes=F]  
 # scale the raw library sizes dgelist  
 dge <- calcNormFactors(dge\_keep)  
 # fit the GLM  
 design.mat <- model.matrix(~ 0 + dge$samples$group)  
 d2 <- estimateGLMCommonDisp(dge, design.mat)  
 d2 <- estimateGLMTagwiseDisp(d2, design.mat)  
 fit <- glmFit(d2, design.mat)  
 #######  
 # if (missing(adjust.method))  
 # adjust.method="fdr"  
 # if (missing(p.value))  
 # p.value=0.05  
 group\_index <- as.character(design$Group[!duplicated(design$Group)])  
 # enrich groups  
 lrt\_1\_2 <- glmLRT(fit, contrast=c(1, -1, 0))  
 lrt\_1\_3 <- glmLRT(fit, contrast=c(1, 0, -1))  
  
 de\_1\_2 <- decideTestsDGE(lrt\_1\_2, adjust.method=adjust.method,   
 p.value=p.value)  
 de\_1\_3 <- decideTestsDGE(lrt\_1\_3, adjust.method=adjust.method,   
 p.value=p.value)  
   
 rich\_1 <- rownames(otu)[de\_1\_2 == 1 & de\_1\_3 == 1]  
 enrich\_1 <- data.frame(OTUs=rich\_1,   
 enrich=rep(group\_index[1], length(rich\_1)))  
 ###############################  
 lrt\_2\_3 <- glmLRT(fit, contrast=c(0, 1, -1))  
 lrt\_2\_1 <- glmLRT(fit, contrast=c(-1, 1, 0))  
   
 de\_2\_3 <- decideTestsDGE(lrt\_2\_3, adjust.method=adjust.method,   
 p.value=p.value)  
 de\_2\_1 <- decideTestsDGE(lrt\_2\_1, adjust.method=adjust.method,   
 p.value=p.value)  
   
 rich\_2 <- rownames(otu)[de\_2\_3 == 1 & de\_2\_1 == 1]  
 enrich\_2 <- data.frame(OTUs=rich\_2,   
 enrich=rep(group\_index[2], length(rich\_2)))  
 ###################  
 lrt\_3\_1 <- glmLRT(fit, contrast=c(-1, 0, 1))  
 lrt\_3\_2 <- glmLRT(fit, contrast=c(0, -1, 1))  
   
 de\_3\_1 <- decideTestsDGE(lrt\_3\_1, adjust.method=adjust.method,   
 p.value=p.value)  
 de\_3\_2 <- decideTestsDGE(lrt\_3\_2, adjust.method=adjust.method,   
 p.value=p.value)  
   
 rich\_3 <- rownames(otu)[de\_3\_1 == 1 & de\_3\_2 == 1]  
 enrich\_3 <- data.frame(OTUs=rich\_3,   
 enrich=rep(group\_index[3], length(rich\_3)))  
 enrich\_index <- rbind(enrich\_1, enrich\_2, enrich\_3)  
 return(enrich\_index)  
}

enrich\_index <- enrich\_data(otutab, design, p.value=0.05)  
plot\_data <- merge(otu\_tern, enrich\_index, by="OTUs", all.x=T)  
p <- ggtern(data=plot\_data,   
 aes(x=KO, y=OE, z=WT)) +   
 geom\_mask() + # 可将超出边界的点正常显示出来  
 geom\_point(aes(size=size, color=enrich),alpha=0.8) +   
 guides(size="none") +theme\_bw() +  
 theme(axis.text=element\_blank(),   
 axis.ticks=element\_blank())  
p

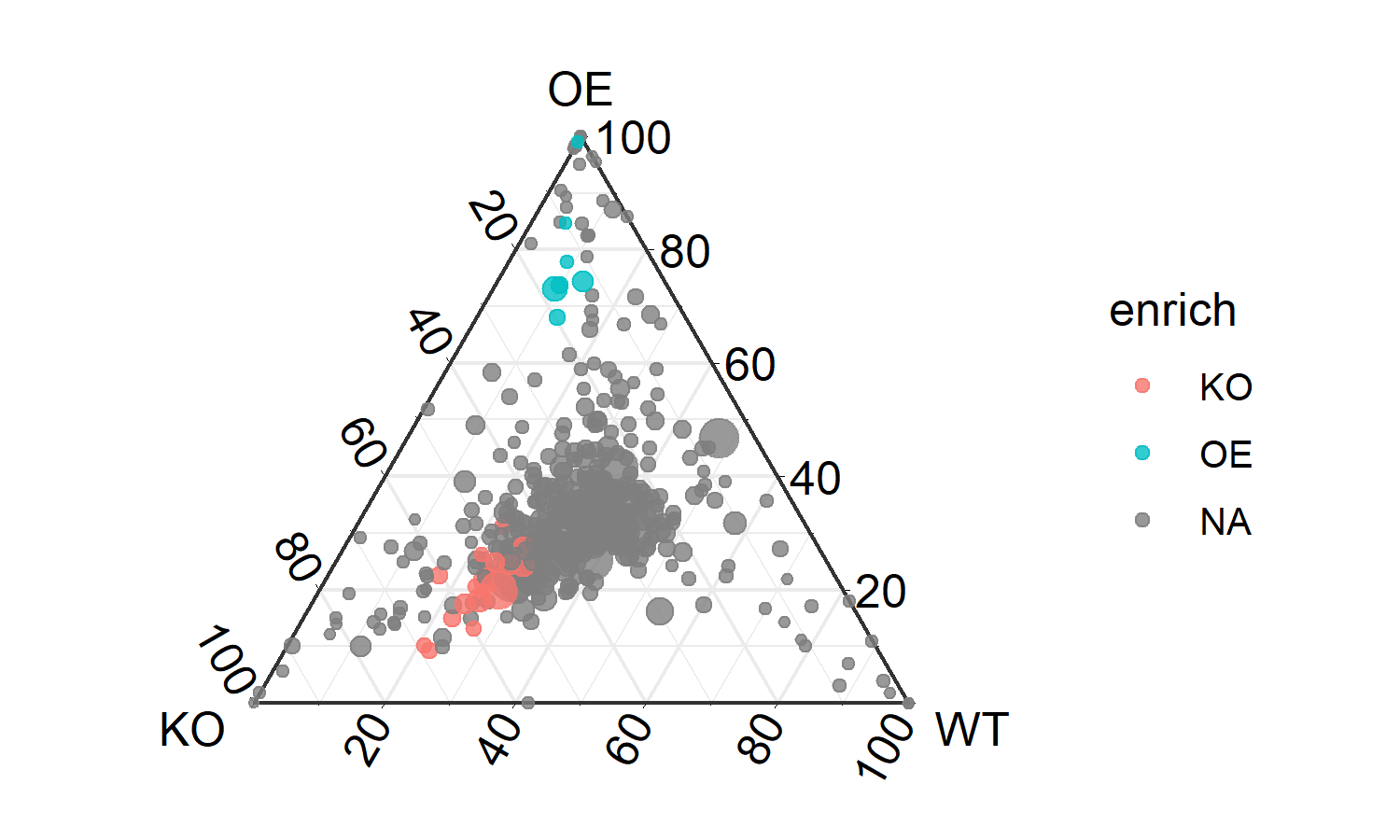


图7. 三元图展示富集OTUs

**图片描述：**

除了图5中的信息外，该图还展示了在KO, OE, WT三个分组中，分别相对于另外两个分组显著性富集的OTUs。

**图注描述注意事项：**

由于该图中并未全部展示所有OTUs，所以在描述的时候，最好对数据统计分析过程的参数进行描述：

1. 相对丰度阈值
2. 点大小对应的平均相对丰度的倍数变化
3. p.value
4. adjust.method

#### 参考文献

Manoj Kumar, Günter Brader, Angela Sessitsch, Anita Mäki, Jan D. van Elsas & Riitta Nissinen. (2017). Plants Assemble Species Specific Bacterial Communities from Common Core Taxa in Three Arcto-Alpine Climate Zones. Frontiers in Microbiology 8, doi: <https://doi.org/10.3389/fmicb.2017.00012>

1. wikipedia. *wikipedia* (2020).

2. Kumar, M. *et al.* Plants assemble species specific bacterial communities from common core taxa in three arcto-alpine climate zones. *Frontiers in Microbiology* **8**, (2017).

3. Zgadzaj, R. *et al.* Root nodule symbiosis in lotus japonicus drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. *Proc Natl Acad Sci U S A* **113**, E7996–E8005 (2016).