微生物宏基因组学

报告人：王怡雯

中国农业科学院深圳农业基因组研究所

主要内容

- 微生物宏基因组以及基因组数据的相关背景
- 批次效应及其处理
  - 处理难点
  - 处理流程
  - 优化处理批次效应的方法
- 校正后无批次效应的微生物宏基因组数据分析

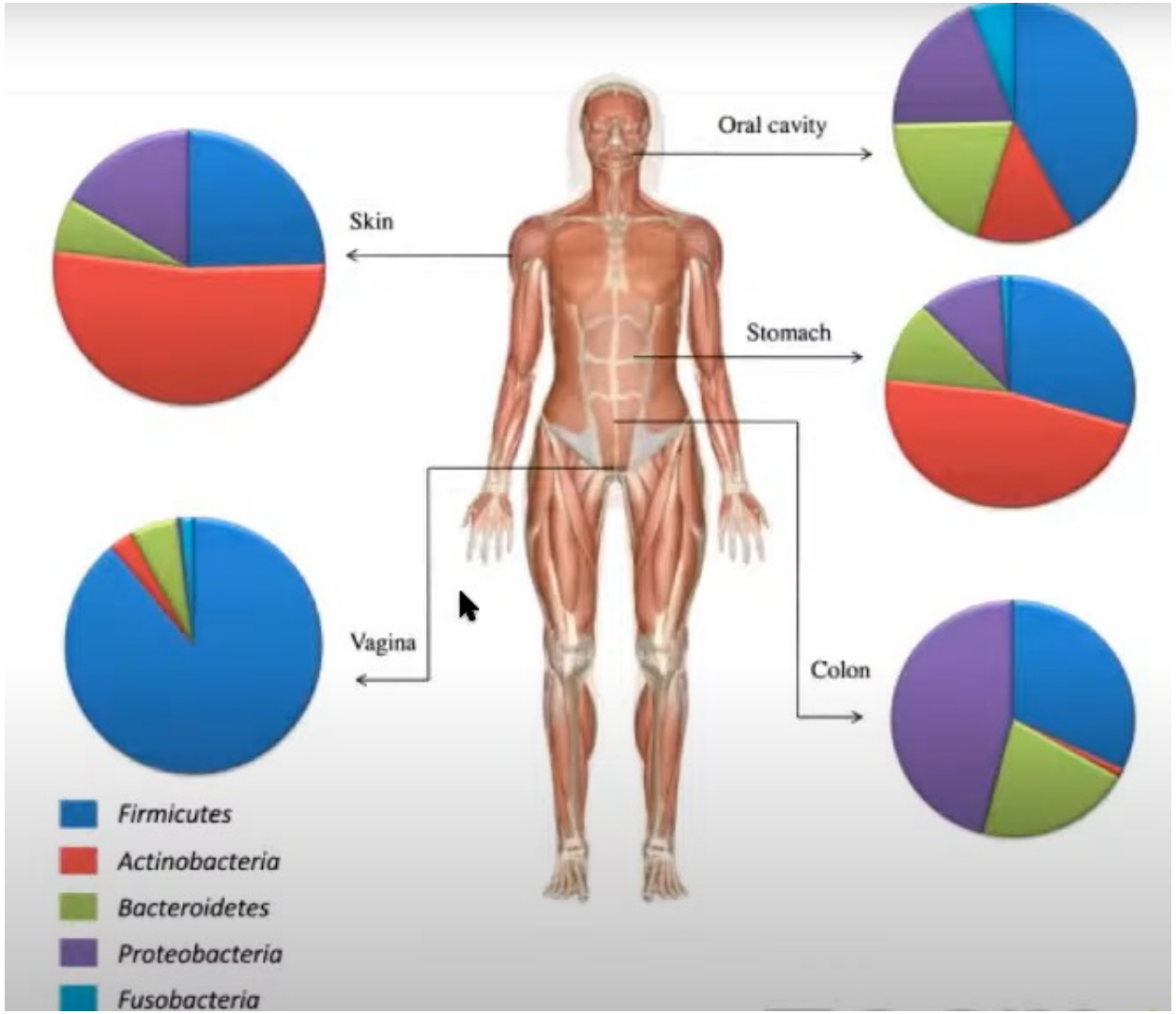
微生物宏基因组以及基因组数据的相关背景

微生物的研究和功能

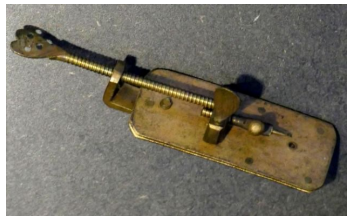
微生物的研究：通过调查研究微生物与其周围环境的联系来了解微生物的功能与角色，从而利用对我们有利的微生物。

- 免疫系统的分化
- 食物的消化和吸收
- 能量的生产和保存
- 新陈代谢调控
- 环境化学品的处理
- 皮肤及粘膜屏障功能的维护
- 预防病原体侵入

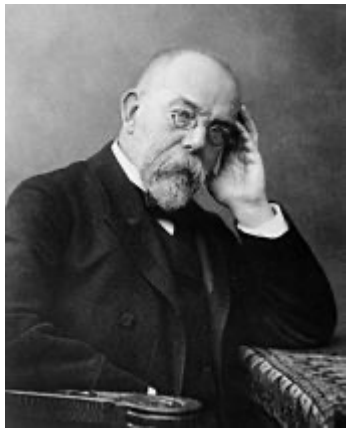
人体不同部位的微生物组成不同



微生物研究的发展史



1665年，安东尼·冯·列文虎克自制显微镜，首次发现微生物



1876年，罗伯特·科赫发现微生物可以致病

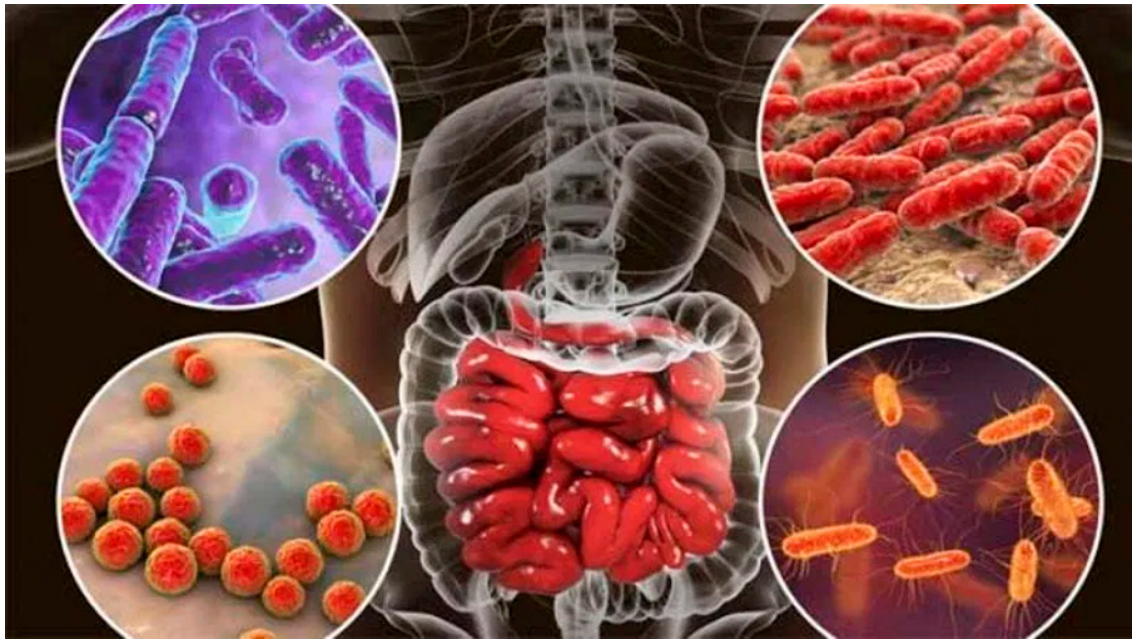
19世纪晚期到现在，人们通过培养、染色、显微镜观察来研究微生物

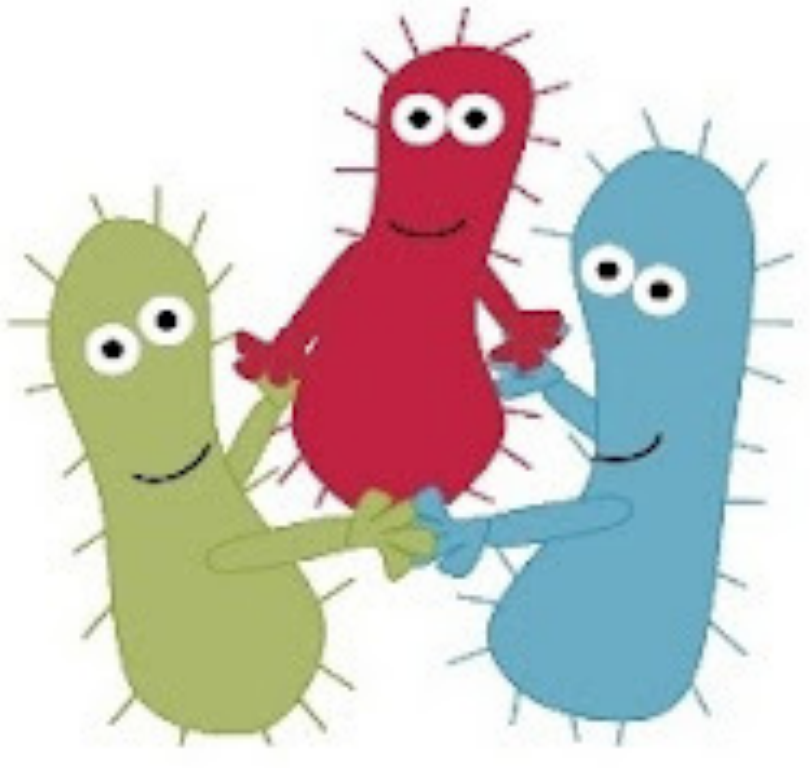
20世纪80年代到现在，难以培养的微生物研究基于DNA测序数据

2007年，人类微生物组计划启动

21世纪，微生物研究基于生物信息学和多组学数据分析

微生物群落





- 同一微环境中不同种的微生物是共存的，有相互关联
- 大部分（> 98% [1]) 的微生物不可分离和在体外培养

[1] Wade (2002). Unculturable bacteria—the uncharacterized organisms that cause oral infections. *Journal of the Royal Society of Medicine*.

微生物宏基因组

- **微生物宏基因组：**在某一个环境中全部的微生物以及这些微生物的基因组整合[1]
- **测序手段：**
  - 标志基因：16S rRNA基因测序

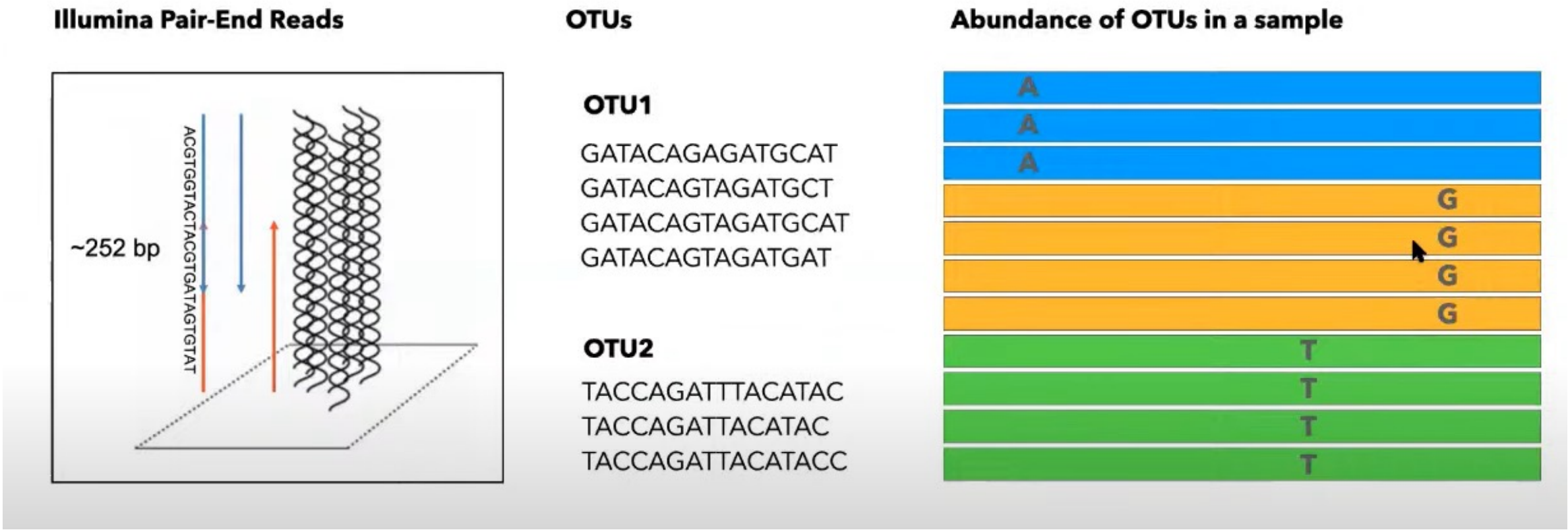
    ---> 属或种水平，~20,000 微生物分类单位[2], ~$80/样本
  - 整个基因组：鸟枪法宏基因组测序

    ---> 种或亚种水平，病毒， ~2,000,000 微生物分类单位[2], ~$200/样本
- **微生物宏基因组数据：**通过测序手段获得的微生物宏基因组信息，一般为OTU表

[1] Marchesi & Ravel (2015). The vocabulary of microbiome research: a proposal. *Microbiome*.

[2] Brumfield, et al (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS One*.

运算分类单位 (OTUs) 的获得

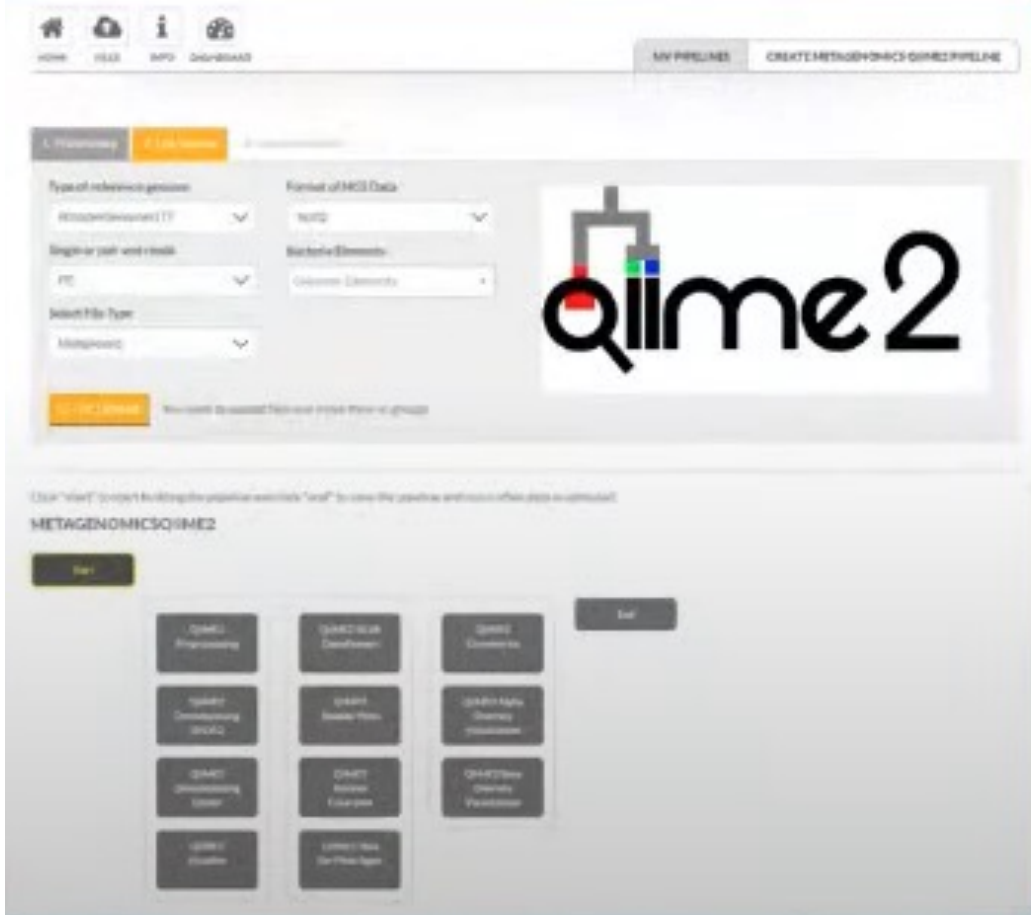


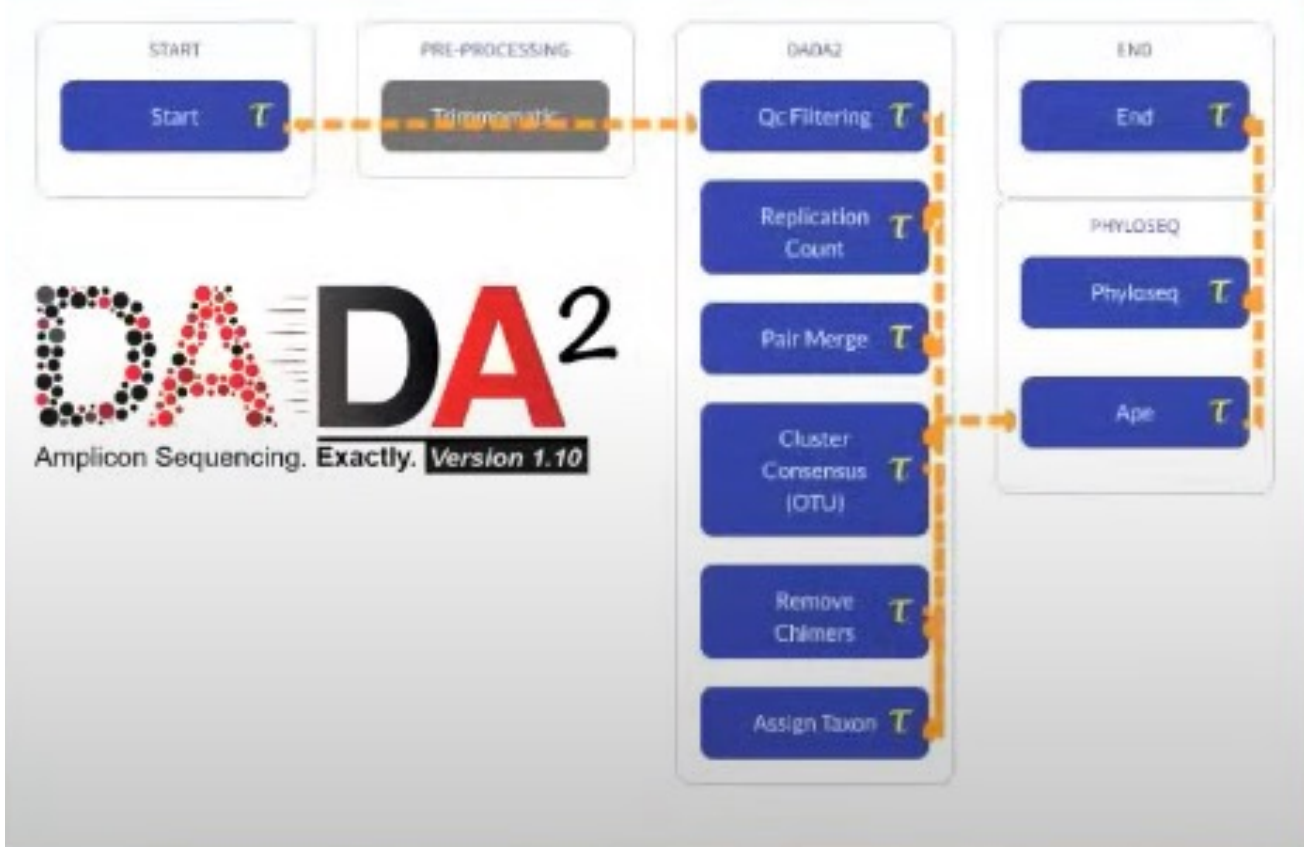
测序数据 (Sequencing reads)

Sequencing reads 聚类:
97% / 99% 相似度

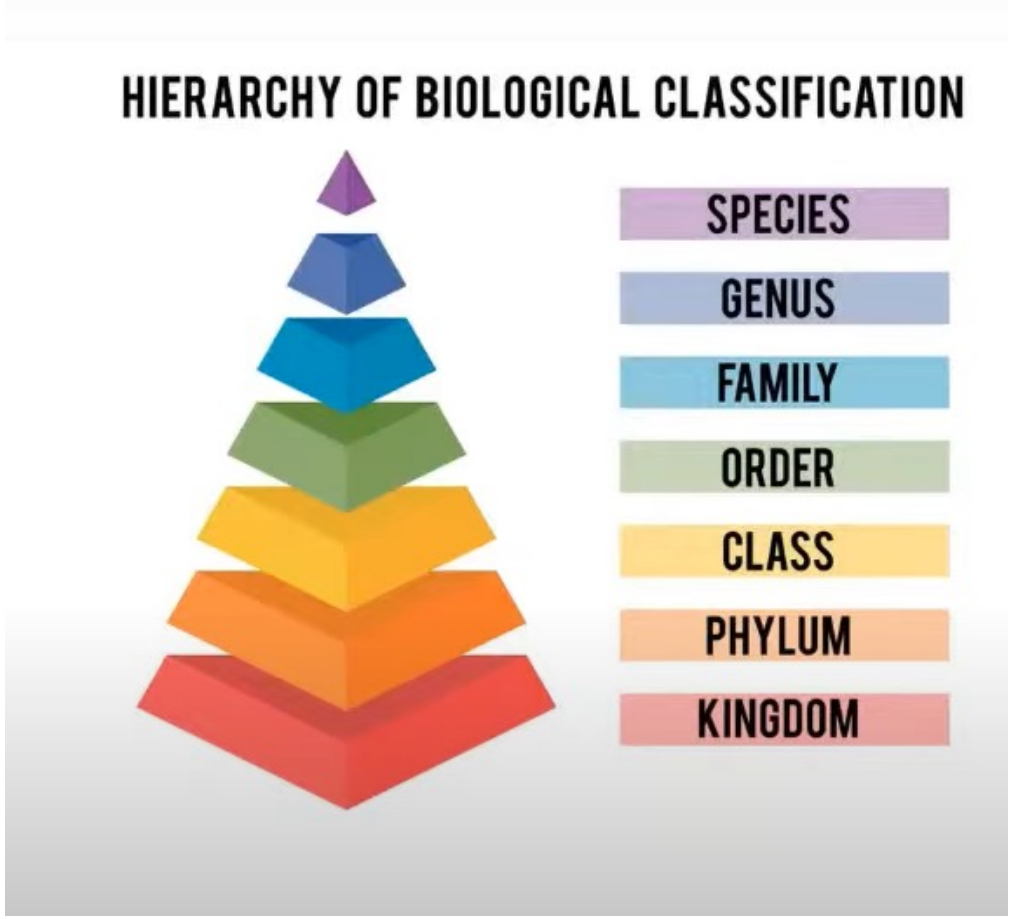
不同 OTUs 的counts

运算分类单位 (OTUs) 的获得

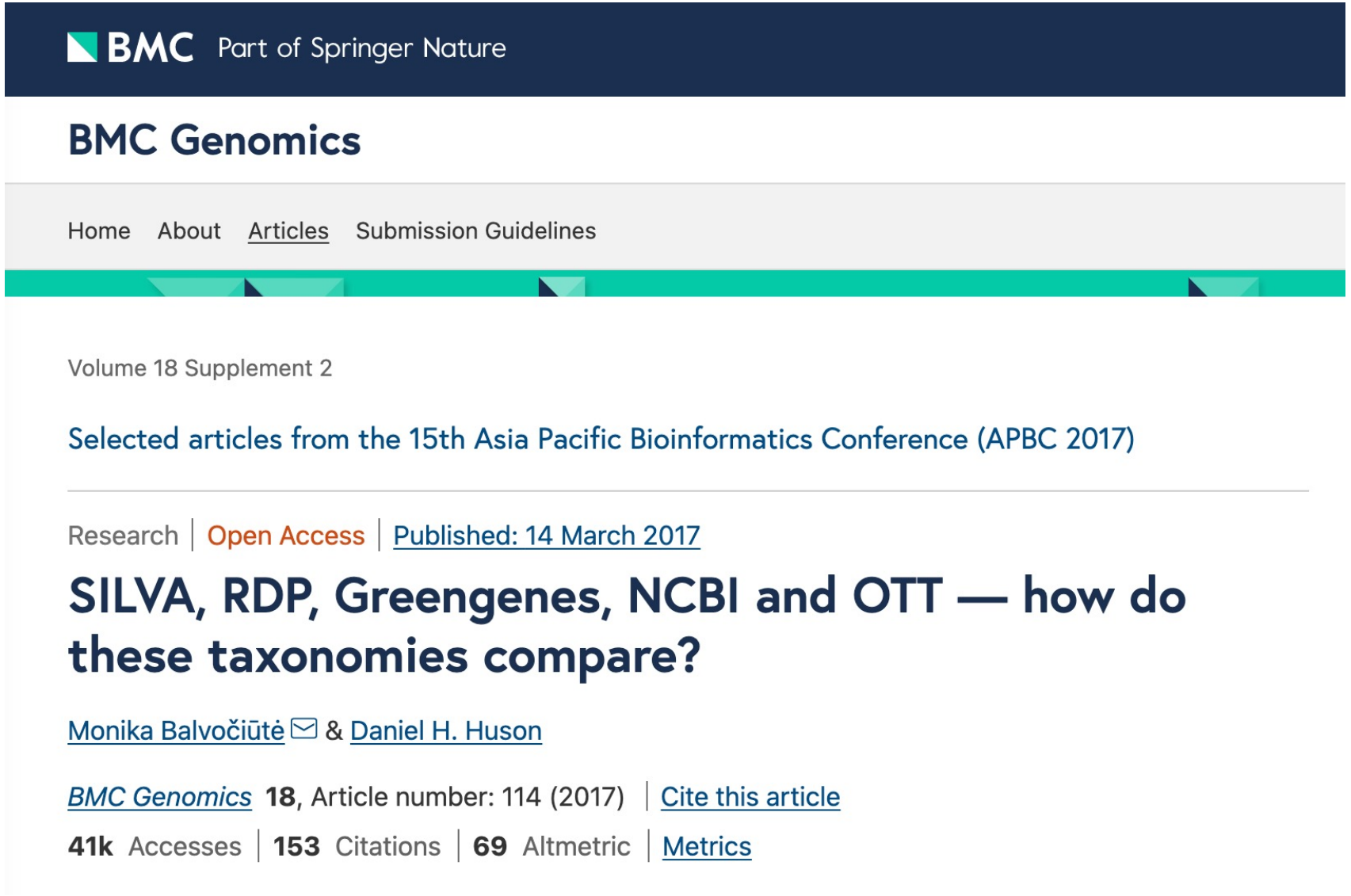




运算分类单位 (OTUs) 匹配数据库 (silva) 找出微生物分类学信息

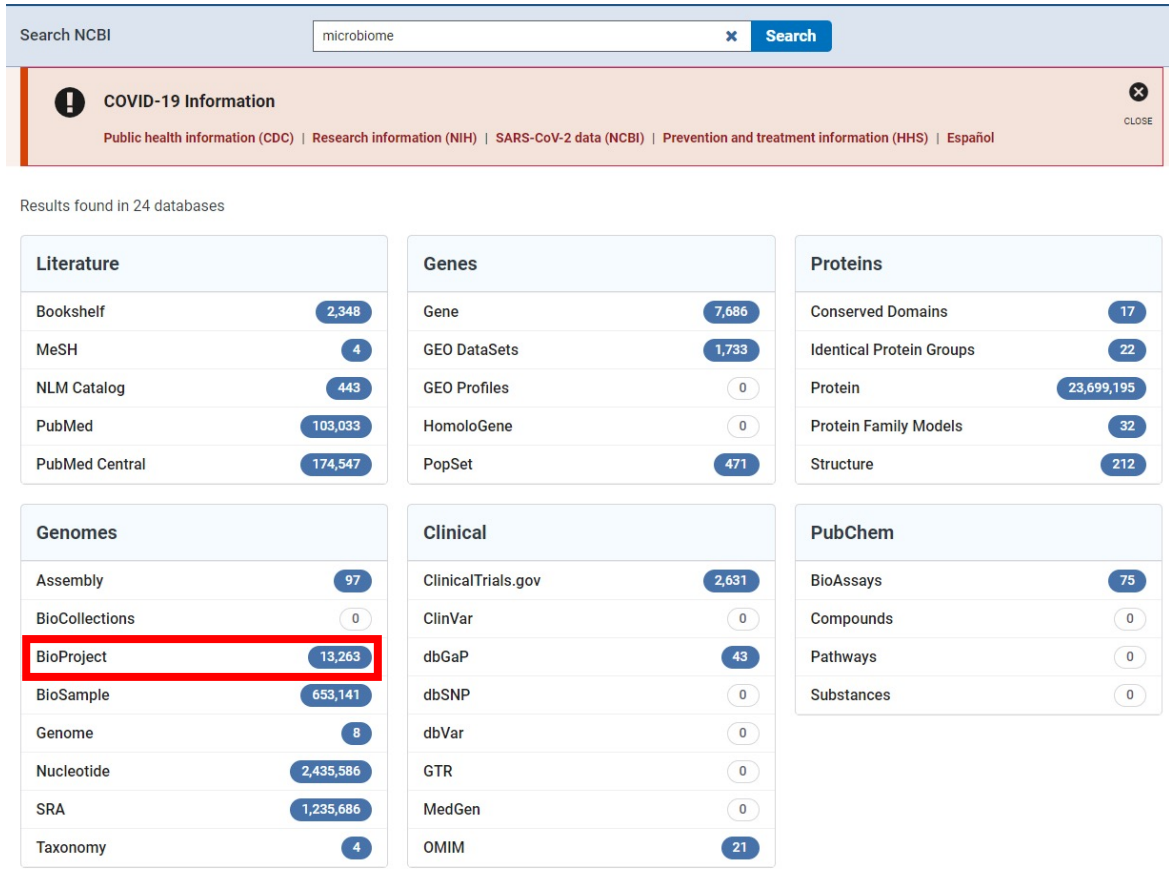


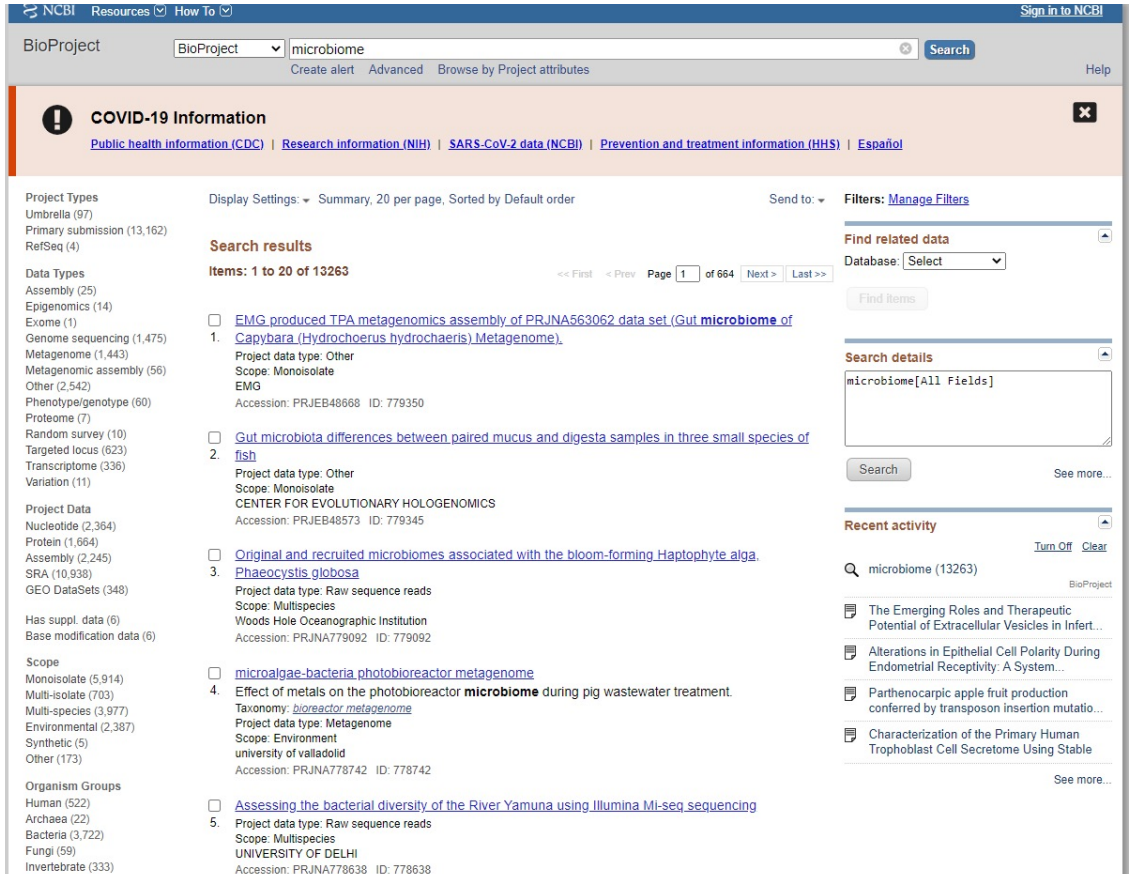
不同的微生物分类学数据库



一起分析的数据推荐选择同一数据库

现有NCBI中收录的所有微生物研究课题

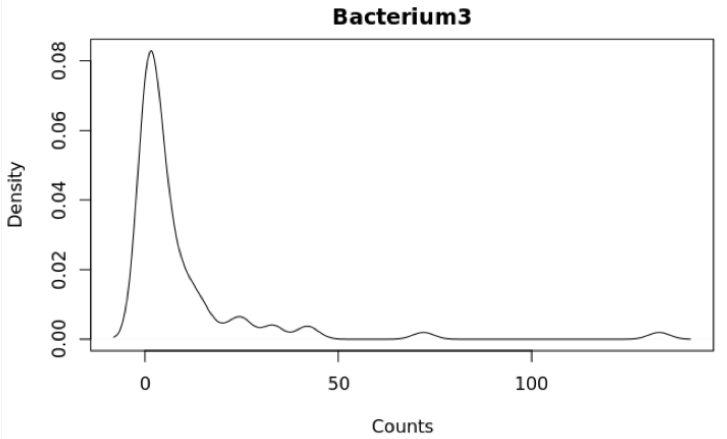




微生物宏基因组数据特点

- 很多 0，过度离散 ---> 偏态分布

	Bacterium1	Bacterium2	Bacterium3	Bacterium4
Sample 1	4	2	0	0
Sample 2	4	1	0	0
Sample 3	5	3	1	0
Sample 4	9	4	0	0
Sample 5	4	8	0	0
Sample 6	2	1	0	1
Sample 7	6	7	0	0

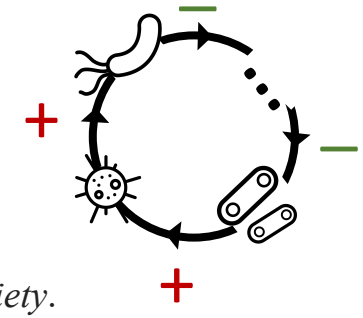


- 文库 (library size) 大小不均一 ---> 样本间比较出现偏差

	Bacterium1	Bacterium2	Bacterium3
Sample A	207	203	590
Sample B	2	3	5

- 存在组分结构 (compositional structure) [1] ---> 数据维度丢失

- 微生物之间存在相互关联 ---> 微生物变量不独立

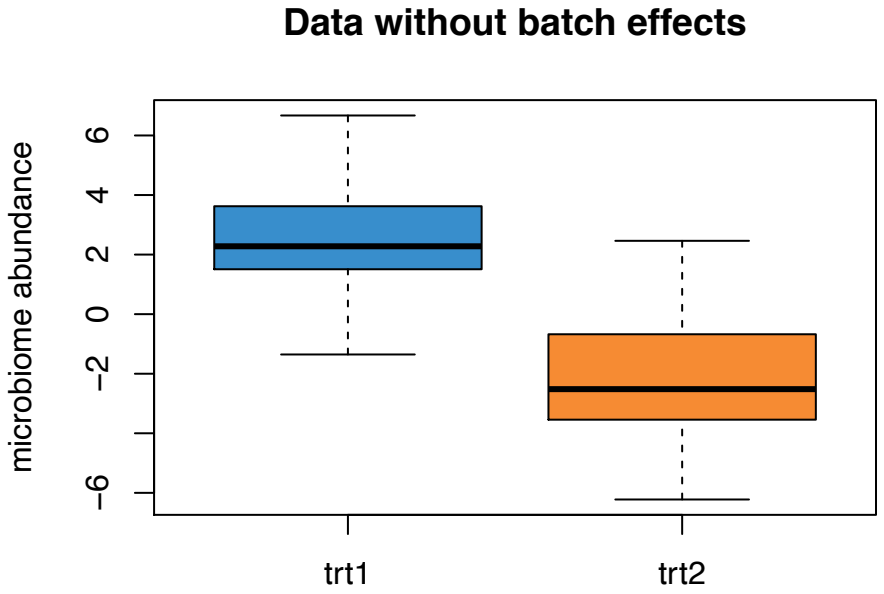


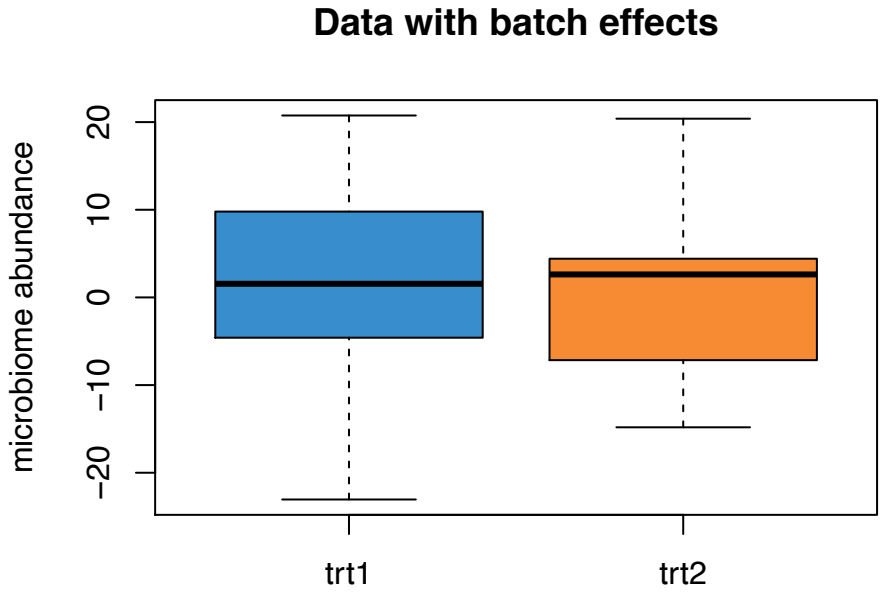
[1] Aitchison (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society*.

批次效应？

批次效应

- **定义**：任何来源的与我们感兴趣的目标差异无关甚至会掩盖目标差异的不必要差异。
- **后果**：





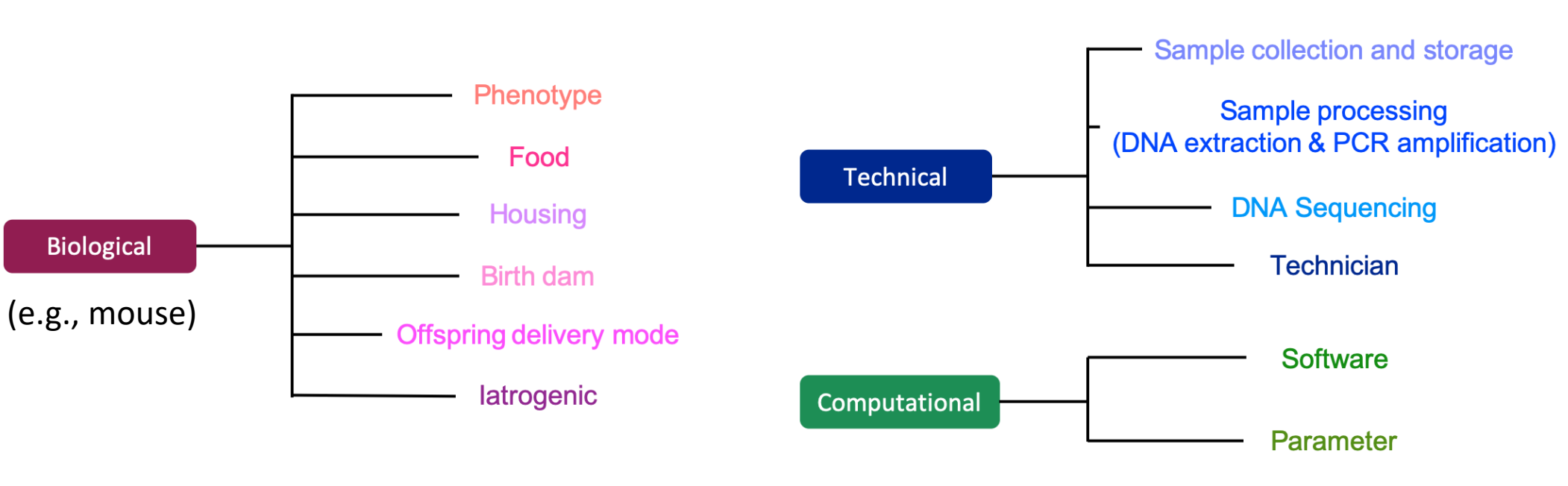
P < 0.001 of the treatment effect in T-test

P > 0.05 of the treatment effect in T-test

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

处理批次效应的难点

- 微生物群落是动态平衡的，所以对批次效应非常敏感
- 批次效应在微生物实验的任何一步都可能发生
- **潜在的批次效应：**



Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

处理批次效应的难点

- 多数方法默认 batch x treatment designs 是平衡的
- 不同的 batch x treatment designs：

Balanced

	Treat 1	Treat 2
Batch 1	10	10
Batch 2	10	10

Nested

	Treat 1	Treat 2
Batch 1	10	0
Batch 2	0	10
Batch 3	10	0
Batch 4	0	10

Unbalanced

	Treat 1	Treat 2
Batch 1	4	16
Batch 2	16	4

Nested ✗

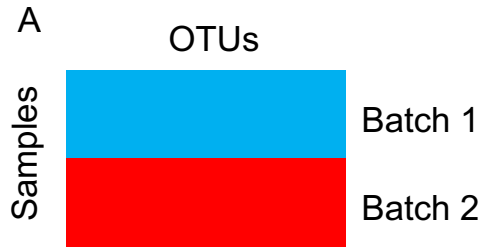
	Treat 1	Treat 2
Batch 1	0	20
Batch 2	20	0

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

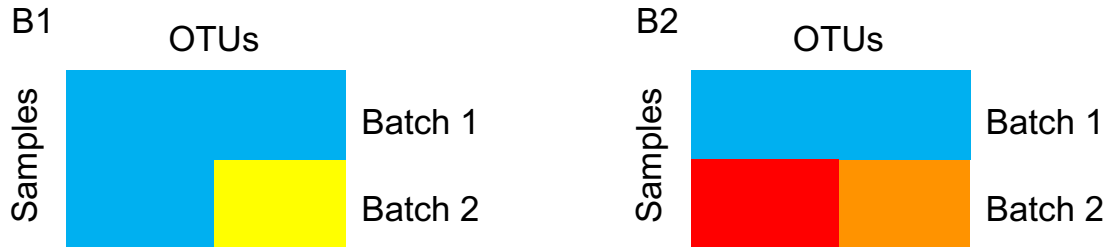
处理批次效应的难点

- 批次效应对不同微生物变量的影响是不同的
- **系统性和非系统性批次效应：**
  - 系统性批次效应对所有微生物变量的影响是一致的
  - 非系统性批次效应对不同微生物变量的影响是不一致的
  - 多数方法默认批次效应是系统性的

Systematic



Non-systematic



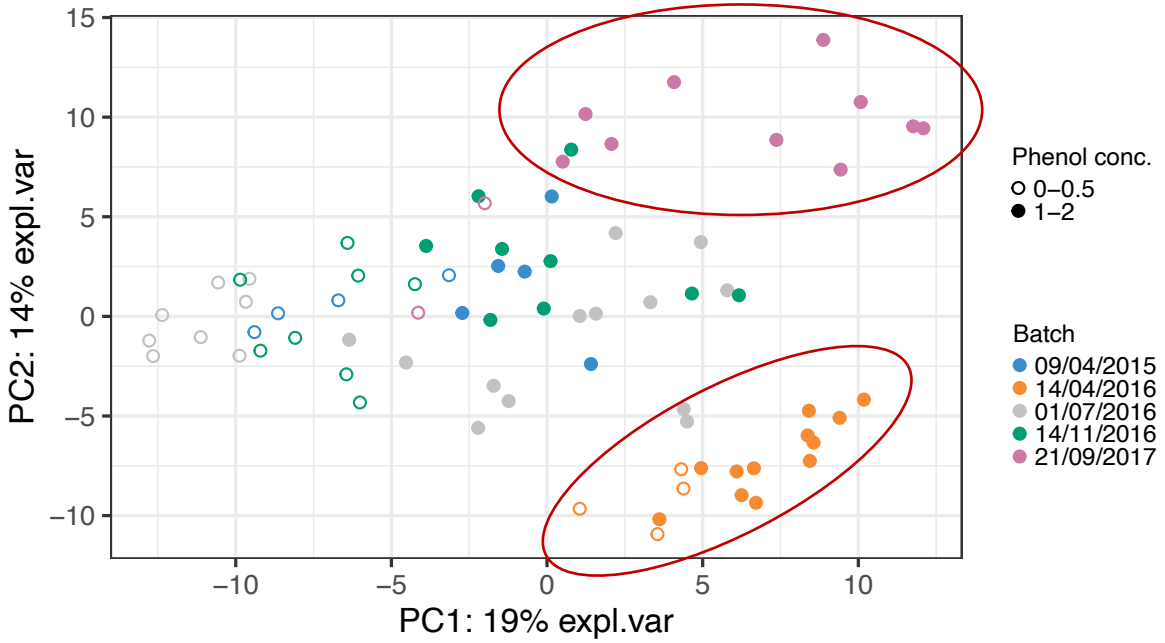
Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

实际案例

厌氧消化数据 (AD data):

- 生物反应器实验：为了提高消化生物垃圾的有效性
- 567个微生物变量和75个样本
- 处理组差异：苯酚浓度不同
- 批次组差异：样本处理日期不同

=> 批次效应检测 & 去除



Chapleur, et al (2016). Increasing concentrations of phenol progressively affect anaerobic digestion of cellulose and associated microbial communities. *Biodegradation*.

海绵数据 (sponge data):

- 不同海绵组织上微生物组分之间的差异
- 24个微生物变量和32个样本
- 处理组差异：海绵组织不同
- 批次组差异：电泳凝胶不同
- 数据特点：完全平衡的batch x treatment design

	Tissue 1	Tissue 2
Batch 1	8	8
Batch 2	8	8

=> 批次效应检测

Sacristan-Soriano, et al (2011). Exploring the links between natural products and bacterial assemblages in the sponge aplysina aerophoba. *Appl. Environ. Microbiol.*

患有 Huntington's disease 的小鼠模型 (HD data):

- 患有HD和健康小鼠的肠道微生物组分之间的差异
- 368个微生物变量和30个样本
- 处理组差异：基因型不同
- 批次组差异：小鼠来自不同鼠笼
- 数据特点：完全不平衡的

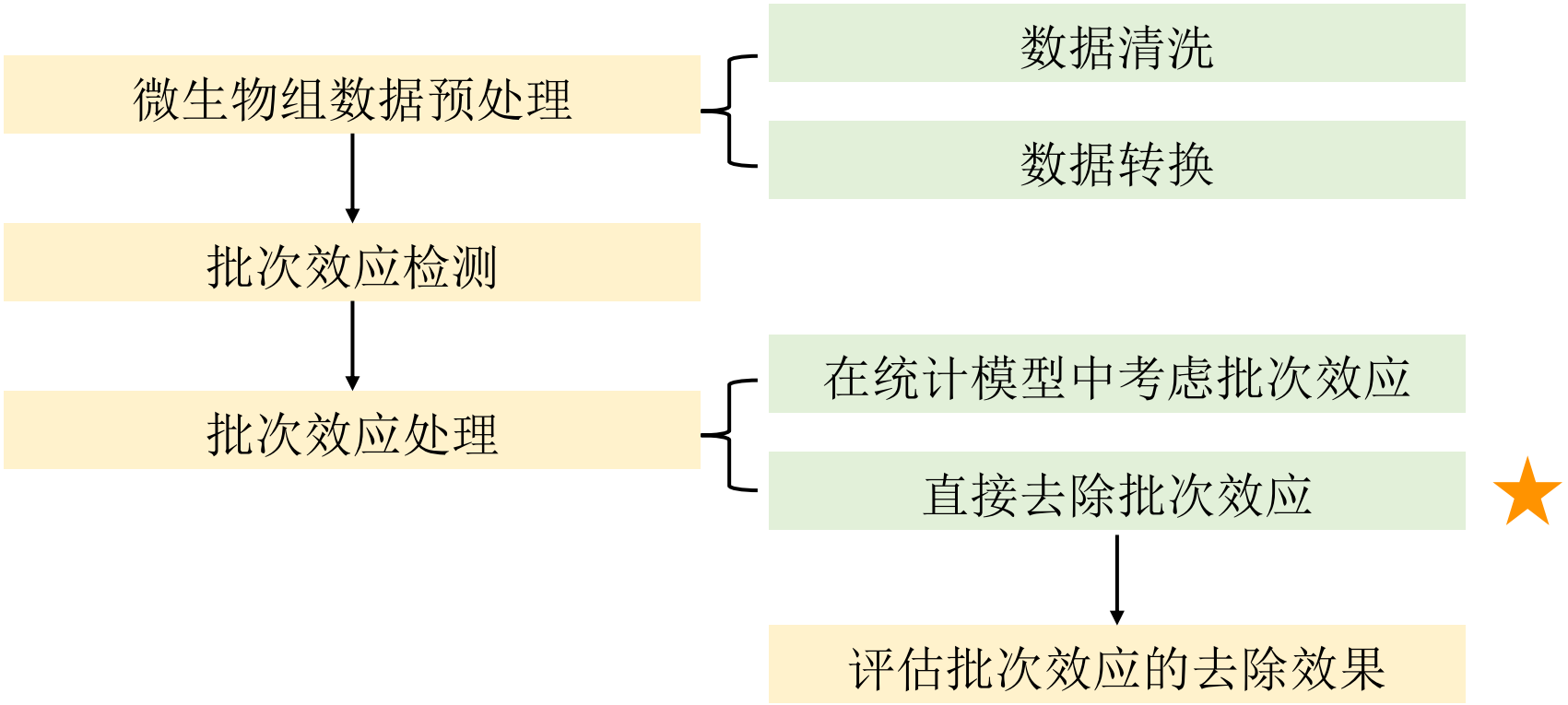
batch x treatment design (nested)

=> 批次效应检测 & 在模型中考虑

Cages\Genotypes	HD	WT
Cage A	2	0
Cage B	3	0
Cage C	2	0
Cage D	0	4
Cage E	0	4
Cage F	0	3
Cage G	3	0
Cage H	3	0
Cage I	2	0
Cage J	0	4

Kong, et al (2016). Microbiome profiling reveals gut dysbiosis in a transgenic mouse model of Huntington's disease. *Neurobiol Dis*.

处理批次效应流程



Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

I. 微生物组数据预处理

目的： 一种务实地应对微生物组数据前三项数据特点的方式

- 很多 0，过度离散
- 文库 (library size) 大小不均一
- 存在组分结构 (compositional structure)

步骤 A： 数据清洗

---> 减轻 0 含量过高的现象和测序误差

方式：去除 counts 始终很低的样本或者变量

例子：AD data  ：含63%的 0 (前)；含38%的 0 (后)

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

I. 微生物组数据预处理

步骤 B：数据转换

---> 应对组分结构 (compositional structure)

微生物组数据存在天然的组分结构：

- 因为一个环境中的资源有限，所以微生物之间存在相互作用
- 测序的样本只是原始生态系统中的一个随机碎片

数据转换方法：去中心对数转换 (Centered log-ratio, CLR)

For n microbial variables, X_j is the abundance of each variable j:

$$CLR(X) = CLR(X_1, ..., X_n) = \left(\log\left(\frac{X_1}{g(X)}\right), ..., \log\left(\frac{X_n}{g(X)}\right)\right), g(X) = (\prod X_j)^{\frac{1}{n}}$$

CLR 附带效果：

- 应对微生物样本文库 (library size) 大小不均一
- 获得接近正态的分布

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

II. 批次效应检测

目的: 检测是否存在批次效应以及是否需要处理批次效应

A. 可视化方法: 很难检测到非常弱的批次效应

- 主成分分析 (PCA)
- 箱线图和密度图
- 热图（Heatmap）

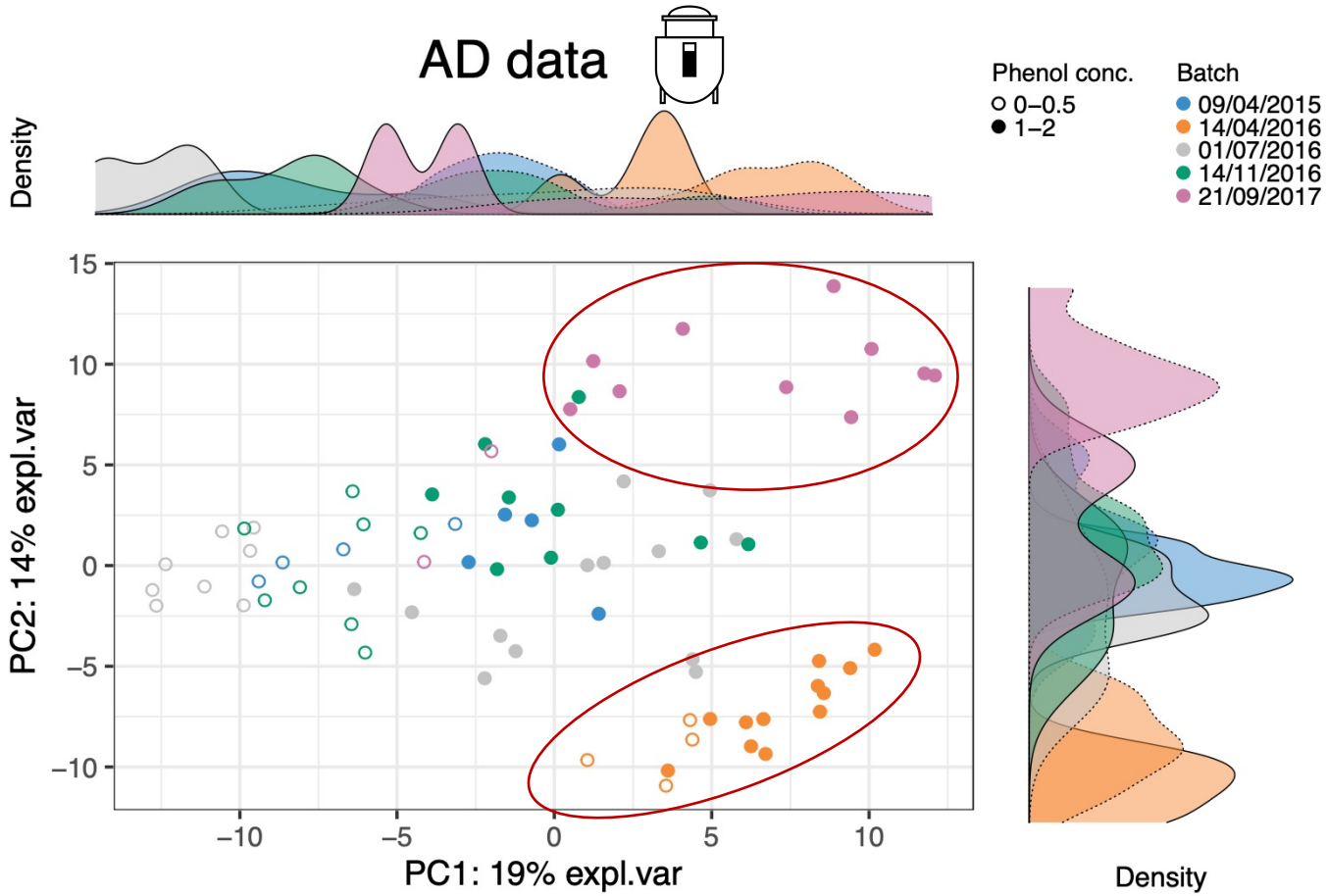
B. 定量方法: 对于批次效应非常敏感

偏冗余分析 (pRDA): 一次性计算涵盖所有微生物变量的批次组及处理组之间的差异在整个数据差异中所占的比例

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

II. 批次效应检测

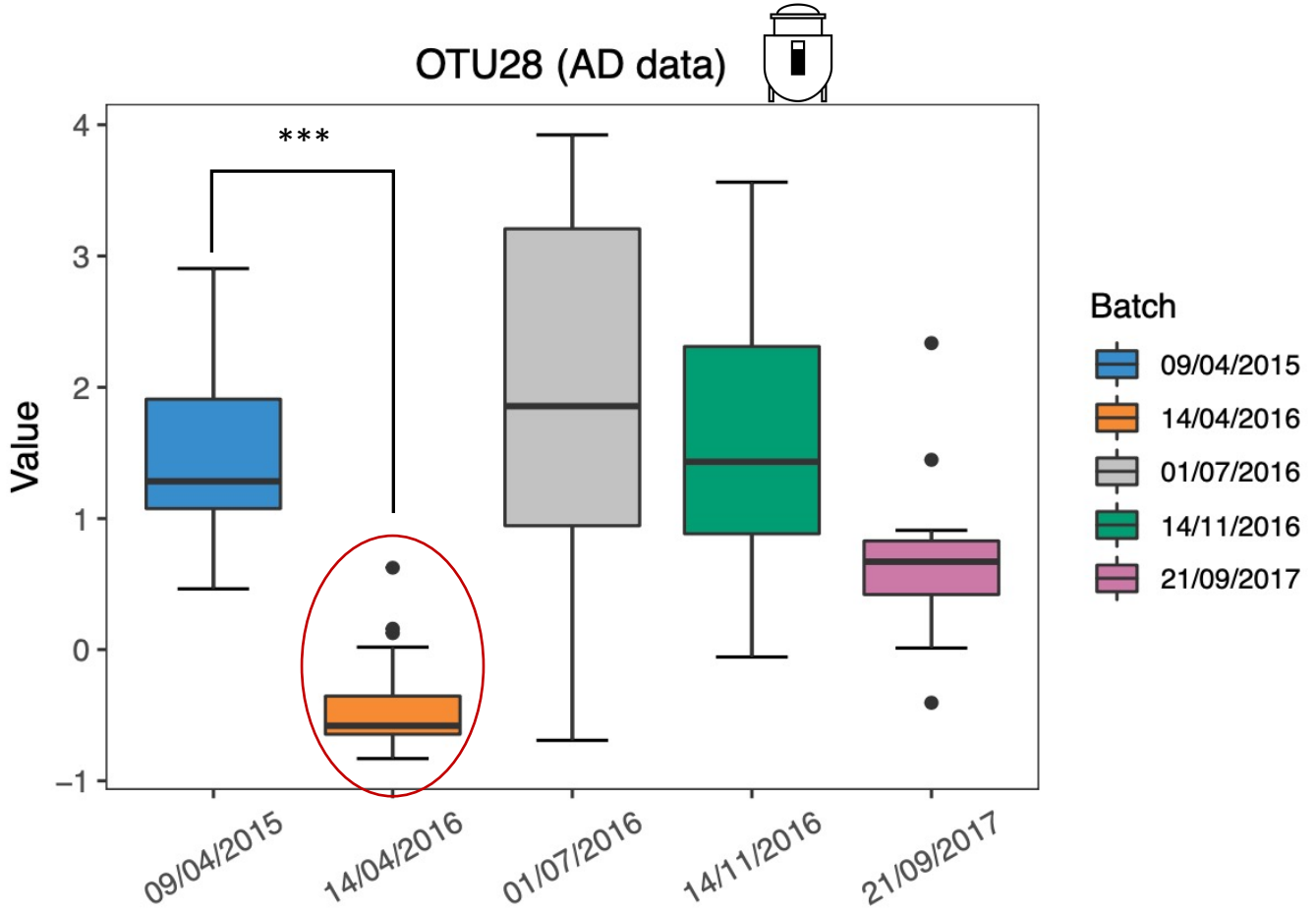
PCA：多元，涵盖所有微生物变量



Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

II. 批次效应检测

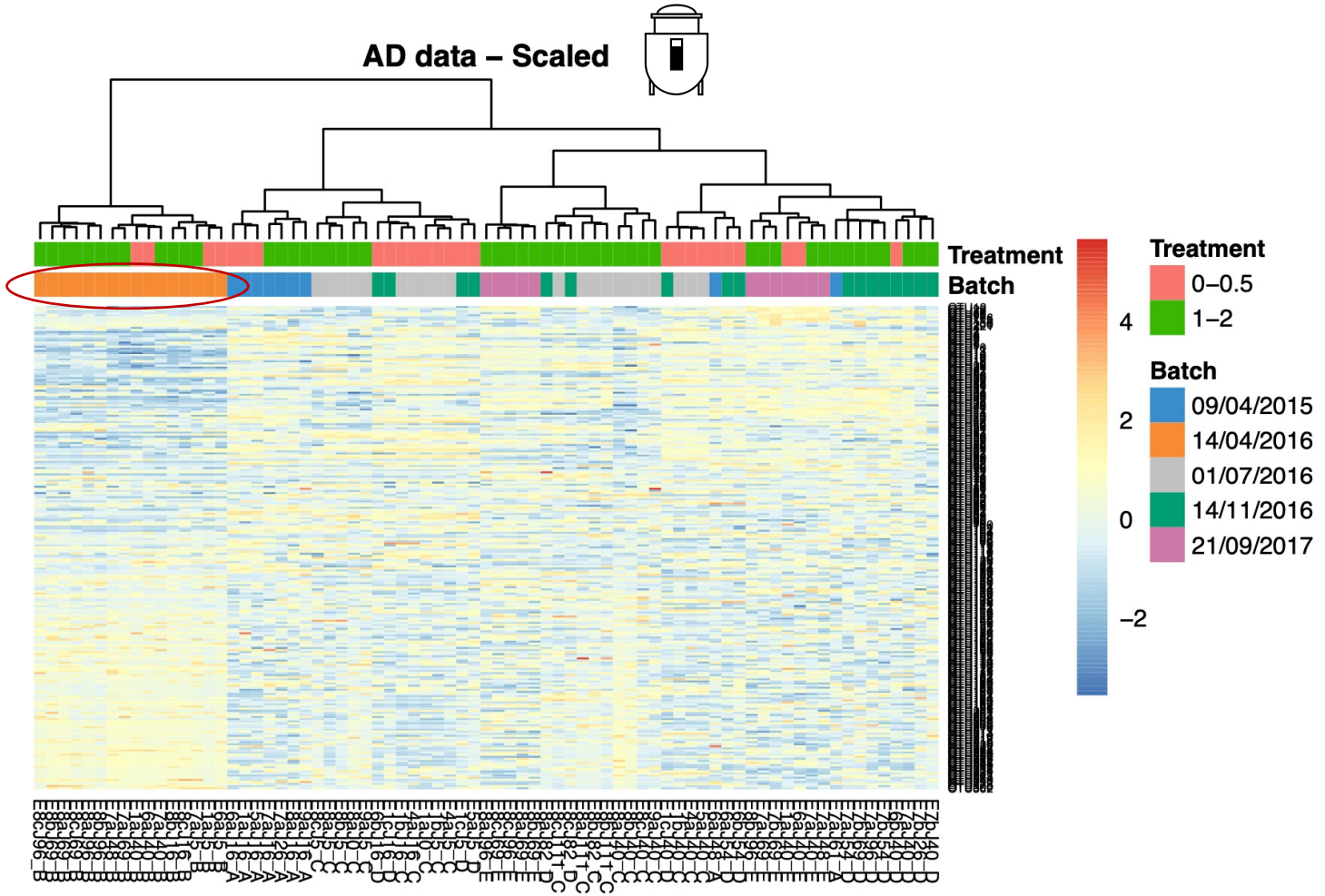
箱线图：一元，单一微生物变量



Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

II. 批次效应检测

热图 (Heatmap)：涵盖所有微生物变量和样本

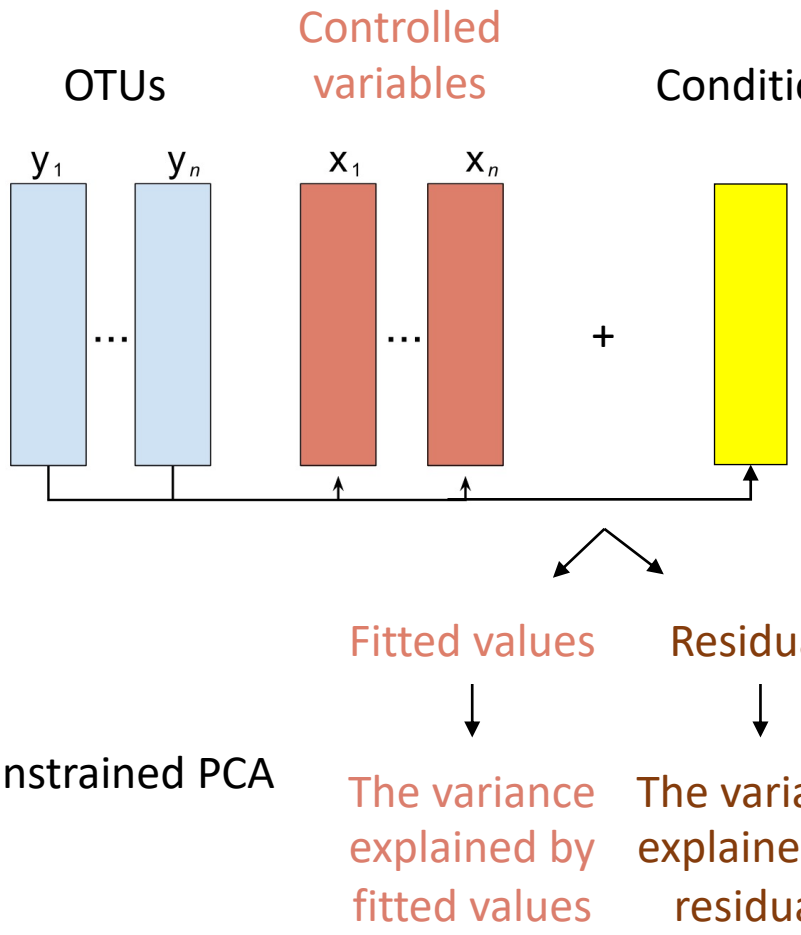


Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

II. 批次效应检测

偏冗余分析(pRDA): 多元

多重线性回归



输出结果:涵盖所有微生物变量的批次组及处理组之间的差异（作为控制变量拟合获得）在整个数据差异中所占的比例

Borcard, et al (1992). Partialling out the spatial component of ecological variation. *Ecology*.

II. 批次效应检测

偏冗余分析(pRDA): 能够暗示 batch x treatment designs 是否平衡

大致平衡的 batch x treatment design (**AD data** )

Dates\Phenol conc.	0-0.5	1-2
09/04/2015	4	5
14/04/2016	4	12
01/07/2016	8	13
14/11/2016	8	9
21/09/2017	2	10

The intersection variance 暗示design的不平衡程度:

=> 批次和处理效应具有相关性

```
##                          Df R.squared Adj.R.squared Testable
## [a] = Treat | Batch       1        NA     0.08943682     TRUE
## [b]                       0        NA     0.01296248    FALSE
## [c] = Batch | Treat       4        NA     0.26604420     TRUE
## [d] = Residuals          NA        NA     0.63155651    FALSE
```

II. 批次效应检测

偏冗余分析(pRDA)结果：

完全平衡的 batch x treatment design (**sponge data**)

	Tissue 1	Tissue 2
Batch 1	8	8
Batch 2	8	8

没有 intersection variance:

=> 批次和处理效应是相互独立的

```
##                       Df R.squared Adj.R.squared Testable
## [a] = Treat | Batch    1        NA     0.16572246     TRUE
## [b]                    0        NA    -0.01063501    FALSE
## [c] = Batch | Treat    1        NA     0.16396277     TRUE
## [d] = Residuals       NA        NA     0.68094977    FALSE
```

II. 批次效应检测

偏冗余分析(pRDA)结果：

完全不平衡 (nested) 的 design (**HD data** ⚖)

Cages\Genotypes	HD	WT
Cage A	2	0
Cage B	3	0
Cage C	2	0
Cage D	0	4
Cage E	0	4
Cage F	0	3
Cage G	3	0
Cage H	3	0
Cage I	2	0
Cage J	0	4

```
##                          Df R.squared Adj.R.squared Testable
## [a] = Treat | Batch       0        NA -2.220446e-16    FALSE
## [b]                       0        NA  9.730583e-02    FALSE
## [c] = Batch | Treat       8        NA  1.608205e-01     TRUE
## [d] = Residuals          NA        NA  7.418737e-01    FALSE
```

没有 treatment variance 和存在大量的 intersection variance:

=> 批次和处理效应是共线的

III. 批次效应处理

- 在模型中考虑批次效应：
  在统计模型中将批次效应设定为协变量

- 直接去除批次效应：
  从原始数据中去除批次效应，从而获得一个没有批次效应的数据

- 多元的方法能够考虑到微生物变量之间的相互关联，而不是完全独立的.

- 多元 vs.一元 :
  - 多元: 所有变量一起处理
    例如 *phyloseq* 中的 PCA, CCA等
  - 一元: 每个变量单独处理.
    例如在 *DESeq2, edgeR* 中的差异性表达分析

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

III. 批次效应处理

在模型中考虑批次效应的方法：

优点:
能够考虑微生物组数据的特点以及批次和处理效应之间的相关性

缺点:
- 局限于差异性表达分析 (只能获得各个微生物变量的P值)
- 很难直接评估批次效应的去除效果

例如线性模型 (linear models)

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

III. 批次效应处理

在模型中考虑批次效应的方法：

A. 为微生物组数据而设计 (应用于 counts 数据):

- Cumulative-Sum Scaling normalisation + Zero-inflated Gaussian mixture model:
  - 能够应对文库 (library size) 大小不均一
  - 能够应对组分结构 (compositional structure)
  - 能够考虑因采样不足而导致的数据缺失
- Bayesian Dirichlet-multinomial regression meta-analysis:
  - 能够拟合微生物变量之间的相互依赖性
  - 能够承受微生物组数据中 0 过多的现象

B. 从其他领域引入并修改后用于微生物组数据分析的方法 (适用于 CLR 转换后的数据):

- Linear regression: 能够处理 nested batch x treatment designs (**HD data** )
- Surrogate variable analysis (SVA): 在不需要额外信息的情况下估算未知的批次效应
- Remove unwanted variation in 4 steps (RUV4): 能够估算未知的批次效应，但是需要能够反映批次差异的阴性对照变量 (negative control variables) 和重复样本 (sample replicates)

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

III. 批次效应处理

直接去除批次效应：

- 适用于 CLR 转换后的数据
- 优点: 校正后的数据可用于后续的所有分析
  - 降维 (dimension reduction)
  - 可视化 (visualisation)
  - 聚类分析 (clustering)
  - 变量选择 (variable selection)
- 缺点:
  - 并不能在模型中考虑微生物组数据的特征
  - 并不能处理批次和处理效应可能存在的相关性

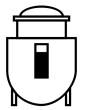
例如 ComBat

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

III. 批次效应处理

直接去除批次效应：

- removeBatchEffect:
  - 从线性模式中通过回归去除掉批次效应
  - 一元
- ComBat:
  - 默认批次效应是系统性的
- Percentile normalisation:
  - 只适用于有空白对照组的实验数据分析
  - 一元
- Remove Uwanted Variation-III:
  - 需要能够反映批次差异的阴性对照变量 (negative control variables) 和重复样本 (sample replicates)
  - 多元

所列方法都应用于实际案例 (AD data )

Wang & Lê Cao (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*.

IV. 评估批次效应的去除效果

- 检测批次效应的方法:
  - 可视化: 主成分分析(PCA), 箱线图(boxplots), 密度图(density plots), 热图(heatmap)
  - 偏冗余分析(pRDA): 涵盖所有变量的各个效应所解释差异在总数据差异中的占比
- 其它方法:

  R^2 from one-way ANOVA: 每个变量中，各个效应所解释的差异占比

用 PLSDA-batch 去除批次效应

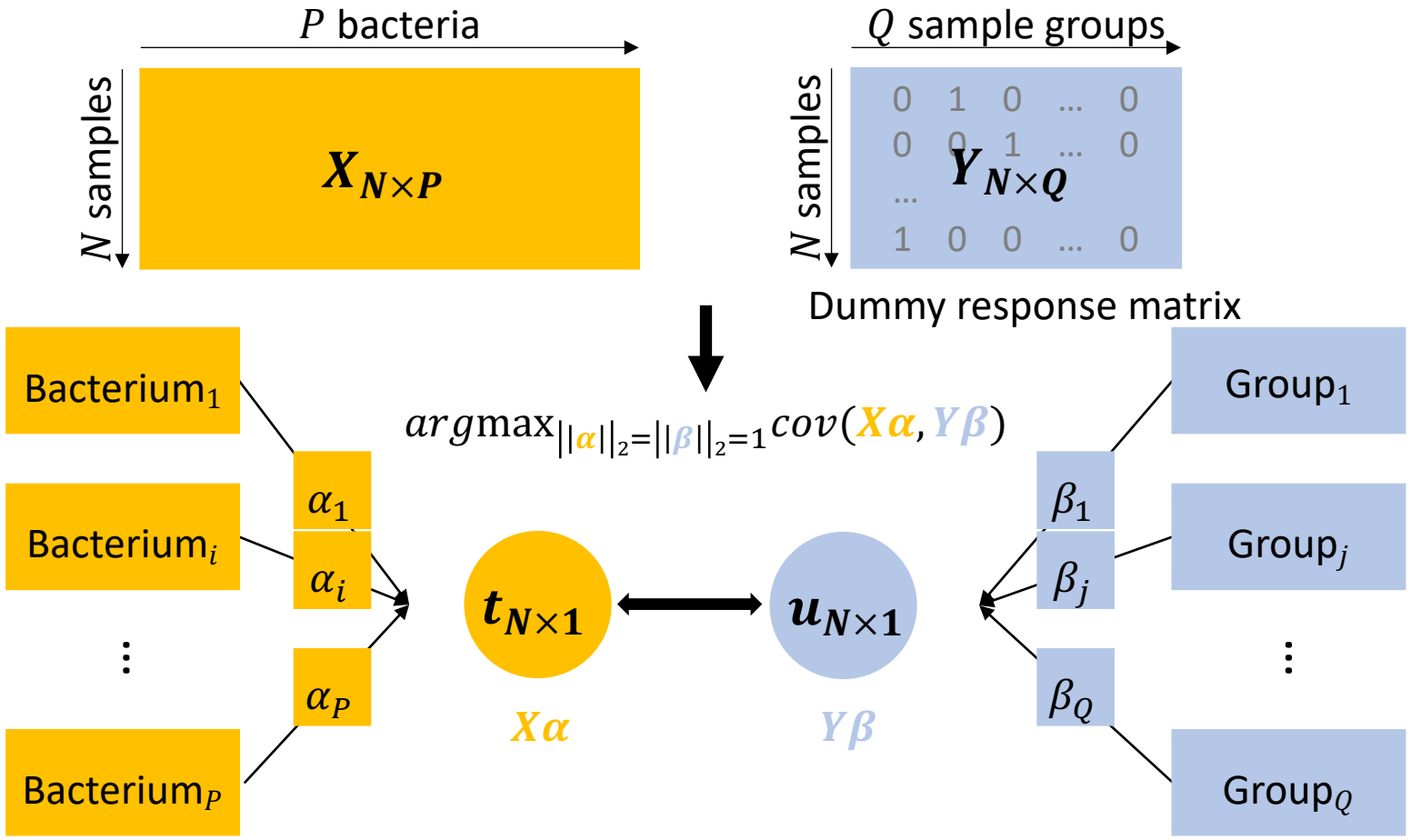
PLSDA-batch

- 基于偏最小二乘法判别分析 (PLS-DA)
- 非参数：可以应对偏态分布
- 多元：可以应对不独立的微生物变量
- 结合centered log ratio (CLR)去中心对数转换：可以应对文库(library size) 大小不均一和组分结构
- 可以应对非系统性的批次效应

Barker & Rayens (2003). Partial least squares for discrimination. *Journal of Chemometrics*

Lê Cao, et al (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics*

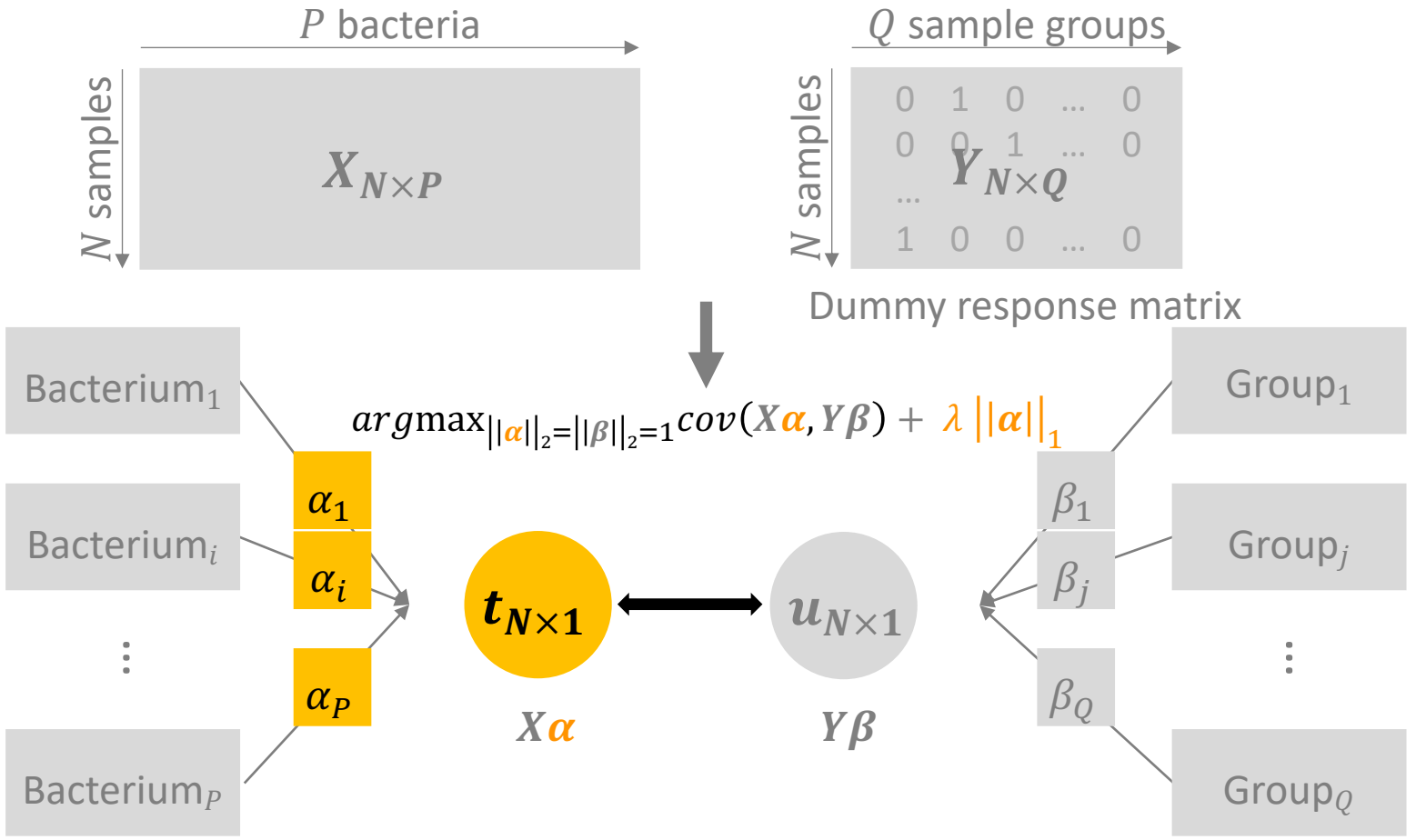
PLSDA: 用于定位隐藏的组分



- 组分不限于一个:
  $\boldsymbol{T} = \boldsymbol{t_1}, \boldsymbol{t_2}, \ldots, \boldsymbol{t_H}$
- 隐藏组分 $\boldsymbol{T}$ 代表的是 $\boldsymbol{X}$ 中存在的组间差异，这个分组信息储存于 $\boldsymbol{Y}$

Barker & Rayens (2003). Partial least squares for discrimination. *Journal of Chemometrics*

Sparse PLSDA

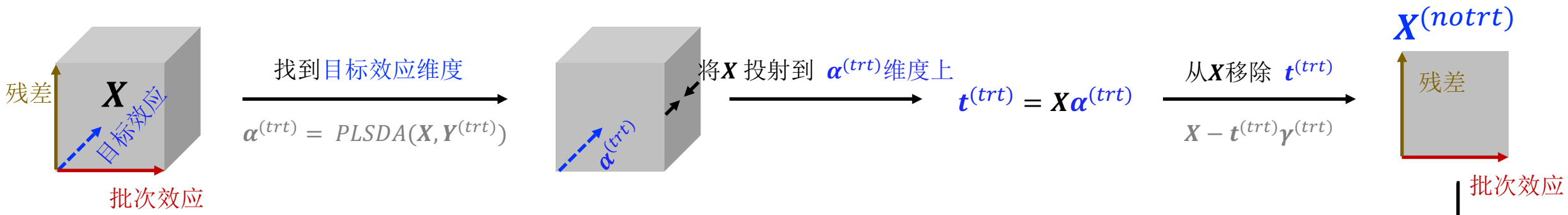


- 加入 l_1 penalty λ 来控制非 0 的α系数数量
- 筛选 X 中最能区分应答变量矩阵 Y 里已给分组之间差异的变量
- 通过多重交叉验证 (repeated cross validation) 选择 λ

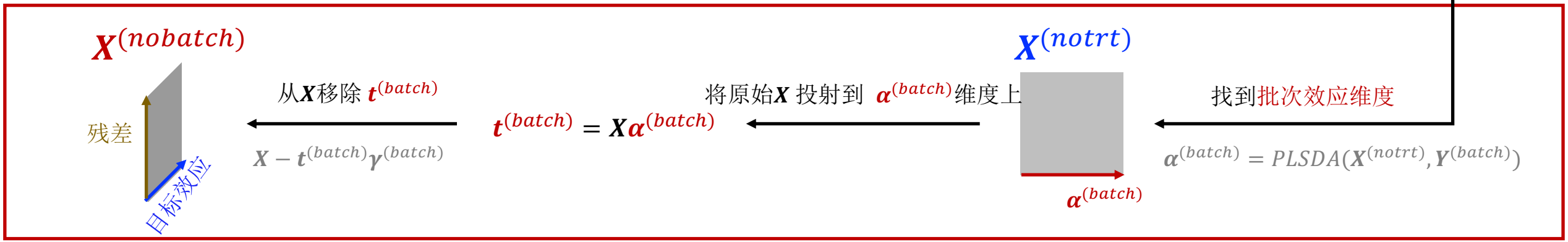
Lê Cao, et al (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics*

PLSDA-batch 基本原理

第1步：预存目标差异



第2步：移除批次差异



$\gamma^{(trt)}$是 X 关于 $t^{(trt)}$的回归系数；$\gamma^{(batch)}$是 X 关于 $t^{(batch)}$的回归系数

Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv*.

PLSDA-batch 衍生方法

- Weighted PLSDA-batch (wPLSDA-batch)

  非平衡的 batch x treatment design : 批次和处理效应存在相关性

	Treat 1	Treat 2
Batch 1	4	16
Batch 2	16	4

  但是 PLSDA-batch 估算的批次组和处理组差异组分是独立和正交的

  => 并不能计算 intersection variance 里批次组和处理组差异各自所占的比例

  wPLSDA-batch 根据每个样本所在的批次组和处理组的 group size 给每个样本增加权重

- Sparse PLSDA-batch (sPLSDA-batch)
  - 降低 PLSDA-batch 中由于过度拟合可能形成的虚假处理组差异
  - sPLSDA-batch 加入 l_1 penalty 来限制对非相关的微生物变量的选择

Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv*.

参数选择

- 需要保留的处理组差异的组分数量:

=>可以100%解释 $\boldsymbol{Y}^{treatment}$ 矩阵的差异

- 需要去除的批次组差异的组分数量:

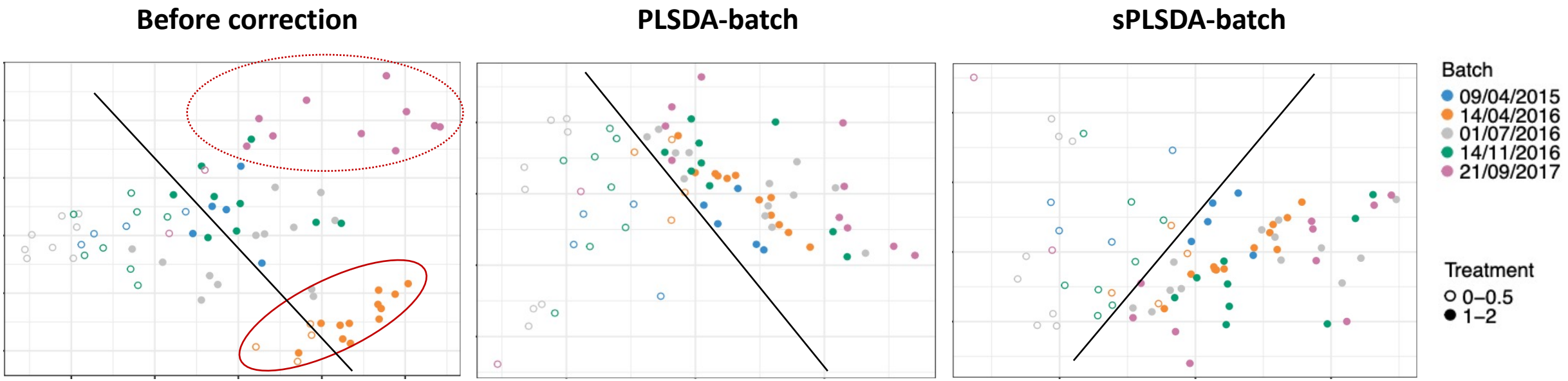
=> 可以100% 解释$\boldsymbol{Y}^{batch}$ 矩阵的差异

- sPLSDA-batch: 每个组分中非零变量的数量:

=> 在多重交叉验证中错误率最低的变量

Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv.*

AD data 分析结果

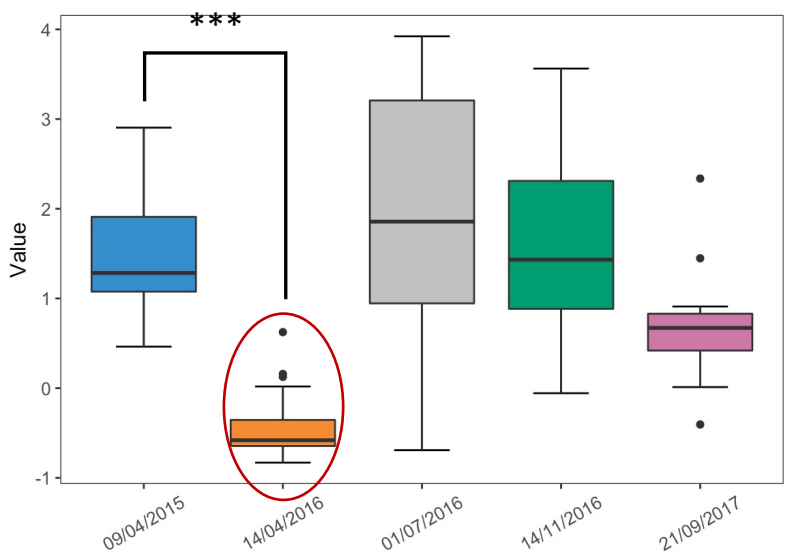


- 批次效应被清除
- 处理组差异更清晰

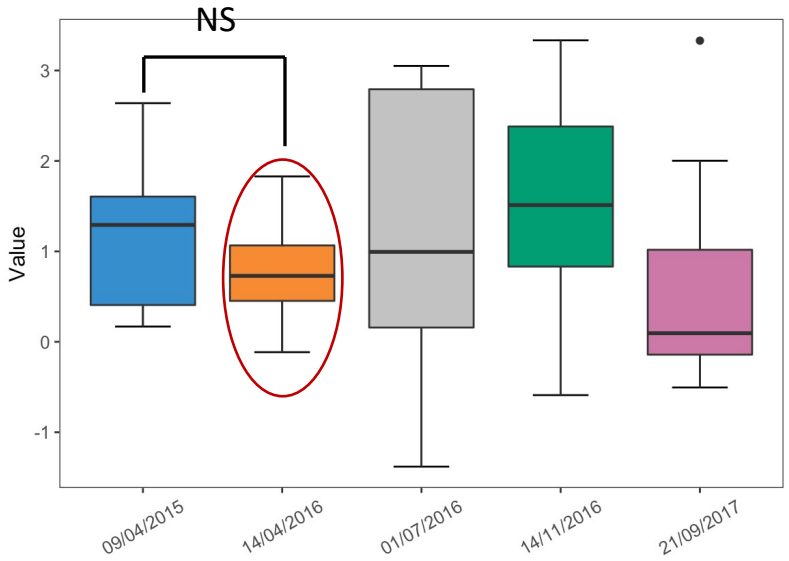
Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv*.

AD data 分析结果

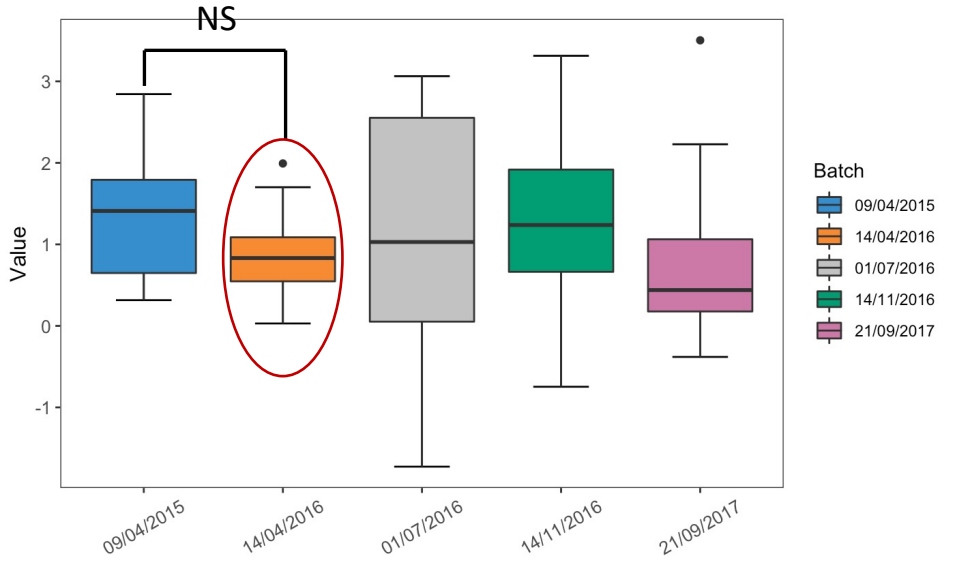
Before correction



PLSDA-batch



sPLSDA-batch

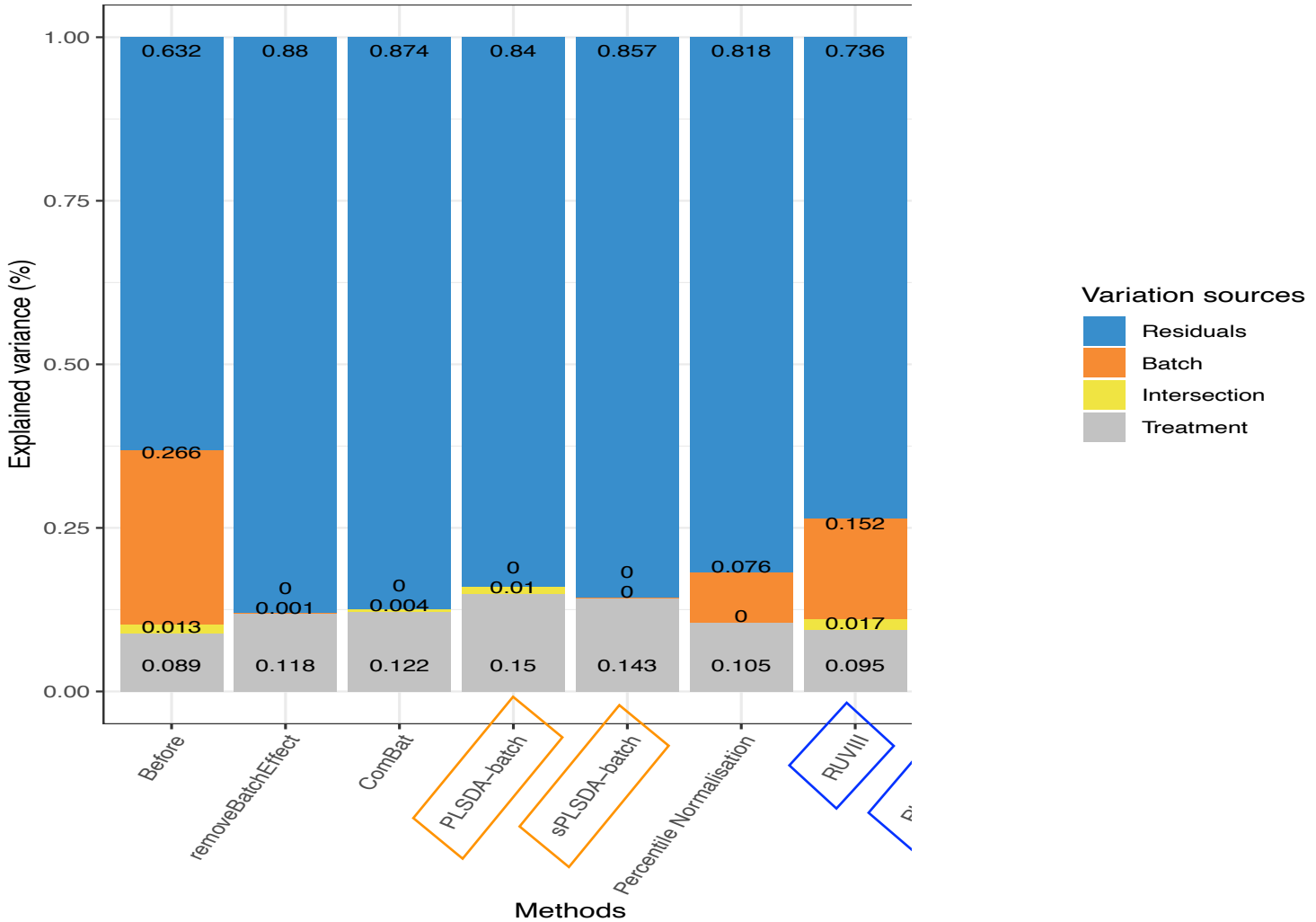


批次效应校正后批次间的差异被去除

Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv.*

AD data 分析结果

用 pRDA 计算的各个来源的差异（涵盖所有微生物变量）

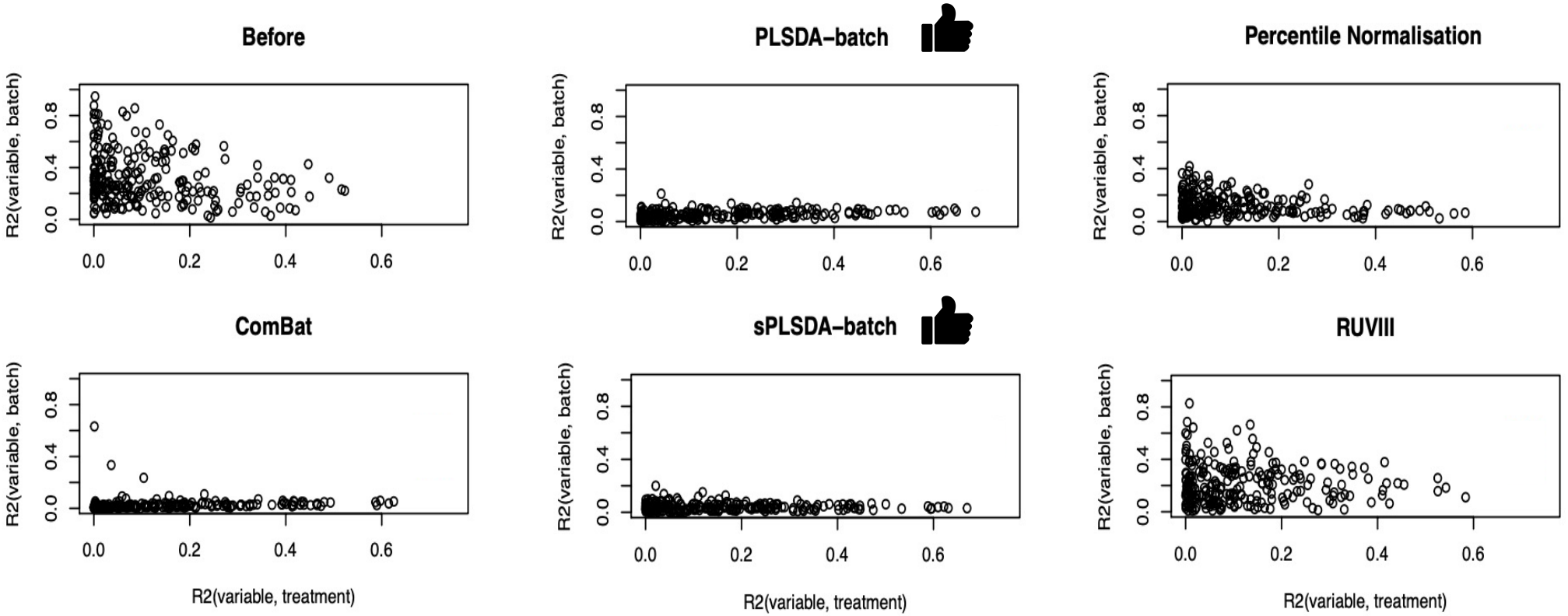


- PLSDA-batch & sPLSDA-batch: 优越于其它方法
- RUVIII: 重复样本 (sample replicates) 起到关键性的作用

Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv*.

AD data 分析结果

每个微生物变量各个来源的差异



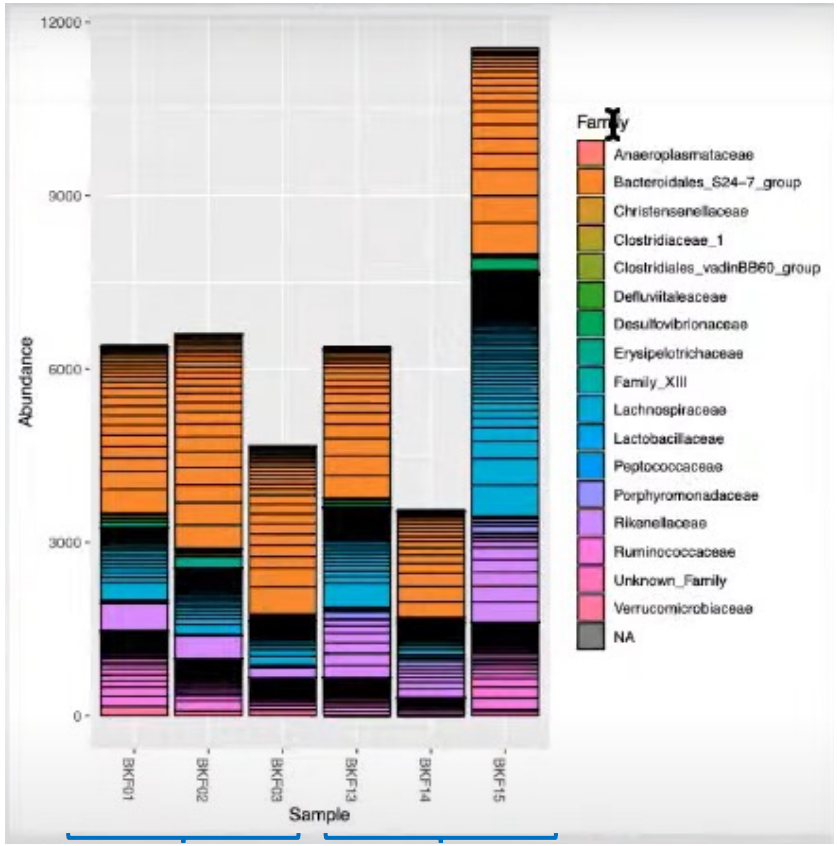
用ComBat, percentile normalisation and RUVIII 校正过的数据仍然存在大量含有批次效应的变量

Wang & Lê Cao (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv*.

后续分析？

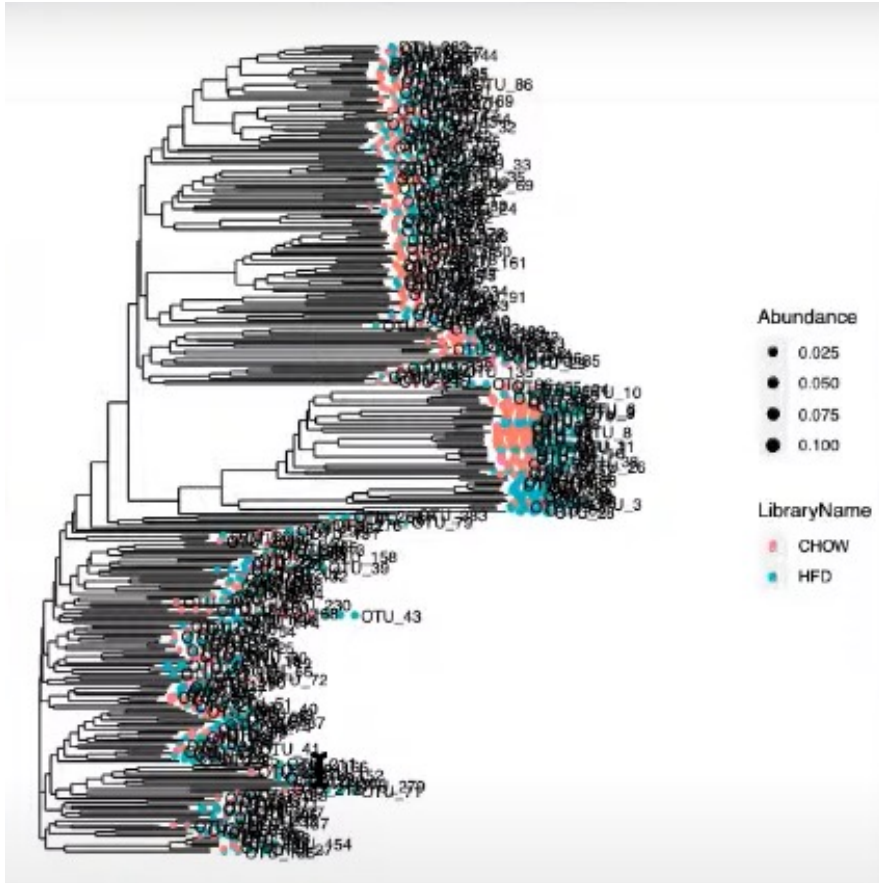
Phyloseq R package

微生物科分布



Regular diet High fat diet

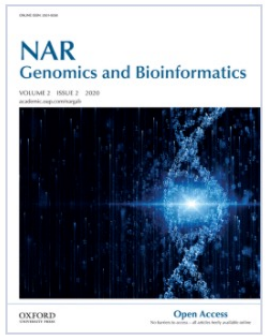
微生物遗传图谱



变量选择

NAR Genomics and Bioinformatics

Issues | More Content ▾ | Submit ▾ | Alerts | About ▾ | All NAR Genomics a



Volume 2, Issue 2
June 2020

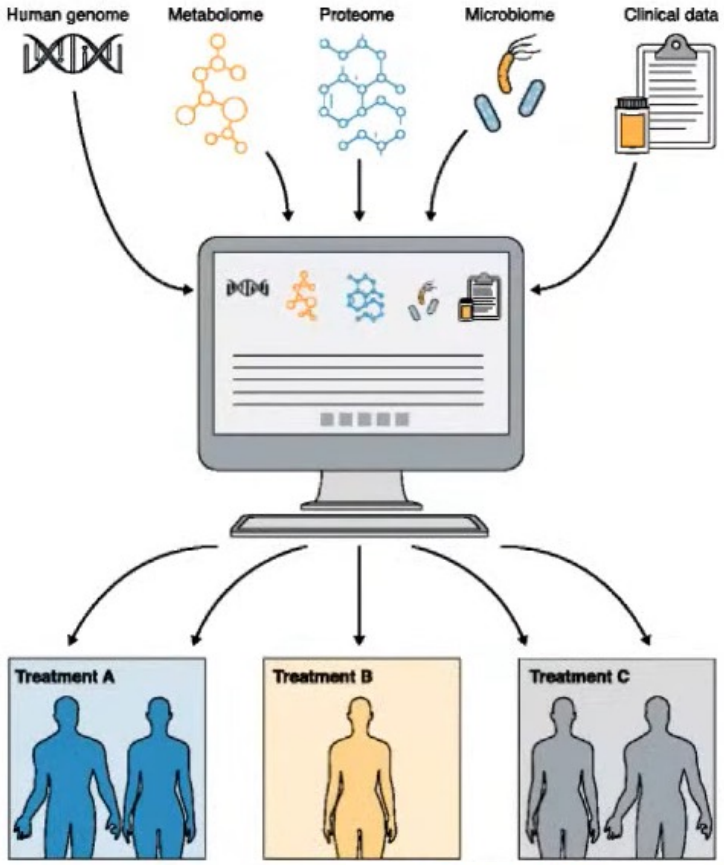
Variable selection in microbiome compositional data analysis

Antoni Susin, Yiwen Wang, Kim-Anh Lê Cao, M Luz Calle

NAR Genomics and Bioinformatics, Volume 2, Issue 2, June 2020, lqaa029,
https://doi.org/10.1093/nargab/lqaa029
Published: 13 May 2020 **Article history** ▾

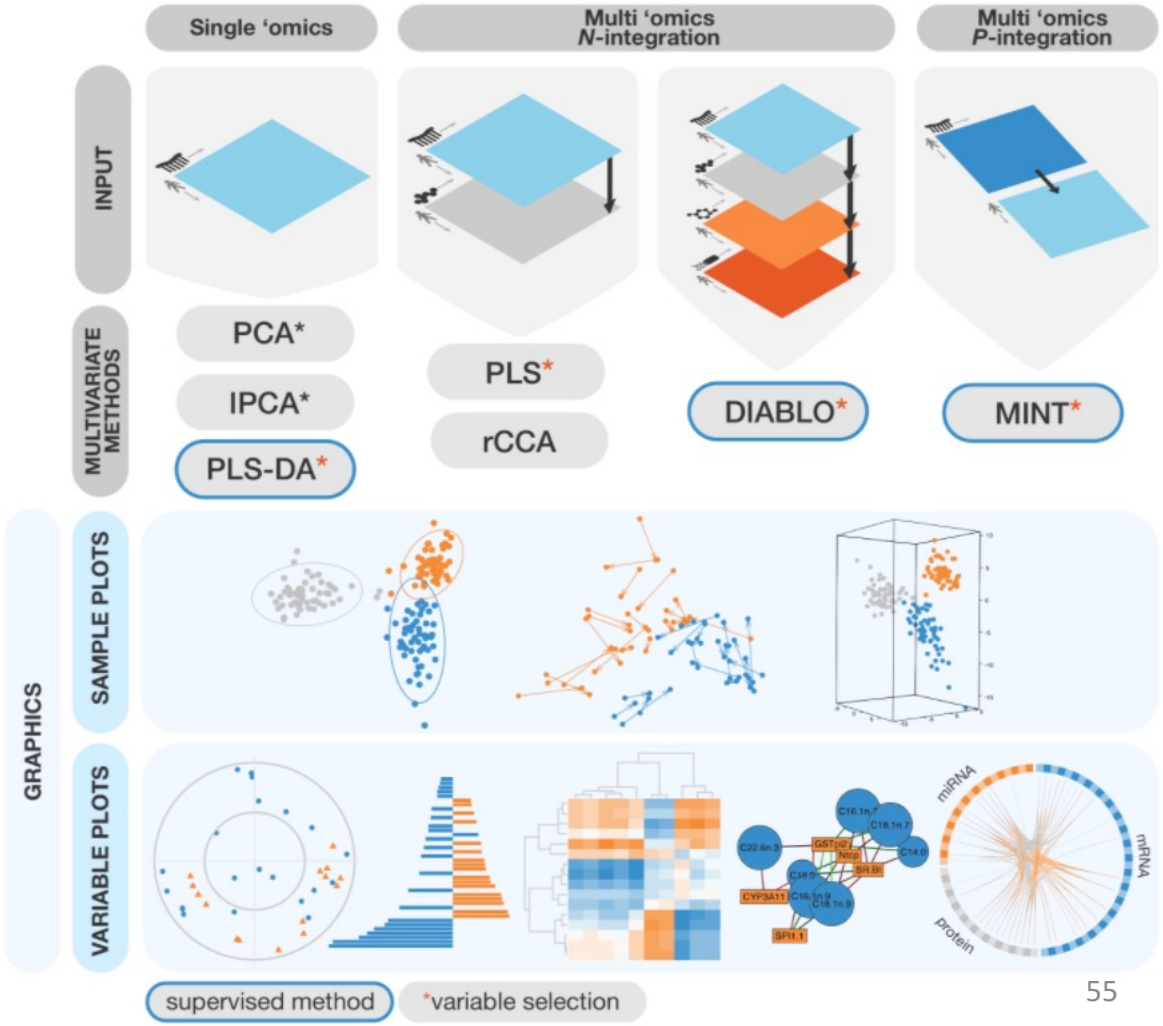
PDF | Split View | Cite | Permissions | Share ▾

多组学数据整合



mixOmics

http://mixomics.org/



总结

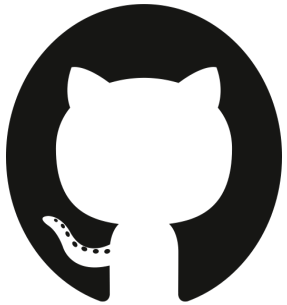
- 微生物宏基因组数据的获取解决了绝大多数微生物难以被分离培养和研究的问题
- 微生物分类学信息可匹配的数据库：SILVA, RDP, Greengenes, NCBI, 注意一起分析的数据选择同一数据库
- 微生物宏基因组数据具有自己独有的特点
- 微生物组数据容易受批次效应的影响
- 处理微生物数据中的批次效应应考虑其内在的数据特征，批次效应的来源，batch x treatment designs 以及对不同微生物变量的不同影响程度

总结

- 与在统计模型中考虑批次效应的方法相比，校正批次效应的方法更灵活
- 检测批次效应和评估批次效应的去除效果是处理批次效应问题时必须要考虑的
- PLSDA-batch是一种多元和非参数的方法
  - 考虑到微生物组数据的特征
  - sparse PLSDA-batch: 避免过度拟合
  - weighted PLSDA-batch : 应对不平衡的 batch x treatment designs
- 校正后的微生物组数据能应用于多种后续分析，比如可视化，变量选择和多组学数据整合等分析

致谢

- A/Prof. Kim-Anh Lê Cao
- Lê Cao 课题组的所有成员
- Melbourne Integrative Genomics 研究所的所有成员
- 周永锋 研究员
- 周永锋课题组的所有成员



github.com/EvaYiwenWang