# 表达差异 - [云]单基因差异分析

| id | logFC | logCPM | F | PValue | padj | gene_name | gene_type |
|---|---|---|---|---|---|---|---|
| ENSG00000168878.19 | 6.5419 | 3.8399 | 46.967 | 1.13e-09 | 6.99e-06 | SFTPB | protein_coding |
| ENSG00000242512.9 | 5.2734 | 5.6526 | 37.77 | 2.57e-08 | 3.54e-05 | LINC01206 | lncRNA |
| ENSG00000233441.6 | 4.5874 | 1.1228 | 35.827 | 5.13e-08 | 5.73e-05 | CYP2AB1P | transcribed_unitary_pseudogene |
| ENSG00000205426.10 | 4.5707 | 3.5171 | 42.495 | 5e-09 | 1.21e-05 | KRT81 | protein_coding |
| ENSG00000185873.8 | 4.5495 | 5.4031 | 31.661 | 2.36e-07 | 0.0001 | TMPRSS11B | protein_coding |
| ENSG00000143536.7 | 4.5377 | 8.4663 | 20.741 | 1.76e-05 | 0.0017 | CRNN | protein_coding |
| ENSG00000204544.5 | 4.5189 | 7.5103 | 20.559 | 1.9e-05 | 0.0018 | MUC21 | protein_coding |
| ENSG00000117983.17 | 4.4415 | 6.4661 | 22.542 | 8.37e-06 | 0.0010 | MUC5B | protein_coding |
| ENSG00000215853.3 | 4.39 | 4.7316 | 30.008 | 4.4e-07 | 0.0002 | RPTN | protein_coding |
| ENSG00000116990.11 | 4.3788 | 7.9519 | 48 | 8.05e-10 | 6.99e-06 | MYCL | protein_coding |
| ENSG00000184956.16 | 4.2205 | 4.7055 | 21.824 | 1.12e-05 | 0.0012 | MUC6 | protein_coding |
| ENSG00000108688.11 | -2.0577 | -1.1377 | 19.882 | 2.53e-05 | 0.0020 | CCL7 | protein_coding |

网址：https://www.xiantao.love

更新时间：2023.03.03

目录

# 基本概念

- 差异分析：通过 R（以及常用 R 包）或者其他手段分析和筛选出表达谱中两组样本间的差异表达分子。
- 单基因差异分析：通过在同一个疾病状态中，对大量样本按照某个基因的表达值分成高低表达组（模拟类似过表达或者敲低这个基因的效果），分析两组中的差异分子，以推断与这个分子可能关联的分子。
- 常用 R 包介绍：
  - DESeq2 包：支持测序数据的 Counts 格式，也支持其他高通量数据的分析。
  - edgeR 包：支持测序数据的 Counts 格式，也支持其他高通量数据的分析。
    - https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf （edgeR 包帮助文档 16 页）



# 应用场景

单基因差异分析，可以分析数据集按照某个分子分成高低表达组后，两组的差异表达情况，从而拿到与这个分子关联的差异表达列表。有了这个差异表达列表，后续可以进行 GO|KEGG 富集分析（差异分子） 或者 GSEA 富集分析（所有分子和对应的 logFC），进一步推断这个分子可能涉及的功能或者通路等。

- 本模块是基于公共数据（云端数据）直接进行单基因差异分析。
- 功能分析

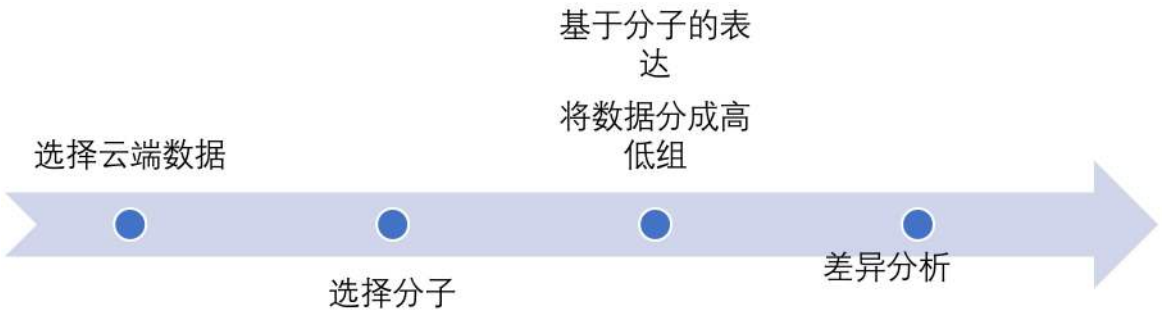
- 很多情况下当一个分子研究较少时，相应的这个分子功能注释会缺少，通过单基因差异联合富集分析，可对这个分子可能涉及的功能和通路有一点的预测作用！
- 单基因 GSEA 就是联合了单基因差异分析+GSEA，先进行单基因差异分析，拿到差异表达基因列表后，再进行 GSEA 分析，得到的结果就是单基因 GSEA 分析！

➢ 注意事项

- 对比单基因相关性筛选，单基因差异分析是将一个连续变量转成了二分类变量，从数值转成二分类是会缺失一些信息，所以有可能会出现单基因相关性筛选中相关性很显著的分子，在单基因差异分析中不显著的情况！
- 单基因差异分析，一般是以低表达组作为参考组，高表达组作为实验组，如果需要调整，可以使用【分组】中的参数修改。



## 分析流程



# 主要结果

基于 DEseq2 流程的主要分析结果

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | id | baseMean | log2FoldCha | lfcSE | stat | pvalue | padj | gene_name | gene_type |
| 2 | ENSG00000000003.15 | 3144.78815 | 0.25974749 | 0.16035866 | 1.61979086 | 0.10527721 | 0.36431377 | TSPAN6 | protein_coding |
| 3 | ENSG00000000005.6 | 1.08853791 | -0.86295651 | 0.51163295 | -1.68667112 | 0.09166661 | | TNMD | protein_coding |
| 4 | ENSG00000000419.13 | 3344.47211 | -0.19976542 | 0.09744258 | -2.05008336 | 0.0403563 | 0.21330769 | DPM1 | protein_coding |
| 5 | ENSG00000000457.14 | 1028.632 | 0.22881476 | 0.08574364 | 2.66859172 | 0.007617 | 0.0785335 | SCYL3 | protein_coding |
| 6 | ENSG00000000460.17 | 882.215575 | 0.16671214 | 0.1287338 | 1.29501448 | 0.19531523 | 0.49867737 | C1orf112 | protein_coding |
| 7 | ENSG00000000938.13 | 493.632802 | -0.38892734 | 0.19422114 | -2.00249749 | 0.04523125 | 0.22705704 | FGR | protein_coding |
| 8 | ENSG00000000971.16 | 4710.70388 | -0.57697168 | 0.23271838 | -2.47926994 | 0.01316516 | 0.1114116 | CFH | protein_coding |
| 9 | ENSG00000001036.14 | 2295.79063 | -0.42766767 | 0.14625002 | -2.92422288 | 0.00345317 | 0.04784439 | FUCA2 | protein_coding |
| 10 | ENSG00000001084.13 | 11106.6683 | -0.10919013 | 0.28404077 | -0.38441709 | 0.70066933 | 0.88369163 | GCLC | protein_coding |
| 11 | ENSG00000001167.14 | 2396.03665 | 0.33575668 | 0.11617113 | 2.89019039 | 0.00385009 | 0.05132044 | NFYA | protein_coding |
| 12 | ENSG00000001460.18 | 1088.6755 | -0.04996663 | 0.15196933 | -0.32879419 | 0.74231125 | 0.90186401 | STPG1 | protein_coding |
| 13 | ENSG00000001461.17 | 2877.03598 | 0.07022648 | 0.11503763 | 0.61046534 | 0.54155359 | 0.79683699 | NIPAL3 | protein_coding |

- id: ensembl 库注释的分子 ID。
- baseMean: 校正后的测序的 read count 的均值。（如果是挑分子，建议也要关注这部分的结果，可以用于判断这个基因的表达情况，如果很低就不建议选择了）
- $log_2$FoldChange: 差异倍数 FoldChange 值 log2 转化，当 $log_2$FoldChange=1 时，即说明有 2 倍的差异。（筛选差异的条件之一）
- lfcSE: $log_2$FoldChange 估计的标准误。
- stat: 统计量，可以不用理解。
- pvalue：统计检验的 p 值。
- padj: 统计检验校正后的 p 值。（筛选差异的条件之一）
- gene_name: ensembl 库注释的分子名
- gene_type: ensembl 库注释的分子类型，其中包括了编码基因、lncRNA、miRNA 以及其他类型的分子

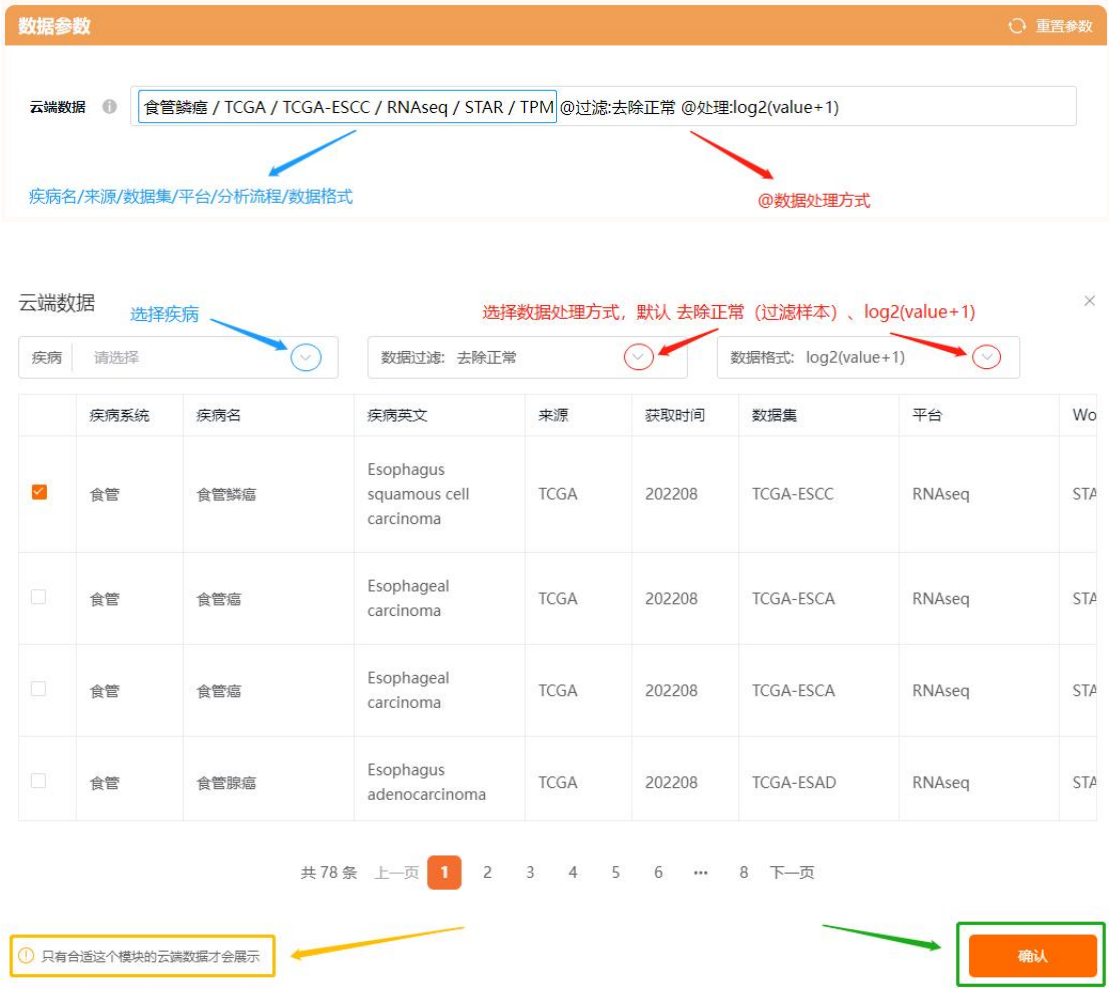
基于 edgeR 流程的主要分析结果

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | id | logFC | logCPM | F | PValue | padj | gene_name | gene_type |
| 2 | ENSG00000000003.15 | 0.25253268 | 5.4099445 | 2.44403829 | 0.12173304 | 0.36469937 | TSPAN6 | protein_coding |
| 3 | ENSG00000000419.13 | -0.20865732 | 5.49994005 | 4.39333066 | 0.03908998 | 0.19394804 | DPM1 | protein_coding |
| 4 | ENSG00000000457.14 | 0.2205781 | 3.79719228 | 6.13795893 | 0.01523536 | 0.11386135 | SCYL3 | protein_coding |
| 5 | ENSG00000000460.17 | 0.15866394 | 3.57780615 | 1.46243221 | 0.22993721 | 0.51108144 | C1orf112 | protein_coding |
| 6 | ENSG00000000938.13 | -0.39274278 | 2.73736753 | 3.95804046 | 0.04990276 | 0.22200316 | FGR | protein_coding |
| 7 | ENSG00000000971.16 | -0.58340359 | 5.99097751 | 5.95834793 | 0.01674901 | 0.12084227 | CFH | protein_coding |
| 8 | ENSG00000001036.14 | -0.44040224 | 4.95853937 | 8.84263992 | 0.00384017 | 0.05014863 | FUCA2 | protein_coding |
| 9 | ENSG00000001084.13 | -0.11420233 | 7.23064556 | 0.14730539 | 0.70209388 | 0.86952867 | GCLC | protein_coding |
| 10 | ENSG00000001167.14 | 0.32057967 | 5.01358941 | 7.73279928 | 0.00669276 | 0.06991712 | NFYA | protein_coding |
| 11 | ENSG00000001460.18 | -0.05933094 | 3.88379065 | 0.14621465 | 0.7031454 | 0.87010342 | STPG1 | protein_coding |
| 12 | ENSG00000001461.17 | 0.06005712 | 5.28004059 | 0.27243344 | 0.60307812 | 0.81469888 | NIPAL3 | protein_coding |
| 13 | ENSG00000001497.18 | -0.15948465 | 5.50575738 | 2.01924912 | 0.15901873 | 0.42113061 | LAS1L | protein_coding |

- id: ensembl 库注释的分子 ID。
- logFC: 差异倍数 FoldChange 值 log2 转化，当 $log_2$FoldChange=1 时，即说明有 2 倍的差异。（筛选差异的条件之一）
- logCPM: 标度转换，CPM（counts per million）是将 counts 转变为 CPM 指数，logCPM 是将 CPM 值 log2 转化。
- F: 检验统计量，可以不用理解。
- PValue：统计检验的 p 值。
- padj: 统计检验校正后的 p 值。（筛选差异的条件之一）
- gene_name: ensembl 库注释的分子名
- gene_type: ensembl 库注释的分子类型，其中包括了编码基因、lncRNA、miRNA 以及其他类型的分子

空值的原因可能是因为该分子在分组间表达不显著导致的无法计算一些值。

# 云端数据




本模块提供预清洗好的云端数据，不同平台的云端数据集的分子可能会有不同。注意查看当前数据参数选中的云端数据。

这里为任务式模块，提交任务后需要到**历史记录**中刷新并等待任务完成，（分析时间大概在几分钟到十几分钟不等，具体要看对应的数据集的样本量，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）

# 参数说明

（说明：标注了颜色的为常用参数。）

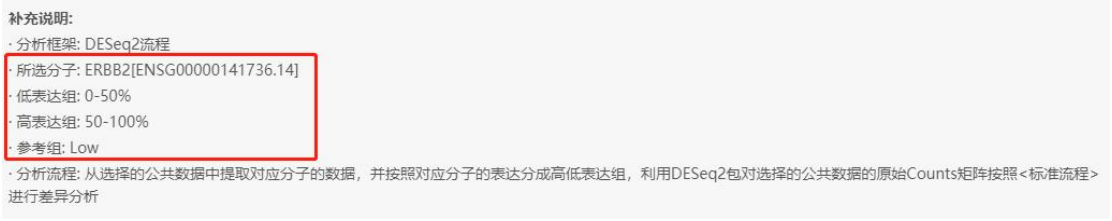
# 特殊参数


- 特殊参数：下拉框将列出对应所选数据集分子，可以输入关键字搜索分子，基因 symbol 或 Ensembl ID，只能选单个分析。

# 分组


- 参考组： 默认对数据按照对应分子的表达进行排序，取前%作为低表达组，后%-100%作为高表达组。根据具体情况可以自由选择参考组的分组。
- 低表达组(0~?%)：按照对应分子表达从低到高进行排序，取前？%作为低表达组，这里就是输入这个前百分之几。默认是 50%，即中位数。可以适当缩小，可能会使得高低表达组之间的差异更加明显！输入的值需要比 0 大并且不能大于高表达设定的值。
- 高表达组(?%~100%)：按照对应分子表达从低到高进行排序，取后%-100%作为高表达组，这里就是输入这个后百分之几。默认是 50%，即中位数。可以适当缩小，可能会使得高低表达组之间的差异更加明显 !输入的值需要比 100 小并且不能小于低表达组设定的值。

参数使用情况：



## 分析参数


- 流程：可以选择 DESeq2 流程、edgeR 流程。
  - edgeR 流程
    - 利用 edgeR 包对原始 Counts 矩阵进行差异分析，按照标准流程对表达丰度低的分子进行过滤，并且用 edgeR 包提供的 logCPM(Counts Per Million)方法对原始 Counts 矩阵进行标准化处理(Normalize)。
  - DESeq2 流程
    - 利用 DESeq2 包对原始 Counts 矩阵进行差异分析，按照标准流程进行分析，并且用 DESeq2 包提供的 VST(Variance Stabilizing Transformations)方法对原始 Counts 矩阵进行标准化处理(Normalize)。

# 结果说明

## 主要结果

**单基因差异分析**: 基于单个分子的表达分成高低表达组进行差异分析, 当前参考组为: Low
分析流程: DESeq2流程
页面中仅仅展示高表达(logFC为正)以及低表达(logFC为负)各30个的结果，更多的结果需要下载差异分析表格

| id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | gene_name | gene_type |
|---|---|---|---|---|---|---|---|---|
| ENSG00000264564.1 | 25.015 | 5.1554 | 1.0276 | 5.017 | 5.25e-07 | 8.47e-05 | AC090897.1 | processed_pseudo |
| ENSG00000185873.8 | 3156.7 | 4.5672 | 0.6246 | 7.3122 | 2.63e-13 | 8.01e-10 | TMPRSS11B | protein_codin |
| ENSG00000143536.7 | 2.628e+04 | 4.5441 | 0.75037 | 6.0559 | 1.4e-09 | 7.85e-07 | CRNN | protein_codin |
| ENSG00000204544.5 | 1.359e+04 | 4.5289 | 0.77141 | 5.8709 | 4.34e-09 | 1.96e-06 | MUC21 | protein_codin |
| ENSG00000096006.12 | 1208.1 | 4.4842 | 0.71425 | 6.2782 | 3.43e-10 | 2.85e-07 | CRISP3 | protein_codin |
| ENSG00000215853.3 | 1939.7 | 4.3844 | 0.62514 | 7.0134 | 2.33e-12 | 4.05e-09 | RPTN | protein_codin |
| ENSG00000258616.5 | 8.3333 | 4.3193 | 0.85371 | 5.0595 | 4.2e-07 | 7.05e-05 | LINC02303 | lncRNA |
| ENSG00000196260.5 | 74.733 | 4.3139 | 0.56467 | 7.6396 | 2.18e-14 | 1.59e-10 | SFTA2 | protein_codin |
| ENSG00000267428.2 | 2.8004 | 4.2495 | 0.8897 | 4.7764 | 1.78e-06 | 0.0002 | AC091135.1 | unprocessed_pseud |
| ENSG00000198488.10 | 191.88 | 4.2238 | 0.50473 | 8.3685 | 5.84e-17 | 1.07e-12 | B3GNT6 | protein_codin |
| ENSG00000203688.6 | 536.79 | 4.2115 | 0.5673 | 7.4237 | 1.14e-13 | 5.27e-10 | LINC02487 | lncRNA |
| ENSG00000145879.11 | 1565.3 | 4.1366 | 0.59173 | 6.9906 | 2.74e-12 | 4.55e-09 | SPINK7 | protein_codin |

单基因差异分析.xlsx

此表格提供单基因-差异分析结果（页面只展示 60 个，即高表达(logFC 为正)以及低表达(logFC 为负)各 30 个的结果），提供 EXCEL 格式下载。

## 补充结果

样本信息

差异分析参考组: Low

| 组别 | 数量 |
|---|---|
| Low | 41 |
| High | 41 |

样本信息.xlsx

此表格提供进行差异分析的样本信息，包括分组情况、对应分组内样本数量和参考组信息。

差异统计

差异分析后一些常见阈值(|logFC|大于2或者1或者是0.58(0.58换算过来就是1.5倍))下的差异分子数量, 也可以根据需要下载差异分析结果用excel表进行过滤

| 筛选条件 | 筛选后的数量 |
| --- | --- |
| \|LogFC\|>2 & p.adj<0.05 | 250 |
| \|LogFC\|>1 & p.adj<0.05 | 1248 |
| \|LogFC\|>0.58 & p.adj<0.05 | 2191 |

此表格提供在差异分析结果中，一些常见阈值的差异分子数量统计。可以根据需要下载差异分析结果后用 excel 表进行过滤。

这里为任务式模块，提交任务后需要到**历史记录**中刷新并等待任务完成，（分析时间大概在几分钟到十几分钟不等，具体要看对应的数据集的样本量，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）任务完成后，提供 Excel 及完整报告下载。

# 方法学

所有分析和可视化均在 R 4.2.1 中进行

涉及的 R 包：DESeq2、edgeR

处理过程：

（1） 从选择的公共数据中提取对应分子的数据，并按照对应分子的表达分成高低表达组，利用 DESeq2/edgeR 包对选择的公共数据的原始 Counts 矩阵按照<标准流程>进行差异分析。

a) edgeR 流程

i. 利用 edgeR 包对原始 Counts 矩阵进行差异分析，按照标准流程对表达丰度低的分子进行过滤，并且用 edgeR 包提供的 logCPM(Counts Per Million)方法对原始 Counts 矩阵进行标准化处理(Normalize)。

b) DESeq2 流程

i. 利用 DESeq2 包对原始 Counts 矩阵进行差异分析，按照标准流程进行分析，并且用 DESeq2 包提供的 VST(Variance Stabilizing Transformations)方法对原始 Counts 矩阵进行标准化处理(Normalize)。

# 如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

# 常见问题

**1. 单基因差异分析 与 单基因相关性筛选有什么区别？分别有什么用？**

答：

- 区别

  单基因差异分析 是按某个基因的表达值分成高低表达组，分析两组的差异表达的分子；

  单基因相关性筛选 是直接把某个基因的表达值 和 其他所有的分子 两两进行相关性分析；

  对比单基因相关性筛选，单基因差异分析是将一个连续变量转成了二分类变量，从数值转成二分类是会缺失一些信息，所以有可能会出现单基因相关性筛选中相关性很显著的分子，在单基因差异分析中不显著的情况！。

- 用途

  单基因差异分析 拿到结果后，可以进行富集分析以及后续差异分析的下游流程，如果进行的是 GSEA，则是单基因 GSEA 富集分析；

  单基因相关性筛选 只是提供了一个相关性筛选的结果，可以更加有目的的选分子进行相关性分析的可视化（散点图、单基因共表达热图）等。

**2. 单基因差异分析的流程是什么样的？为什么测序平台还有 TPM 和 FPKM 之分？**

答：

分组排序的数据是在 TPM 或者 FPKM 中进行的，分完组（高低表达）后再对测序平台的 read counts 格式的数据利用 DESeq2 包进行两组的差异分析 !所以用 TPM 或者 FPKM 均会影响分组排序的情况，影响两组的样本，最终影响差异分析的结果。

**3. 为什么不用 counts 直接进行排序，而要用 TPM 或者 FPKM？**

答：

counts 由于受到测序深度、基因组长度比对的影响，不同的基因差别很大，即便是同一个基因在不同的样本中差别也可能会很大，一般需要先进行校正后才能进行比较。而 TPM 和 FPKM 都在一定程度上考虑到这个，所以可以用 FPKM 和 TPM 来进行比较排序！

**4. 在云端数据框内看到的例数和分析时候的例数不同，这个是什么情况？**

答：

云端数据的例数一般是对应组学所有的例数，单基因差异分析是会去除正常样本的，也就是只针对肿瘤样本。具体需要看说明文本中对于数据的处理情况的说明。