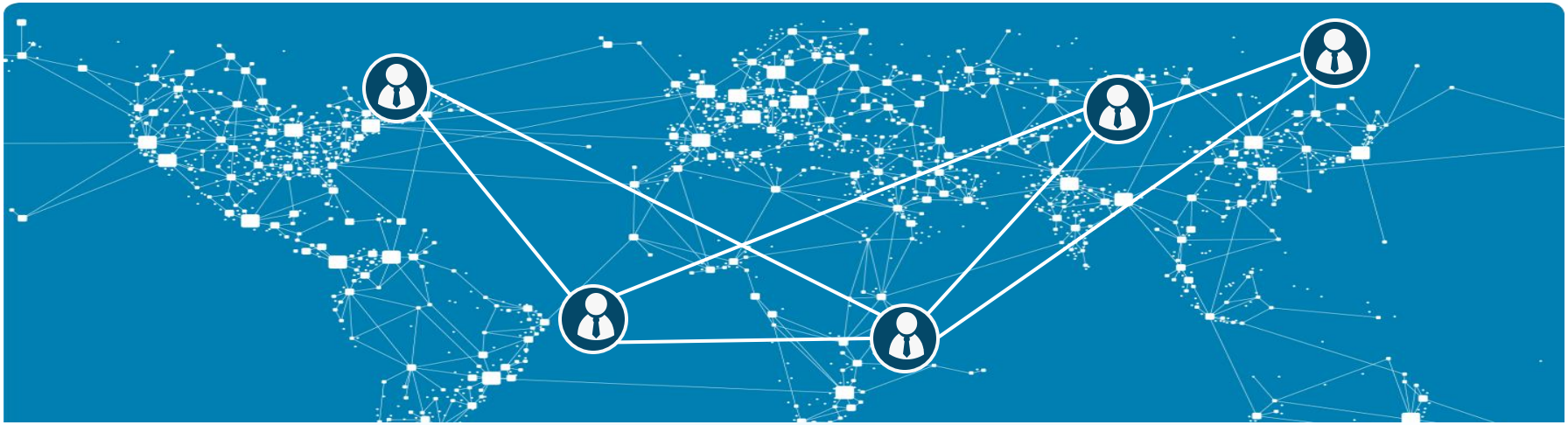


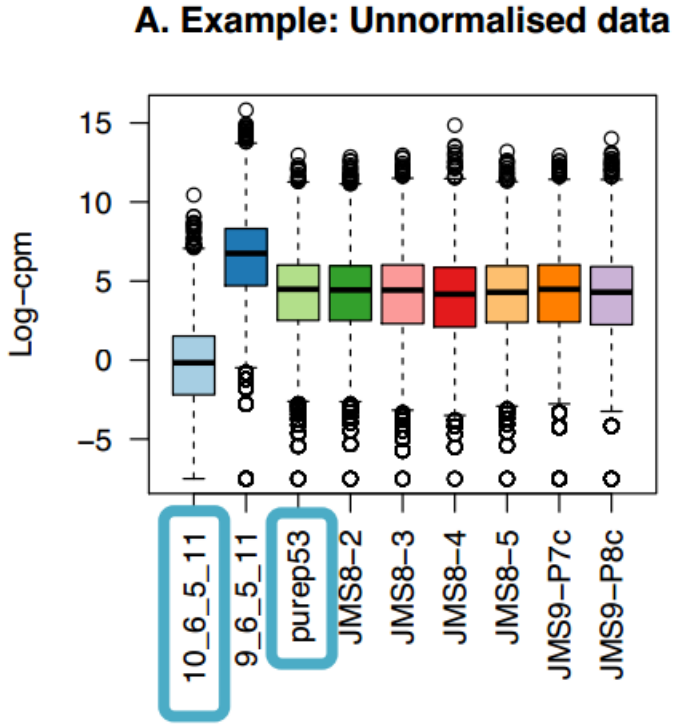
差异基因分析理论和实战

易生信

差异基因分析的几个问题

- 什么是差异基因分析?
  不同组间稳定表达差异的基因筛选
- 为什么要做差异基因分析?
  筛选靶点
- 怎么做差异基因分析?
  标准化
  统计差异检测
  统计筛选

如果不进行标准化会怎样？ - 系统偏差



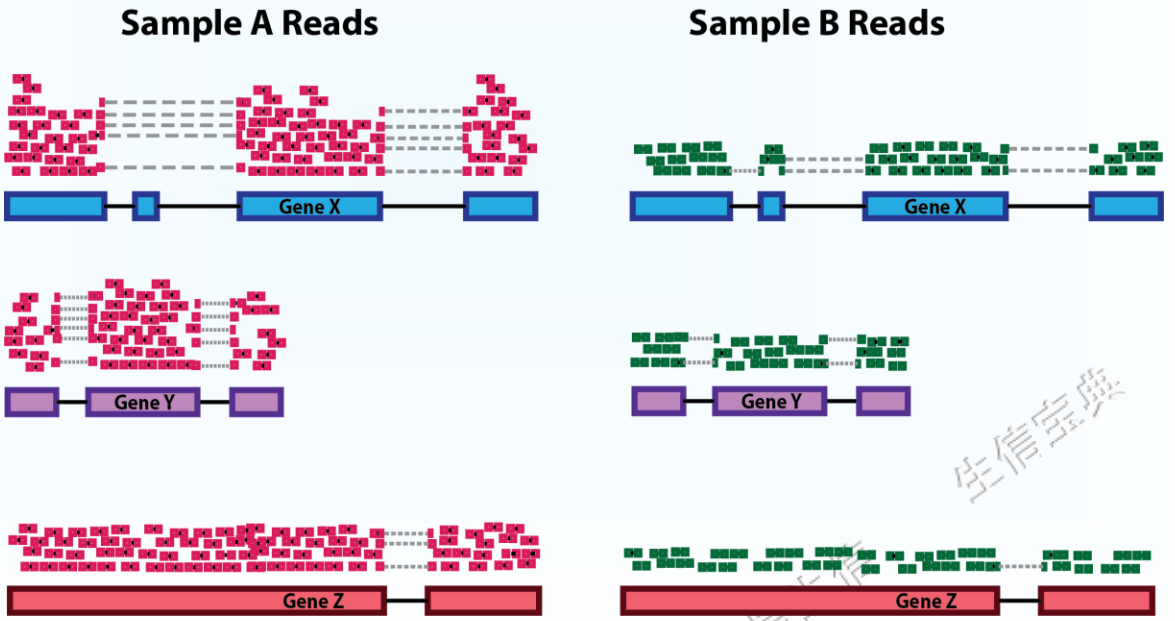
If we ran a DE analysis on Sample 1 and Sample 3, almost all genes will be down-regulated in Sample 1!!

基因表达标准化

- 每个基因的reads计数是与基因的表达量（我们感兴趣的因素）和其他非目标因素的影响成比例的。
- 基因表达标准化就是试图抹去非目标因素的影响，获得在样品内和样品间可比较的基因表达量。

标准化时需要考虑的因素之一：测序深度

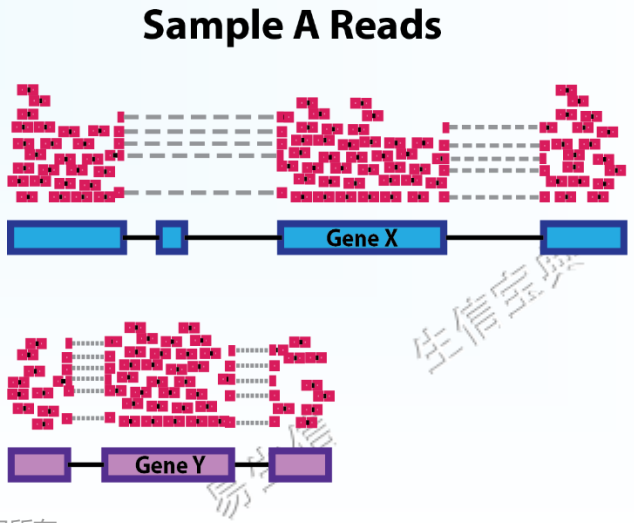
- 样本A中每个基因表达量看上去都是样本B的2倍，这是因为样本A的测序深度是样本B的2倍。



标准化时需要考虑的因素之一：基因长度

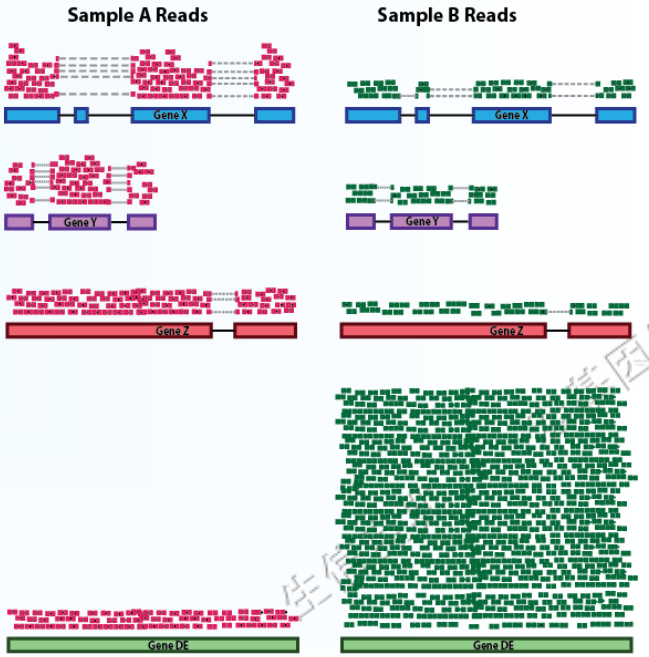
- 样本A中Gene X与Gene Y的表达量相近，但Gene X的reads计数约是Gene Y的2倍，是因为Gene X的长度是Gene Y的2倍？

- Gene X的长度怎么计算？



标准化时需要考虑的因素之一：RNA构成

- 如图，Gene DE在Sample B中有极高的reads count。
- 如果每个基因的标准化方式是原始reads count除以该样本的总reads count，那么原本没有差异的Gene X,Y,Z在Sample A中就会被标准化出更高的丰度，可能会导致其显著高于Sample B。



基因表达标准化

- 不同样品的测序量会有差异，最简单的标准化方式是计算 counts per million (CPM)，即每个样品基因的原始reads count除以样品总可用reads数乘以1,000,000。
- 缺点是容易受到极高表达且在不同样品中存在差异表达的基因的影响；这些基因的打开或关闭会影响到细胞中总的分子数目，可能导致这些基因标准化之后就不存在表达差异了，而原本没有差异的基因标准化之后却有差异了。

$$RPKM = \frac{\text{Read counts} * 10^6 * 1000}{\text{Total reads} * \text{Gene length}}$$

$$RPK = \frac{\text{Read counts} * 1000}{\text{Gene length}}$$

$$TPM = \frac{RPK * 10^6}{\text{sum (RPK)}}$$

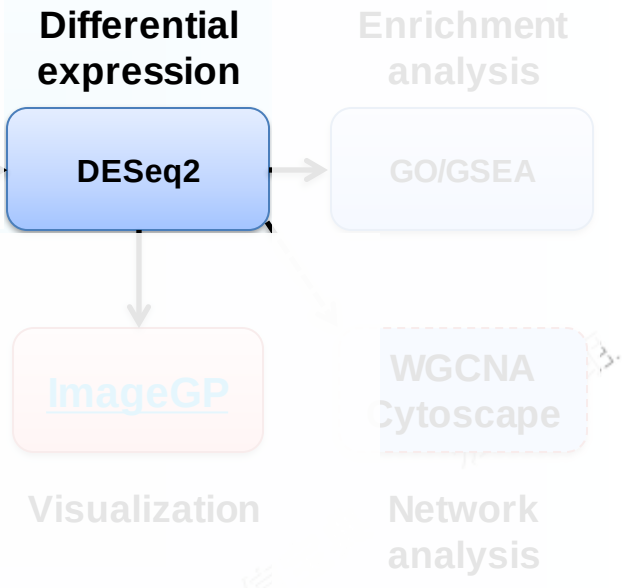
常用标准化方式和适用场景

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons** Except Salmon output.
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

DESeq2: 量化因子标准化

量化因子 (size factor, SF)是由 `DESeq` 提出的。其方法是首先计算每个基因在所有样品中表达的几何平均值。每个细胞的量化因子(size factor)是所有基因与其在所有样品中的表达值的几何平均值的比值的中位数。由于几何平均值的使用，只有在所有样品中表达都不为0的基因才能用来计算。这一方法又被称为 **RLE (relative log expression)**。

```
calc_sf <- function (expr_mat, spikes=NULL){
  geomeans <- exp(rowMeans(log(expr_mat[-spikes,])))
  SF <- function(cnts){
    median((cnts/geomeans)[(is.finite(geomeans) & geomeans >0)])
  }
  norm_factor <- apply(expr_mat[-spikes,],2,SF)
  return(t(t(expr_mat)/norm_factor))
}
```



DESeq2标准化步骤拆解 — 生成测试数据，获取gene的几何平均值

```r
matrix <- matrix(sample(0:10,16, replace=T),nrow=4)
rownames(matrix) <- paste0("Gene",1:nrow(matrix))
colnames(matrix) <- paste0("Samp",1:ncol(matrix))
matrix
```

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

DESeq2标准化步骤拆解 — 生成测试数据，获取gene的几何平均值

```r
matrix <- matrix(sample(0:10,16, replace=T),nrow=4)
rownames(matrix) <- paste0("Gene",1:nrow(matrix))
colnames(matrix) <- paste0("Samp",1:ncol(matrix))
matrix
```

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

对数转换

```r
matrix_log <- log(matrix)
matrix_log
```

	Samp1	Samp2	Samp3	Samp4
Gene1	1.609438	-Inf	0.6931472	2.079442
Gene2	0.000000	2.197225	2.1972246	1.791759
Gene3	1.386294	1.945910	2.1972246	1.609438
Gene4	2.197225	2.302585	0.6931472	-Inf

DESeq2标准化步骤拆解 – 生成测试数据，获取gene的几何平均值

```r
matrix <- matrix(sample(0:10,16, replace=T),nrow=4)
rownames(matrix) <- paste0("Gene",1:nrow(matrix))
colnames(matrix) <- paste0("Samp",1:ncol(matrix))
matrix
```

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

⬇ 对数转换

```r
matrix_log <- log(matrix)
matrix_log
```

	Samp1	Samp2	Samp3	Samp4
Gene1	1.609438	-Inf	0.6931472	2.079442
Gene2	0.000000	2.197225	2.1972246	1.791759
Gene3	1.386294	1.945910	2.1972246	1.609438
Gene4	2.197225	2.302585	0.6931472	-Inf

⬇ 行平均值

```r
matrix_rowmeans <- rowMeans(matrix_log)
matrix_rowmeans
```

Gene1	Gene2	Gene3	Gene4
-Inf	1.546552	1.784717	-Inf

DESeq2标准化步骤拆解 – 生成测试数据，获取gene的几何平均值

```r
matrix <- matrix(sample(0:10,16, replace=T),nrow=4)
rownames(matrix) <- paste0("Gene",1:nrow(matrix))
colnames(matrix) <- paste0("Samp",1:ncol(matrix))
matrix
```

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

对数转换

```r
matrix_log <- log(matrix)
matrix_log
```

	Samp1	Samp2	Samp3	Samp4
Gene1	1.609438	-Inf	0.6931472	2.079442
Gene2	0.000000	2.197225	2.1972246	1.791759
Gene3	1.386294	1.945910	2.1972246	1.609438
Gene4	2.197225	2.302585	0.6931472	-Inf

行平均值

```r
matrix_rowmeans <- rowMeans(matrix_log)
matrix_rowmeans
```

Gene1	Gene2	Gene3	Gene4
-Inf	1.546552	1.784717	-Inf

转换为与原始矩阵同样的标度，
获得每个基因在所有样本的几何平均值

```r
geo_means <- exp(matrix_rowmeans)
geo_means
```

Gene1	Gene2	Gene3	Gene4
0.000000	4.695254	5.957892	0.000000

DESeq2标准化步骤拆解 – 获得量化因子

原始矩阵

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

几何平均值

Gene1	Gene2	Gene3	Gene4
0.000000	4.695254	5.957892	0.000000

DESeq2标准化步骤拆解 – 获得量化因子

原始矩阵

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

每个样品的基因对应除以相应的几何均值

```
apply(matrix, 2, function(cnts){(cnts/geo_means)})
```

几何平均值

Gene1	Gene2	Gene3	Gene4
0.000000	4.695254	5.957892	0.000000

	Samp1	Samp2	Samp3	Samp4
Gene1	Inf	NaN	Inf	Inf
Gene2	0.2129810	1.916829	1.916829	1.277886
Gene3	0.6713784	1.174912	1.510601	0.839223
Gene4	Inf	Inf	Inf	NaN

DESeq2标准化步骤拆解 – 获得量化因子

原始矩阵

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

几何平均值

Gene1	Gene2	Gene3	Gene4
0.000000	4.695254	5.957892	0.000000

每个样品的基因对应除以相应的几何均值

```
apply(matrix, 2, function(cnts){(cnts/geo_means)})
```

	Samp1	Samp2	Samp3	Samp4
Gene1	Inf	NaN	Inf	Inf
Gene2	0.2129810	1.916829	1.916829	1.277886
Gene3	0.6713784	1.174912	1.510601	0.839223
Gene4	Inf	Inf	Inf	NaN

对每个样本求取中位数，获得size factor

```
norm_factor = apply(matrix, 2,
function(cnts){median((cnts/geo_means)[(is.finite(geo_means) & geo_means >0)])})
```

Samp1	Samp2	Samp3	Samp4
0.4421797	1.5458707	1.7137153	1.0585546

DESeq2标准化步骤拆解 – 除以量化因子获得标准化的数据

原始矩阵

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

每个样品的基因对应除以相应的几何均值

```
apply(matrix, 2, function(cnts){(cnts/geo_means)})
```

对每个样本求取中位数，获得size factor

```
norm_factor = apply(matrix, 2,
function(cnts){median((cnts/geo_means)[(is.finit
e(geo_means) & geo_means >0)])})
```

原始矩阵除以自己样品对应的 size factor

```
t(t(matrix)/norm_factor)
```

几何平均值

Gene1	Gene2	Gene3	Gene4
0.000000	4.695254	5.957892	0.000000

	Samp1	Samp2	Samp3	Samp4
Gene1	Inf	NaN	Inf	Inf
Gene2	0.2129810	1.916829	1.916829	1.277886
Gene3	0.6713784	1.174912	1.510601	0.839223
Gene4	Inf	Inf	Inf	NaN

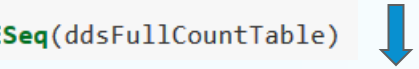
Samp1	Samp2	Samp3	Samp4
0.4421797	1.5458707	1.7137153	1.0585546

	Samp1	Samp2	Samp3	Samp4
Gene1	11.307620	0.000000	1.167055	7.557475
Gene2	2.261524	5.821962	5.251747	5.668106
Gene3	9.046096	4.528192	5.251747	4.723422
Gene4	20.353715	6.468846	1.167055	0.000000

DESeq2中标准化只需要一个函数即可

	Samp1	Samp2	Samp3	Samp4
Gene1	5	0	2	8
Gene2	1	9	9	6
Gene3	4	7	9	5
Gene4	9	10	2	0

```
dds <- DESeq(ddsFullCountTable)
```



	Samp1	Samp2	Samp3	Samp4
Gene1	11.307620	0.000000	1.167055	7.557475
Gene2	2.261524	5.821962	5.251747	5.668106
Gene3	9.046096	4.528192	5.251747	4.723422
Gene4	20.353715	6.468846	1.167055	0.000000

DESeq2输入文件 (salmon)

salmon.output

Samp	SalmonOutput
untrt_N61311	untrt_N61311/untrt_N61311.salmon.count/quant.sf
untrt_N052611	untrt_N052611/untrt_N052611.salmon.count/quant.sf
untrt_N080611	untrt_N080611/untrt_N080611.salmon.count/quant.sf
untrt_N061011	untrt_N061011/untrt_N061011.salmon.count/quant.sf
trt_N61311	trt_N61311/trt_N61311.salmon.count/quant.sf
trt_N052611	trt_N052611/trt_N052611.salmon.count/quant.sf
trt_N080611	trt_N080611/trt_N080611.salmon.count/quant.sf
trt_N061011	trt_N061011/trt_N061011.salmon.count/quant.sf

genome/GRCh38.tx2gene

txname	gene
ENST00000456328	ENSG00000223972
ENST00000450305	ENSG00000223972
ENST00000488147	ENSG00000227232
ENST00000619216	ENSG00000278267

一定注意quant.sf文件是否存在

txiimport

sampleFile

Samp	conditions
untrt_N61311	untrt
untrt_N052611	untrt
untrt_N080611	untrt
untrt_N061011	untrt
trt_N61311	trt
trt_N052611	trt

DESeqDataSetFromTximport

DESeq2

DESeq2输入文件 (reads count)

Reads count

ENSG	untrt_N61311	untrt_N052611	untrt_N080611	untrt_N061011	trt_
ENSG00000223972	1	0	0	0	1	0	0
ENSG00000227232	13	25	23	24	12	12	22	22
ENSG00000278267	0	5	3	4	2	4	3	1
ENSG00000237613	1	0	0	0	0	0	0	0
ENSG00000238009	0	0	0	0	1	0	0	0
ENSG00000268903	0	2	0	0	0	2	0	3
ENSG00000269981	0	3	0	1	0	2	0	0
ENSG00000241860	3	11	1	5	3	2	0	2
ENSG00000279928	0	0	0	0	1	0	0	0
ENSG00000279457	46	90	73	49	52	46	89	31
ENSG00000273874	1	0	0	0	0	0	0	1
ENSG00000228463	5	4	13	6	5	0	11	4
ENSG00000237094	0	16	7	2	1	3	3	2
ENSG00000230021	5	3	2	0	2	1	0	1
ENSG00000225972	5	0	3	0	1	2	5	0
ENSG00000225630	762	868	992	525	735	633	1215	669
ENSG00000276171	0	1	2	1	1	1	0	0
ENSG00000237973	1082	1010	1361	925	1237	821	1632	976

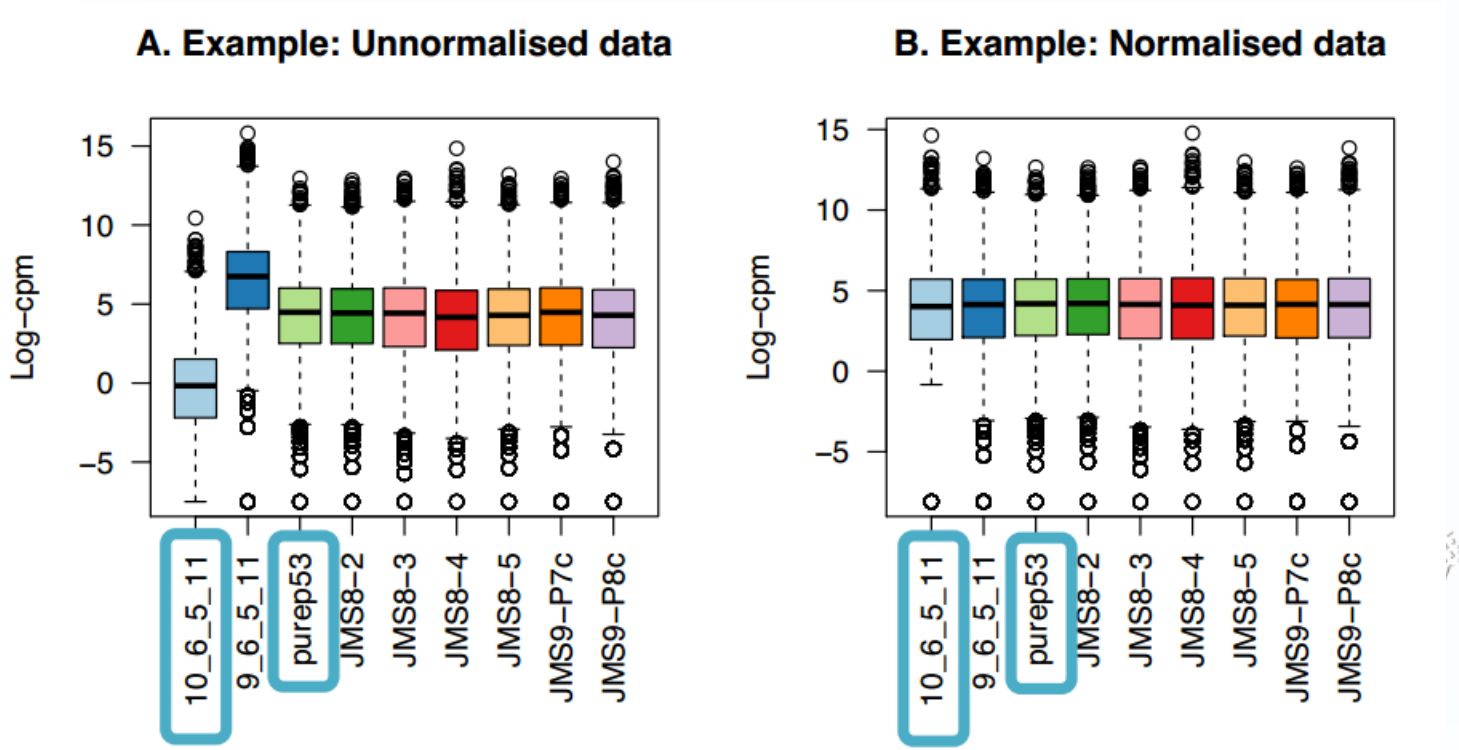
sampleFile

Samp	conditions
untrt_N61311	untrt
untrt_N052611	untrt
untrt_N080611	untrt
untrt_N061011	untrt
trt_N61311	trt
trt_N052611	trt

DESeq2

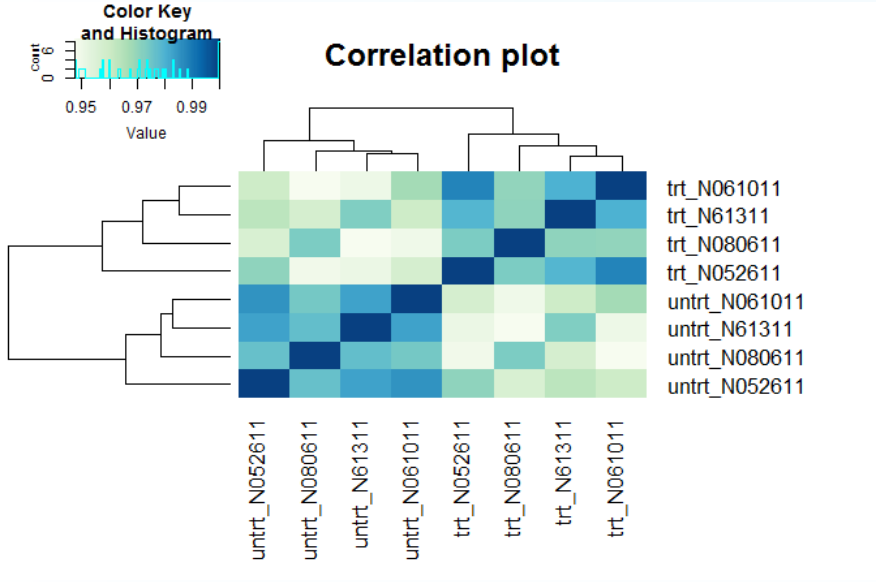
DESeqDataSetFromMatrix

标准化前后数据分布比较: 好的标准化分布一致

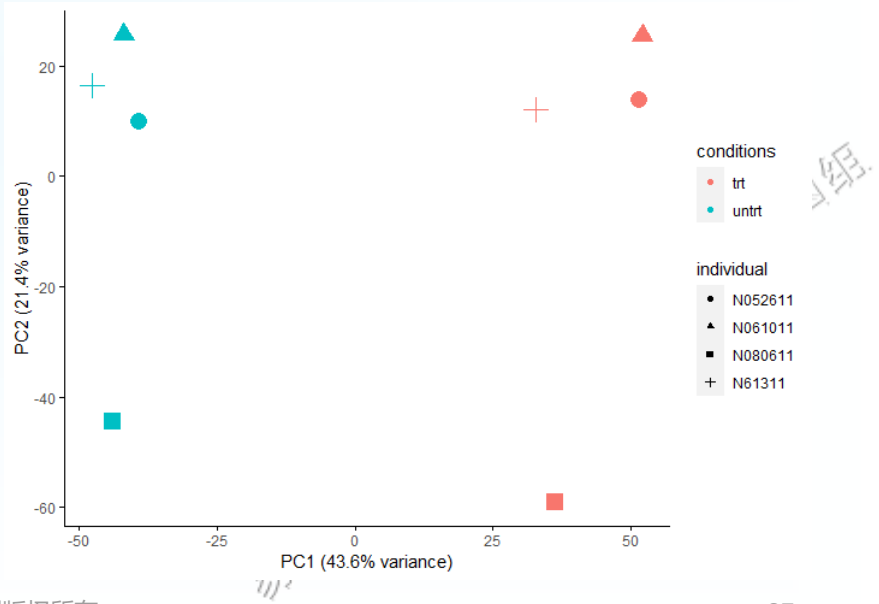


标准化表达矩阵相关性评估 – 样品聚类和PCA

样品整体表达谱相关性，横纵轴都代表样品，图是对称的。两个样品对应的热图区块颜色越深，相关性越大。



把基因根据其贡献度映射到两个或多个主成分，利用主成分降维从而可以在平面上展示样品之间的相似程度。距离越近越相似。



聚类和PCA结果的影响因素

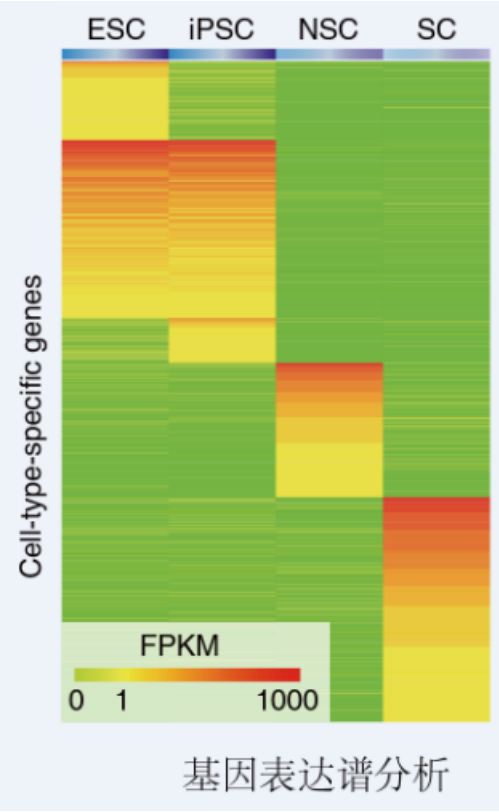
- 样品相似性计算方式 (默认log转换后的数据近似服从正态分布)
- 聚类算法
- PCA默认计算样本欧式距离，也可以采用其它距离(PCoA)
- 选取Top 50、100、500、1000 变化最大的基因 (选择标准方差、中位绝对偏差)
- 用全部基因、用差异基因
- PCA计算中是否归一化（scale）
- 表达值是否进行log转换、sqrt转换等

DESeq2输出结果解释

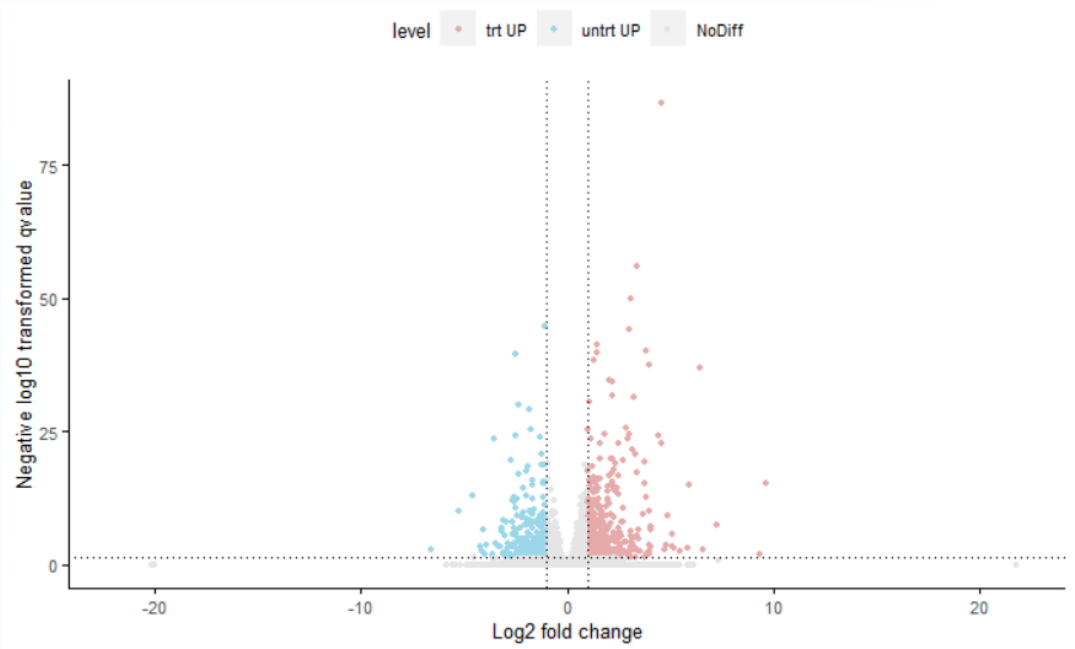
ID	untrt	trt	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000152583	77.9	1900.4	989.1	-4.6	0.2	-21.8	7.80E-106	1.33E-101
ENSG00000148175	7233.0	19483.6	13358.3	-1.4	0.1	-16.5	1.67E-61	1.43E-57
ENSG00000179094	151.7	1380.9	766.3	-3.2	0.2	-15.9	1.01E-56	5.72E-53
ENSG00000134686	1554.0	4082.4	2818.2	-1.4	0.1	-15.2	7.07E-52	3.02E-48
ENSG00000125148	1298.0	5958.9	3628.5	-2.2	0.1	-14.9	5.70E-50	1.94E-46
ENSG00000120129	774.0	5975.5	3374.7	-2.9	0.2	-14.8	1.66E-49	4.73E-46
ENSG00000189221	426.5	4223.5	2325.0	-3.3	0.2	-14.6	3.52E-48	8.59E-45
ENSG00000109906	5.5	745.8	375.7	-7.1	0.5	-14.4	3.45E-47	7.36E-44
ENSG00000178695	4482.6	790.3	2636.4	2.5	0.2	14.1	3.80E-45	7.20E-42
ENSG00000101347	1630.2	23605.6	12617.9	-3.9	0.3	-13.9	9.73E-44	1.66E-40
ENSG00000196517	374.9	79.0	226.9	2.2	0.2	13.7	9.40E-43	1.46E-39
ENSG00000162614	2056.6	8395.4	5226.0	-2.0	0.1	-13.5	9.45E-42	1.34E-38
ENSG00000096060	311.5	4779.3	2545.4	-3.9	0.3	-13.4	7.39E-41	9.70E-38
ENSG00000162616	751.9	2167.2	1459.6	-1.5	0.1	-12.7	9.38E-37	1.14E-33
ENSG00000166741	1762.7	7600.1	4681.4	-2.1	0.2	-12.4	1.46E-35	1.66E-32
ENSG00000116584	2978.4	1479.0	2228.7	1.0	0.1	12.1	9.40E-34	1.00E-30
ENSG00000183044	305.0	681.6	493.3	-1.2	0.1	-12.0	3.12E-33	3.14E-30
ENSG00000144369	1869.9	733.6	1301.7	1.4	0.1	11.8	6.36E-32	6.03E-29

- 绘制火山图需要哪几列?

表达图谱热图和差异基因火山图



基因表达谱分析



差异基因火山图

番外：不同分析到底要用什么数据？

- 原始矩阵：原始表达表/Count矩阵
- 标准化：CPM、TPM、RPKM（FPKM）、DESeq2/edgeR
- 样本聚类分析、相关性热图等除差异基因分析的全部场景都可以用以上任意一种标准化的数据。
- 原始Count用于提供给DESeq2/edgeR/limma进行差异基因分析

原始矩阵（sum不等）

ID	Samp1	Samp2
Gene1	5	20
Gene2	10	10
SUM	15	30

CPM/TPM矩阵（sum相等）

ID	Samp1	Samp2
Gene1	3.3	6.7
Gene2	6.7	3.3
SUM	10	10

RPKM/FPKM（sum不等）

ID	Samp1	Samp2
Gene1	3.3	6.7
Gene2	3.4	1.6
SUM	6.7	8.3

番外：不同分析到底要用什么数据？

- 数据转换：log转换(log需要注意只能对正数做转换)、 sqrt转换、Hellinger转换、scale归一化、Wisconsin转换
- 是否转换影响结果好坏，无对错之分

Hellinger转换

任意矩阵

ID	Samp1	Samp2
OTU1	5	10
OTU2	10	5
SUM	15	15

→ 除以每一列的总和

ID	Samp1	Samp2
OTU1	0.33	0.67
OTU2	0.67	0.33
SUM	1	1

→ 然后开平方

ID	Samp1	Samp2
OTU1	0.57	0.82
OTU2	0.82	0.57
SUM		

Wisconsin转换

抽平矩阵（sum相等）

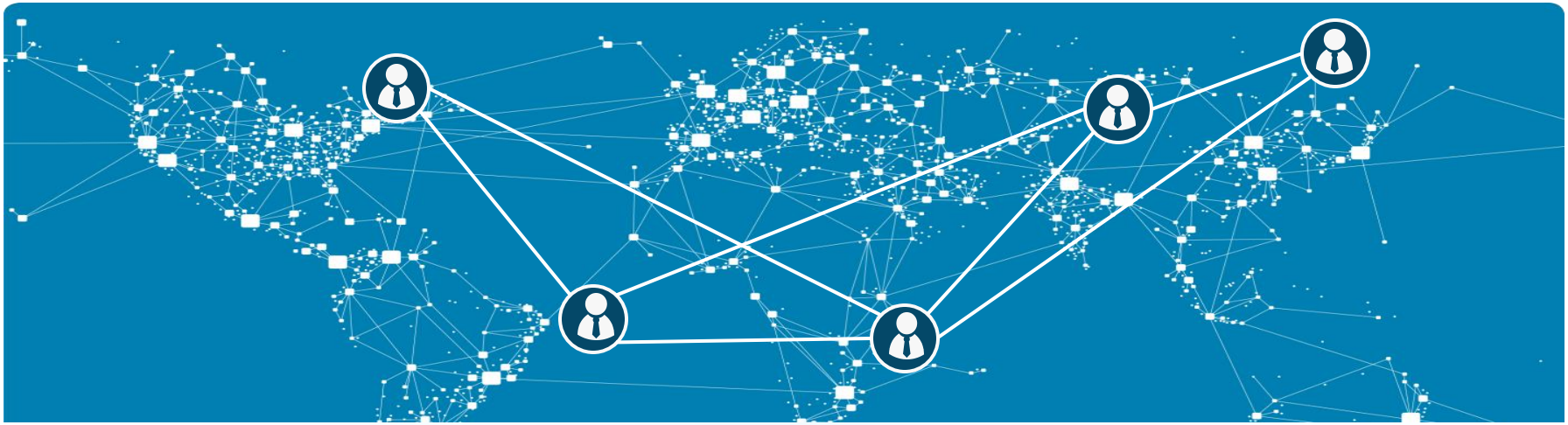
ID	Samp1	Samp2
OTU1	5	10
OTU2	10	5
SUM	15	15

→ 除以每一行最大值

ID	Samp1	Samp2	Max
OTU1	0.5	1	10
OTU2	1	0.5	10
SUM	1.5	1.5	

→ 除以每一列的总和

ID	Samp1	Samp2
OTU1	0.33	0.67
OTU2	0.67	0.33
SUM	1	1



总被审稿人提起的多重假设检验校正是什么？

单次检验的I类错误

假设检验的基本方法是提出一个空假设（`null hypothesis`），也叫做原假设或无效假设，符号是`H0`。一次检验有四种可能的结果，用下面的表格表示：

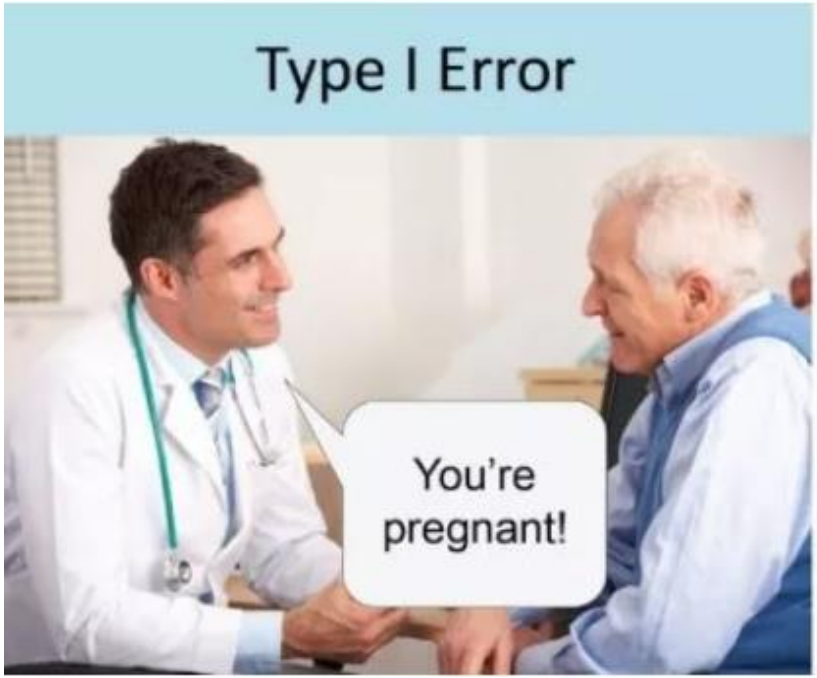
- `Type I error`，I类错误，也叫做`α`错误，假阳性。
- `Type II error`，II类错误，也叫做`β`错误，假阴性。

Actual Situation "Truth"

Decision	H_0 True	H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$$\alpha = P(Type\ I\ Error) \quad \beta = P(Type\ II\ Error)$$

Type I Error：假阳性



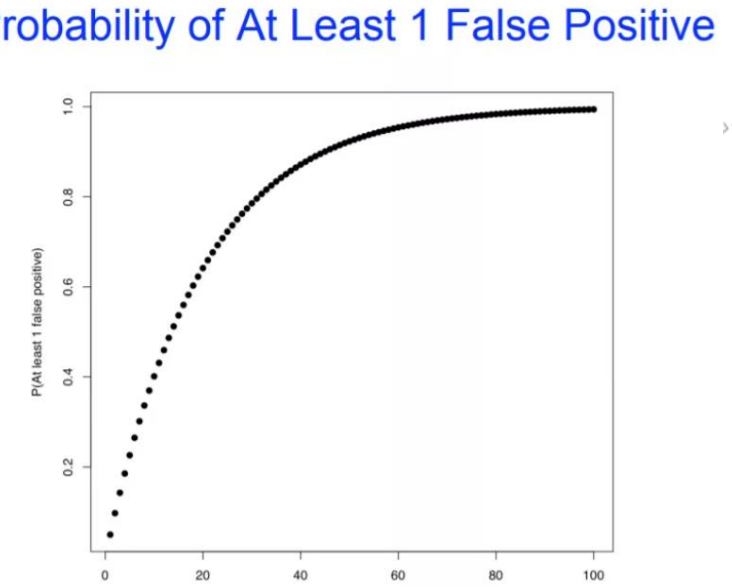
总被审稿人提起的多重假设检验校正是什么？

检验次数增加使得犯I类错误概率增大

在传统的假设检验中，单个检验的显著性水平或 `I` 型错误率 (错误拒绝原假设的概率)为计算出的 `P-value` 。但随着检验次数的增加，错误拒绝原假设的概率即I型错误率大大增加。

例如：如果我们进行了 `m` 次假设检验，至少有 `1` 个假阳性的概率是多少?

```
错误拒绝原假设的概率 P(Reject H0|H0=True) = α
决策正确的概率 P(No Reject H0|H0=True) = 1-α
P(在m次检验全部决策正确)=(1-α)^m
P(在m次检验中至少一次决策错误) = 1-(1-α)^m
```



总被审稿人提起的多重假设检验校正是

控制整体的I型错误率

- 随着检验次数的增多，出现至少一次决策错误的概率快速提高。当说起“根据假设检验的次数校正p值”时，意思是控制整体的I型错误率。
- 例如：当做差异基因检测时，每个基因分别进行检测生成一个p值。如果p值设置为0.05，每个差异基因识别出错的概率为5%。
- 如果同时分析100个基因，按照p<0.05筛选的差异基因中有5个可能是差异不显著的。
- 如果对一组10000个基因进行检测，按照p<0.05筛选的差异基因中有500个可能是差异不显著的。
- 因此，同时进行多次统计检验时，校正每个基因的p值是很重要的。多重检验校正调整每个基因的p值，以使总体错误率小于或等于用户指定的p-cutoff value。

Family Wise Error Rate校正法控制假阳性率为0

- Family Wise Error Rate是控制全部比较中至少出现一次Type I error的概率，也就是控制假阳性率为0。这是很严格的方式。

Bonferroni correction方法

如果要维持整个检测 (做了 `m` 次检测)的 `Type I error rate < 0.05` ，则需要设定 `p-value` 为 `0.05/m` 作为筛选标准。反过来，如果我们做了 `10000` 次统计检测，采用 `Bonferroni correction` 方法校正后的 `p` 值就是 `原始P-value * 10000` 。

当然，我们也只是借这个方法理解校正的计算方式，实际却不用这个方法。

这对其中任何一个检测是否差异统计显著是不公平的，因为它取决于检测的总数目。一个检测放在有 `100` 次检测的操作集合中可能统计显著，而放在有 `1000` 次检测的操作集合中可能统计就不显著了，这是不合适的。

> Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference.
> Perneger (1998)

FDR校正法：允许一定的假阳性率

Benjamini and Hochberg FDR (BH)

这是我们最常用的校正 `P-value` 控制假阳性率的方式。假设针对 `10000` 个基因进行了统计检验，对所有的原始 `P-value` 进行由*小到大*的排序分别为 `p1, p2, ..., p10000`，校正后的 `FDR` 为：`p1*10000/1, p2*10000/2, ..., p10000*10000/10000`。与 `Bonferroni correction` 一致的地方是都乘以了检测总数，不一致的地方是 `BH` 算法在此基础上除去了各个原始 `p-value` 的排序值。

具体计算方式见下表（总检测次数为 `10` 次；控制 `FDR` 小于 `0.1`）

BH计算方法

具体计算方式见下表（总检测次数为 `10` 次；控制 `FDR` 小于 `0.1` ）

Rank	P-value	FDR	FDR_formula	Reject H0	Reject_formula
1	0.0008	0.008	=B2*10/A2	TRUE	=C2<0.1
2	0.009	0.045	=B3*10/A3	TRUE	=C3<0.1
3	0.165	0.55	=B4*10/A4	FALSE	=C4<0.1
4	0.205	0.5125	=B5*10/A5	FALSE	=C5<0.1
5	0.396	0.792	=B6*10/A6	FALSE	=C6<0.1
6	0.45	0.75	=B7*10/A7	FALSE	=C7<0.1
7	0.641	0.915714286	=B8*10/A8	FALSE	=C8<0.1
8	0.781	0.97625	=B9*10/A9	FALSE	=C9<0.1
9	0.9	1	=B10*10/A10	FALSE	=C10<0.1
10	0.993	0.993	=B11*10/A11	FALSE	=C11<0.1

`BH` 法有时也称 `fdr` 法，是我们最常用的多重假设检验校正方法，可以很好的控制假阳性率和维持统计检出力。R函数 `p.adjust` 可用来计算一组 `p-value` 校正后的 `fdr` 值。(`DESeq2` 中返回的 `padj` 也是用 `BH` 方法控制的 `FDR`)

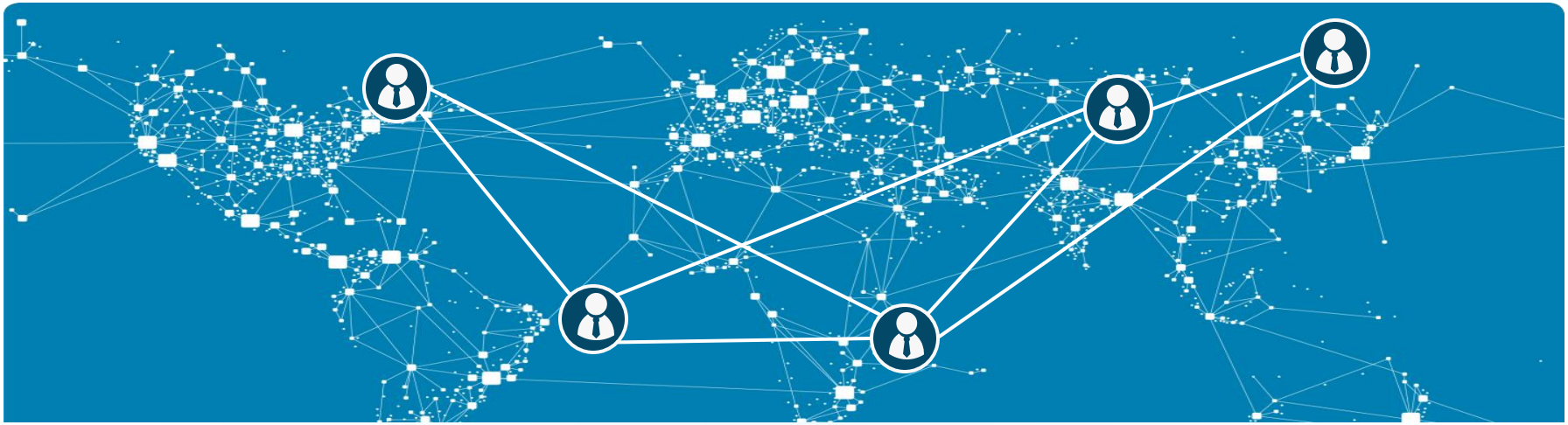
如何尽量减少统计检验次数?

我们看到上面的校正方法多于统计检测次数有关，统计检测次数越多，校正也会越强烈。有没有合适的办法来规避一些无意义的统计检验呢?

- WGCNA方法通过把基因聚类为模块再进行统计分析，大大降低了统计检验次数，具体见WGCNA分析，简单全面的最新教程
- GSEA、GO等富集分析时合并相似的GO/KEGG通路再进行富集分析，如一文掌握GSEA，超详细教程中提到的合并共有基因数目超过70%的通路。
- 差异基因分析时过滤掉极低表达的基因 (*低表达基因通常生物意义小或检测噪声大，即便有差异也难分清是生物差异还是技术差异*)，如高通量数据中批次效应的鉴定和处理 - 系列总结和更新提到的方法。

  `DESeq2` 中还额外进行了 `independent filtering` 进行进一步过滤提高统计检出率。

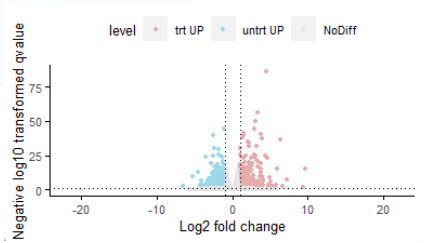
  没有通过过滤标准的基因校正后的 `padj` 赋值为 `NA` (这也是之前总被问起的 `DESeq2` 结果中 `NA` 的来源)。



火山图

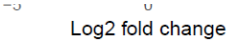
差异OTU火山图展示

- 火山图是散点图的一种，类似图的解析方式都是先看横纵坐标轴代表什么含义，图例是什么含义，图形解释是什么。



- 每个点代表一个检测到的**基因**。
- 横轴和纵轴用于固定点在空间的位置。
- 一般横轴是 `Log2(fold change)`，点越偏离中心，表示差异倍数越大。
- 纵轴是 `-Log 10 (adjusted P-value)`，点越靠图的顶部表示差异越显著。
- 点的大小和颜色也可以表示更多的属性，如下图中点的颜色标记其对应的基因是 `上调`，`下调` 还是 `无差异`。

  大小也可用于展示基因表达的平均丰度，一般我们关注表达水平较高且差异较大的基因用于后续的分析和验证。



火山图详细解读：Volcano plot | 别再问我这为什么是火山图

差异OTU火山图展示

1. **什么是 `fold change` ?**

   翻译成中文是 `差异倍数` ，简单来说就是基因在一组样品中的表达值的均值除以其在另一组样品中的表达值的均值。所以火山图只适合展示两组样品之间的比较。

2. **为什么要做 `Log 2` 转换?**

   两个数相除获得的结果 （ `fold change` ）要么 `大于` 1，要么 `小于` 1，要么 `等于` 1。这是一句正确的废话吧？那么对应于基因差异呢？简单说，大于1表示上调（可以描述为上调多少倍），小于1表示下调（可以描述为下调为原来的多少分之多少）。大于1可以到多大呢？多大都有可能。小于1可以到多小呢？最小到0。用原始的 `fold change` 描述上调方便，描述下调不方便。绘制到图中时，上调占的空间多，下调占的空间少，展示起来不方便。所以一般会做 `Log 2` 转换。默认我们都会用两倍差异 （ `fold change == 2 | 0.5` ）做为一个筛选标准。 `Log2` 转换的优势就体现出来了，上调的基因转换后 `Log2 (fold change)` 都大于等于 `1` ，下调的基因转换后 `Log2 (fold change)` 都小于等于 `-1` 。无论是展示还是描述是不是都更方便了。

差异OTU火山图展示

3. `P-value` 都比较熟悉，统计检验获得的是否统计差异显著的一个衡量值，约定成俗的 `P-value<0.05` 为统计检验显著的常规标准。

4. **什么是 `adjusted P-value` ？**

   这里面就涉及到一个统计学问题了。做差异基因检测时，要对成千上万的基因分别做差异统计检验。统计学家认为做这么多次的检验，本身就会引入假阳性结果，需要做一个多重假设检验校正。

   这个校正怎么做呢？最简单粗暴的方法是每一次统计检验获得的 `P-value` 都乘以总的统计检验的次数获得 `adjusted P-value` (这就是 `Bonferroni correction`) 。

   但这样操作太严苛了，很容易降低统计检出力，找不到有差异的基因。后续又有统计学家提出相对不这么严苛的计算方法，如 `holm` , `hochberg` , `hommel` , `BH` , `BY` , `fdr` 等。 `BH` 是我们比较常用的一个校正方法，获得的值是 `假阳性率 FDR` (`false discovery rate`) 。

   `FDR` 筛选时就可以不用遵循 `0.05` 这个标准了。我们可以设置 `FDR<0.05` 表示我们容许数据中存在至多 `5%` 假阳性率； `FDR<0.1` 表示我们对假阳性率的容忍度至多是 `10%` 。当然如果说我们设置FDR<0.5，即数据中最多可能有一半是假阳性就说不过去了。

5. **同样为什么做 `-Log 10` 转换呢?**

   因为FDR值是 `0-1` 之间，数值越小越是统计显著，也越是我们关注的。 `-Log 10 (adjusted P-value)` 转换后正好是反了多来，数值越大越显著，而且以 `10` 为底很容易换算回去。

差异火山图数据格式

- 根据火山图展示哪些信息，就可以选择对应的数据列。如上面第一幅图用到了logFC, logCPM, level信息；第二幅图用到了logFC,FDR,level信息。
- 差异分析结果可以用edgeR或DESeq2生成。

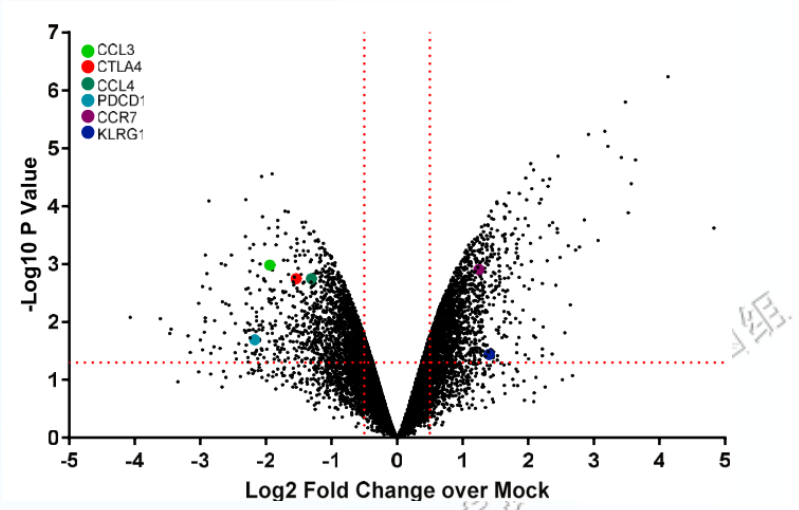
Table 6.1: 差异分析表格

	logFC	logCPM	PValue	FDR	level
OTU_5	-3.016	13.971	0.0e+00	0.0000466	Depleted
OTU_1890	-4.980	6.724	2.0e-06	0.0019933	Depleted
OTU_1763	-3.521	7.314	2.7e-06	0.0019933	Depleted
OTU_248	-2.394	9.074	3.4e-06	0.0019933	Depleted
OTU_1574	-4.940	6.705	3.5e-06	0.0019933	Depleted

R 学习 - 火山图

差异基因火山图展示

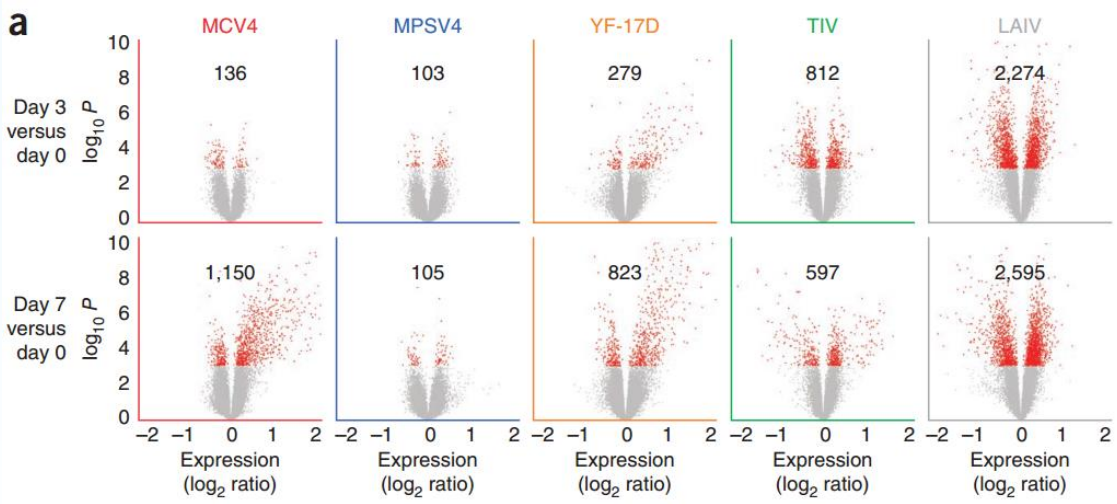
- (I) Volcano plot of relative RNA expression for $CD45^{+}CD3^{+}CD8^{+}$ FACS-obtained lymphocytes isolated from tumor-bearing lungs of 3-month Aza + ITF-2357-treated LSL-$Kras^{G12D}$ mice as compared to mock mice. Genes in the upper left and right quadrants are significantly differentially expressed (microarray, n = 2 per group). Highlighted genes are involved in T cell fate determination.



Molecular signatures of antibody responses derived from a systems biology study of five human vaccines

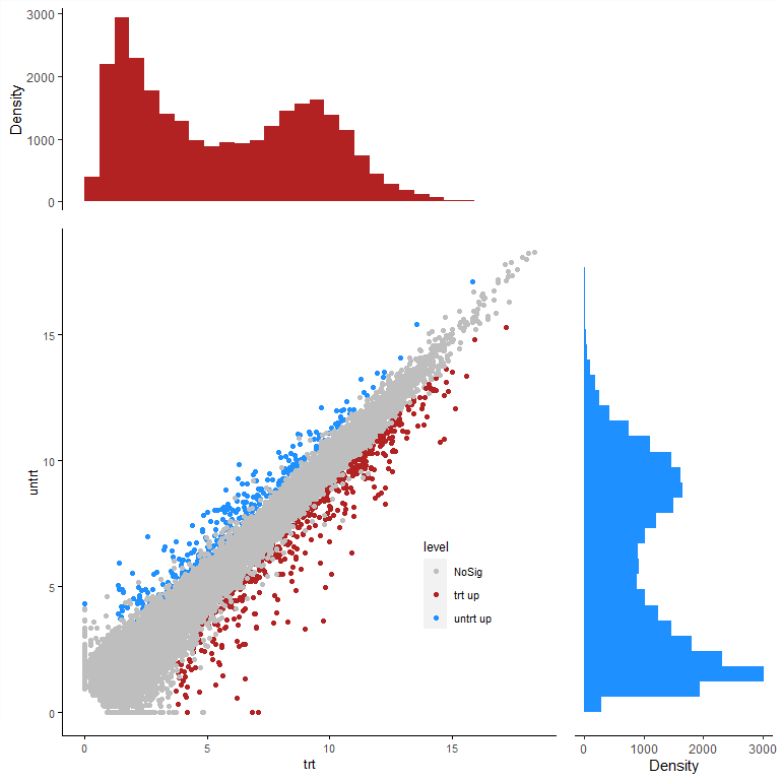
差异基因火山图展示

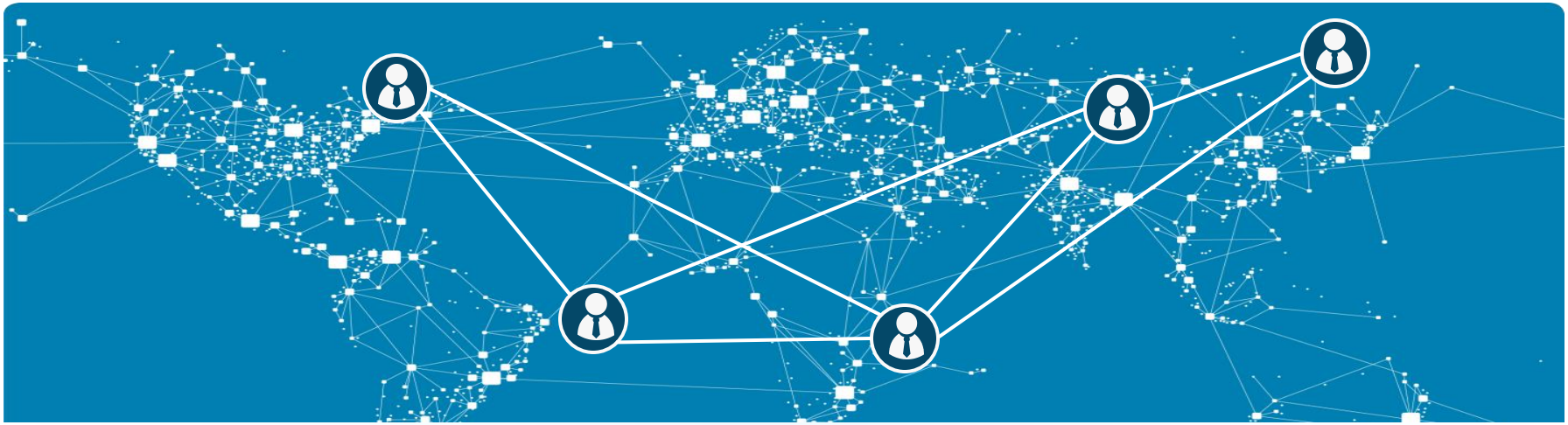
- Analysis of blood transcriptomic data from five human vaccines. (a) Differential expression analysis was performed using paired t-test for each vaccine and each time point (day 3 or day 7 compared to baseline). The red dots in volcano plots show differentially expressed genes (DEGs, $p < 0.001$), with the numbers of DEGs.



Molecular signatures of antibody responses derived from a systems biology study of five human vaccines

什么是倾斜45度的火山图?

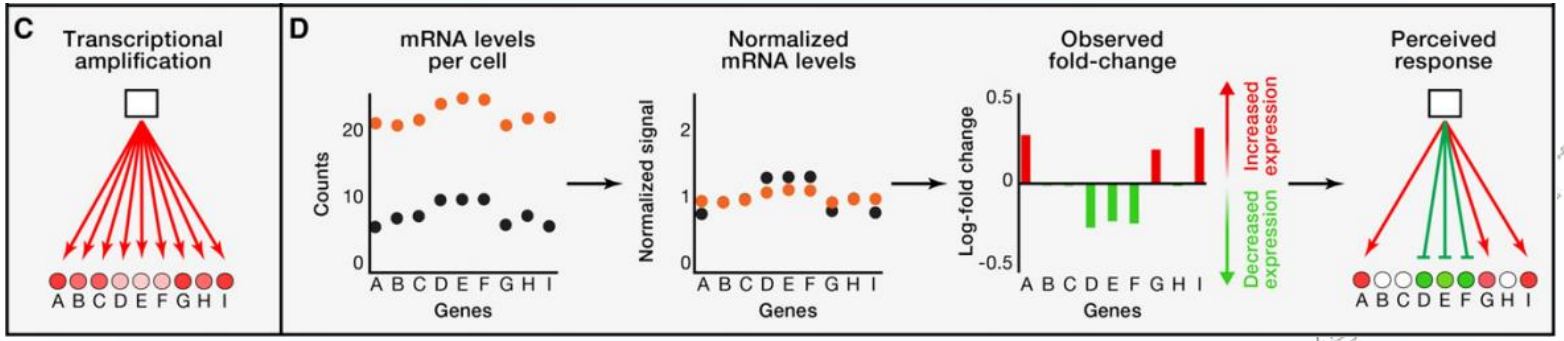




其他差异基因相关技术

细胞RNA总量差异影响分析结果

- 但是有些细胞，如癌细胞中的RNA总量具有极大差异，常规的归一化方法计算的差异基因会导致错误结果。
- cMyc的激活会使大部分基因的表达都提升。

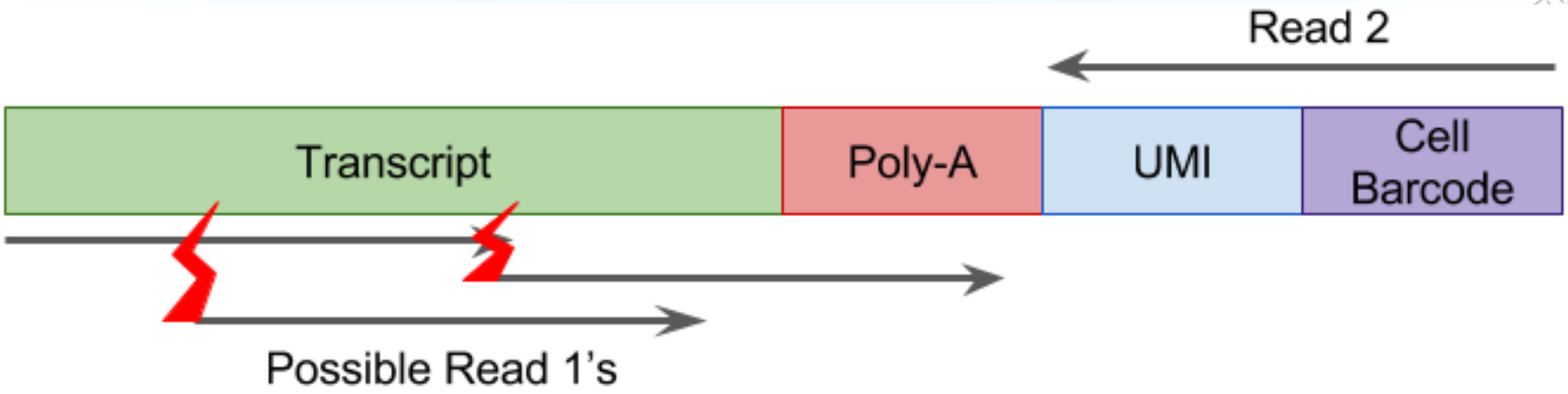


Revisiting Global Gene Expression Analysis

Jakob Lovén,[1,5] David A. Orlando,[1,5] Alla A. Sigova,[1] Charles Y. Lin,[1,2] Peter B. Rahl,[1] Christopher B. Burge,[3] David L. Levens,[4] Tong Ihn Lee,[1,6,*] and Richard A. Young[1,3,6,*]

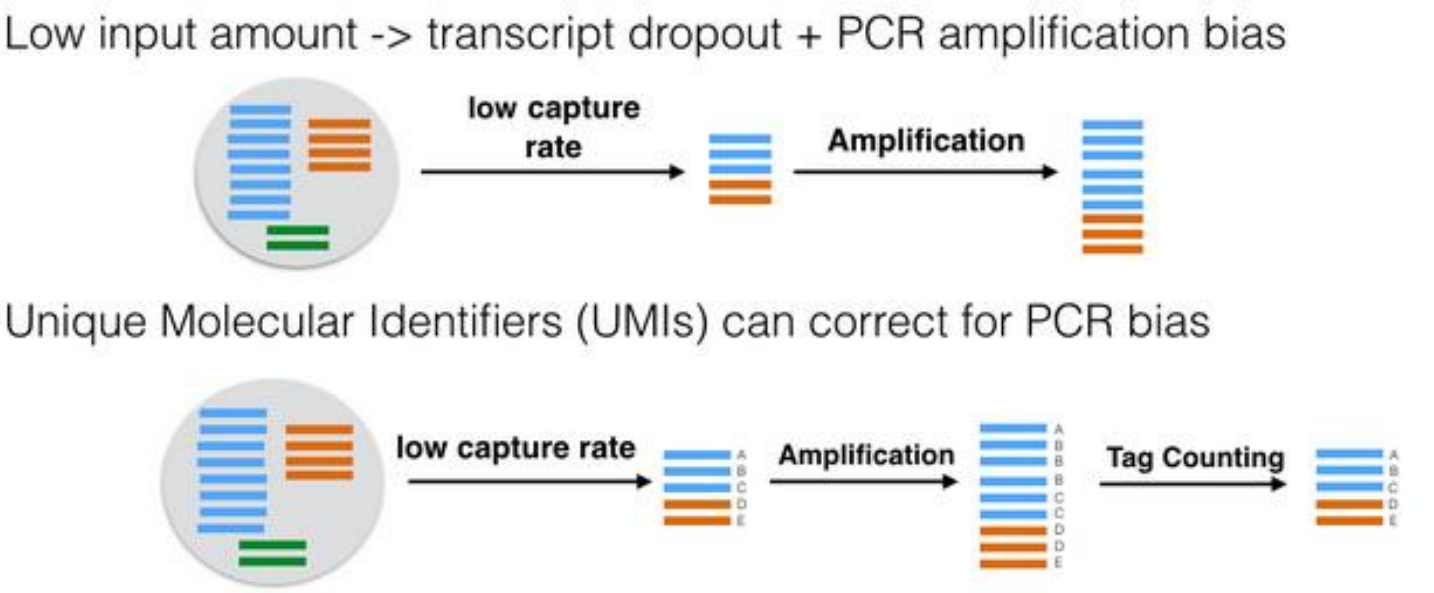
利用UMI去除PCR duplicate的影响，获得更准确定量

- UMI (unique molecular identifier): 反转录过程中添加到转录本上的4-10 bp的随机条形码序列，使得测序reads可以对应到单个转录本，去除扩增噪声和偏好性。



UMI校正PCR偏好性的机制

- UMI校正PCR偏好性



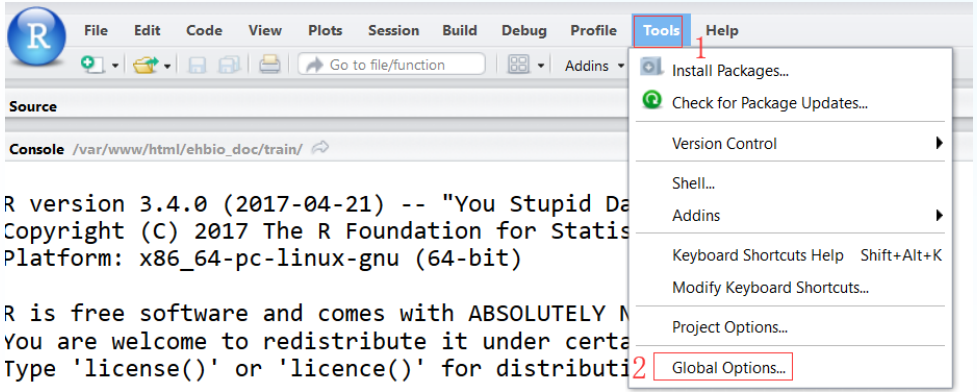
Broad workshop

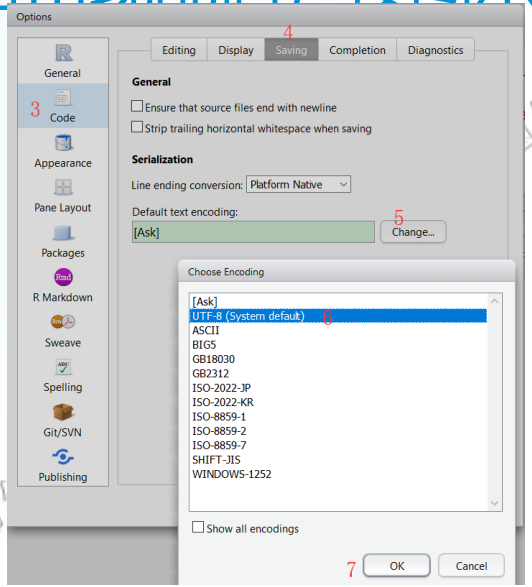
差异基因分析的工具组合

Task	Input data	Software (examples)	Post-processing
DGE	Aggregated transcript counts + average transcript length offsets, or simple counts + average transcript length offsets	Salmon, kallisto, BitSeq, RSEM	
		tximport	
		DESeq2, edgeR, voom/ limma	
DTE	Transcript counts	Salmon, kallisto, BitSeq, RSEM	Optional gene-level aggregation
		tximport	
		DESeq2, edgeR, sleuth, voom/limma	
DTU/DEU	Transcript counts or bin counts, depending on interpretation potential[18]	Salmon, kallisto, BitSeq, RSEM	Optional gene-level aggregation
		DEXSeq	

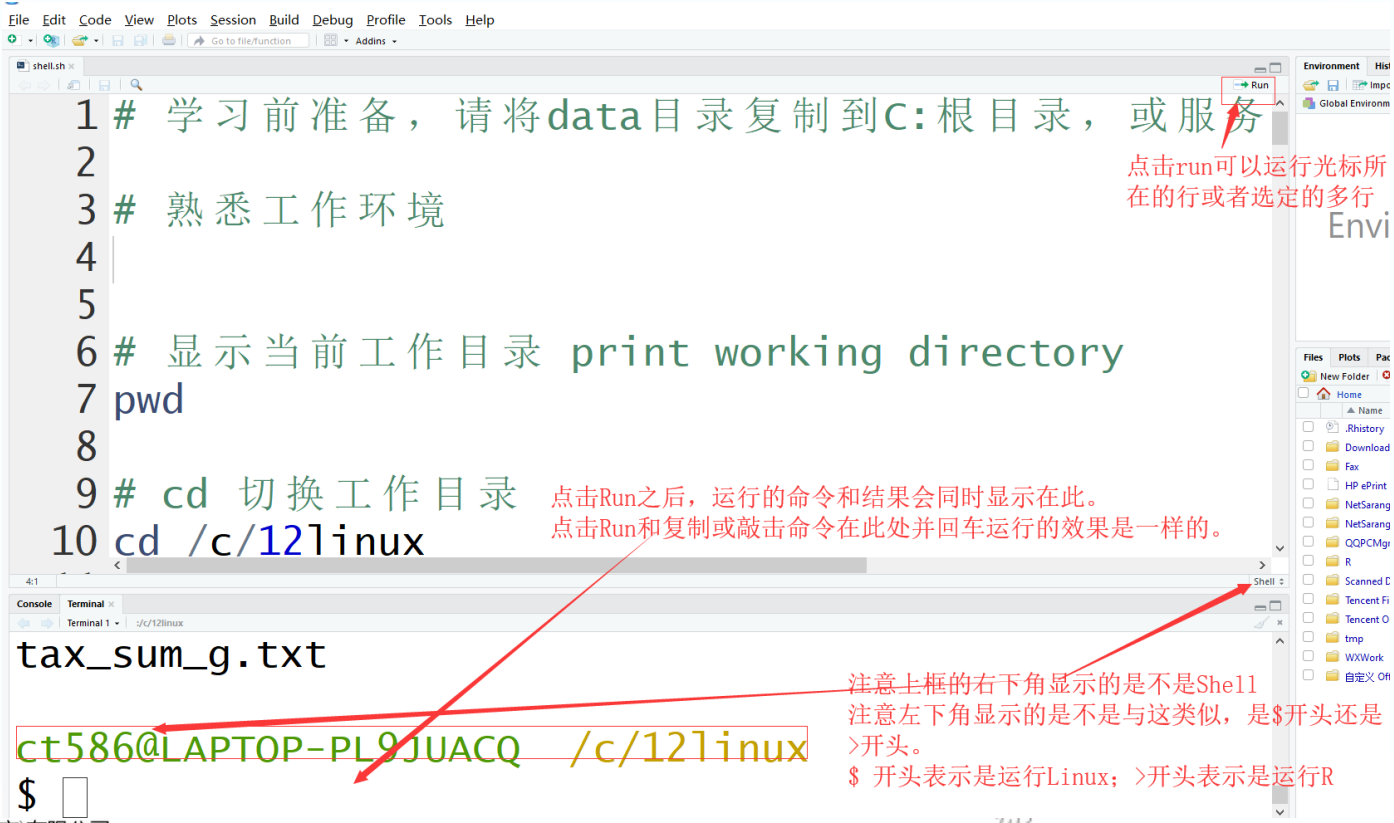
实际操作 – 打开文件和乱码解决

- 拷贝U盘中transcriptome文件夹到C盘根目录下 （任意一个盘不含中文和空格的目录都可以，为了统一都放在C盘下）。
- File – Open file – 打开pipelineStar.sh；若出现乱码，按图示操作。





实际操作 – 注意下面几个关键点



实际操作 – 文件路径

```
ct586@LAPTOP-PL9JUACQ   /c/12linux   1
$ cd test

ct586@LAPTOP-PL9JUACQ   /c/12linux/test   2
$ ls # 查看文件夹内容
test.txt

ct586@LAPTOP-PL9JUACQ   /c/12linux/test
$ # 切换至上级目录

ct586@LAPTOP-PL9JUACQ   /c/12linux/test
$ cd ..

ct586@LAPTOP-PL9JUACQ   /c/12linux   3
$
```

注意目录的变化

开始接触命令行，一个难跨越的概念是文件路径。Linux系统虽然好用，但还没智能到只给文件名就能判断路径的地步，实际也没必要，而且会引发危险。在Windows下访问文件时，会一层层打开文件夹去查看，Linux下也类似，只不过用cd代替了打开操作。如果碰到文件找不到的错误，一定先看下当前所在目录和文件所在目录。

参考资料

- https://bioconductor.org/packages/devel/bioc/vignettes/tximport/inst/doc/tximport.html
- https://f1000research.com/articles/4-1521/v1
- https://f1000research.com/articles/7-952/v1

Sequencing costs a lot and gains more

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