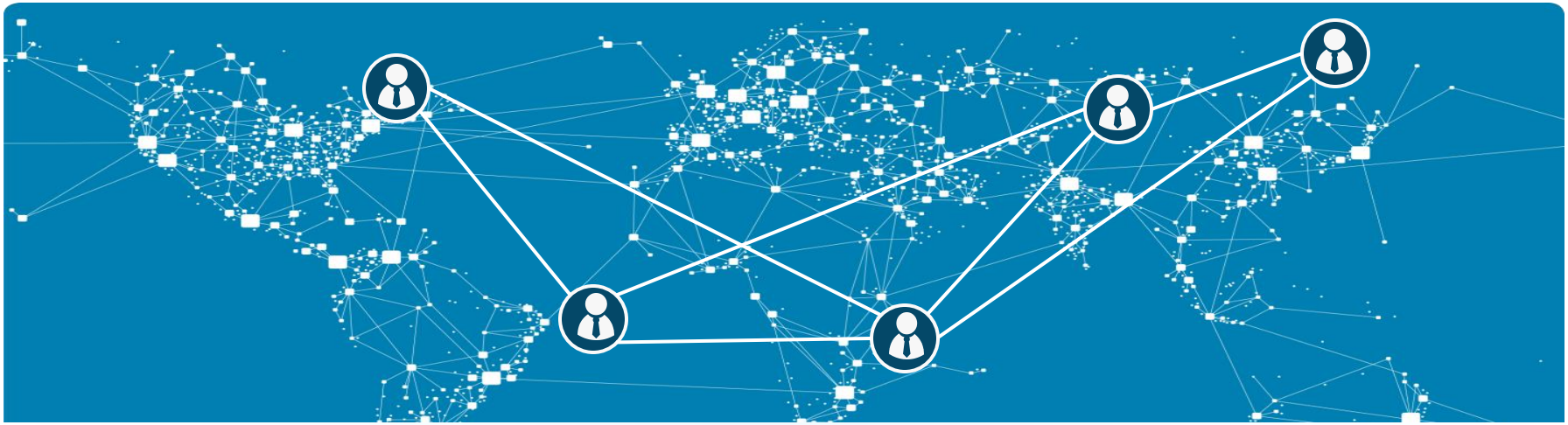


生物大数据整合分析中的批次效应处理

什么是批次效应?

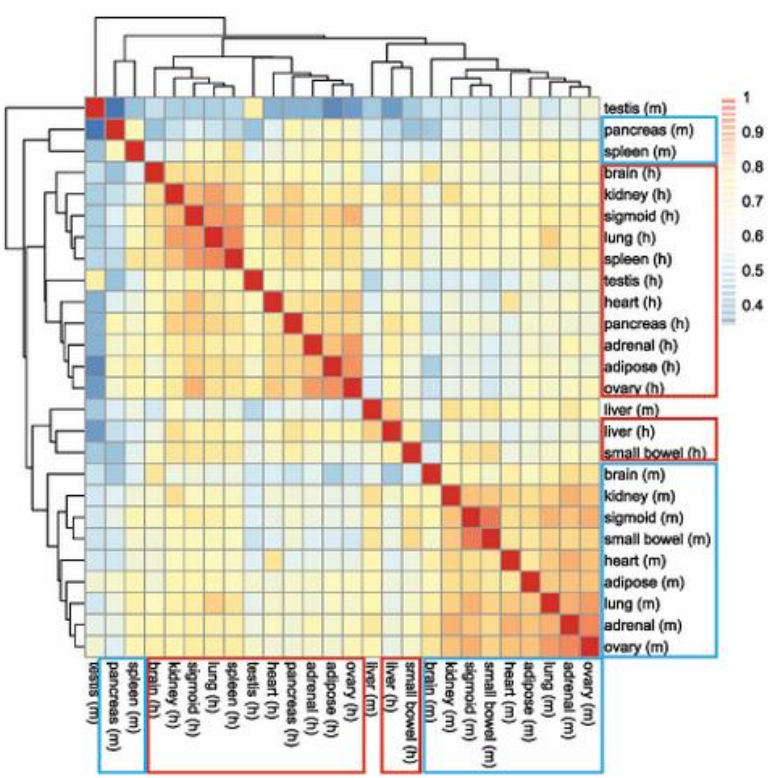
- 批次效应表示样品在不同的批次处理和测量时引入的与**生物状态**不相关的**系统性**的技术偏差。
- 很多因素都可能导致批次效应的产生，如不同实验条件、不同操作者、不同公司的试剂、不同批的试剂、实验开展的时间、检测设备、不同的测序批次、昼夜节律、细胞周期等。

评估21种不同DNA提取方法对检测到菌群构成的影响

小鼠的脑与小鼠的肾脏的相似性大于小鼠的脑与人的脑的相似性？

2014年生信领域的大牛Michael P Snyder在PNAS发文比较了人和小鼠不同组织和器官中表达谱的异同。

研究发现不同物种之间组织特异表达的基因是一致的，但很多基因在同一物种不同组织的表达相似度大于它们在不同物种同一组织的表达相似度。



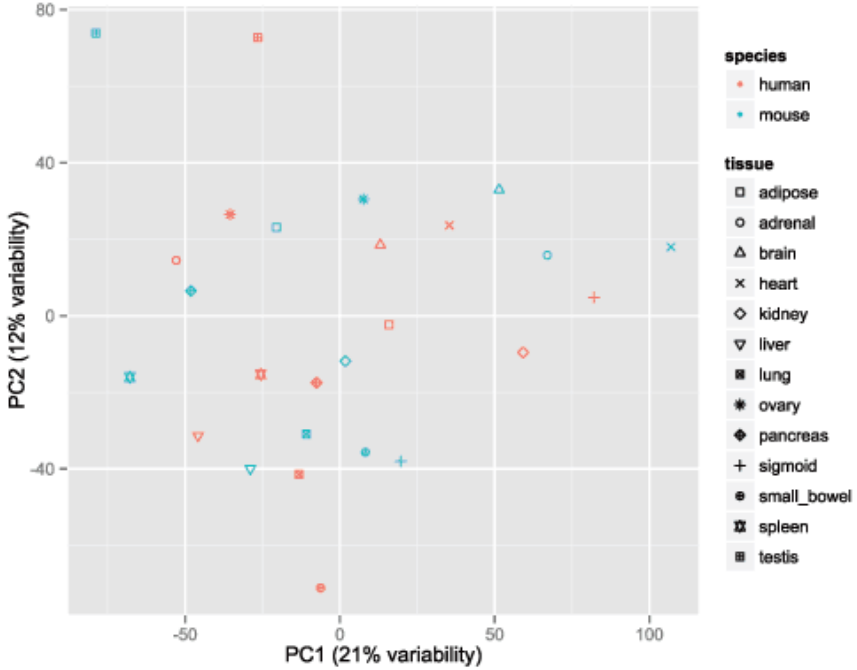
Comparison of the transcriptional landscapes between human and mouse tissues PBAS 2014

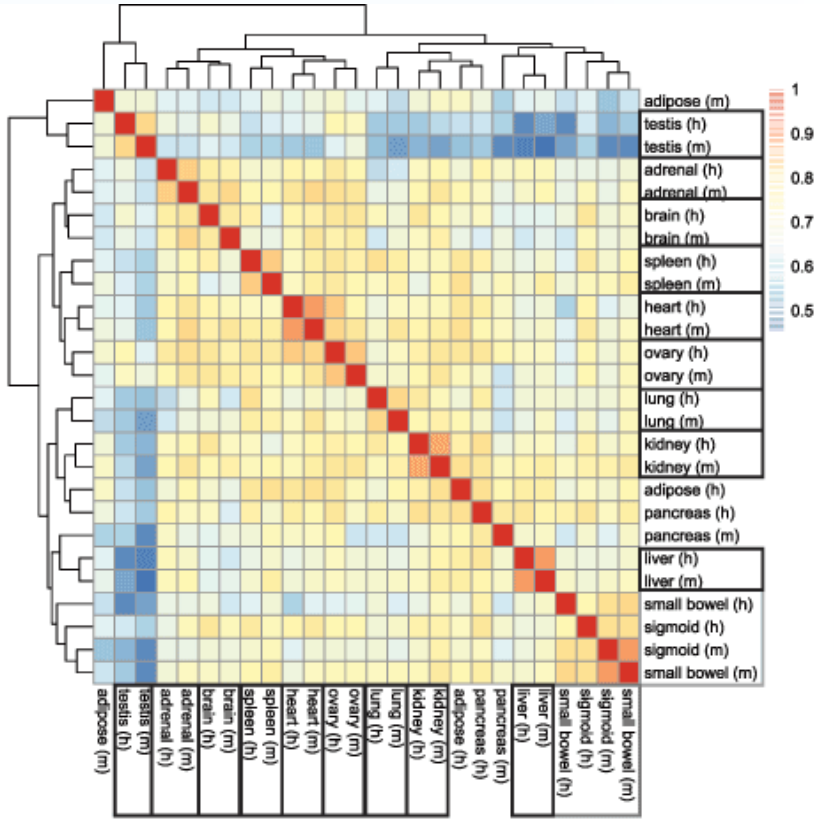
这样的实验设计很无奈

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

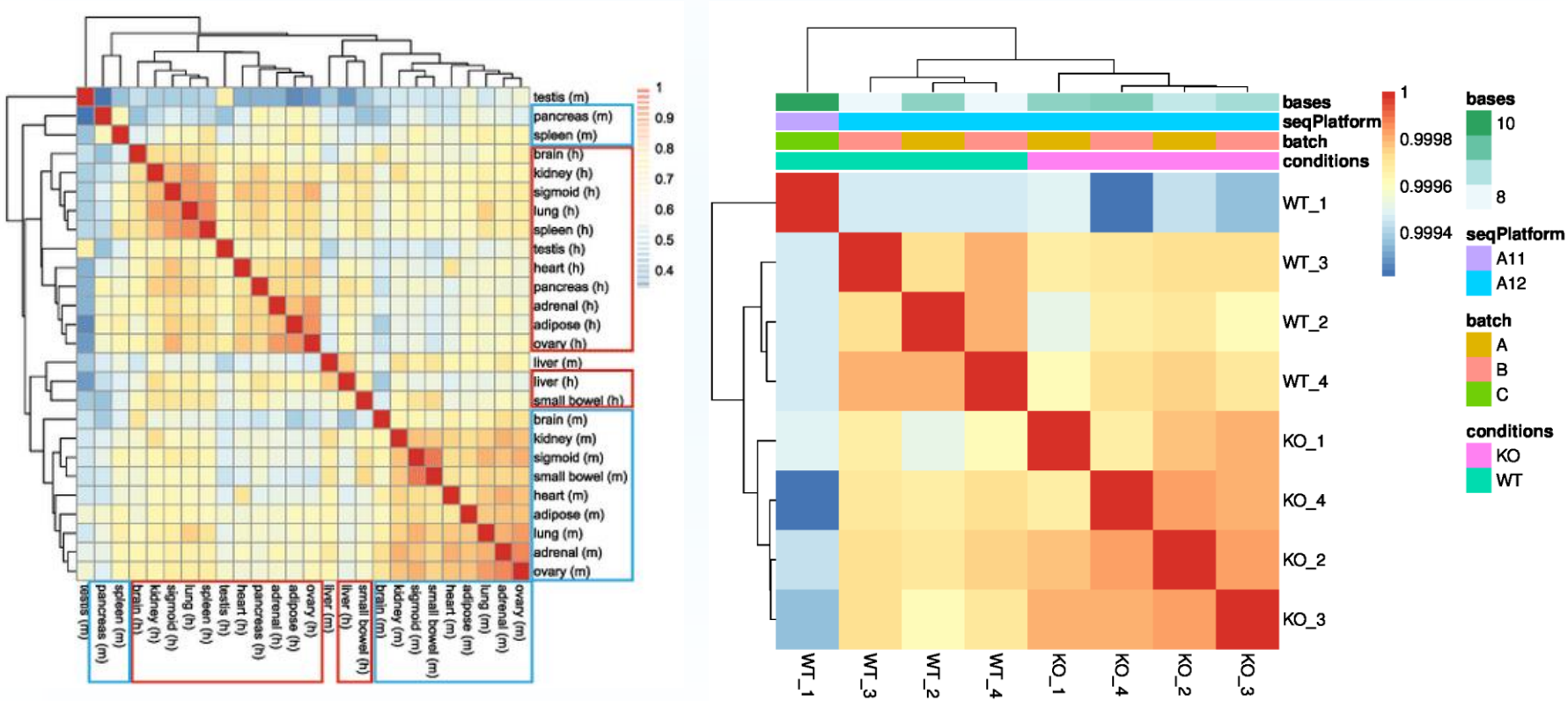
分析批次影响

批次校正后，表达谱按组织类型而非物种聚在一起





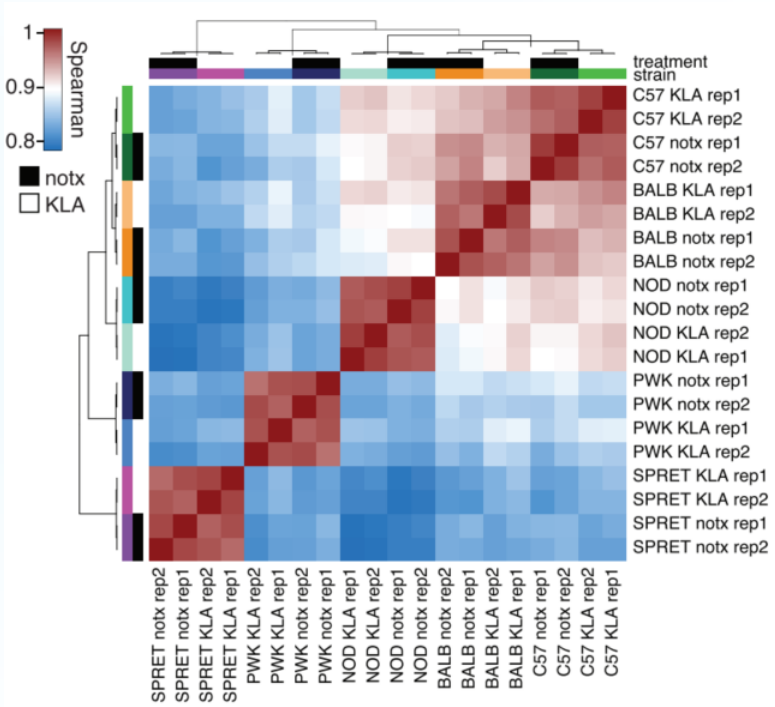
从聚类结果查看批次效应（不同聚类算法可能会改变聚类结构）



Comparison of the transcriptional landscapes between human and mouse tissues

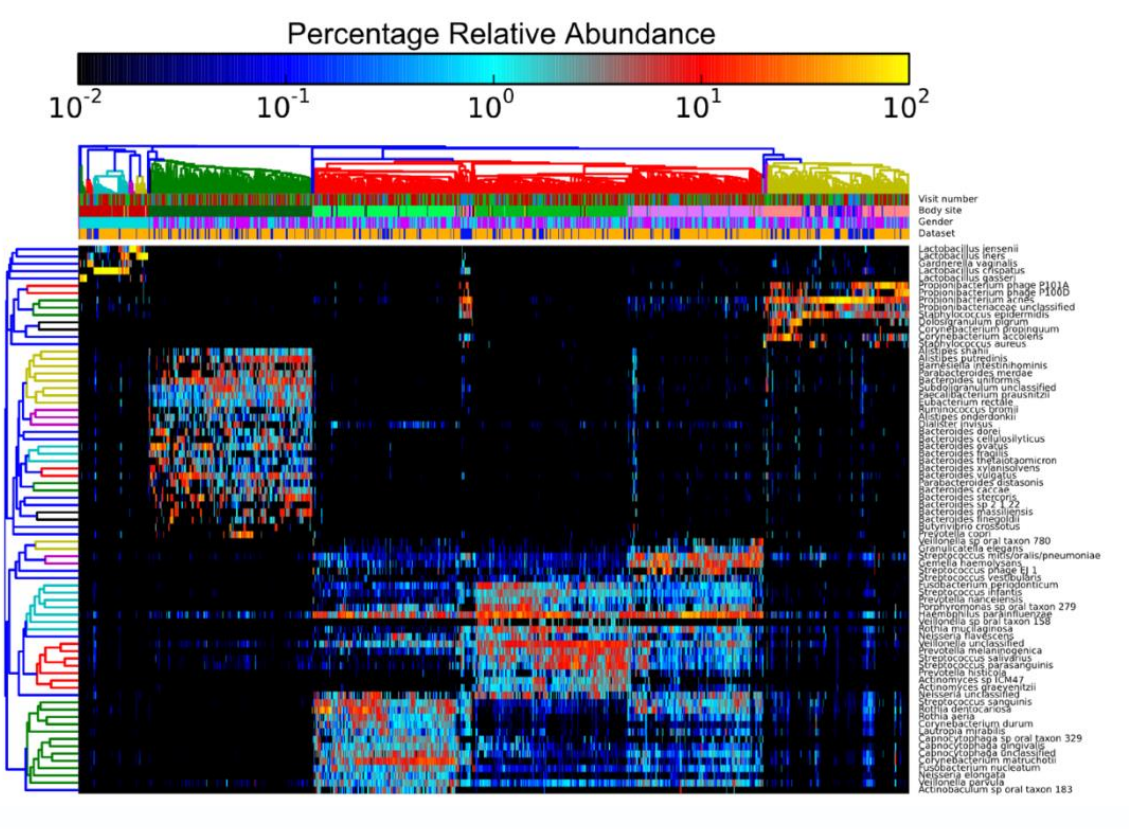
相关性热图加样本属性标记

- 与PCA聚类相似的是相关系数层级聚类，也可以展示样品之间的相关性。
- 同时可以增加样品的属性信息借以查看不同的属性信息与分组的关系，从而确定有没有哪些意料外的信息影响到了样品分组。
- 小鼠不同株野生型和KLA处理组基因表达谱相似性热图。
- Figure legend: Clustered Spearman correlation matrix for different RNA-seq replicates for no treatment and KLA 1h.



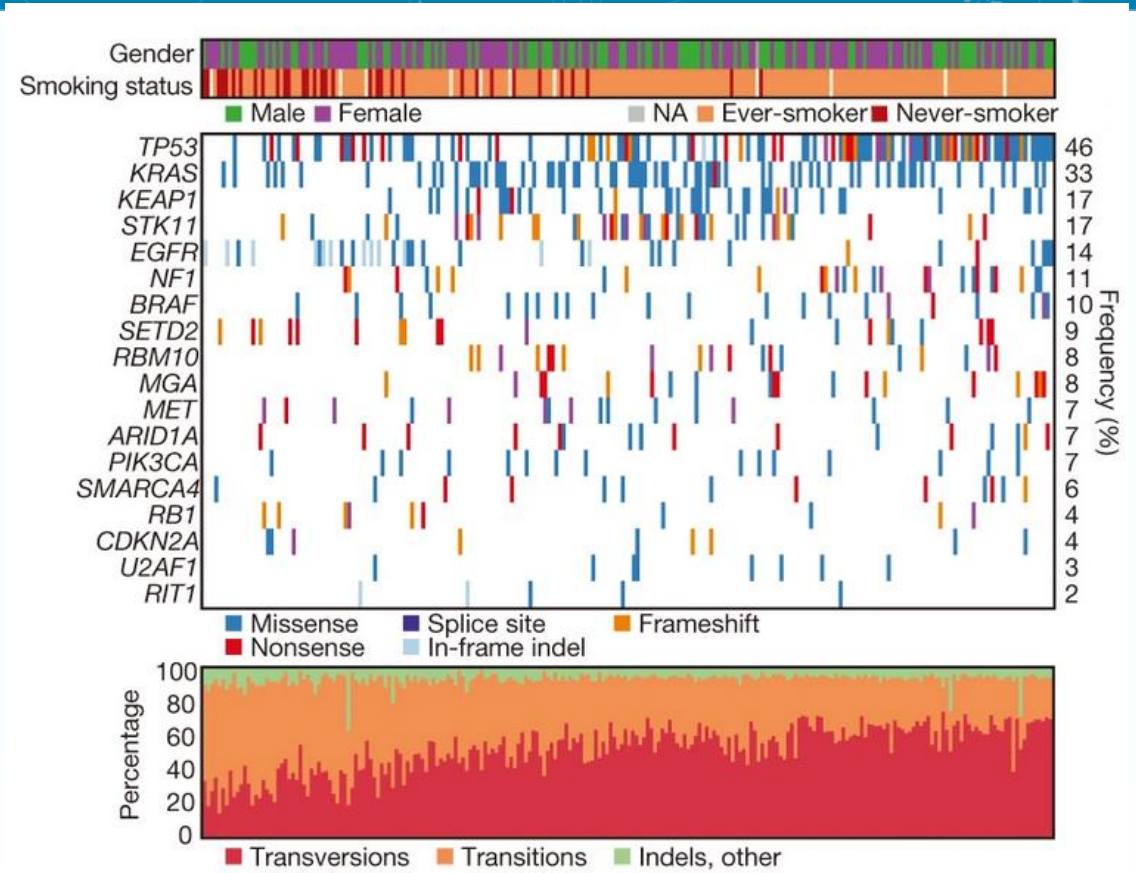
Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide mechanisms that Control Transcription Factor Binding and Function Cell (Figure S1 C)

展示结果时标记样品的属性和数据来源



MetaPhlAn2 for enhanced metagenomic taxonomic profiling

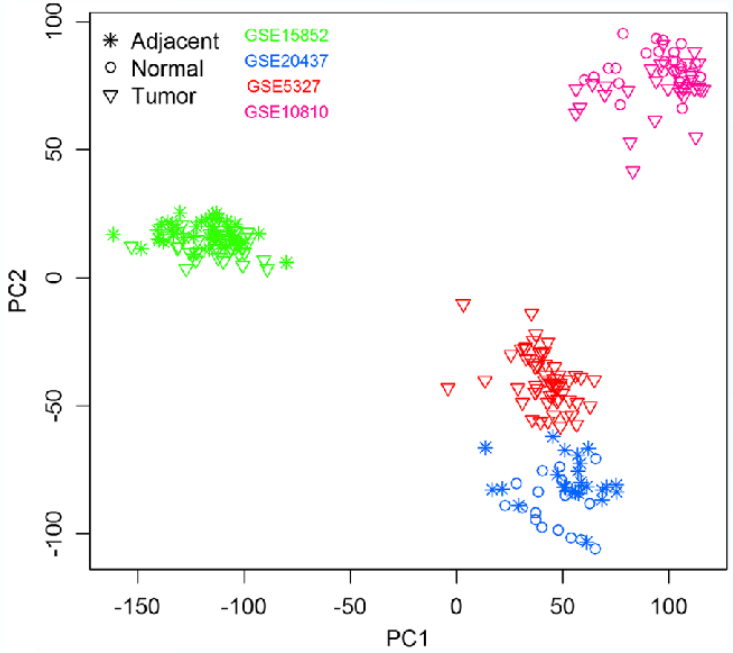
展示结果时标记样本属性



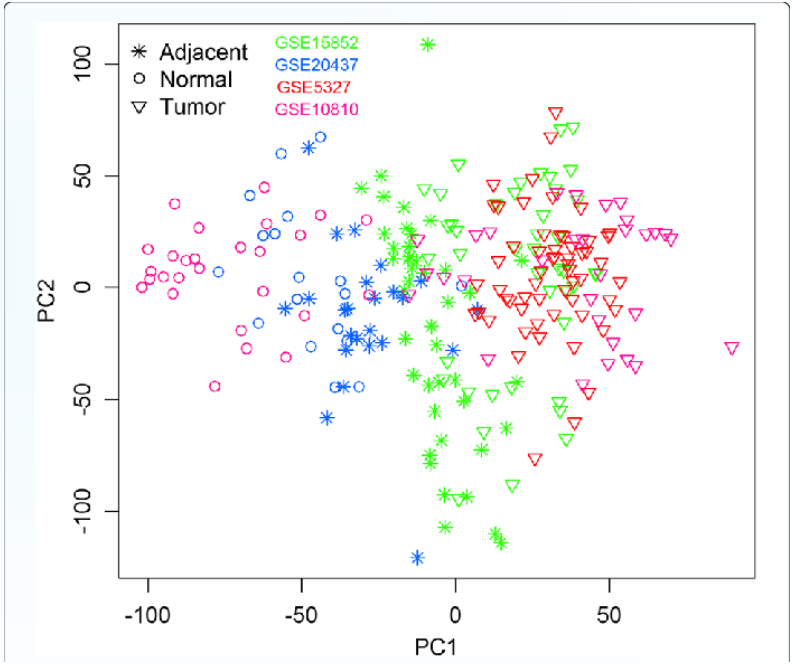
Nature

从PCA结果查看批次效应

批次校正前



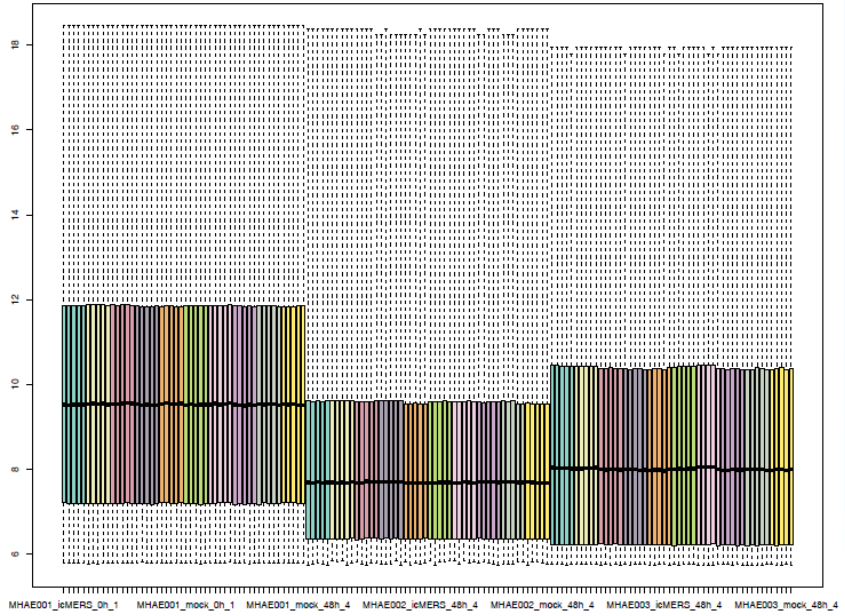
批次校正后

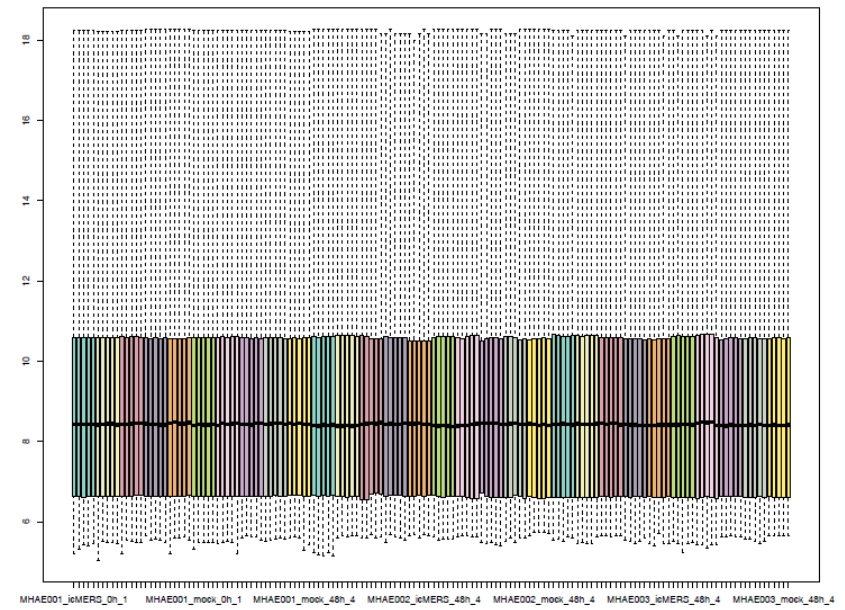


PCA plot for combat batch

从表达分布查看批次效应

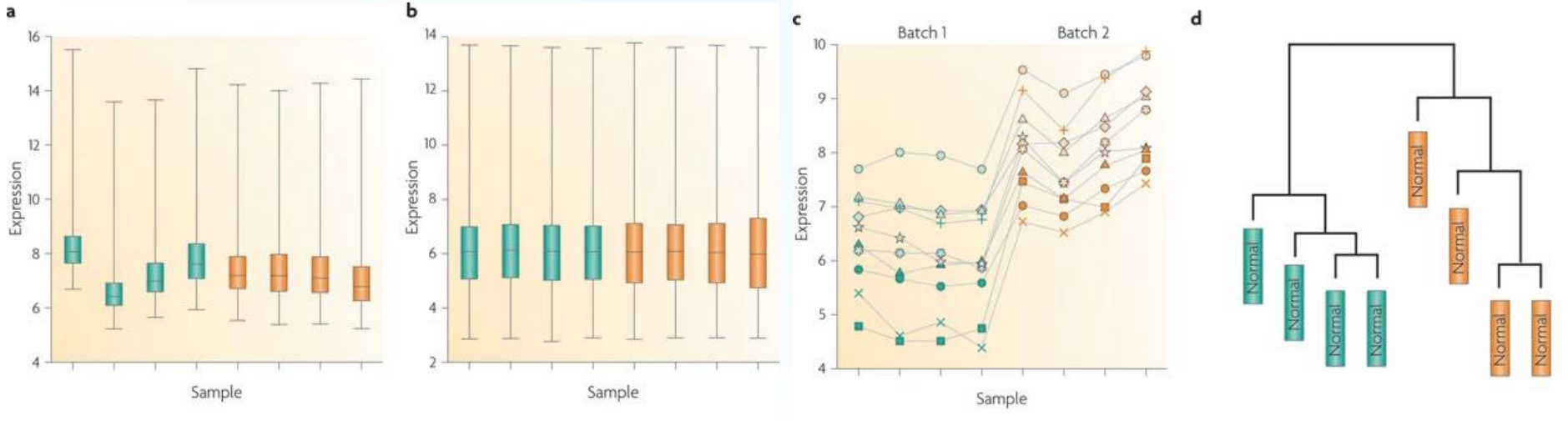
- 不同来源的样本表达整体分布不同（左图为批次校正前，右图为批次校正后）





从单基因表达水平查看批次效应

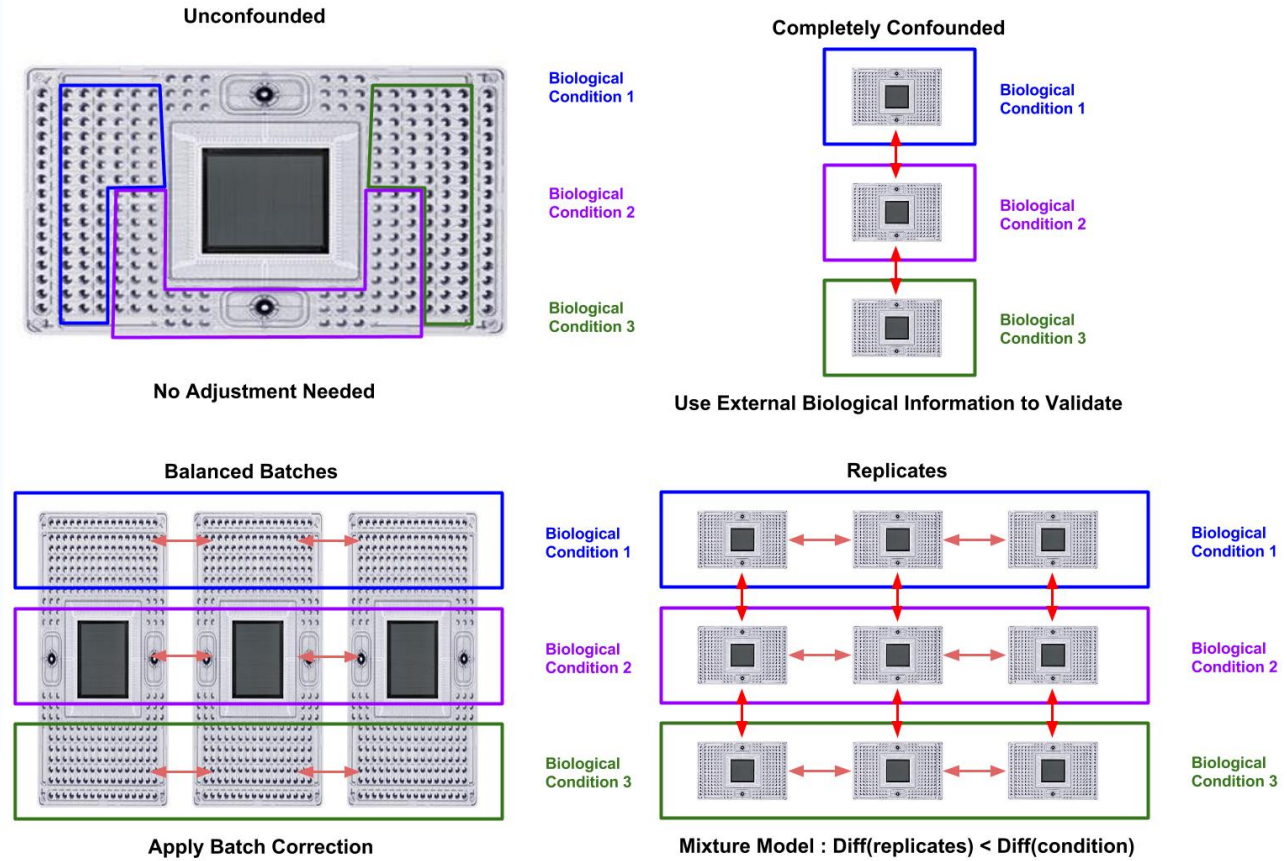
- 蓝色和黄色为不同时间点处理的正常样品，一起标准化之后，整体表达分布看不出差异，但有成百上千的基因表达明显受到批次影响，且从聚类可看出两组正常样品构成2大分支。



Nature Review Genetics Batch

合理的实验设计 – 尽量避免批次效应影响

试验设计指南



合理的实验设计 – 尽量避免批次效应影响

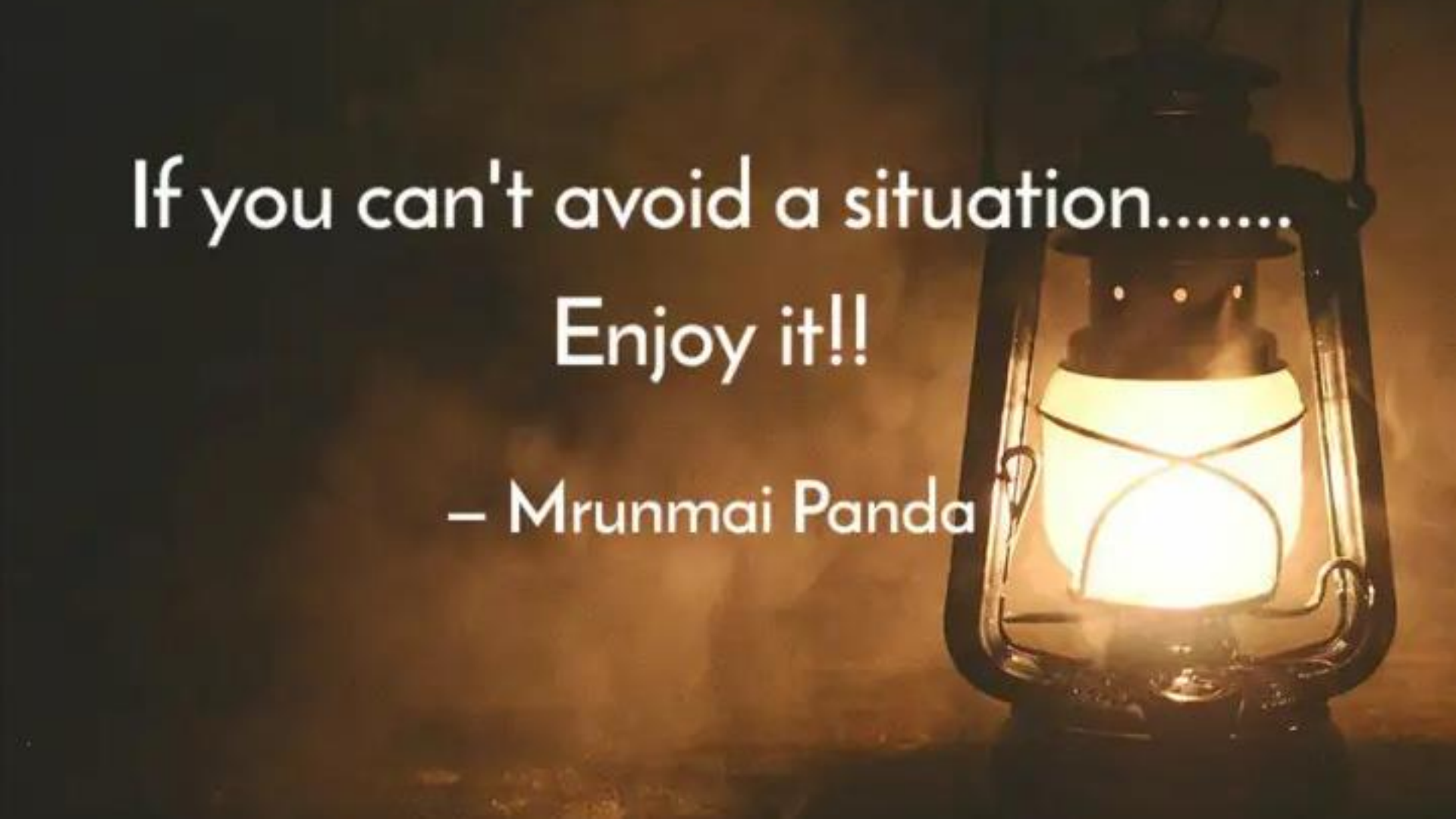
易拉罐莉莉娅 7

我感觉comBat不能那样用，comBat是基于几组样本彼此是平衡设计、理论分布近似，只是实验和仪器分析引入批次效应，才可以使用它来去除这样的效应。人和鼠不同的物种在不同的批次里，生物学差异很可能大于实验批次差异。

> 作者
>
> 作者也是想通过这个方式说明该实验设计的缺陷。如此设计，校正是没有好的方式去校正，只能尽量去从数学角度去抹平那些来源于系统的一致的变化。

> 易拉罐莉莉娅
>
> 抱歉我没有仔细读文章，不过看起来实验设计确实有缺陷。通常我们实验室会在不同批次中加入一个相同的QC样本，可以用来测试是否有批次效应，校正的时候如果QC重合了校正就是有效的

> 作者 3
>
> 正解，一般加上2-3个QC，作为校正依据，赞严谨的设计。


If you can't avoid a situation.....
Enjoy it!!
– Mrunmai Panda

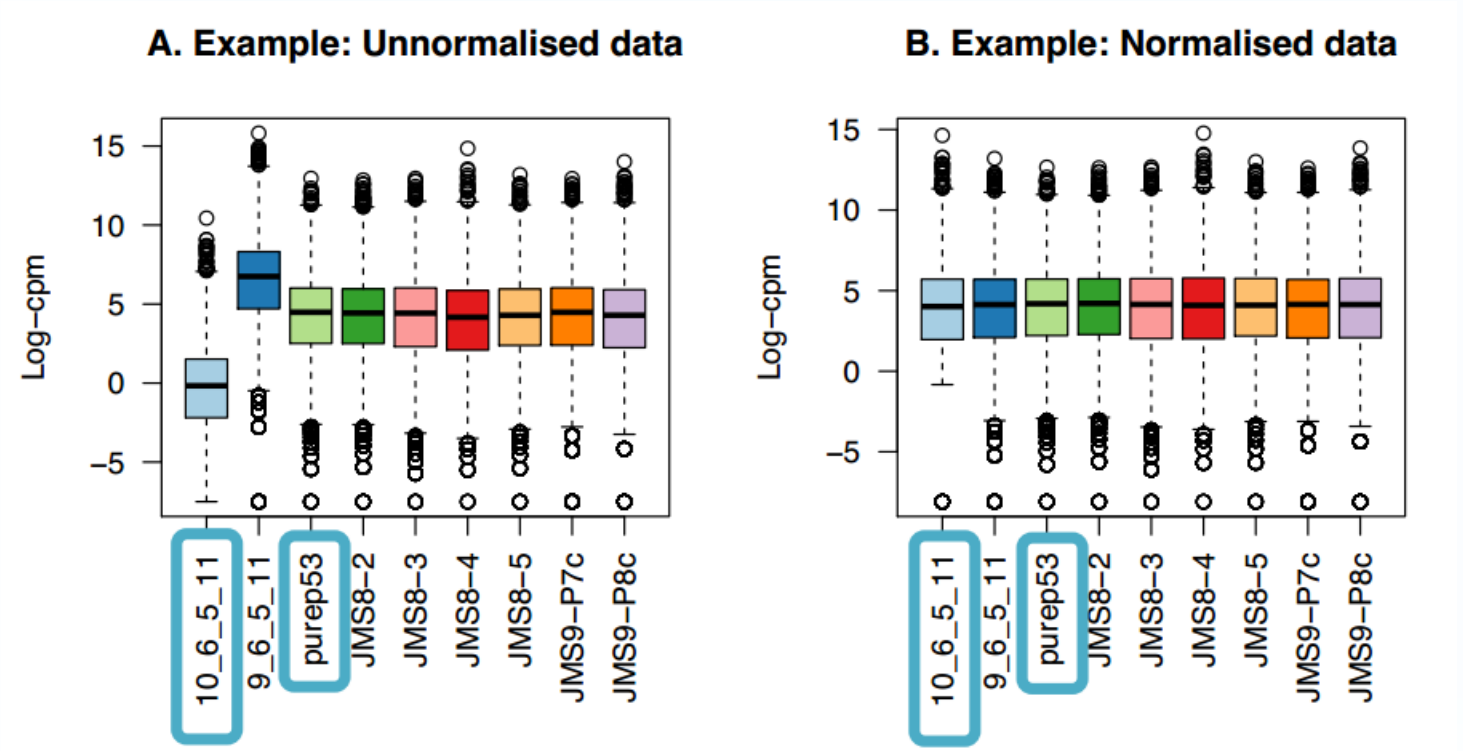
药物处理实验设计： 4个对照组，4个处理组

Samp	conditions	individual
untrt_N61311	untrt	N61311
untrt_N052611	untrt	N052611
untrt_N080611	untrt	N080611
untrt_N061011	untrt	N061011
trt_N61311	trt	N61311
trt_N052611	trt	N052611
trt_N080611	trt	N080611
trt_N061011	trt	N061011

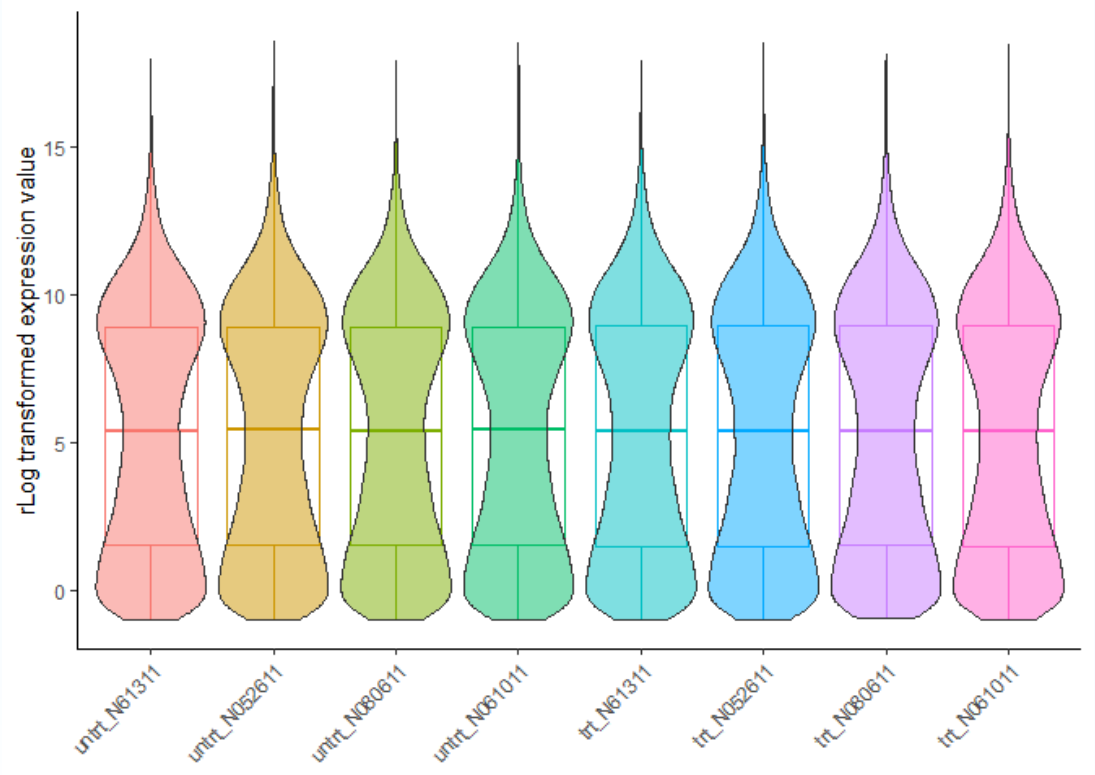
数据读入和标准化 (DESeq2)

- [1] “Read in 32799 genes”
- [2] “23936 genes remained after filtering of genes with all counts less than 4 in all samples.”
- [3] “Perform DESeq on given datasets.”
  - estimating size factors
  - estimating dispersions
  - gene-wise dispersion estimates
  - mean-dispersion relationship
  - final dispersion estimates
  - fitting model and testing
- [4] “Output normalized counts”
- [5] “Output rlog transformed normalized counts”

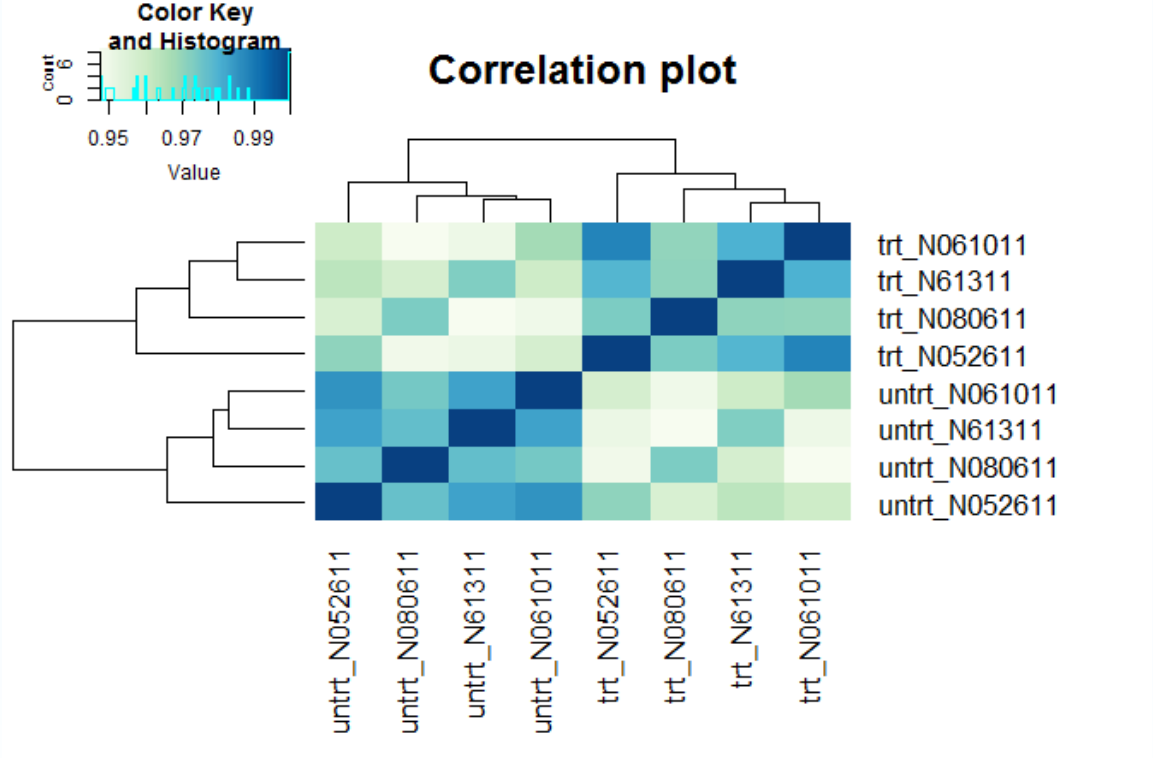
标准化前后数据分布比较：好的标准化分布一致



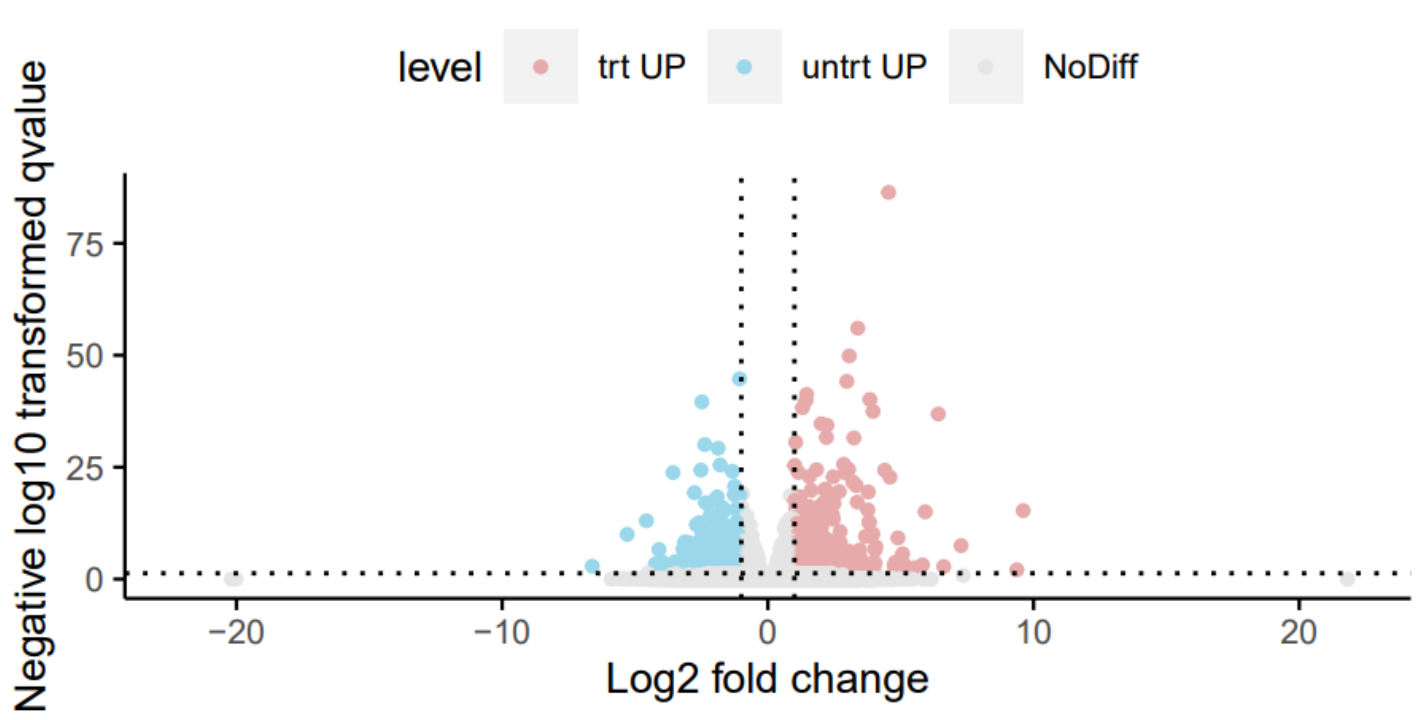
案例数据标准化后整体表达分布一致



聚类热图可把处理组和对照组分成2个大的分支



有差异基因万事大吉?

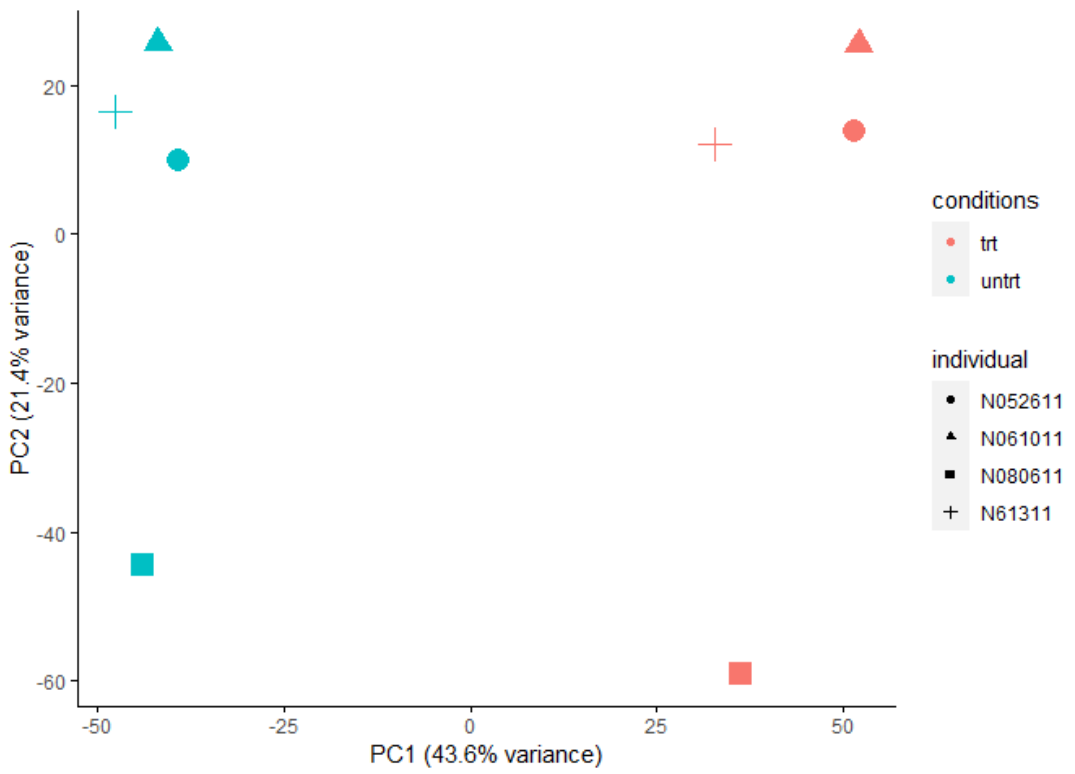

level
trt UP
untrt UP
NoDiff
Negative log10 transformed qvalue
75
50
25
0
−20
−10
0
10
20
Log2 fold change

PCA图显示来源个体信息可能是一个批次因素

PC1轴上，样品按处理条件分开；

PC2轴上，样品按来源个体分开；

不同的个体是影响样品基因表达差异的一个重要因素。



如何在差异基因鉴定过程中移除批次效应?

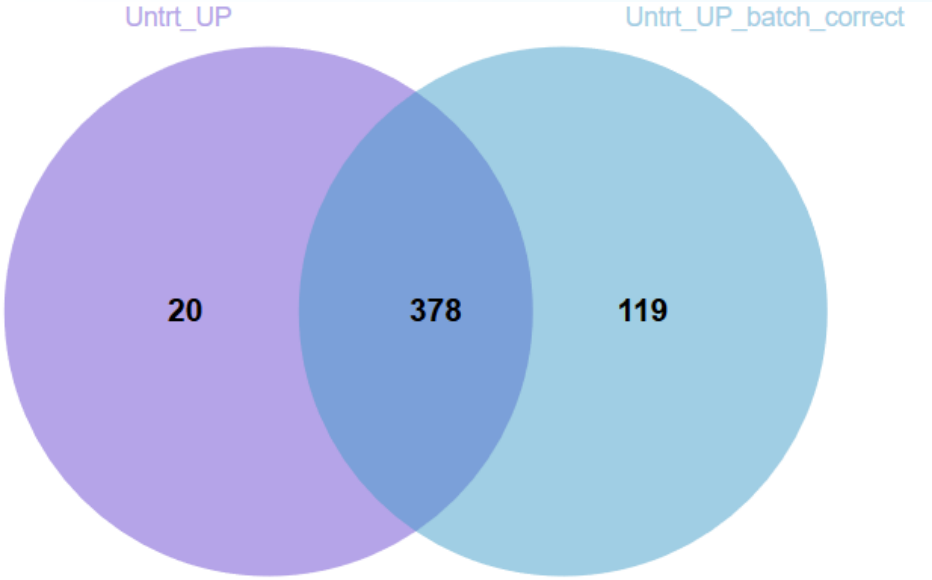
- DESeq2分析时设定批次变量如下面代码中的batch即可，需要注意的是目标分组信息建议放在最后一位，方便后期处理。

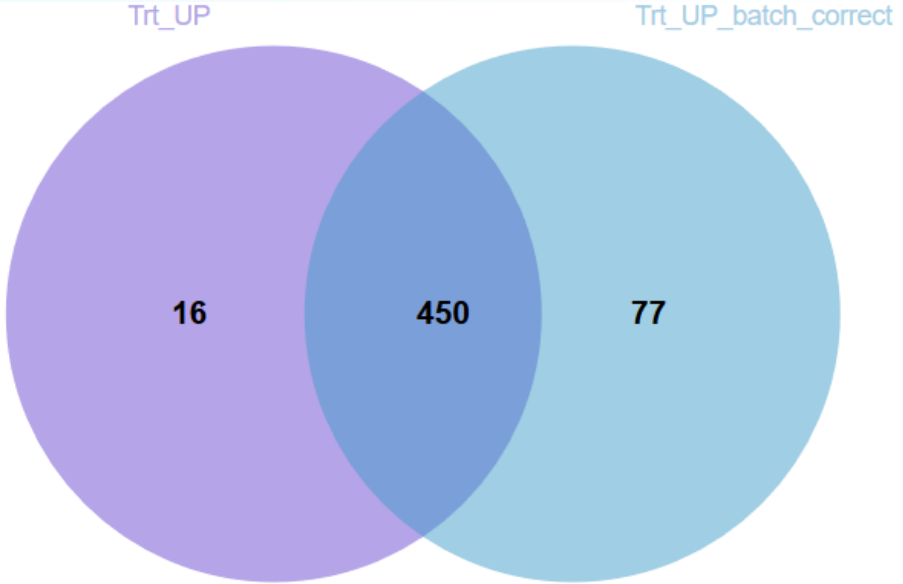
```
ddsFullCountTable <- DESeqDataSetFromMatrix(countData = data,
        colData = sample,  design= ~ batch + conditions)


dds <- DESeq(ddsFullCountTable)
```

DESeq2分析指南

批次校正后获得的差异基因数目变多了

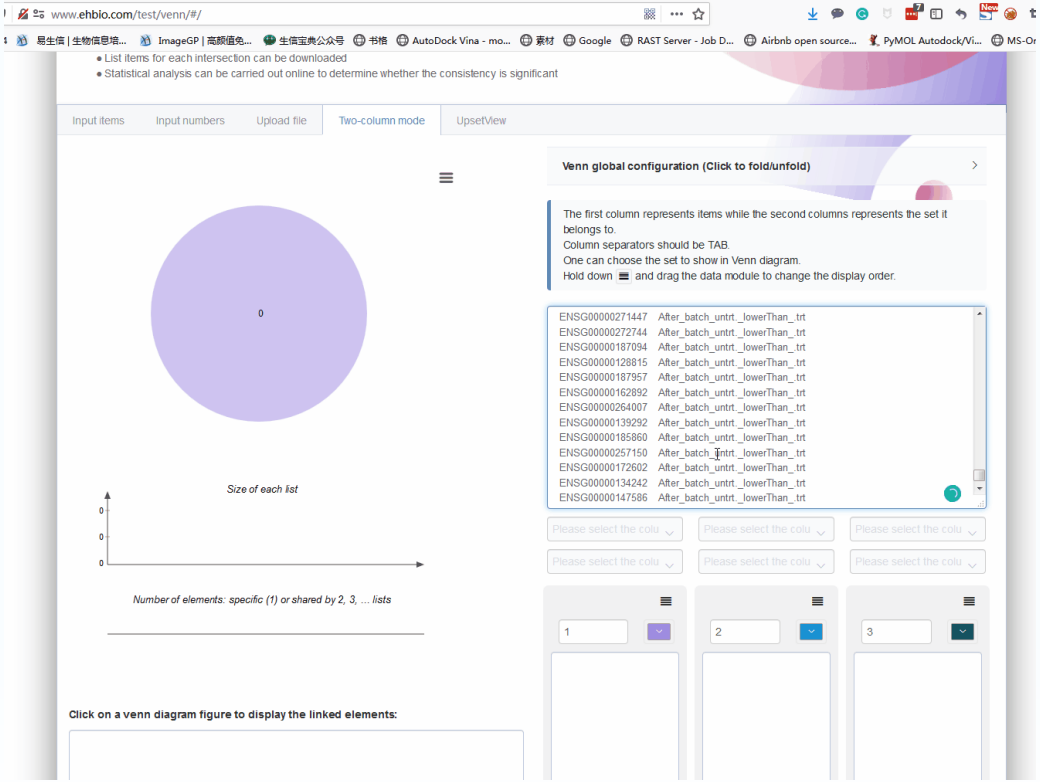
- 紫色是未考虑批次因素时鉴定出的上下调基因
- 绿色是　考虑批次因素时鉴定出的上下调基因。





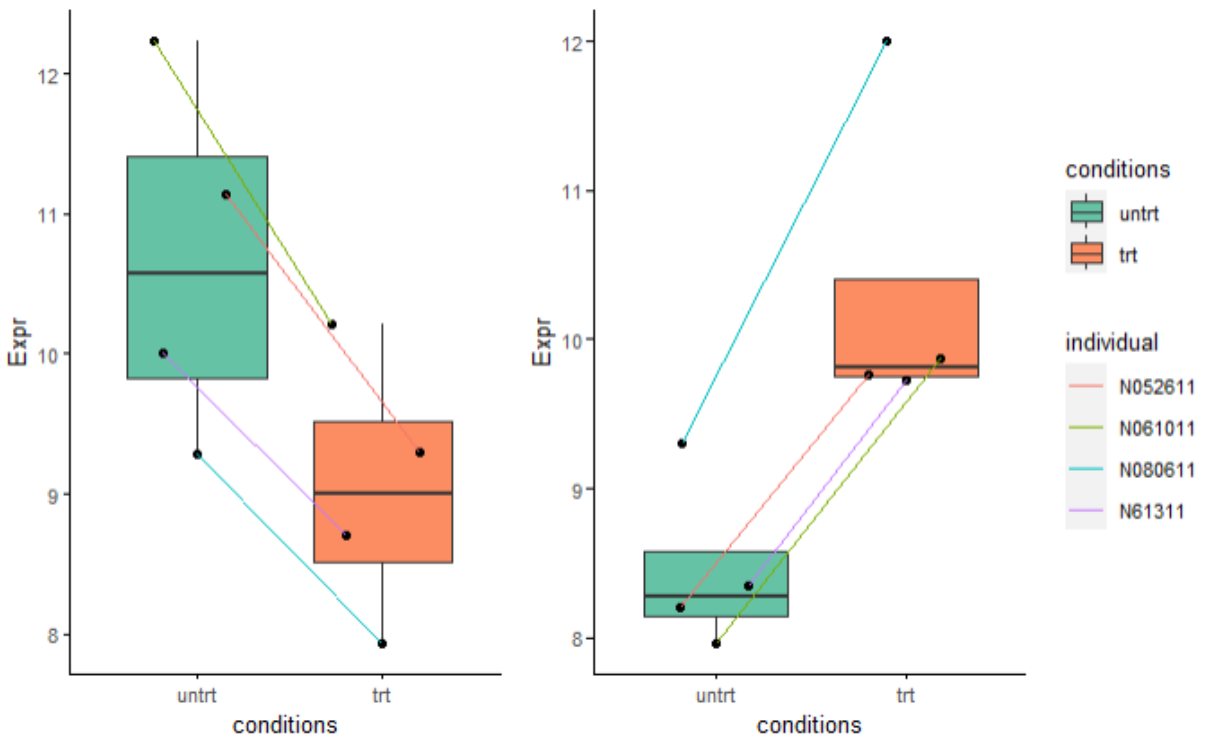
在线网站轻松绘制5种交互式Venn图并输出各个集合的基因

交互式Venn图动图展示



在线网站轻松绘制5种交互式Venn图并输出各个集合的基因

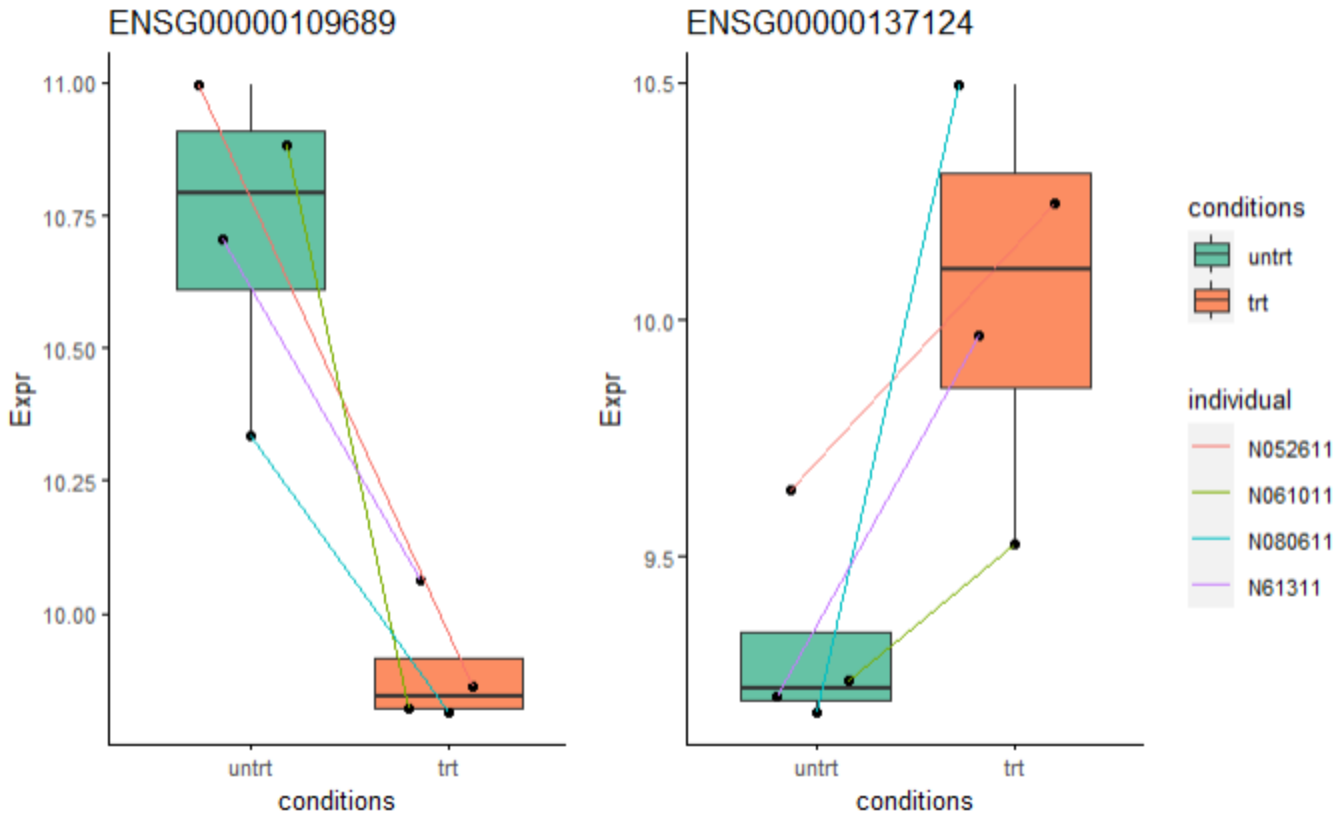
批次校正后鉴定为差异的基因有明显的个体表达偏好性

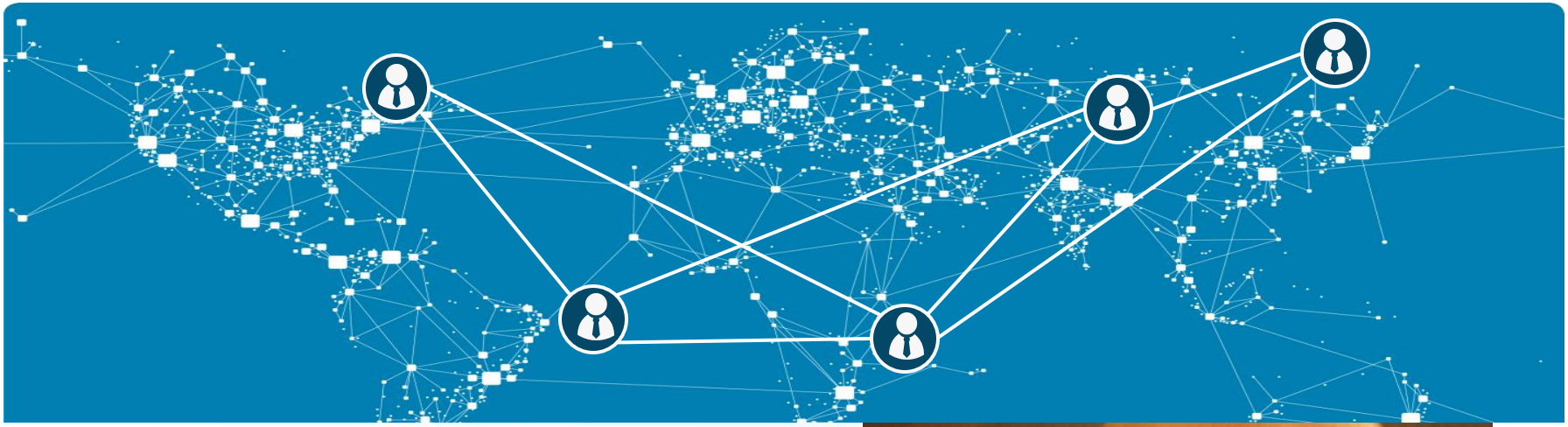


这些基因在同一个体处理前后变化倍数一致。

考虑到每个个体的基准表达水平不同，最终获得的差异倍数会有较高的方差。批次校正后解决了样品个体来源基因本底表达差异的影响，获得的差异基因倍数方差会变小，所以检测出更多差异基因。

批次校正后不再判断为差异的基因表达更聚集





如果批次效应未知呢？

SVA预测可能存在的混杂因素（含批次效应）

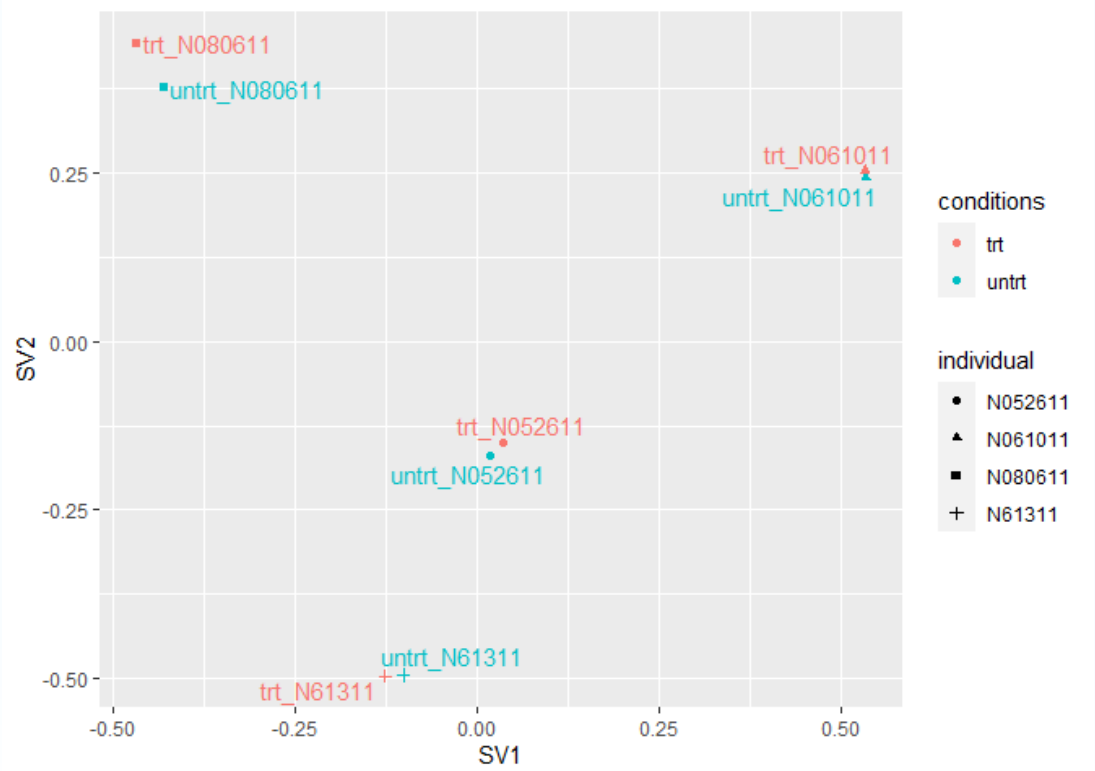
```r
# 获取标准化后的表达矩阵
dat <- normexpr$rlog
# 根据关键生物表型构建设计矩阵
mod  <- model.matrix(as.formula(paste0("~ ", design)), colData(dds))
# 构建对照设计矩阵
mod0 <- model.matrix(~ 1, colData(dds))
# 指定混杂因素的数目为 2，也可以让 sva 自己预测
svseq2 <- sva(dat, mod, mod0)
```

```
## Number of significant surrogate variables is:  3
## Iteration (out of 5 ):1  2  3  4  5
```

```r
svseq2$sv
```

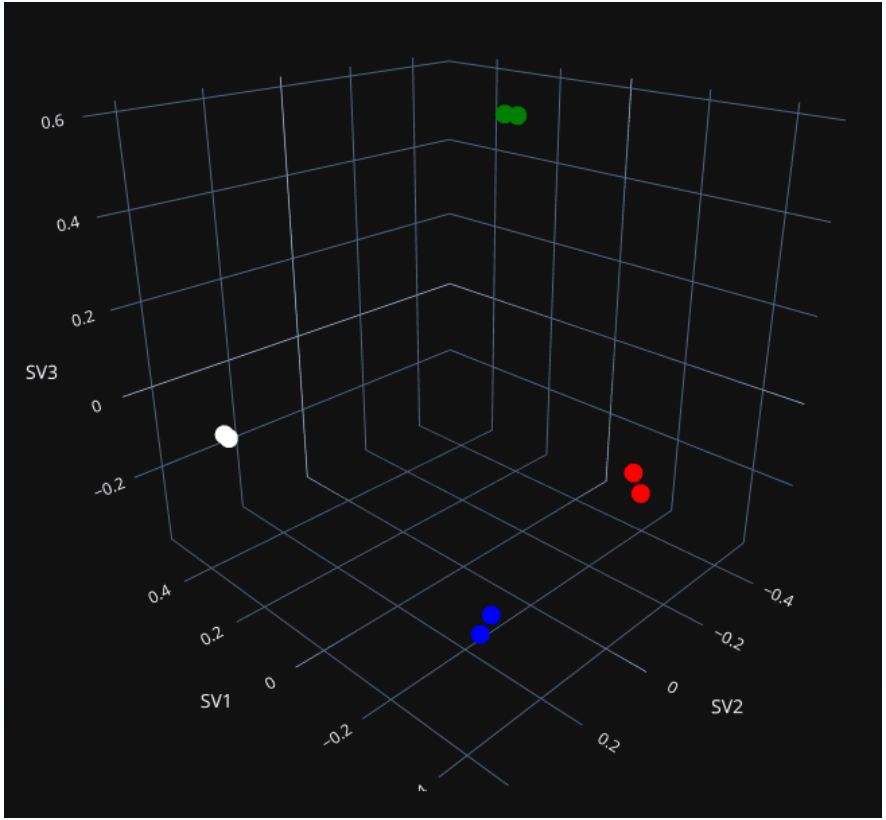
```
##              [,1]       [,2]        [,3]
## [1,] -0.10060276 -0.4943515 -0.31643414
## [2,]  0.01827805 -0.1701072  0.58841449
## [3,] -0.42949246  0.3756333 -0.08929489
## [4,]  0.53452344  0.2413745 -0.17649122
## [5,] -0.12535571 -0.4956603 -0.36550128
## [6,]  0.03588340 -0.1512014  0.59141777
## [7,] -0.46668468  0.4413426 -0.07016838
## [8,]  0.53345071  0.2529700 -0.16194236
```

SVA预测的混杂因素与个体信息比较一致



高颜值免费在线绘图平台

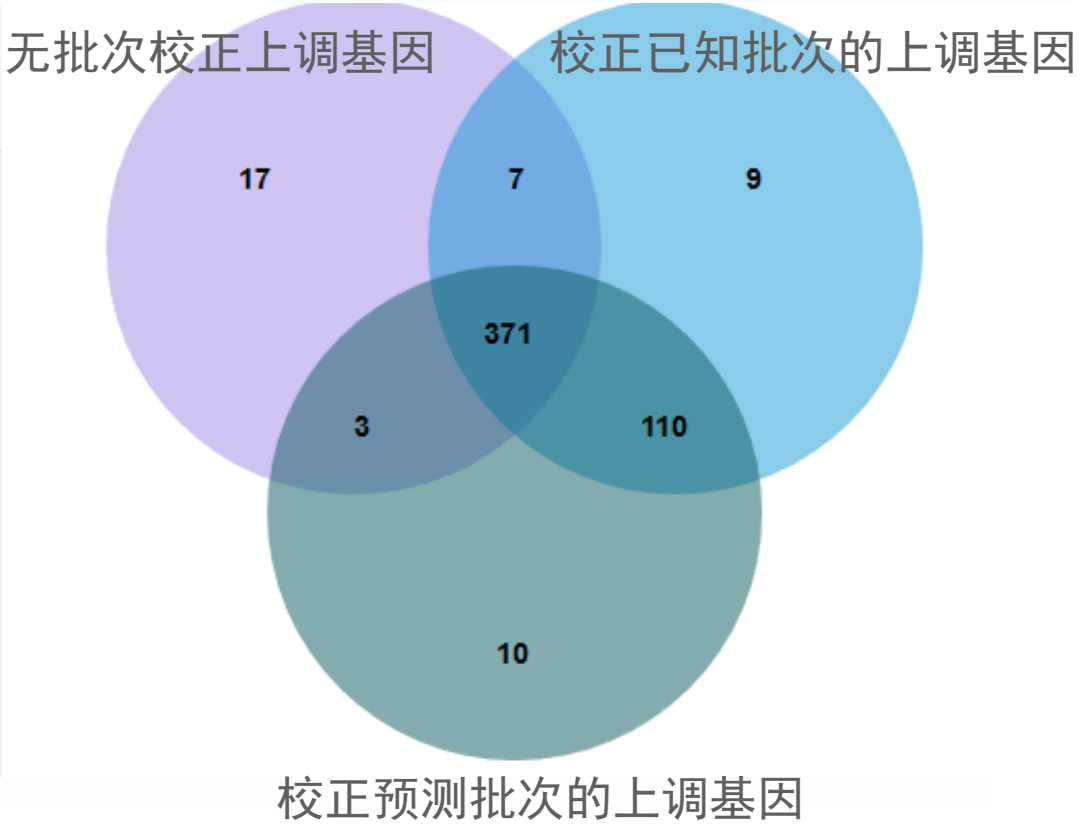
SVA预测的混杂因素与个体信息比较一致



3D 散点图

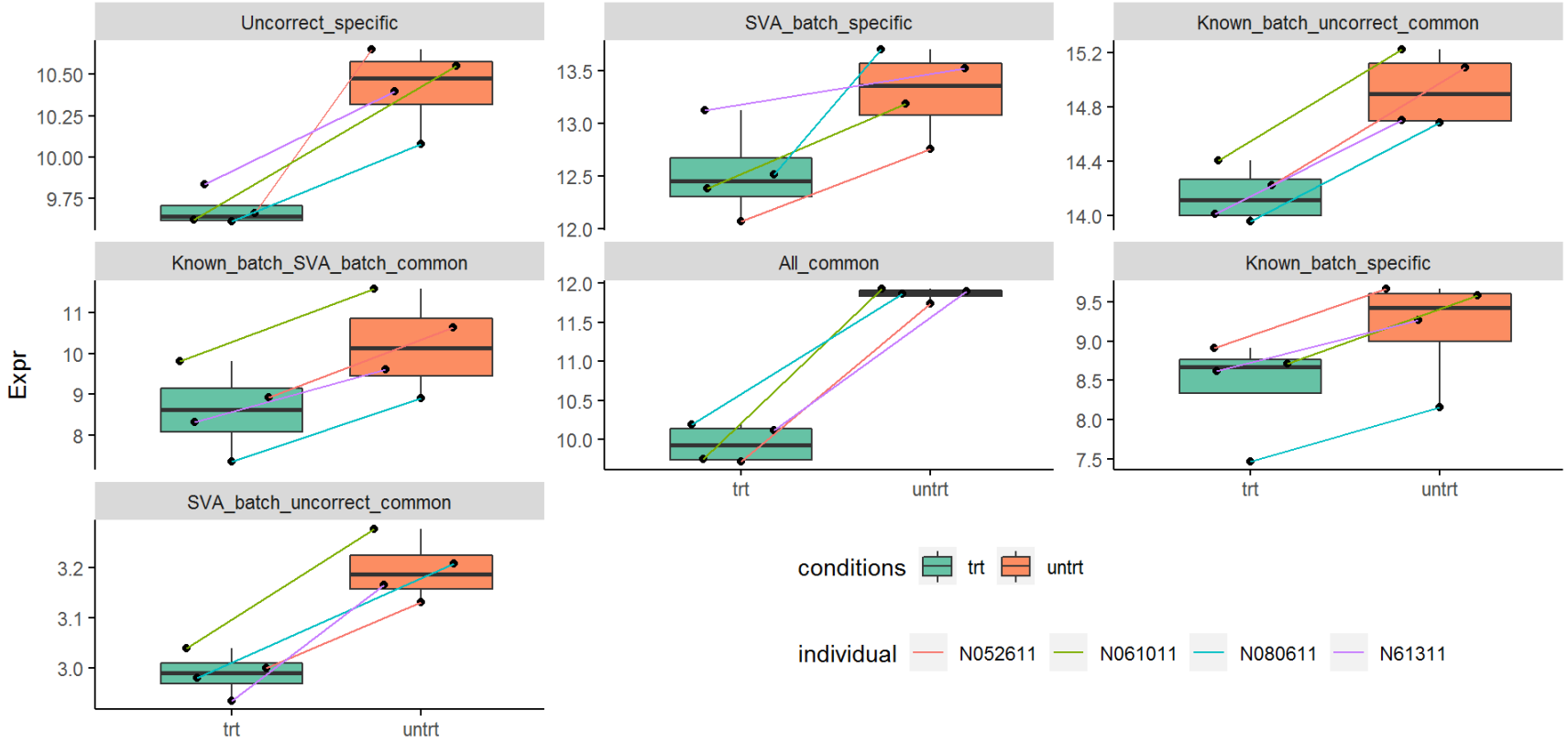
无批次校正、校正已知批次、校正预测批次结果比较

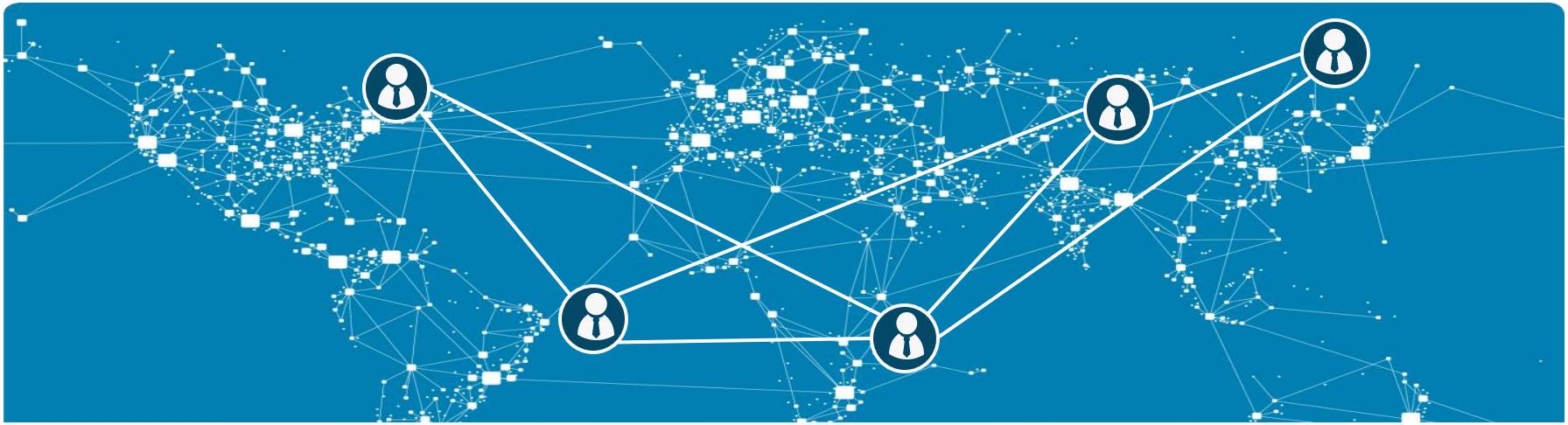
- 批次校正后差异基因增多
- 两种类型的批次校正差异基因结果整体一致



在线Venn图工具

不同类型基因的表达可视化

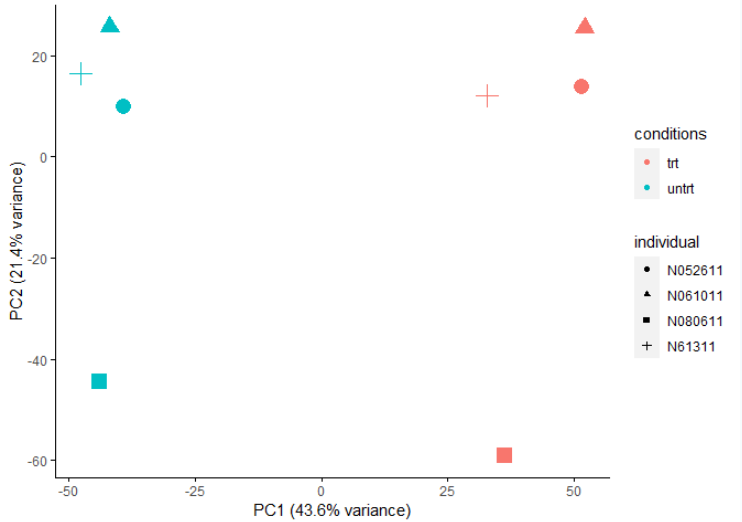




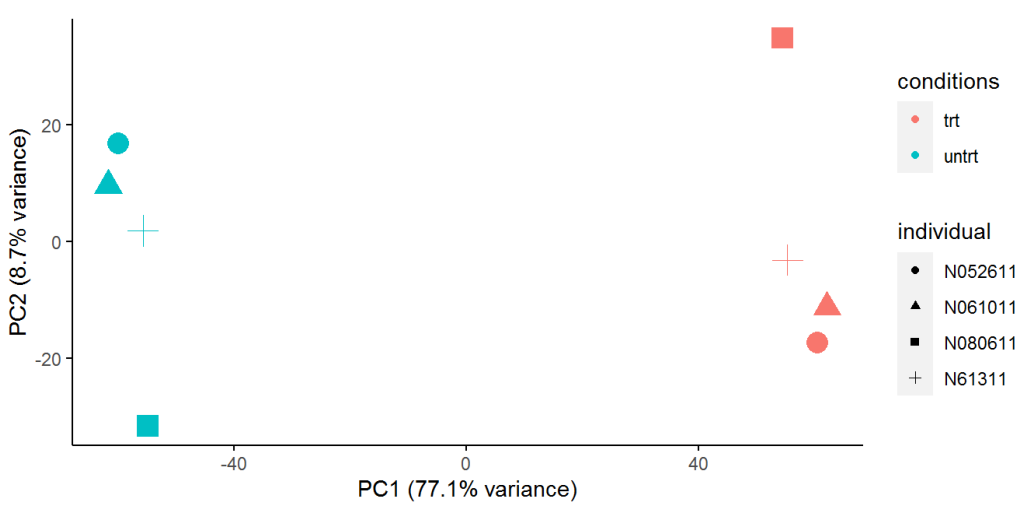
校正表达矩阵

表达矩阵批次校正前后样品PCA可视化结果

批次校正前



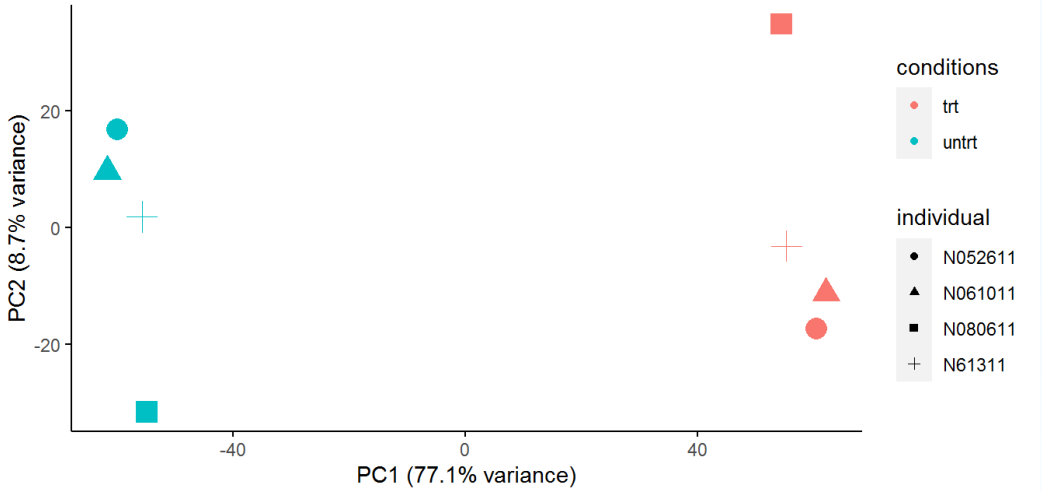
ComBat批次校正后



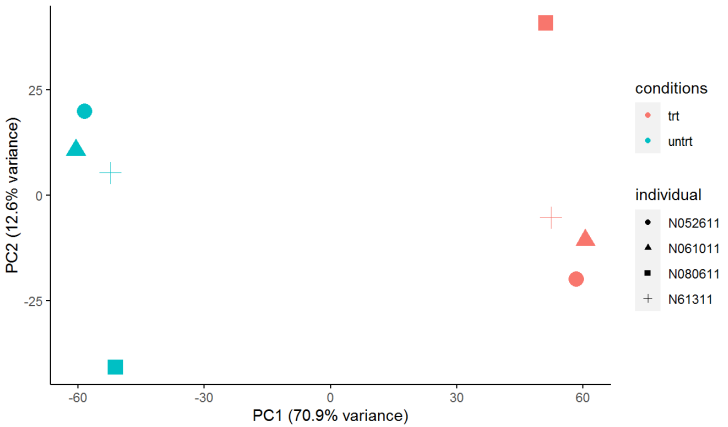
- 校正后PC1解释的差异增大了近2倍
- 校正后不同个体在PC2轴上无一致的分布规律了

ComBat校正和limma校正比较

ComBat校正已知批次后

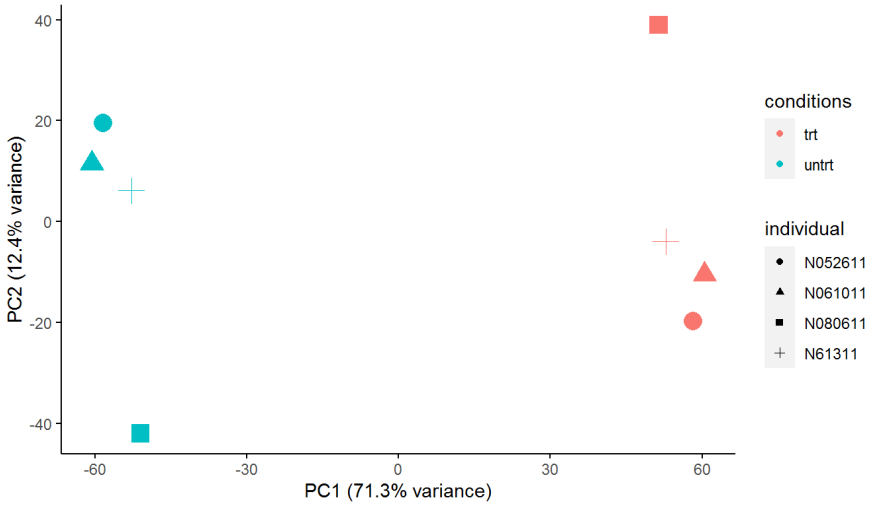


Limma校正已知批次后

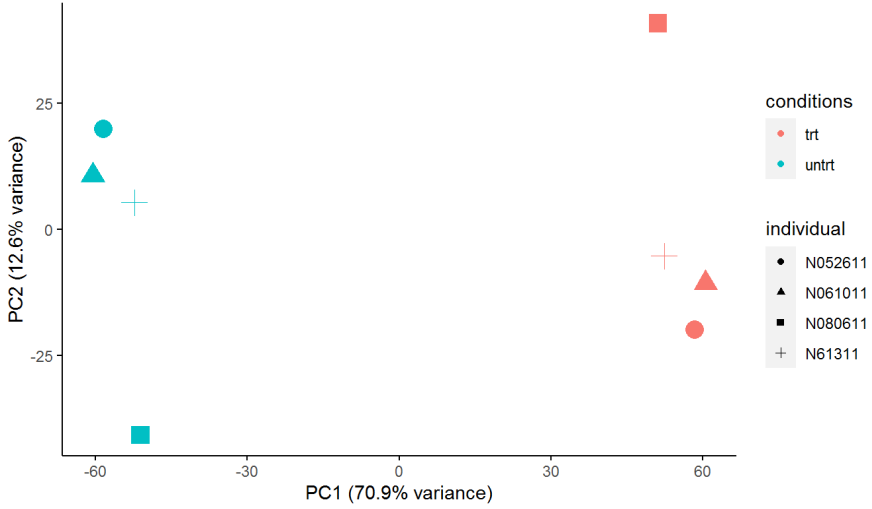


limma校正已知批次和预测批次结果比较

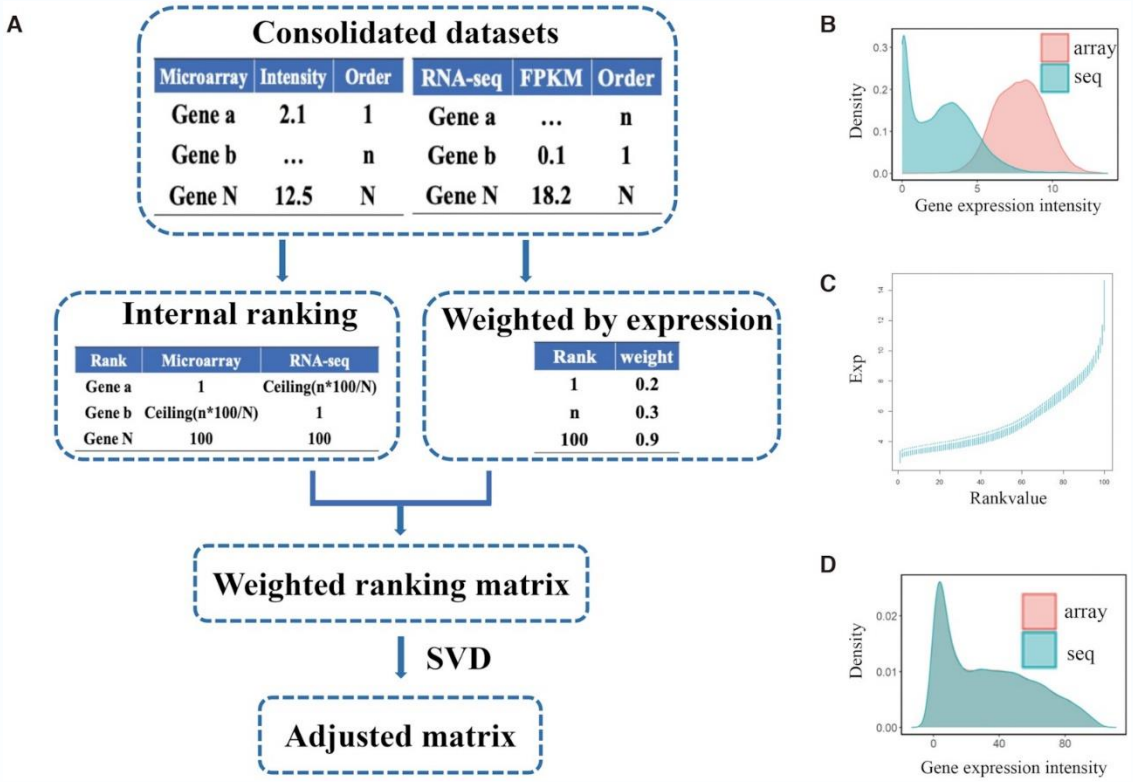
Limma校正预测批次



Limma校正已知批次

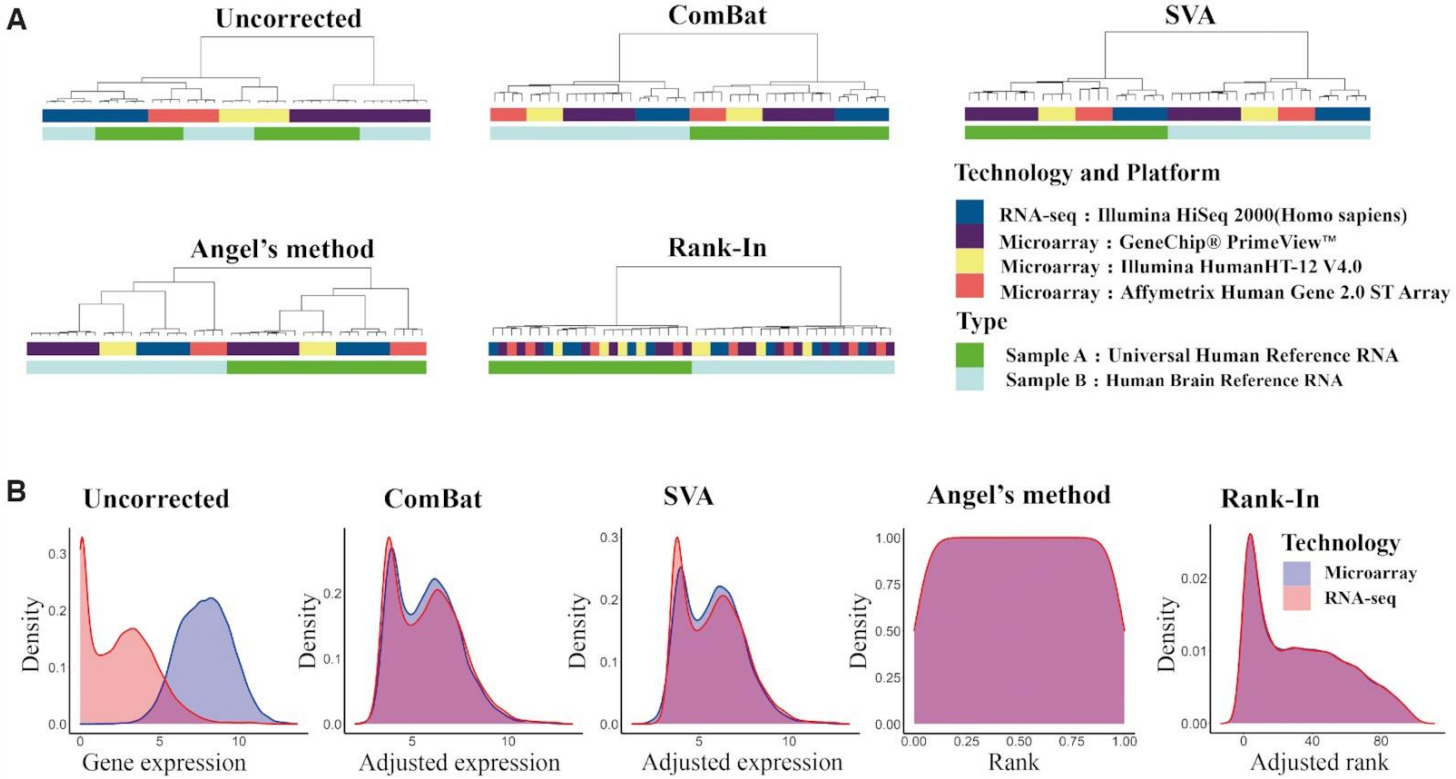


Rankin – 整合芯片和转录组数据



RankIN校正批次

Rank-in更好的抹去了不同数据集的批次



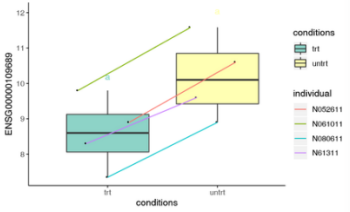
[RankIN在线工具]

相关数据、代码和软件

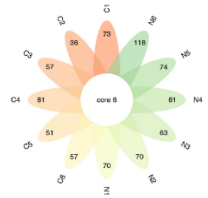
- 数据和代码：http://www.ehbio.com/Bioinfo_R_course/
- Venn图在线绘制：http://www.ehbio.com/test/venn/#/
- PCA、箱线图绘制：http://www.ehbio.com/Cloud_Platform/front/
- 绘图视频：https://space.bilibili.com/362709786

高颜值免费在线绘图 ImageGP

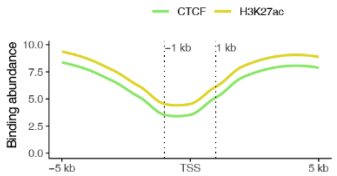
Boxplot



Flower plot



Line plot



Long to wide matrix

Long matrix → Wide matrix

ID	Gene	Expr
ESC	Pou5f1	5.1
iPSC	Pou5f1	5.0
ESC	Sox2	6.2
iPSC	Sox2	6.1
ESC	Nanog	7.8
iPSC	Nanog	7.7

ID	Pou5f1	Sox2	Nanog
ESC	5.1	6.2	7.8
iPSC	5.0	6.1	7.7

Column used as rownames: Gene
Column used as headers: ID
Column used as numeric values: Expr

Merge matrixes

Left matrix

ID	Pou5f1	Sox2
ESC	5.1	6.2
iPSC	5.0	6.1

Right matrix

ID	Stem	Count
ESC	10	600
NSC	9.8	610

Left merge

ID	Pou5f1	Sox2	Stem	Count
ESC	5.1	6.2	10	600
iPSC	5.0	6.1	NA	NA

Inner merge

ID	Pou5f1	Sox2	Stem	Count
ESC	5.1	6.2	10	600

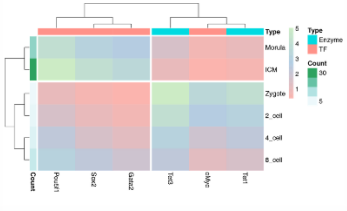
Right merge

ID	Pou5f1	Sox2	Stem	Count
ESC	5.1	6.2	10	600
NSC	NA	NA	9.8	610

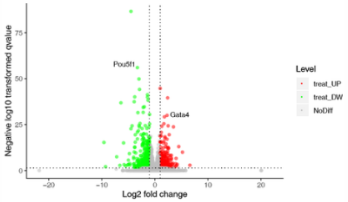
Outer merge

ID	Pou5f1	Sox2	Stem	Count
ESC	5.1	6.2	10	600
iPSC	5.0	6.1	NA	NA
NSC	NA	NA	9.8	610

Pretty heatmap



Volcano plot



Wide to long matrix

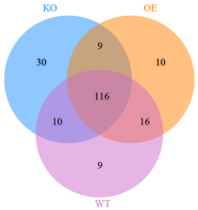
Wide matrix → Long matrix

ID	Pou5f1	Sox2	Nanog
ESC	5.1	6.2	7.8
iPSC	5.0	6.1	7.7

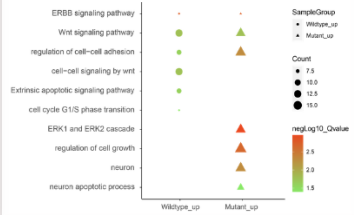
ID	Gene	Expr
ESC	Pou5f1	5.1
iPSC	Pou5f1	5.0
ESC	Sox2	6.2
iPSC	Sox2	6.1
ESC	Nanog	7.8
iPSC	Nanog	7.7

ID variable: ID
Column name for value: Expr
Column name for variable: Gene

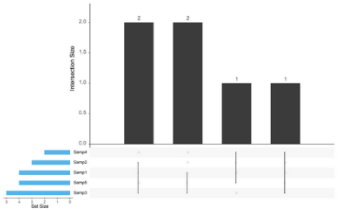
VennDiagram plot



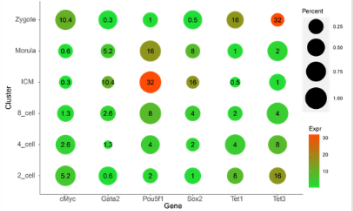
Enrichment plot



Upsetview plot



Scatter plot



高颜值免费在线绘图 ImageGP

ImageGP与iMeta

iMeta 大记事 2022.2-2023.12

01 创刊文章发表 2022.2.21

iMeta：高颜值高被引绘图网站imageGP在线发表

SHORT COMMUNICATION Open Access

ImageGP: An easy-to-use data visualization web server for scientific researchers

Tong Chen, Yong-Xin Liu, Luqi Huang

First Published: 21 February 2022

Representative visualization results of ImageGP. ImageGP supports 16 types of images (including heatmap, volcano plot, enrichment bubble plot) and four types of online analysis with up to 26 parameters for customization.

Abstract | Full text | PDF | References | Request permissions

[iMeta | 23/11/28起被ESCI收录创刊起所有文章，24年将获得首个影响因子]

跟17万朋友一起学习生物信息和微生物组知识

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识