交互网络 - [免疫浸润] Cibersort 算法

Sample	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting
TCGA-23-1120	0	0.004965	0	0.026006	0	0.18072
TCGA-29-1695	0	0.0034876	0	0.015348	0	0.056311
TCGA-61-2003	0.001246	0.010855	0	0.03552	0	0.16754
TCGA-13-1404	0.085022	0	0.10642	0.04998	0	0.085255
TCGA-13-1512	0	0.057859	0.022579	0.067313	0	0.0067551
TCGA-23-1022	0.0029911	0	0	0	0	0.10825
TCGA-61-2088	0	0.031858	0.10739	0	0	0.16438
TCGA-13-1507	0.015874	0	0.19867	0	0	0.05262
TCGA-25-1628	0.084811	0	0.040915	0.28276	0	0.028213
TCGA-31-1944	0.043617	0	0.065277	0.045486	0	0.25531

网址：https://www.xiantao.love

更新时间：2023.12.05

目录

基本概念

- 免疫浸润分析：利用转录组或者其他组学的数据，通过算法计算组织中免疫细胞的分数情况，推测组织中免疫细胞的构成情况。
- CIBERSORT 算法
  - 该算法由 CIBERSORT.R 脚本分析，是基于线性支持向量回归（linear support vector regression）的原理对人类免疫细胞亚型的表达矩阵进行反卷积。
  - 22 种免疫细胞亚型的基因表达特征集：LM22.txt，由 547 个基因组成，包括 7 种 T 细胞类型、naïve 和记忆 B 细胞、浆细胞、 NK 细胞和骨髓亚群，来源于 CIBERSORTx 网站（https://cibersortx.stanford.edu/），如果想要了解更多使用和细节，可以到该网站进一步了解。

应用场景

- 免疫浸润分析主要是基于组织样本转录组测序数据或者微阵列芯片数据的分析。
- 在肿瘤研究中，所采集的组织样本并不止含有肿瘤细胞，还会有正常细胞、免疫细胞、基质细胞等，而不同的细胞具有一些标志性的 marker，免疫细胞也是一样。因此，可以根据这些 marker 基因在组织中的表达量，结合一些生物信息学的算法，评估和量化免疫细胞浸润、免疫微环境。
- 免疫浸润分析，正是为了弄清楚肿瘤组织当中免疫细胞的构成而产生的。
- 不同算法对应所使用的 marker 不同，如果手上的数据并未包含对应的 marker 基因，那么将由于匹配不全而无法进行免疫浸润分析。

主要结果

	A	B	C	D	E	F	G
1	Sample	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 n	T cells CD4 m
2	TCGA-23-1120	0	0.004965017	0	0.02600566	0	0.18072038
3	TCGA-29-1695	0	0.003487573	0	0.01534778	0	0.05631147
4	TCGA-61-2003	0.00124601	0.010855121	0	0.03552043	0	0.16754249
5	TCGA-13-1404	0.08502239	0	0.106417842	0.04998018	0	0.08525493
6	TCGA-13-1512	0	0.057858739	0.022579289	0.06731278	0	0.00675509
7	TCGA-23-1022	0.002991076	0	0	0	0	0.10825467
8	TCGA-61-2088	0	0.031858069	0.107388116	0	0	0.16438331
9	TCGA-13-1507	0.015874101	0	0.198667004	0	0	0.05262036
10	TCGA-25-1628	0.084811414	0	0.040915066	0.28275785	0	0.02821264
11	TCGA-31-1944	0.043617133	0	0.065277156	0.04548606	0	0.25531301

- 表格中的行名（第一列）代表样本，列名（从第二列开始）代表免疫细胞类型，数值代表每个样本在不同细胞的预测比例。
- 列信息
  - B 细胞：B cells naïve、B cells memory
  - PCs：Plasma cells
  - T 细胞：T cells CD8、T cells CD4 naive、T cells CD4 memory resting、T cells CD4 memory activated、T cells follicular helper、T cells regulatory (Tregs) 、T cells gamma delta
  - NK 细胞：NK cells resting、NK cells activated
  - 髓系细胞：Monocytes、Macrophages M0、Macrophages M1、Macrophages M2、Dendritic cells resting、Dendritic cells activated、Mast cells resting、Mast cells activated
  - 粒系：Eosinophils、Neutrophils
  - P-value：置换检验（蒙特卡罗方法），在所有细胞子集上反卷积结果的统计显着性，越小越可信。
  - Correlation：相关系数，是通过将实际 bulk RNA 基因表达谱（GEP）与预测的 bulk RNA GEP 进行比较而得出的相关性，后者 GEP 是使用估算的细胞比例和来自签名矩阵 signature matrix 的相应表达谱进行计算出的。

- RMSE：均方根误差，实际 bulk RNA 基因表达谱（GEP）与预测的 bulk RNA GEP 的均方根误差，越小效果越好。

主要结果包含了 22 种细胞类型和相关统计检验结果的数据，如果后续需要进行可视化，可以先删除统计检验结果列信息后，作为可视化输入数据进行下游分析。

数据格式

	A	B	C	D	E	F	G
1	gene_id	TCGA-23-1120	TCGA-29-1695	TCGA-61-2003	TCGA-13-1404	TCGA-13-1512	TCGA-23-1022
2	RAB4B	2.021775728	4.908403094	5.433856696	2.388734048	6.33655999	5.046220328
3	C12orf5	1.952965422	10.24395474	3.017626555	1.882028238	2.606324505	0.830262666
4	RNF44	28.59748336	21.00536905	28.13665217	28.41160017	14.89508077	16.6884242
5	DNAH3	0.084499389	0.016197327	0.168182336	0.167447976	0.511683969	0.059074999
6	RPL23A	391.8151425	254.0200414	223.8741998	562.7939575	277.7897572	742.1034512
7	ARL8B	27.91084812	24.8952554	23.64663025	16.26606641	15.37472808	11.97626001
8	CALB2	0.840168409	1.252111397	1.725005886	4.362976611	0.394821136	7.950225258
9	MFSD3	18.51669087	16.81430036	9.568504292	21.97428333	33.25160649	36.52365769
10	PIGV	4.920701107	2.779924919	4.31818145	10.45757245	6.785078658	5.325567493
11	ZNF708	1.564035738	1.816362239	1.77258742	1.404986122	0.980998278	1.351677936
12	MYADML2	0.039560332	0.024645241	0.011436867	0.172733911	0.056112361	0.028763612
13	PHEX	0.093566277	0.267590803	0.317050584	0.179568589	0.507709635	1.655669383

数据要求：

- 数据至少有 2 列以上，至少需要 100 行数据。第一行为样本编号，第一列为基因名，不能含有缺失、重复及特殊字符。
- 数值部分为不同基因在各样本中的表达量，需要非 log 转换的数据。
- 数据中不能含有单个样本（单列）的数据都是一样的，即样本方差不能为 0。
- Cibersort 分析会先从表达谱中取在内置参考数据（LM22）中的 marker 基因，再使用核心算法分析，因此需要尽量多的匹配到内置 marker，否则结果可能达不到预期效果。
- 最多支持 500 列，70000 行。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。

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 十几分钟不等，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

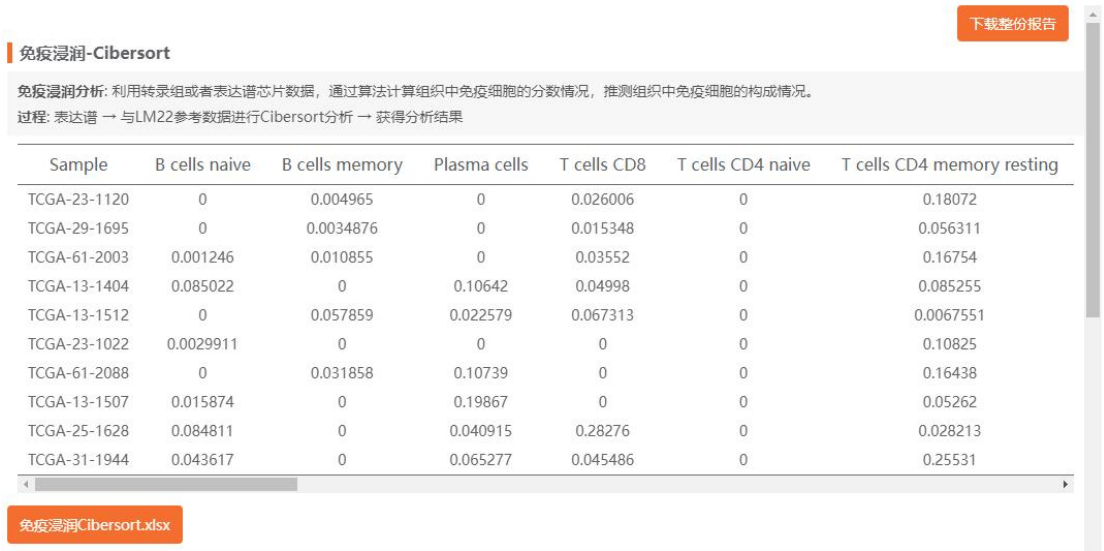
（说明：标注了颜色的为常用参数。）

分析参数

- 种子号：设置种子数可以保证统计检验 p 值结果可重复，默认为 2022，此参数请输入非零整数。
- 置换次数：置换检验显著性分析的排列次数，主要影响 p 值结果，目前可以选 1 次。
- 分数位归一化：是否使用分位数标准化，芯片数据可选，建议对 RNA-Seq 数据禁用。
- 物种：物种选择，目前可以选人源。

结果说明

主要结果



下载整份报告

免疫浸润-Cibersort

免疫浸润分析: 利用转录组或者表达谱芯片数据，通过算法计算组织中免疫细胞的分数情况，推测组织中免疫细胞的构成情况。
过程: 表达谱 → 与LM22参考数据进行Cibersort分析 → 获得分析结果

Sample	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting
TCGA-23-1120	0	0.004965	0	0.026006	0	0.18072
TCGA-29-1695	0	0.0034876	0	0.015348	0	0.056311
TCGA-61-2003	0.001246	0.010855	0	0.03552	0	0.16754
TCGA-13-1404	0.085022	0	0.10642	0.04998	0	0.085255
TCGA-13-1512	0	0.057859	0.022579	0.067313	0	0.0067551
TCGA-23-1022	0.0029911	0	0	0	0	0.10825
TCGA-61-2088	0	0.031858	0.10739	0	0	0.16438
TCGA-13-1507	0.015874	0	0.19867	0	0	0.05262
TCGA-25-1628	0.084811	0	0.040915	0.28276	0	0.028213
TCGA-31-1944	0.043617	0	0.065277	0.045486	0	0.25531

免疫浸润Cibersort.xlsx

主要结果格式为表格格式，提供 Excel 格式下载，结果报告可以下载包括说明文本的内容。

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 十几分钟不等，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 格式下载。

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：CIBERSORT（脚本）

处理过程：

1) 基于 CIBERSORT(CIBERSORT.R 脚本分析)核心算法
2) 利用 CIBERSORTx 网站(https://cibersortx.stanford.edu/)提供的 22 种免疫细胞的 signature matrix 基因表达谱
3) 计算各样本的免疫浸润情况。

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. 芯片数据中只有探针信息怎么办？或者没有对应基因表达谱？

答：

Cibersort 分析会先从表达谱中取在内置参考数据（LM22）中的 marker 基因，再使用核心算法分析，因此需要尽量多的匹配到内置 marker，否则结果可能达不到预期效果。

芯片数据只有探针信息时，需要自行进行 ID 转换，获得基因 Symbol，才能进行免疫基因润-Cibersort 算法分析。

2. 如何进行可视化的操作？

答：

提交分析任务完成后，历史记录中会有一条对应的结果记录，可以下载对应的结果表格（免疫浸润 Cibersort.xlsx）。对数据进行相应绘图模块的整理后，即可进行对应的可视化。

3. 免疫浸润可以做什么实验验证？

答：

可以通过免疫组化检测对应的免疫细胞的 markers，也可以对组织做流式分析分析细胞的情况等等。具体要根据研究情况进行安排。

4. 仙桃使用的算法（其他一些算法）给到的结果 跟 其他方法（或者别的数据库 TIMER 等）趋势不一样，这个是什么原因？

答：

不同算法之间可能是会存在有一定的差别，况且算法只是一种推测手段，实际是什么情况还是需要通过做实验来确定的。所以，如果只是单纯想要拿一些结果来充实自己的研究，那么可以只放满足自己想要的趋势的数据。

5. 怎么上传的样本数和分析后的样本数不对应？是什么原因？

答：

Cibersort 分析会先从表达谱中取在内置参考数据（LM22）中的 marker 基因，再使用核心算法分析，因此需要尽量多的匹配到内置 marker，否则结果可能达不到预期效果。

Cibersort 分析的核心算法中，使用 svm 方法进行多个 NU 值线性回归，从中选出最优模型，从而获取模型系数。当匹配到得内置 marker 过少，可能会导致模型无法预测到样本的回归系数，使得对应的样本没有结果的情况。对于这些样本，分析后将会被过滤掉，这是原始数据导致的问题，与算法本身无关。