表达差异 - 转录组 Counts 数据(两组)差异分析

id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ALPG	340.46	10.686	1.681	6.3571	2.06e-10	1.55e-08
IBSP	114.97	9.842	1.1802	8.3393	7.47e-17	1.7e-14
MMP3	1.631e+04	9.6103	1.2308	7.8081	5.81e-15	1.03e-12
AC005550.2	284.68	8.8288	1.2257	7.203	5.89e-13	6.79e-11
DMBT1	3.546e+04	8.63	0.74249	11.623	3.15e-31	2.86e-28
REG4	2.326e+04	8.6015	1.336	6.4382	1.21e-10	9.76e-09
LYPD8	43.459	8.4383	3.384	2.4936	0.0126	0.0642
ATOH1	323.07	8.3476	1.5366	5.4326	5.55e-08	2.43e-06
ADGRG7	752.93	8.3468	0.83084	10.046	9.55e-24	4.73e-21
MMP12	5455.2	8.3371	1.023	8.1495	3.66e-16	7.42e-14
AC073365.1	36.149	8.1726	3.001	2.7233	0.0065	0.0387
CXCL6	4931.4	7.9654	1.1747	6.7805	1.2e-11	1.11e-09

网址：https://www.xiantao.love

更新时间：2023.02.22

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。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）
- 测序数据差异分析：测序数据的差异分析一般是下游分析开始（下游分析一般以拿到表达谱作为起点，在这之前是上游分析），上述 R 包基于的统计模型基本都是 Counts 格式，所以进行差异分析一般是用 Counts 格式，其他格式包括转化都是不太合适的（需要提供原始的）。
- 测序数据 Counts 格式：原始的 Counts 都是正整数，满足负二项分布。请注意，并不是将数据转换成正整数就可以，这是不正确的，需要确认数据是不是原始的 Counts 格式。

应用场景

对高通量测序数据进行差异分析，旨在找出不同样本（组）间存在差异表达的分子，得到差异表达分子之后，可以再对其做 GO 功能显著性和 KEGG Pathway 显著性分析。

主要结果

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Description	setSize	enrichmentS	NES	pvalue	p.adjust	qvalue	rank	leading_edg	core_enrichment	
2	REACTOME_	REACTOME_	237	0.66674999	3.04379824	1E-10	7.5813E-09	6.0164E-09	1898	tags=43%, lis	CDC45/CDCA8/MCM10/CD	
3	REACTOME_	REACTOME_	142	0.70095021	3.00209513	1E-10	7.5813E-09	6.0164E-09	1151	tags=41%, lis	CDC45/MCM10/MYBL2/TOI	
4	REACTOME_	REACTOME_	137	0.69582082	2.95608573	1E-10	7.5813E-09	6.0164E-09	1763	tags=49%, lis	CDC45/MCM10/UBE2C/UBE	
5	REACTOME_	REACTOME_	458	0.6073098	2.94715166	1E-10	7.5813E-09	6.0164E-09	1763	tags=36%, lis	CDC45/CDCA8/MCM10/CD	
6	REACTOME_	REACTOME_	134	0.68661939	2.90126992	1E-10	7.5813E-09	6.0164E-09	1898	tags=49%, lis	CDC45/MCM10/CCNB2/CDI	
7	REACTOME_	REACTOME_	110	0.71134747	2.90065947	1E-10	7.5813E-09	6.0164E-09	1898	tags=54%, lis	CDC45/UBE2C/UBE2S/CCN	
8	REACTOME_	REACTOME_	201	0.65023395	2.88715897	1E-10	7.5813E-09	6.0164E-09	1850	tags=40%, lis	CDCA8/CDC20/CENPE/CCN	
9	WP_RETINOI	WP_RETINOI	84	0.73015214	2.81434613	1E-10	7.5813E-09	6.0164E-09	1329	tags=51%, lis	CDC45/CCNB2/TOP2A/RRM	
10	REACTOME_	REACTOME_	145	0.65531327	2.79876078	1E-10	7.5813E-09	6.0164E-09	1763	tags=46%, lis	CDC45/UBE2C/UBE2S/CCN	
11	REACTOME_	REACTOME_	92	0.70154933	2.77241955	1E-10	7.5813E-09	6.0164E-09	449	tags=27%, lis	CDCA8/CDC20/CENPE/NDC	
12	REACTOME_	REACTOME_	111	0.68070952	2.75987156	1E-10	7.5813E-09	6.0164E-09	1898	tags=46%, lis	CDC45/MCM10/UBE2C/UBE	
13	REACTOME_	REACTOME_	162	0.64291611	2.75628681	1E-10	7.5813E-09	6.0164E-09	1762	tags=38%, lis	CDCA8/CDC20/CENPE/NDC	
14	NABA_CORE	NABA_CORE	201	-0.66811865	-2.75251153	1E-10	7.5813E-09	6.0164E-09	1537	tags=50%, lis	MFAP3/IGFBP7/THSD4/VW	
15	KEGG_CELL_	KEGG_CELL_	114	0.66878969	2.72918483	1E-10	7.5813E-09	6.0164E-09	1230	tags=40%, lis	CDC45/CDC20/CCNB2/CCN	

- id：对应上传数据中第一列的 id，可以理解成测序数据的分子。
- 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）
- log2FoldChange：差异倍数 FoldChange 值 log2 转化，当 log2FoldChange=1 时，即说明有 2 倍的差异。（筛选差异的条件之一）
- lfcSE: log2FoldChange 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）

空值的原因可能是因为该分子在样本间表达不显著导致的无法计算一些值。

数据格式

转录组-Counts 数据

	A	B	C	D	E	F	G	H	I
1	id	TCGA-L5-A43C-01A-	TCGA-L5-A43C-11A-	TCGA-L5-A4OG-01A-	TCGA-L5-A4OG-11A-	TCGA-L5-A4OJ-01A-	TCGA-L5-A4OJ-11A-	TCGA-L5-A4OO-01A-	TCGA-L5-A4OO-11A-1
2	TSPAN6	3420	1220	2135	1754	1829	2635	3067	1024
3	TNMD	1	0	1	2	0	6	1	3
4	DPM1	2602	900	6094	1392	2556	1452	3223	1106
5	SCYL3	671	479	1142	1025	1920	828	1617	621
6	C1orf112	443	70	868	199	1146	225	374	109
7	FGR	3855	149	904	522	181	318	1252	365
8	CFH	9476	1484	5260	7614	2238	2658	8874	3405
9	FUCA2	3495	1223	8052	2085	5487	1499	3065	1129
10	GCLC	1495	1599	2982	3437	4364	2239	4132	1747
11	NFYA	1494	501	3435	1123	4968	1240	2859	1026
12	STPG1	276	179	830	339	693	373	748	284
13	NIPAL3	1786	1872	5751	3895	5379	3641	3746	2571

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。
- 第一列以后的数值部分为不同分子在各样本中的原始 Counts 数值。
  - 原始的 Counts 都是正整数，满足负二项分布。请注意，并不是将数据转换成正整数就可以，这是不正确的，需要确认数据是不是原始的 Counts 格式。
- 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。
- 文件不能大于 80M（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）

样本信息

	A	B
1	sample	group
2	TCGA-L5-A43C-11A-	Normal
3	TCGA-L5-A4OG-11A	Normal
4	TCGA-L5-A4OJ-11A-	Normal
5	TCGA-L5-A4OO-11A	Normal
6	TCGA-L5-A43C-01A-	Tumor
7	TCGA-L5-A4OG-01A	Tumor
8	TCGA-L5-A4OJ-01A-	Tumor
9	TCGA-L5-A4OO-01A	Tumor

数据要求：

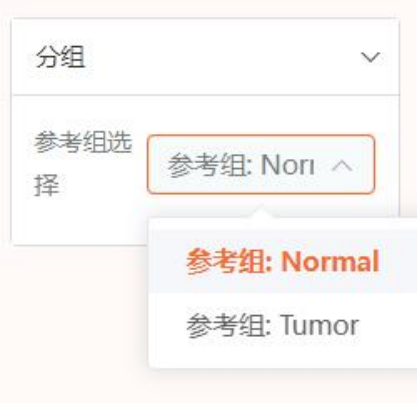
- 第一列为样本编号，请与表达谱数据第一行样本保持一致。
- 第二列为对应的分组信息，第一行为样本分组信息，需要提供两个分组，每个分组至少含有 2 个样本以上的生物学重复。（重复不是指复制样本，是平行实验样本）。
  - 注意：上传数据后会自动返回分组信息到【参考组选择】参数中。

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

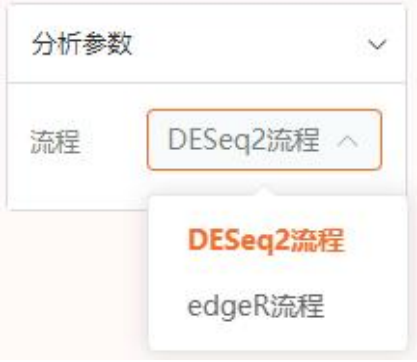
（说明：标注了颜色的为常用参数。）

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。

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

差异分析: 基于转录组Counts数据进行两组差异分析, 当前参考组为: Normal
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
ALPG	340.46	10.686	1.681	6.3571	2.06e-10	1.55e-08
IBSP	114.97	9.842	1.1802	8.3393	7.47e-17	1.7e-14
MMP3	1.631e+04	9.6103	1.2308	7.8081	5.81e-15	1.03e-12
AC005550.2	284.68	8.8288	1.2257	7.203	5.89e-13	6.79e-11
DMBT1	3.546e+04	8.63	0.74249	11.623	3.15e-31	2.86e-28
REG4	2.326e+04	8.6015	1.336	6.4382	1.21e-10	9.76e-09
LYPD8	43.459	8.4383	3.384	2.4936	0.0126	0.0642
ATOH1	323.07	8.3476	1.5366	5.4326	5.55e-08	2.43e-06
ADGRG7	752.93	8.3468	0.83084	10.046	9.55e-24	4.73e-21
MMP12	5455.2	8.3371	1.023	8.1495	3.66e-16	7.42e-14
AC073365.1	36.149	8.1726	3.001	2.7233	0.0065	0.0387
CXCL6	4931.4	7.9654	1.1747	6.7805	1.2e-11	1.11e-09
CCL20	3825	7.801	1.0022	7.7842	7.02e-15	1.22e-12

差异分析.xlsx　标准化数据.csv

此表格提供转录组-差异分析结果（页面只展示 60 个），提供 EXCEL 格式以及标准化数据的 CSV 格式下载。

补充结果

样本信息

差异分析参考组: Normal

组别	数量
Normal	4
Tumor	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。

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	3606
|LogFC|>1 & p.adj<0.05	5624
|LogFC|>0.58 & p.adj<0.05	6091

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。

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 、Docx 格式及完整报告下载。

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） 对原始 Counts 数据进行差异分析，同时输出标准化(Normalize)好的数据。

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. 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. 如果没有原始 Counts 数据，是否能进行差异分析?

答:

大部分的 R 包都不支持经过转换或者非 Counts 格式的差异分析，请尽可能获取到原始的 Counts 数据。

3. 如何挑选分子?

答:

首先关注 log2FodlChange 和校正后的 p 值，在满足设定的阈值下，另外再关注 baseMean，即平均表达，这个也不能少，如果一个分子表达小，可能一点表达量的改变就会对 logFC 有很大的影响。

4. 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。

5. 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导致的无法计算一些值。