# 表达差异 - 差异分析-Limma

| id | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|---|
| WP_CELLULAR_PROTEOSTASIS | 1.006 | -0.024089 | 7.8382 | 2.4e-09 | 5.929120e-06 | 11.328 |
| WP_REGULATION_OF_SISTER_CHROM... | 0.93748 | -0.016468 | 7.1144 | 2.12e-08 | 1.603647e-05 | 9.2544 |
| BIOCARTA_SM_PATHWAY | 0.93004 | -0.035681 | 6.7568 | 6.31e-08 | 1.603647e-05 | 8.2116 |
| REACTOME_E2F_ENABLED_INHIBITION... | 0.92952 | -0.048675 | 6.7197 | 7.07e-08 | 1.603647e-05 | 8.1029 |
| REACTOME_CDC6_ASSOCIATION_WIT... | 0.9285 | -0.035593 | 6.7081 | 7.33e-08 | 1.603647e-05 | 8.0688 |
| REACTOME_SUMO_IS_CONJUGATED_T... | 0.92597 | -0.039604 | 6.2566 | 2.95e-07 | 2.309654e-05 | 6.7378 |
| REACTOME_UNWINDING_OF_DNA | 0.92489 | 0.0059298 | 6.6053 | 1e-07 | 1.612337e-05 | 7.7667 |
| REACTOME_PROCESSIVE_SYNTHESIS_... | 0.91728 | -0.010868 | 7.2223 | 1.53e-08 | 1.550306e-05 | 9.5669 |
| REACTOME_DNA_STRAND_ELONGATI... | 0.90144 | -0.015447 | 7.0176 | 2.84e-08 | 1.603647e-05 | 8.9731 |
| REACTOME_SLBP_DEPENDENT_PROCE... | 0.89603 | -0.0078671 | 6.2981 | 2.59e-07 | 2.303074e-05 | 6.8606 |
| REACTOME_APOPTOSIS_INDUCED_DN... | 0.89457 | -0.050595 | 7.6763 | 3.89e-09 | 5.929120e-06 | 10.869 |
| REACTOME_ACTIVATION_OF_THE_PRE_... | 0.88842 | -0.033117 | 6.6901 | 7.74e-08 | 1.603647e-05 | 8.0159 |
| BIOCARTA_BARD1_PATHWAY | 0.88785 | -0.034045 | 6.492 | 1.42e-07 | 1.737610e-05 | 7.433 |

网址：https://www.xiantao.love

更新时间：2023.06.14

目录

# 基本概念

- 差异分析：通过 limma 包实现（两组）组间比较的差异分析，筛选出数据矩阵中两组样本间的差异表达分子。
  - limma 包：支持微阵列芯片、转录组表达谱（如 FPKM/TPM ）、GSVA 富集分析结果等。
    - https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf （limma 包分析手册）

# 应用场景

主要是对芯片或高通量测序数据进行差异分析，旨在找出不同样本（组）间存在差异表达的变量，得到差异表达变量之后，可以进行相应的可视化分析。本模块适用于除转录组原始 counts 以外的数据。

同时支持 GSVA 输出的结果进行差异分析，得到对应差异的 t 值，再用双向柱状图模块进行可视化

# 主要结果

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | id | logFC | AveExpr | t | P.Value | adj.P.Val | B |
| 2 | WP_CELLULAR_PROT | 1.006045966 | -0.024088756 | 7.838179993 | 2.40111E-09 | 5.92912E-06 | 11.32778634 |
| 3 | REACTOME_METALLC | -0.993168931 | -0.014324441 | -6.412829525 | 1.81869E-07 | 1.91213E-05 | 7.199618145 |
| 4 | REACTOME_RESPONS | -0.941020945 | -0.002044488 | -6.621196292 | 9.56918E-08 | 1.61234E-05 | 7.813487992 |
| 5 | WP_REGULATION_OI | 0.937479224 | -0.016468378 | 7.11436755 | 2.11672E-08 | 1.60365E-05 | 9.254373402 |
| 6 | BIOCARTA_SM_PATH | 0.93004496 | -0.035680981 | 6.756813903 | 6.3089E-08 | 1.60365E-05 | 8.211571236 |
| 7 | REACTOME_E2F_ENA | 0.929524336 | -0.048675334 | 6.719739717 | 7.06906E-08 | 1.60365E-05 | 8.102872052 |
| 8 | REACTOME_CDC6_AS | 0.928504963 | -0.035592553 | 6.708141335 | 7.32532E-08 | 1.60365E-05 | 8.068846364 |
| 9 | REACTOME_SUMO_IS | 0.925970114 | -0.039604237 | 6.256583205 | 2.94774E-07 | 2.30965E-05 | 6.737832977 |
| 10 | REACTOME_UNWIND | 0.924888263 | 0.005929772 | 6.605292631 | 1.0049E-07 | 1.61234E-05 | 7.766725085 |
| 11 | REACTOME_PROCESS | 0.917283144 | -0.010867972 | 7.222285109 | 1.52539E-08 | 1.55031E-05 | 9.566929507 |
| 12 | REACTOME_DNA_STI | 0.901439998 | -0.015446567 | 7.017550601 | 2.84229E-08 | 1.60365E-05 | 8.973062608 |
| 13 | REACTOME_SLBP_DE | 0.896031119 | -0.007867053 | 6.298094406 | 2.59254E-07 | 2.30307E-05 | 6.860623684 |

- id：对应上传数据中第一列的 id，示例数据为通路名称。
- logFC：组间差异值，limma 包内置的计算方式为 case 组数据均值减去 control 组数据均值。（由于 limma 包是针对芯片数据开发，默认分析的数据是经过 log2 处理的，因此在使用 limma 的函数计算时，如果输入的矩阵没有经过 log2 处理，则会把 FC 当成 log2FC 输入，比较组间相减以得到 log2FC，数值很大时会导致计算出来的差异表达高达几百甚至上千。）（筛选差异的条件之一）
- AveExpr：所有样本的均值。
- t：t 统计量，即 LogFC 除以它的标准差（未标度的标准差/后验方差的开方）。（筛选差异的条件之一）
- P.Value：统计检验的 p 值，表示是否显示差异。
- adj.P.Val：统计检验校正后的 p 值。（筛选差异的条件之一）
- B：对数概率值，可以不理解

空值的原因可能是因为该分子在样本间表达不显著或者空值(缺失)较多导致的无法计算一些值。

# 数据格式

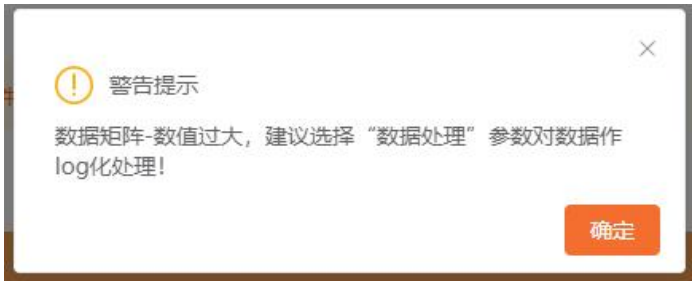
**数据矩阵**

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | IDs | TCGA-IC- | TCGA-IC- | TCGA-L5- | TCGA-L5- | TCGA-L5- | TCGA-L5- | TCGA-L5- | TCGA-L5- |
| 2 | BIOCARTA_41BB_PATHWAY | 0.211852 | -0.01391 | -0.35759 | -0.1707 | -0.30799 | -0.53615 | -0.28159 | 0.160148 |
| 3 | BIOCARTA_ACE2_PATHWAY | -0.09039 | 0.091342 | 0.136498 | -0.0504 | 0.237997 | 0.315002 | 0.467162 | 0.592167 |
| 4 | BIOCARTA_ACETAMINOPHEN | 0.555595 | 0.636198 | 0.469264 | -0.35334 | -0.11738 | 0.039783 | 0.59599 | 0.207359 |
| 5 | BIOCARTA_ACH_PATHWAY | -0.21212 | 0.187524 | -0.14525 | -0.25659 | 0.426841 | -0.20685 | 0.173573 | 0.262117 |
| 6 | BIOCARTA_ACTINY_PATHWAY | 0.505618 | 0.548071 | -0.70908 | -0.38222 | -0.52349 | -0.6224 | -0.15485 | 0.681342 |
| 7 | BIOCARTA_AGPCR_PATHWAY | -0.45248 | -0.40712 | 0.018402 | -0.19835 | 0.128957 | 0.019715 | 0.653225 | 0.50157 |
| 8 | BIOCARTA_AGR_PATHWAY | 0.056678 | 0.208978 | -0.30969 | 0.261647 | 0.030904 | -0.01824 | 0.160053 | 0.538372 |
| 9 | BIOCARTA_AHSP_PATHWAY | 0.040799 | 0.402187 | 0.141087 | -0.13419 | 0.300914 | 0.354476 | 0.177173 | 0.56987 |
| 10 | BIOCARTA_AKAP13_PATHWAY | 0.218686 | 0.119437 | -0.17333 | 0.063406 | 0.279936 | -0.18817 | 0.304189 | 0.15315 |
| 11 | BIOCARTA_AKAP95_PATHWAY | -0.1643 | 0.07299 | -0.49215 | -0.15275 | -0.38566 | -0.37139 | -0.1667 | -0.17683 |
| 12 | BIOCARTA_AKAPCENTROSOM | 0.113422 | -0.09741 | -0.43967 | 0.091486 | -0.26725 | -0.44256 | 0.215399 | 0.210213 |
| 13 | BIOCARTA_AKT_PATHWAY | 0.238612 | 0.184665 | -0.46528 | -0.14311 | 0.090153 | -0.4151 | 0.363489 | 0.584714 |

数据要求：

- 第一行为样本编号，第一列为分子，不能含有缺失、重复及特殊字符。第一列的 ID 不需要是 ENSG，任何都可以，但是不能含有缺失、特殊字符或重复 ID，重复的 ID 会在分析过程中过滤掉，保留第一个出现的对应数据。
- 数据至少有 5 列以上，至少需要 2 行数据。
- 第一列以后的数值部分为不同分子在各样本中的数值，可以是 GSVA 富集得分矩阵、微阵列芯片表达谱、转录组 FPKM/TPM 等数据。
- 若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。
- 最多支持 600 列，70000 行。文件不能大于 100M（文件过大有可能会上传失败，如果上传 1 分钟后没有反应，可能是上传失败了）

注意：limma 包是针对微阵列芯片数据开发，通常使用该包处理时，需要经过 log2 处理后的矩阵作为输入。但并非所有的数据都已经经过 log 处理，对于没有 log 处理，如转录组测序标准化数据 FPKM/TPM/RPKM 等，会建议通过参数【数据处理】进行 log2 处理，也可以自行先做处理后上传数据进行分析。



**样本信息**

| | A | B |
|---|---|---|
| 1 | sample | group |
| 2 | TCGA-IC-A6RE-11A-12R-A336-31 | Normal |
| 3 | TCGA-IC-A6RF-11A-21R-A336-31 | Normal |
| 4 | TCGA-L5-A43C-11A-11R-A24K-31 | Normal |
| 5 | TCGA-L5-A4OF-11A-12R-A260-31 | Normal |
| 6 | TCGA-2H-A9GL-01A-12R-A37I-31 | Tumor |
| 7 | TCGA-2H-A9GM-01A-11R-A37I-31 | Tumor |
| 8 | TCGA-2H-A9GN-01A-11R-A37I-31 | Tumor |
| 9 | TCGA-2H-A9GO-01A-11R-A37I-31 | Tumor |

数据要求：

- 第一列为样本编号，请与数据矩阵第一行样本保持一致。
- 第二列为对应的分组信息，第一行为样本分组信息，需要提供两个分组，每个分组至少含有 2 个样本以上的生物学重复。（重复不是指复制样本，是平行实验样本）。
  - 注意：上传数据后会自动返回分组信息到【参考组选择】参数中。

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。）

## 分组


- 参考组选择 ： 上传【样本信息】数据后会自动读取对应的内容作为分组信息，可以自由选择参考组的分组名。

## 数据处理

- 转换：可以对数据进行 log2 转换，默认“无”，即不进行处理。（注意：数据中含有小于等于 0 的数值时，将会被转换为空值，此时的空值过多可能会影响后续分析，请按需选择数据转换模式。）

## 分析参数


- p 值校正方法：即 padj，多重检验校正后的 p 值。可选择 BH、holm、hochberg、hommel、bonferroni、BY，其中 BH 方法又称 fdr。

# 结果说明

## 主要结果

**差异分析-limma**: 基于limma包标准差异分析流程, 对数据进行两组差异分析
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 | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|---|
| WP_CELLULAR_PROTEOSTASIS | 1.006 | -0.024089 | 7.8382 | 2.4e-09 | 5.929120e-06 | 11.328 |
| WP_REGULATION_OF_SISTER_CHROM... | 0.93748 | -0.016468 | 7.1144 | 2.12e-08 | 1.603647e-05 | 9.2544 |
| BIOCARTA_SM_PATHWAY | 0.93004 | -0.035681 | 6.7568 | 6.31e-08 | 1.603647e-05 | 8.2116 |
| REACTOME_E2F_ENABLED_INHIBITION... | 0.92952 | -0.048675 | 6.7197 | 7.07e-08 | 1.603647e-05 | 8.1029 |
| REACTOME_CDC6_ASSOCIATION_WIT... | 0.9285 | -0.035593 | 6.7081 | 7.33e-08 | 1.603647e-05 | 8.0688 |
| REACTOME_SUMO_IS_CONJUGATED_T... | 0.92597 | -0.039604 | 6.2566 | 2.95e-07 | 2.309654e-05 | 6.7378 |
| REACTOME_UNWINDING_OF_DNA | 0.92489 | 0.0059298 | 6.6053 | 1e-07 | 1.612337e-05 | 7.7667 |
| REACTOME_PROCESSIVE_SYNTHESIS_... | 0.91728 | -0.010868 | 7.2223 | 1.53e-08 | 1.550306e-05 | 9.5669 |
| REACTOME_DNA_STRAND_ELONGATI... | 0.90144 | -0.015447 | 7.0176 | 2.84e-08 | 1.603647e-05 | 8.9731 |
| REACTOME_SLBP_DEPENDENT_PROCE... | 0.89603 | -0.0078671 | 6.2981 | 2.59e-07 | 2.303074e-05 | 6.8606 |
| REACTOME_APOPTOSIS_INDUCED_DN... | 0.89457 | -0.050595 | 7.6763 | 3.89e-09 | 5.929120e-06 | 10.869 |
| REACTOME_ACTIVATION_OF_THE_PRE_... | 0.88842 | -0.033117 | 6.6901 | 7.74e-08 | 1.603647e-05 | 8.0159 |
| BIOCARTA_BARD1_PATHWAY | 0.88785 | -0.034045 | 6.492 | 1.42e-07 | 1.737610e-05 | 7.433 |

差异分析.xlsx

此表格为差异分析结果（页面只展示 60 个），提供 EXCEL 格式下载。

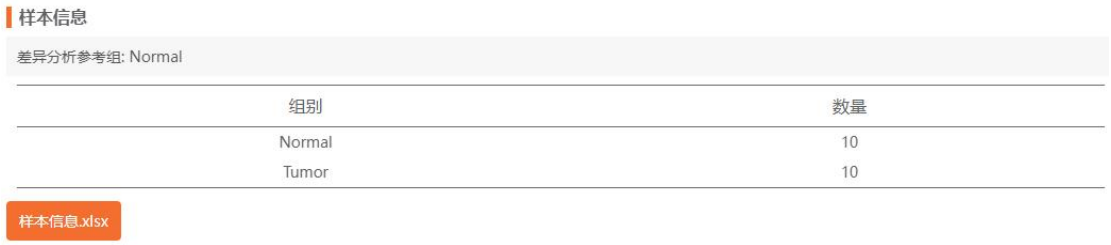
## 补充结果

**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 | 0 |
| \|LogFC\|>1 & p.adj<0.05 | 1 |
| \|LogFC\|>0.58 & p.adj<0.05 | 358 |

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用 excel 表进行过滤。



样本信息

差异分析参考组: Normal

| 组别 | 数量 |
| --- | --- |
| Normal | 10 |
| Tumor | 10 |

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。

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：limma

处理过程：

（1） 使用 limma 包，根据样本分组对数据进行标准组间差异分析。

# 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. 上传数据后为什么没有反应，点击确认总是跳出"没能识别数据"?**

答:

- 上传数据 1 分钟内，如果没有弹出下图"验证成功"，则说明可能失败了。
- 当文件过大时，有可能会上传或者数据验证失败。

**2. 如何挑选分子?**

答:

首先关注校正后的 p 值，再结合 log2FodlChange 或 t 值，看是否满足设定的阈值。由于 limma 差异分析中，logFC 为组间 log(均值) 相减，如果数据矩阵中分子的数值过大或过小（未经 log 处理），可能就会对 logFC 有很大的影响。

**3. 分析后的校正后 p 值很大，能否用 p 值替代? 校正后 p 值很大的原因是什么?**

答:

常见的文章里面是用校正后的 p 值,如果直接用 p 值,很可能会被审稿人提问,也相对不好回答,所以是不建议直接用 p 值。如果出现校正后 p 值都很大的情况，一般是样本差异很小或者样本过少导致的，建议是先做一个 PCA 图看看两个分组的样本差异情况。

**4. 为什么有一些分子在表格中对应的统计学值都是空的?**

答:

空值的原因可能是因为该分子在样本间表达不显著或者数据中缺失过多导致的无法计算一些值。

**5. logFC 的计算是不是不对呢?**

答:

关于 limma 包差异分析结果的 logFC 解释。

- 首先，limma 包接受的输入数据为表达矩阵，而且是 log（以 2 为底）化后的表达矩阵；也就是说 limma 包针对芯片数据进行分析，默认认为输入数据即为 log 化和标准化的数据。
- FoldChange 正常的理解是（组间）差异倍数，即如果 case 组的平均表达量是 8，control 组是 2，那么 FoldChange 就是 4。而 limma 包里面，计算方式是 case 组和 control 组的均值相减，输入的时候 case 组的平均表达量被 log 后是 3，control 是 1，那么差值是 2，就是说分析所获得的 logFC 就是 2，是 log 均值的差值为 2，不是差异倍数。