物种丰度热图

WuYiLei

2020/6/6

Table of Contents

## 简介

热图是使用颜色来展示数值矩阵的图形，图中每一个小方格都是一个数值，按照预设的颜色值对应不同的数值。通常还会结合行、列的聚类分析，以表达实验数据多方面的结果。

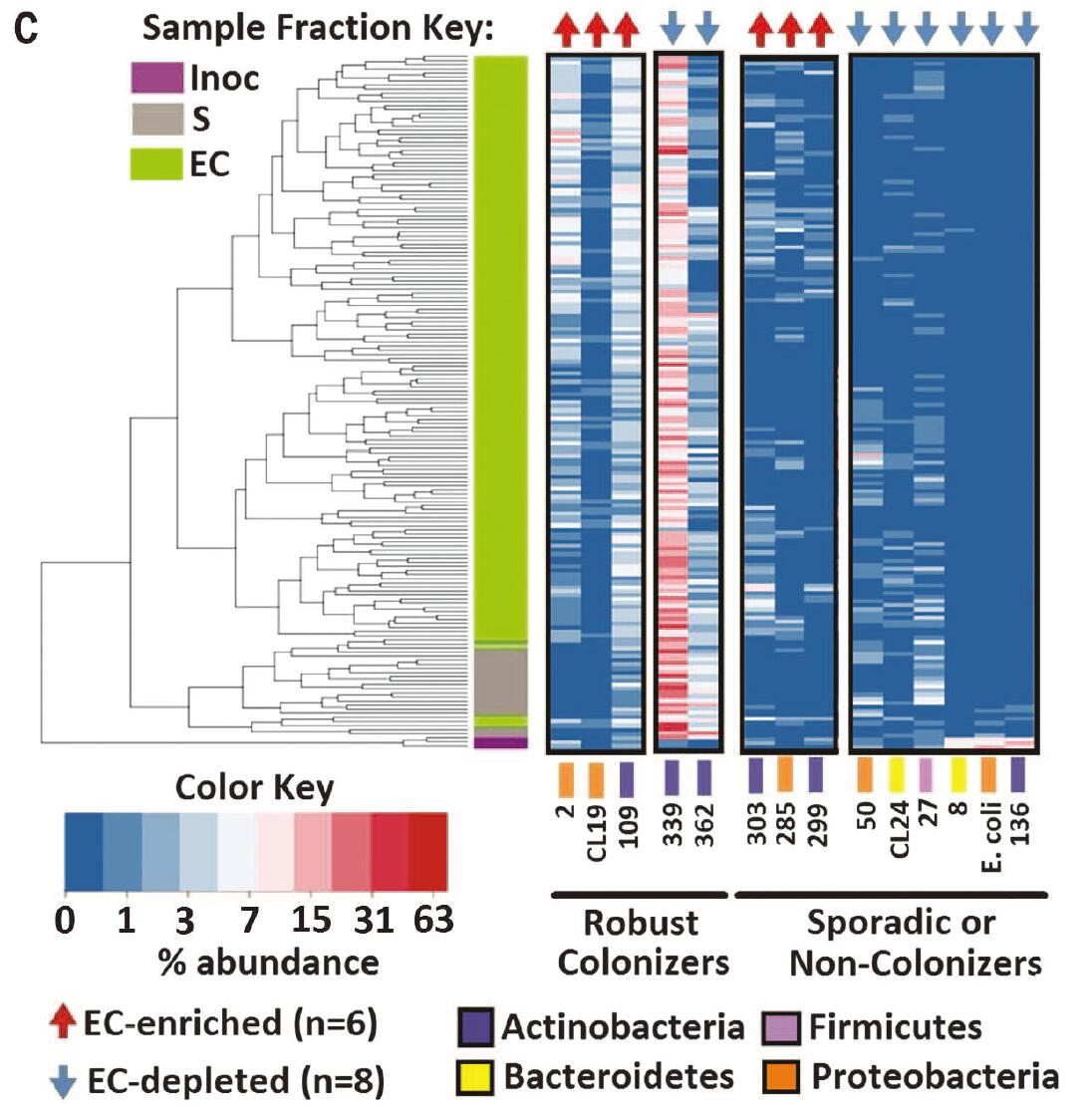
热图在生物学领域应该广泛，尤其在高通量测序的结果展示中很流行，如样品-基因表达，样品-OTU相对丰度矩阵非常适合采用热图呈现。

在16s rDNA下游分析中，一般根据所有样本在属水平的物种注释及丰度信息，选取丰度排名前35的种属，从物种和样本两个层面进行聚类并绘制成热图，便于发现哪些物种在哪些样本中聚集较多或含量较低。

由于阅读数字时需要思考和比较，无法形成大范围的感官印象，而热图采用颜色的深浅代替数据表使得很多规律性的结果更加明显，而且热图在非常小的区域展示了大量的基因表达/细菌丰度数据，既可以快速比较组间的变化，同时还可以显示组内每个样品的的丰度，以及组内各样品间的重复情况,便于从中挖掘规律。结合聚类结果，使得整个实验的结果更加清晰和易于解释1。

补充热图原理：热图示意图，图形中元素的含义，相关变体等

## 实例解读

[Lebeis, S. L., et al. (2015)](https://science.sciencemag.org/content/349/6250/860)2是最早植物人工重组菌群的文章，研究了植物水杨酸对微生物组的影响，开山之作值得阅读。 

图C. 热图展示丰度显著差异的菌在所有样品中的相对丰度(相对丰度百分数%经log2对数变换)

1. 图中元素解读
   * 左侧聚类图为所有样品聚类的结果，左上角的图例代表三大类样品，紫、灰和绿它们分别代表接种菌、土壤和根样品，颜色标签在热图中第一列，用以区分样品组；
   * 右侧为图的主图区，展示左侧样品中对应筛选的14个差异丰度菌的相对丰度值，丰度值百分比采用log2转换来缩小数据范围，并按从小到大对应的颜色梯度为蓝、白、红，即越红越高，越蓝越低。对应的图例为下方左上角的Color Key;
   * 右侧正文区上方红上或蓝下箭头，代表这些菌的表达样式，为上调或下调，对应的图例为下方图例区的左下方(EC-enriched/depleted)；
   * 右侧正文区下方菌的标签上还有颜色，对应最下面图例区的菌门信息；同时菌还继续分为两类，稳定定殖者(Robust Colonizers)和偶然或非定殖者(Sporadic or Non-Colonizers)。
2. 图表结果：图中展示了人工重组的菌在接种后，也可以形成丰度各异的微生物群体，并与自然条件下很多样式保持一致。
3. 图表结论或规律：受水杨酸调控差异表达的菌，可以在人工重组实验中得到验证。
4. 图片优点：配色采用红白蓝，比较严肃；图中添加了聚类信息、分组信息和菌分类为信息，极大的增加了图片的信息和可读性。有些热图信息量大，标签太小或根本无标签导致理解困难，此图的做法值得学习。
5. 实例解读需要中英文对照，其次解读部分需要补充文中其他部分的描述信息，例如结果部分描述信息，中英文对照。
6. 实例数量不够，需要至少两个例子，并且尽量为最新文献，热图在文章中的使用相当多，要求使用最近一年内的高水平文献做解读。

## 绘图实战

热图代码可以跑通，需要使用刘老师示例数据，也就是真实数据再演示一遍。图片出现问题重叠，坐标轴丢失现象，需要修改。

建议：

热图一般挑选的差异OTU进行展示，建议补充差异OTU挑选部分？

热图在不同分类层次上都有使用，纵坐标标记为OTU，似乎没有直接标记门类信息更加直观，例如属水平注释信息。

一般很坐标为样本，但是如今扩增子测序样本数重复量非常多，为了读者更好展示，建议将每个分组测的的重复合并，并使用样本分组作为横坐标。

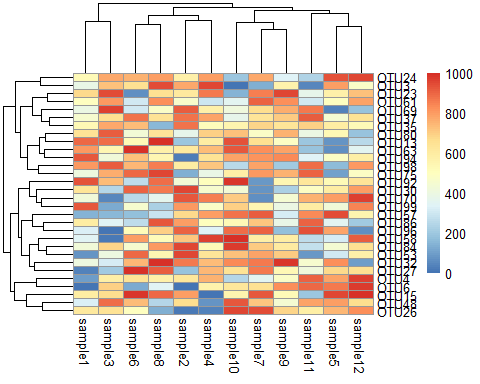
### pheatmap 主要参数

pheatmap 包似乎只有pheatmap 一个函数，应该是用grid底层系统构建的，因此可以利用grid系统的相关函数进一步添加组分。由于只有一个函数，所有作者将所有的参数都压缩到了一起，可以通过?pheatmap查看，常用的有：

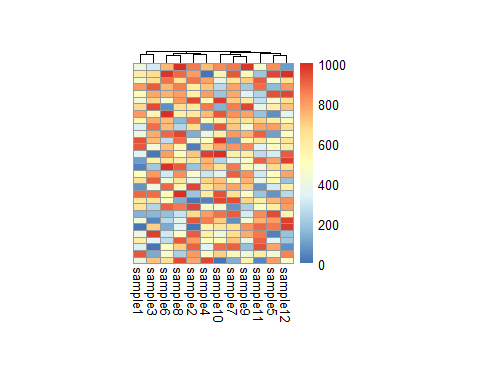
* mat：用于可视化的数据矩阵
* color: 配色要求
* cellwidth/cellheight: 矩形色块的规格
* treeheight\_row/col：聚类树的规格
* cluster\_rows/cluster\_cols：是否对或列聚类(TRUE/FALSE)
* clustering\_distance\_rows/cols：聚类时使用的距离类型，和dist()函数相同
* clustering\_method：聚类方法，与hclust 相同
* show\_colnames/rownames: 是否显示行/列名(TRUE/FALSE)
* annotation\_row/col：给行或列添加注释信息

### 基础热图

# load package  
library(pheatmap)  
  
# set sample data  
set.seed(13)  
otu <- matrix(sample(c(0:1000), 1200, replace = TRUE),   
 ncol = 12, nrow = 100,   
 dimnames = list(row\_names = paste0("OTU",seq(1:100)),   
 col\_names = paste0("sample",seq(1:12))))  
  
# head(otu)  
  
# 按每个OTU的总丰度排序  
otu <- as.data.frame(otu)  
otu$sum <- rowSums(otu)  
otu\_order <- otu[order(otu$sum, decreasing = TRUE), ]  
# head(otu\_order)  
  
# 取丰度前30的OTUs  
mat <- otu\_order[1:30, -ncol(otu)]  
  
# create heatmap using pheatmap  
pheatmap(mat = mat)



# 比较完整的参数  
pheatmap(mat = mat,   
 border\_color = 'grey60',   
 treeheight\_col = 5,  
 cellwidth=10, cellheight=5,   
 cluster\_row = F, cluster\_col= T,   
 show\_rownames = F, show\_colnames = T,  
 clustering\_method = "complete")

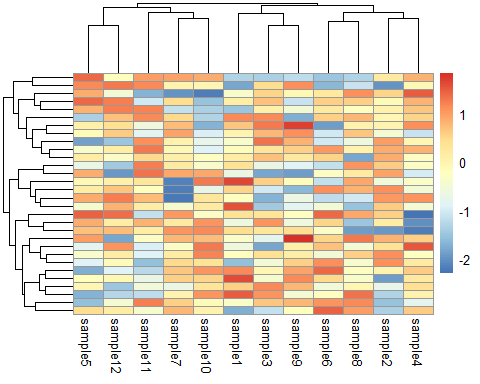


### 数据转换(归一化/标准化)

如果使用原始相对丰度或表达值，范围通常为0-100或0-1000000，而大部分的OTU或基因较低，做出的图会使绝大数据的数量颜色处于低丰度区，很难发现规律；因此需要数据变换，常用的方法有两类：

1. log2(x+1) x为丰度或表达值
   * 给原始值+1是为了保证结果仍为正值，因为2的0次方为1；
   * 为什么要使用log变换，以log2为例，0-1000的表达范围，经变化为0-10的范围，颜色梯度范围更容易使人与数值建立对应关系。
   * 为什么常用log2对数变化，因为筛选差异的标准通常为两倍，log2对数变化后，每相差1的两个值都有两倍差异，选择目标很方便；有时也会根据具体情况，选择ln， log10等转换方式；
2. Z-score标准化：标准分数（standard score）也叫z分数（z-score）,是一个分数与平均数的差再除以标准差的过程。用公式表示为：z=(x-μ)/σ。其中x为某一具体分数，μ为平均数，σ为标准差。此种方法可以使有差异且稳定变化的两组明显区分为不同的颜色，但却丢失了原始相对丰度、差异倍数的信息。但由于结果比较美观，规律明显，使用较多。

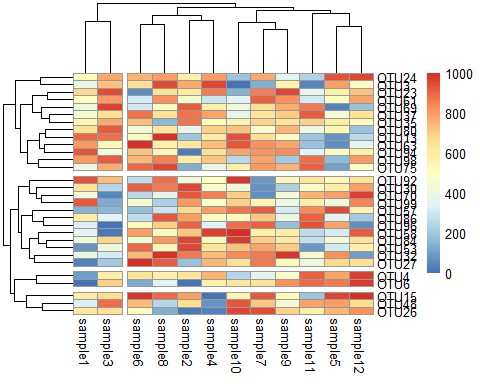
# log2 转换  
# scale\_test <- apply(mat, 2, function(x) log2(x+1))  
  
  
# scale 转换  
scale\_test <- apply(mat, 2, scale)  
pheatmap(mat = scale\_test)



### 根据聚类结果分割热图

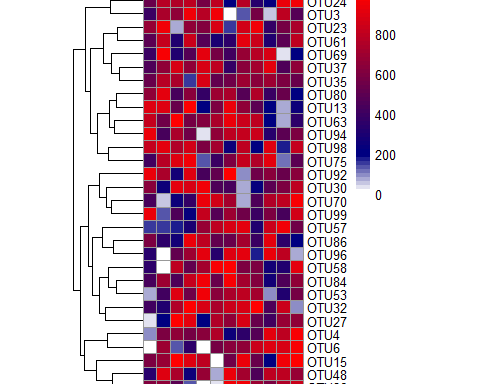
可以依据聚类簇将热图板块分开，这样我们就可以在热图主体中直接获得不同聚类簇的信息，而不会分心去查看聚类情况，在大量数据聚集在一起的时候，非常好用。

pheatmap(mat = mat, cutree\_rows = 4, cutree\_cols = 2)



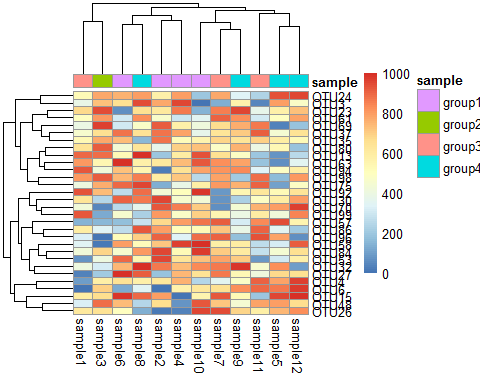
### 热图色键分配

# color palette and cell size set  
# divide the data range into several parts, say 50   
# set color palette according to the divided data range  
# 利用colorRampPalette 进行色键分割，参考?colorRampPalette  
# 要求 col2rgb()  
first <- colorRampPalette(c("white", "navy"))(10)  
second <- colorRampPalette(c("navy", "red"))(40)  
palette <- c(first, second)  
  
pheatmap(mat = mat,   
 color = palette,   
 border\_color = 'grey60', cellwidth=10,   
 cellheight=10, cluster\_row = T,   
 cluster\_col= T, show\_colnames = T,   
 clustering\_method = "complete")



### 添加注释信息

# add annotation to heatmap   
  
# add annotation data out of scratch  
set.seed(13)  
# 随机生成的分组x信息  
sample <- paste0("group",   
 sample(seq(1:4), 12, replace = TRUE ))  
  
# 如果分组信息是一个表格，则用merge()依据sample 合并即可  
annot\_data <- data.frame(row.names = colnames(mat),   
 sample)  
# head(annot\_data)  
pheatmap(mat = mat, annotation\_col = annot\_data)



## Reference

1. 刘永鑫. 扩增子图表解读3热图：差异菌、OTU及功能.

2. Lebeis, S. L. *et al.* Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. **349**, 860–864 (2015).