表达差异 - 芯片数据-差异分析

id	logFC	AveExpr	t	P.Value	adj.P.Val	B
226847_at	5.8481	6.6834	4.0192	0.0034	0.8739	-2.2638
225387_at	5.3013	4.5282	3.9593	0.0037	0.8739	-2.306
242277_at	5.0225	4.7807	3.9539	0.0037	0.8739	-2.3098
237056_at	4.9518	2.6626	9.5603	7.53e-06	0.1372	-0.28938
204948_s_at	4.8729	8.0047	4.4926	0.0017	0.7432	-1.9526
207345_at	4.8655	5.8567	4.1674	0.0027	0.8319	-2.1622
224579_at	4.7975	3.5842	2.619	0.0291	0.9559	-3.4104
209656_s_at	4.7663	4.4505	2.2042	0.0565	0.9559	-3.7985
226610_at	4.7662	4.1399	4.2	0.0026	0.8200	-2.1403
204049_s_at	4.7636	5.7103	3.7616	0.0049	0.8739	-2.4499
204036_at	4.6106	6.6124	3.179	0.0120	0.9339	-2.914
220658_s_at	4.5853	4.0069	5.5787	0.0004	0.6268	-1.3757

网址：https://www.xiantao.love

更新时间：2023.08.04

目录

基本概念 3
应用场景 3
主要结果 4
数据格式 5
参数说明 10
    分组 10
    数据处理 10
    分析参数 11
结果说明 12
    主要结果 12
    补充结果 12
    方法学 14
如何引用 15
常见问题 16

基本概念

- 差异分析：基于芯片表达谱数据，通过 limma 包实现（两组）组间比较的差异分析，筛选出表达谱数据中两组样本间的差异表达分子。
  - limma 包
    - 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。本模块适用于微阵列芯片数据。

注意：与【数据集检索】工具不同，本模块**无**探针处理和注释，以及可视化内容。

主要结果

	A	B	C	D	E	F	G
1	id	logFC	AveExpr	t	P.Value	adj.P.Val	B
2	226847_at	5.848144159	6.683367449	4.019242785	0.00336124	0.87386499	-2.26380868
3	209848_s_at	-5.618199888	10.53832772	-3.15220908	0.012478246	0.933945904	-2.936677779
4	225387_at	5.301315535	4.528223729	3.959326945	0.003668425	0.87386499	-2.306011269
5	206696_at	-5.095975319	8.030714694	-2.594687375	0.030219646	0.955905677	-3.432821193
6	242277_at	5.022548066	4.780740886	3.953924721	0.003697557	0.87386499	-2.309848007
7	1552367_a_at	-5.013454871	5.042817758	-3.423297229	0.00819769	0.90808425	-2.712180277
8	237056_at	4.951847854	2.662599625	9.56030012	7.53012E-06	0.137236452	-0.289383972
9	204948_s_at	4.872926451	8.004736925	4.492629981	0.001715205	0.743166955	-1.952641263
10	207345_at	4.865475245	5.856701257	4.167391456	0.002713584	0.831922242	-2.162206292
11	224579_at	4.797491486	3.584235574	2.618981671	0.029066166	0.955905677	-3.410418153
12	237070_at	-4.779365684	7.306657022	-2.371418707	0.043235137	0.955905677	-3.640930754
13	209656_s_at	4.766265828	4.450514263	2.204222293	0.056520197	0.955905677	-3.798452383

- id：对应上传数据中第一列的 id，示例数据为探针名称。
- 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

芯片表达谱

	A	B	C	D	E	F	G	H	I
1	id	GSM162929	GSM162932	GSM162933	GSM162935	GSM162904	GSM162905	GSM162928	GSM162930
2	1007_s_at	8.637327672	9.304647931	7.904002316	8.211679052	8.268939608	9.18151506	8.014578465	9.448388915
3	1053_at	9.056282755	9.469157446	8.93414198	8.691816786	8.83412021	9.066124181	8.676895009	8.523902988
4	117_at	5.762045786	5.709668356	5.326655986	5.980815914	5.132441121	6.768091672	6.095618553	5.119526335
5	121_at	7.81690045	7.523412974	7.355518483	7.369797979	7.11953256	7.26746089	7.289225838	7.542915781
6	1255_g_at	8.177524186	4.274462953	1.516282475	4.220956741	2.148579487	4.418081978	4.148527425	1.751562268
7	1294_at	6.555402057	6.946438456	6.692106328	6.185620791	6.89788175	5.505786153	5.55310549	7.319708244
8	1316_at	5.391186869	5.1505678	4.866636026	5.477236941	4.533681699	4.620662648	4.189571571	4.663088396
9	1320_at	4.834953869	1.942660879	3.738119289	4.916065752	4.94602899	3.944708548	2.825419311	3.967251625
10	1405_i_at	0.452342401	4.406210109	1.19839357	0.990461175	3.633152177	3.913013846	4.204117235	4.560268628
11	1431_at	3.686085868	3.272330812	3.119578624	3.883865155	3.031668675	3.56524377	1.233458623	4.339906997
12	1438_at	4.742070904	4.88945894	4.701332803	2.918907114	2.361860981	7.301615081	3.537793034	6.181857899
13	1487_at	8.441159753	8.445970806	7.92110304	8.183848731	8.969092916	7.660951428	7.196302123	7.25641629

数据要求：

- 第一行为样本编号，第一列为探针号，不能含有缺失、重复及特殊字符。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 100 行数据。
- 第一列以后的数值部分为不同探针在各样本中的表达量。
- 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 处理的情况，本模块会自动进行 log 化处理，也可以自行先做处理后上传数据进行分析。（注意，当数据中存在小于等于 0 时，不建议再进行 log 化处理，且这类型的数据过多，如单个样本中所有表达值都较小时，建议删除样本或者更换数据）

样本信息

	A	B
1	sample	group
2	GSM162929	control
3	GSM162932	control
4	GSM162933	control
5	GSM162935	control
6	GSM162904	case
7	GSM162905	case
8	GSM162928	case
9	GSM162930	case

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。
- 可以提供样本对应的配对信息或者批次信息，必须在前两列的基础上补充，且只允许使用"pair", "batch"作为列名，否则无法识别。
  - pair 列：不允许含有缺失和特殊字符；所有的分组（group）中配对数量和信息都要一致；所有的分组中含有的 pair 内容不能重复，并且要求每个 pair 内容在所有的分组都有唯一的记。

	A	B	C	D
1	sample	group	pair √	
2	GSM162929	control	1	
3	GSM162932	control	2	
4	GSM162933	control	3	正确信息
5	GSM162935	control	4	
6	GSM162904	case	1	
7	GSM162905	case	2	
8	GSM162928	case	3	
9	GSM162930	case	4	

	A	B	C	D
1	sample	group	pair ✗	
2	GSM162929	control		含有缺失
3	GSM162932	control	2	
4	GSM162933	control	3	
5	GSM162935	control	5	
6	GSM162904	case	1	
7	GSM162905	case	2	
8	GSM162928	case	3	
9	GSM162930	case		

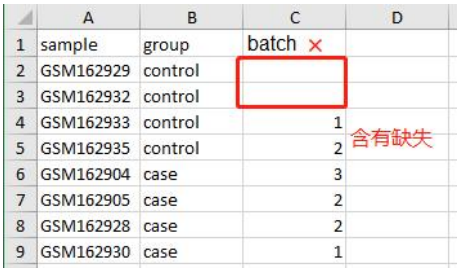
	A	B	C	D
1	sample	group	pair ✗	
2	GSM162932	control	2	
3	GSM162933	control	3	
4	GSM162935	control	4	group间数量不一致
5	GSM162904	case	1	
6	GSM162905	case	2	
7	GSM162928	case	3	
8	GSM162930	case	4	

	A	B	C	D
1	sample	group	pair ✗	
2	GSM162929	control	1	
3	GSM162932	control	2	
4	GSM162933	control	5	含有不同的配对信息内容
5	GSM162935	control	6	
6	GSM162904	case	1	
7	GSM162905	case	2	
8	GSM162928	case	3	
9	GSM162930	case	4	

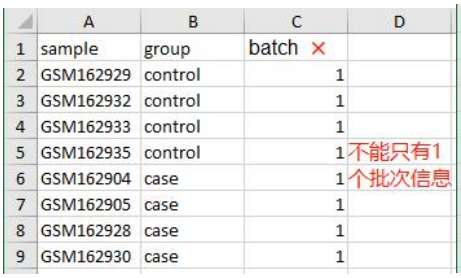
	A	B	C	D
1	sample	group	pair ✗	
2	GSM162929	control	1	
3	GSM162932	control	2	
4	GSM162933	control	5	
5	GSM162935	control	6	
6	GSM162904	case	1	配对信息不允许重复
7	GSM162905	case	2	
8	GSM162928	case	1	
9	GSM162930	case	2	

- batch 列：不允许含有缺失和特殊字符；当提供了 batch 列时，batch 至少是 2 个批次，且至多 4 个批次；同一批次内至少含有 2 个样本；注意，应该每个批次都能覆盖所有的分组，这样校正批次的效果才好；不建议批次只含有 1 个分组，批次校正可能会部分掩盖掉分组间的差别。

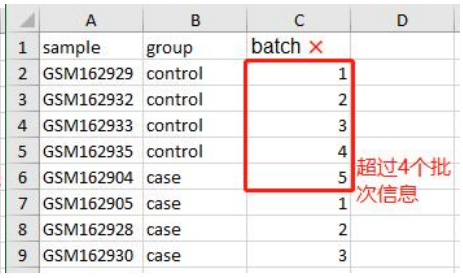
	A	B	C	D
1	sample	group	batch √	
2	GSM162929	control	3	
3	GSM162932	control	1	
4	GSM162933	control	1	
5	GSM162935	control	2	
6	GSM162904	case	3	正确信息
7	GSM162905	case	2	
8	GSM162928	case	2	
9	GSM162930	case	1	



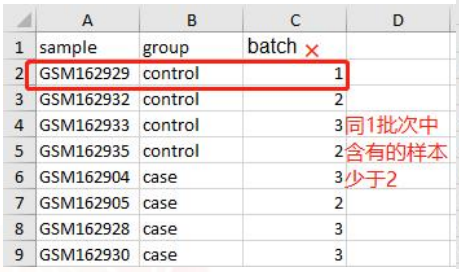
	A	B	C	D
1	sample	group	batch ×	
2	GSM162929	control		
3	GSM162932	control		
4	GSM162933	control	1	
5	GSM162935	control	2	含有缺失
6	GSM162904	case	3	
7	GSM162905	case	2	
8	GSM162928	case	2	
9	GSM162930	case	1	



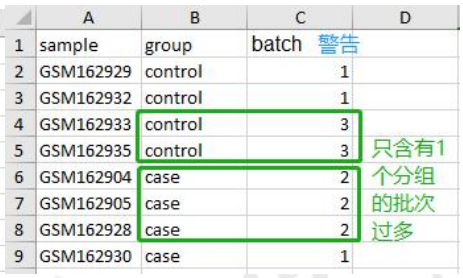
	A	B	C	D
1	sample	group	batch ×	
2	GSM162929	control	1	
3	GSM162932	control	1	
4	GSM162933	control	1	
5	GSM162935	control	1	不能只有1
6	GSM162904	case	1	个批次信息
7	GSM162905	case	1	
8	GSM162928	case	1	
9	GSM162930	case	1	



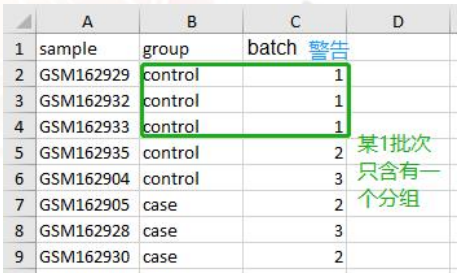
	A	B	C	D
1	sample	group	batch ×	
2	GSM162929	control	1	
3	GSM162932	control	2	
4	GSM162933	control	3	
5	GSM162935	control	4	
6	GSM162904	case	5	超过4个批
7	GSM162905	case	1	次信息
8	GSM162928	case	2	
9	GSM162930	case	3	



	A	B	C	D
1	sample	group	batch ×	
2	GSM162929	control	1	
3	GSM162932	control	2	
4	GSM162933	control	3	同1批次中
5	GSM162935	control	2	含有的样本
6	GSM162904	case	3	少于2
7	GSM162905	case	2	
8	GSM162928	case	3	
9	GSM162930	case	3	



	A	B	C	D
1	sample	group	batch 警告	
2	GSM162929	control	1	
3	GSM162932	control	1	
4	GSM162933	control	3	
5	GSM162935	control	3	只含有1
6	GSM162904	case	2	个分组
7	GSM162905	case	2	的批次
8	GSM162928	case	2	过多
9	GSM162930	case	1	



	A	B	C	D
1	sample	group	batch 警告	
2	GSM162929	control	1	
3	GSM162932	control	1	
4	GSM162933	control	1	
5	GSM162935	control	2	某1批次
6	GSM162904	control	3	只含有一
7	GSM162905	case	2	个分组
8	GSM162928	case	3	
9	GSM162930	case	2	

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

- 缺失值处理：可以选择用“插补法”来处理缺失值，对于数据中含有少量缺失值情况的处理，一般情况下芯片是不含有缺失值，个别数据集会出现有。可以选择“不处理”。
  - 注意：“插补法”使用 impute 包的 impute.knn 函数进行缺失值填充。使用该方法的前提，样本（列）中不应该存在超过 80%%探针数据为缺失，或者，探针（行）数据不应该存在超过 50%%的样本为缺失。
- 标准化处理：可以选择使用 limma 包 normalizeBetweenArrays 来对数据进行标准化处理。可以选择“不处理”。
  - 注意：一般情况从 Series Matrix 文件开始的文件基本都是标准化好的，如果觉得标准化还不够（比如箱式图查看分布的时候看到箱式图的上下四分位和中位数差别很大），可选择该参数处理，把不同样本都拉到一个水平线上。

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: 基于芯片表达谱数据进行两组差异分析
分析流程: limma包标准差异分析流程
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	logFC	AveExpr	t	P.Value	adj.P.Val	B
226847_at	5.8481	6.6834	4.0192	0.0034	0.8739	-2.2638
225387_at	5.3013	4.5282	3.9593	0.0037	0.8739	-2.306
242277_at	5.0225	4.7807	3.9539	0.0037	0.8739	-2.3098
237056_at	4.9518	2.6626	9.5603	7.53e-06	0.1372	-0.28938
204948_s_at	4.8729	8.0047	4.4926	0.0017	0.7432	-1.9526
207345_at	4.8655	5.8567	4.1674	0.0027	0.8319	-2.1622
224579_at	4.7975	3.5842	2.619	0.0291	0.9559	-3.4104
209656_s_at	4.7663	4.4505	2.2042	0.0565	0.9559	-3.7985
226610_at	4.7662	4.1399	4.2	0.0026	0.8200	-2.1403
204049_s_at	4.7636	5.7103	3.7616	0.0049	0.8739	-2.4499
204036_at	4.6106	6.6124	3.179	0.0120	0.9339	-2.914
220658_s_at	4.5853	4.0069	5.5787	0.0004	0.6268	-1.3757
213652_at	4.5391	2.4259	3.1131	0.0133	0.9478	-2.97

差异分析.xlsx　标准化数据.csv

此表格为差异分析结果（页面只展示 60 个），提供 EXCEL 格式下载。另外，还提供 normalizeBetweenArrays 标准化处理后的数据，提供 CSV 格式下载。

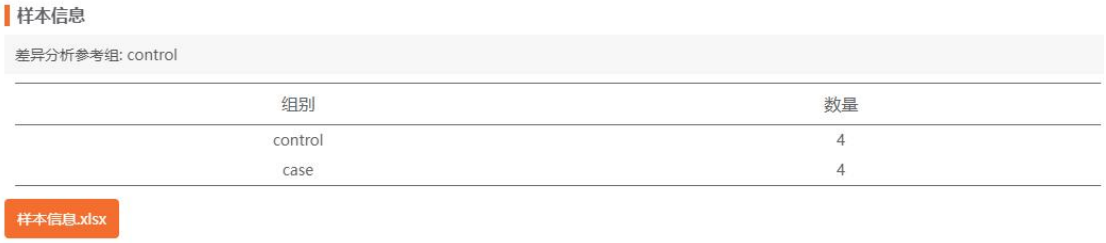
补充结果

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
当校正后p值均不满足<0.05时会考虑使用p值来作为统计

筛选条件	筛选后的数量
|LogFC|>2 & pvalue<0.05	914
|LogFC|>1 & pvalue<0.05	2525
|LogFC|>0.58 & pvalue<0.05	2810

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: control

组别	数量
control	4
case	4

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，根据样本分组对芯片数据进行标准组间(2 组)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。

6. 为什么要对数据做 log 化处理?

答:

因为，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 处理的情况，本模块会自动进行 log 化处理，也可以自行先做处理后上传数据进行分析。（注意，当数据中存在小于等于 0 时，不建议再进行 log 化处理，且这类型的数据过多，如单个样本中所有表达值都较小时，建议删除样本或者更换数据）

7. 什么是“插补法”?

答:

从不同数据库平台所下载的数据中，多多少少都会存在数据缺失情况。如果你的表达谱数据中的分子/基因含有部分样本缺失问题，可以把整个基因（行）都删除，但是如果基因缺失比例很大（行数太多），这个时候强行删除就会带来后续分析的偏差。

本模块使用的"插补法"，是使用 impute R 包的 impute.knn 函数进行缺失值填充。使用该方法的前提，数据中含有少量缺失，样本（列）中不应该存在超过 80%%探针数据为缺失，或者，探针（行）数据不应该存在超过 50%%的样本为缺失。

8. 模块是怎么解决批次效应问题的？

答：

只有当用户提供样本的准确批次信息时，才会进行批次矫正（具体填写标准，可以到【数据格式】中查看）。本模块使用 sva R 包的 ComBat 函数进行批次矫正。使用该方法的前提，数据中不应该含有过多缺失，探针（行）数据不应该存在过多样本为缺失，此时，请尝试选择删除批次信息或剔除掉这些探针。