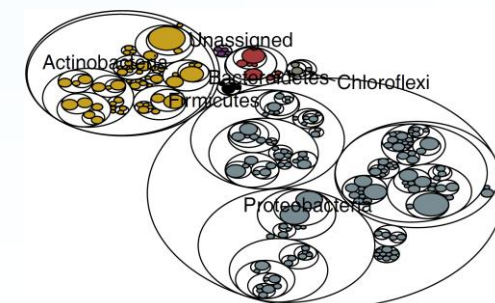
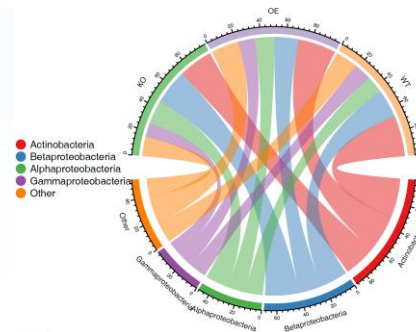
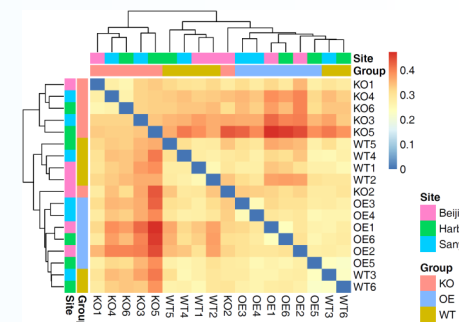
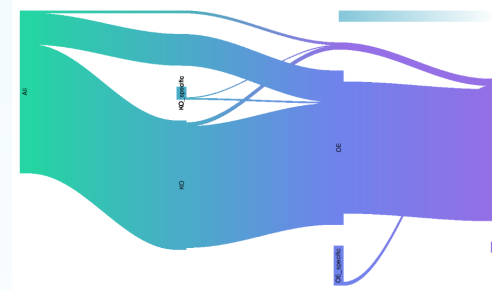




23物种多样性分析Alpha, Beta & Taxonomy

易生信
2023年6月10日



- Alpha多样性——样品自身多样性
- Beta多样性——样品/组间差异PCoA, CPCoA
- 物种组成——不同分类级别相对丰度

○ 总结



- Alpha多样性——样品自身多样性
- Beta多样性——样品/组间差异PCoA, CPCoA
- 物种组成——不同分类级别相对丰度

○ 总结

Alpha多样性指数

- 评价Alpha多样性的方法
- 常用richness(observed_otus)、chao1、ACE估计样本物种数量

$$chao1 = S_{obs} + \frac{F_1^2}{2F_2}$$

F_1 and F_2 are the count of singletons and doubletons

- Shannon-Wiener diversity计算每个特征的香农熵

$$H = - \sum_{i=1}^s (p_i \log_2 p_i)$$

P_i 代表第*i*个OTU的比例



- Dominance优势度指数, Simpson = 1 - dominance

$$\sum p_i^2$$

Pi代表第i个OTU的比例

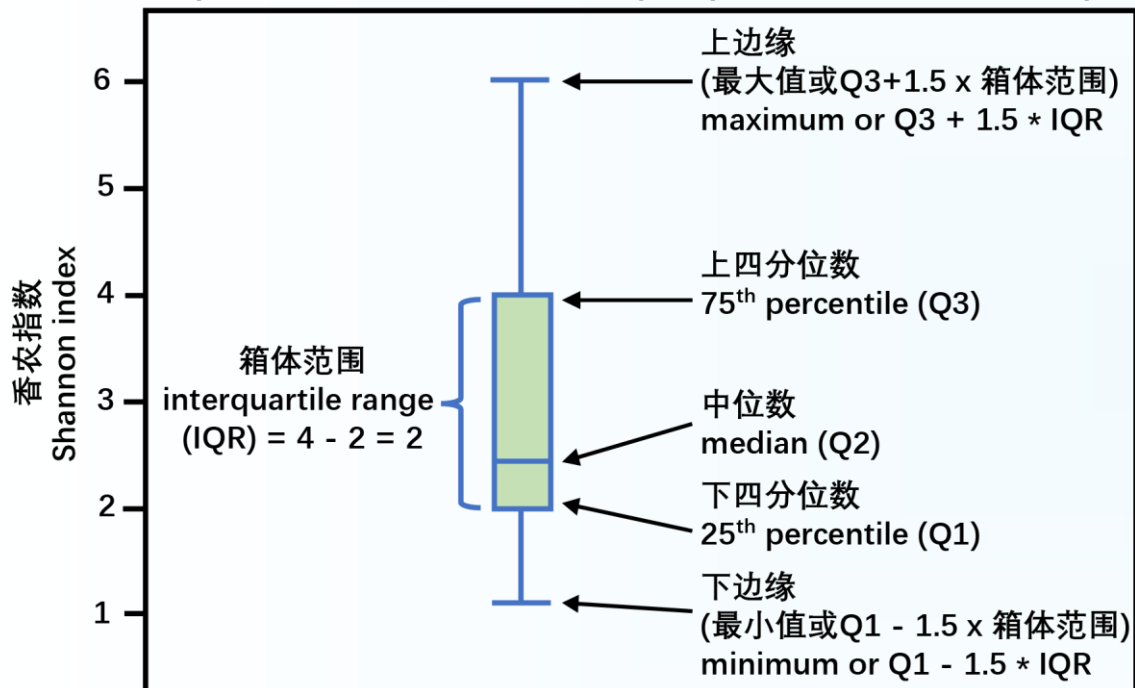
- alpha_div命令基于标准化OTU表计算14种alpha多样性指数计算
- Python3中的skbio.diversity.alpha包提供34种alpha多样性指数

<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html>

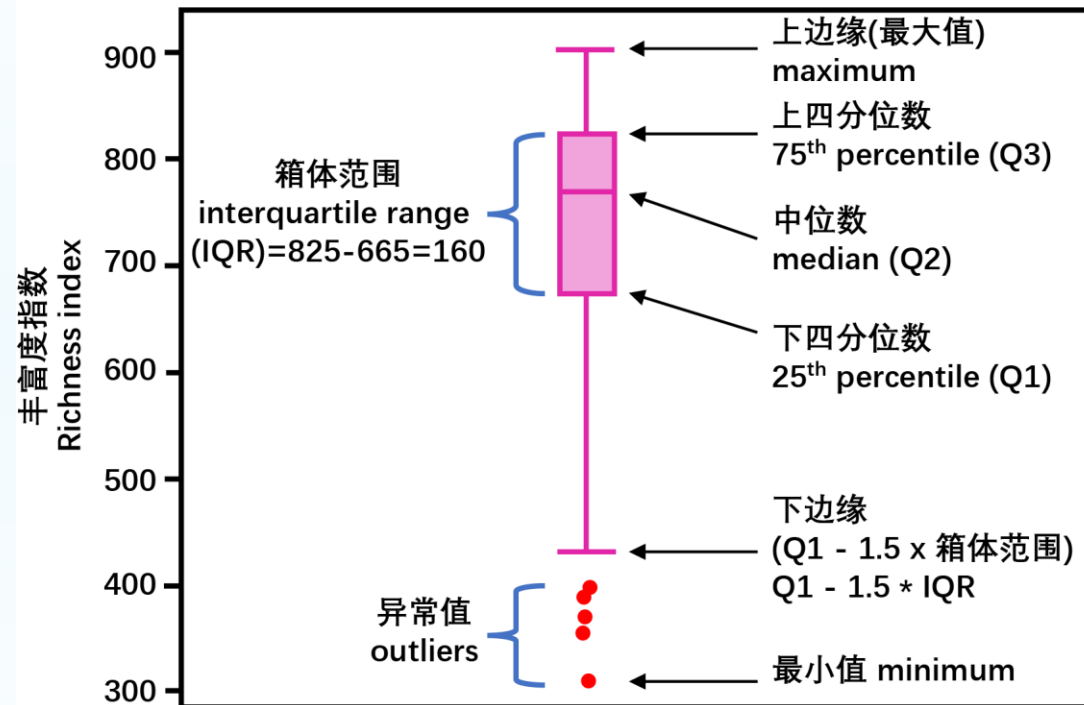
提供基本描述、计算公式和参考文献

箱线图的基本知识

Alpha多样性香农指数箱线图(Boxplot of Shannon index)

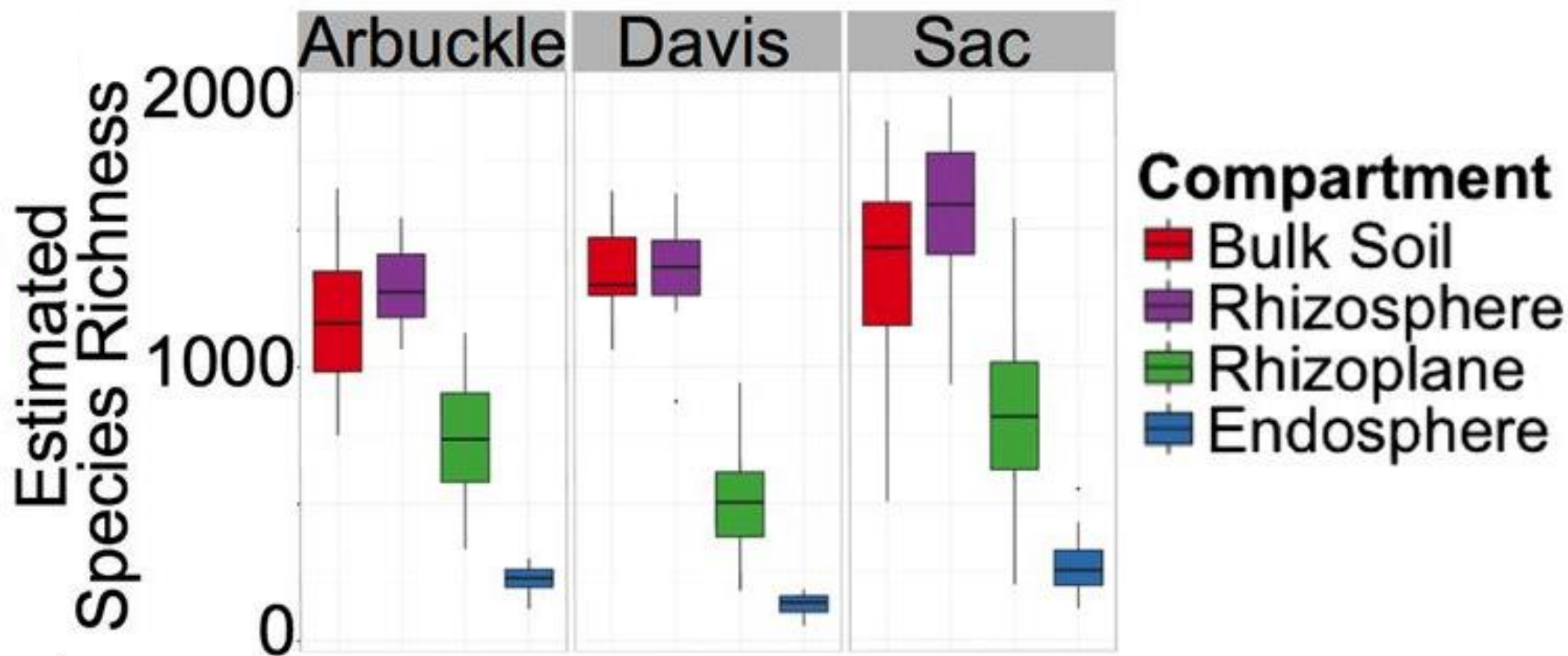


Alpha多样性丰富度指数箱线图(Boxplot of richness index)

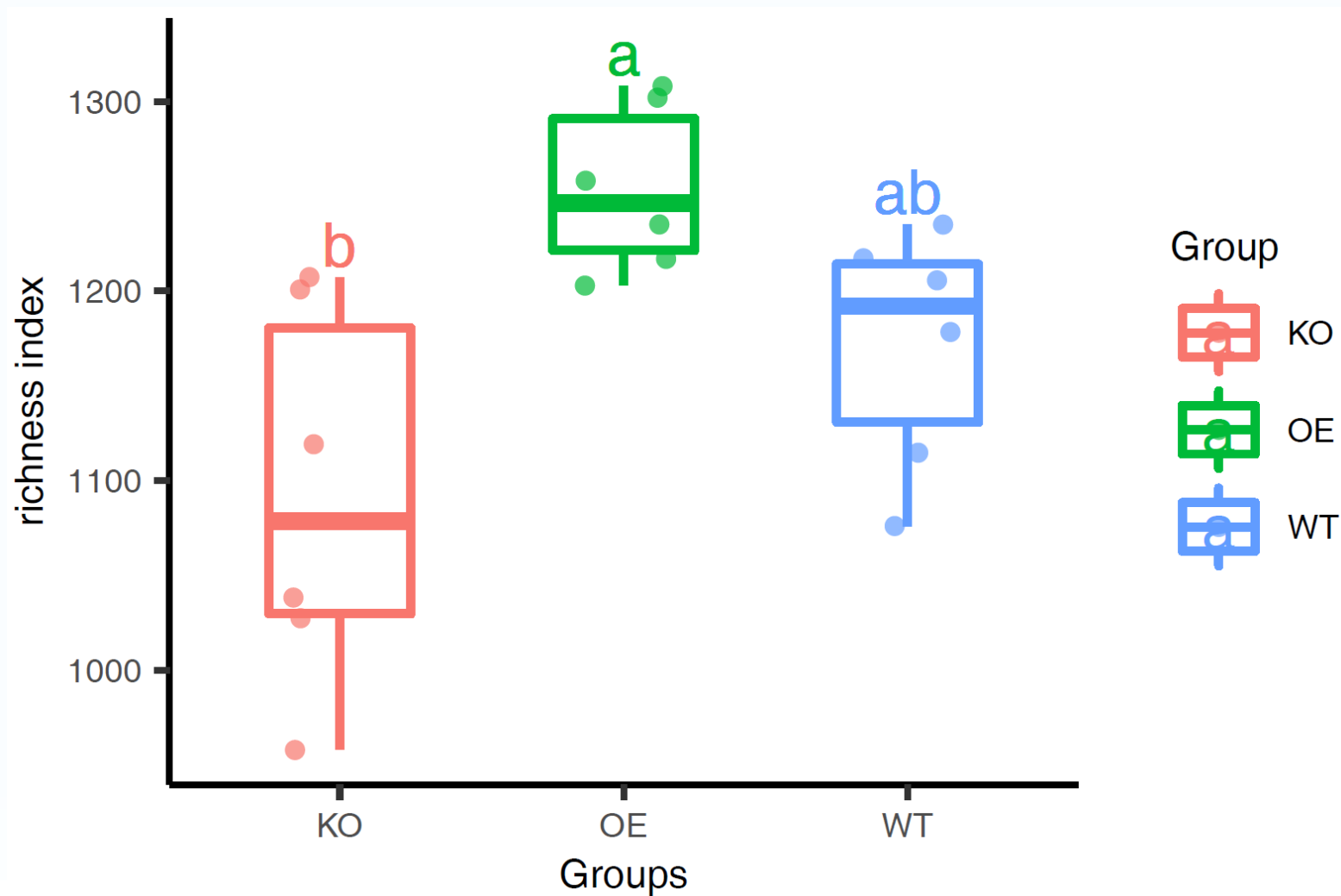


[Alpha多样性箱线图\(样章, 11图2视频\)](#)

1. 箱线图比较不同取样位置/品种间物种数量



Richness/Observed_features展示物种数量分布



宏基因组

易生信

制作Alpha多样性表格(pipeline.sh # 5.3和# 6)

- # 5.3 样品抽平至最小值，才能评估alpha多样性

```
Rscript ${bin}/script/otutab_rare.R --input result/otutab.txt \  
--depth 10000 --seed 1 \  
--normalize result/otutab_rare.txt \  
--output result/alpha/vegan.txt # (常用6种alpha多样性指数)
```

- # 6 计算15种多样性指数(chao1指数有误，不要使用)

```
usearch -alpha_div result/otutab_rare.txt -output result/alpha/alpha.txt
```

- # 稀释曲线：取1%-100%的序列中OTUs数量

```
usearch -alpha_div_rare result/otutab_rare.txt \  
-output result/alpha/alpha_rare.txt -method without_replacement
```



如何绘制上述箱线图——文件准备alpha目录下

- Alpha多样性指数：alpha/vegan.txt

SampleID	richness	chao1	ACE	shannon	simpson	invsimpson
KO1	2353	2713.802	2700.834	6.133361	0.990311	103.2051
KO2	2307	2641.016	2643.029	6.17347	0.991873	123.0445
KO3	1946	2324.613	2318.931	5.83144	0.989605	96.19992

- 实验设计：metadata.txt

SampleID	Group	Site	Date	BarcodeSequence
KO1	KO	Beijing	2017/6/30	ACGCTCGACA
KO2	KO	Beijing	2017/6/30	ATCAGACACG
KO3	KO	Sanya	2017/7/2	ATATCGCGAG
KO4	KO	Sanya	2017/7/2	CACGAGACAG
KO5	KO	Harbin	2017/7/4	CTCGCGTGTC
KO6	KO	Harbin	2017/7/5	TAGTATCAGC

基因组
信



方法1. 命令行模式下绘制Alpha多样性箱线图

查看帮助

```
Rscript ${bin}/script/alpha_boxplot.R -h
```

完整参数，程序、指数选择、输入文件、实验设计、分组列名、输出目录、图片宽(mm, 89为半栏)、图片高

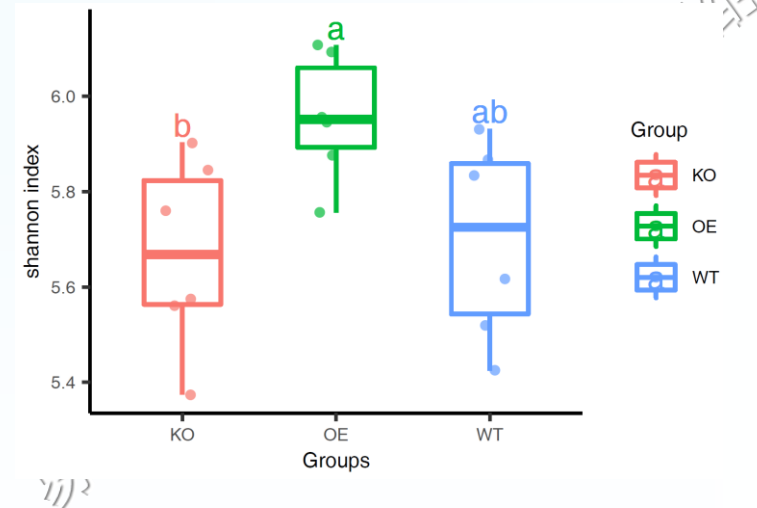
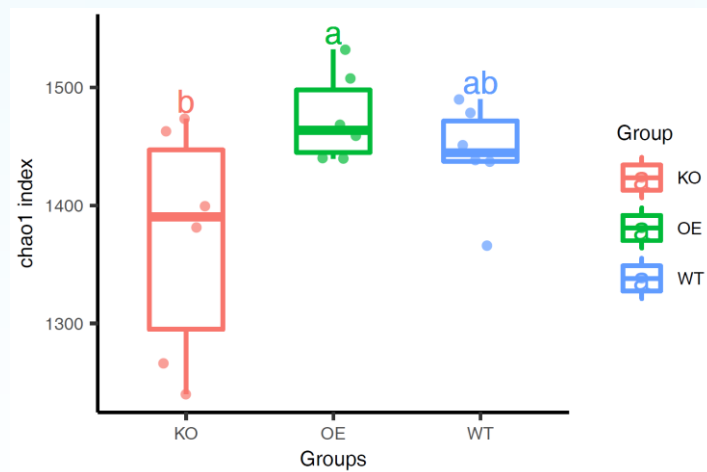
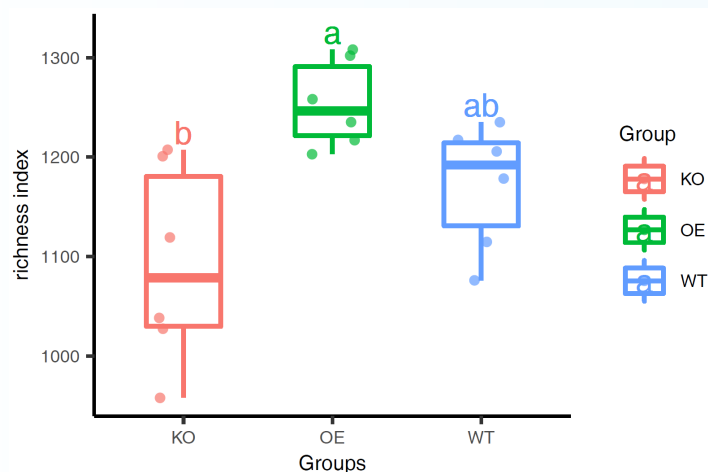
```
Rscript ${bin}/script/alpha_boxplot.R --alpha_index richness \  
--input result/alpha/vegan.txt --design result/metadata.txt \  
--group Group --output result/alpha/ \  
--width 89 --height 59
```



使用循环绘制6种常用指数

```
for i in `head -n1 result/alpha/vegan.txt|cut -f 2-`;do
  Rscript ${db}/script/alpha_boxplot.R --alpha_index ${i} \
    --input result/alpha/vegan.txt --design result/metadata.txt \
    --group Group --output result/alpha/ \
    --width 89 --height 59
done
```

比较不同指数Y轴的刻度值



方法2. Rmd模式下绘图箱线图(result目录下)

- 复制EasyAmplicon(正对照)中result/Diversity.Rmd至新项目result/, 需要有输入文件alpha/vegan.txt和metadata.txt
- Rstudio打开Diversity.Rmd文件
- 检查“## 实验设计”段落实验设计metadata、分组group和图片宽width高height(半版88 x 59 mm)
- 检查“# α 多样性”文件位置和指数类型, 默认为"Richness"
- 从头开始逐行运行代码并查看右侧变量, 也可右箭头按顺序逐个运行代码块, 或Knit运行整个床生成结果和计算过程网页报告
- 结果位于alpha目录下alpha/alpha_boxplot_Richness.pdf



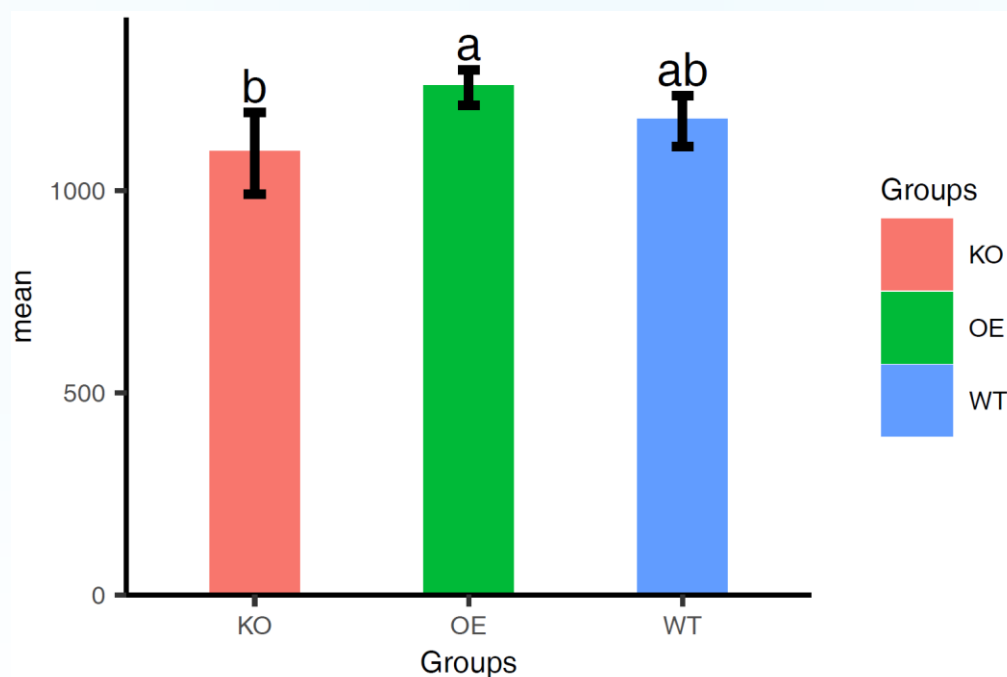
方法3. 在线绘制箱线图(选学)

- 用Excel打开metadata.txt和vegan.txt文件，分别复制group和richness列至新文件相邻两列，并选中新数据并复制
- <https://www.bic.ac.cn/ImageGP/> —— Boxplot
- Ctrl+A全选文本框，按Ctrl+C粘贴
- Legend选择group，Y-axis选择richness，点击PLOT
- 此外Legend order调整组显示顺序，Data preprocess设置异常值处理和数据变换；Layout设置方向，刻痕，图形样式，图例位置；title设置图、x、y轴标题；Picture attribute设置图片输出大小；
- 下方预览为png格式可单击放大和右键另存，PDF可下载，方便修改。



Alpha多样性柱状图+标准差

```
Rscript ${db}/script/alpha_barplot.R --alpha_index richness \  
--input result/alpha/vegan.txt --design result/metadata.txt \  
--group Group --output result/alpha/ \  
--width 89 --height 59
```



2. 稀释曲线展示OTUs数量随深度变化和测序饱和度

OTUs数量稀释曲线

a 根际

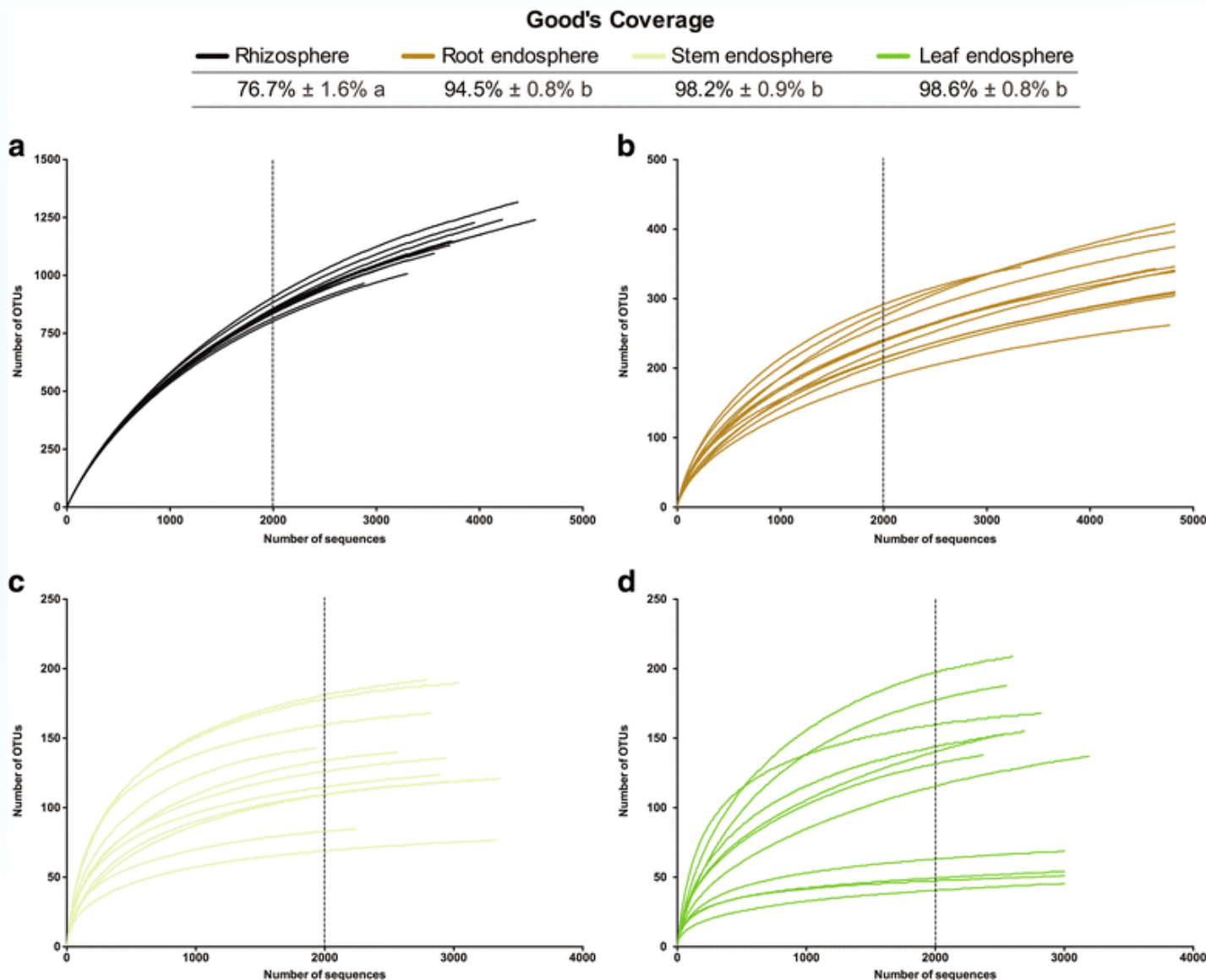
b 根

c 茎

d 叶

Beckers, Bram, et al. "Structural variability and niche differentiation in the rhizosphere and endosphere bacterial microbiome of field-grown poplar trees." Microbiome 5.1 (2017): 25.

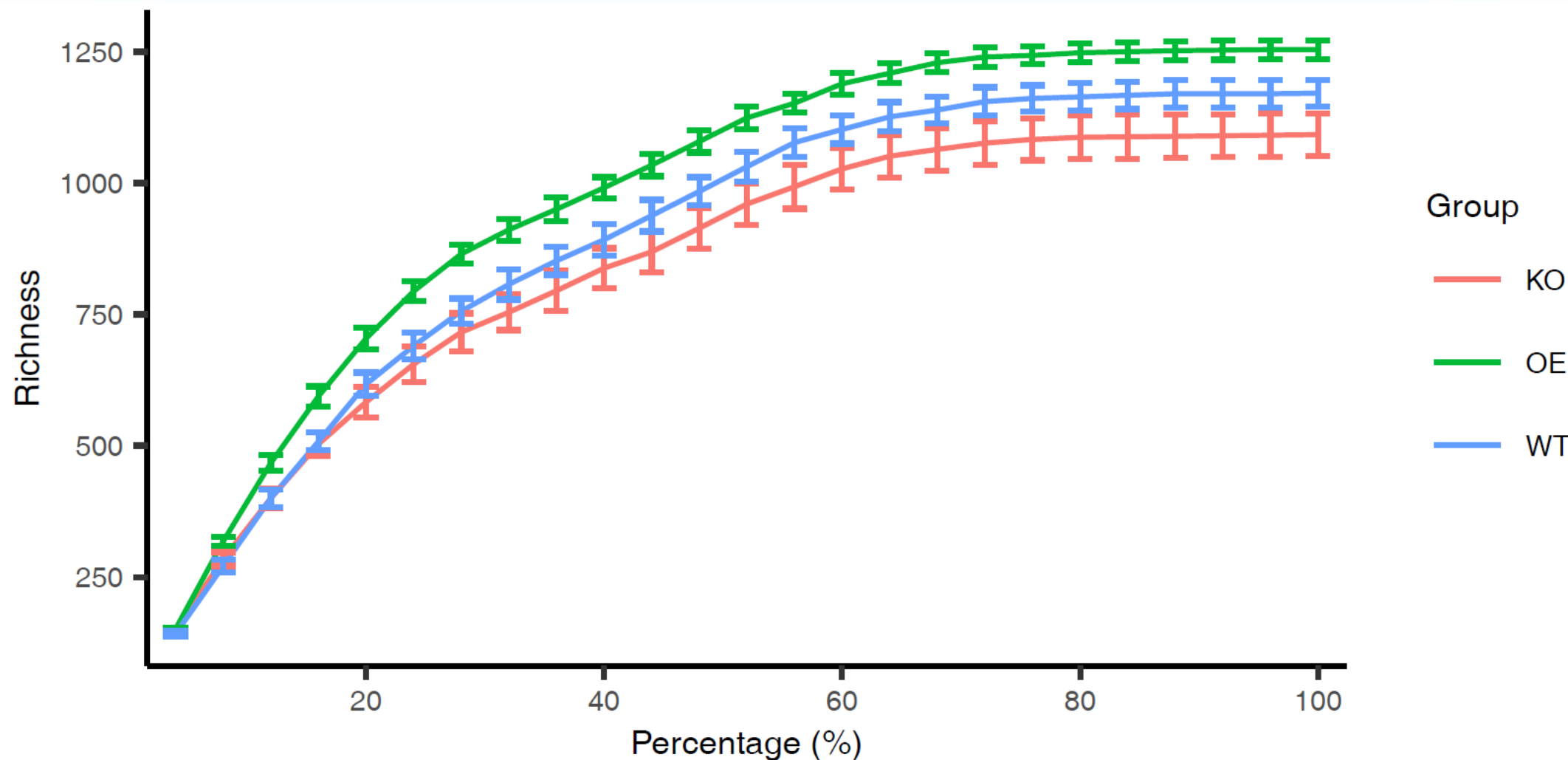
[Microbiome: 简单套路发高分文章--杨树微生物组](#)



基因组



稀释曲线展示组测序深度均值和标准误与特征(OTU/ASV)数量关系



基因组

命令行或RStudio中绘图稀释曲线

- 输入文件alpha_rare.txt, 实验设计和分组, 输出目录和图片宽高
Rscript \${db}/script/alpha_rare_curve.R \
--input result/alpha/alpha_rare.txt --design result/metadata.txt \
--group Group --output result/alpha/ \
--width 89 --height 59
- RStudio打开Diversity.Rmd文件
- 检查“# 稀释曲线”段落文件输入和输出位置
- 逐行运行, 或绿右三角运行代码块
- 结果位于alpha目录下alpha/alpha_rarefaction_curve.pdf

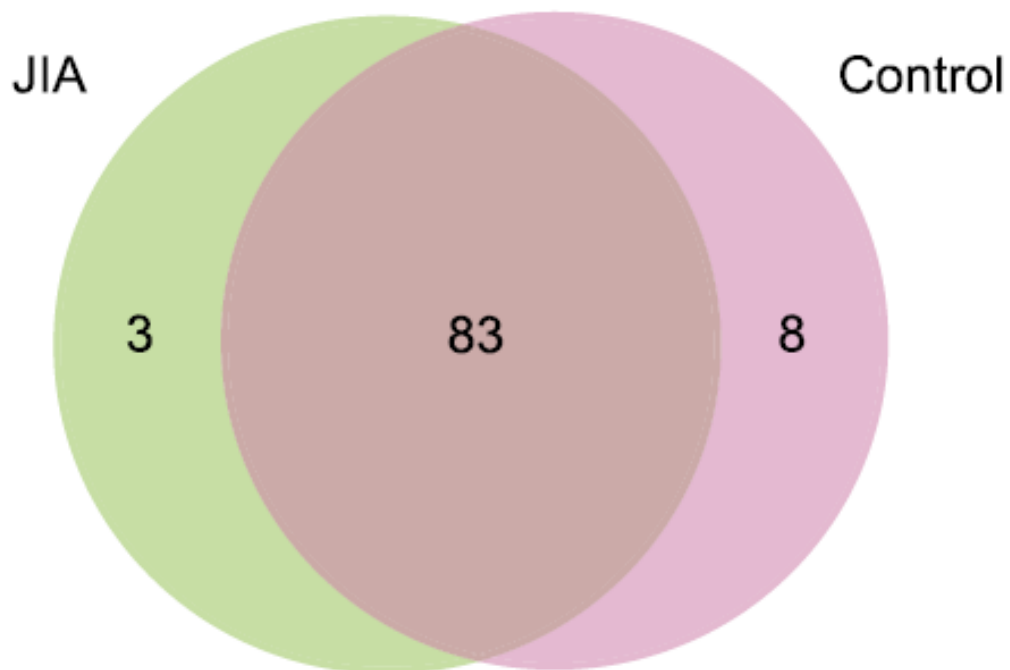


在Excel或在线网页中画折线图(选学)

- 稀释曲线数据: alpha/alpha_rare.txt
- Excel中绘制: 打开文件——全选——插入——折线图
- Excel计算组均值, 并纯文本粘贴制作一个新表格;
- 可Excel中全选插入折线图
- 在线绘制: 需要Excel中整理好数据格式, 然后可直接复制
- <http://www.bic.ac.cn/ImageGP/> —— Line plot
- Ctrl+A 全选文本框, 按 Ctrl+C复制, 再Ctrl+V 粘贴—— PLOT

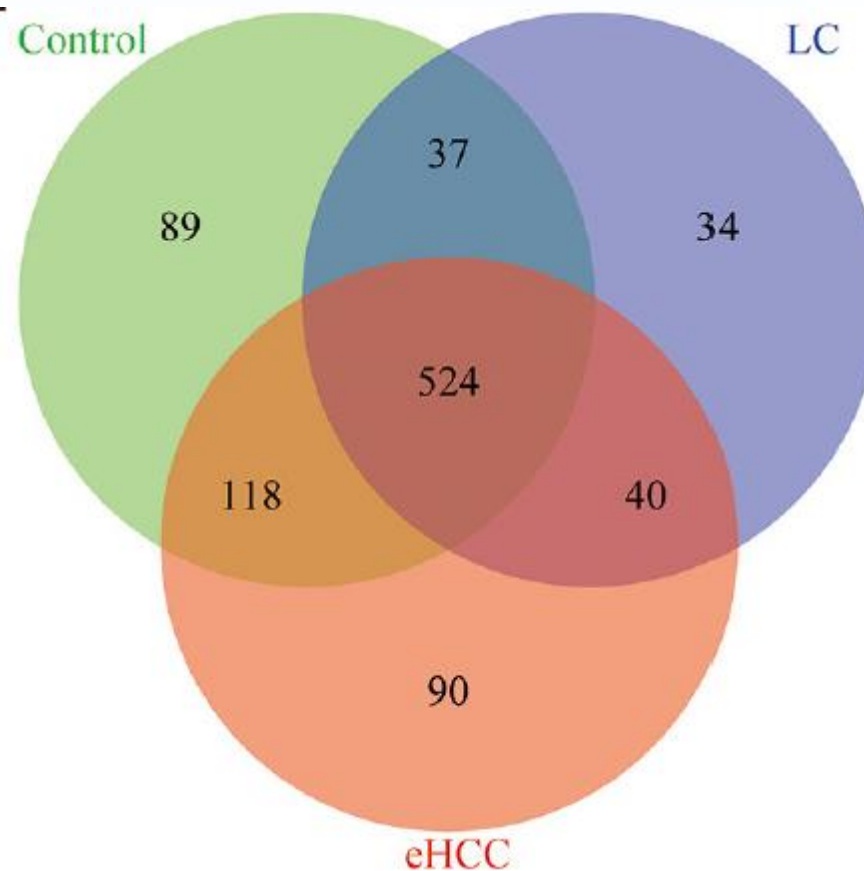


3. 多样性组间共有、特有比较维恩图



韦恩图，显示了两组有83个属是相同的，但是JIA组有3个属是独有的，对照组有8个属是独有的

[BMC: 幼儿关节炎患儿肠道菌群的特征](#)



Venn图展示了Control健康人、LC肝硬化、eHCC早期肝癌三组间独有和共有的OTU

[Gut: 早期肝癌肠道生物标志物鉴定](#)

3. 多样性组间共有、特有比较维恩图

- 输入文件otu_group_exist.txt

- 两列：分别为ID和分组

- 命令行绘制

- # -f输入文件,-a/b/c分组名,-w/u为宽高,-p输出名位置，默认同输入

```
bash ${db}/script/sp_vennDiagram.sh \
-f result/alpha/otu_group_exist.txt \
-a WT -b KO -c OE \
-w 3 -u 3 \
-p WT_KO_OE
```

ASV_657

ASV_657

ASV_657

ASV_2

ASV_2

ASV_2

KO

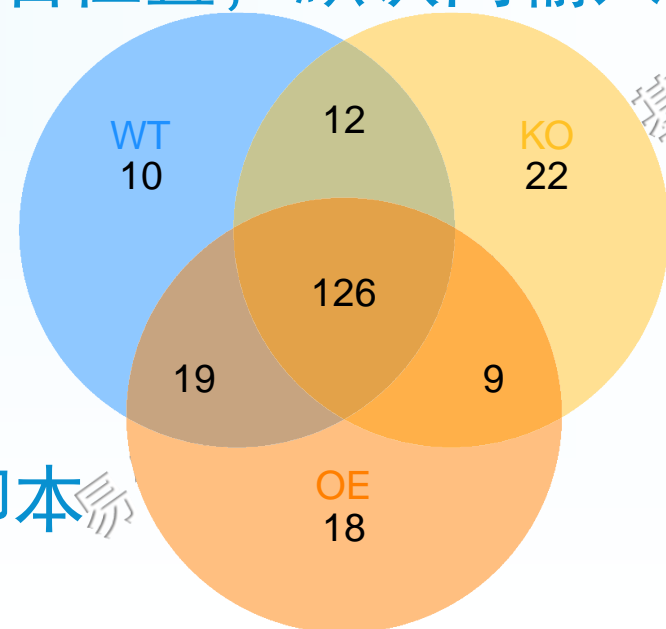
OE

WT

KO

OE

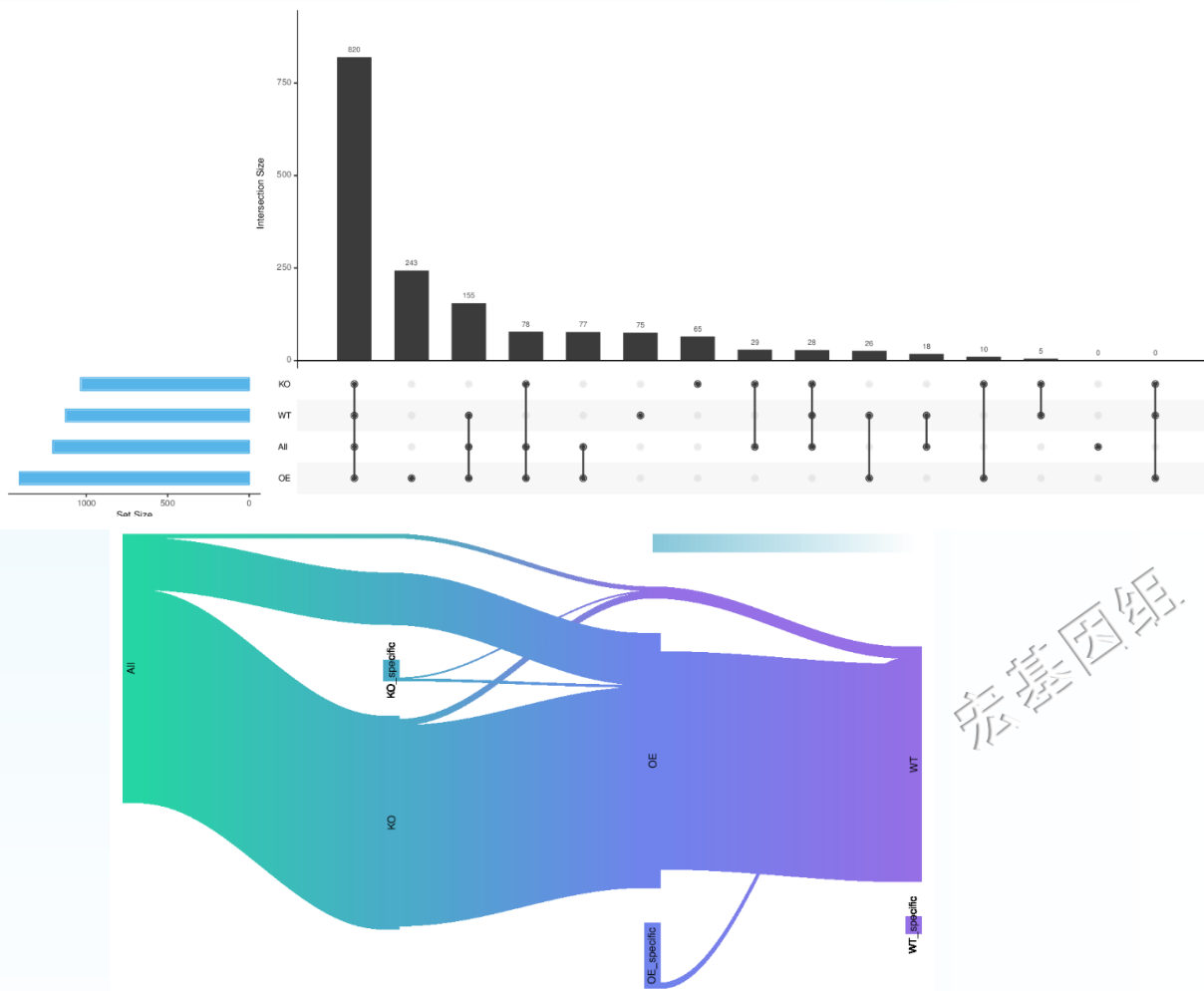
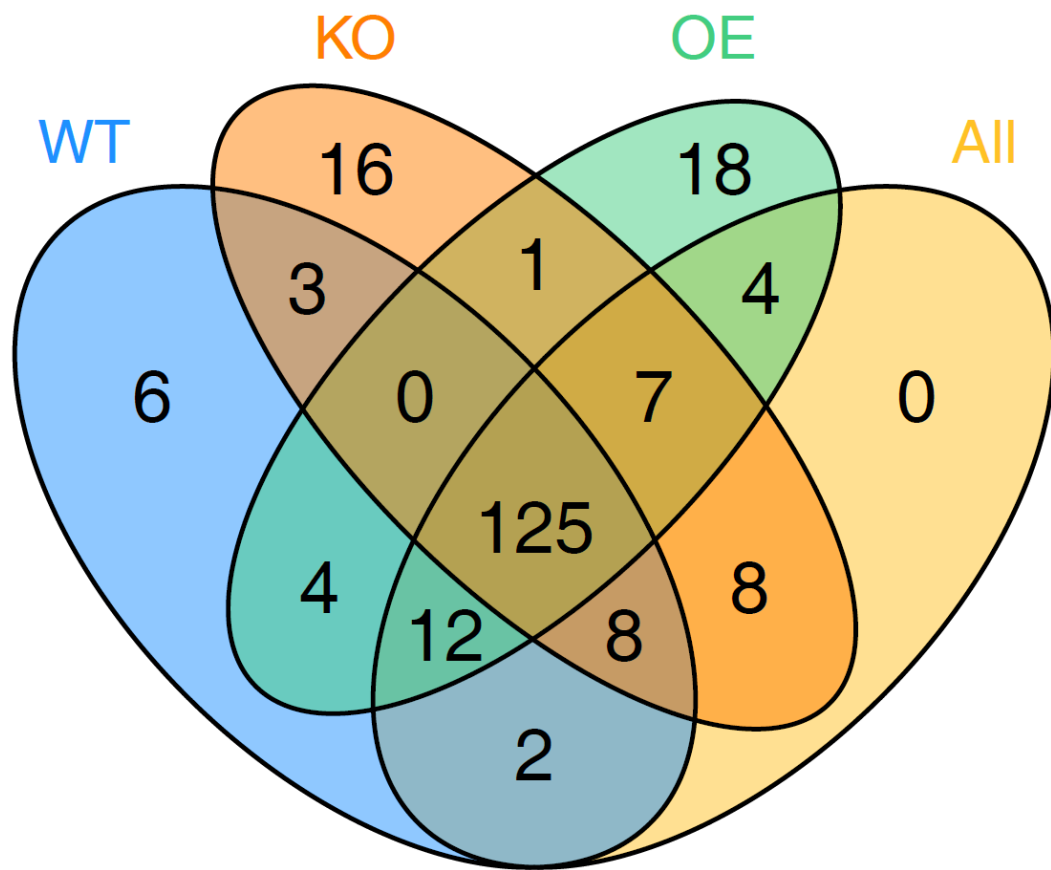
WT



- 输出：以输入文件开头的矢量PDF图片和R脚本



otu_group_exist.txt在线绘制Venn/Upset/Sanky



<http://www.bic.ac.cn/ImageGP/>

Tong Chen, Yong-Xin Liu, Luqi Huang. 2022. ImageGP: An easy-to-use data visualization web server for scientific researchers. *iMeta* 1: e5. <https://doi.org/10.1002/imt2.5>



Alpha多样性箱线图、折线图进一步学习：

- 理论讲解：
- 图表解读1箱线图：Alpha多样性
- 实战代码：
- 统计绘图1箱线图：Alpha多样性
- Alpha多样性稀释曲线rarefaction curve
- R语言学习 - 线图一步法
- R语言学习 - 箱线图（小提琴图、抖动图、区域散点图）
- R语言学习 - 箱线图一步法



- Alpha多样性——样品自身多样性
- **Beta多样性——样品/组间差异PCoA, CPCoA**
- 物种组成——不同分类级别相对丰度

○ 总结

- Beta多样性是生态学概念，专指不同组或生态位间物种组成的差异。
- 在宏基因组领域，散点图常用于展示样品组间的Beta多样性，常用的分析方法有主成分分析(PCA)，主坐标轴分析(PCoA/MDS)和限制性/有监督的主坐标轴分析(CPCoA/CCA/RDA)。
- 在阅读文章中经常可以看到PCA、PCoA、NMDS、CPCoA、CCA/RDA和LDA。它们在本质上是排序(ordination)分析。排序的过程就是在一个可视化的低维空间(通常是二维)重新排列这些样品，使得样方之间的距离最大程度地反映出平面散点图内样品间的关系信息。
- **Constrained和Unconstrained区别：分组 vs 样品**

[221.Beta多样性PCoA和NMDS排序](#)

[223.主成分分析PCA](#)

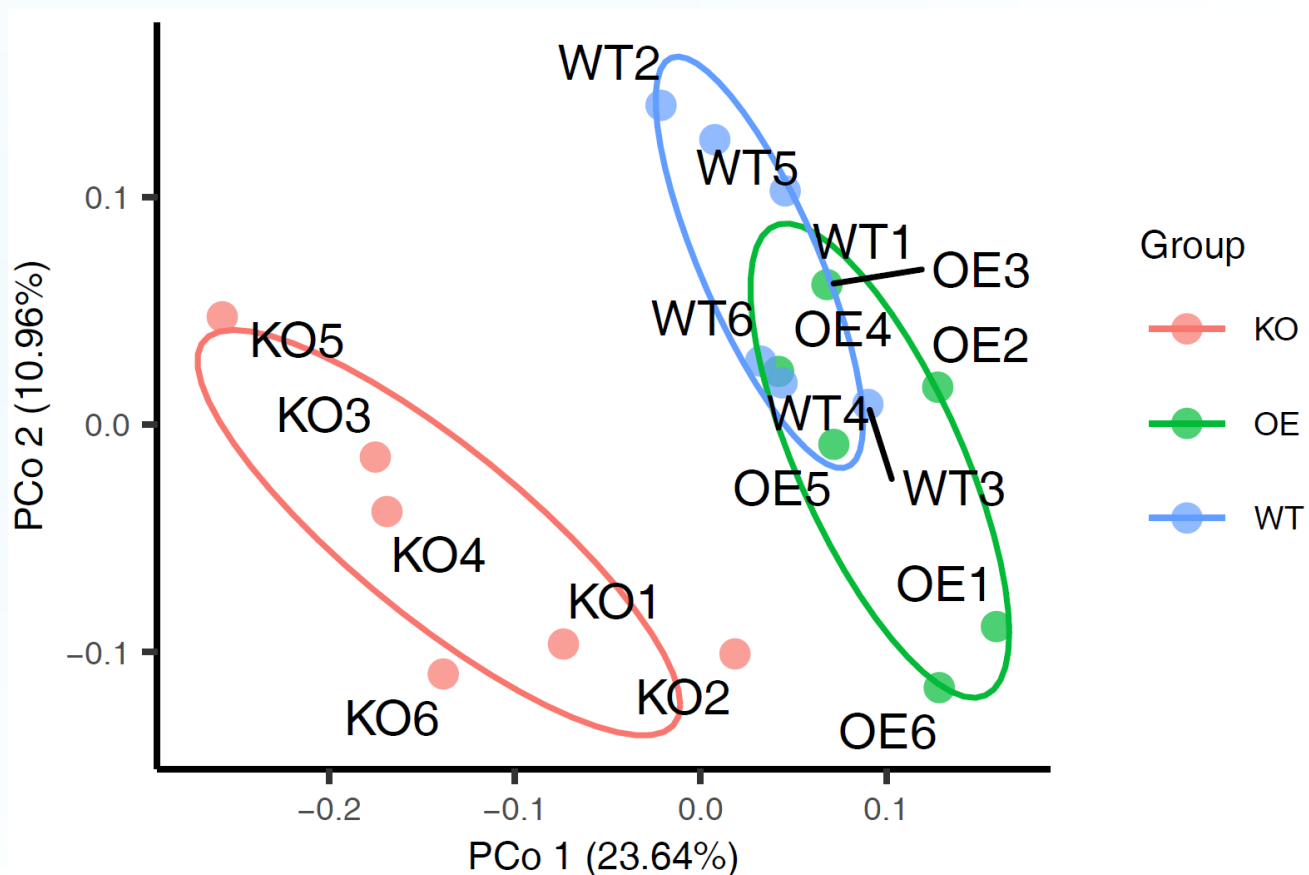
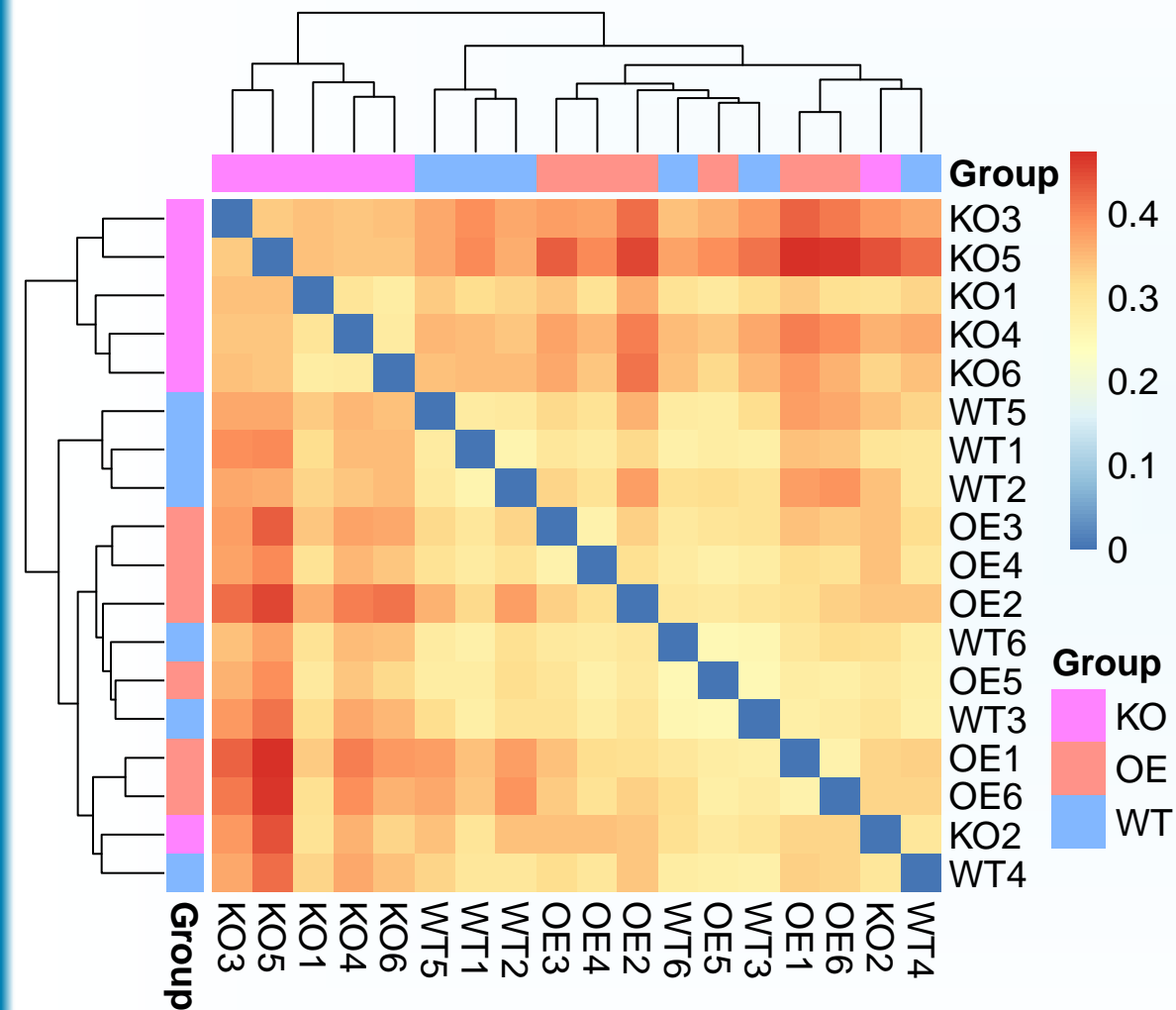


排序分析降维-举个栗子

假如你是一本养花工具宣传册的摄影师，你正在拍摄一个水壶。水壶是三维的，但是照片是二维的，为了更全面的把水壶展示给客户，你需要从不同角度拍几张图片。下图是你从水壶背面，正面，正上方，斜上方的照片：



热图或主坐标轴分析(PCoA)散点图展示样品间差异



Beta多样性距离矩阵计算(pipeline.sh ## 7. Beta)

- # PCoA计算基于距离矩阵，Unifrac距离矩阵依赖进化树

```
mkdir -p result/beta/
```

- # 基于OTU构建进化树

```
usearch -cluster_agg result/otus.fa -treeout result/otus.tree
```

- # 生成5种距离矩阵：bray_curtis, euclidean, jaccard, manhattan, unifrac

```
usearch -beta_div result/otutab_rare.txt -tree result/otus.tree \
-filename_prefix result/beta/
```

Liu et al. 2023. EasyAmplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. *iMeta* 2: e83. <https://doi.org/10.1002/imt2.83>



输入文件——距离矩阵

- 距离矩阵: result/beta/*.txt, 有7种距离矩阵, 常用bray_curtis, unifrac, unifrac_binary(unweighted_unifrac), jaccard

Bray-Curtis	KO1	KO2	KO3	KO4	KO5	KO6
KO1	0	0.271	0.306	0.256	0.323	0.244
KO2	0.271	0	0.346	0.328	0.423	0.3
KO3	0.306	0.346	0	0.303	0.319	0.307
KO4	0.256	0.328	0.303	0	0.318	0.25
KO5	0.323	0.423	0.319	0.318	0	0.315
KO6	0.244	0.3	0.307	0.25	0.315	0

- 添加分组颜色需要: 实验设计样品-组

Liu et al. 2023. EasyAmplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. *iMeta* 2: e83. <https://doi.org/10.1002/imt2.83>



1. 聚类热图+行列注释

- # 以bray_curtis为例，-f输入文件,-h是否聚类,-u/v为宽高英寸

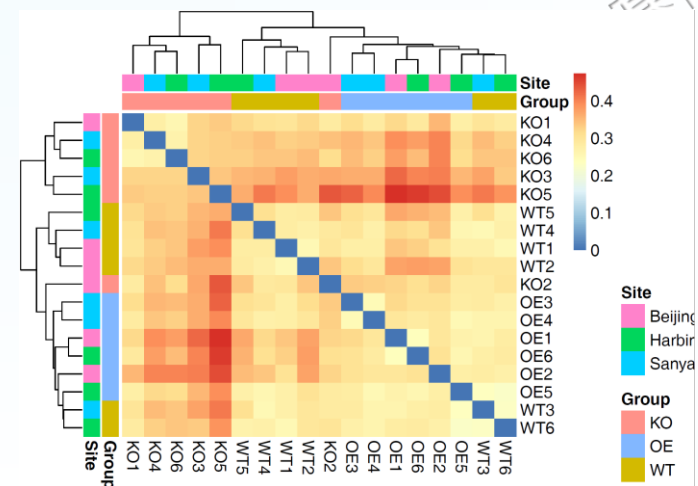
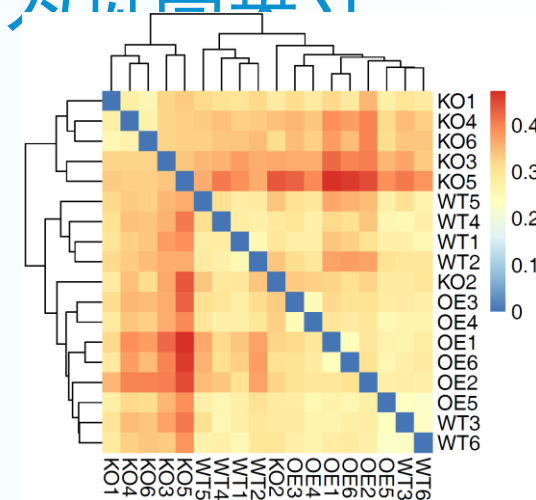
```
bash ${db}/script/sp_pheatmap.sh \
-f result/beta/bray_curtis.txt \
-H 'TRUE' -u 5 -v 5
```

- # 添加分组注释，如2，4列的基因型和地点

```
cut -f 1-2,4 result/metadata.txt > temp/group.txt
```

- # -P添加行注释文件，-Q添加列注释

```
bash ${db}/script/sp_pheatmap.sh \
-f result/beta/bray_curtis.txt \
-H 'TRUE' -u 8 -v 6 \
-P temp/group.txt -Q temp/group.txt
```



- [illegible]

- 水稻微生物组时间序列分析 2a相关分析
- 绘图相关系数矩阵corrplot 相关矩阵可视化ggcorrplot

2. PCoA绘图方法

- 命令行绘制：输入文件，选择分组，输出文件，图片尺寸

```
Rscript ${db}/script/beta_pcoa.R \  
  --input result/beta/bray_curtis.txt --design result/metadata.txt \  
  --group Group --output result/beta/bray_curtis.txt.pcoa.pdf \  
  --width 89 --height 59
```

- <http://www.bic.ac.cn/ImageGP/> —— PCoA plot

- 本地

Rstudio打开Diversity.Rmd文件

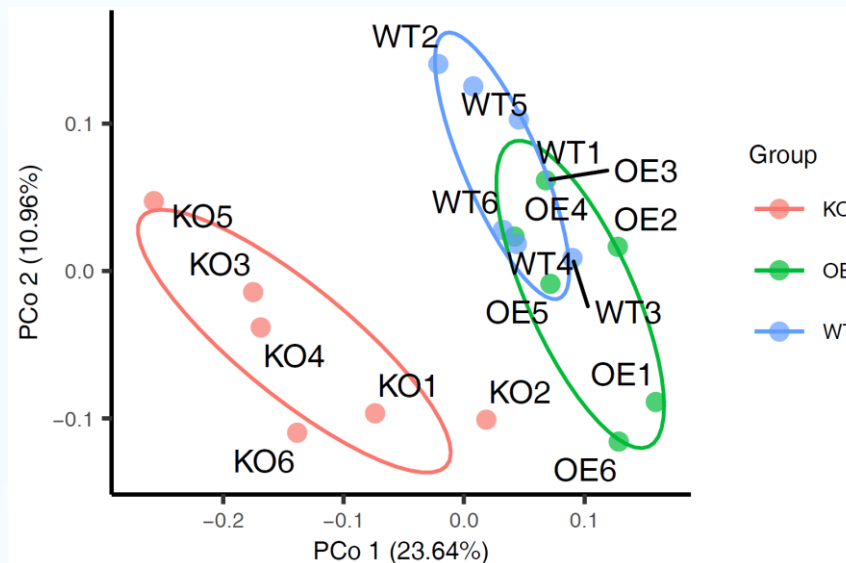
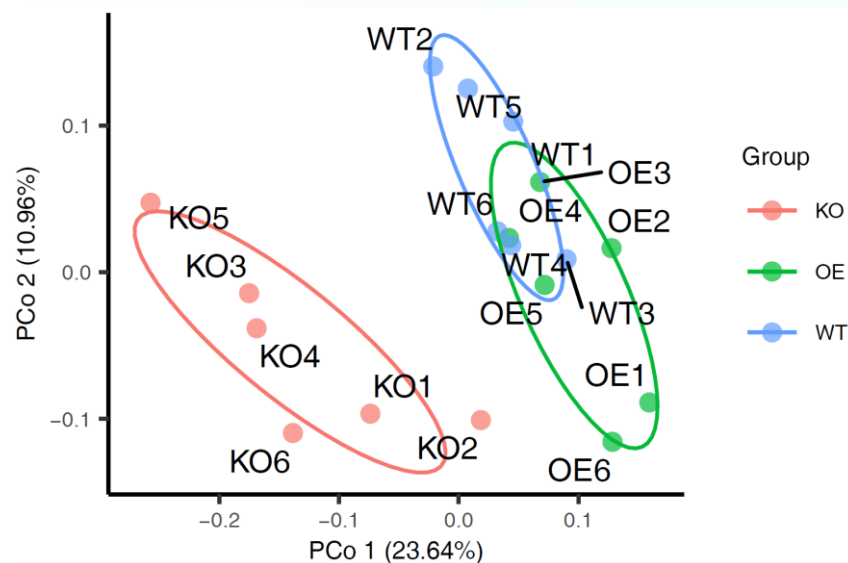
检查“# β 多样 —— ## 主坐标轴分析”段落文件输入和输出位置

Knit生成结果和计算过程网页结果

结果有 **beta/pcoa_bray_curtis.pdf**和 **beta_pcoa_stat.txt**



○ beta/pcoa_bray_curtis.pdf # 绘图结果



添加标签

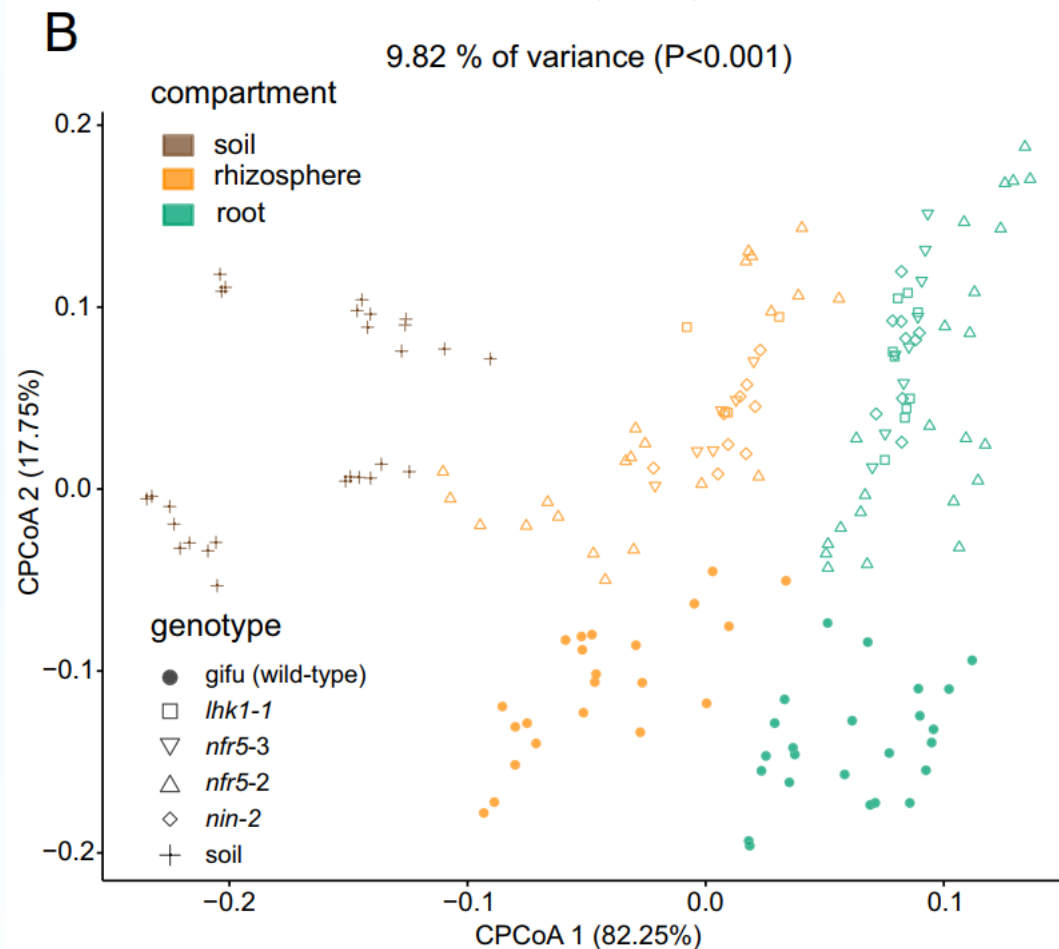
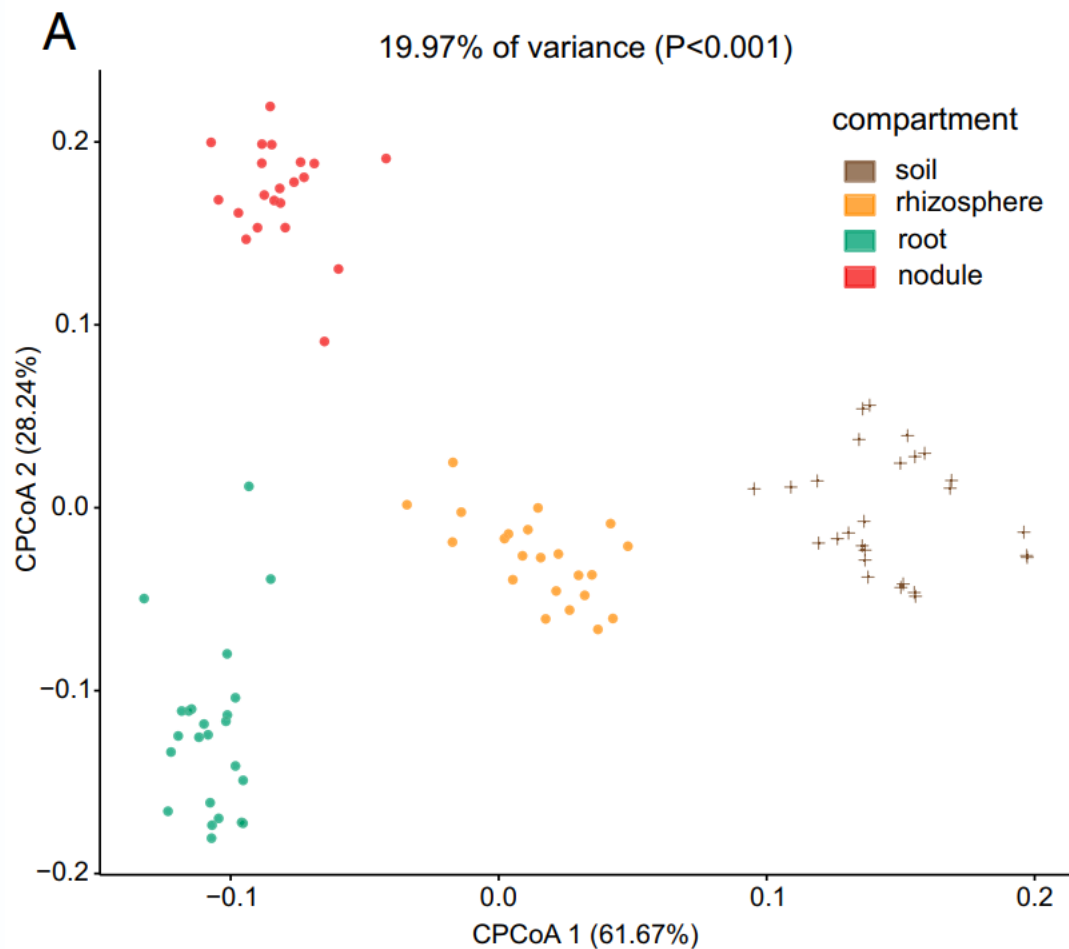
命令行 -l T 或
--label TRUE

R语言中, 添加参
数, label=T

○ beta_pcoa_stat.txt # Adonis统计组间置换检验 *P*-value

KO	OE	0.0025
KO	WT	0.0031
OE	WT	0.0052

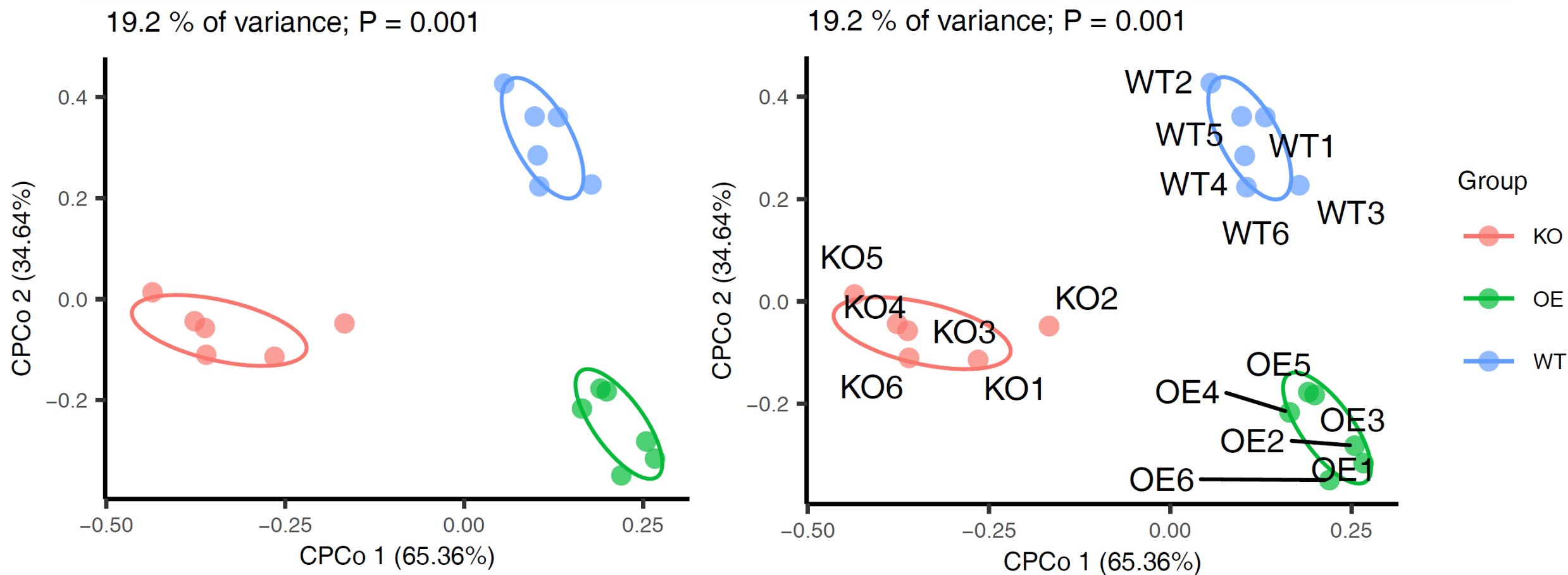
3. 限制性主坐标分析——展示间组最大差异



Zgadzaj, Rafal, et al. "Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities." *Proceedings of the National Academy of Sciences* 113.49 (2016): E7996-E8005.

图表解读2散点图: Beta多样性, PCoA, CPCoA

限制性主坐标轴分析(CPCoA)展示样品间差异



```
Rscript ${db}/script/beta_cpcoa.R \  
  --input result/beta/bray_curtis.txt --design result/metadata.txt \  
  --group Group --output result/beta/bray_curtis.txt.cpcoa.pdf \  
  --width 89 --height 59
```

- 打开Diversity.Rmd —— 阅读“# β 多样 — ## 有监督PCoA(CPCoA)”部分并修改文件位置 —— Knit 运行 —— 结果为网页和beta/cpcoa_bray_curtis.pdf
- 本地版支持距离矩阵，可绘制bray_curtis、unifrac等距离
- <http://www.bic.ac.cn/ImageGP/> ——CPCoA plot
- 在线版支持非进化的十多种距离，但无法计算unifrac相关距离



方差分解PERMANOVA(Diversity.Rmd)

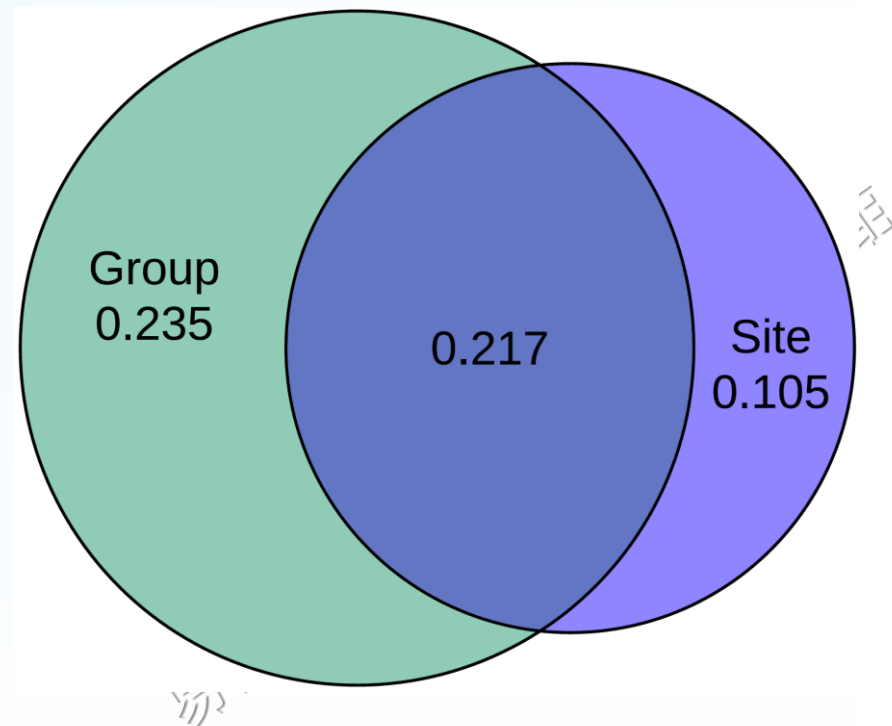
```
# 计算元数据中样本分组(Group)和取样地点(Site)对群落结构的变异解释率和显著性
adonis_var <- adonis (as.dist(distance_mat) ~ Group*Site, data = metadata, by=NULL, parallel=4)
# 预览结果
adonis_var$aov.tab
# 保存表格
write.table(adonis_var$aov.tab, file="beta/adonis_var.txt", append = F, sep="\t", quote=F,
row.names=T, col.names=T)
```

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Group	2	0.235	0.117	2.678	0.247	0.001
Site	2	0.105	0.053	1.202	0.111	0.176
Group:Site	4	0.217	0.054	1.239	0.228	0.085
Residuals	9	0.394	0.044	NA	0.414	NA
Total	17	0.951	NA	NA	1.000	NA

整理数据为右侧格式，在线EVENN
用Euler diagram – intersection
count 可视为右图

<http://www.ehbio.com/test/venn/#/>

Intersection	Count
Group	0.235
Site	0.105
Group&Site	0.217



Beta多样性进一步学习:

- 理论讲解：
- 图表解读2散点图：Beta多样性, PCoA, CPCoA
- 排序方法比较大全PCA、PCoA、NMDS、CPCoA
- 实战代码：
- 统计绘图2散点图：Beta多样性, PCoA, CPCoA
- LDA分析、作图及添加置信-ggord
- R语言学习 - 散点图绘制
- 一文看懂PCA主成分分析

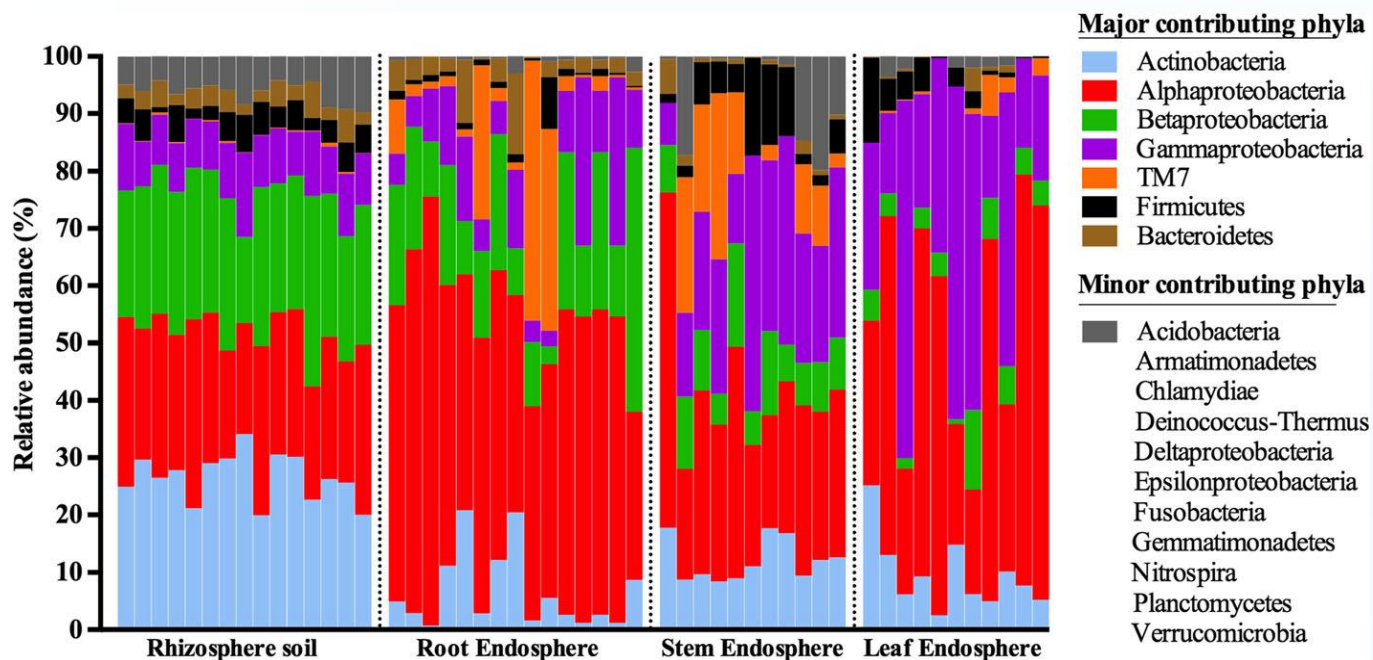


- Alpha多样性——样品自身多样性
- Beta多样性——样品/组间差异PCoA, CPCoA
- 物种组成——不同分类级别相对丰度
- 总结

易生信 生信宝典 宏基因组



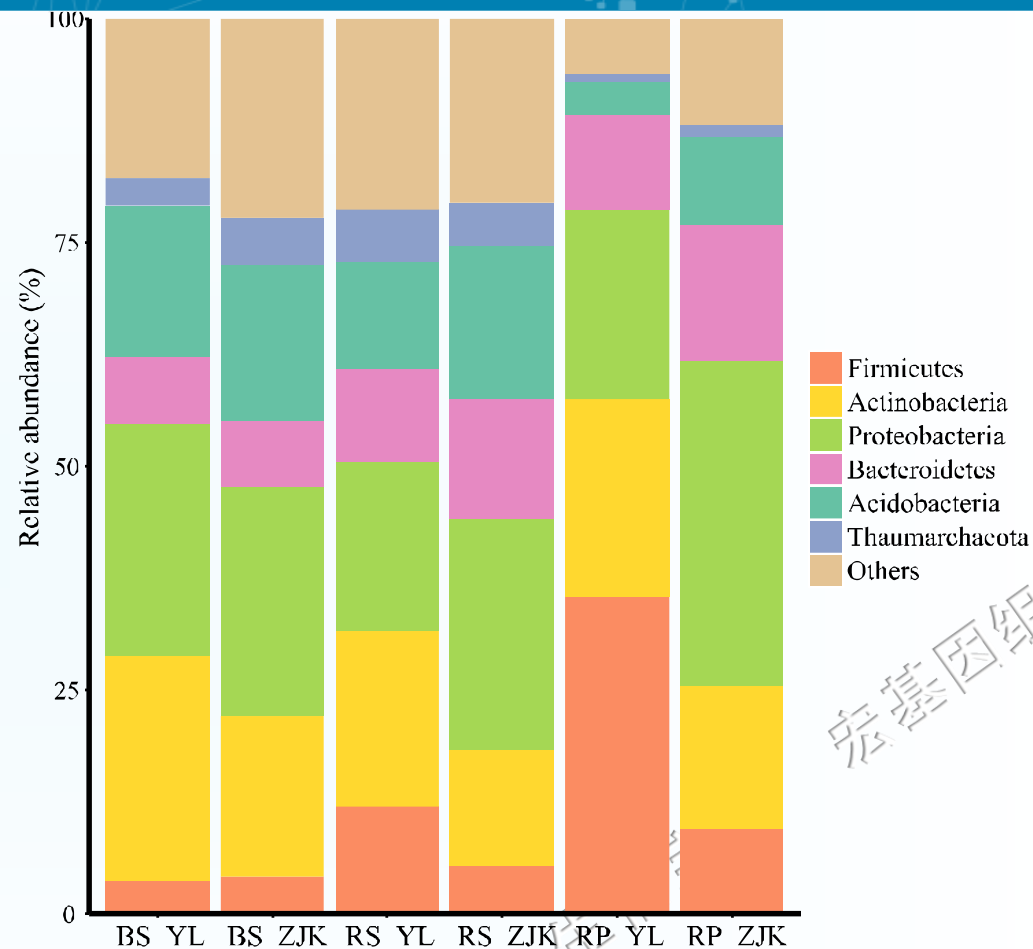
物种组成堆叠柱状图举例



Microbiome图4. 各样品OTU在门水平分类的相对丰度柱状图

图中展示了每个Compartment的每个样品门水平的相对丰度；其中Proteobacteria由于组比较大，也将其分成了apha, beta, gamma, delta, epsilon五类展示；高丰度的前7类用彩色显示，其它低丰度的门用灰色显示。

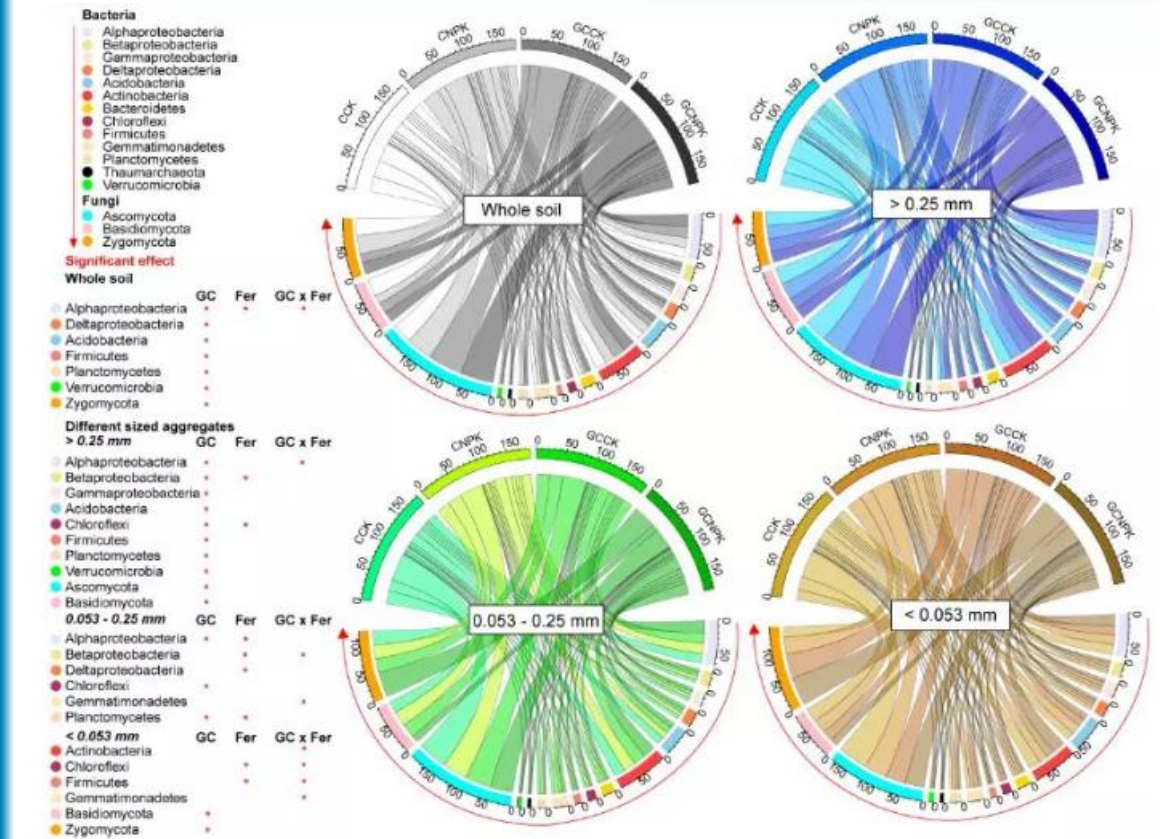
[Microbiome: 简单套路发高分文章--杨树微生物组](#)



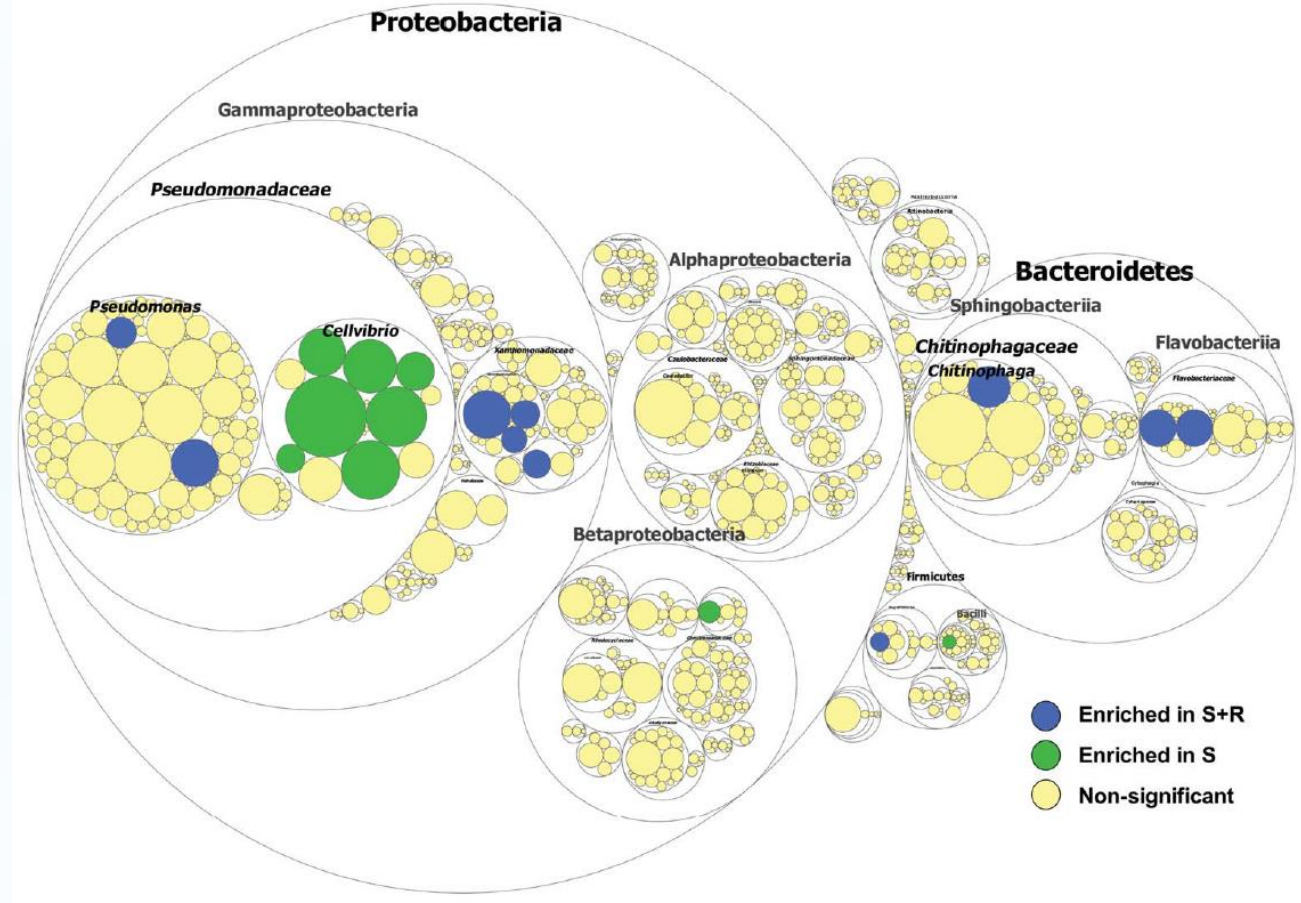
GigaScience图1. 谷子根和土壤细菌主要菌门。杨凌(YL)、张家口(ZJK)、非根际土(BS)、根际(RS)、根表(RP)

[GigaScience: 谷子产量与微生物组关联分析](#)

弦图和树图(TreeMap)



R包circlize: 物种组成可视为弦状图



Science: 病原菌激活植物内生菌群的抑病功能

- 标准化的分类级：tax/sum_*.txt

Phylum	KO1	KO2	KO3	KO4	KO5	KO6	OE1
Proteobacteria	62.1	46.4	57.9	59.7	71.3	64.1	50.6
Actinobacteria	24.8	39.2	27.2	27.3	16.7	22.7	26
Unassigned	6.92	7.55	4.01	5.44	4.34	7.18	7.65
Firmicutes	1.64	2.05	1.55	3.25	0.775	1.74	5
Bacteroidetes	3	2.73	7.7	3.07	5.23	2.42	3.8
Acidobacteria	0.365	0.334	0.246	0.227	0.192	0.357	0.998
Verrucomicrobia	0.203	0.218	0.151	0.111	0.117	0.192	0.534
Planctomycetes	0.0349	0.268	0.161	0.0284	0.063	0.0757	0.458

- 实验设计：metadata.txt

1. 堆叠柱状图绘图方法：命令行/Rmd

- 以门(p)水平为例，结果包括output.sample/group.pdf两个文件

```
Rscript ${db}/script/tax_stackplot.R \  
  --input result/tax/sum_p.txt --design result/metadata.txt \  
  --group Group --output result/tax/sum_p.stackplot \  
  --legend 5 --width 89 --height 59
```

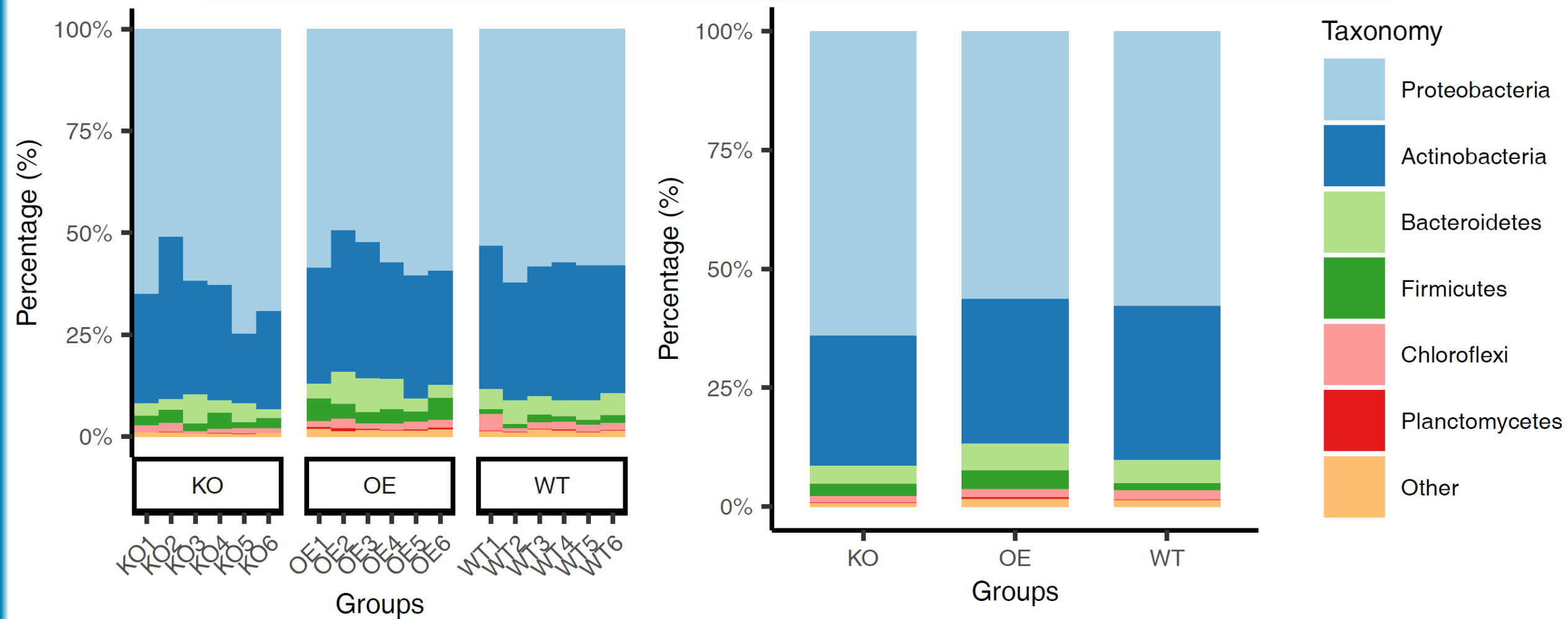
- 本地绘制

- 打开Diversity.Rmd
- 检查“# 物种组成 Taxonomy stackplot”段落参数，调整分类级、输入输出文件位置和图片大小等
- Knit生成结果和计算过程网页

- 可选在线绘制 <http://www.bic.ac.cn/ImageGP/> —— Bar plot



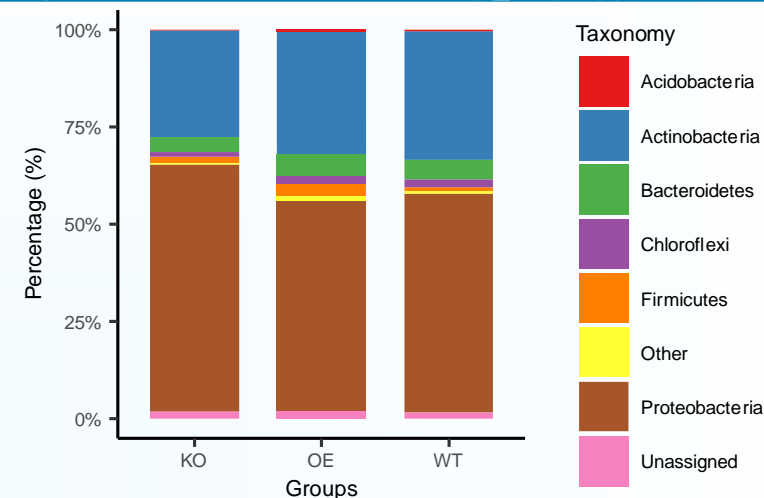
堆叠柱状图展示各样品/组主要门水平相对丰度



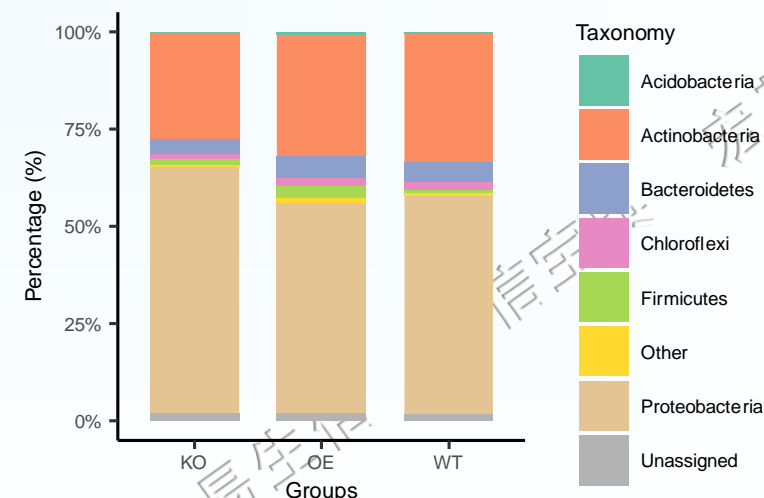
RColorBrewer修改本色方案



```
library(RColorBrewer)
display.brewer.all()
```



p + scale_fill_brewer(palette = "Set1")



p + scale_fill_brewer(palette = "Set2")

2.弦图(圈图)circlize

- 以纲(c)水平为例，结果包括circlize*.pdf两个文件

```
Rscript ${db}/script/tax_circlize.R \  
  --input result/tax/sum_c.txt --design result/metadata.txt \  
  --group Group --legend 5
```

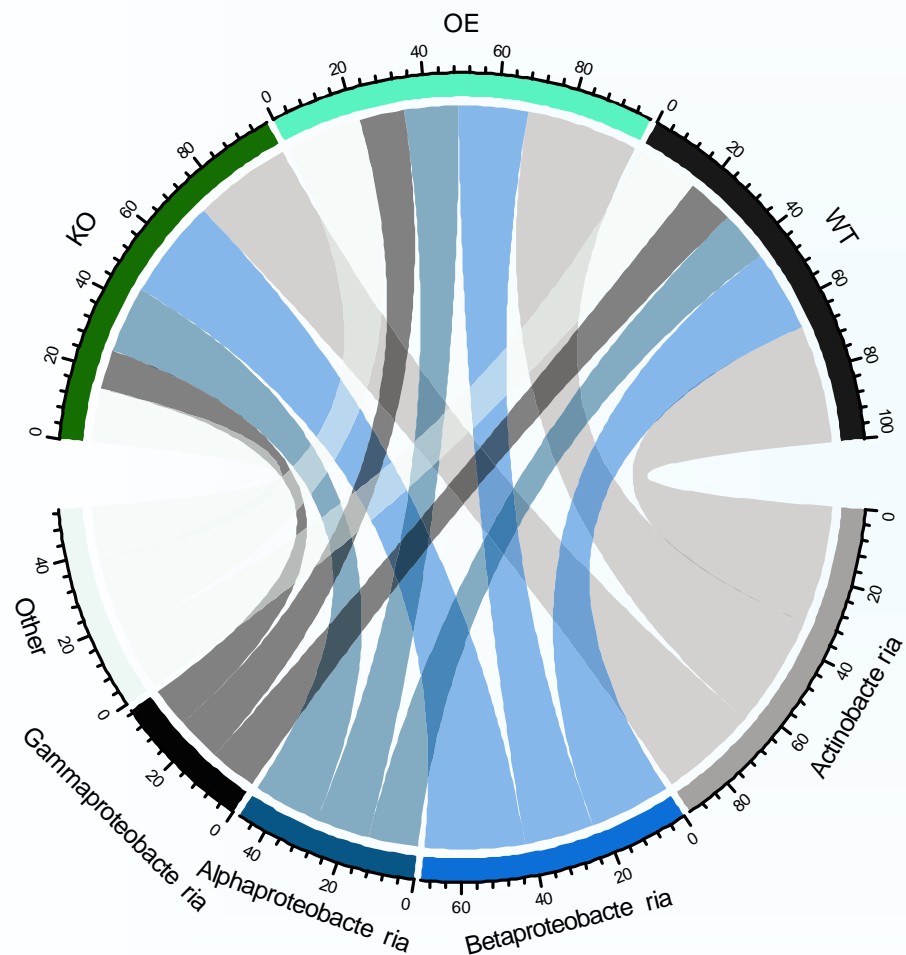
- 本地绘制

- 打开Diversity.Rmd
- 检查“## 弦图circlize plot”段落参数，调整分类级、图例数量和分组列名
- 右键头运行段落，在代码文件内查看结果circlize.pdf和circlize_legend.pdf

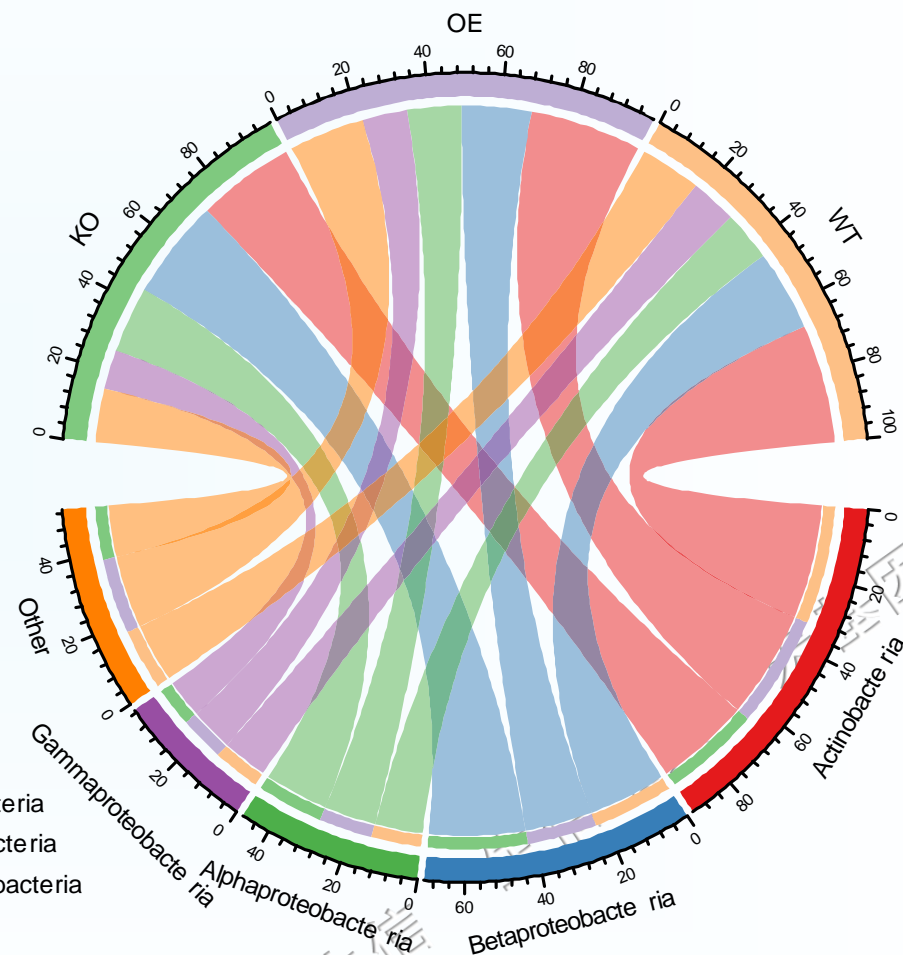
- 可选在线绘制 <http://www.bic.ac.cn/ImageGP/> ——Bar plot



弦图的两默认输出：随机颜色和固定颜色+图例



- KO
- OE
- WT



- Actinobacteria
- Betaproteobacteria
- Alphaproteobacteria
- Gammaproteobacteria
- Other

- Alpha多样性——样品自身多样性
- Beta多样性——样品/组间差异PCoA, CPCoA
- 物种组成——不同分类级别相对丰度
- **总结**

易生信 生信宝典 宏基因组



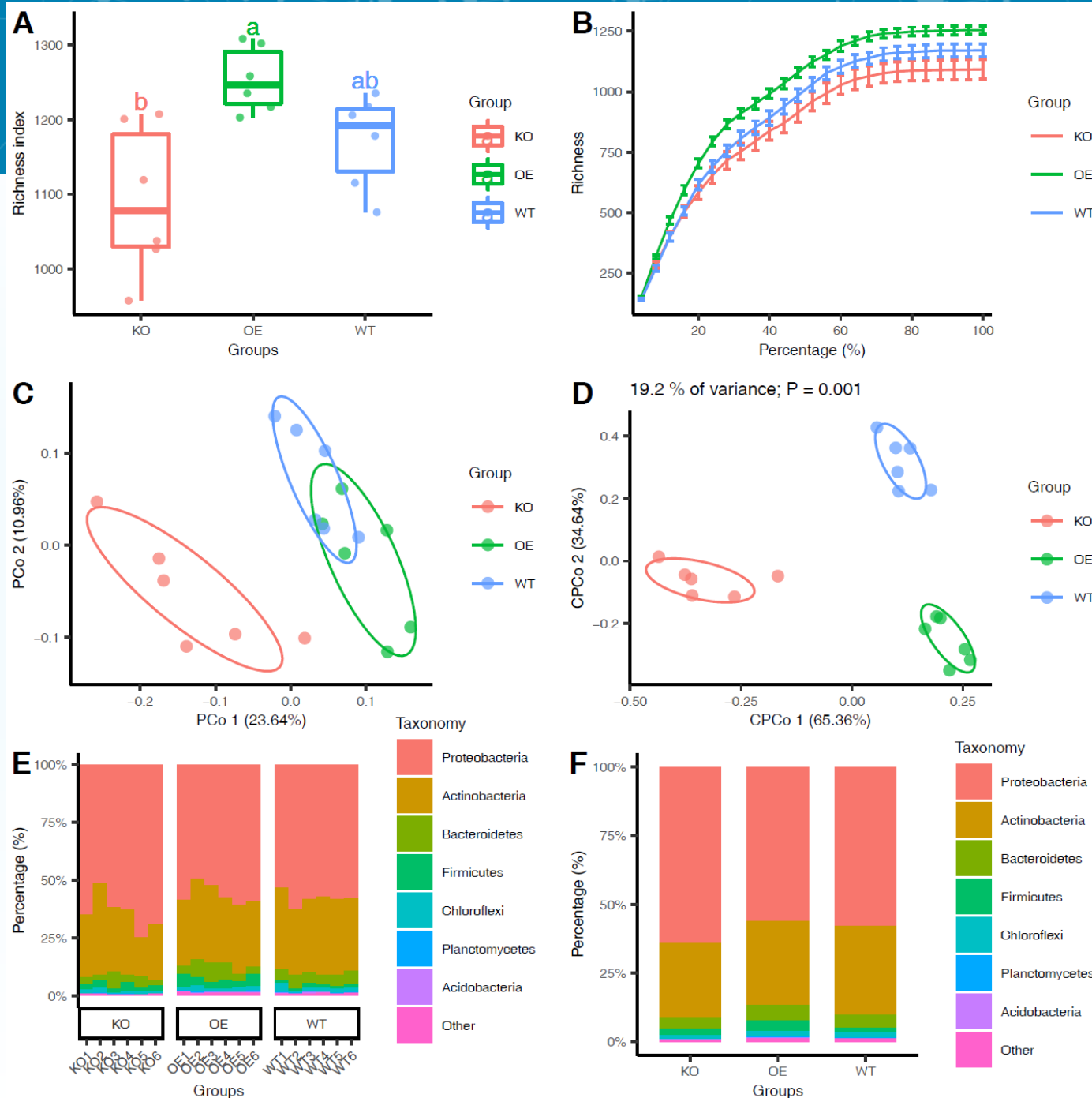


常见多样性展示方法

- Alpha多样性: 箱线图+统计分组、折线图、折线图+误差棒、维恩图及衍生图(包括维恩图Venn diagram, 集合图UpSetView和桑基图Sanky diagram)
- Beta多样性: PCoA散点图+置信区间+Adonis统计、限制性PCoA散点图+置信区间+anova.cca统计、热图或聚类图
- 物种组成: 堆叠状图 样本 vs 组, 控制图例数据6-8为宜, 尽量不要超过10类, 可选弦图
- 想想你还见到过、或喜欢使用的展示样式? 使用Excel或ImageGP还可画那些种类的图?

cowplot排版

- library(cowplot)
- (p0 = plot_grid(p1, p2, p3, p4, p5, p6, labels = c("A", "B", "C", "D", "E", "F"), ncol = 2))
- ggsave("diversity.pdf", p0, width = width * 2, height = height * 3, units = "mm")



进一步阅读

- [宏基因组公众号文章目录](#) [生信宝典公众号文章目录](#)
- [科学出版社《微生物组数据分析与可视化实战》——30+篇](#)
- [Bio-protocol《微生物组实验手册》计划——200+篇](#)
- [Protein Cell: 扩增子和宏基因组数据分析实用指南](#)
- [CMJ: 人类微生物组研究设计、样本采集和生物信息分析指南](#)
- [扩增子图表解读 分析流程 统计绘图](#)
- [QIIME2中文教程-把握分析趋势](#)
- [扩增子16S分析专题研讨讨论会——背景介绍](#)





扫码关注生信宝典，学习更多生信知识



扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识

