表达差异 - [云]筛选分子

id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	gene_name	gene_type
ENSG00000169297.8	243.39	6.7276	1.5266	4.4068	1.05e-05	0.0120	NR0B1	protein_coding
ENSG00000155897.10	94.581	6.2112	1.237	5.021	5.14e-07	0.0026	ADCY8	protein_coding
ENSG00000145920.15	409.21	6.112	1.1295	5.4113	6.26e-08	0.0005	CPLX2	protein_coding
ENSG00000104755.15	11.883	5.9813	1.9392	3.0844	0.0020	0.1089	ADAM2	protein_coding
ENSG00000249196.6	19.174	5.9336	1.5833	3.7475	0.0002	0.0387	TMEM132D-AS1	lncRNA
ENSG00000285894.2	8.6506	5.8862	2.1203	2.7761	0.0055	0.1683	AL136372.2	lncRNA
ENSG00000276399.1	15.515	5.6507	1.8301	3.0876	0.0020	0.1085	FLJ36000	lncRNA
ENSG00000272342.1	12.058	5.6374	1.4159	3.9814	6.85e-05	0.0297	AC116609.3	lncRNA
ENSG00000180053.7	9.2138	5.6023	1.7525	3.1968	0.0014	0.0936	NKX2-6	protein_coding
ENSG00000232258.6	7.0551	5.5944	1.3562	4.1251	3.71e-05	0.0248	TMEM114	protein_coding
ENSG00000256597.3	13.972	5.4962	1.4938	3.6794	0.0002	0.0436	LINC02393	lncRNA
ENSG00000258986.7	15.446	5.3323	1.1044	4.8283	1.38e-06	0.0052	TMEM179	protein_coding
ENSG00000250954.6	5.4829	5.2282	1.5312	3.4144	0.0006	0.0675	AC016687.3	lncRNA

网址：https://www.xiantao.love

更新时间：2023.03.03

目录

基本概念

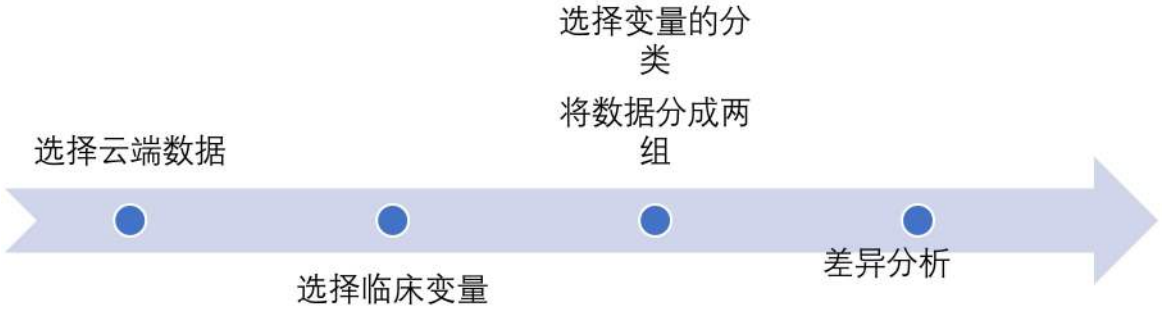
- 差异分析：通过 R（以及常用 R 包）或者其他手段分析和筛选出表达谱中两组样本间的差异表达分子。
- 常用 R 包介绍：
  - DESeq2 包：支持测序数据的 Counts 格式，也支持其他高通量数据的分析。
  - edgeR 包：支持测序数据的 Counts 格式，也支持其他高通量数据的分析。
    - https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf （edgeR 包帮助文档 16 页）

应用场景

筛选公共数据（云端数据）中的差异分子，能更加针对性筛选出有研究价值的分子。

- 本模块可以利用云端数据，基于清理好的临床数据，自定义差异分析的两组（如下图所示），从而获得两组的差异分子列表。

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	2921.8778	-0.25352085	0.33699753	-0.75229291	0.45187493	0.81534244	TSPAN6	protein_coding
3	ENSG00000000005.6	1.09856918	0.11430172	1.26017362	0.09070315	0.92772846		TNMD	protein_coding
4	ENSG00000000419.13	3158.67128	0.13024632	0.18317989	0.71102956	0.47706592	0.82639238	DPM1	protein_coding
5	ENSG00000000457.14	1027.69697	-0.20525717	0.14633326	-1.40266932	0.16071547	0.58089565	SCYL3	protein_coding
6	ENSG00000000460.17	889.936521	-0.10060193	0.26021964	-0.38660392	0.69904946	0.91732502	C1orf112	protein_coding
7	ENSG00000000938.13	403.913243	-0.09386566	0.32334282	-0.29029763	0.77158855	0.94104909	FGR	protein_coding
8	ENSG00000000971.16	4848.88998	0.49739917	0.405799	1.22572792	0.22030108	0.65026568	CFH	protein_coding
9	ENSG00000001036.14	2164.53742	0.06196385	0.19696432	0.31459431	0.7530697	0.93594935	FUCA2	protein_coding
10	ENSG00000001084.13	13681.7974	0.20466566	0.49598563	0.41264433	0.67986722	0.90986956	GCLC	protein_coding
11	ENSG00000001167.14	2157.02496	-0.04890496	0.16333158	-0.29942131	0.76461861	0.93887341	NFYA	protein_coding
12	ENSG00000001460.18	1108.17927	-0.08070597	0.27301035	-0.29561507	0.76752406	0.940045	STPG1	protein_coding
13	ENSG00000001461.17	2844.54161	-0.18483972	0.22570047	-0.81896025	0.41280909	0.79015526	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_2FoldChange: 差异倍数 FoldChange 值 log2 转化，当 log_2FoldChange=1 时，即说明有 2 倍的差异。（筛选差异的条件之一）
- lfcSE: log_2FoldChange 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	-0.26476254	5.32219349	0.64244786	0.42795214	0.82622061	TSPAN6	protein_coding
3	ENSG00000000419.13	0.12386139	5.43268745	0.43166601	0.51524572	0.86417251	DPM1	protein_coding
4	ENSG00000000457.14	-0.2138766	3.81179377	1.82140835	0.1853597	0.66923279	SCYL3	protein_coding
5	ENSG00000000460.17	-0.10719816	3.60566408	0.16896204	0.68341104	0.92349689	C1orf112	protein_coding
6	ENSG00000000938.13	-0.10659618	2.46586865	0.10721341	0.74518781	0.93992657	FGR	protein_coding
7	ENSG00000000971.16	0.49089217	6.05113073	1.36031166	0.25095897	0.72366526	CFH	protein_coding
8	ENSG00000001036.14	0.05692165	4.88868077	0.07927953	0.77984762	0.9511992	FUCA2	protein_coding
9	ENSG00000001084.13	0.19956385	7.55439145	0.14774338	0.70290533	0.92832892	GCLC	protein_coding
10	ENSG00000001167.14	-0.05472827	4.87979082	0.10714372	0.74526764	0.93994204	NFYA	protein_coding
11	ENSG00000001460.18	-0.08685002	3.92616436	0.0994864	0.75422205	0.94284936	STPG1	protein_coding
12	ENSG00000001461.17	-0.19046153	5.28222677	0.72830628	0.39893329	0.8114833	NIPAL3	protein_coding
13	ENSG00000001497.18	-0.12697976	5.45028978	0.36636933	0.54869106	0.8771228	LAS1L	protein_coding

- id: ensembl 库注释的分子 ID。
- logFC: 差异倍数 FoldChange 值 log2 转化，当 log_2FoldChange=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可选临床变量可能会有不同。注意查看当前数据参数选中的云端数据。

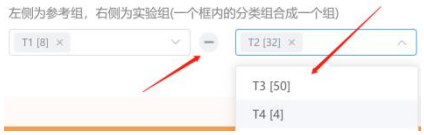
这里为任务式模块，提交任务后需要到**历史记录**中刷新并等待任务完成，（分析时间大概在几分钟到十几分钟不等，具体要看对应的数据集的样本量，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）

参数说明

（说明：标注了颜色的为常用参数。）

特殊参数

- 临床变量：下拉框将列出对应所选数据集的临床变量，选中变量后，右侧可选关联的分类信息，如 Pathologic_T_stage 对应 T1-T4 分类。



- 分组：在变量对应的分类中自定义比较分组，左侧为参考组，右侧为实验组。

加减号修改分组，一个框内的分类组可以合成一个组，如 T1 和 T2 分类作为一组等等。注意，选择的分类（单个组）中样本数需要大于等于 2 才能进行差异分析。根据具体情况可以自由选择参考组的分组。

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

筛选分子-云

筛选分子-云: 基于云端数据筛选分子
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
ENSG00000169297.8	243.39	6.7276	1.5266	4.4068	1.05e-05	0.0120	NR0B1	protein_coding
ENSG00000155897.10	94.581	6.2112	1.237	5.021	5.14e-07	0.0026	ADCY8	protein_coding
ENSG00000145920.15	409.21	6.112	1.1295	5.4113	6.26e-08	0.0005	CPLX2	protein_coding
ENSG00000104755.15	11.883	5.9813	1.9392	3.0844	0.0020	0.1089	ADAM2	protein_coding
ENSG00000249196.6	19.174	5.9336	1.5833	3.7475	0.0002	0.0387	TMEM132D-AS1	lncRNA
ENSG00000285894.2	8.6506	5.8862	2.1203	2.7761	0.0055	0.1683	AL136372.2	lncRNA
ENSG00000276399.1	15.515	5.6507	1.8301	3.0876	0.0020	0.1085	FLJ36000	lncRNA
ENSG00000272342.1	12.058	5.6374	1.4159	3.9814	6.85e-05	0.0297	AC116609.3	lncRNA
ENSG00000180053.7	9.2138	5.6023	1.7525	3.1968	0.0014	0.0936	NKX2-6	protein_coding
ENSG00000232258.6	7.0551	5.5944	1.3562	4.1251	3.71e-05	0.0248	TMEM114	protein_coding

筛选分子.xlsx

此表格提供差异分析结果（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

分组信息

差异分析参考组: Pathologic_T_stage(T1)

组别	数量
T1	8
T2	27

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	131
|LogFC|>1 & p.adj<0.05	200
|LogFC|>0.58 & p.adj<0.05	206

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选择临床变量，并按照对应变量的分类自定义比较组，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. 某个肿瘤癌旁和癌的差异分子列表如何获得?

答:

在变量中找到 status 临床变量，这个变量会提供癌旁和癌的分组信息。

注意，不是所有的云端数据都有癌旁样本的，即可能会没有这个变量或者这个变量内没有 Normal 选项，这个情况就说明这个云端数据没有癌旁样本，建议选择其他的变量进行分析。