213维恩图和集合图

徐俊 北京大学人民医院

2020-07-07

Table of Contents

### 维恩图和集合图

本节作者：徐俊，北京大学人民医院

版本1.0.5，更新日期：2020年6月30日

本项目永久地址： <https://github.com/YongxinLiu/MicrobiomeStatPlot> ，本节目录 213Venn，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。

#### 背景介绍

在微生物数据分析过程中，经常需要对某几组样本中共有或特有的OTU或微生物进行可视化。基于此需求，通常可以选择维恩图(Venn diagram，或韦恩图)等进行可视化。然而，当分组信息过多，维恩图的展示能力及可读性则有所下降，因此推荐使用维恩图的升级版本——集合图(Upset plot)。

在本教程中，我将从菌群特征表(OTU/ASV)开始，无需在代码外整理数据，从而直接实现用维恩图或集合图对集合信息进行可视化。可视化过程使用VennDiagram和UpsetR等R包本地实现可重复分析。

除了以上两个R包以外，还有ImageGP、yyplot、VennPainter、VennMaster以及TBtools的WonderfulVenn等软件或网站可供选择。详见：<https://cloud.tencent.com/developer/article/1423035>

此外，需要指出的是，本文撰写过程中涉及到的描述和代码参考了诸多生信和可视化方面专家的前期工作，不具备开创性特色，不过是搜集整理并让代码适用于直接对微生物分析中的OTU表格的分析。可以说本文的撰写，确实是站在巨人的肩上。文尾将对参考内容进行整理，此处不一一致谢了。

其他：本文所有示例数据和更新代码均已经上传Github：<https://github.com/JerryHnuPKUPH/OTU2V-UpPlot> 。

##### 维恩图简介

维恩图（英语：Venn diagram），又称韦恩图。是十九世纪英国数学家约翰·维恩（John Venn）发明，用于展示集合之间大致关系的一类图形。其中圈或椭圆重合（overlap）的部分就是集合与集合间元素的交集，非重叠部分则为特定集合的特有元素。

##### 维恩图和集合图对比

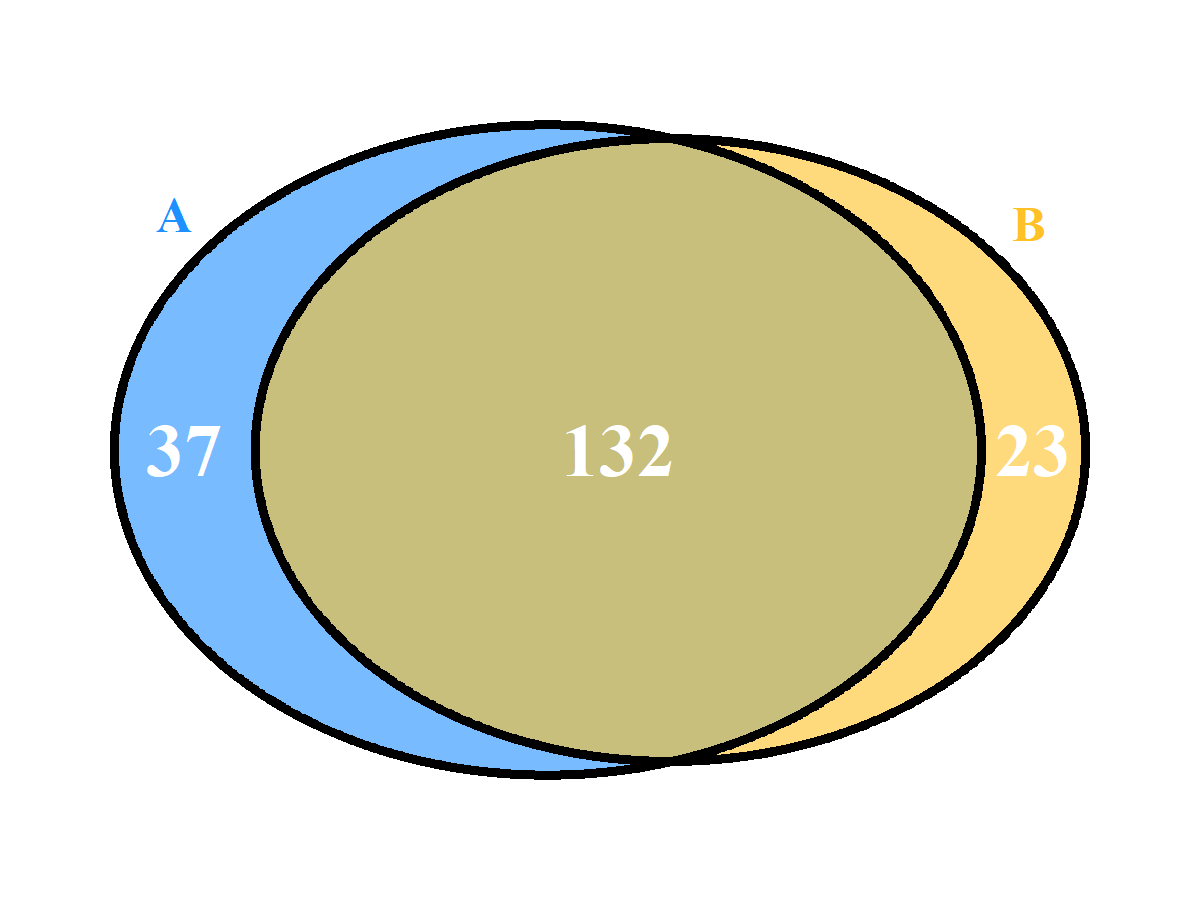
维恩图和集合图均可用于对集合共有和特有元素信息进行可视化，但是当数据分组过多（>4）时，维恩图看起来会非常杂乱，而集合图可以展示≥5个分组的集合元素共有和特有信息。

**总结起来：**

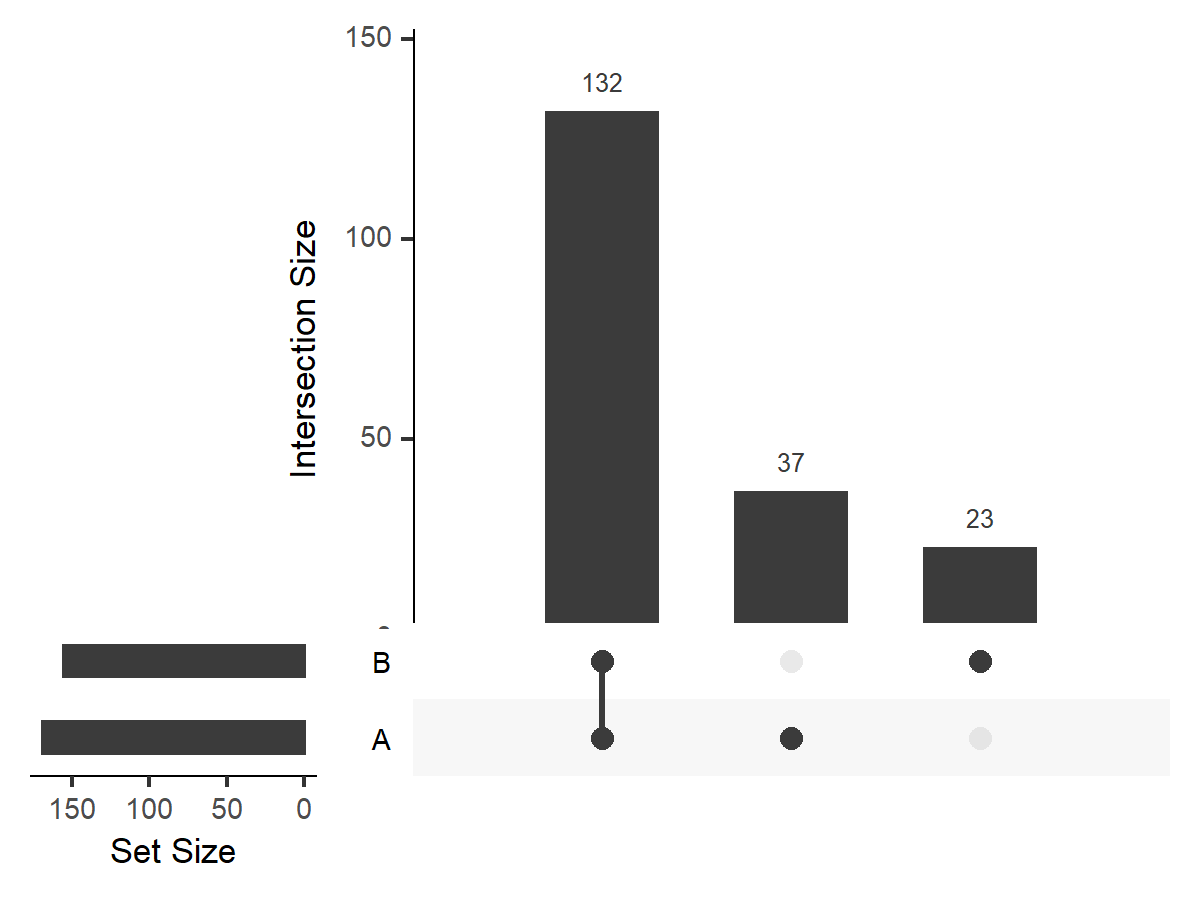
1）分组<5，维恩图更清晰；

2）分组≥5，集合图更清晰；

3）集合图展示方式更多元。

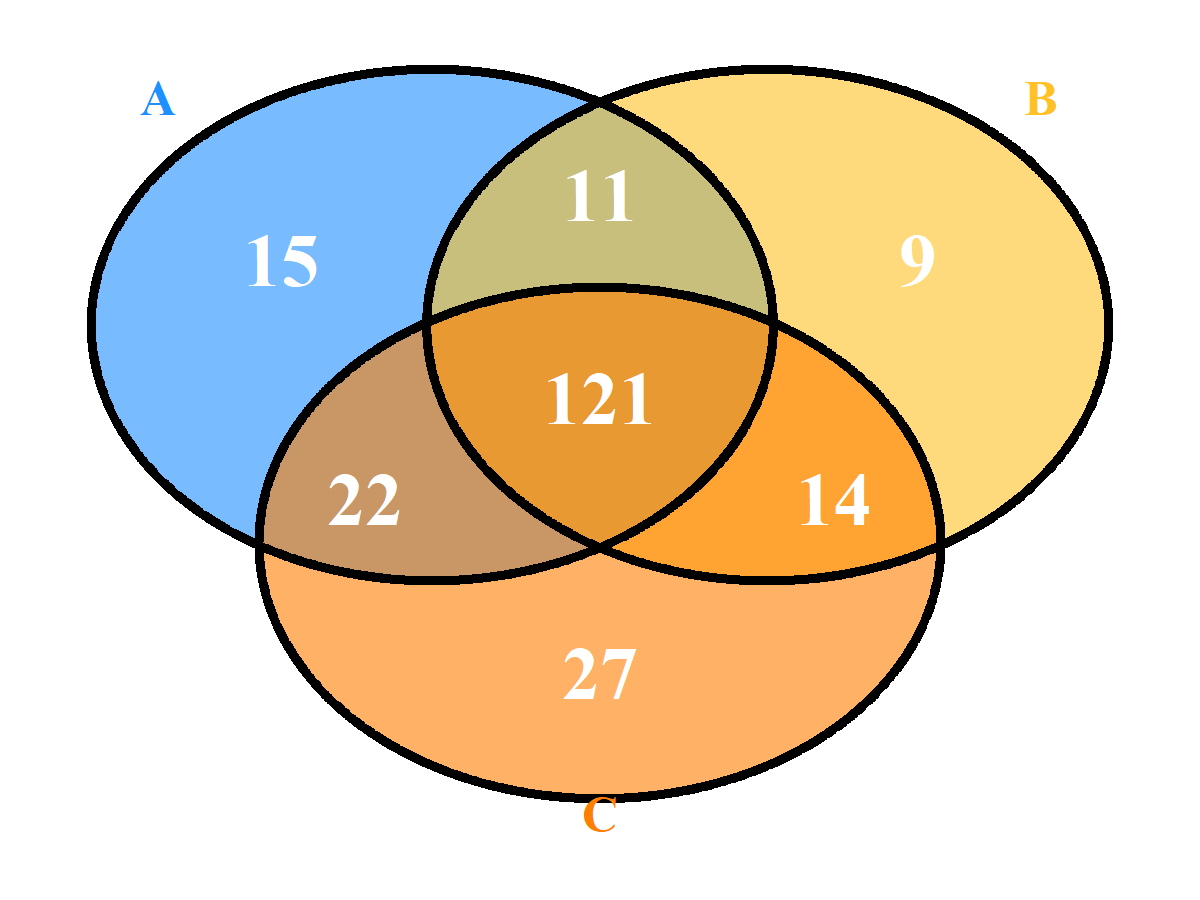


图b1. 维恩图使用示例，展示A/B两组间共有和特有元素的数量

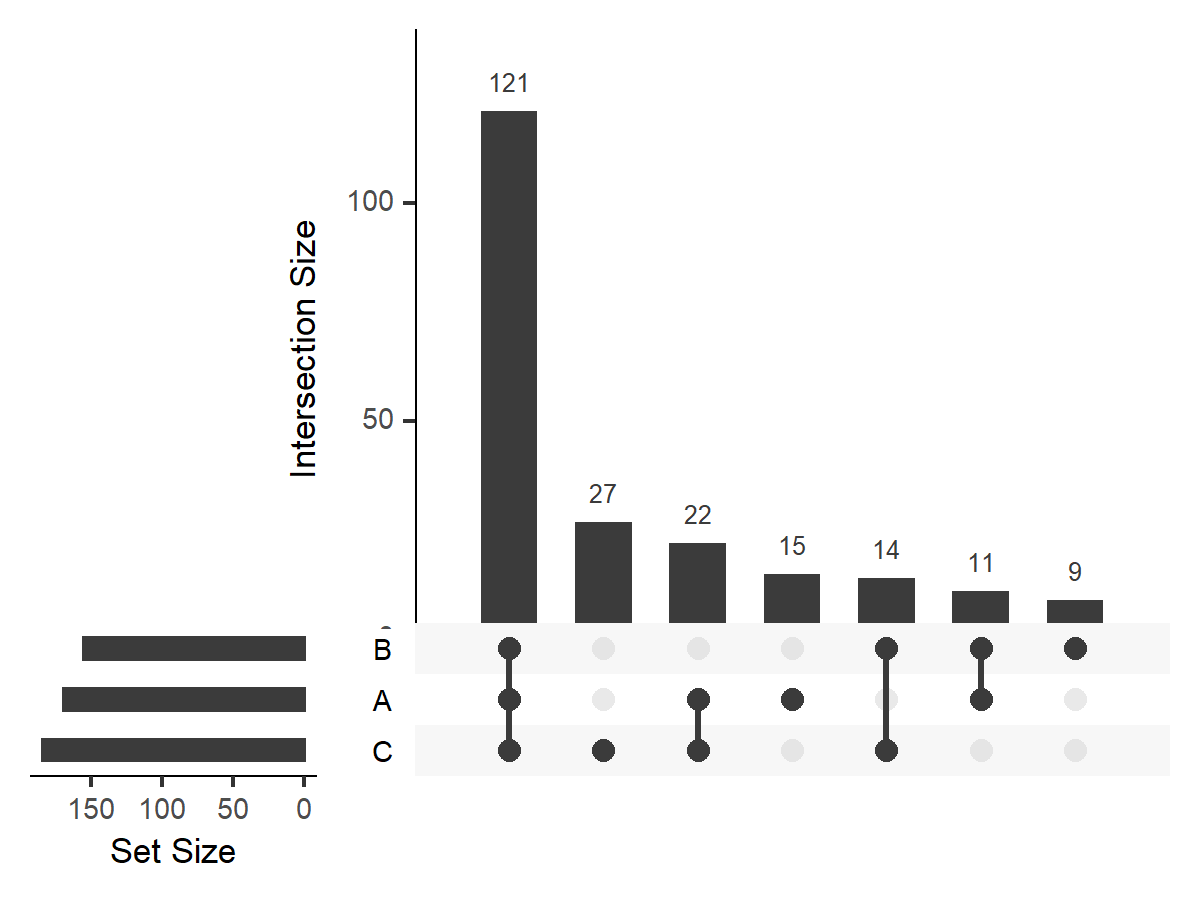


图b2. 集合图使用示例，展示A/B两组间共有和特有元素的数量

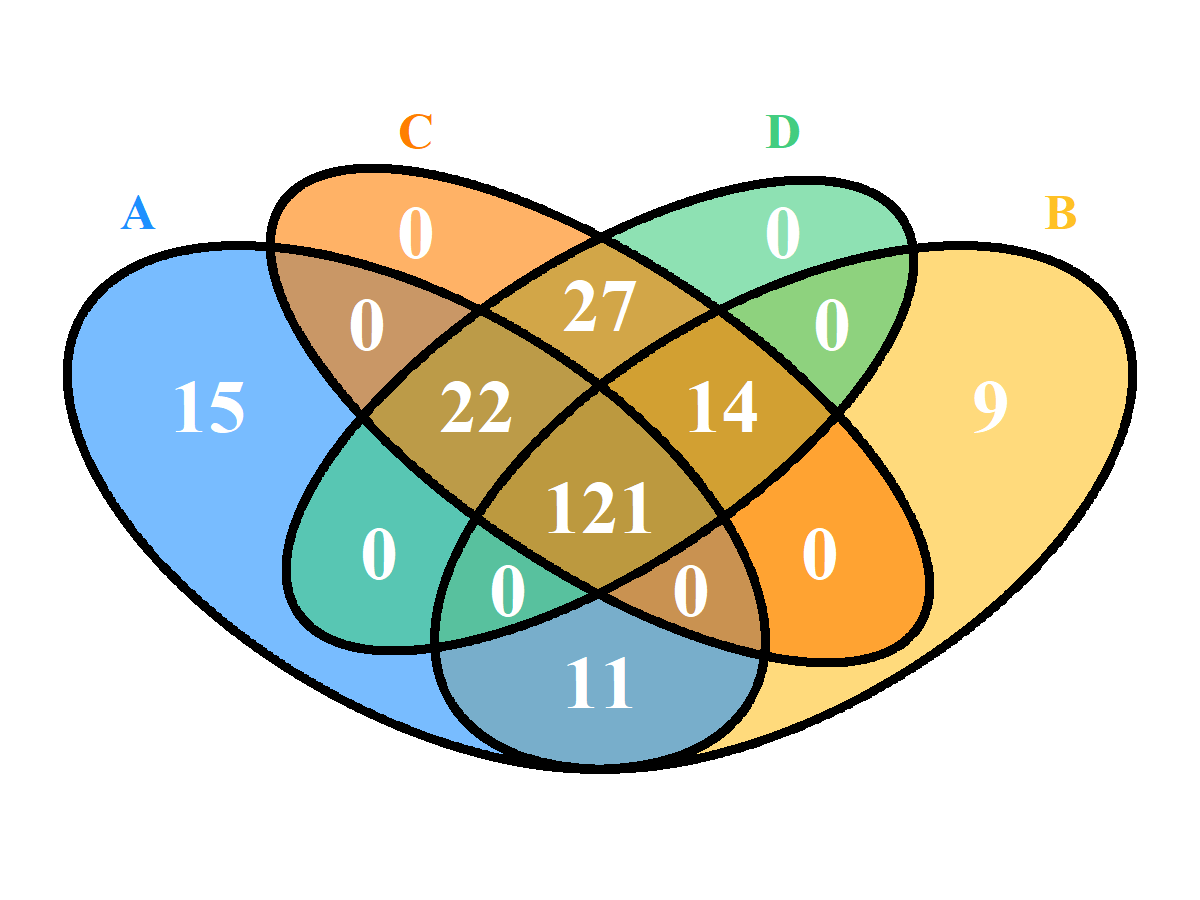
***集合较少时，集合图略显单调。***



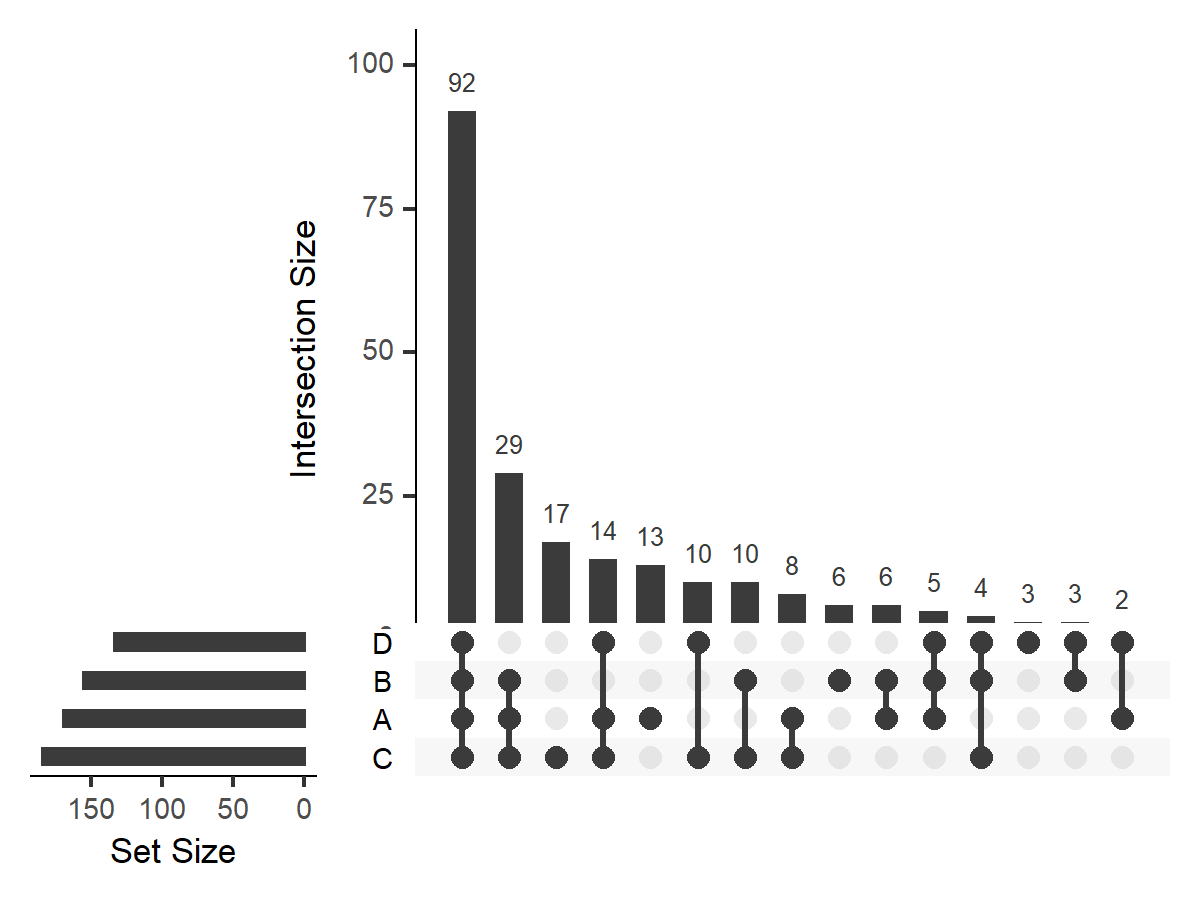
图b3. 维恩图使用示例，展示A/B/C三组间共有和特有元素的数量



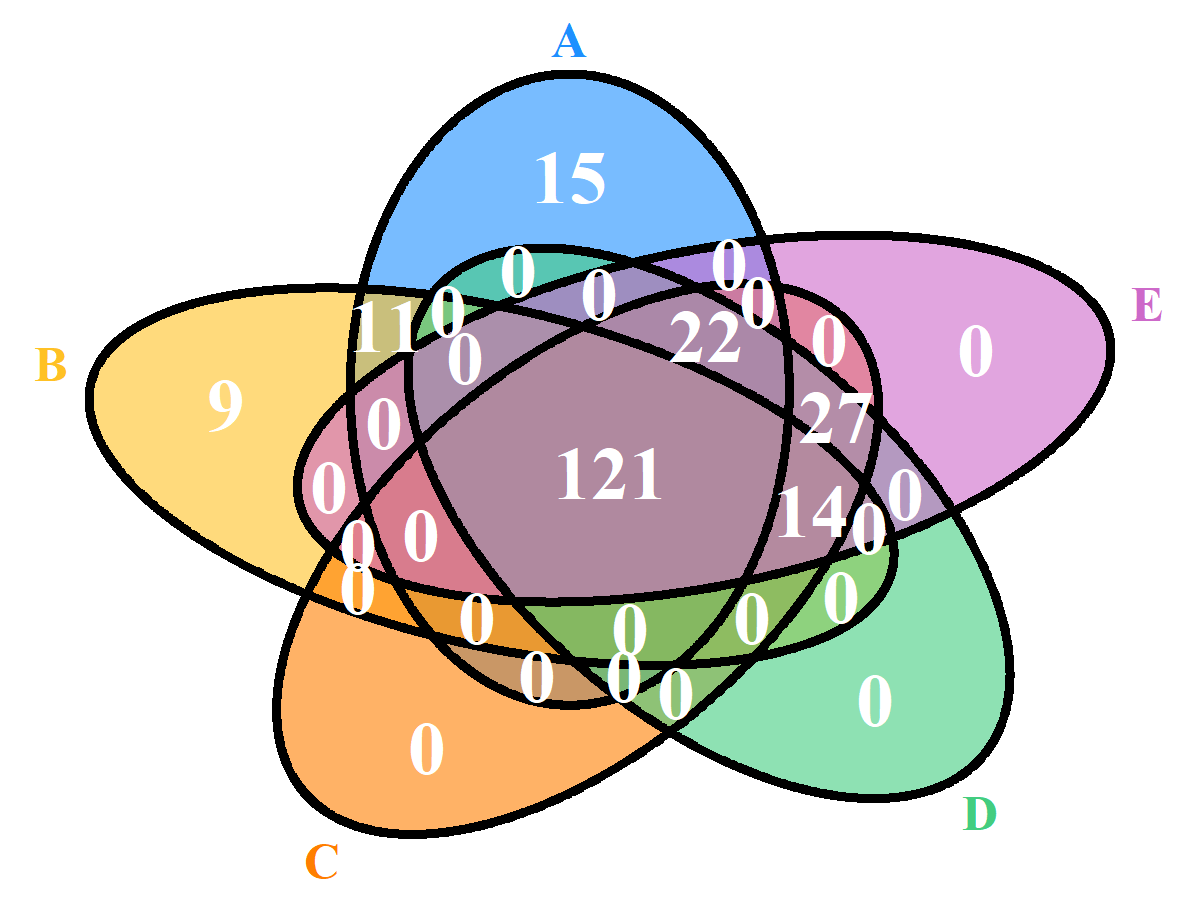
图b4. 集合图使用示例，展示A/B/C三组间共有和特有元素的数量



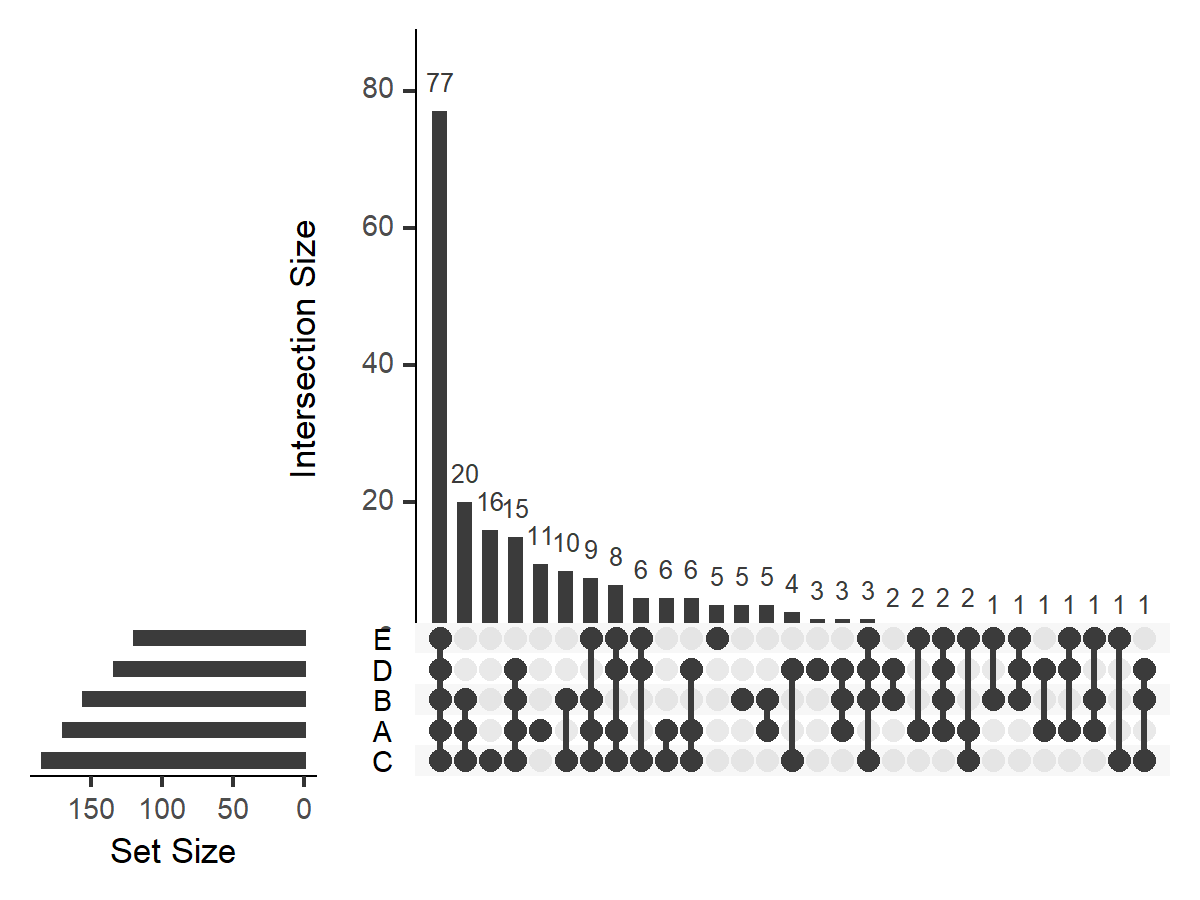
图b5. 维恩图使用示例，展示A/B/C/D四组间共有和特有元素的数量



图b6. 集合图使用示例，展示A/B/C/D四组间共有和特有元素的数量



图b7. 维恩图使用示例，展示A/B/C/D/E五组间共有和特有元素的数量



图b8. 集合图使用示例，展示A/B/C/D/E五组间共有和特有元素的数量

***可以看出，当分组信息>5时，韦恩图的结果有些杂乱；而集合图的结果依然较为清晰。***

#### 文献解读

##### 例1. 两组属水平多样性比较

本文是浙江大学医学院附属儿童医院风湿免疫-变态反应科和中科院遗传发育所合作完成，于2020年4月7日发表在《BMC Genomics》上的论文。文章的主要内容为幼年特发性关节炎患儿肠道菌群的相关研究。文章的详细解读，详见《[BMC：幼儿关节炎患儿肠道菌群的特征](https://mp.weixin.qq.com/s/UunMxA8grkeQ8AMUTtMwRA)》。示例如图1，（该示例来自原文图1B）。

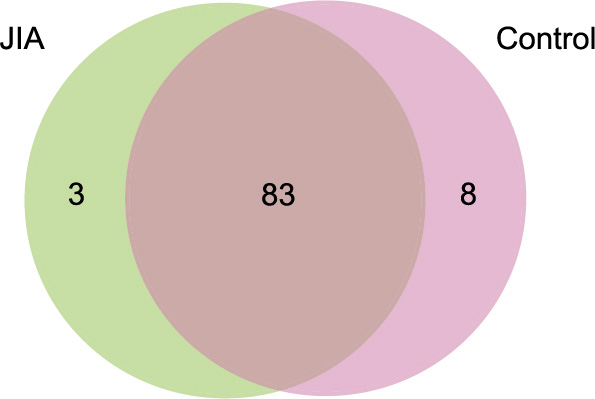


图1. 两组多样性比较。韦恩图，显示了两组有83个属是相同的，但是幼年特发性关节炎（juvenile idiopathic arthritis，JIA）组有3个属是独有的，对照组有8个属是独有的。

Fig. 1 Diversity analyses show that the differences in the α- and β-diversities of the gut microbiota differ between the JIA and the control groups. b Venn diagram based on genera. The two groups have 83 shared genera, with 3 unique genera in the JIA group and 8 unique genera in the control group.

##### 例2. 三组上下调OTUs分别比较

本文是2019年5月由中科院遗传发育所白洋组与JIC的Anne Osbourn组合作在拟南芥代谢物调控根系微生物组领域取得重大突破，成果以Article形式在线发表于Science杂志，海南大学罗杰教授对本工作的意义进行点评。详细报导请点击下方链接：- [Science：拟南芥三萜化合物特异调控根系微生物组\*](https://mp.weixin.qq.com/s/KmMDEmptBKz8Fv7VSdz2Jg) [专家点评](https://mp.weixin.qq.com/s/xlYx9Jb9cBVNs1QrW48RAg)。示例如图2，（该示例来自原文图4C/D）。

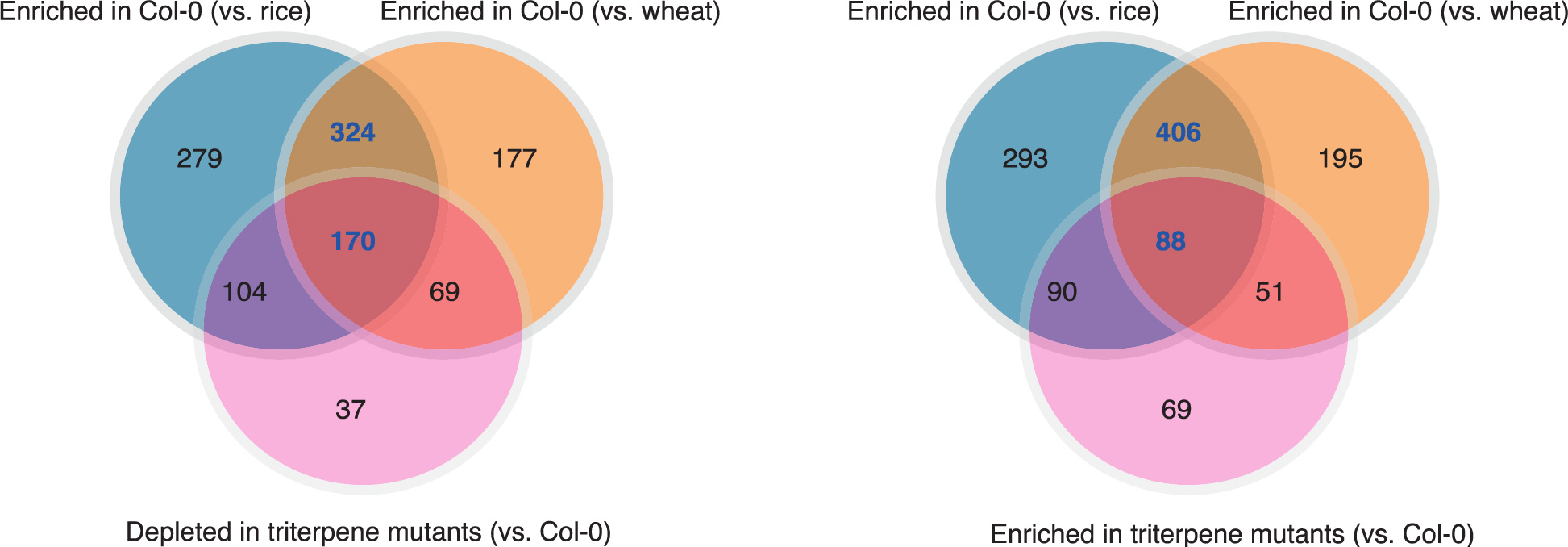


图2. 三萜通路的突变体特异调控的根系细菌类群。 维恩图展示了的拟南芥三萜突变体中下调(左)或富集(右)的OTUs，与水稻和小麦与拟南芥野生型相比变化的OTUs大量重叠。

Fig. 2. Modulation of specific root bacterial taxa in triterpene pathway mutants. Venn diagrams showing substantial overlap of OTUs (left) depleted or (right) enriched in the root microbiota of A. thaliana triterpene mutant lines as compared with the wild type (Col-0) (pink circles), compared with those depleted in the root microbiota of rice (blue circles) and wheat (orange circles) versus the A. thaliana Col-0 wild type.The OTU numbers specifically enriched in the root microbiota of A. thaliana Col-0 compared with rice and wheat are highlighted in blue and bold in the Venn diagram overlaps.

##### 例3. 四组样本比较共有和特有

本文是北京大学人民医院消化内科刘玉兰教授课题组于2020年2月发表在炎症性肠病研究的专业杂志《Inflammatory bowel disease》的文章。主要内容是关于溃疡性结肠炎患者使用5-氨基水杨酸（5-ASA）治疗后肠道真菌菌群的变化特点研究。

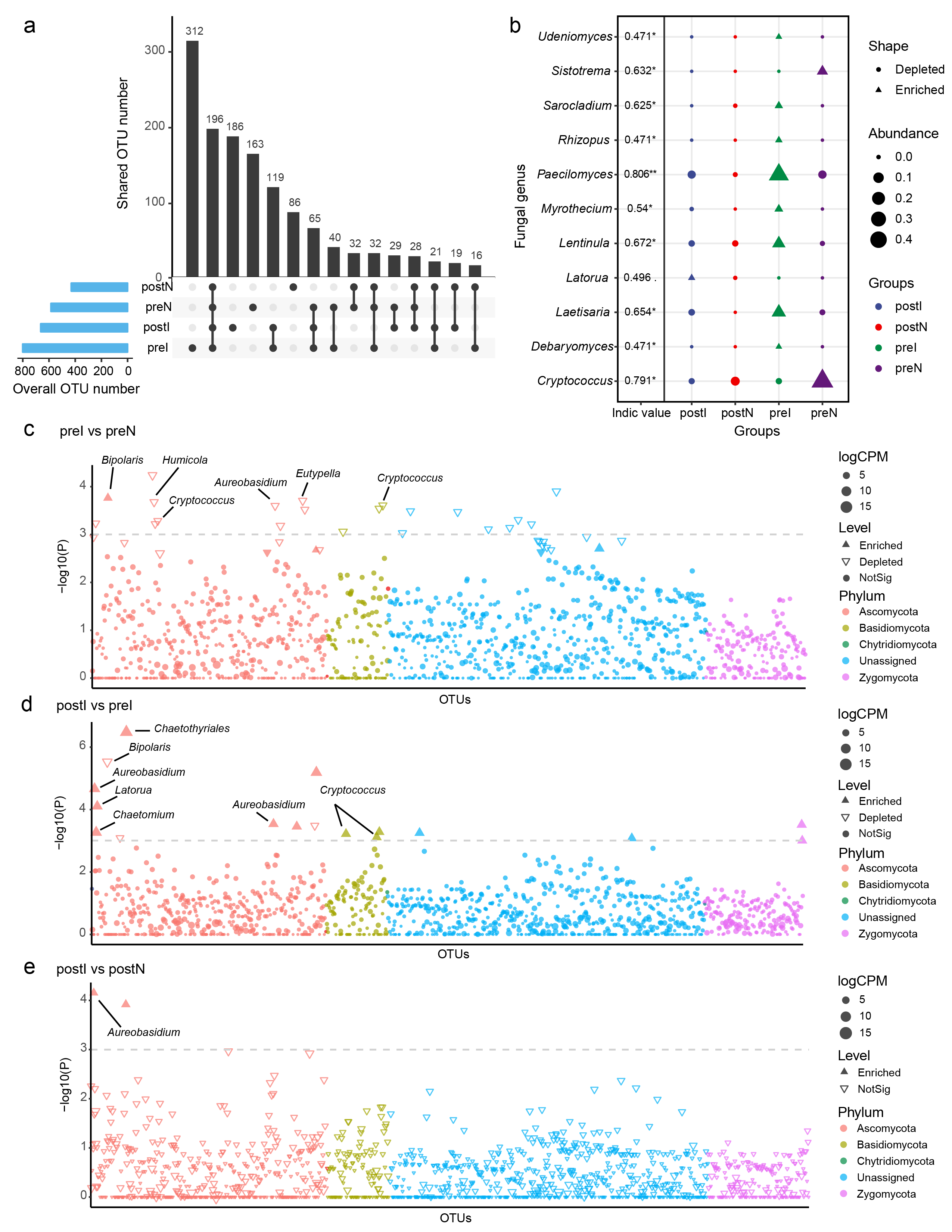


图3. 在验证性研究中组特异性的真菌微生物群。A、集合图显示每组的OTU数量。B、各组的指示物种。利用indicspecies软件包对属水平真菌丰度进行了分析。点的形状代表组内otu的富集（三角）或减少（圆点），点的大小代表真菌属水平的丰度。（C-E）组间OTU丰度差异分析：（C）治疗前的炎性粘膜（preI）与治疗前的非炎性粘膜（preN）；（D）5-ASA治疗后的炎性粘膜（postI）与治疗前的炎性粘膜（preI）；（E）5-ASA治疗后的炎性粘膜（postI）与5-ASA治疗后的非炎性粘膜（postN）。采用EdgeR包进行差异分析,两组之间的差异用曼哈顿图展示。点形状显示前一组的OTUs较后一组增加（Enriched）、降低(Depleted)或不显著(NotSig)。点大小表示OTU的丰度。

FIGURE 3. Group-specific fungal microbiota in the validation study. A, The upset plot shows the OTU count in each group. B, Indicator species in each group. Fungal abundances at the genus level were analyzed using the indicspecies package. The shape of the point represents OTUs enriched or depleted in the group, and point size represents the abundance of OTUs. Comparative analysis of OTU abundance between 2 groups ((C) preI vs preN; (D) postI vs preI; (E) postI vs postN). The EdgeR package was used for comparative analysis. The difference between the 2 groups is shown as a Manhattan diagram. Point shape indicates OTUs enriched, depleted, or not significant in the former group compared with the latter. Point size indicates the abundance of OTUs.

#### 绘图实战

我们在R语言环境下，使用RStudio进行绘图，主要使用VennDiagram和UpsetR包分别用于绘制维恩图和集合图。

##### VennDiagram和UpsetR的数据要求

值得注意的是，VennDiagram和UpsetR的数据要求并不相同，前者要求适用list输入以各组为集合的元素变量名,有几个分组就输入几个集合；而后者以元素变量名为行名，用数字0和1代表元素在分组集合中存在与否，数据输入是以数据框的形式输入。因此在数据准备和图形绘制过程中需要准备对应参数。

**VennDiagram的数据类型**

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Set1 | Set2 | Set3 |
| 1 | A | A | A |
| 2 | B |  | B |
| 3 | C | C | C |
| 4 | D | D |  |
| 5 | E |  | E |
| 6 | F |  |  |

**UpsetR的数据类型**

|  |  |  |  |
| --- | --- | --- | --- |
| Var. | Set1 | Set2 | Set3 |
| A | 1 | 1 | 1 |
| B | 1 | 0 | 1 |
| C | 1 | 1 | 1 |
| D | 1 | 1 | 0 |
| E | 1 | 0 | 1 |
| F | 1 | 0 | 0 |

##### 软件安装

# 设置清华源镜像，国内加速下载(可选)  
site="https://mirrors.tuna.tsinghua.edu.cn/CRAN"  
# 判断R包加载是否成功来决定是否安装后再加载  
package\_list = c("VennDiagram","UpSetR")  
for(p in package\_list){  
 if(!requireNamespace(p, quietly = TRUE))  
 install.packages(p, repos=site)  
}

##### 数据前处理

# 读取属水平相对丰度表  
Data = read.table("d1.genus.profile.txt", header=T, row.names= 1, sep="\t", comment.char="")   
  
# 读取实验设计(元数据，至少3列)，第三列信息随意，没有回报错  
design = read.table("d2.design.txt", header=T, row.names= 1, sep="\t", comment.char="")  
  
# 匹配样本名(design行名和Data的列名)，两表交叉筛选共有  
index = rownames(design) %in% colnames(Data)   
design = design[index,]  
Data = Data[,rownames(design)]   
  
# 数据表中添加分组信息  
Data\_t = t(Data)  
Data\_t2 = merge(design, Data\_t, by="row.names")  
# 删除来自design的非分组信息  
Data\_t2 = Data\_t2[,c(-1,-3)]  
  
# 按组求均值  
# 定义求各组均值函数  
Data\_mean = aggregate(Data\_t2[,-1], by=Data\_t2[1], FUN=mean)  
# 各组求均值   
Data4Pic = as.data.frame(do.call(rbind, Data\_mean)[-1,])  
# 为均值表添加列名（分组信息）  
colnames(Data4Pic) = Data\_mean$group  
Data4Pic=as.data.frame(Data4Pic)  
  
# 将数据表中>0的数值替换为1，数值>0则OTU或各级数据在分组中有出现，替换为1是用于可视化的数据准备  
# 数据为百分比，也可按一定丰度筛选作为有意义或高丰度的特征，如0.1%  
Data4Pic[Data4Pic>0]=1  
  
# 保存数据，用于以后分析需要  
write.table(Data4Pic,"d3.data4venn.txt", sep="\t", quote=F)

##### 使用VennDiagram绘制维恩图

**注意:** VennDiagram的图形绘制结果无法在Rstudio中直接呈现，而是直接生成图形并保存于工作目录。

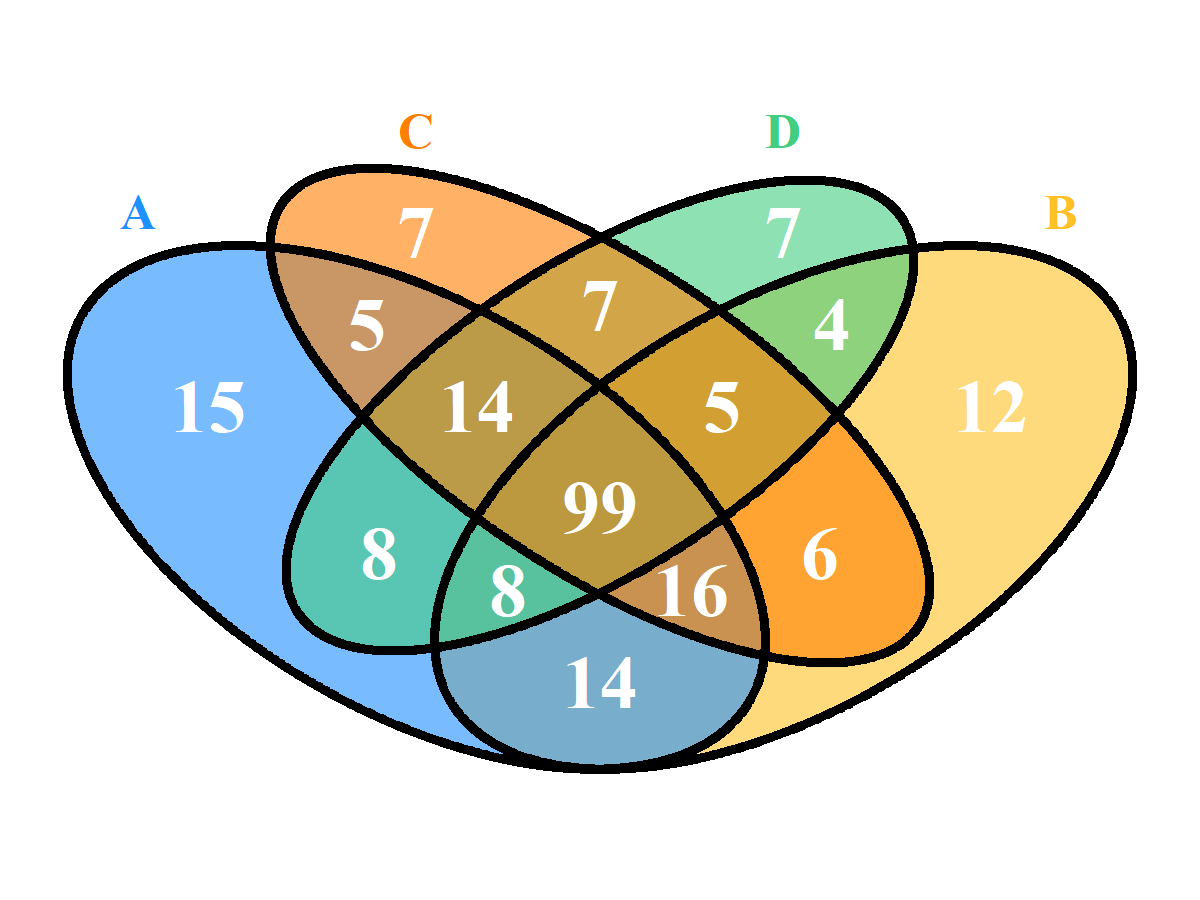
**参数设置：**x=list()指定集合，由于VennDiagram要求输入以各组为集合的元素变量名，因此，作者将提取Data4Pic各组中数值=1的变量名作为数据输入的集合。 filename=指定图形绘制的结果保存的名称。 imagetype=参数设置图片生成的类型，但遗憾的是它只能指定png,tiff等非矢量图格式。 为了能够将图形绘制的结果保存为pdf格式，作者将filename=指定为NULL,并使用grid.draw函数输出图像。

library(VennDiagram)  
Data4Pic = read.table("d3.data4venn.txt", header=T, row.names=1)  
# 设置pdf文件名，用于存储绘制的图形  
pdf(file="p1.GenusVenn.pdf", width=4, height=3, pointsize=8)   
p1 <- venn.diagram(  
 #提取各组中`=1`的行名。需根据自己的分组，调整list中的分组情况  
 x=list(A=row.names(Data4Pic[Data4Pic$A==1,]),  
 B=row.names(Data4Pic[Data4Pic$B==1,]),  
 C=row.names(Data4Pic[Data4Pic$C==1,]),  
 D=row.names(Data4Pic[Data4Pic$C==1,])),  
 # 不指定保存的文件名，也不指定`imagetype`  
 filename = NULL, lwd = 3, alpha = 0.6,  
 #设置字体颜色  
 label.col = "white", cex = 1.5,  
 #设置各组的颜色  
 fill = c("dodgerblue", "goldenrod1", "darkorange1", "seagreen3"),   
 cat.col = c("dodgerblue", "goldenrod1", "darkorange1", "seagreen3"),  
 #设置字体  
 fontfamily = "serif", fontface = "bold",  
 cat.fontfamily = "serif",cat.fontface = "bold",  
 margin = 0.05)  
#使用`grid.draw`函数在`venn.diagram`绘图函数外绘制图形  
grid.draw(p1)   
dev.off()

png   
 2

png(file="p1.GenusVenn.png", width=4, height=3, res=300, units="in")   
grid.draw(p1)   
dev.off()

png   
 2



##### 使用UpsetR绘制集合图

为方便代码兼容R3.6和R4.0版本，再进行一次数据处理

# 目的是将'Data4Pic'的行名转换为'numeric'的行号  
row.names(Data4Pic) <- 1:nrow(Data4Pic)

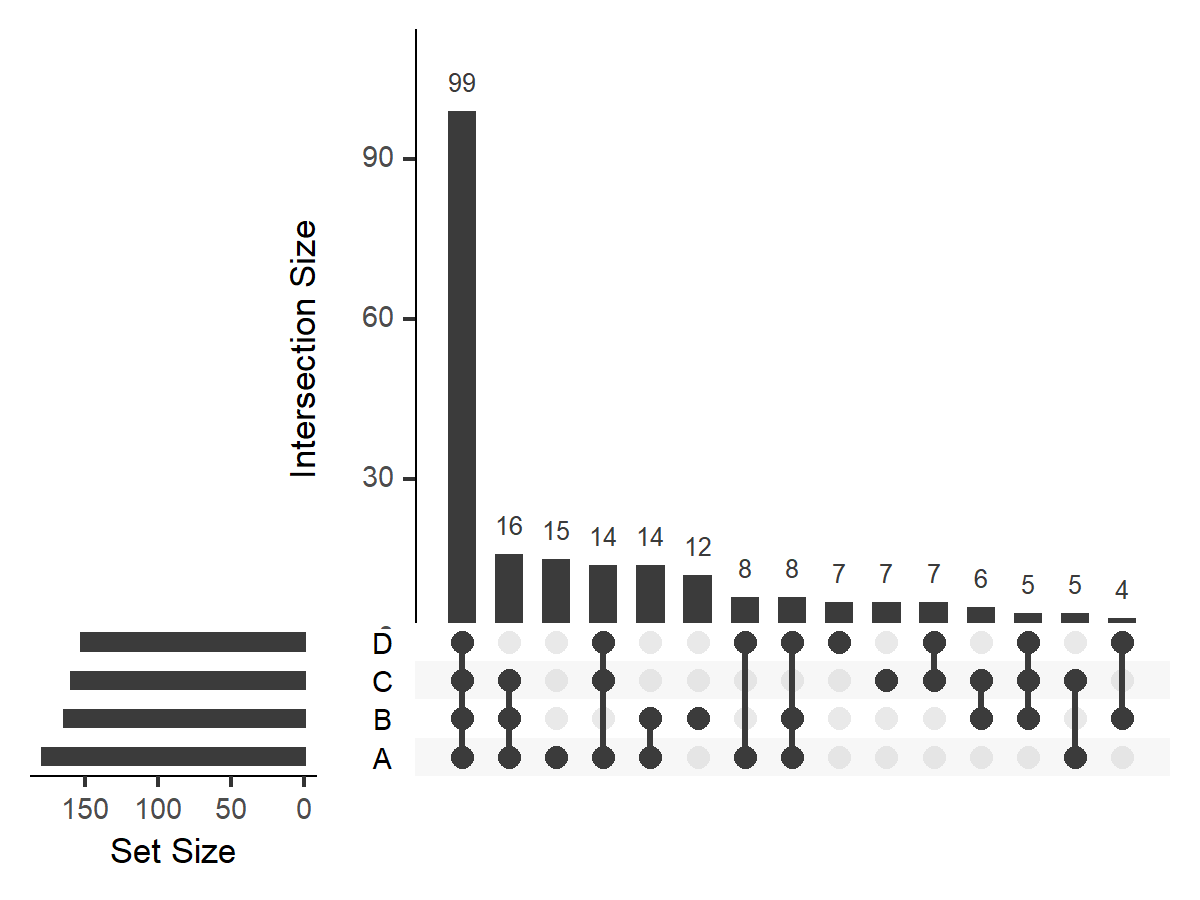
随后进行数据可视化

library(UpSetR)  
# 集合图的基本图形绘制  
pdf(file="p2.GenusUpset.pdf", width=4, height=3, pointsize=8)  
(p2 <-upset(Data4Pic, sets = colnames(Data4Pic),order.by = "freq"))  
dev.off()

png   
 2

png(file="p2.GenusUpset.png", width=4, height=3, res=300, units="in")   
p2  
dev.off()

png   
 2



**Upset常用参数**

* example：存放矩阵的变量名称；
* set：所需要的集合名称；
* mb.ratio：调整上下两部分的比例；
* order.by：排序方式，freq为按频率排序；
* queries：查询函数，用于对指定列添加颜色；
  + param: list：query作用于哪个交集；
  + color：每个query都是一个list，里面可以设置颜色，没设置的话将调用包里默认的调色板；
  + active：被指定的条形图：TRUE显示颜色，FALSE在条形图顶端显示三角形;
* nset：集合数量，也可用set参数指定具体集合；
* number.angles：上方条形图数字角度，0为横向，90为竖向，但90时不在正上方；
* point.size：下方点阵中点的大小；
* line.size：下方点阵中每个线的粗细；
* mainbar.y.label：上方条形图Y轴名称;sets.x.label：左下方条形图X轴名称；
* text.scale：六个数字控制关系见；
* query.legend：指定query图例的位置……

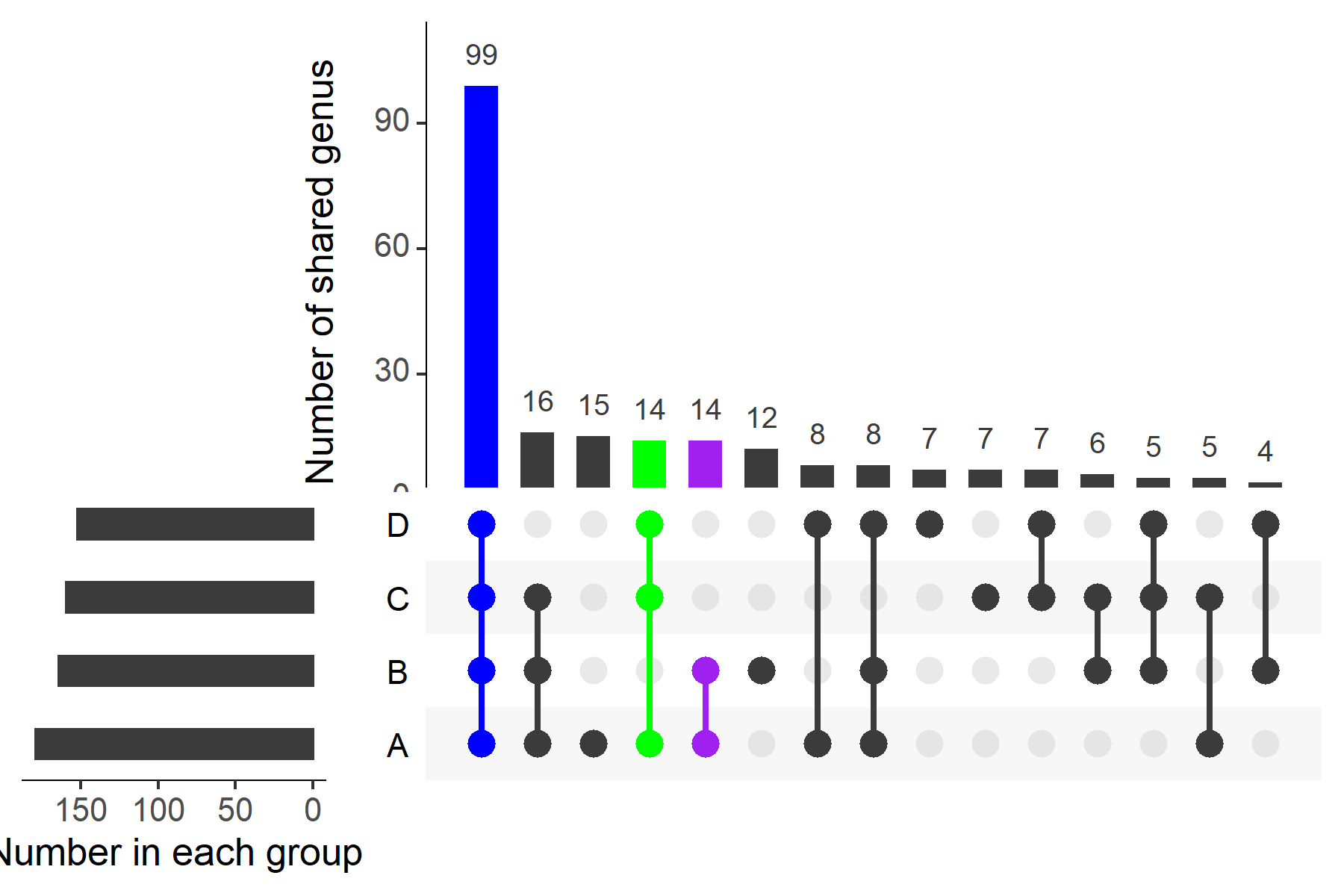
使用Queries参数对集合图进行修饰

pdf(file="p3.GenusUpsetIndiv.pdf", width=6, height=4, pointsize=8)  
p3<-upset(Data4Pic, sets = colnames(Data4Pic), mb.ratio = c(0.55, 0.45), order.by = "freq",  
 queries = list(list(query=intersects, params=list("A", "B"), color="purple", active=T),   
 list(query=intersects, params=list("C", "D", "A"), color="green", active=T),   
 list(query=intersects, params=list("B", "C", "A", "D"), color="blue", active=T)),   
 nsets = 3, number.angles = 0, point.size = 4, line.size = 1,   
 mainbar.y.label = "Number of shared genus",  
 sets.x.label = "Number in each group", text.scale = c(1.5, 1.5, 1.5, 1.5, 1.5, 1.5))  
p3  
dev.off()

png   
 2

png(file="p3.GenusUpsetIndiv.png", width=6, height=4, res=300, units="in")   
p3  
dev.off()

png   
 2



**更多参数信息**

此外，UpsetR还提供了attribute.plots参数，可绘制histogram，scatter\_plot，boxplot.summary等图形的绘制。可以根据自己数据的需要进行配置数据，详细可见：<https://cran.r-project.org/web/packages/UpSetR/vignettes/>

#### 参考文献

Xubo Qian, Yong-Xin Liu, Xiaohong Ye, Wenjie Zheng, Shaoxia Lv, Miaojun Mo, Jinjing Lin, Wenqin Wang, Weihan Wang, Xianning Zhang & Meiping Lu. (2020). Gut microbiota in children with juvenile idiopathic arthritis: characteristics, biomarker identification, and usefulness in clinical prediction. BMC Genomics 21, 286, doi: <https://doi.org/10.1186/s12864-020-6703-0>

Ancheng C. Huang, Ting Jiang, Yong-Xin Liu, Yue-Chen Bai, James Reed, Baoyuan Qu, Alain Goossens, Hans-Wilhelm Nützmann, Yang Bai & Anne Osbourn. (2019). A specialized metabolic network selectively modulates Arabidopsis root microbiota. Science 364, eaau6389, doi: <https://doi.org/10.1126/science.aau6389>

Xu Jun, Chen Ning, Song Yang, Wu Zhe, Wu Na, Zhang Yifan, Ren Xinhua & Liu Yulan. (2019). Alteration of Fungal Microbiota After 5-ASA Treatment in UC Patients. Inflammatory Bowel Diseases 26, 380-390, doi: <https://doi.org/10.1093/ibd/izz207>

venn.diagram保存pdf格式文件？，<https://www.cnblogs.com/jessepeng/p/11610055.html>

使用VennDiagram包绘制韦恩图，<https://www.jianshu.com/p/285b4ac66768>

UpSetR官方帮助文档，<https://cran.r-project.org/web/packages/UpSetR/vignettes/>

进阶版Venn plot：Upset plot入门，<https://blog.csdn.net/tuanzide5233/article/details/83109527>

责编：刘永鑫 中科院遗传发育所

版本更新历史 - 1.0.0，2020/6/8，徐俊，初稿 - 1.0.1，2020/6/8，刘永鑫审阅建议：R 4.0下建议有错误，补充2-5个实例 - 1.0.2，2020/6/9，腾讯会议沟通统一思想和格式 - 1.0.3，2020/6/23，徐俊，大修，添加背景、实例和规范代码讲解 - 1.0.4，2020/7/2，刘永鑫，大修，整合背景、实例、实战为Rmd格式，全文修改 - 1.0.5，2020/7/3，徐俊，背景图添加2-5组样式展示，文章补充矢量图 - 1.0.6，2020/7/7，席娇，内容修改，格式微调