**基于微生物成分数据的差异zOTU分析流程**

**Workflow for discriminative zOTU analysis based on compositional data**

焦金真1，张小丽1, 2，王敏1，谭支良1, \*

1中国科学院亚热带农业生态研究所，亚热带农业生态过程重点实验室，畜禽养殖污染控制与资源化技术国家工程实验室，动物营养生理和代谢过程湖南省重点实验室，湖南长沙 410125

2 中国科学院大学，北京，100049

\*通讯作者邮箱: zltan@isa.ac.cn.

**摘要**

近年来，得益于测序技术的革新及成本的大幅度下降，一系列微生物组（microbiome）研究技术，如扩增子、宏基因组和宏转录组等，被用于医学、工农业和环境等各个领域的研究，推动微生物组研究进入了黄金发展期[1,2]。面对如此海量的测序数据，如何从中有效且准确地挖掘信息显得尤为重要。值得注意的是，测序时核酸通常被设定到一定浓度，结果提供的总序列数均为一有限值，使得任一条序列均为相对比例，称为成分数据（compositional data）[3]。本文在Gloor等（2017）[4] 开发的基于成分数据的分析流程的基础上，以山羊空肠细菌16S rRNA测序数据为示例（食靡细菌 vs. 粘膜细菌），在获得试验设计、zOTU丰度表、zOTU物种注释信息的基础上，运用R和R Studio软件，详述基于微生物成分数据的差异zOTU分析流程。zOTU指zero-radius OTU，是通过unosie3命令去燥输出的有效生物学序列，类似于DADA2和Deblur输出的扩增序列变体（amplicon sequence variants，ASV）。

**关键词:** 扩增子测序，成分数据，差异zOTU

**分析策略**

一般而言，微生物16S rRNA扩增子的分析流程是从将测序序列数归一到同一测序深度，即Normalization开始的（表1）。主流的Normalization是将测序序列进行抽平[5]，或者用R的DESeq包对序列数进行均一化[6, 7]。基于成分数据的Normalization是通过centered log-ratio (CLR) 将测序数据转化为比例数据[8]，极大地降低了下游分析受测序深度的影响程度。下一步的分析通常是计算样品之间的β多样性（PCoA），降维出图（Ordination）和多元比较（Multivariate comparison）。主流的方法中这些分析是基于样品间的Bray-Curtis、UniFrac和Jensen-Shannon距离等[4，7]。成分数据的下游分析是基于通过CLR计算的Aitchison距离，之后进行PCA（Principal Component Analysis）降维，这可以最大程度地保证成分数据的降维是通过数据集中变异程度最大的zOTU实现的。主流的和基于成分数据的分析均通过ANOSIM、ADONIS等方式进行多元比较，来统计两组的微生物之间是否存在显著差异。最后，在筛选差异微生物（如差异zOTU）时，与主流流程通过LEfSe、DESeq2等不同的是，基于成分数据的差异zOTU分析是通过R包中的ANCOM和ALDEx2等来实现的[4,7,8]。本流程最大的优势是可以极大程度降低测序深度对分析的影响程度，使得分析结果更准确可信。

表1. 微生物数据分析的主流和成分数据分析流程

|  |  |  |
| --- | --- | --- |
| 步骤 | 主流分析 | 成分数据分析 |
| 均一化 | Rarefaction  DESeq | CLR  ILR  ALR |
| 距离 | Bray-Curtis  UniFrac  Jenson-Shannon | Aitchison |
| 降维 | PCoA  (Abundance) | PCA  (Variance) |
| 多元比较 | ADONIS  ANOSIM | ADONIS  ANOSIM |
| 差异分析 | LEfSe  DESeq2 | ANCOM  ALDEx2 |

**软件**

R (version 3.6.2)

R studio (https://github.com/rstudio/rstudio/)

**输入数据**

试验设计表格：conds.txt

zOTU表格：zotu.txt

zOTU注释信息表格：taxonomy.txt

**实验步骤**

**1. R包安装及加载**

本分析需要用到5个R包：

ALDEx2（version 1.14.0，差异分析）、

CoDaSeq（version 0.99.2，成分数据分析）、

factoextra（version 1.0.6, PCA图）、

ggplot2（version 3.2.1，作图）、

vegan （version 2.5-6，ADONIS分析）。

1.1. R studio 中可以通过菜单安装R包。选择“packages”窗口的“install”进行安装。

注：ALDEx2和CoDaSeq可以分别用以下命令进行安装。

source("https://bioconductor.org/biocLite.R")

biocLite("ALDEx2")

library(devtools)

devtools::install\_github('ggloor/CoDaSeq/CoDaSeq')

1.2. 加载R 包

library (R包名）

**2. 数据的载入**

通过以下三条命令分别载入试验设计表格、zOTU表格、zOTU物种注释信息表格。

conds <- read.delim ("conds.txt", header=T, row.names= , sep="\t")

otu = read.delim ("zotu.txt", header=T, row.names= 1, sep="\t")

tax = read.delim ("taxonomy.txt", header=T, row.names= 1, sep="\t")

注：本流程数据来源于山羊空肠食靡和粘膜细菌16S rRNA扩增子数据，分别测定了其食靡（Digesta， n=6）和粘膜 （Mucosa, n=6）的细菌组成。前期通过usearch的unosie3命令将拼接的Tags在99%的相似度上聚类为zOTU。之后执行SINTAX命令比对了RDP数据库，对zOTU进行了物种注释。本流程旨在筛选空肠不同生态位细菌的差异zOTU。

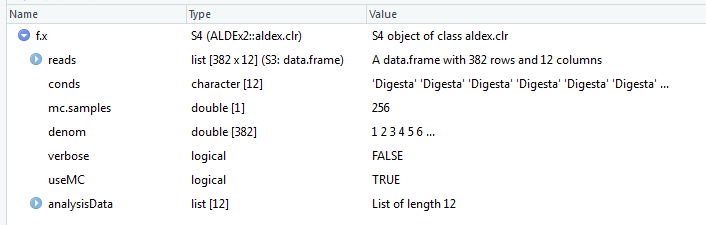
**3.** **运用ALDEx2包的相关命令对zOTU表进行CLR变换，获得每个zOTU的效应量大小和 E（CLR）矩阵**

3.1运用codaSeq.filter命令过滤zotu表格，之后通过aldex.clr进行中心对数（CLR）变换。

f <- codaSeq.filter(zotu, min.reads=5000, min.prop=0.001, min.occurrence = 0.1, samples.by.row=FALSE)

f.x <- aldex.clr(f, as.vector(conds$Region), mc.samples=256, verbose=FALSE, useMC = TRUE)

注：aldex.clr 将数据存储在s4类型变量f.x 对象中，包括参数reads、 conds、mc.samples、denom、verbose、useMD和analysisData等（图1）。$reads为CLR的表达矩阵，conds为样品的分组信息。



**图1.** CLR变换后的f.x的数据结构

3.2 通过aldex.effect命令获得每个zOTU的效应量大小（expected CLR value）。

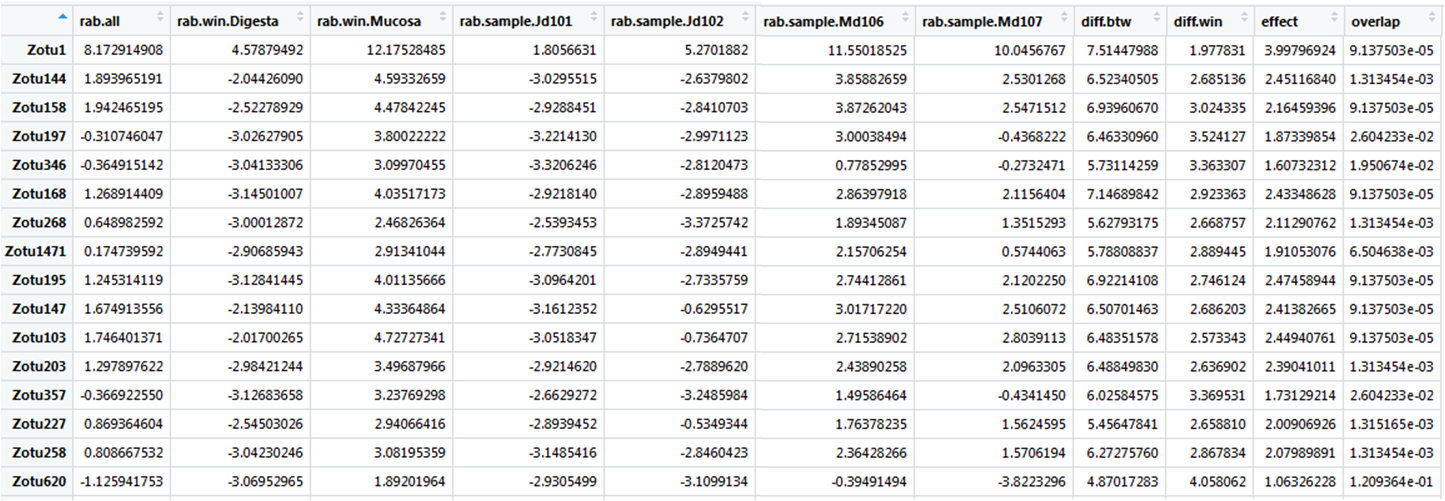
f.e <- aldex.effect(f.x, as.vector(conds$Region),

include.sample.summary=TRUE,

verbose=FALSE,

useMC = TRUE)

注：aldex.effect的输出结果f.e（图2）包括: rab.sample.\*\*\*（zOTU在各个样品的CLR值）；diff.btw （zOTU在两组间的中位数差异）；diff.win ，每个zOTU（zOTU在组内Dirichlet instances的中位数差异）；effect（zOTU的效应量）。effect值可用于后续的差异zOTU筛选标准之一。



**图2.** aldex.effect的输出结果f.e

3.3 获得E（CLR）值的矩阵（E.clr），用于后续的PCA降维和作图。

E.E.clr <- t(f.e[,grep("rab.sample", colnames(f.e))])

rownames(E.E.clr) <- gsub("rab.sample.", "", rownames(E.E.clr))

exp <- apply(E.E.clr, 1, function(x) 2^x)

E.clr <- t(apply(exp, 2, function(x) log2(x) - mean(log2(x)) )

注：E.clr矩阵中行是12个样品，列是382个zOTU的E.clr值。

**4. PCA降维及作图**

4.1 通过prcomp命令进行PCA分析

pcx <- prcomp(E.clr)

4.2 执行factoextra 包的fviz\_pca\_ind命令进行PCA作图

fviz\_pca\_ind(pcx,

geom.ind = c("point", "text"),

fill.ind = conds$Region, col.ind = conds$Region,

pointshape = 21, pointsize = 2,

pallette = "jco",

addEllipses = TRUE, mean.point = FALSE,

ellipse.level = 0.95, ellipse.alpha = 0,

repel = TRUE,

title = "Compositional PCA sample plot Digeta Mucosa",

labelsize = 4,

ggtheme = theme\_minimal()) +

theme(plot.title = element\_text(hjust = 0.5)) +

labs(fill = "Regions") + guides(color=FALSE)

ggsave("Compostional\_PCA\_sampleplot\_Digesta\_Mucosa.jpg", width = 297, height = 210, units = "mm", dpi = 600)

**5. 基于****Aitchison距离进行ADONIS分析**

5.1 获得样品间的Aitchison距离矩阵

dist.clr <- dist(E.clr, method = "euclidean")

注：基于CLR向量的欧氏距离的样品间的比例距离，称为Aitchison距离。

5.2 利用vegan包的adonis2和betadisper命令进行ADONIS分析

adonis.At = adonis2(dist.clr ~ Region, data = conds, contr.unordered = "contr.sum")

adonis.At

adonis.At.disp = betadisper(dist.clr, conds$Region)

adonis.At.disp

permutest(adonis.At.disp, pairwise = TRUE)

TukeyHSD(adonis.At.disp)

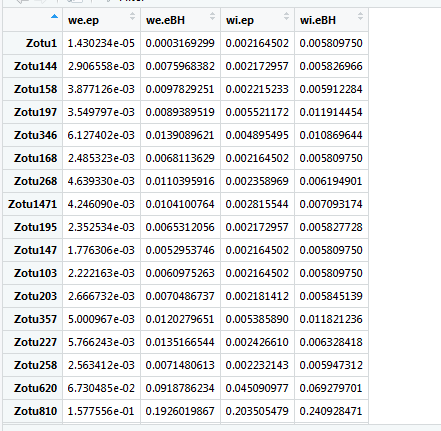
注：函数 adonis2研究两组间的Aitchison距离的平均值的差异，跟多元方差分析的功能相同；而函数 betadisper 研究组内同质性，统计两组内的Aitchison距离是否有差异，类似于Levene’s 等方差检验。

**6. 运用ALDEx2包的相关命令进行差异zOTU分析**

6.1 通过aldex.ttest命令将变换后的值，用参数或非参数检验进行单变量统计检验，并返回Benjamini-Hochberg 校正后的 *P*值。

f.t <- aldex.ttest(f.x, as.vector(conds$Region))

**注：**aldex.ttest输出结果f.t（图4）为矩阵型变量，包括：we.ep (Welch's t 检验的*P*值)；we.eBH ( Welch's t 检验校正后的 *P*值)；wi.ep (秩和检验的 *P*值)；wi.eBH (秩和检验校正后的*P*值)。



**图3.** aldex.ttest的输出结果f.t

6.2 基于effect效应量值和校正后的*P*值进行差异zOTU筛选，获得在Mucosa中富集的zOTU（high.e.Mucosa）和在Digesta中富集的zOTU（high.e.Digesta）。

low.p <- which(f.t$we.eBH < 0.05)

high.e <- which(abs(f.e$effect) > 1)

high.e.Mucosa <- which(f.e$effect >= 1 & f.t$we.eBH < 0.05)

high.e.Digeta <- which(f.e$effect <= -1 & f.t$we.eBH < 0.05)

**注：**本文中差异zOTU要同时满足以下两个标准：Welch's t 检验校正后的 *P*值<0.05；effect值>1 或者effect值<-1。

6.3 差异zOTU的effect 图。

pdf (file = "Discrimitive zOTUs.pdf")

par(mfrow=c(1,2))

plot(f.e$diff.win, f.e$diff.btw,

pch=19, col=rgb(0,0,0,0.3),

cex=0.5, main="effect plot",

xlab="Dispersion", ylab="Difference")

points(f.e$diff.win[high.e],

f.e$diff.btw[high.e],

col=rgb(1,0,0,0.5), cex=0.8)

points(f.e$diff.win[low.p],

f.e$diff.btw[low.p],

pch=19, col=rgb(0,0,1,0.5),

cex=0.5)

abline(0,1, lty=2, lwd=2,col="grey")

abline(0,-1, lty=2, lwd=2,col="grey")

plot(f.e$effect, f.t$we.eBH, log="y",

pch=19, col=rgb(0,0,0,0.3),

main="E vs p", xlab="effect size",

ylab="E(p.adjust)", cex=0.5)

points(f.e$effect[high.e], f.t$we.eBH[high.e],

col=rgb(1,0,0,0.5), cex=0.8)

points(f.e$effect[low.p], f.t$we.eBH[low.p],

pch=19, col=rgb(0,0,1,0.5),cex=0.5)

abline(v=1, lty=2, lwd=2,col="grey")

abline(v=-1, lty=2, lwd=2,col="grey")

abline(h = 0.05, lty=2, lwd=2,col="grey")

dev.off();

6.4 基于zOTU注释信息，获得两组差异zOTU的物种信息表格，并写入csv文件。

high.e.Mucosa.tax <- tax[rownames(f)[high.e.Mucosa],]

high.e.Digesta.tax <- tax[rownames(f)[high.e.Digeta],]

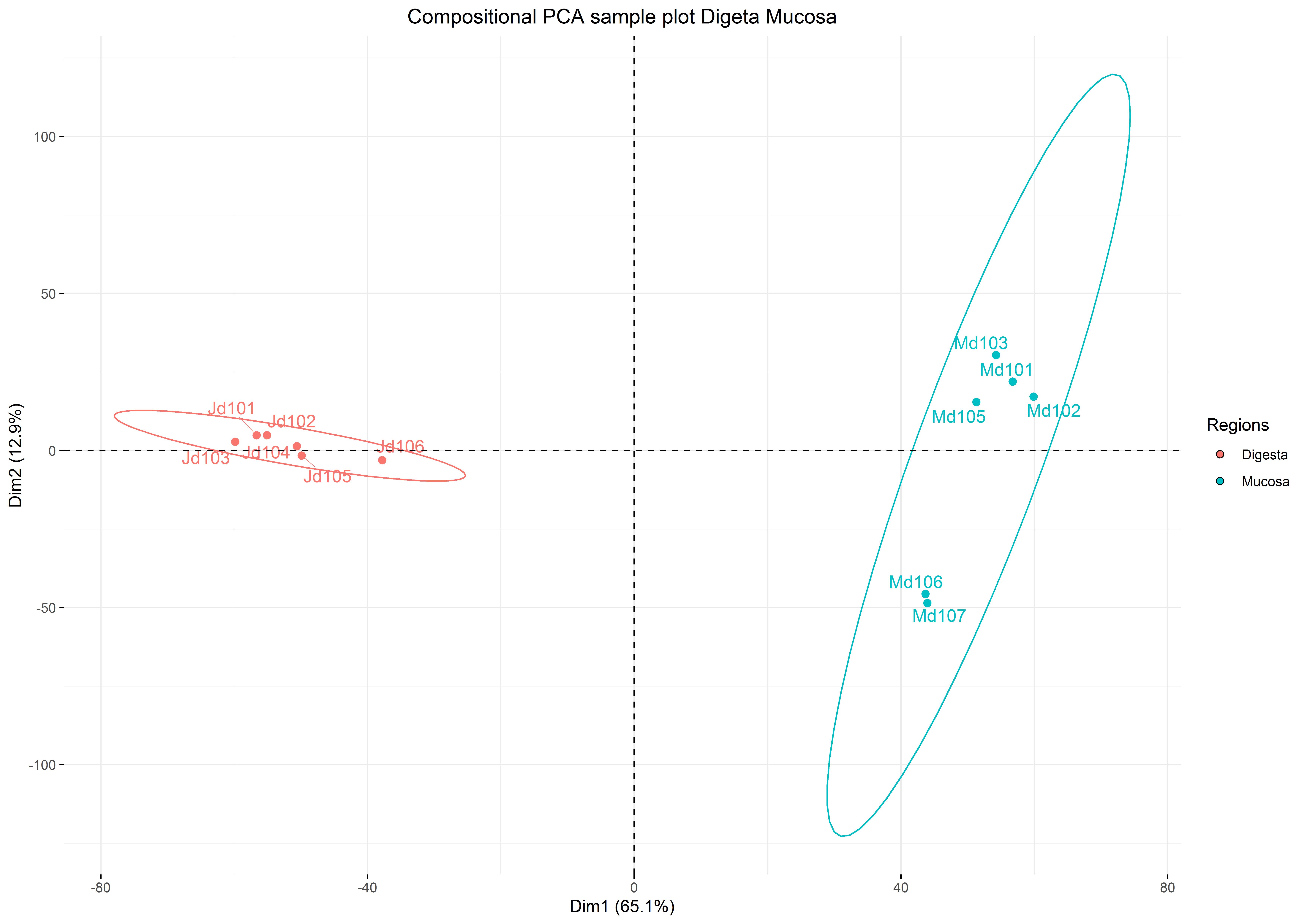
write.table (high.e.Mucosa.tax,"Mucosa enriched OTUs.csv", row.names = T, quote = F, sep=",")

write.table (high.e.Digesta.tax,"Digesta enriched OTUs.csv", row.names = T, quote = F, sep=",")

**结果与分析**

**1. PCA主成分分析图**

如图**4**所示，基于Aitchison距离的PCA图表明，空肠食靡（Digesta）和粘膜（Mucosa）的细菌组成聚成两个相互独立的簇，即：食靡细菌聚成一簇，而粘膜细菌聚成另外一簇。坐标轴中PC1/2的数值为总体差异的解释率分别为65.1%和12.9%。

****

**图4.** 基于Aitchison距离的PCA图

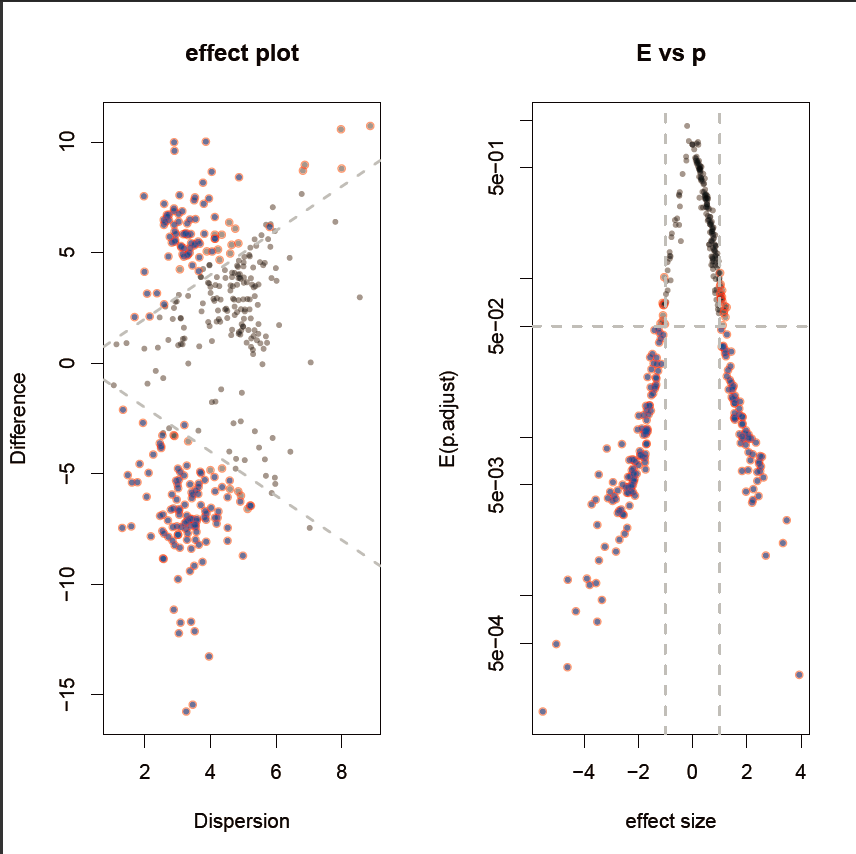
**2. ADONIS分析结果**

ADONIS分析是一种判断不同组的同质性和差异的统计方法。本例中adonis2函数的*P* = 0.002，表明山羊空肠食靡和粘膜细菌组成显著不同。betadisper函数的*P* = 0.026，提示粘膜细菌的组间差异显著高于食靡细菌的组间差异。

**3. 差异zOTUs**

3.1 差异zOTUs 的effect图

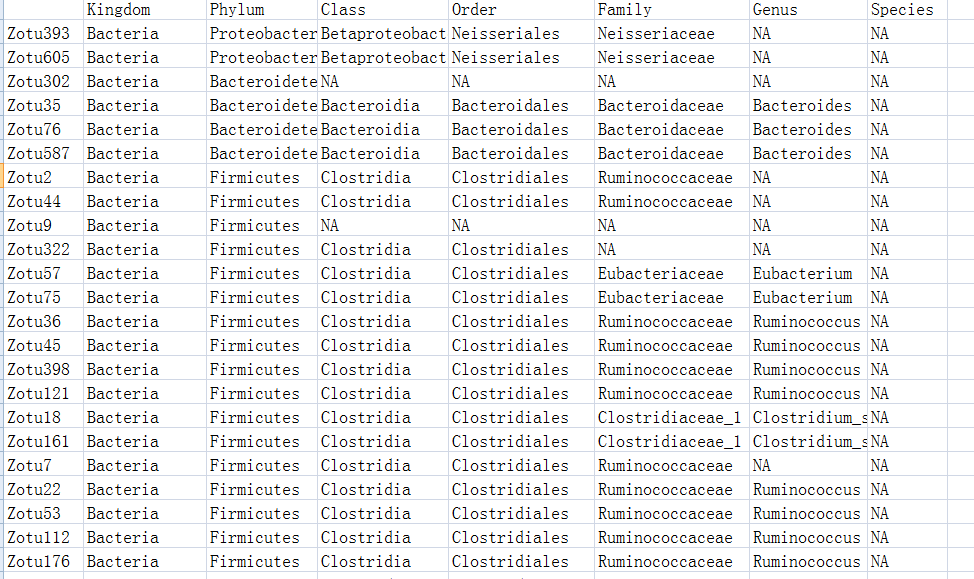
以Welch's t 检验校正后的 *P*值<0.05；effect值>1 或者effect值<-1为标准，本示例共筛选出179个差异zOTUs, 其中70个zOTUs在空肠粘膜样品中显著富集，109个zOTUs在空肠食靡样品中显著富集。如图5所示，图中的每个点代表1个zOTU。左侧的effect图是基于Dispersion（x轴，每个zOTU的组内中位数差异）和Difference（Y轴，每个zOTU的组间中位数差异）而做的。当zOTUs的Difference的绝对值大于Dispersion时（即effect值>1 或者effect值<-1 ），图中的点外围被标红色圆圈；当zOTUs的Welch's t 检验校正后的 *P*值<0.05，图中的点被标蓝。由此判断，左图中的蓝色圆圈且外围标红的圆点即为差异zOTUs。右图中的x轴是每个zOTU的effect size（效应量大小），y轴是Welch's t 检验校正后的 *P*值<0.05。与左图类似，右图中的的蓝色圆圈且外围标红的圆点即为差异zOTUs。



**图5.** 差异zOTUs的effect图

3.2 差异zOTUs 的物种注释

差异zOTUs的文件格式如图6。山羊空肠食靡中（Digesta enriched OTUs.csv）富集的细菌zOTUs主要为*Eubacterium*、*Ruminococcus*、*Romboutsia*、*Lactobacillus*、*Streptococcus*、*Desulfovibrio*、*Halomonas*、*Shewanella*等，而山羊粘膜中（Mucosa enriched OTUs.csv）富集的细菌zOTUs主要为*Porphyromonas*、*Bacteroides*、*Campylobacter*和一些未注释到物种信息的zOTUs。



**图6**. 差异zOTUs的物种注释结果

**致谢**

感谢国家自然科学基金委青年项目（31601967）对本分析流程的资助。本试验方案内容改编于Gloor等（2017）开发的流程，特此致谢。本分析流程已应用于Zhang等（2020）的文章中，特此致谢。

**参考文献**

Claesson MJ, Clooney AG and O'Toole PW. 2017, A clinician's guide to microbiome analysis. Nat Rev Gastroenterol Hepatol. 14(10): 585-595.

Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. 2019, Separating host and microbiome contributions to drug pharmacokinetics and toxicity. Science. 363(6427).

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bahler J. 2015. Proportionality: a valid alternative to correlation for relative data . PLoS Comput Biol. 11(3): e1004075.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017, Microbiome datasets are compositional: and this is not optional. Front Microbiol. 8: 2224.

Jiao J, Huang J, Zhou C, Tan Z. 2015, Taxonomic identification of ruminal epithelial bacterial diversity during rumen development in goats. Appl Environ Microbiol. 81(10): 3502-3509.

Love MI, Huber W and Anders S. 2014, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12): 550.

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vazquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017, Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 5.

Thorsen J, Brejnrod A, Mortensen M, Rasmussen Morten A, Stokholm Jakob, Abu Al-Soud Waleed, Sorensen Soren, Bisgaard Hans, Waage Johannes. 2016, Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 4.

Zhang X, Wu J, Zhou C, Tan Z, Jiao J. 2020. Spatial and temporal organization of jejunal microbiota in goats during animal development process. J Appl Microbiol. doi: 10.1111/jam.14961.