**物种差异统计分析工具——ALDEx2**

物种差异分析是指通过统计软件查找组间有统计学差异的物种，它是菌群研究的重要内容之一。物种丰度数据具有二个显著的特点，第一是其具有稀疏性的特点，即丰度表中含有大量的0值，而且数据离散性也很明显；第二是丰度表中的数值（reads数）受文库制备和测序深度的影响，无法代表实际环境中的绝对值。这些特点给差异统计带来了巨大挑战。

首先，由于丰度表具有稀疏性的特点，数据不可能服从正态分布，甚至不服从泊松分布，这种数据如果不进行预先处理，是无法用传统的t检验和方差分析等统计方法进行检验的，否则很容易产生I类错误（两组间本无差异，统计出来确有差异的现象叫I类错误）[1]。统计学家常用二种方法来解决丰度表的统计问题，一是对数据进行转换然后再统计，二是采用负二项分布模型进行统计。本文主要介绍第一种，这也是Rob Knight团队最认可的方法。

第二个挑战来自丰度表中的reads数所代表的意义。丰度表中的reads数值除了受样本中实际物种丰度的影响外，还受到测序深度的影响，所以丰度表中的reads数并不是环境中实际物种丰度。我们无法使用丰度表中的数据直接用于样本间和组间差异比较。就因为这个原因，在进行差异统计前必须对丰度表中的数据进行normalization，也叫标准化或归一化，否则统计结果根本就不可信。常用的标准化方法有百分比化、抽平至100万、稀疏化（rarefying）和对数转换等[2]。百分比化就是把每个样本数的reads先求百分比，使每个样本的reads合计数为1；抽平至100万方法类似于百分比化，只不过reads值没有小数点了，每个样本reads合计数为100万；对数转换是目前最流行的方法，具体方法是先用很小的数值替换丰度表中的0值，然后在求对数[2]。百分比化和抽平至百万的方法会产生明显的数据扭曲，产生负相关偏倚（negative correlation bias），导致相关或基于相关分析的方法均无效，这些方法包括PCA、相关网络等[3]，所以如果是用于物种差异分析，这二种标准化方法均不可取。下图中间的小图就是百分比转换，可以看出数据出现扭曲。对数转换是目前认为比较靠谱的方法之一，也是Rob Knight团队最推崇的方法，下文介绍的ALDEx2方法就是典型的对数转换方法之一。

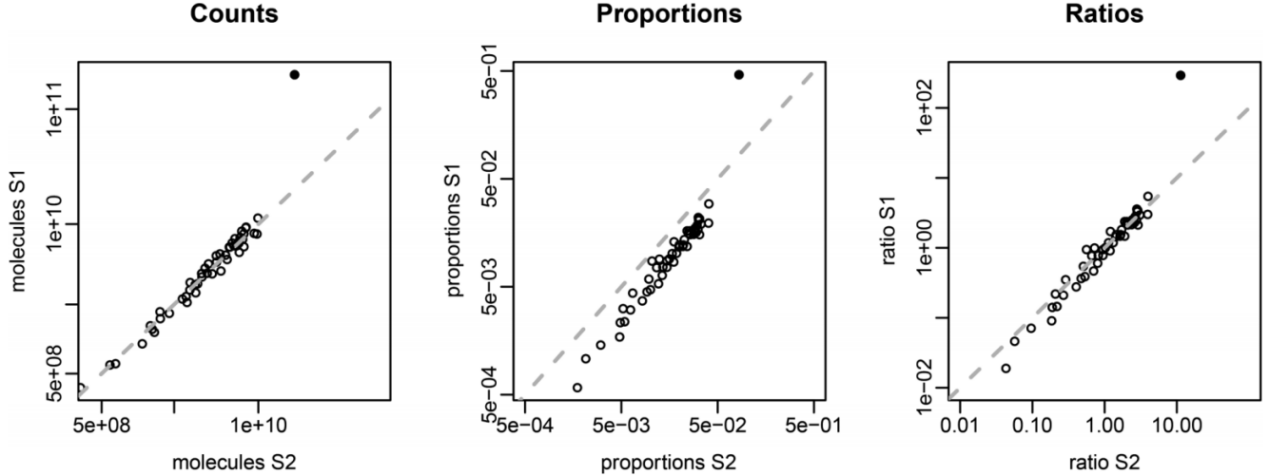


图1：不同标准化方法对数据的影响。

下面详细介绍ALDEx2用于物种差异分析的代码及注意事项。

**一、加载包和数据初步处理**

本文的演示数据为Qiime 2教程中的粪菌移植演练数据，后台回复ALDEx2即可获得演示数据和R代码下载方式。ALDEx2需要的数据必须是行名为OTU，列名为样本名。

###加载包

library(ALDEx2)

library(ggplot2)

###读入数据，将数据转置为行名为OTU名，列名为样本名

mydata <- read.csv('fmt.csv', skip=1, row.names=1, head=T)

mydata[1:10, 1:20]

otu <- data.frame(t(mydata)) #ALDEx2必须行名为OTU，列名为样本名

otu[1:10, 1:20]

###分组文件

group <- substring(colnames(otu), 1, 1) #截取列名前1个字母作为组别名

**二、统计过程**

set.seed(1000)

otu.log <- aldex.clr(otu, group, mc.samples=128, verbose=TRUE) #log转换，输入文件行必须是OTU，列是样本名

otu.test <- aldex.ttest(otu.log, group, paired.test=FALSE) #Welch和Wilcoxon检验

head(otu.test) #we.ep为welch的P值，we.eBH为校正P值；wi.ep为WilcoxonP值，wi.eBH为校正P值

###计算Effect Size

otu.effect <- aldex.effect(otu.log, include.sample.summary=FALSE, verbose=FALSE)

**三、查看统计结果**

###将P值文件和效应文件合成一个文件

otu.p.all <- data.frame(otu.test, otu.effect)

head(otu.p.all)

###筛选出P值＜0.05的数据

subset(otu.p.all, otu.p.all$we.ep < 0.05) #Welch检验P值＜0.05

subset(otu.p.all, otu.p.all$we.eBH < 0.05) #Welch检验校正P值＜0.05

subset(otu.p.all, otu.p.all$wi.ep < 0.05) #Wilcoxon检验P值＜0.05

subset(otu.p.all, otu.p.all$wi.eBH < 0.05) #Wilcoxon检验校正P值＜0.05

subset(otu.p.all, otu.p.all$effect < -0.5) #查看有哪些otu的effect＜-0.5

subset(otu.p.all, otu.p.all$effect > 0.5) #查看有哪些otu的effect＜-0.5

**四、可视化**

将效应值作为X轴，校正后的Wilcoxon检验P值作为Y轴绘制散点图。

otu.p.all$sig.otu <- ifelse(otu.p.all$wi.eBH < 0.05, "P ＜ 0.05", "P ＞ 0.05") #把Wilcoxon校正P值<0.5的标记为”P ＜ 0.05“，其余标记为”>0.5“

ggplot(data=otu.p.all, aes(x=effect, y=-log10(wi.eBH), color=sig.otu)) + geom\_point(size=2, alpha=0.5) +

scale\_x\_continuous(name='Effect size', limits=c(-0.9, 0.9)) +

scale\_y\_continuous(name='Corrected P values (-log10)', limits=c(0,4.8), breaks=c(seq(0,4.8,1))) +

geom\_vline(xintercept=c(-0.5, 0.5), linetype=2, size=1, color='grey') +

scale\_color\_manual(name="", values=c('red', 'black')) +

theme(plot.margin=unit(c(3,10,5,5), "mm"), #上、右、下、左

panel.background=element\_rect(fill="white"), #背景

axis.line=element\_line(color='black'), #坐标轴

legend.position=c(0.95, 0.92), #图例位置

legend.title=element\_text(color='black', size=13), #图例名称文字大小

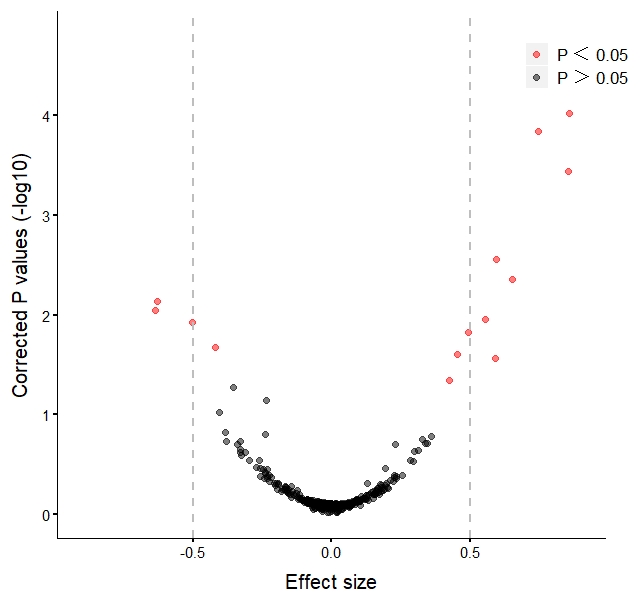
legend.text=element\_text(color='black', size=13), #图标文字大小

axis.ticks=element\_line(color='black', size=1), #坐标轴刻度大小

axis.text=element\_text(color='black', size=11), #坐标轴刻度文字大小

axis.title.x=element\_text(size=15, vjust=-1), #X轴标签

axis.title.y=element\_text(size=15, vjust=3)) #Y轴标签



**五、图片解读**

图片X轴代表效应值，小于0.5点是在两组间有差异的物种，可以认为是生物标记。Y轴是经过Benjamini-Hochberg法校正后的P值，红色的点就是校正P值＜0.05的物种。你可以发现，红色的点有部分在-0.5或0.5效应界值虚线里面，这种情况下应该优先按照效应值来判断某物种是否为生物标记，因为效应值的可靠性优于P值[4]。

**六、注意事项和展望**

ALDEx2这种物种差异分析方法最大的优势是其假阳性率比较低，低于EdgeR和DESeq2[2,5]。

ALDEx2的统计结果与Qiime 2教程中的ANCOM比较类似，统计结果很多情况下也一致。但是ALDEx2和ANCOM的方法强行地将0值替换为其他数值，也会导致一定的偏倚，而且它们无法解决0值膨胀的问题。

Rob Knight团队最近在研究差异分析的其他方法，比如用songbird软件，借助谷歌大脑的Tensorflow人工智能方法筛选差异物种，但是目前仍在探索阶段，尚无法用于实战[6]。

参考文献

1. Patricio S La Rosa, J Paul Brooks, Elena Deych, et al. *Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data*. PLoS One, 7 (12), e52078. DOI: 10.1371/journal.pone.0052078
2. Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, ..., Rob Knight. *Normalization and microbial differential abundance strategies depend upon data characteristics*. Microbiome, 5 (1), 27. DOI: 10.1186/s40168-017-0237-y
3. Gregory B Gloor, Gregor Reid. *Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data*. Can J Microbiol, 62 (8), 692-703. DOI: 10.1139/cjm-2015-0821
4. Xia Y, Sun J, Chen D. *Community diversity measures and calculations*. In: Xia Y, Sun J, Chen D, eds. *Statistical Analysis of Microbiome Data with R*. Singapore: Springer Singapore; 2018: 167-190.
5. Thomas P Quinn, Tamsyn M Crowley, Mark F Richardson. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics, 19 (1), 274. DOI: 10.1186/s12859-018-2261-8
6. James T Morton, Clarisse Marotz, Alex Washburne, ..., Rob Knight. *Establishing Microbial Composition Measurement Standards With Reference Frames*. Nat Commun, 10 (1), 2719. DOI: 10.1038/s41467-019-10656-5