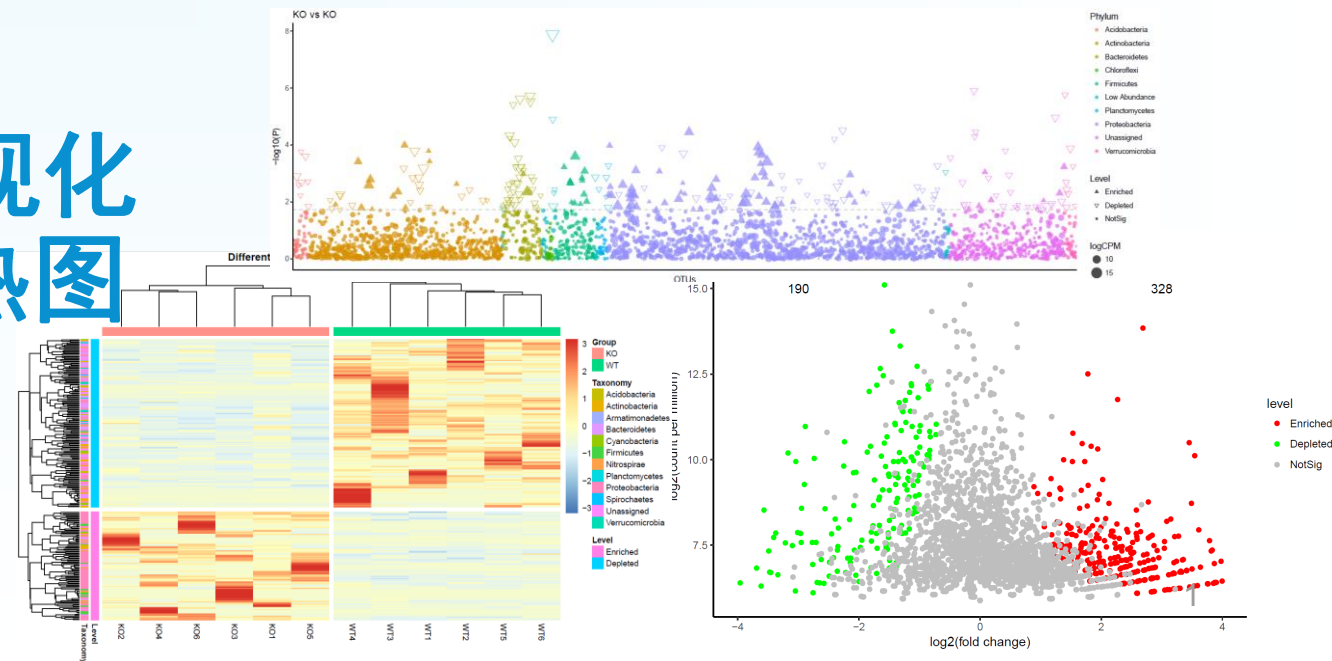




24组间差异比较及可视化 曼哈顿图、火山图和热图

易生信
2022年1月8日



- edgeR包计算差异OTUs/ASVs
- 火山图展示差异OTUs/ASVs
- 热图展示差异OTUs/ASVs
- 曼哈顿图展示差异OTUs/ASVs

易生信 生信宝典 宏基因组



- **edgeR包计算差异OTUs/ASVs**
- 火山图展示差异OTUs/ASVs
- 热图展示差异OTUs/ASVs
- 曼哈顿图展示差异OTUs/ASVs

易生信 生信宝典 宏基因组



edgeR: 数字基因表达数据差异表达分析包



Mark D. Robinson

FOLLOW

Associate Professor of Statistical Genomics,
[University of Zurich](#)

Verified email at imls.uzh.ch - [Homepage](#)

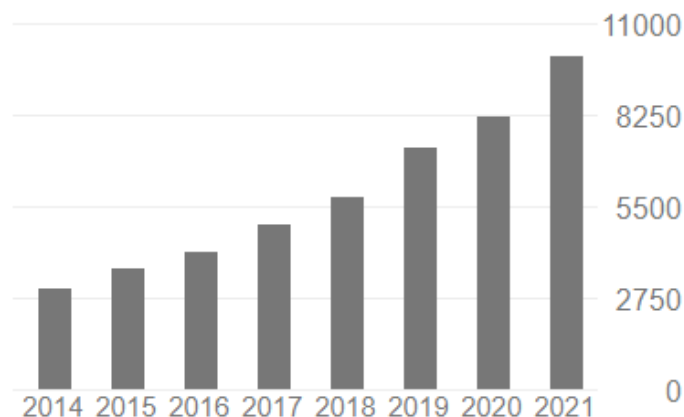
Statistics Bioinformatics Transcriptomics
Single-Cell Data Analysis Epigenomics

TITLE	CITED BY	YEAR
edgeR: a Bioconductor package for differential expression analysis of digital gene expression data MD Robinson, DJ McCarthy, GK Smyth Bioinformatics 26 (1), 139-140	23298	2010
Comprehensive genomic characterization defines human glioblastoma genes and core pathways Cancer Genome Atlas (TCGA) Research Network Nature 455 (7216), 1061	5910	2008
<pre>if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager") BiocManager::install("edgeR")</pre>		2010

Cited by

[VIEW ALL](#)

	All	Since 2016
Citations	59541	40594
h-index	64	54
i10-index	115	109



Public access

[VIEW ALL](#)

0 articles

85 articles

not available

available

Robinson MD, McCarthy DJ, Smyth GK. [edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26\(1\):139-40.](#)

软件 <https://www.bioconductor.org/packages/release/bioc/html/edgeR.html>



如何使用edgeR

主页 工具

edgeRUsersGuide.p... x

53 / 107

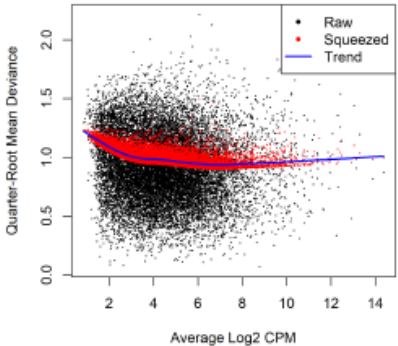
49.4%

The square root of dispersion is the coefficient of biological variation (BCV). The common BCV is on the high side, considering that this is a designed experiment using genetically identical plants. The trended dispersion shows a decreasing trend with expression level. At low logCPM, the dispersions are very large indeed.

Note that only the trended dispersion is used under the quasi-likelihood (QL) pipeline. The tagwise and common estimates are shown here but will not be used further.

The QL dispersions can be estimated using the `glmQLFit` function, and then be visualized with the `plotQLDisp` function.

```
> fit <- glmQLFit(y, design, robust=TRUE)
> plotQLDisp(fit)
```



```
AT2G11230      3.50      -1.532    5.60  98.7 1.22e-08 3.36e-05
AT2G07782      3.48      -1.616    5.28  93.5 1.70e-08 4.01e-05
AT2G18193      3.05      -2.396    5.08  84.8 3.09e-08 6.04e-05
AT2G23910      3.59      -0.384    5.13  83.9 3.29e-08 6.04e-05
AT5G54830      3.07      -0.367    6.07  79.7 4.51e-08 7.31e-05

> FDR <- p.adjust(qlf$table$PValue, method="BH")
> sum(FDR < 0.05)
[1] 1628
```

Now conduct QL F-tests for the pathogen effect and show the top genes. By default, the test is for the last coefficient in the design matrix, which in this case is the treatment effect:

```
> qlf <- glmQLFTest(fit)
> topTags(qlf)
```

	Coefficient:	Treathrcc					
	logFC	logCPM	F	PValue	FDR		
AT2G19190	4.50	7.37	304	1.83e-10	2.62e-06		
AT2G39530	4.34	6.71	278	3.17e-10	2.62e-06		
AT3G46280	4.78	8.10	247	6.70e-10	2.78e-06		
AT2G39380	4.94	5.77	247	6.72e-10	2.78e-06		
AT1G51800	3.97	7.71	232	9.92e-10	3.28e-06		
AT1G51850	5.32	5.42	209	1.89e-09	4.30e-06		
AT5G48430	6.32	6.73	203	2.30e-09	4.30e-06		
AT2G44370	5.41	5.20	200	2.50e-09	4.30e-06		
AT1G51820	4.34	6.37	198	2.64e-09	4.30e-06		
AT3G55150	5.78	4.90	196	2.80e-09	4.30e-06		

Here's a closer look at the individual counts-per-million for the top genes. The top genes are very consistent across the three replicates:

```
> top <- rownames(topTags(qlf))
> cpm(y)[top,]
              mock1 mock2 mock3 hrcc1 hrcc2 hrcc3
AT2G19190 16.696  12.0 13.29 341.3 254.7 351.1
```



准备输入文件：特征表和实验设计

原始序列记数(read counts)的特征表：otutab.txt

#OTUID	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2	OE3	OE4	OE5	OE6	WT1	WT2	WT3	WT4	WT5	WT6
ASV_657	1073	1926	810	1356	1064	1069	1259	1610	1337	944	1245	1013	2303	2512	1698	1974	1441	1544
ASV_2	1965	1233	2368	2241	2901	1835	641	497	1225	1271	948	638	1286	1499	843	1122	1496	1177
ASV_3	567	460	898	902	1224	854	606	455	1057	1035	837	673	1039	1795	1019	1200	1202	765
ASV_12	240	250	1823	677	1459	491	324	1918	1487	991	557	337	1075	1199	776	408	1037	1424
ASV_4	1434	401	536	759	1289	507	516	592	440	622	661	429	1125	1449	547	577	1115	925
ASV_6	890	680	829	900	1503	1026	254	265	372	384	400	359	561	543	466	546	509	507
ASV_8	508	504	608	424	190	327	335	535	1578	780	507	516	634	763	553	1053	457	514
.....																
ASV_N	122	440	509	90	191	293	74	81	71	132	170	69	159	231	125	493	150	117

实验设计：metadata.txt

SampleID	Group	Date	Site	CRA	CRR	BarcodeSequence	LinkerPrimerSequence	ReversePrimer
KO1	KO	2017/6/30	Chaoyang	CRA002352	CRR117575	ACGCTCGACA	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
KO2	KO	2017/6/30	Chaoyang	CRA002352	CRR117576	ATCAGACACG	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
KO3	KO	2017/7/2	Changping	CRA002352	CRR117577	ATATCGCGAG	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
KO4	KO	2017/7/2	Changping	CRA002352	CRR117578	CACGAGACAG	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
KO5	KO	2017/7/4	Haidian	CRA002352	CRR117579	CTCGCGTGTC	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
KO6	KO	2017/7/4	Haidian	CRA002352	CRR117580	TAGTATCAGC	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
OE1	OE	2017/6/30	Chaoyang	CRA002352	CRR117581	TCTCTATGCG	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC
OE2	OE	2017/6/30	Chaoyang	CRA002352	CRR117582	TACTGAGCTA	AACMGGATTAGATACCCCKG	ACGTCATCCCCACCTTCC

要求：第一列为样品名，其名列分为分组和其他样本信息

命令行模式分步计算实现灵活的差异分析和可视化 (pipeline.sh # 24、R语言差异分析)

输入特征表、元数据；指定分组列名、比较组和丰度

○ # 可选统计方法 **wilcox / t.test / edgeR**、pvalue和fdr和输出目录

```
compare="KO-WT"
```

```
Rscript ${db}/script/compare.R \
```

```
--input result/otutab.txt --design result/metadata.txt \
```

```
--group Group --compare ${compare} --threshold 0.1 \
```

```
--method edgeR --pvalue 0.05 --fdr 0.2 \
```

```
--output result/compare/
```

输出你选择的方法

```
[1] "Your are using edgeR test!"
```

输出差异比较的统计结果

```
Depleted Enriched NotSig
```

```
20      22      129
```

输出结果为\${compare}.txt，如KO-WT.txt



方法2. Rmd模式下运行差异比较

- 准备输入文件metadata.txt和result目录下otutab.txt、taxonomy.txt、
- Rstudio打开Compare.Rmd文件
- 阅读14 – 38行阅读帮助文档
- 检查“# 解析命令行”段落(53-74行)
- 默认比较KO-WT组差异，并绘制火山图、曼哈顿图和热图
- Knit生成结果和计算过程网页结果
- 结果位于当前目录下的统计结果 KO-WT_all/sig.txt, volcano/heatmap/manhattan.pdf



- 主要有两类文件，以KO-WT为例说明

- 2个表格

所有OTUs/ASVs比较结果：KO-WT_all.txt

筛选的显著差异OTUs/ASVs：KO-WT_sig.txt

- 3个图片

火山图：KO-WT_volcano.pdf

曼哈顿图：KO-WT_manhattan.pdf

热图：KO-WT_heatmap.pdf

易生信 生信宝典 宏基因组



OTUs/ASVs比较结果表格

- 使用Excel打开，选择文本文件；或将文件拖拽入Excel中打开

OTUID	logFC	logCPM	PValue	FDR	level	Phylum	Order	Genus	Mean A	Mean B
ASV_251	2.417	9.569	3.38E-08	5.57E-05	Enriched	Proteobacteria	Burkholderiales	Massilia	0.106	0.024
ASV_201	3.008	9.929	5.33E-08	5.57E-05	Enriched	Proteobacteria	Burkholderiales	Unassigned	0.144	0.022
ASV_2907	-3.205	7.853	6.58E-08	5.57E-05	Depleted	Planctomycetes	Planctomycetales	Blastopirellula	0.003	0.035
ASV_1191	-5.413	7.063	8.46E-08	5.57E-05	Depleted	Proteobacteria	Myxococcales	Unassigned	0	0.017

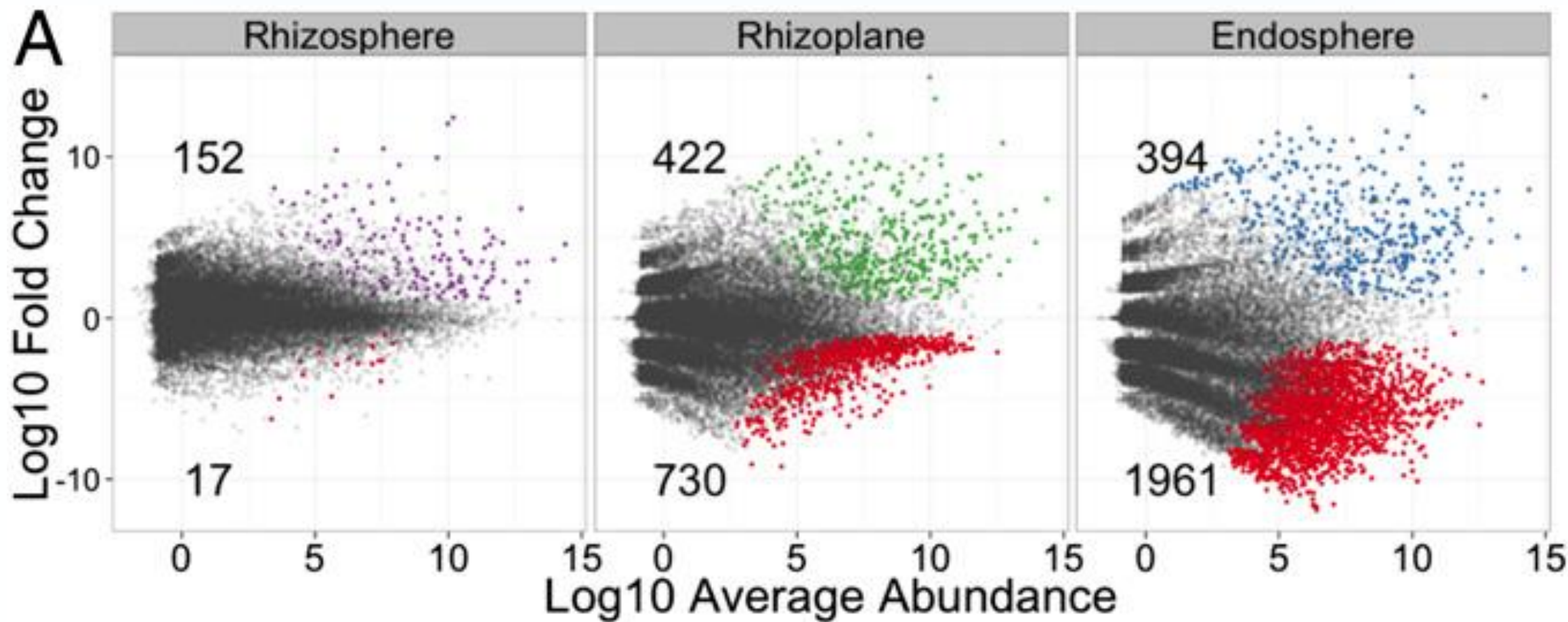
- logFC是Fold Change差异倍数取log2对数，正数为富集，负为下降
- logCPM是所有样品均值的百万分比+1取log2，范围是0~20
- MeanA和MeanB表示两个组的百分比均值
- Level是根据pvalue+FDR筛选的显著上调或下降的特征(OTUs/ASVs)

- edgeR包计算差异OTUs/ASVs
- 火山图展示差异OTUs/ASVs
- 热图展示差异OTUs/ASVs
- 曼哈顿图展示差异OTUs/ASVs

易生信 生信宝典 宏基因组

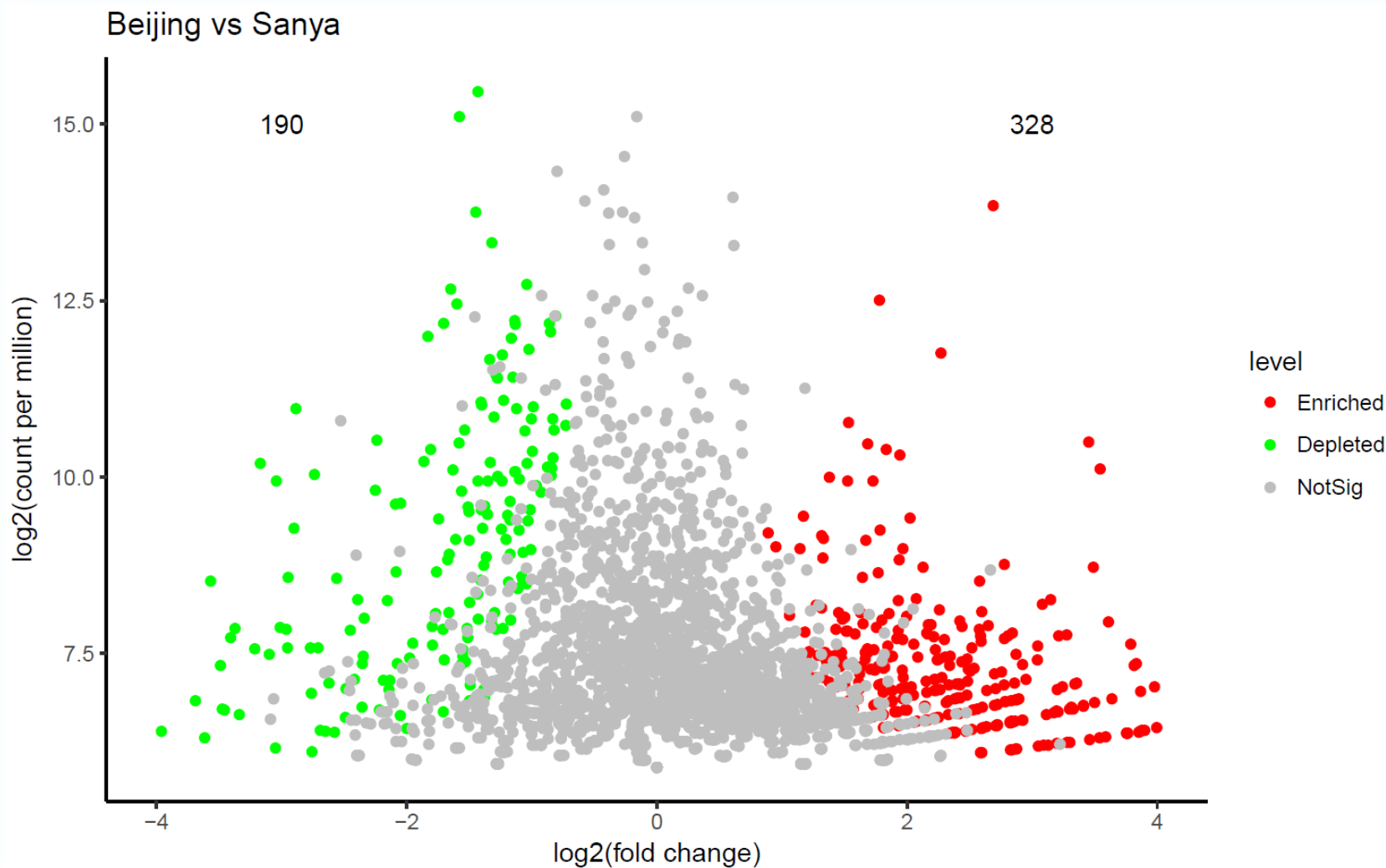


火山图展示根际-根表-根内与土差异OTUs/ASVs数量



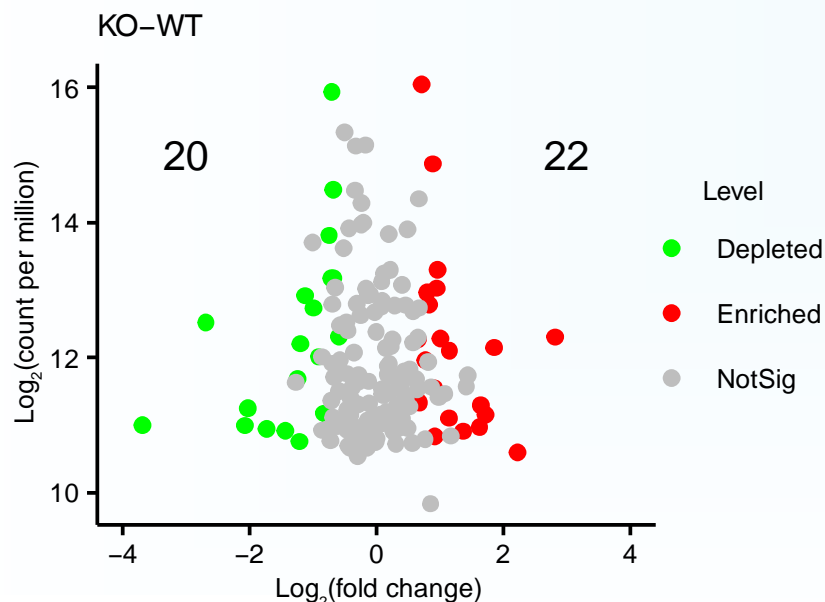
Edwards, Joseph, et al. "Structure, variation, and assembly of the root-associated microbiomes of rice." PNAS 112.8 (2015): E911-E920.
PNAS: 水稻微生物组

火山图展示OTUs/ASVs组间差异倍数和相对丰度



- # 输入compare.R的结果，输出火山图带数据标签，可指定图片大小

```
Rscript ${db}/script/compare_volcano.R \  
--input result/compare/${compare}.txt \  
--output result/compare/${compare}.txt.volcano.pdf  
--width 89 --height 59
```



此图和前面示例的结果数量和差异上有很大区别

我们在差异比较步骤进行了千分之一丰度筛选，一般只会保留100个左右的特征，抓住主要矛盾。实际分析中，可根据需求优化丰度筛选阈值，选择更符合预期和具有合理生物学意义的结果

Rmd文档中ggplot绘制火山图代码

```
p = ggplot(output, aes(x=logFC, y=logCPM, color=level)) +  
  geom_point() + xlim(-4, 4) + theme_classic()+  
  scale_colour_manual(values=c("red","green","grey")) +  
  labs(x="log2(fold change)", y="log2(count per million)",  
       title=paste(group_list[1], "vs", group_list[2], sep=" "))+  
  annotate("text",x=-3,y=15,label=paste(NoD,sep=""))+  
  annotate("text",x=3,y=15,label=paste(NoE,sep=""))
```

设置x,y轴和颜色数据列

散点图，x轴范围，经典主题

手动控制颜色

坐标轴标签

图标题

添加显著下降数量标签

添加显著上升数量标签

ggplot2绘图就是搭积木，需要一步一个脚印，和我们正常画图一样，美图是一层一层画出来的，想画哪里画哪里！



在线绘图：点点鼠标就完成

- 访问 <http://www.ehbio.com/ImageGP/> —— 选择左侧 Volcano plot
- 我们的结果正好符合第二种多列数据，粘贴进文本框
- 必须参数 Fold 选 logFC，Statistical 选 logCPM，去掉 -Log10 勾选，Color 选择 level 并右侧可选择显示顺序
- Layout 中可修改 X 轴是否对称、图例位置、透明度和点大小
- 修改图标题可设置标题、X/Y 轴标签
- 修改大小可设置图片长宽，单位为 cm

易生信
生信宝典
宏基因组

火山图进一步学习：

- 理论讲解：
- 5火山图：差异OTU数量及变化规律
- 实战代码：
- 5火山图：差异OTU数量及变化规律
- R语言学习 - 火山图

易生信 生信宝典 宏基因组

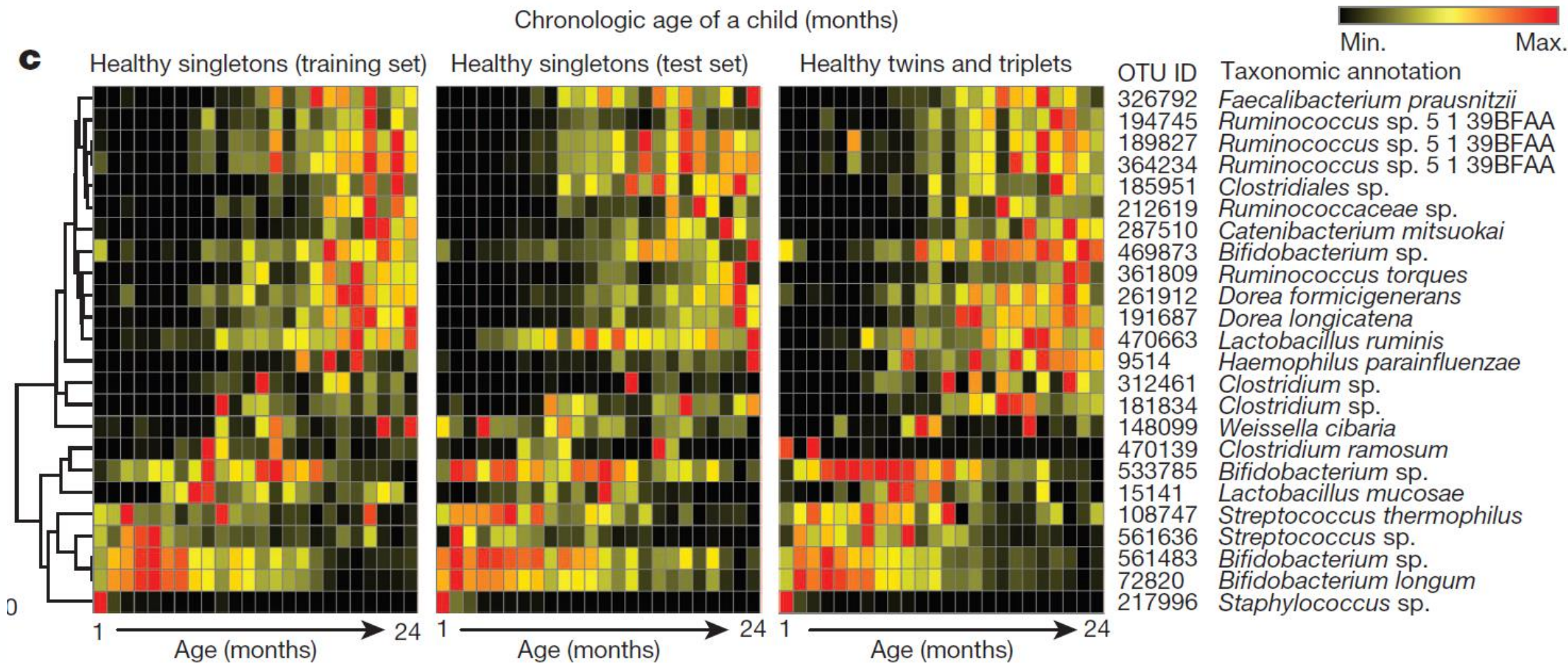


- edgeR包计算差异OTUs/ASVs
- 火山图展示差异OTUs/ASVs
- **热图展示差异OTUs/ASVs**
- 曼哈顿图展示差异OTUs/ASVs

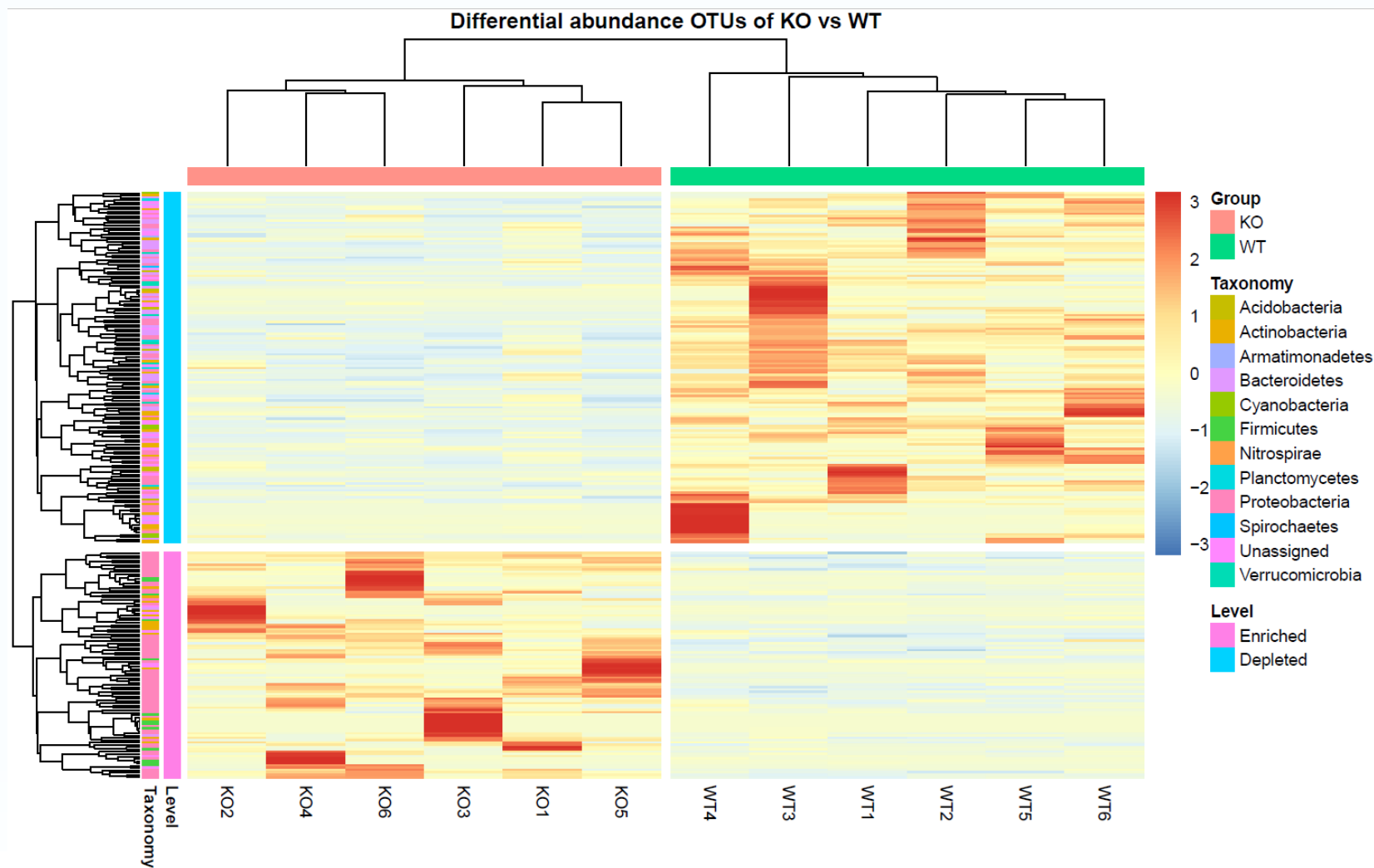
易生信 生信宝典 宏基因组



热图在文章中实例



组间差异热图绘制



默认生成火山图: KO-WT_heatmap.pdf

Rmd中Pheatmap包代码绘制热图

```
pheatmap(norm[rownames(output),],
```

```
  scale = "row",
```

```
  cutree_rows=2,cutree_cols = 2,
```

```
  annotation_col = anno_col, annotation_row = anno_row,
```

```
  filename = paste(opts$output, "_heatmap.pdf", sep=""),
```

```
  width=opts$width, height=opts$height,
```

```
  annotation_names_row= T,annotation_names_col=F,
```

```
  show_rownames=F,show_colnames=T,
```

```
  main = paste("Differential abundance OTUs/ASVs of",group_list[1], "vs", group_list[2],sep=""),
```

```
  fontsize=7,display_numbers=F)
```

标准化矩阵再筛选

行Z-Score标准化，差异更明显

行、列分组

行、列加额外注释标签

输出文件名，paste添加变量

图片输出宽和高

注释列是否显示名称

是否显示行/列名称

标题，

字体，是否显示单元格数值



<http://www.ehbio.com/ImageGP/index.php/Home/Index/PHeatmap.html>

- 主体——相对丰度矩阵

KO-WT_sig.txt中制作有行、列名称的相对丰度矩阵

- 行注释——变化类型、物种丰度信息(可选)

KO-WT_sig.txt中制作有行、列名称的物种信息，可多层注释

- 列注释——分组信息(可选)

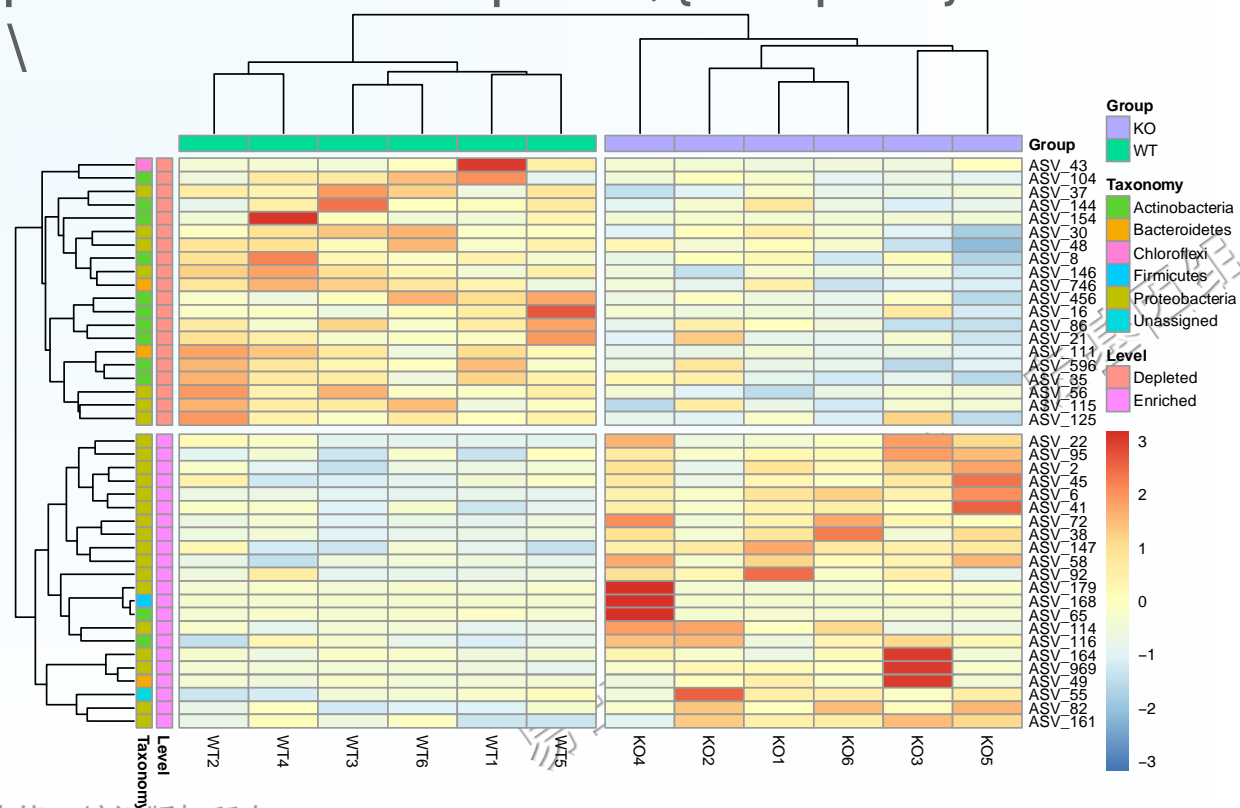
metadata.txt筛选对应样品的信息，可多种分组方式同时显示



命令行绘制差异特征热图+分组和物种注释

- 输入compare.R的结果，筛选列数，指定元数据和分组、物种注释，图大小英寸和字号

```
bash ${db}/script/compare_heatmap.sh -i result/compare/${compare}.txt -l 7 \  
-d result/metadata.txt -A Group \  
-t result/taxonomy.txt \  
-w 8 -h 5 -s 7 \  
-o result/compare/${compare}
```



热图进一步学习：

- 理论讲解：
- 3热图：差异菌、OTU及功能
- 实战代码：
- 3热图：差异菌、OTU及功能
- R语言学习 - 热图绘制 (heatmap)
- R语言学习 - 热图简化
- R语言学习 - 热图美化

易生信 生信宝典 宏基因组

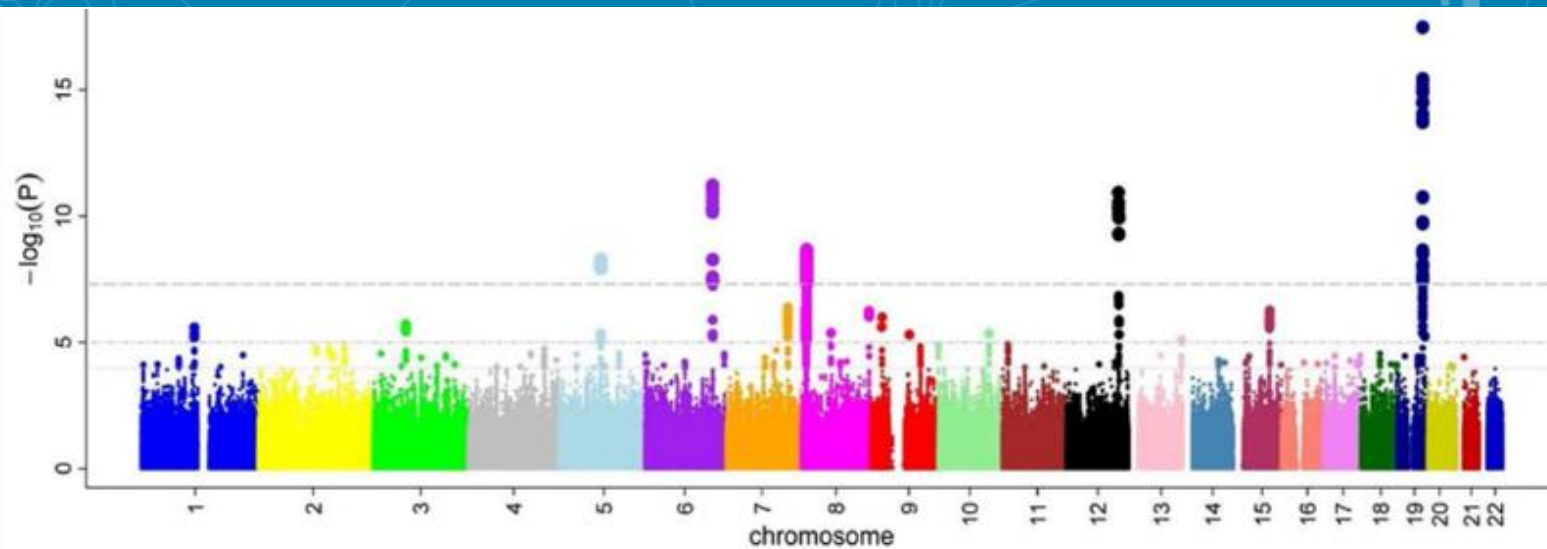


- edgeR包计算差异OTUs/ASVs
- 火山图展示差异OTUs/ASVs
- 热图展示差异OTUs/ASVs
- **曼哈顿图展示差异OTUs/ASVs**

易生信 生信宝典 宏基因组



曼哈顿图 Manhattan Plot



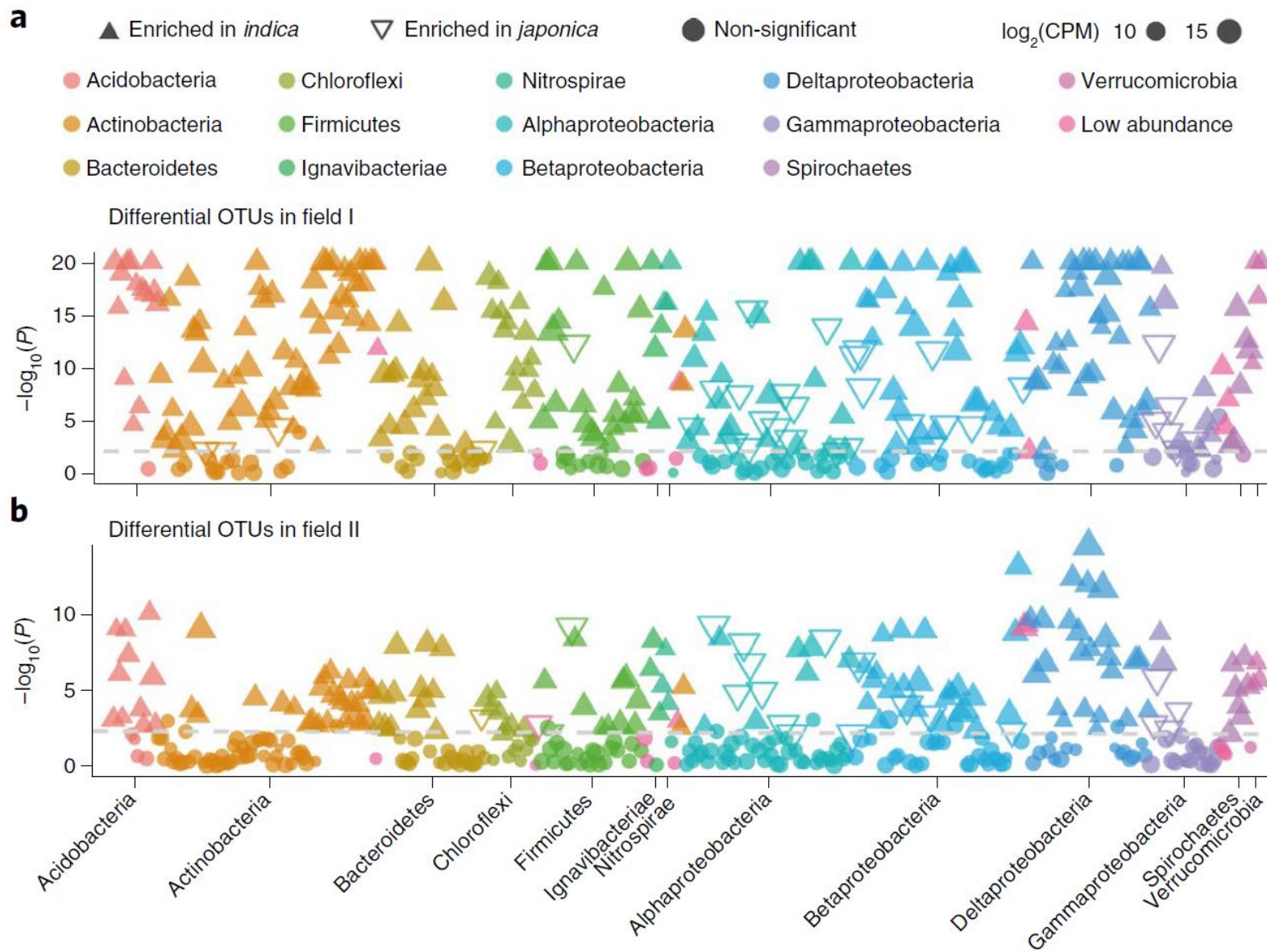
宏基因组



曼哈顿图展示变化OTUs/ASVs分布特点-相同

图3. 籼粳稻根系微生物组物种的差异

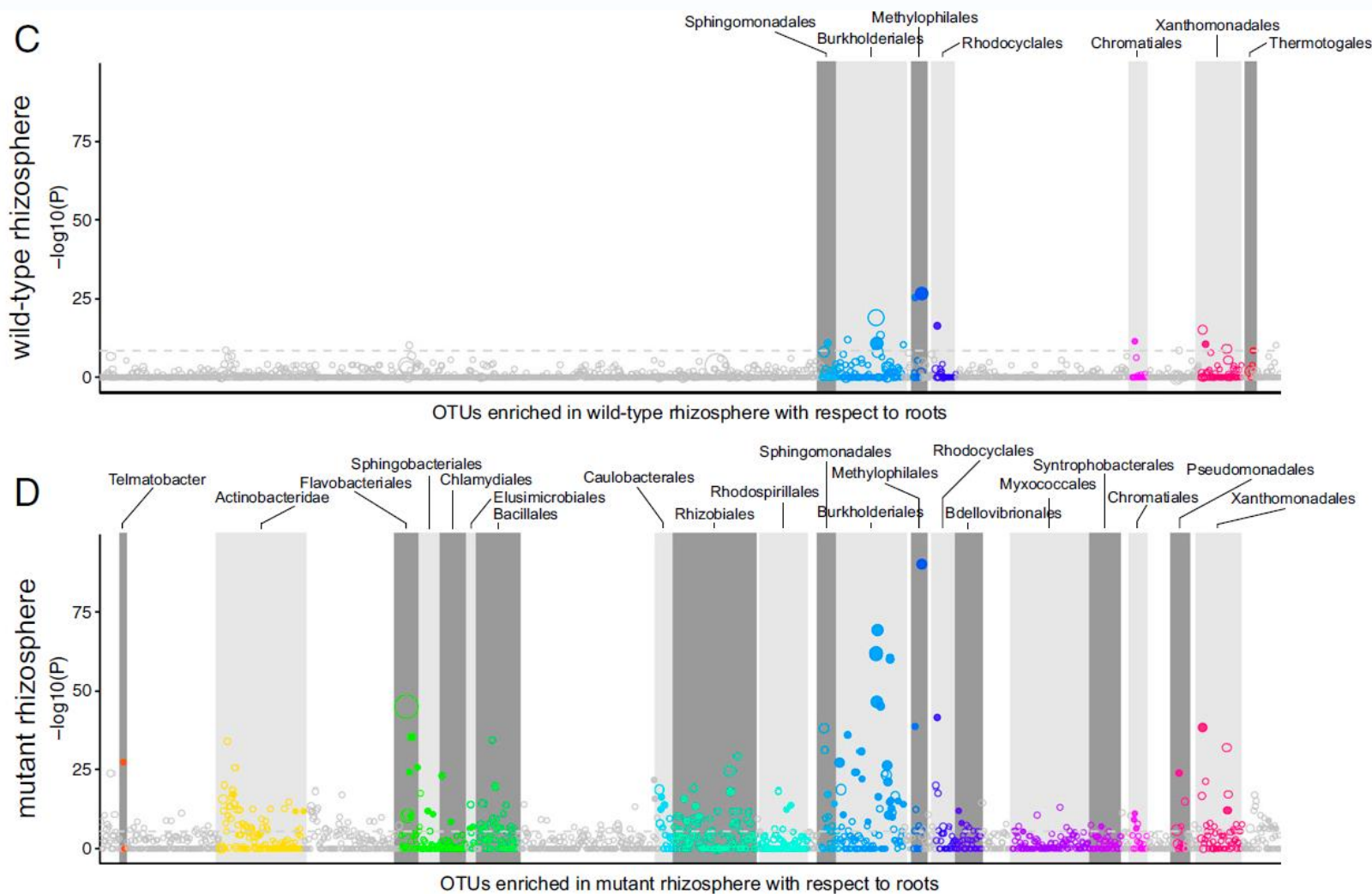
a/b. 曼哈顿图展示地块I(a)和地块II(b)籼粳稻间差异的OTUs。圆形或三角形代表OTUs，籼稻显著富集的为实心上三角，粳稻显著富集的为空心下三角，采用Wilcoxon秩和检验，阈值为FDR校正的P值 < 0.05 。OTUs在图中按物种注释字母顺序排列，按门和变形菌纲着色。



• [NBT封面：水稻NRT1.1B基因调控根系微生物组参与氮利用](#)



曼哈顿图展示变化OTUs/ASVs分布特点-不同



宏基因组

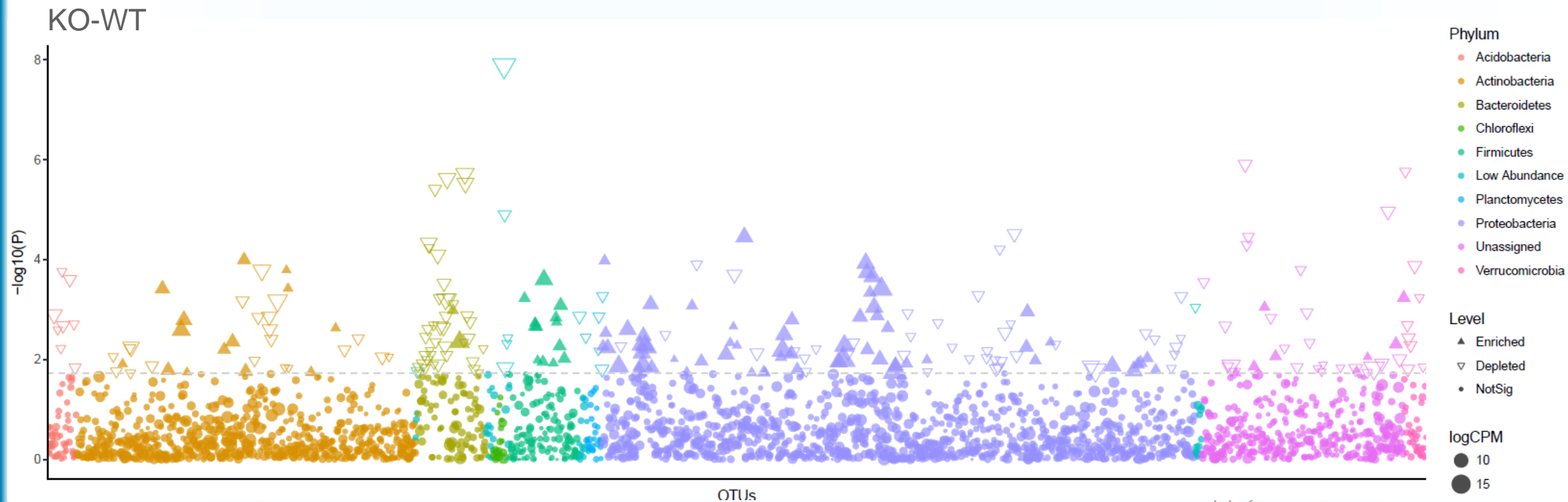
测序

易生信

Zgadzaj, R., **Garrido-Oter, R.**, Jensen, D.B., Koprivova, A., Schulze-Lefert, P. and Radutoiu, S., 2016. 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), pp.E7996-E8005.



Rmd绘制的曼哈顿图结果展示和解读



默认生成曼哈顿图: KO-WT_manhattan.pdf

Rmd绘制中ggplot绘制曼哈顿图的核心代码

```
p = ggplot(x, aes(x=otu, y=neglogp, color=Phylum, size=logCPM, shape=Level)) +  
  geom_point(alpha=.7) +  
  geom_hline(yintercept=FDR, linetype=2, color="lightgrey") +  
  scale_shape_manual(values=c(17, 25, 20))+  
  scale_size(breaks=c(5, 10, 15)) +  
  labs(x="OTUs/ASVs", y="-log10(P)", title=paste(group_list[1], "vs", group_list[1], sep=" ")) +  
  theme_classic() +  
  theme(axis.ticks.x=element_blank(),axis.text.x=element_blank(),legend.position="right")
```

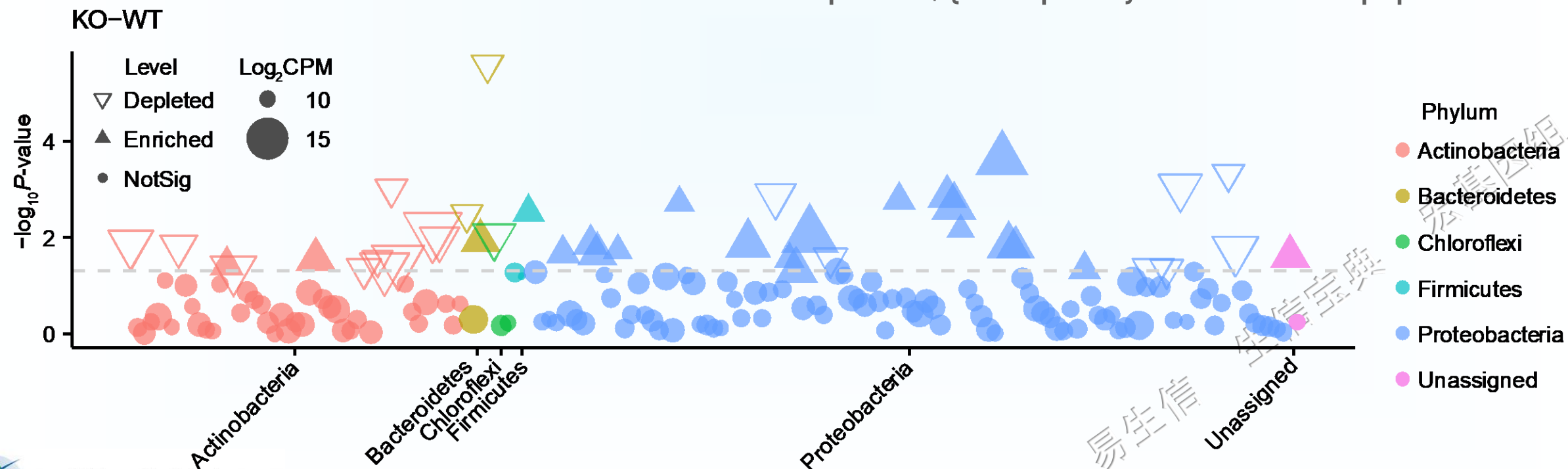
作业：大家说说每部分分别对应图中的什么？
可回顾前面的知识，R中帮助和网络搜索



命令行绘制曼哈顿图，按门水平着色

- # i差异比较结果,t物种注释,p图例,w宽,v高,s字号,l图例最大值

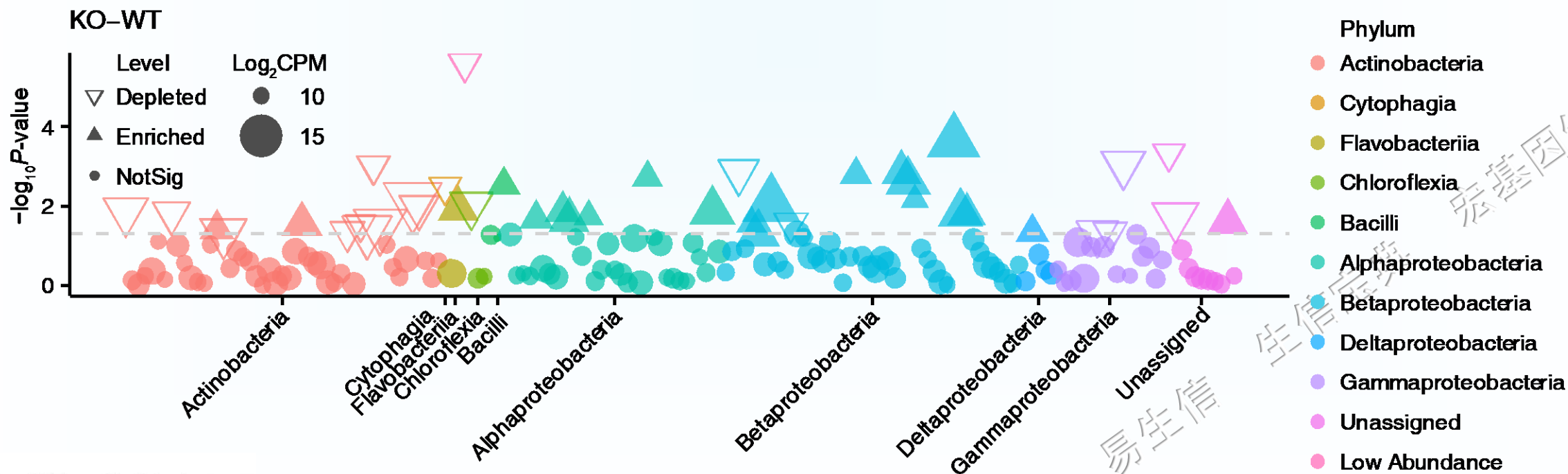
```
bash ${db}/script/compare_manhattan.sh -i result/compare/${compare}.txt \  
-t result/taxonomy.txt -p result/tax/sum_p.txt \  
-w 183 -v 59 -s 7 -l 10 -o result/compare/${compare}.manhattan.p.pdf
```



命令行绘制曼哈顿图，按纲水平着色

- # 上图只有6个门，切换为纲sum_c.txt和-L Class展示细节

```
bash ${db}/script/compare_manhattan.sh -i result/compare/${compare}.txt \
-t result/taxonomy.txt -p result/tax/sum_c.txt \
-w 183 -v 59 -s 7 -l 10 -L Class -o result/compare/${compare}.manhattan.c.pdf
```



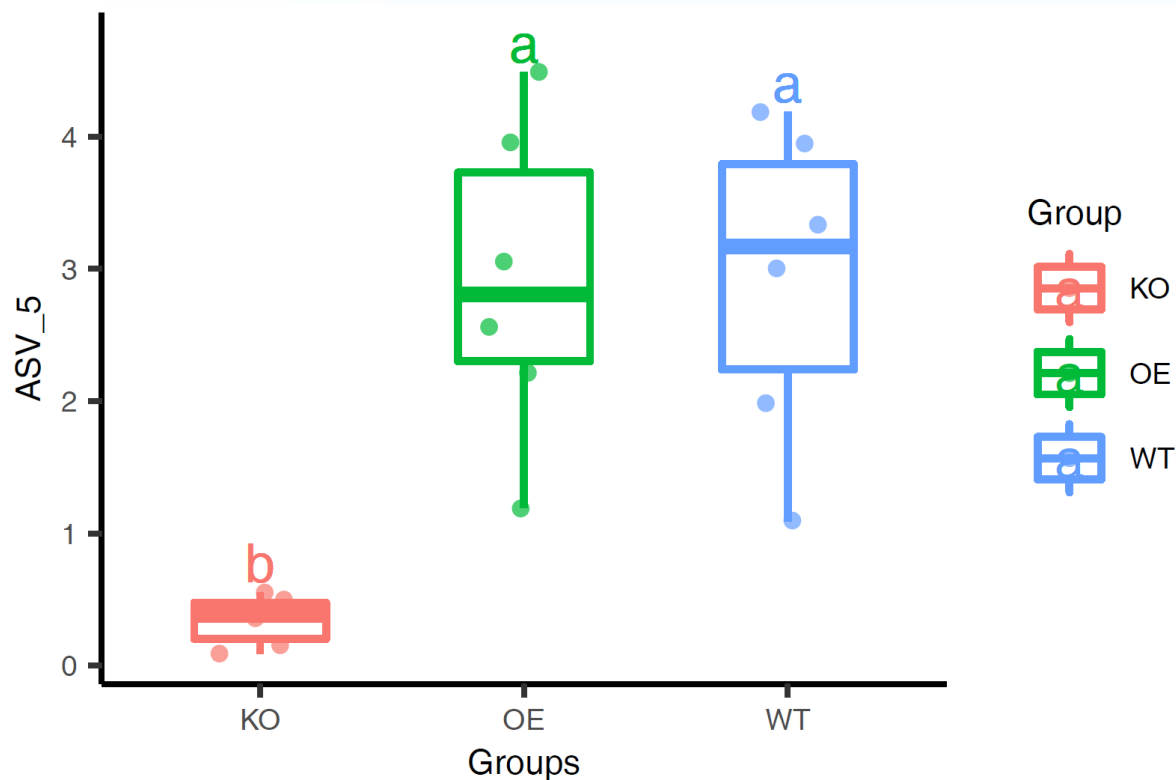
曼哈顿进一步学习：

- 理论讲解：
- 4曼哈顿图：差异OTU或Taxonomy
- 实战代码：
- 4曼哈顿图：差异OTU或Taxonomy

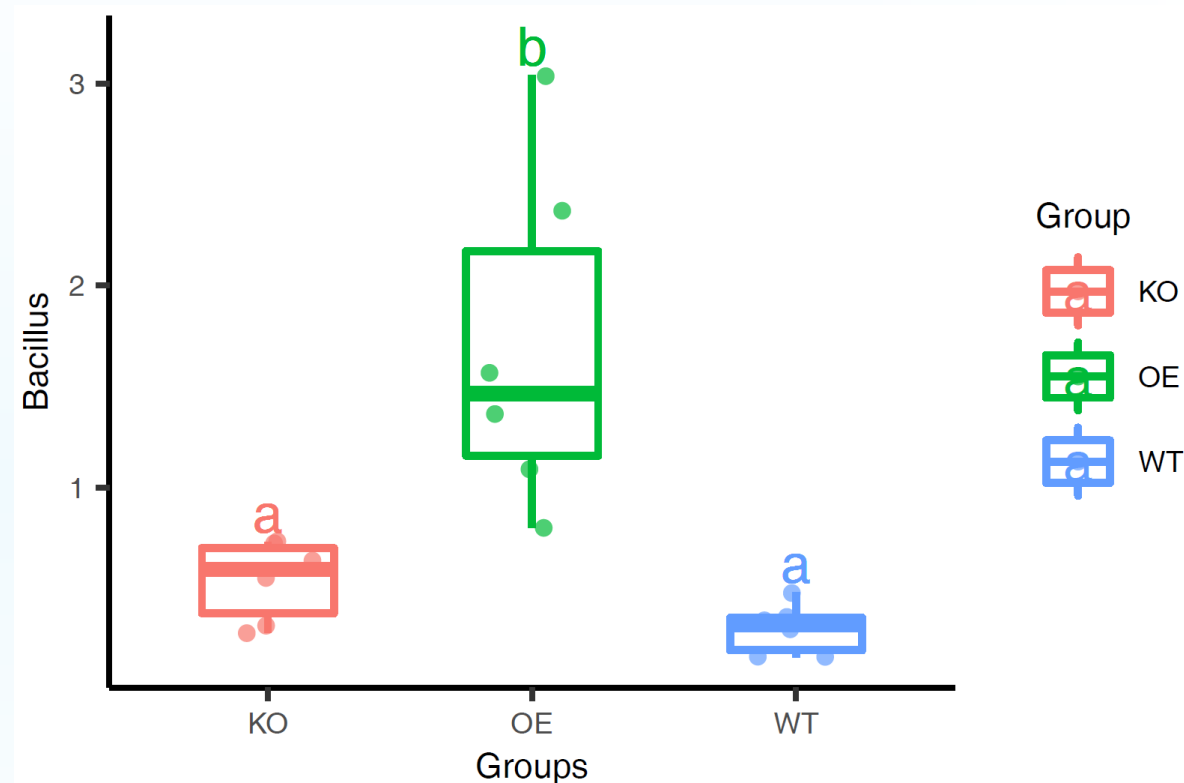
易生信 生信宝典 宏基因组



单个ASV或属展示



```
Rscript ${db}/script/alpha_boxplot.R --alpha_index ASV_5 \
--input result/otutab.txt --design result/metadata.txt \
--transpose TRUE --scale TRUE \
--width 89 --height 59 \
--group Group --output result/compare/feature_
```



```
Rscript ${db}/script/alpha_boxplot.R --alpha_index Bacillus \
--input result/tax/sum_g.txt --design result/metadata.txt \
--transpose TRUE \
--width 89 --height 59 \
--group Group --output result/compare/feature_
```

不同批次, ASV编号会变, 有时不存在ASV_5。
不同版本的数据库注释结果也不同, 新版RDP18注释没有Bacillus。

如果想用edgeR比较高分类级， 汇总计数型值

- 要根据 otutab.txt 、 taxonomy.txt 和 taxonomy_summary_count.Rmd 制作界、门、纲、目、科、属、种的汇总值，原理同23STAMP节类似

taxonomy_summary_count.Rmd,
结果文件：count*, 9个新矩阵文件。使用compare.Rmd修改otutab.txt为如下文件，并删除taxonomy.txt文件即可。

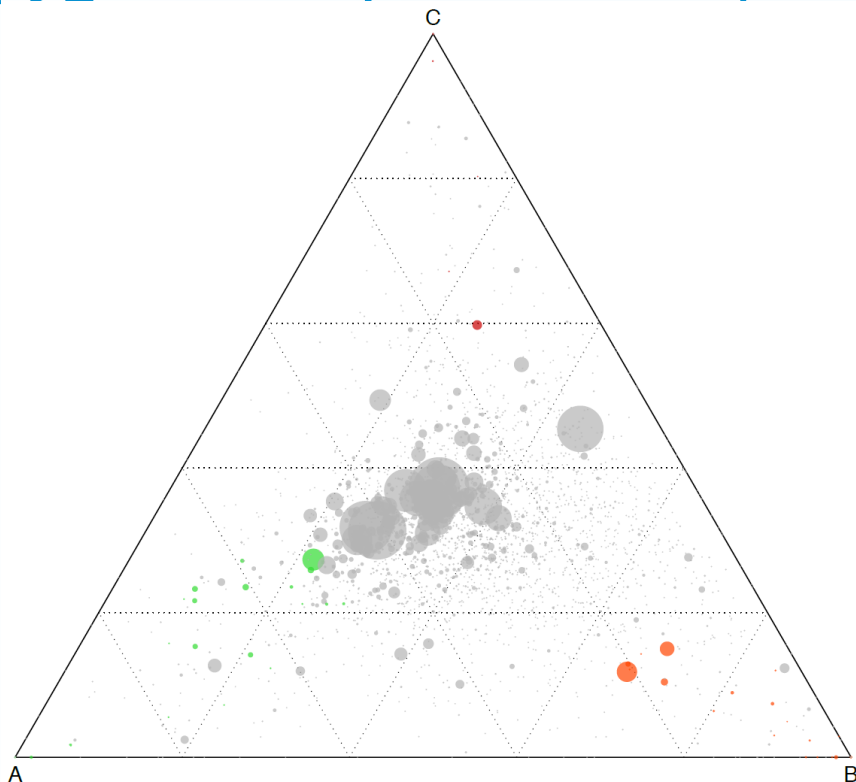
415 count_0LEfSe.txt
2 count_1Kingdom.txt
16 count_2Phylum.txt
33 count_3Class.txt
53 count_4Order.txt

111 count_5Family.txt
205 count_6Genus.txt
205 count_7Species.txt
1115 count_8OTU0.01.txt

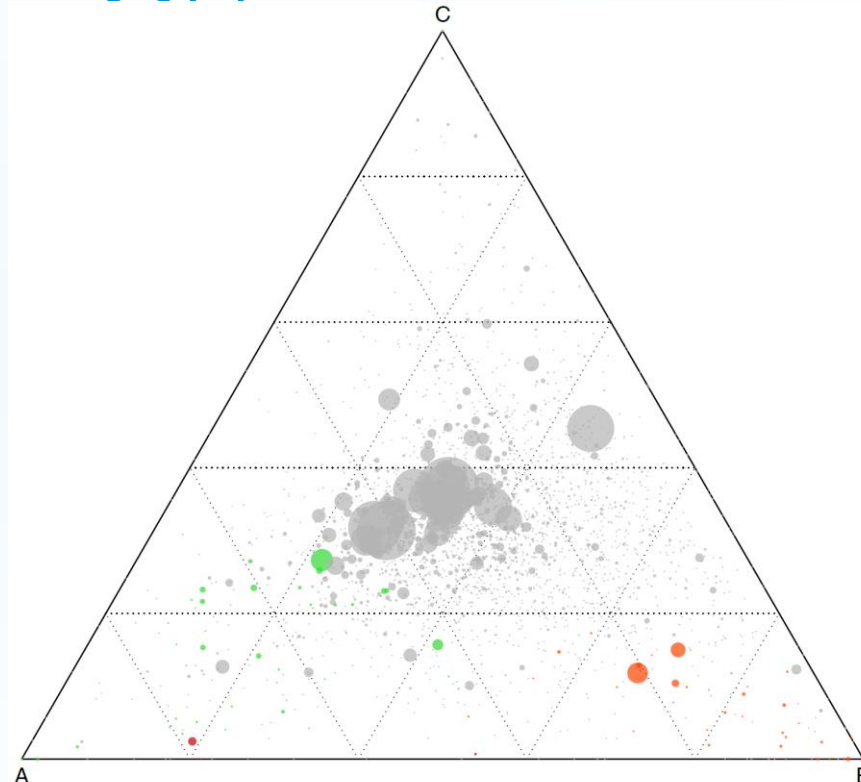


此外还有三元图示例代码参考

- 详见 [24Compare\ternary\ternary.Rmd](#) 文档



展示3组中各自相对于另外两组特征



展示下面两组特有和共有的特征，顶部为对照

- [246.三元图的应用与绘图实战](#)
- [扩增子 图片解读7三元图 统计绘图7三元图](#)

- R语言差异比较的统计方法：常用t检验(t.test)、秩和检验(Wilcoxon test)或基于负二项分布的edgeR/DESeq2等；秩和检验没有前提假设无错但统计功效低；基于负二项分布的方法前提为大部分不变，在差异小时更敏感，差异大时不适用；
- 火山图：整体上展示差异特征的分布，以及显著差异的数量；
- 热图：可展示微生物样本和组特征细节，常用组均值展示减少波动；
- 曼哈顿图：展示特征丰度、组间差异类型、分类学分布规律等；
- 三元图：仅三组可用有较大局限性，但图片颜值高，需根据数据实际情况筛选和优化参数。

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

