群落差异与环境因子分析

王　鹏

2018.8

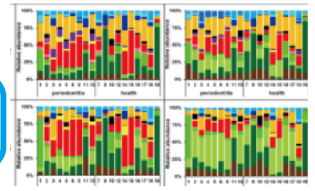
目录

CONTENTS

扩增子分析简述

Diversity & Dynamics

Types | Distribution | Changes



Orgnisms

Genes

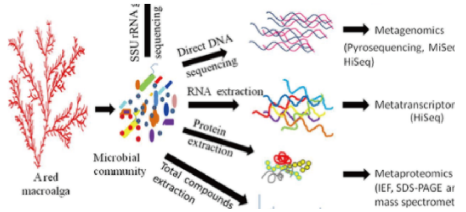
Community structure

Who | Abundances | Composition

Microbial Community

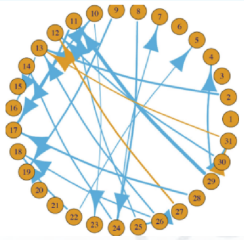
Ecosystem function

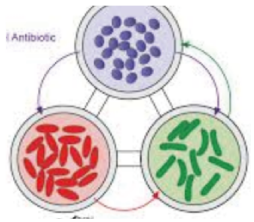
Who | Abundances | Composition



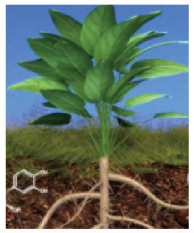
Interaction & Communication

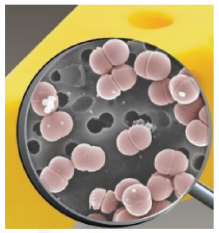
Competition | Parters | Signaling



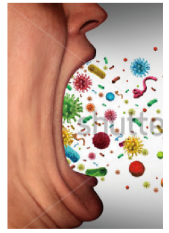


Soil Water Food





Gut Skin Oral



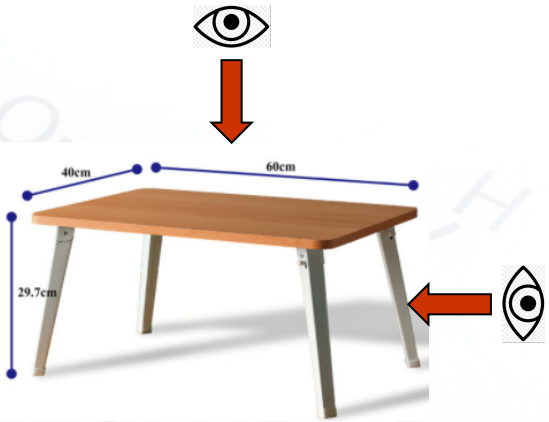
Novogene 诺禾致源

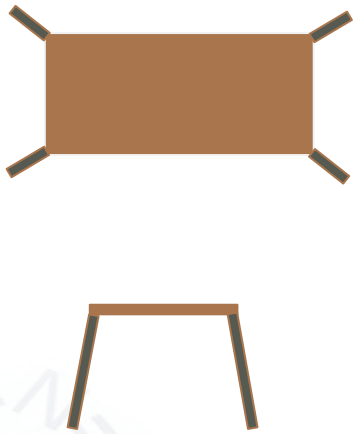
Providing advanced genomic solutions!

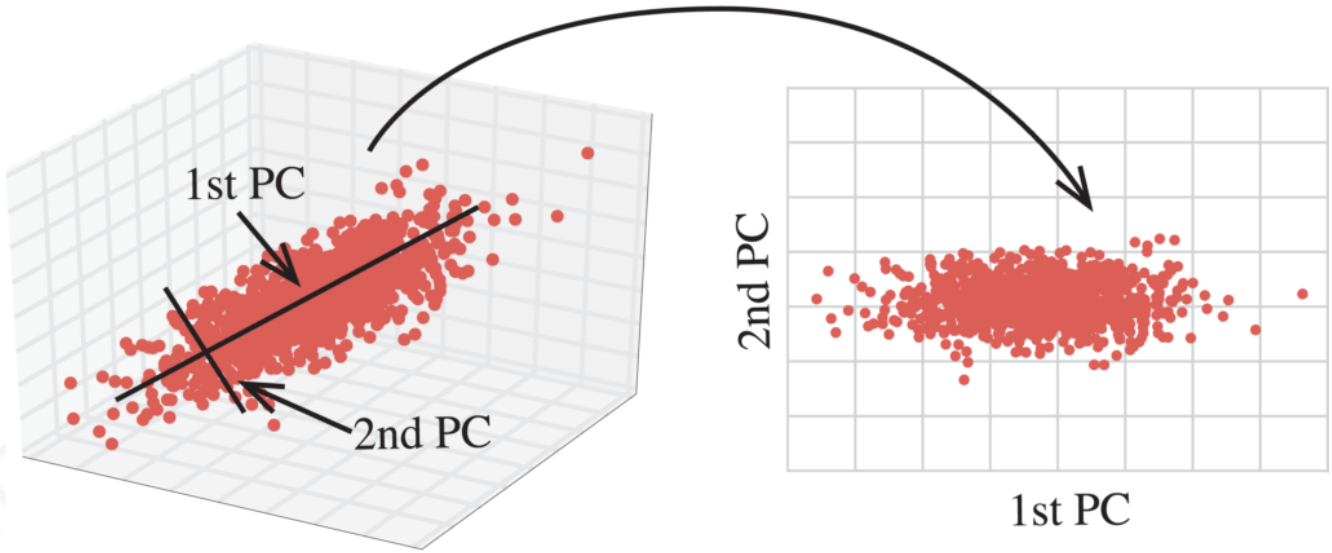
降维分析

High dimension → **Low dimension**

Complexity — **Simplicity**







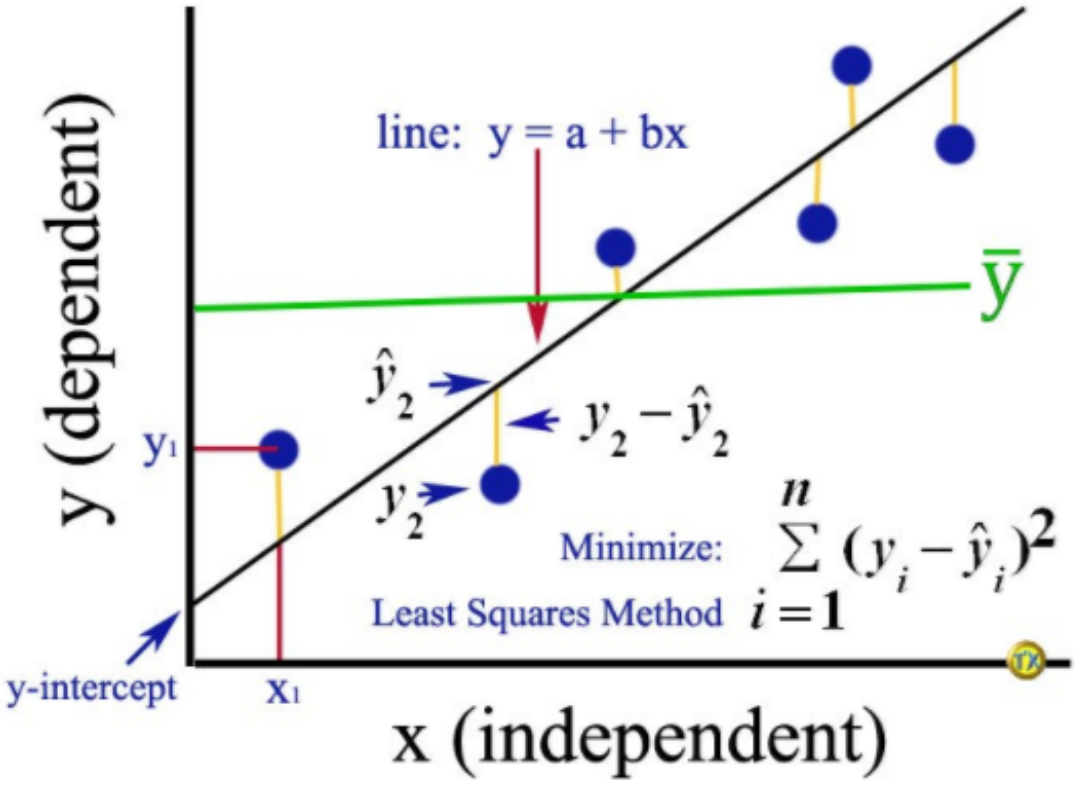
PCA ⊕ **PCoA** ⊕ **NMDS**

Novogene 诺禾致源

Providing advanced genomic solutions!

Principal component analysis

线性回归



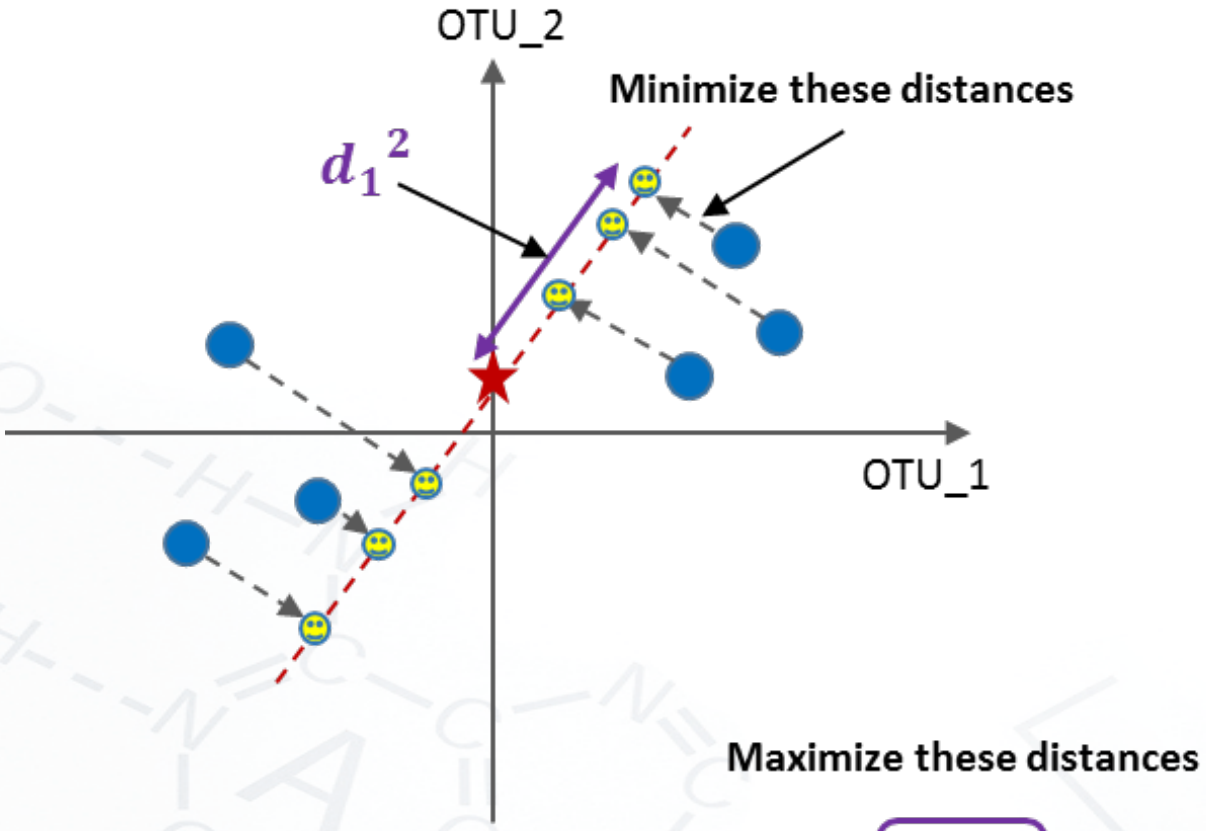
线性回归：

Step1：计算中心点（x和y的平均值）

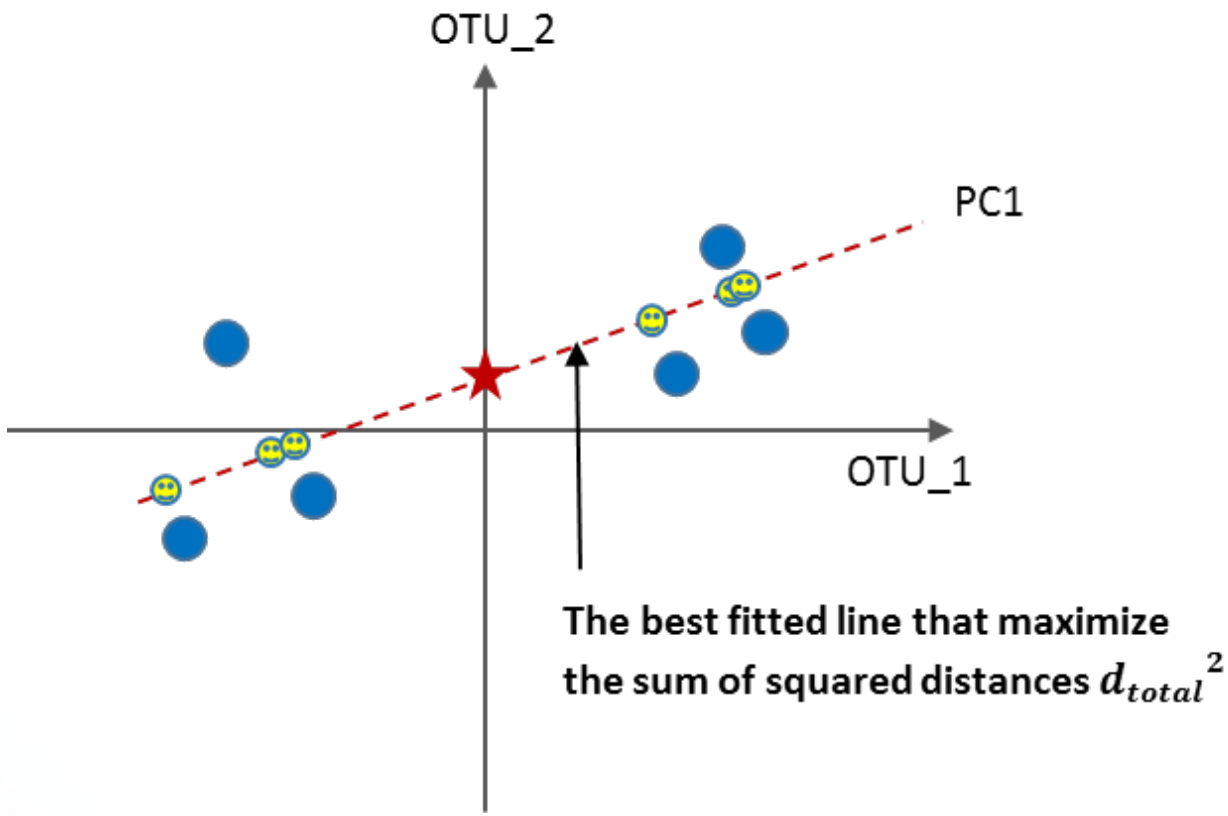
Step2：线性拟合（最小二乘）

Step3：计算斜率b、截距a和决定系数R^2

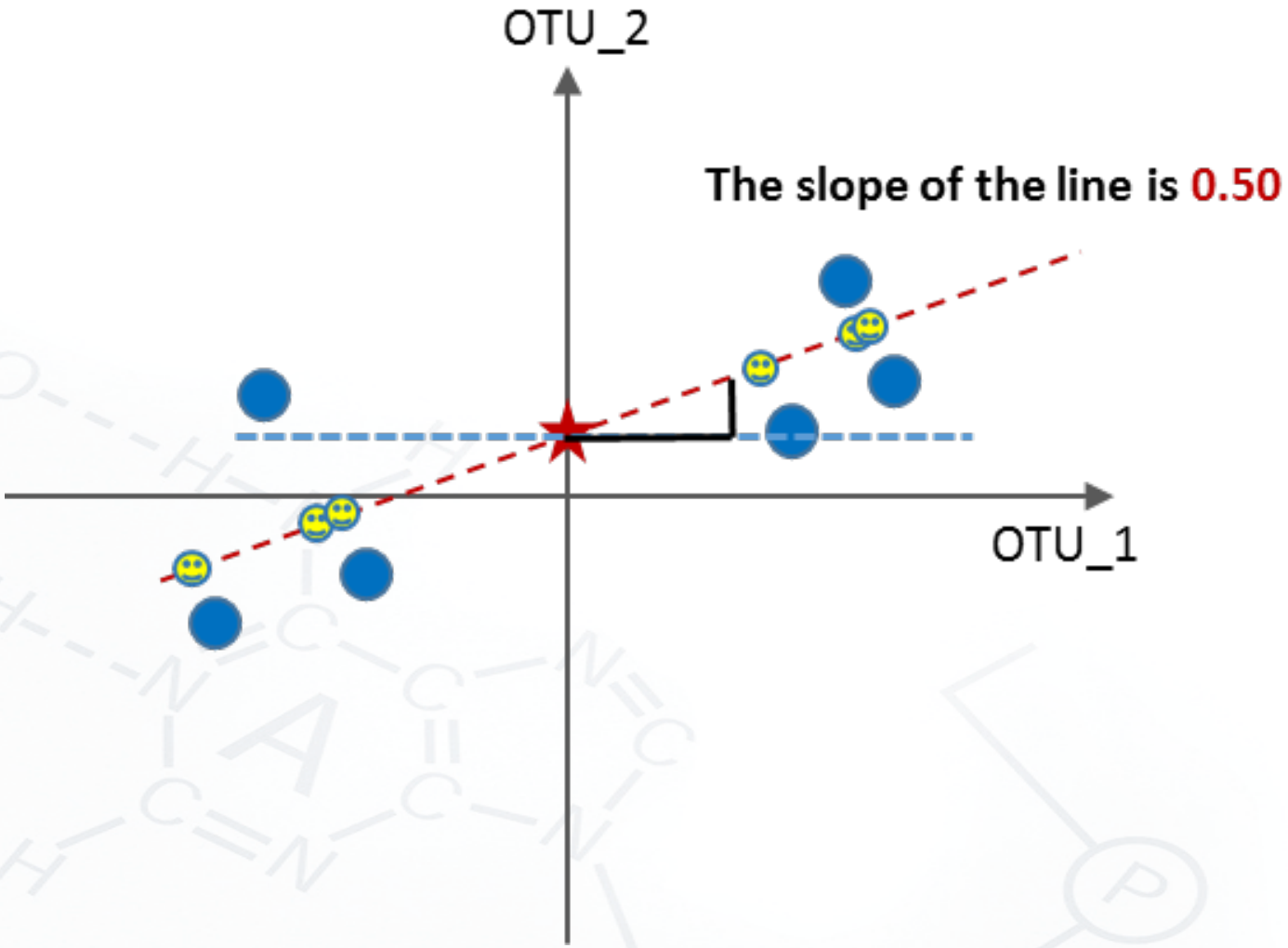
Principal component analysis

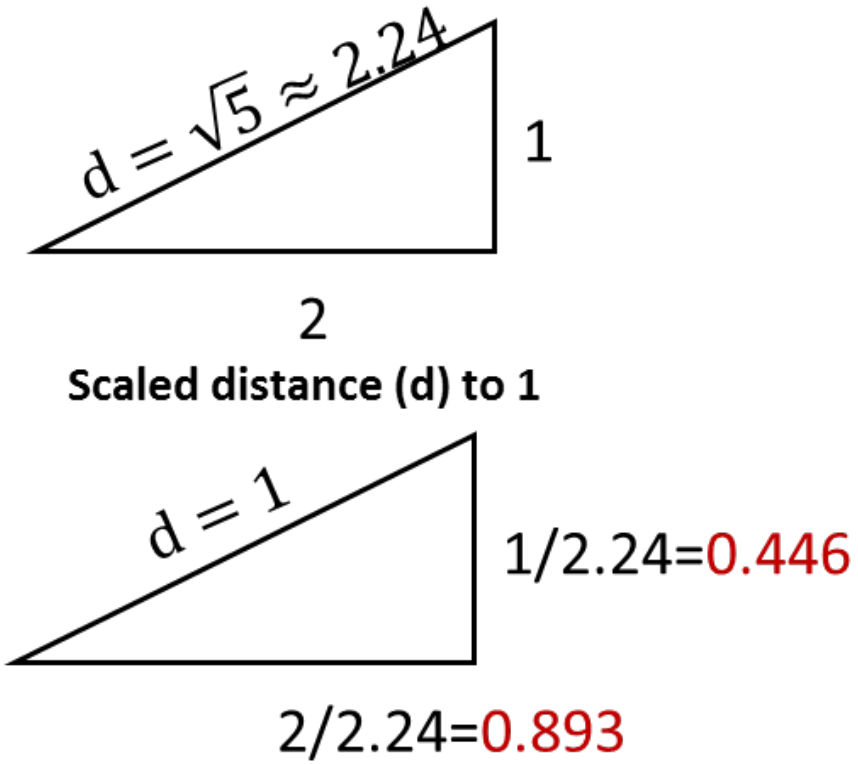


$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = d_{total}^2$$



Principal component analysis

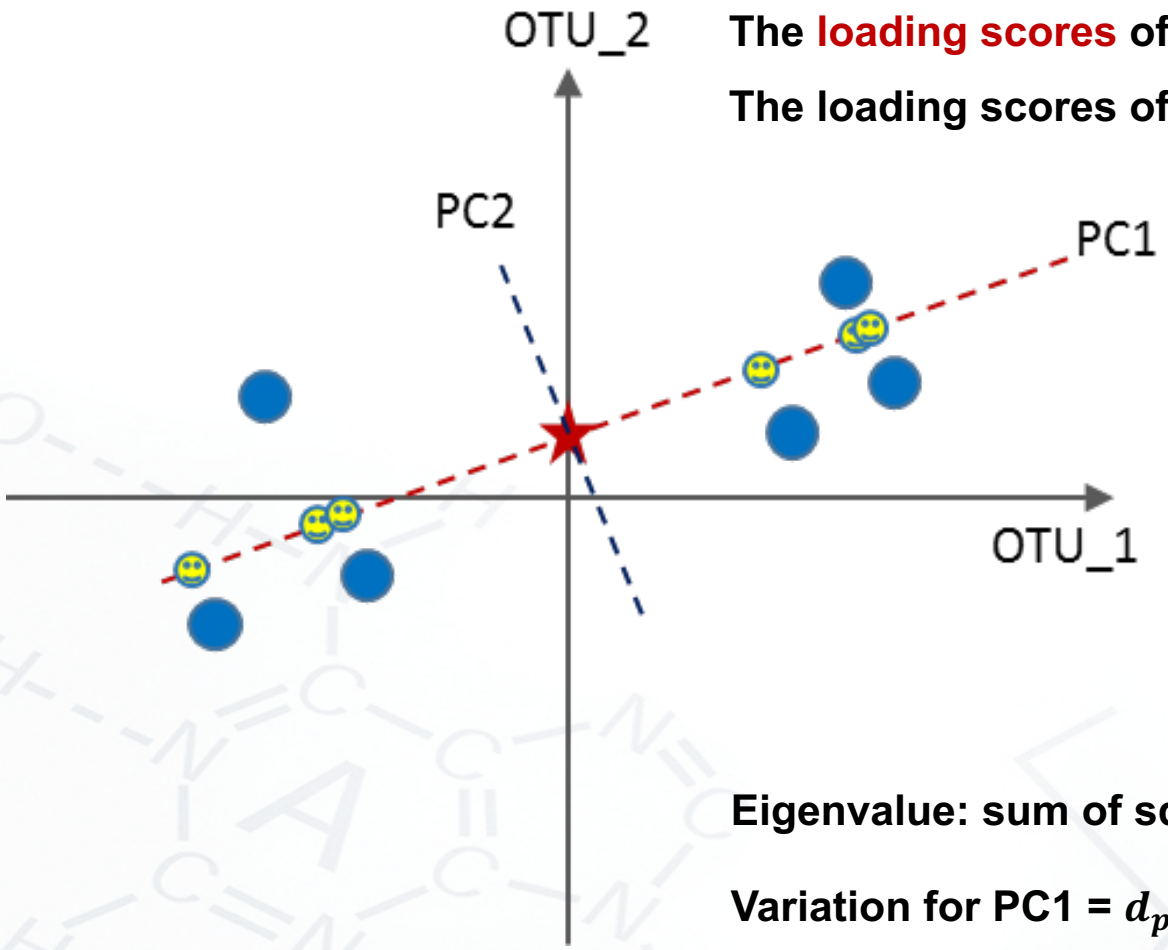




The loading scores of OTU_1 in PC1: 0.446

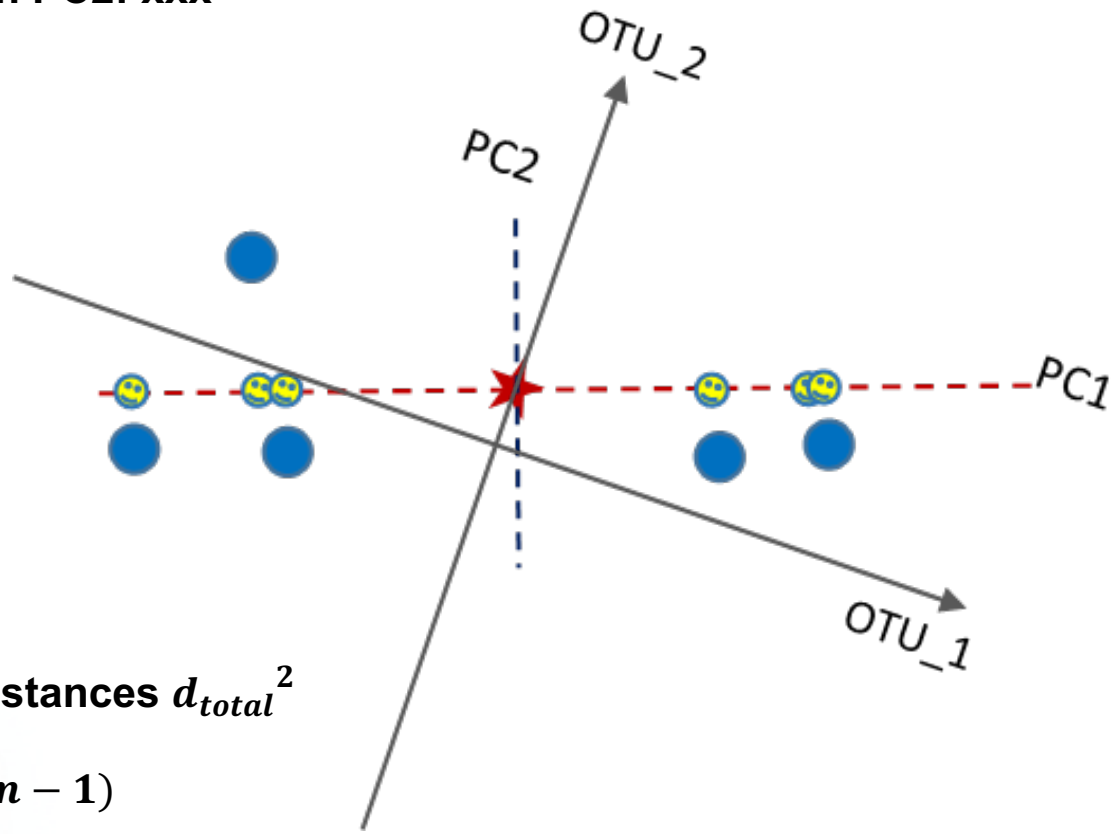
The loading scores of OTU_2 in PC1: 0.893

Principal component analysis



The loading scores of OTU_1 in PC2: xxx

The loading scores of OTU_2 in PC2: xxx



Eigenvalue: sum of squared distances $d_{total}{}^2$

Variation for PC1 = $d_{pc1_{total}}{}^2/(n-1)$

Variation for PC2 = $d_{pc2_{total}}{}^2/(n-1)$

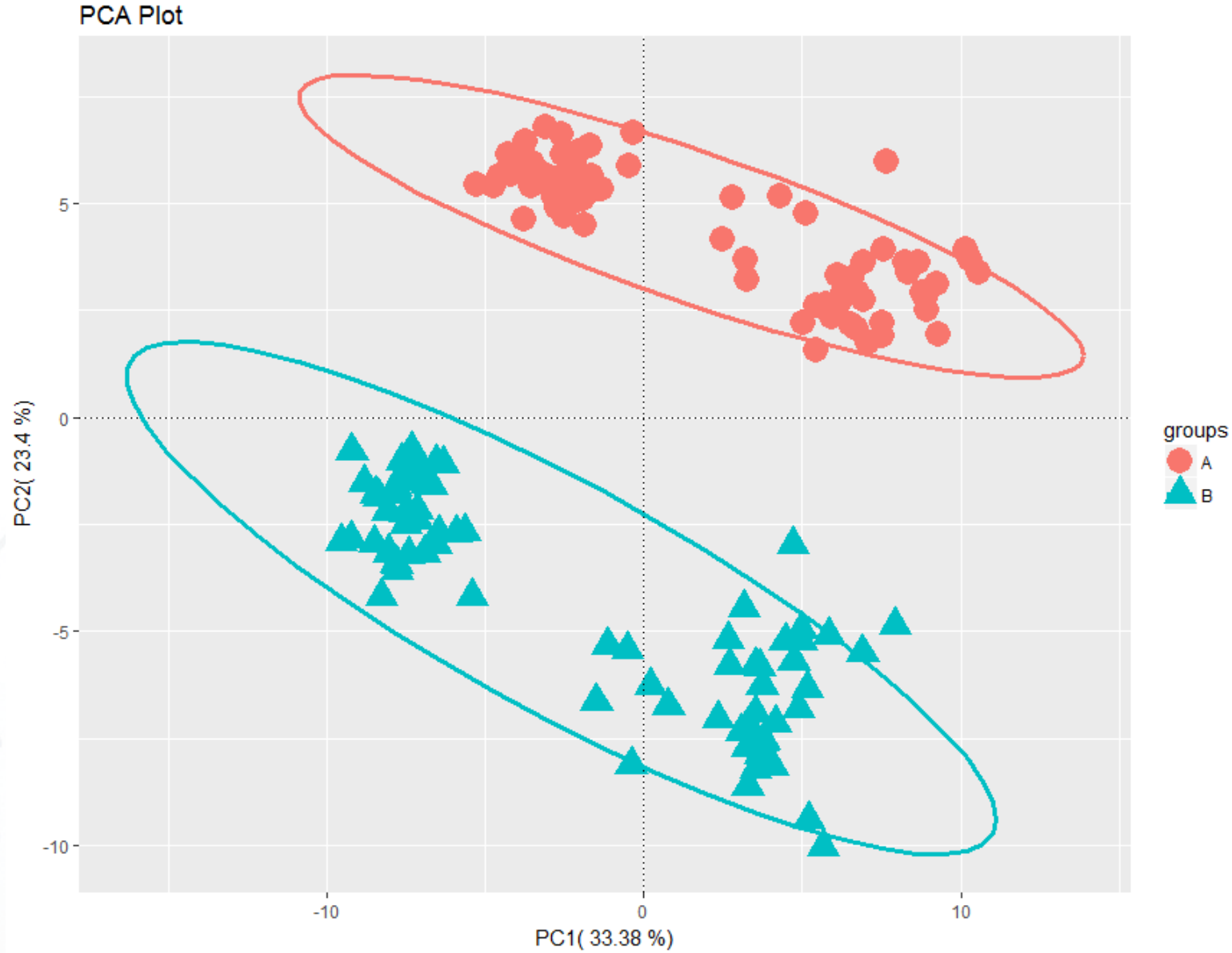
How to perform PCA in R?

prcomp() ade4---dudi,pca() vegan---rda()

```
library(ade4);library(ggplot2)
data <- read.table("otu_table.relative.xls", stringsAsFactors = F,
                   head=T, row.names=1,sep="\t", comment.char = "")
data <- data[, c(1:(ncol(data)-1))]
data <- t(data)
groups = read.table("group.list",
                    head=F,colClasses=c("character","character"))
pca<-dudi.pca(data[,1:ncol(data)], scannf=F, nf=5)
PC1 <-pca$li[,1]
PC2 <-pca$li[,2]
plotdata<-data.frame(rownames(pca$li),PC1,PC2,groups$V2)
colnames(plotdata) <-c("sample","PC1","PC2","groups")
pc1 <-round(pca$eig[1],2)
pc2 <-round(pca$eig[2],2)
```

How to perform PCA in R?

```
P<-ggplot(plotdata, aes(PC1, PC2))

P<-P+geom_point(aes(colour=groups,shape=groups),size=4)

P<-P+geom_text(aes(label=sample),size=3,family="serif", hjust=0.5,vjust=-1)

P<-P+labs(title="PCA Plot",x=paste("PC1(",pc1,"%)"),

y=paste("PC2(",pc2,"%)"))

P<-P+geom_vline(xintercept=0,linetype="dotted")

P<-P+geom_hline(yintercept=0,linetype="dotted")

P+stat_ellipse(aes(group=groups, colour = groups), size=1.2)
```



Novogene 诺禾致源

Providing advanced genomic solutions!

Principal coordinate analysis (PCoA)

	otu1	otu2	otu3
1	0	1	1
2	1	0	0
3	0	4	4

Euclidean distance

	1	2	3
1	0.000000	1.732051	4.242641
2	1.732051	0.000000	5.744563
3	4.242641	5.744563	0.000000

Euclidean distance is usually not appropriate for ecological data

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{21} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

eigenvalues decomposition

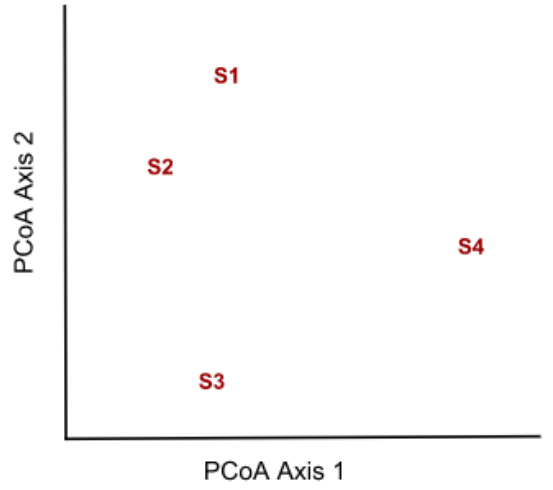
$$\min_{\tilde{\mathbf{x}}_1,\cdots,\tilde{\mathbf{x}}_n} \left(\sum_{i<j} \left(\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\| - d_{ij} \right)^2 \right)^{1/2}$$

a

Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

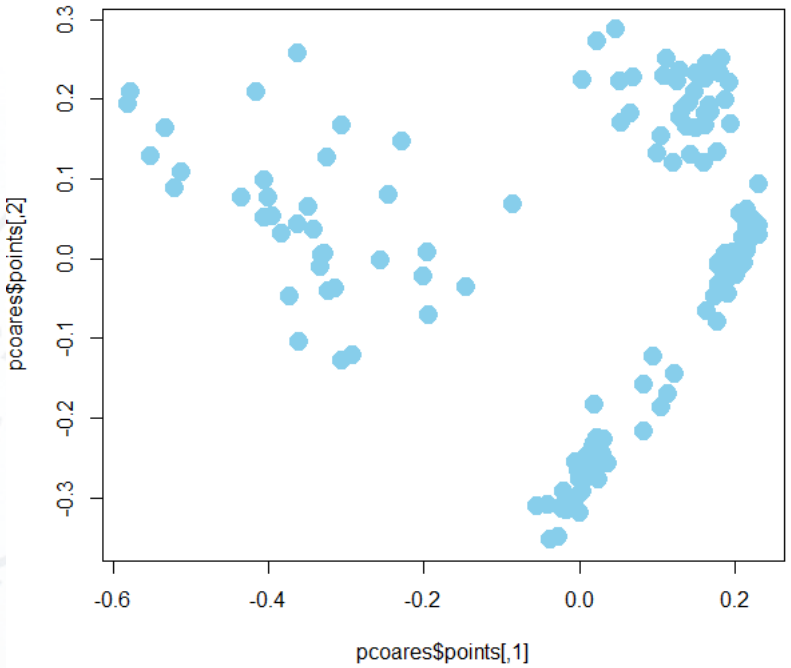
b

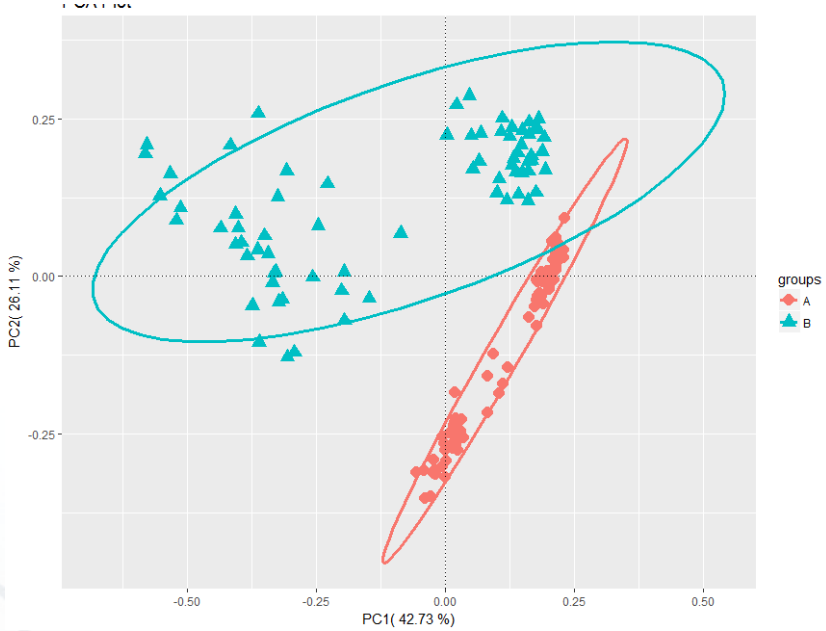


How to perform PCoA?

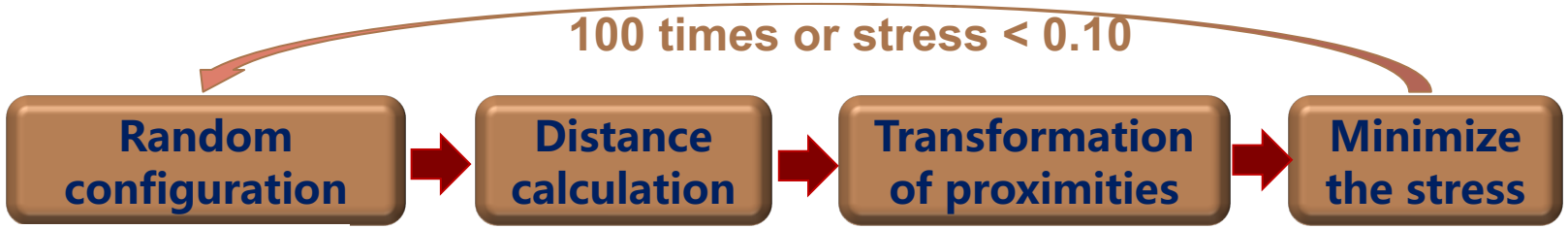
```
library(vegan)
dst <- vegdist(data, method = 'bray')
pcoares <- cmdscale(dst, eig = TRUE, x.ret = TRUE)
plot(pcoares$points, cex=2, pch=19, col='skyblue')
```

Classical multi-demensional scaling





Non-metric multi-demensional scaling (NMDS)



Species

Sites	X1	X2	X3	X4
S1	14	2	14	14
S2	10	14	0	8
S3	0	5	0	2
S4	0	0	1	0

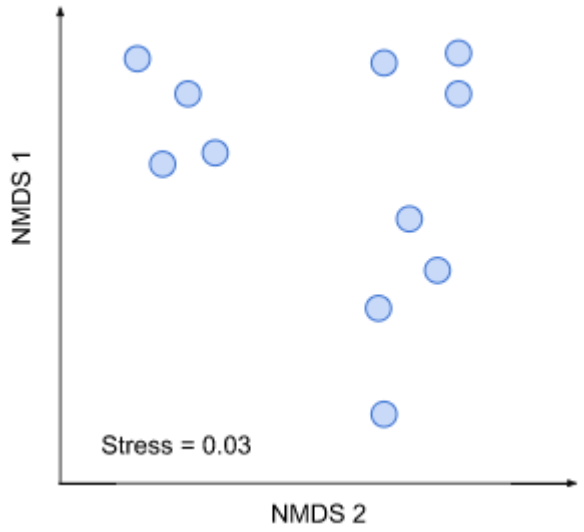
Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

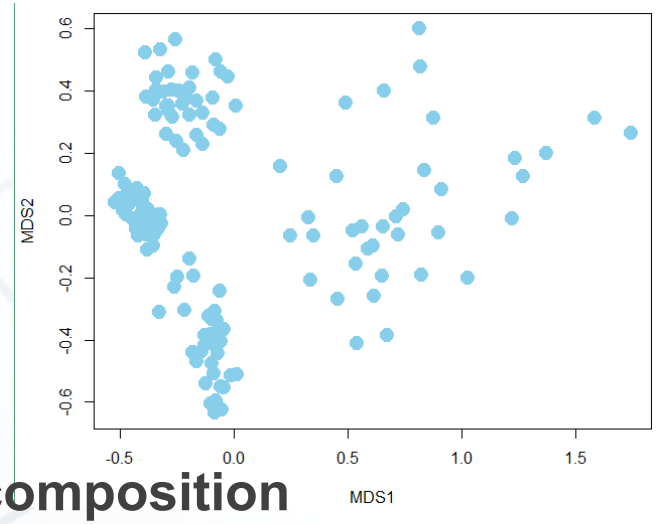
Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	1	0	...	...
S3	3	2	0	...
S4	4	5.5	5.5	0

$$\text{Stress} = \sqrt{\frac{\sum_{ij} (d_{ij} - \hat{d}_{ij})^2}{\sum_{ij} d_{ij}^2}}$$




```
library(vegan)
nmdsres <- metaMDS(data, k = 2,
distance = 'bray', trace = FALSE)
plot(nmdsres$points, col='skyblue',
pch=19, cex=2)
```

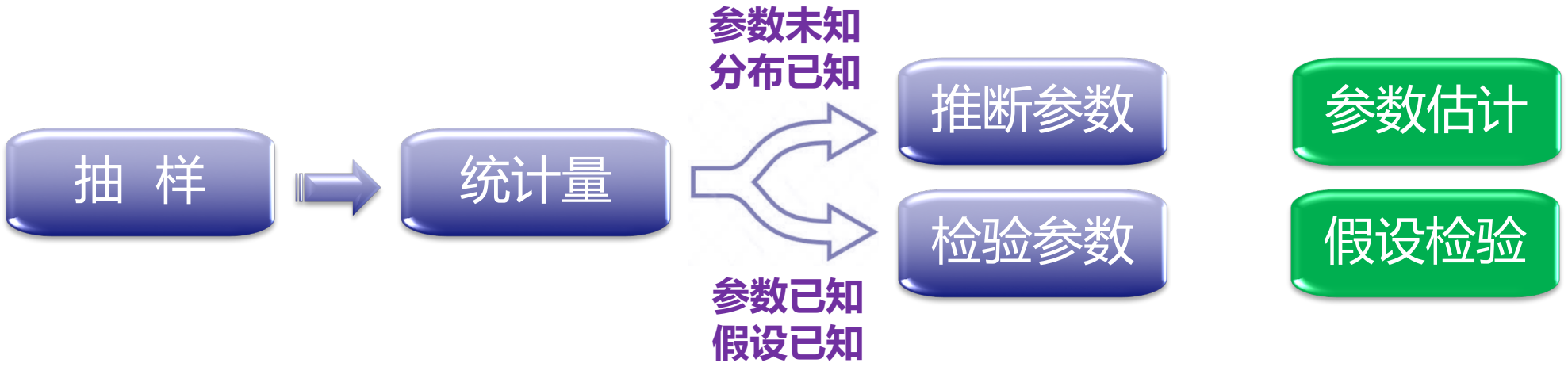


- No eigenvalue decomposition
- Non-linear model

Novogene 诺禾致源

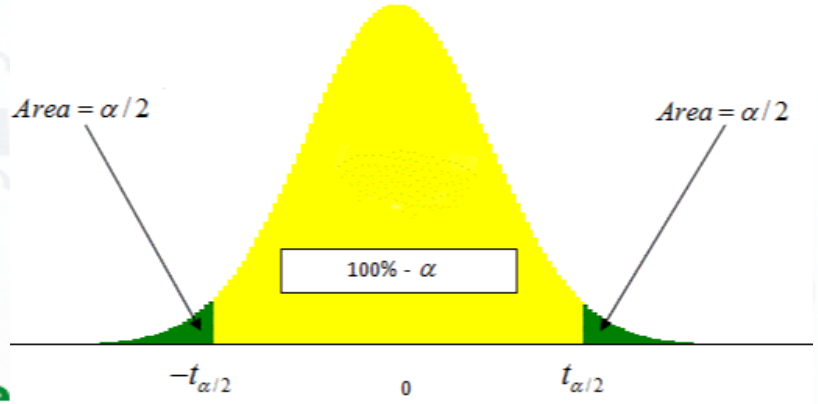
Providing advanced genomic solutions!

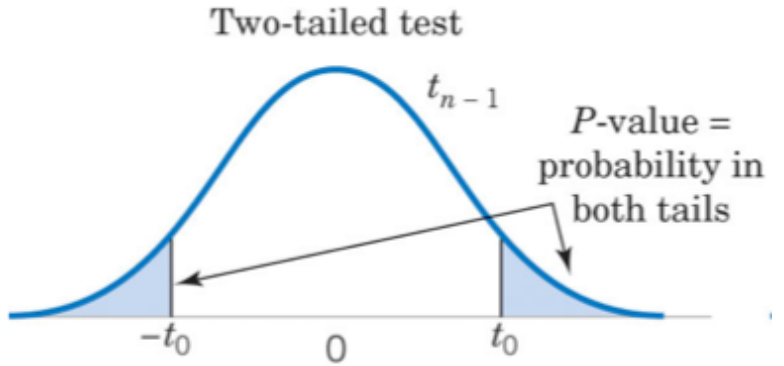
假设检验基础



显著性水平α：发生第I类错误的概率（经验值一般为0.01和0.05）

P-value（Probability）：在原假设成立的情况下，出现极端情况的概率





Novogene 诺禾致源

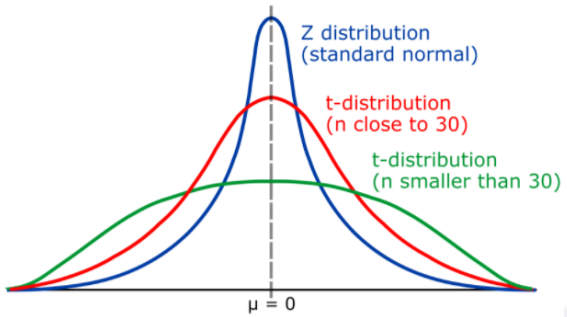
Providing advanced genomic solutions!

差异物种分析

T-test

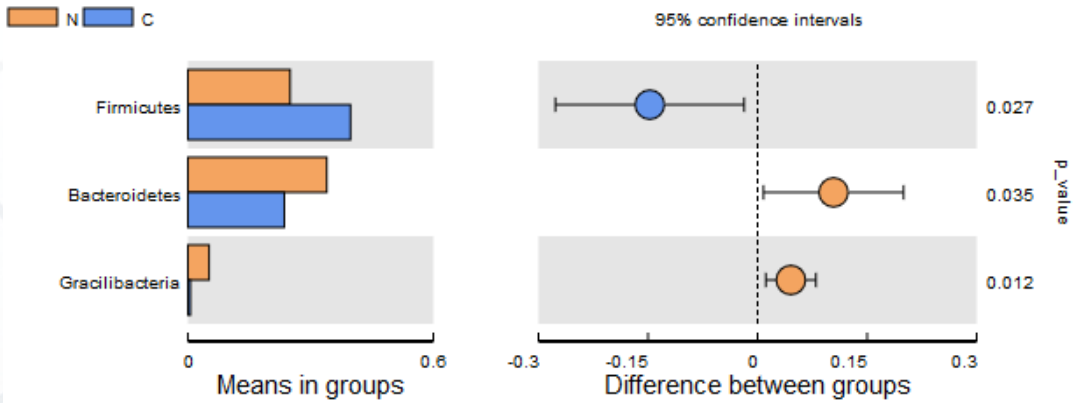
T-test：examine whether the two samples are drawn from the same population

t-distribution



- One sample T-test
- Two Independent Samples T-Test
- Paired T-test

```
sci <- c(8.3, 8.9, 8.8, 8.1, 7.3, 7.5, 5.8, 6.9)
t.test(sci, mu=5, alternative = 'two.sided', conf.level = 0.95)
#----------------------------------------
before <- c(5, 4, 5, 4.5, 6, 6.5, 6.3)
after <- c(10, 11, 11.1, 10.9, 9.8, 9.4, 11.3)
t.test(before, after, alternative = 'two.sided', paired=FALSE,
var.equal = FALSE, conf.level = 0.95)
#----------------------------------------
t.test(before, after, alternative = 'two.sided', paired=TRUE, var.equal
= FALSE, conf.level = 0.95)
```



Novogene 诺禾致源

Providing advanced genomic solutions!

差异物种分析

MetaStats

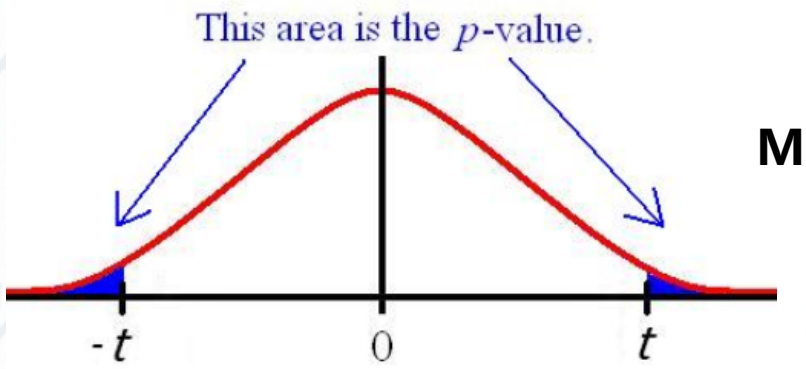
OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples

James Robert White[1], Niranjan Nagarajan[2], Mihai Pop[3]*

T-test



Permutation

Multiple test

p-value

p_1
p_2
p_3
p_4
....
....
....

FDR

Benjamini & Hochberg,

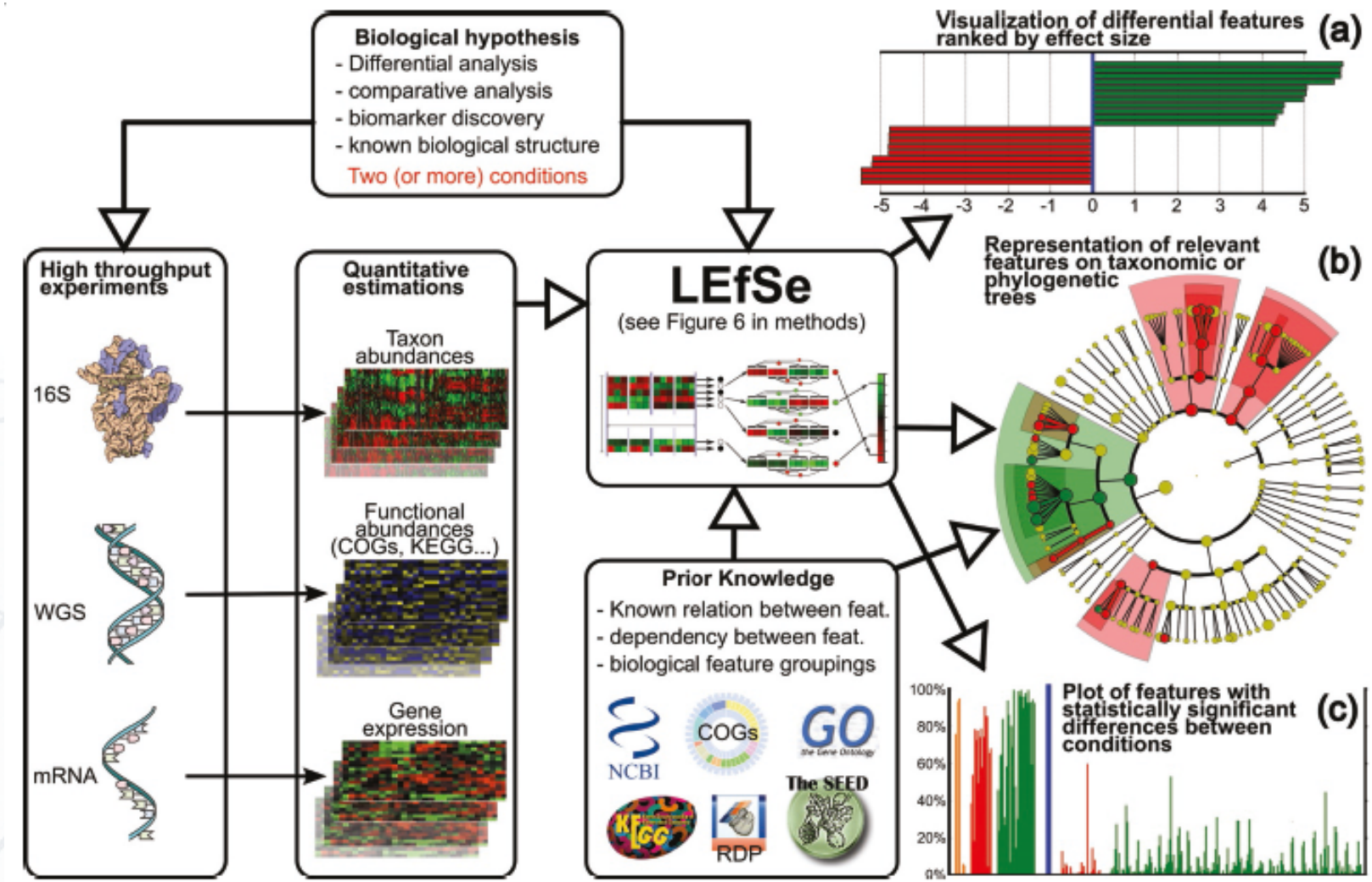
q-value

q_1
q_2
q_3
q_4
....
....
....

```
pval <- c(0.8401, 0, 0.0271, 0.0411, 0.0137, 0.1272)
p.adjust(pval, method = 'BH')
```

差异物种分析

LefSe (LDA Effect Size)



Kruskal-Wallis (KW) sum-rank test

Wilcoxon rank-sum test

Linear Discriminant Analysis

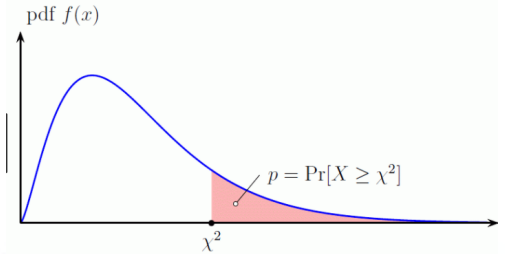
差异物种分析

LefSe (LDA Effect Size)

Kruskal-Wallis (KW) sum-rank test

A	rank	B	rank	C	rank
1.2	1	3.2	7	6.1	14
1.5	2	3.4	9	6.3	15
1.7	3	3.6	10	6.4	16
2.1	4	3.3	8	6	13
2.5	6	3.8	11	7.1	17
2	5	4	12	7.7	18
	21		57		93

$$H = \frac{12}{n\cdot(n+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3\cdot(n+1)$$



$$H = \frac{12}{18*19}*\left(\frac{21^2}{6}+\frac{51^2}{6}+\frac{93^2}{6}\right) - 3*(18+1) = 15.16$$

```
        Kruskal-Wallis rank sum test

data:  x and g
Kruskal-Wallis chi-squared = 15.158, df = 2, p-value = 0.0005111
```

```
x <- c(1.2, 1.5, 1.7, 2.1, 2.5, 2,
       3.2, 3.4, 3.6, 3.3, 3.8, 4,
       6.1, 6.3, 6.4, 6, 7.1, 7.7)
g <- factor(rep(c('A', 'B', 'C'), each=6))
kruskal.test(x, g)
```

Wilcoxon rank-sum test

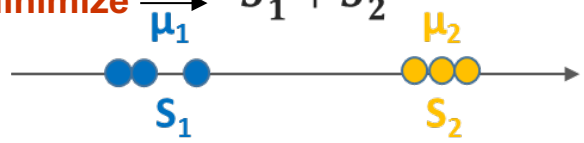
```
#================
x <- c(1.2, 1.5, 1.7, 2.1, 2.5, 2)
y <-  c(3.2, 3.4, 3.6, 3.3, 3.8, 4)
wilcox.test(x,y)
```

Linear Discriminant Analysis

Maximize this function

Maximize $\longrightarrow \dfrac{(\mu_1-\mu_2)^2}{S_1^2+S_2^2}$

Minimize $\longrightarrow$

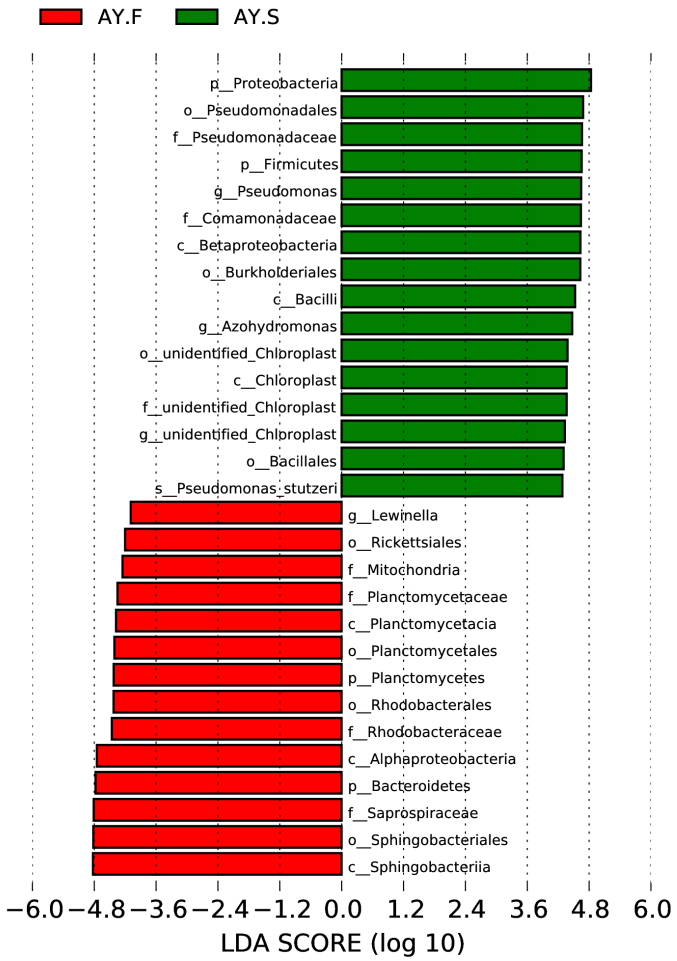


Novogene 诺禾致源

Providing advanced genomic solutions!

差异物种分析

LefSe (LDA Effect Size)



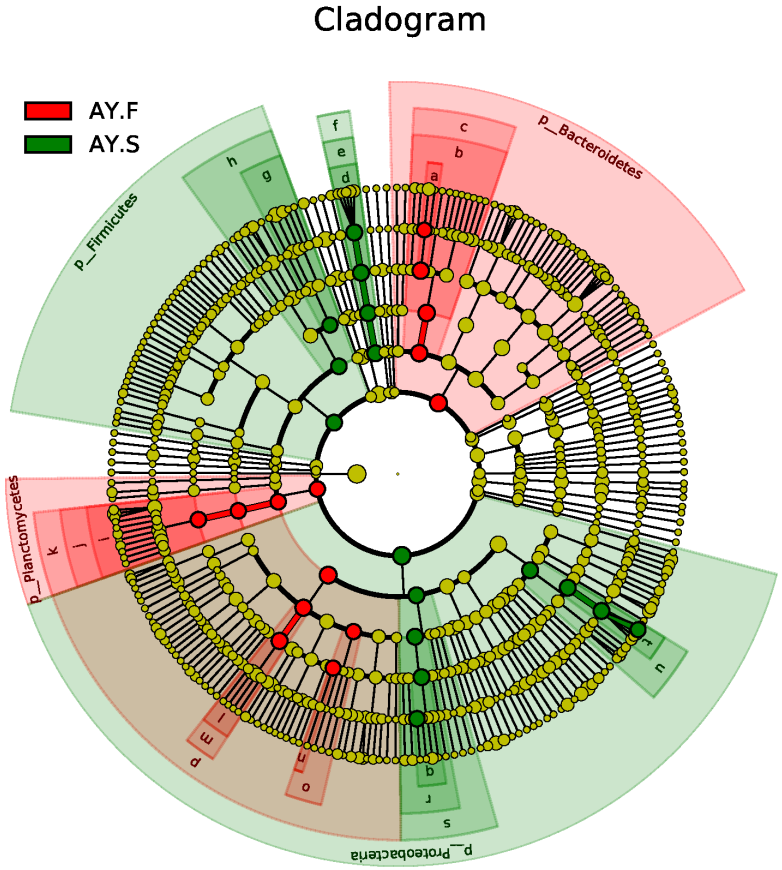
柱子长度：LDA SCORE

柱子颜色：高丰度分组

圆圈：颜色和大小

扇形：颜色和面积

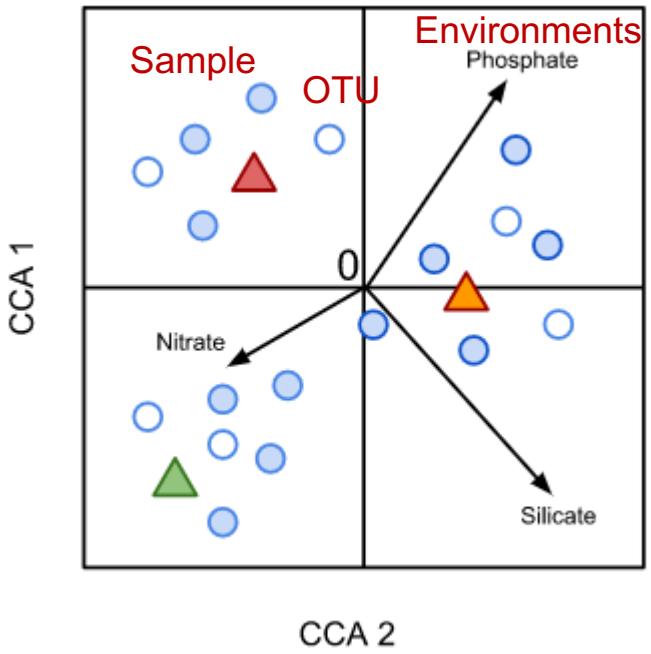
颜色：分组信息



环境因子分析 Constrained Correspondence Analysis (CCA)

Canonical Correlation Analysis

Response variables & Explanatory variables → PCA → u1,u2,u3..... & v1,v2,v3..... → argmax ρ(u1,v1)



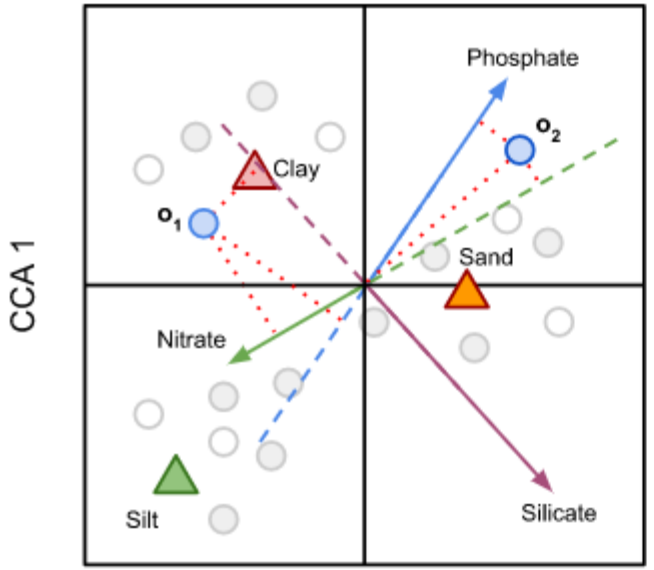
Sample-OTU & Sample-Sample & OTU-OTU

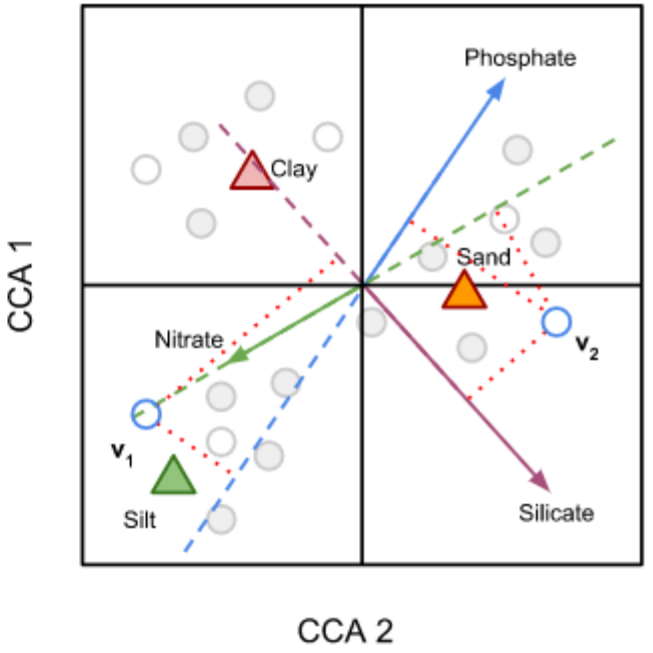
Env-Sample & Env-OUT & Env-Env

Type 1

Object (Samples)

Explanatory variables





Type 2:

Explanatory variables

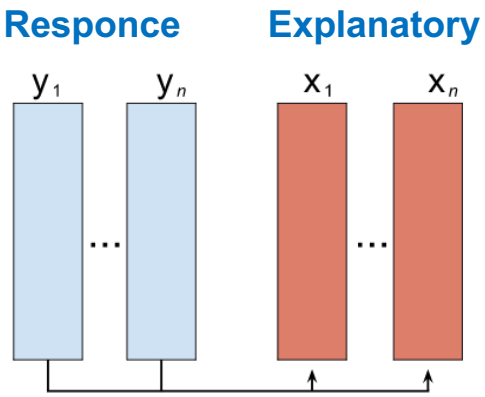
Response variables

Novogene 诺禾致源

...roviding advanced genomic solutions!

Reference:https://mb3is.megx.net/gustame/constrained-analyses/cca

Redundancy Analysis (RDA)

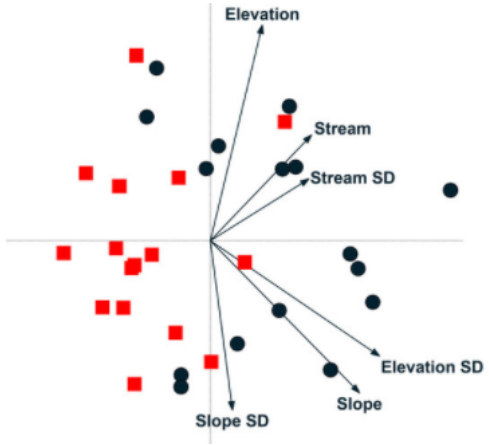


Multiple linear regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

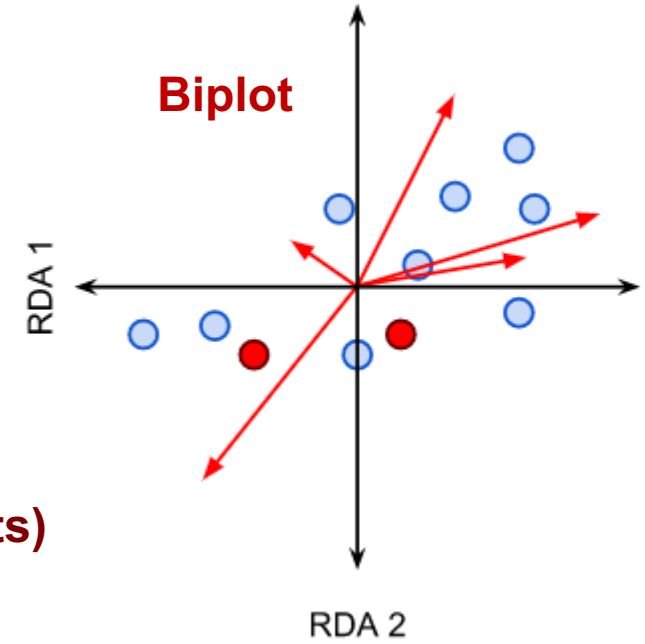
Fitted value matrix of response variables

PCA

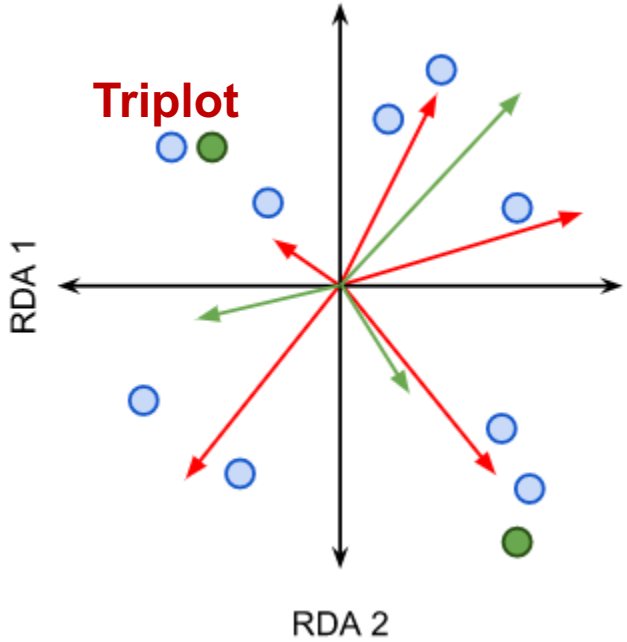


- Constrained analysis
- Linear model

Biplot



Triplot



Object focused (Distance plots)

Responce var focused (Correlation plots)

Novogene 诺禾致源

Providing advanced genomic solutions!

Referrence:https://sites.google.com/site/mb3gustame/constrained-analyses/rda

Distance based Redundancy Analysis (dbRDA)

Samples

	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

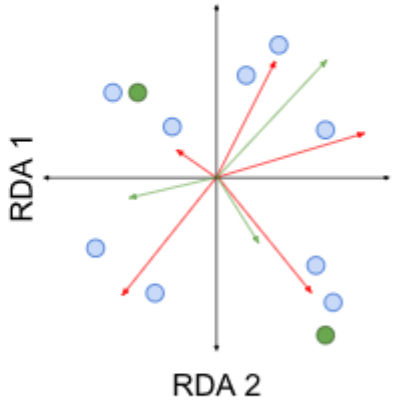
Explanatory variables

	E1	E2	E3
S1	A	3.2	12
S2	A	2.1	6
S3	B	0.5	11
S4	B	1.6	22

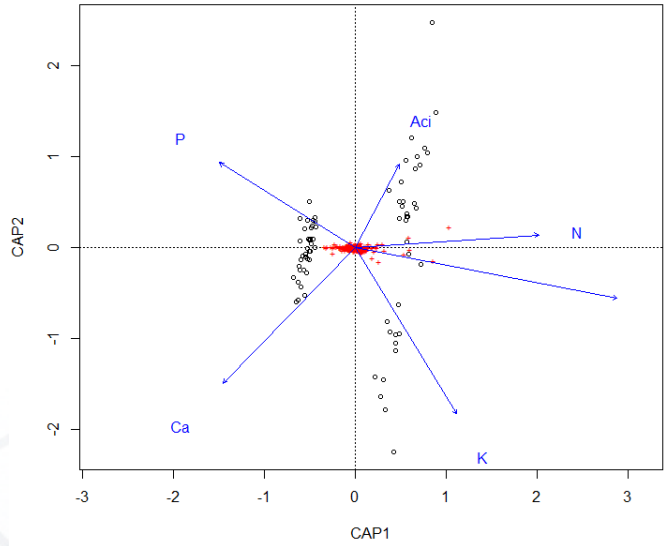
PCoA

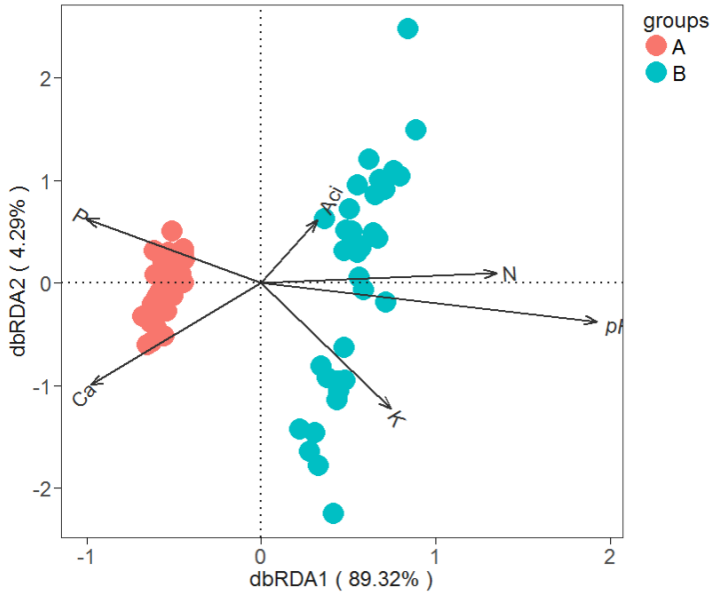
RDA

Samples	PCoA Axis 1	...	PCoA Axis *n*
S1	-1.23		0.47
S2	0.21	...	-0.02
S3	0.50		0.20
S4	-0.11		0.33




```
env <- read.table("env.list", stringsAsFactors = F,
                  head=T, row.names=1,sep="\t")
dbres <- capscale(data~., env, distance = 'bray')
plot(dbres)
```





- Constrained analysis
- Non-linear model
- Distance based

Object focused (Distance plots)

Responce var focused (Correlation plots)

Novogene 诺禾致源

Providing advanced genomic solutions!

Reference:https://sites.google.com/site/mb3gustame/constrained-analyses/rda/dbrda

How to perform CCA/RDA/dbRDA

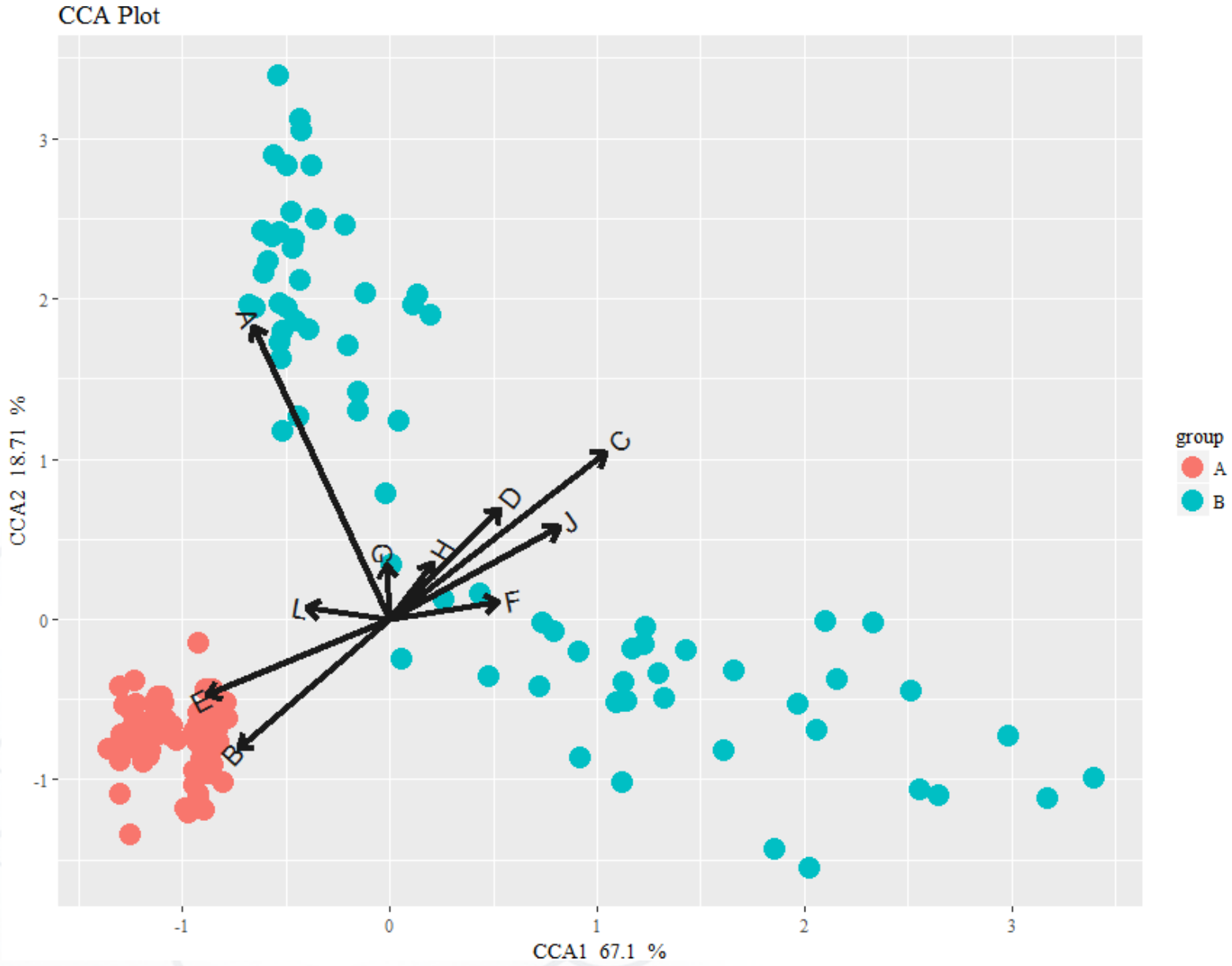
```
library(ggplot2);library(vegan)
#read data
data <- read.table("otu_table.relative.xls", stringsAsFactors = F,
                   head=T, row.names=1,sep="\t", comment.char = "")
data <- data[, c(1:(ncol(data)-1))]
data <- t(data)
env <- read.table("env.list", stringsAsFactors = F,
                  head=T, row.names=1,sep="\t", comment.char = "")
groups = read.table("group.list",
                    head=F,colClasses=c("character","character"))
```

How to perform CCA/RDA/dbRDA

```
cca<-cca(data, env, scale=T)

scorcca <- scores(cca)

sam <- data.frame(scorcca$sites, groups$V2)   #提取样本得分

colnames(sam) <- c("CCA1","CCA2","group")

spec <- scorcca$species   #物种得分

spec <- as.data.frame(spec)

env <- cca$CCA$biplot[,c(1,2)] #环境因子得分

env <- as.data.frame(env)

cca1 =round(cca$CCA$eig[1]/sum(cca$CCA$eig)*100,2) #第一轴标签

cca2 =round(cca$CCA$eig[2]/sum(cca$CCA$eig)*100,2) #第二轴标签
```

How to perform CCA/RDA/dbRDA

```
p <- ggplot(data=sam,aes(CCA1,CCA2))
p <- p + geom_point(aes(colour=group,shape=group),size=5) +
  scale_shape_manual(values=c(19,19)) +
  labs(title="CCA Plot",x=paste("CCA1 ",cca1," %"),y=paste("CCA2 ",cca2," %")) +
  theme(text=element_text(family="serif"))
p + geom_segment(data = env,aes(x=0,y=0,xend = env[,1], yend = env[,2]),
colour="gray10", size=1.5, arrow=arrow(angle=35, length=unit(0.3, "cm"))) +
  geom_text(data=env, aes(x=env[,1], y=env[,2], label=rownames(env)),
            size=5, colour="gray10",hjust = (1 - 2 * sign(env[ ,1])) / 3,
            angle = (180/pi) * atan(env[ ,2]/env[ ,1]))
```



Correlation analysis

Pearson's correlation coefficient (PCC)

总体相关系数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

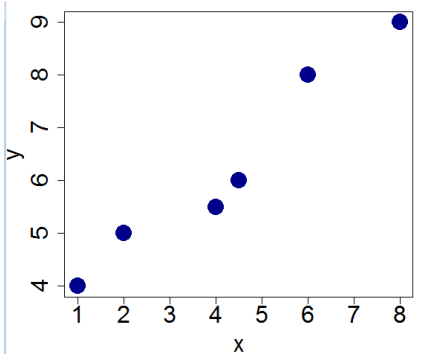
where:

- n is the sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

样本相关系数

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

```
x <- c(1,2,4,4.5,6,8)
y <- c(4,5,5.5,6,8,9)
cor(x,y,method='pearson')
plot(x,y,cex=3,col='darkblue',pch=19,cex.axis=2,cex.lab=2)
```



Correlation analysis

Spearman's rank correlation coefficient (SCC)

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$$

where

- ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables.
- $\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)$ is the covariance of the rank variables.
- σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

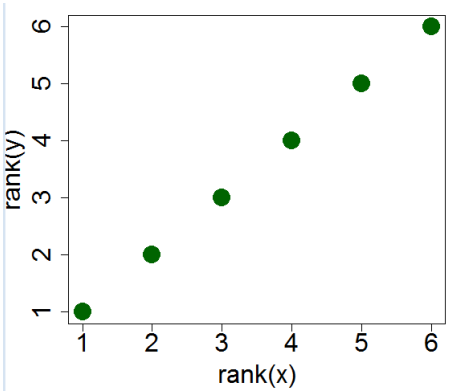
```
x <- c(1,2,4,4.5,6,8)
y <- c(4,5,5.5,6,8,9)
cor(x,y,method='spearman')
plot(rank(x),rank(y),col='darkgreen',pch=19,cex=3,cex.axis=2,cex.lab=2)
```

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

where

- $d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$, is the difference between the two ranks of each observation.
- n is the number of observations

x	1	2	4	4.5	6	8
rg(x)	1	2	3	4	5	6
y	4	5	5.5	6	8	9
rg(y)	1	2	3	4	5	6
rg(x)-rg(y)	0	0	0	0	0	0



x <- c(1, 2, 2, 3)

rank(x)

Novogene 诺禾致源

Providing advanced genomic solutions!

Correlation analysis

Kendall rank correlation coefficient (KCC)

The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}.$$

If there are tied (same value) observations then τ_b is used:

$$\tau_b = \frac{S}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i(t_i-1)/2\right]\left[n(n-1)/2 - \sum_{i=1}^{u} u_i(u_i-1)/2\right]}}$$

- where t_i is the number of observations tied at a particular rank of x and u_i is the number tied at a rank of y.

x	1	3	2	4.5	6	8
rg(x)	1	3	2	4	5	6
y	4	5	5.5	6	8	9
rg(y)	1	2	3	4	5	6

一致的个数：5+3+3+2+1=14
不一致的个数：1

```
a <- c('High', 'Middle', 'Low', 'High', 'Low', 'Middle')
af <- factor(a, ordered=TRUE, levels=c('Low', 'Middle', 'High'), labels=c(1,2,3))
b <- c('High', 'Low', 'Low', 'High', 'Low', 'Middle')
bf <- factor(b, ordered=TRUE, levels=c('Low', 'Middle', 'High'), labels=c(1,2,3))
cor(as.numeric(af), as.numeric(bf), method='kendall')
```

三种相关系数比较

Pearson：连续变量、双变量正态分布、线性相关

Spearman：数据分布未知、连续或离散数据均可

Kendall：数据分布未知、适用于类别变量

cor()计算相关性值

cor.test()检验相关性的显著性

psych包中的corr.test()

```
x <- c(2,2.3,2.5,2.7,3,4)
y <- c(4,4.6,4.8,5.3,6,6.5)
cor(x,y,method = 'pearson')
cor.test(x,y,method = 'pearson')
library(psych)
dt <- data.frame(OTU_1 = c(1,2,2,3,4,6,8), OTU_2 =
c(2,4,4,5,6,8,10), OTU_3 = c(5,6,7,6,8,9,11))
rp <- corr.test(dt, method = 'pearson')
rp$r;rp$p
```

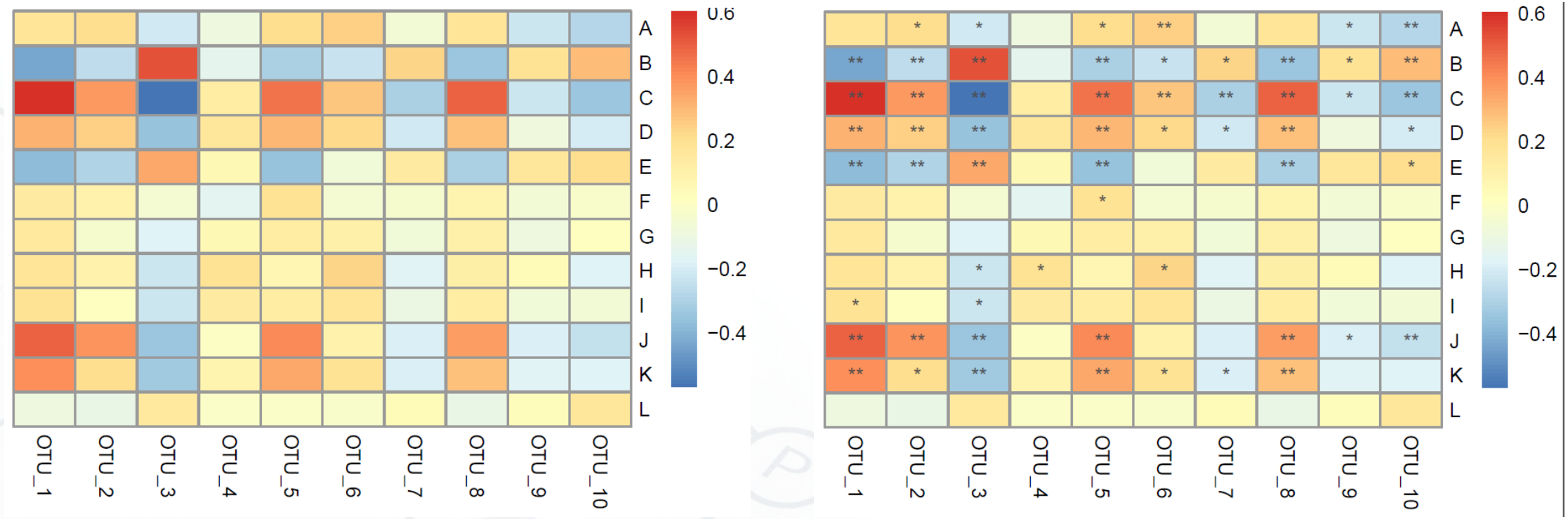
$$t_{corr} = \sqrt{\frac{r^2}{(1-r^2)/(n-2)}}$$

```
STATISTIC <- c(t = sqrt(df) * r/sqrt(1 - r^2))
```

```
        Pearson's product-moment correlation

data:  x and y
t = 6.1542, df = 4, p-value = 0.003537
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6112816 0.9947929
sample estimates:
      cor
0.9510394
```

How to perform correlation analysis

```
library(psych); library(pheatmap)

res <- corr.test(env,otu10, method="spearman",adjust="BH");

pheatmap(res$r, cluster_rows=F, fontsize_number=8, cluster_col=F, fontsize=8)
```



Summary

非约束降维排序

- PCA
- PCoA
- NMDS

差异物种分析

- T-test
- MetaStats
- LefSe

约束性降维排序

- CCA
- RDA
- db-RDA
- Spearman

数据处理 统计分析

图形展示

Novogene 诺禾致源

Providing advanced genomic solutions!

Thanks for your attention!

更多关注，敬请留意：www.novogene.cn