(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117912552 B
(45)授权公告日 2025.03.28

(21)申请号 202410054361.8
(22)申请日 2023.01.07
(65)同一申请的已公布的文献号
申请公布号 CN 117912552 A
(43)申请公布日 2024.04.19
(62)分案原申请数据
202310021477.7 2023.01.07
(73)专利权人 杭州链康医学检验实验室有限公司
地址 310018 浙江省杭州市杭州经济技术开发区白杨街道6号大街260号1幢2楼西
(72)发明人 葛长利 韩斐然 郎秋蕾
(74)专利代理机构 杭州信与义专利代理有限公司 33450
专利代理师 万景旺
(51)Int.Cl.
G16B 20/30(2019.01)
G16B 30/10(2019.01)
G16B 25/10(2019.01)
G16B 5/00(2019.01)
G16B 40/00(2019.01)
(56)对比文件
CN 101180389 A,2008.05.14
CN 110191948 A,2019.08.30
审查员 胡佳

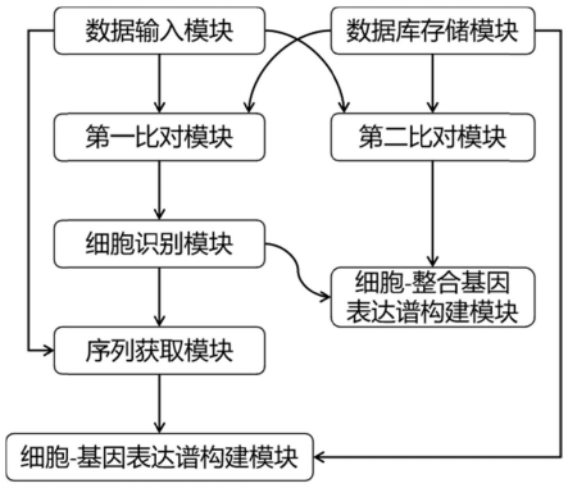
权利要求书2页 说明书12页 附图6页

(54)发明名称

一种PDX模型单细胞转录组数据的分析方法、设备和介质

(57)摘要

本发明公开了一种PDX模型单细胞转录组数据的分析方法、设备和介质,属于生物信息学技术领域。所述分析方法包括:将PDX模型的单细胞转录组测序数据与混合基因组库进行比对,获得细胞-基因表达谱矩阵;根据人或小鼠基因的比例识别细胞为人细胞、小鼠细胞或双细胞;基于细胞的barcode,从所述单细胞转录组测序数据中提取细胞序列,并与来源物种的参考基因组进行对比,获得相应的细胞-基因表达谱,可用于进行差异基因分析、功能富集分析等,获得更多的数据,为临床治疗癌症提供技术支持。进一步还可以基于整合基因序列集获得多物种的细胞-整合基因表达谱,进一步通过细胞聚类分析,可获知人和小鼠细胞的相互作用机制。



CN 117912552 B

1.一种PDX模型单细胞转录组数据的分析方法,其特征在于,包括以下步骤:

S1,将获得的PDX模型的单细胞转录组测序数据与人和小鼠的混合基因组库进行比对,获得基于所述混合基因组库的细胞-基因表达谱矩阵,其中,所述混合基因组库是将人和小鼠的参考基因组文件和基因注释文件进行合并得到的;

S2,根据细胞中表达人基因的比例或表达小鼠基因的比例是否大于等于第一预设阈值P1,识别细胞为人细胞、小鼠细胞或双细胞;

S3,基于细胞的barcode,从所述单细胞转录组测序数据中提取识别为人细胞的序列和识别为小鼠细胞的序列;

S4,将获得的人细胞的序列与人的参考基因组进行对比,并将获得的小鼠细胞的序列与小鼠的参考基因组进行比对,获得相应的细胞-基因表达谱,

S5,将所述单细胞转录组测序数据与人和小鼠同源基因的整合基因序列集进行比对,获得与整合基因的比对结果;

S6,基于步骤S2识别的人细胞和小鼠细胞barcode,从步骤S5得到的比对结果中获得细胞-整合基因表达谱,

其中,所述同源基因的整合基因序列集基于以下步骤得到:

(1)将人和小鼠的每个同源基因的序列进行拼接,得到每个同源基因的整合基因序列;人和小鼠的基因序列之间利用60~100个N碱基填充;

根据整合基因序列,构建对应的注释文件:将同源基因作为一条独立的染色体,来自人和小鼠的同源基因序列作为2个转录本,

P1的设置使得双细胞率与多细胞率的一半相差不超过5%,由下面公式计算双细胞率和多细胞率:

双细胞率=(双细胞数目/(人细胞数目+小鼠细胞数目+双细胞数目))×100%;

多细胞率=(捕获的细胞数目×$7.589\times10^{-6}+5.272\times10^{-4}$)×100%。

2.根据权利要求1所述的一种PDX模型单细胞转录组数据的分析方法,其特征在于,步骤S2具体包括:

S21,统计细胞中表达人基因的数目Nh以及表达小鼠基因的数目Nm;

S22,计算细胞中表达的人基因的比例Ph以及表达的小鼠基因的比例Pm,其中Ph=Nh/(Nh+Nm),Pm=Nm/(Nh+Nm);

S23,若Ph大于等于第一预设阈值P1,则细胞识别为人细胞,若Pm大于等于第一预设阈值P1,则细胞识别为小鼠细胞,其余为双细胞。

3.根据权利要求1所述的一种PDX模型单细胞转录组数据的分析方法,其特征在于,步骤S3中,所述提取识别为人细胞的序列和识别为小鼠细胞的序列具体包括:

S31,识别测序序列的barcode,与细胞的barcode进行对比,获得碱基匹配系数Mi,Mi=Lm/Lb,其中,Lm为测序序列的barcode与细胞的barcode匹配的碱基数目,Lb为细胞的barcode的碱基数目;

S32,根据Mi与第二预设阈值P2进行序列提取:

若Mi=100%,则直接提取对应的序列;若P2≤Mi<100%时,并且未完全匹配上的测序reads对应碱基的测序质量值<10,则将测序reads校正为正确的碱基后提取序列;若Mi<P2,则不提取,

其中,P2≥80%。

4.根据权利要求3所述的一种PDX模型单细胞转录组数据的分析方法,其特征在于,所述将测序reads校正为正确的碱基是指将测序序列的barcode中测序质量值<10的碱基校正为匹配的细胞的barcode对应位置的碱基。

5.根据权利要求1所述的一种PDX模型单细胞转录组数据的分析方法,其特征在于,在步骤S5和S6中,进一步包括比对结果过滤的步骤:

如果比对上一个整合基因的唯一位置,则保留对应的比对信息,提取一条比对信息;如果比对上同一个整合基因的多个位置,则只保留其中一条比对信息;如果比对上不同的整合基因,则过滤对应的比对信息。

6.一种计算机设备,其特征在于,包括:

存储器,用于存储计算机程序;

处理器,用于执行所述计算机程序时实现如权利要求1-5任一所述的一种PDX模型单细胞转录组数据的分析方法的步骤。

7.一种计算机可读存储介质,其特征在于,

所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时实现如权利要求1-5任一所述的一种PDX模型单细胞转录组数据的分析方法的步骤。

一种PDX模型单细胞转录组数据的分析方法、设备和介质

[0001] 相关专利

[0002] 本申请是申请号为2023100214777,申请日为2023年01月07日,发明名称为“基于PDX的单细胞转录组数据分析方法、系统、设备和介质”的中国发明专利申请的分案申请。

技术领域

[0003] 本发明属于生物数据处理技术领域,具体地,涉及一种PDX模型单细胞转录组数据的分析方法设备和介质。

背景技术

[0004] PDX(病人来源肿瘤异种移植,Patient-derived tumor xenograft)模型是将来源于患者的肿瘤组织或原代细胞移植到NSG(免疫缺陷)小鼠的体内而构建的肿瘤模型。该模型是将肿瘤组织直接移植到NSG小鼠体内,未经过任何人工培养,所以在组织病理学、分子生物学和基因水平上保留了大部分原代肿瘤的特点,与临床的相似度更高。PDX模型是目前为止最接近临床样本的肿瘤模型,这种模型对于临床肿瘤评估治疗和预后具有重要的意义。

[0005] 使用PDX模型进行单细胞转录组测序,可以深入分析肿瘤各个阶段在不同阶段下细胞类型以及基因表达特征,从而为肿瘤的治疗提供指导。目前,10×官方分析软件cellranger虽然可以针对PDX模型的多个物种混合库进行数据分析,并得到相应的表达谱矩阵和聚类结果。但是由于人和小鼠物种间存在大量的同源基因,即使来自人的细胞通过人和小鼠的混合基因组分析后,也会有部分reads可以比对上小鼠的基因组,直接基于该表达谱得到的细胞聚类结果以及下游数据分析挖掘得到的结果都会不准确。且由于人和小鼠基因组存在同源序列,也不能直接用比对上人基因组的序列进行分析,会导致部分来自小鼠的细胞最终被错误识别成人的细胞。

[0006] 另外,在PDX模型研究中,除了需要研究人相关的肿瘤细胞在不同治疗方案下基因表达变化的差异以寻找相应的药物治疗靶点等,还需要研究小鼠体内的细胞与导入的人细胞如何进行相互作用的。因此,如何从PDX模型得到的细胞测序数据中分离单个物种的细胞基因表达谱以及如何获得多个物种的细胞基因表达谱都是至关重要的。然而,目前还没有系统分析基于PDX模型的单细胞转录组的方法。

发明内容

[0007] 为了解决上述技术问题中的至少一个,本发明采用的技术方案如下:

[0008] 本发明第一方面提供一种基于PDX的单细胞转录组数据分析方法,包括以下步骤:

[0009] S1,将获得的PDX模型的单细胞转录组测序数据与人和小鼠的混合基因组库进行比对,获得基于所述混合基因组库的细胞-基因表达谱矩阵,其中,所述混合基因组库是将人和小鼠的参考基因组文件和基因注释文件进行合并得到的;

[0010] S2,根据细胞中表达人基因的比例或表达小鼠基因的比例是否大于等于第一预设

阈值P1,识别细胞为人细胞、小鼠细胞或双细胞;

[0011] S3,基于细胞的barcode,从所述单细胞转录组测序数据中提取识别为人细胞的序列和识别为小鼠细胞的序列;

[0012] S4,将获得的人细胞的序列与人的参考基因组进行对比,并将获得的小鼠细胞的序列与小鼠的参考基因组进行比对,获得相应的细胞-基因表达谱,

[0013] 其中,P1的设置使得双细胞率与多细胞率的一半相差不超过5%,由下面公式计算双细胞率和多细胞率:

[0014] 双细胞率=(双细胞数目/(人细胞数目+小鼠细胞数目+双细胞数目))×100%;

[0015] 多细胞率=(捕获的细胞数目×7.589×10^{-6}+5.272×10^{-4})×100%。

[0016] 由于多细胞率除双细胞(人和小鼠细胞混合)外,还包括人和人细胞、小鼠和小鼠细胞的混合,因此,理论上,双细胞率等于多细胞率的1/2。在本发明中,所述双细胞率与多细胞率的一半相差不超过5%,是指:|(双细胞率-多细胞率/2)|/(多细胞率/2)×100%≤5%。如果P1设置过高,会导致过多的细胞被判断为双细胞,与实际情况不符。

[0017] 在本发明的一些具体实施方案中,P1=70%。

[0018] 在本发明的一些实施方案中,步骤S1中,将人和小鼠的基因组文件和基因注释文件进行合并时,为了避免基因和染色体重复,分别在基因ID、基因名、染色体前加上特异性标签进行区分。例如在人的基因ID、基因名、染色体前加上“human”,在小鼠的基因ID、基因名、染色体前加上“mouse”。进一步地,基于合并的基因组文件和基因注释文件生成可用于比对的库文件。

[0019] 在本发明的一些实施方案中,步骤S2具体包括:

[0020] S21,统计细胞中表达人基因的数目Nh以及表达小鼠基因的数目Nm;

[0021] S22,计算细胞中表达的人基因的比例Ph以及表达的小鼠基因的比例Pm,其中Ph=Nh/(Nh+Nm),Pm=Nm/(Nh+Nm);

[0022] S23,若Ph大于等于第一预设阈值P1,则细胞识别为人细胞,若Pm大于等于第一预设阈值P1,则细胞识别为小鼠细胞。

[0023] 在本发明的一些实施方案中,对于既不满足Ph大于等于第一预设阈值P1也不满足Pm大于等于第一预设阈值P1的细胞,判定为双细胞,即既具有人基因表达又具有小鼠基因表达的细胞。

[0024] barcode也叫index,即条形码或称标签,在测序技术中通常用于区分序列的不同来源。在本发明中,barcode用于区别不同的细胞,即测序结果中具有相同的barcode的测序序列意味着来自同一细胞,从而不同的barcode可以代表不同的细胞。在本发明的一些描述中,barcode和细胞具有相同的含义。

[0025] 在本发明的一些实施方案中,步骤S3中,所述提取识别为人细胞的序列和识别为小鼠细胞的序列具体包括:

[0026] S31,识别测序序列的barcode,与细胞的barcode进行对比,获得碱基匹配系数Mi,Mi=Lm/Lb,其中,Lm为测序序列的barcode与细胞的barcode匹配的碱基数目,Lb为细胞的barcode的碱基数目;

[0027] S32,根据Mi与第二预设阈值P2进行序列提取:

[0028] 若Mi=100%,则直接提取对应的序列;若P2≤Mi<100%时,并且未完全匹配上的

测序reads对应碱基的测序质量值<10,则将测序reads校正为正确的碱基后提取序列;若Mi<P2,则不提取,

[0029] 其中,P2≥80%。

[0030] 利用步骤S2识别出人细胞和小鼠细胞后,虽然也可以根据步骤S1的比对结果获得人细胞或小鼠细胞相应的细胞-基因表达谱。然而,步骤S1的比对是基于所述混合基因组库进行比对的,受同源基因的影响,每个细胞的基因表达谱可能并不准确。因此,基于步骤S3提取来自人和小鼠细胞原始测序数据,可进一步分别各自与人和小鼠的参考基因组进行比对,获得正确的人或小鼠的细胞-基因表达谱矩阵。

[0031] 不同测序平台通常具有不同的barcode长度,例如Illumina 10×单细胞测序平台,barcode长度为16,墨卓平台的barcode长度为28。根据不同的barcode长度选择不同的P2值,通常选择的标准是仅接受1~2个碱基错配。在本发明的一些实施方案中,barcode长度为16,P2设为90%,仅允许1个碱基无法匹配。

[0032] 在本发明的一些实施方案中,步骤S32中,所述将测序reads校正为正确的碱基是指将测序序列的barcode中测序质量值<10的碱基校正为匹配的细胞的barcode对应位置的碱基。如果测序质量会上≥10,则不能进行校正,该测序reads不属于该细胞,应当利用其他的barcode进行提取或者舍弃。

[0033] 在本发明的一些实施方案中,步骤S4中,在构建细胞-基因表达谱矩阵时,仅需要执行基因组比对、UMI校正即可,无需再进行细胞识别的过程。

[0034] 在本发明的一些实施方案中,在步骤S2之后,进行如下步骤:

[0035] S3',将所述单细胞转录组测序数据与人和小鼠同源基因的整合基因序列集进行比对,获得与整合基因的比对结果;

[0036] S4',基于步骤S2识别的人细胞和小鼠细胞barcode,从步骤S3'得到的比对结果中获得细胞-整合基因表达谱,

[0037] 其中,所述同源基因的整合基因序列集基于以下步骤得到:

[0038] (1)将人和小鼠的每个同源基因的序列进行拼接,得到每个同源基因的整合基因序列;人和小鼠的基因序列之间利用60~100个N碱基填充;

[0039] (2)根据整合基因序列,构建对应的注释文件:将同源基因作为一条独立的染色体,来自人和小鼠的同源基因序列作为2个转录本。

[0040] 在本发明的一些实施方案中,整合基因序列中设置60~100个N碱基是为了在整理的注释文件中加入物种信息,以便于后续比对reads过滤。在本发明的一些具体实施方案中,整合基因序列中设置80个N碱基。即在一条整合基因序列中,在来自人的同源基因序列和来自小鼠的同源基因序列之间插入80个N碱基。另一方面,通过设计N碱基,也可能防止后续对比过程中reads跨物种进行比对,即某reads一部分比对到人的同源基因,一部分比对到小鼠的同源基因。

[0041] 在本发明的一些实施方案中,在步骤S3'和S4'中,进一步包括比对结果过滤的步骤:

[0042] 如果比对上一个整合基因的唯一位置,则保留对应的比对信息,提取一条比对信息;如果比对上同一个整合基因的多个位置则只保留其中一条比对信息;如果比对上不同的整合基因,则过滤对应的比对信息。

[0043] 本发明第二方面提供一种基于PDX的单细胞转录组数据分析系统,包括:

[0044] 数据输入模块,用于获得PDX模型的单细胞转录组测序数据;

[0045] 数据库存储模块,用于存储人参考基因组、小鼠参考基因组以及人和小鼠的混合基因组库,其中,所述混合基因组库是将人和小鼠的参考基因组文件和基因注释文件进行合并得到的;

[0046] 第一比对模块,分别与所述数据输入模块和所述数据库存储模块连接,用于将所述单细胞转录组测序数据与所述混合基因组库进行比对;

[0047] 细胞识别模块,与所述第一比对模块连接,用于根据细胞中表达人基因的比例或表达小鼠基因的比例是否大于等于第一预设阈值P1,识别细胞为人细胞、小鼠细胞或双细胞;

[0048] 序列获取模块,分别与所述细胞识别模块和所述数据输入模块连接,用于基于细胞的barcode,从所述单细胞转录组测序数据中提取识别为人细胞的序列和识别为小鼠细胞的序列;

[0049] 细胞-基因表达谱构建模块,分别与所述序列获取模块和所述数据库存储模块连接,用于将获得的人细胞的序列与人的参考基因组进行对比,并将获得的小鼠细胞的序列与小鼠的参考基因组进行比对,获得相应的细胞-基因表达谱,

[0050] 其中,P1的设置使得双细胞率与多细胞率的一半相差不超过5%,由下面公式计算双细胞率和多细胞率:

[0051] 双细胞率=(双细胞数目/(人细胞数目+小鼠细胞数目+双细胞数目))×100%;

[0052] 多细胞率=(捕获的细胞数目$\times 7.589\times 10^{-6}+5.272\times 10^{-4}$)×100%。

[0053] 由于多细胞率除双细胞(人和小鼠细胞混合)外,还包括人和人细胞、小鼠和小鼠细胞的混合,因此,理论上,双细胞率等于多细胞率的1/2。在本发明中,所述双细胞率与多细胞率的一半相差不超过5%,是指:|(双细胞率-多细胞率/2)|/(多细胞率/2)×100%≤5%。

[0054] 在本发明的一些具体实施方案中,P1=70%。

[0055] 在本发明的一些实施方案中,所述数据库存储模块中,在将人和小鼠的基因组文件和基因注释文件进行合并时,为了避免基因和染色体重复,分别在基因ID、基因名、染色体前加上特异性标签进行区分。例如在人的基因ID、基因名、染色体前加上“human”,在小鼠的基因ID、基因名、染色体前加上“mouse”。进一步地,基于合并的基因组文件和基因注释文件生成可用于比对的库文件。

[0056] 在本发明的一些实施方案中,所述细胞识别模块基于以及步骤进行细胞识别:

[0057] 统计细胞中表达人基因的数目Nh以及表达小鼠基因的数目Nm;

[0058] 计算细胞中表达的人基因的比例Ph以及表达的小鼠基因的比例Pm,其中Ph=Nh/(Nh+Nm),Pm=Nm/(Nh+Nm);

[0059] 若Ph大于等于第一预设阈值P1,则细胞识别为人细胞,若Pm大于等于第一预设阈值P1,则细胞识别为小鼠细胞,其余为双细胞。

[0060] 在本发明的一些实施方案中,对于既不满足Ph大于等于第一预设阈值P1也不满足Pm大于等于第一预设阈值P1的细胞,判定为双细胞,即既具有人基因表达又具有小鼠基因表达的细胞。

[0061] 在本发明的一些实施方案中,所述序列获取模块通过以下步骤提取识别为人细胞的序列和识别为小鼠细胞的序列:

[0062] 识别测序序列的barcode,与细胞的barcode进行对比,获得碱基匹配系数Mi,Mi=Lm/Lb,其中,Lm为测序序列的barcode与细胞的barcode匹配的碱基数目,Lb为细胞的barcode的碱基数目;

[0063] 根据Mi与第二预设阈值P2进行序列提取:

[0064] 若Mi=100%,则直接提取对应的序列;若P2≤Mi<100%时,并且未完全匹配上的测序reads对应碱基的测序质量值<10,则将测序reads校正为正确的碱基后提取序列;若Mi<P2,则不提取,

[0065] 其中,P2≥80%。

[0066] 在本发明的一些实施方案中,所述数据库存储模块还用于存储人和小鼠同源基因的整合基因序列集,所述系统还包括或者不包括所述序列获取模块和所述细胞-基因表达谱构建模块但包括:

[0067] 第二比对模块,分别与所述数据输入模块和所述数据库存储模块连接,用于将所述单细胞转录组测序数据与人和小鼠同源基因的整合基因序列集进行比对,获得与整合基因的比对结果;

[0068] 细胞-整合基因表达谱构建模块,分别与所述细胞识别模块和所述第二比对模块连接,用于基于所述细胞识别模块识别的人细胞和小鼠细胞barcode,从所述第二比对模块得到的比对结果中获得细胞-整合基因表达谱。

[0069] 其中,所述同源基因的整合基因序列集基于以下步骤得到:

[0070] (1)将人和小鼠的每个同源基因的序列进行拼接,得到每个同源基因的整合基因序列;人和小鼠的基因序列之间利用60~100个N碱基填充;

[0071] (2)根据整合基因序列,构建对应的注释文件:将同源基因作为一条独立的染色体,来自人和小鼠的同源基因序列作为2个转录本。

[0072] 本发明第三方面提供一种计算机设备,包括:

[0073] 存储器,用于存储计算机程序;

[0074] 处理器,用于执行所述计算机程序时实现如本发明第一方面任一所述的一种基于PDX的单细胞转录组数据分析方法的步骤。

[0075] 本发明第四方面提供一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时实现如本发明第一方面任一所述的一种基于PDX的单细胞转录组数据分析方法的步骤。

[0076] 本发明的有益效果

[0077] 相对于现有技术,本发明取得了以下有益效果:

[0078] 利用本发明的方法和系统可以构建PDS小鼠模型处理(比如药物或其他治疗方式)前后的样本的人和/或小鼠的细胞-基因表达谱,进一步可以进行差异基因分析和功能富集分析等,获得更多的数据,为临床治疗癌症提供技术支持。

[0079] 利用本发明的方法和系统获得细胞-整合基因的表达谱,可以进行下游的细胞聚类及差异基因寻找分析。通过这种方式获得的聚类结果,人和小鼠的细胞也能很好地聚类在一起,可用于获知人和小鼠细胞的相互作用机制。

附图说明

[0080] 图1示出了捕获的细胞数目与多细胞率关系图。

[0081] 图2示出了本发明实施例1中PDX模型单细胞转录组测序结果示意图。

[0082] 图3示出了本发明实施例1中基于混合库分析的聚类结果。

[0083] 图4示出了本发明实施例1中基于提取人的细胞对应的测序数据进行聚类分析的结果。

[0084] 图5示出了本发明实施例1中基于混合库分析和基于提取人的细胞对应的测序数据分析得到的基因数目。

[0085] 图6示出了利用基于混合数据库构建的细胞-基因表达谱得到的聚类结果。

[0086] 图7示出了人和小鼠的部分同源基因信息。

[0087] 图8示出了小鼠Cd3d基因序列。

[0088] 图9示出了人和小鼠同源Cd3d基因序列拼接得到的整合基因序列。

[0089] 图10示出了人和小鼠同源Cd3d基因序列拼接得到的整合基因序列的注释文件。

[0090] 图11示出了本发明实施例2中基于细胞-整合基因的表达谱矩阵聚类分析的结果。

[0091] 图12示出了本发明实施例3中单物种的细胞-基因表达谱构建系统示意图。

[0092] 图13示出了本发明实施例4中多物种的细胞-基因表达谱构建系统示意图。

[0093] 图14示出了本发明实施例5中PDX模型的单细胞转录组测序数据分析系统示意图。

具体实施方式

[0094] 除非另有说明、从上下文暗示或属于现有技术的惯例,否则本申请中所有的份数和百分比都基于重量,且所用的测试和表征方法都是与本申请的提交日期同步的。在适用的情况下,本申请中涉及的任何专利、专利申请或公开的内容全部结合于此作为参考,且其等价的同族专利也引入作为参考,特别这些文献所披露的关于本领域中的相关术语的定义。如果现有技术中披露的具体术语的定义与本申请中提供的任何定义不一致,则以本申请中提供的术语定义为准。

[0095] 为了使本发明所解决的技术问题、技术方案及有益效果更加清楚明白,以下结合实施例,对本发明进行进一步详细说明。

[0096] 实施例

[0097] 以下例子在此用于示范本发明的优选实施方案。本领域内的技术人员会明白,下述例子中披露的技术代表发明人发现的可以用于实施本发明的技术,因此可以视为实施本发明的优选方案。但是本领域内的技术人员根据本说明书应该明白,这里所公开的特定实施例可以做很多修改,仍然能得到相同的或者类似的结果,而非背离本发明的精神或范围。

[0098] 除非另有定义,所有在此使用的技术和科学的术语,和本发明所属领域内的技术人员所通常理解的意思相同,在此公开引用及他们引用的材料都将以引用的方式被并入。

[0099] 下述实施例中的实验方法,如无特殊说明,均为常规方法。下述实施例中所用的仪器设备,如无特殊说明,均为实验室常规仪器设备;下述实施例中所用的试验材料,如无特殊说明,均为自常规生化试剂商店购买得到的。

[0100] 实施例1基于PDX模型的单细胞转录组数据构建单个物种的细胞-基因表达谱

[0101] 本实施例提供基于PDX模型的单细胞转录组数据构建单个物种的细胞-基因表达

谱的方法,包括以下步骤:

[0102] 1.获得PDX模型的单细胞转录组数据

[0103] 将目标组织解离获得单个细胞悬液,然后将含有barcode信息的凝胶珠与细胞和酶的混合物结合,然后被位于微流体系统中的油表面活性剂液滴包裹,形成GEMs(Gel Beads-In-Emulsions)。GEM内的凝胶珠发生溶解,细胞裂解释放mRNA,通过逆转录产生用于测序的带条形码的cDNA,液体油层破坏后,进行cDNA扩增、纯化和文库构建。对完成的文库进行测序,获得单细胞转录组测序数据。

[0104] 2.获得基于小鼠和人混合数据库的细胞-基因表达谱

[0105] 构建人和小鼠的混合基因组库:将人和小鼠的基因组文件(GRCh38和GRCm39)和基因注释文件(ensemble V105版本:Homo_sapiens.GRCh38.105.chr.gtf.gz,Mus_musculus.GRCm39.105.chr.gtf.gz)进行合并,为避免基因和染色体重复,分别在基因I D、基因名、染色体前加上“human”和“mouse”进行区分。然后,基于合并的基因组文件和基因注释文件利用cellranger软件生成可用于star比对的库文件。

[0106] 针对步骤1获得的PDX模型的单细胞转录组数据,利用cellranger软件进行以下过程:

[0107] (1)barcode提取校正

[0108] 将测序R1端的前16bp的序列与10×白名单进行比对,如果序列来自白名单则相应的双端的reads保留,如果测序得到的reads仅有1个碱基与白名单不一样,且不一样位置的测序碱基质量值<10,则将错误匹配的碱基按照白名单进行修改并保留reads,否则该reads不纳入后续的表达谱矩阵计算。

[0109] (2)序列基因组比对

[0110] 将测序的R2 reads,使用star软件与参考基因组进行比对,获得比对信息,并将仅唯一匹配上正义转录本的reads保留纳入统计barcode-基因表达矩阵。

[0111] (3)细胞识别

[0112] 根据barcode-基因表达矩阵,统计每个barcode所有测到的UMI count,并从高到低进行排序。对于高RNA含量的细胞识别来讲,以捕获10000个细胞为例,如果bar code的UMI count数大于第101个barcode的UMI Count值的1/10则认为是细胞。有些低RNA含量的细胞可能无法满足刚才识别的条件,选出一个表达量较低的barco de作为背景,统计该barcode基因表达情况,然后将第一步未识别为细胞的barcode对应的基因表达情况与背景作对比,如果基因表达情况一致则认为是背景,否则认为该barcode为细胞。

[0113] 由此,获得基于混合基因组库的细胞-基因表达谱矩阵。

[0114] 具体的表达谱数据格式如下:

[0115]

基因	barcode 1	barcode 2	barcode 3	barcode 4	……	barcode n
human_TNFRSF4	0	1	0	0	……	2
human_CD3E	1	0	0	3	……	0
……	……	……	……	……	……	……
mouse_Gm28417	0	0	1	0		0

[0116] 3.识别人细胞和小鼠细胞

[0117] 根据基因名前缀统计每个细胞(barcode)表达human和mouse的基因数目,分别记

为Nh和Nm。计算每个barcode中来自人和小鼠基因的比例,分别记为Ph和Pm,其中,Ph=Nh/(Nh+Nm),Pm=Nm/(Nh+Nm)。

[0118] 由于使用的是人和小鼠的混合基因组库进行的比对,而人和小鼠存在大量的同源基因,即使是来自人的细胞也可能少量表达小鼠的基因。如果某个barcode的Ph>70%,则该barcode来自人的细胞,如果某个barcode的Pm>70%,则该barcode来自小鼠的细胞,其余的barcode则认为是双细胞。这里的“70%”为识别barcode来自人或小鼠的细胞所设定的阈值。

[0119] 在10×单细胞测序过程中,多细胞率与捕获的细胞数目有关,如图1所示,拟合得到多细胞率与捕获的细胞数目的关系如下面公式所示:

[0120] 多细胞率=(捕获的细胞数目×7.589×10^{-6}+$5.272\times10^{-4)}$×100%

[0121] 然而,上述多细胞中,除了人和小鼠的细胞混在一个油滴(双细胞),还存在人和人的细胞、小鼠和小鼠的细胞混在一起的情况。因此,理论上,双细胞率应该是多细胞率的一半,即((捕获的细胞数目×7.589×10^{-6}+5.272×10^{-4})/2)。发明人发现当阈值为70%时,得到的双细胞率十分接近多细胞率的一半。

[0122] 具体得到的统计表格如下表所示(展示部分):

[0123]

barcode	Nh	Nm	Ph(%)	Pm(%)	细胞类型
AAACCTGAGCTAACAA	1871	61	96.84	3.16	人细胞
AAACCTGAGTTGCAGG	645	42	93.89	6.11	人细胞
AAGGCAGAGCACACAG	897	1284	41.13	58.87	双细胞
AAGGCAGGTATCTGCA	102	1053	8.83	91.17	小鼠细胞
ACGAGCCGTAGCCTAT	1082	192	84.93	15.07	人细胞
……	……	……	……	……	……
TATGCCCCAGCATGAG	321	1673	16.10	83.90	小鼠细胞
TGCACCTTCAATCACG	1295	825	61.08	38.92	双细胞

[0124] 受限于PDX模型、样本取样部位、取样大小、单细胞解离、细胞捕获效率等影响,不同的样本实际分析得到的来自人和小鼠的细胞比例并不完全一致。下表示出了来自两个不同PDX样本得到的来自人和小鼠的细胞实际数目及占比情况:

[0125]

样本	人细胞个数	小鼠细胞个数	双细胞个数	人细胞率(%)	小鼠细胞率(%)	双细胞率(%)	细胞捕获数目	多细胞率(%)
样本1	2869	5385	272	33.65	63.16	3.19	8565	6.55
样本2	5212	3146	281	60.33	36.42	3.25	8639	6.61

[0126] 由上表可知,样本1捕获的细胞数目(人细胞个数+小鼠细胞个数+双细胞个数)共计8565个,计算得到多细胞率为(8565×7.589×10^{-6}+5.272×10^{-4})×100%=6.55%,当阈值设为70%时,得到的双细胞率为3.19%,与多细胞率的一半3.28%相差仅|3.19-3.28)/3.28×100%=2.7%;样本2捕获的细胞数目(人细胞个数+小鼠细胞个数+双细胞个数)共计8639个,计算得到多细胞率为(8636×7.589×10^{-6}+5.272×10^{-4})×100%=6.61%,当阈值设为70%时,得到的双细胞率为3.25%,与多细胞率的一半3.31%相差仅|3.25-3.31)/

3.31×100%＝1.8%。可见,利用上述方法识别细胞更加精准。

[0127] 4.序列提取

[0128] 分别从原始测序数据中提取上述判定的来自人和小鼠细胞的barcode对应的序列信息。根据单细胞转录组的文库结构可知,barcode序列来自双端测序的R1端的前n个碱基(比如10×单细胞平台,n＝16,如图2所示,国内的单细胞平台对应的细胞b arcode更长,比如墨卓平台的细胞barcode长度为28)。具体的序列提取方式如下:

[0129] 1)计算每条序列对应的碱基匹配系数:细胞barcode长度记为Lb,实际匹配上的碱基长度记为Lm,碱基匹配系数为Mi,Mi＝Lm/Lb。

[0130] 2)考虑到不同单细胞平台设置的细胞barcode长度不同,因此设置碱基匹配系数阈值为90%,即当细胞barcode为16个碱基时,仅允许1个碱基无法匹配。当Mi＝100%,直接输出对应的reads pair。当90%≤Mi<100%时,计算未完全匹配上的测序re ads对应碱基的测序质量值,如果该位置的测序质量值<10,考虑是测序错误导致的无法匹配,将测序reads校正为正确的碱基,即将测序reads校正为已经识别为细胞的b arcode后输出对应的reads pair。

[0131] 5.重新获得单个物种的细胞-基因表达谱矩阵

[0132] 虽然基于步骤3就可以直接得到来自人的细胞对应的表达谱矩阵,但由于该细胞-基因表达谱是基于人和小鼠混合库得到的,受同源基因的影响,每个细胞的基因表达谱可能并不准确。

[0133] 因此,发明人基于步骤4得到的来自人和小鼠的细胞的原始数据,重新以人或小鼠的参考基因组作为库文件进行分析,获得正确的人和小鼠的细胞-基因表达谱。

[0134] 由于步骤4提取的序列全部来自实际的细胞,与正常单细胞转录组测序的情况下还可能存在空载的油包水数据不同,在生成表达谱矩阵时,仅需要执行基因组比对、UMI校正即可,无需再进行细胞识别。如果执行了细胞识别,就很有可能把相同的一类细胞作为背景barcode去除掉,会影响后续的数据挖掘。

[0135] 利用上述方法,对于PDX小鼠模型,可以在处理前后分别获得样本中来自人或小鼠细胞的基因表达谱矩阵,进一步可以进行差异基因分析和功能富集分析等数据挖掘。

[0136] 图3展示了基于混合库分析的聚类结果,可见人和小鼠的细胞分为左右两侧。提取来自人的细胞对应的测序数据,进行聚类分析的结果如图4所示。

[0137] 下表示出了基于混合库分析的和提取来自人细胞原始数据分析得到的差异基因结果:

[0138]

基于混合库	簇	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	上调基因数量	1704	1655	1744	1698	2692	2393	1648	1238	2719	2438	877	2746	782	1839

提取来自人细胞原始数据	簇	0	1	2	3	4	5	6	7
	上调基因数量	22	28	29	62	99	174	106	50

[0139] 从统计结果发现,基于混合库得到的差异上调的基因特别多,由于里面同时存在来自人细胞和来自小鼠细胞,以所有的细胞做差异可能会导致差异基因偏多。因此仅对来自人的cluster 3、4、5、8、9、10、11这个7个cluster进行差异分析,分析结果发现,共找到5066个差异基因,其中有1209个差异基因来自小鼠的基因,这种明显是由于混合库导致的不符合实际情况的差异基因。

[0140] 对于同一个barcode(来自人细胞),使用混合库分析,由于比对上小鼠基因组的缘故,会导致基因数偏高,如图5所示,图上为混合库分析,图下为提取reads使用人的库分析的统计结果。

[0141] 实施例2多物种细胞-整合基因表达谱构建

[0142] 实施例1仅是针对单个物种的细胞表达谱进行分析,在实际应用中会发现,有些PDX模型仅导入了肿瘤原代细胞,如果想获知小鼠的免疫细胞是如何对肿瘤细胞发挥作用的,以及细胞之间是如何相互作用的等信息,仍需要生成基于混合数据的表达谱矩阵。基于实施例1步骤1获得的小鼠和人混合数据库的细胞-基因表达谱数据,由于不同物种间的基因名不同,在细胞聚类时,会导致出现明显分成2个物种群的情况(如图6)。

[0143] 针对这种情况,使用以下步骤获得校正后的多物种细胞-基因表达谱矩阵:

[0144] (1)从Ensemble数据库中获得人、小鼠两个物种的同源基因信息,如图7所示,前两列分别为人基因ID和基因name信息,后两列为对应的小鼠基因ID和基因name信息。人和小鼠共有18973(以人的基因为准,共有18973个基因ID可以找到来自小鼠的同源基因)个同源基因。

[0145] (2)根据基因组注释文件中的基因染色体位置信息以及基因正负链信息,分别从人、小鼠基因组中提取每个基因对应的序列信息。以Cd3d基因为例,提取后的序列信息如图8所示。

[0146] (3)根据步骤(1)获得的同源基因信息对序列进行合并(两个物种基因之间用80个N碱基填充,目的是在整理的注释文件中加入物种信息),得到整合基因。以Cd3d基因为例,其整合基因序列如图9所示。

[0147] (4)根据整合基因序列,构建对应的注释文件(同源基因作为一条独立的染色体,来自两个物种的同源基因作为2个转录本),以merge0000001为例,如图10所示。

[0148] (5)将测序得到的reads以整合基因序列集作为参考基因组,进行比对分析获得与整合基因序列的比对结果。

[0149] (6)比对结果过滤:对于比对上一个整合基因的唯一位置(且比对上的序列仅来自同一个物种,即不跨越N碱基比对)上的,保留对应的比对信息,提取一条reads多重比对的信息;如果比对上同一个整合基因的多个位置(且比对上的多个位置均来自同一个物种,即不跨越N碱基比对),保留其中一条比对信息;如果比对上不同的整合基因,则过滤掉对应的比对信息。本步骤共获得40%左右的比对信息。

[0150] (7)根据实施例1获得的人和小鼠的barcode信息(去除双细胞信息),从步骤(6)得到的过滤后的比对信息中获得细胞-整合基因的表达谱矩阵。

[0151] 基于细胞-整合基因的表达谱矩阵,可以进行下游的细胞聚类及差异基因寻找分析。通过这种方式获得的聚类结果,人和小鼠的细胞也能很好地聚类在一起(如图11所示),可用于获知人和小鼠细胞的相互作用机制。

[0152] 实施例3单物种的细胞-基因表达谱构建系统

[0153] 本实施例提供一种基于PDX模型的单细胞转录组数据构建单物种的细胞-基因表达谱的系统,如图12所示,包括:

[0154] 数据输入模块,用于获得PDX模型的单细胞转录组测序数据;

[0155] 数据库存储模块,用于存储人参考基因组、小鼠参考基因组以及人和小鼠的混合基因组库,其中,该混合基因组库是将人和小鼠的参考基因组文件和基因注释文件进行合并得到的,如实施例1所描述的方法;

[0156] 第一比对模块,分别与数据输入模块和数据库存储模块连接,用于将单细胞转录组测序数据与混合基因组库进行比对;

[0157] 细胞识别模块,与第一比对模块连接,用于根据细胞中表达人基因的比例或表达小鼠基因的比例是否大于等于70%,识别细胞为人细胞或小鼠细胞;

[0158] 序列获取模块,分别与细胞识别模块和数据输入模块连接,用于基于细胞的barcode,从单细胞转录组测序数据中提取识别为人细胞的序列和识别为小鼠细胞的序列;

[0159] 细胞-基因表达谱构建模块,分别与序列获取模块和数据库存储模块连接,用于将获得的人细胞的序列与人的参考基因组进行对比,并将获得的小鼠细胞的序列与小鼠的参考基因组进行比对,获得相应的细胞-基因表达谱。

[0160] 实施例4多物种的细胞-整合基因表达谱构建系统

[0161] 本实施例基于实施例3提供一种多物种的细胞-整合基因表达谱构建系统,如图13所示,包括:

[0162] 数据输入模块,用于获得PDX模型的单细胞转录组测序数据;

[0163] 数据库存储模块,用于存储人参考基因组、小鼠参考基因组以及人和小鼠的混合基因组库,其中,该混合基因组库是将人和小鼠的参考基因组文件和基因注释文件进行合并得到的,如实施例1所描述的方法;

[0164] 第一比对模块,分别与数据输入模块和数据库存储模块连接,用于将单细胞转录组测序数据与混合基因组库进行比对;

[0165] 细胞识别模块,与第一比对模块连接,用于根据细胞中表达人基因的比例或表达小鼠基因的比例是否大于等于70%,识别细胞为人细胞、小鼠细胞或双细胞;

[0166] 第二比对模块,分别与数据输入模块和数据库存储模块连接,用于将单细胞转录组测序数据与人和小鼠同源基因的整合基因序列集进行比对,获得与整合基因的比对结

果,其中,整合基因序列集的构建方法参考实施例2。

[0167] 细胞-整合基因表达谱构建模块,分别与细胞识别模块和第二比对模块连接,用于基于细胞识别模块识别的人细胞和小鼠细胞barcode,从第二比对模块得到的比对结果中获得细胞-整合基因表达谱。

[0168] 实施例5PDX模型的单细胞转录组测序数据分析系统

[0169] 本实施例提供PDX模型的单细胞转录组测序数据分析系统,结合的实施例3和4的系统,如图14所示,包括:

[0170] 数据输入模块,用于获得PDX模型的单细胞转录组测序数据;

[0171] 数据库存储模块,用于存储人参考基因组、小鼠参考基因组以及人和小鼠的混合基因组库,其中,该混合基因组库是将人和小鼠的参考基因组文件和基因注释文件进行合并得到的,如实施例1所描述的方法;

[0172] 第一比对模块,分别与数据输入模块和数据库存储模块连接,用于将单细胞转录组测序数据与混合基因组库进行比对;

[0173] 细胞识别模块,与第一比对模块连接,用于根据细胞中表达人基因的比例或表达小鼠基因的比例是否大于等于70%,识别细胞为人细胞、小鼠细胞或双细胞;

[0174] 序列获取模块,分别与细胞识别模块和数据输入模块连接,用于基于细胞的barcode,从单细胞转录组测序数据中提取识别为人细胞的序列和识别为小鼠细胞的序列;

[0175] 细胞-基因表达谱构建模块,分别与序列获取模块和数据库存储模块连接,用于将获得的人细胞的序列与人的参考基因组进行对比,并将获得的小鼠细胞的序列与小鼠的参考基因组进行比对,获得相应的细胞-基因表达谱,

[0176] 还包括:

[0177] 第二比对模块,分别与数据输入模块和数据库存储模块连接,用于将单细胞转录组测序数据与人和小鼠同源基因的整合基因序列集进行比对,获得与整合基因的比对结果,其中,整合基因序列集的构建方法参考实施例2。

[0178] 细胞-整合基因表达谱构建模块,分别与细胞识别模块和第二比对模块连接,用于基于细胞识别模块识别的人细胞和小鼠细胞barcode,从第二比对模块得到的比对结果中获得细胞-整合基因表达谱。

[0179] 在本发明提及的所有文献都在本申请中引用作为参考,就如同每一篇文献被单独引用作为参考那样。此外应理解,在阅读了本发明的上述讲授内容之后,本领域技术人员可以对本发明作各种改动或修改,这些等价形式同样落于本申请所附权利要求书所限定的范围。

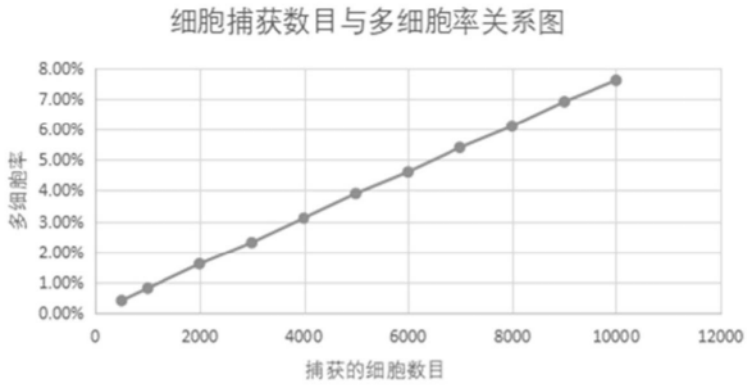

细胞捕获数目与多细胞率关系图
多细胞率
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
0
2000
4000
6000
8000
10000
12000
捕获的细胞数目

图1


@A01890:81:H27CLDSX5:2:1101:11948:1016 1:N:0:TCGGCTCT+AGACCATC
CACTCCATCCAGTATGTTATGAAGAT
+
FFFFFFFFFFFFFFFFFFFFFFFFF
@A01890:81:H27CLDSX5:2:1101:12057:1016 1:N:0:TCGGCTCT+AGACCATC
GGGCATCGTAAACCTCGTACGCTCAA
+
FFFFFFFFFFFFFFFFFF:FFFFFF
barcode序列
测序质量值

图2

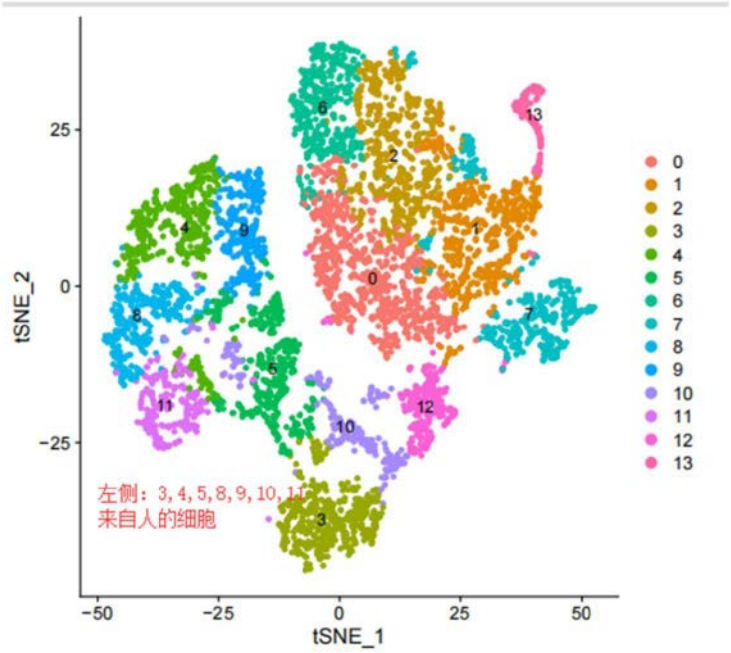

tSNE_2
tSNE_1
25
0
-25
-50
-25
0
25
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
左侧：3,4,5,8,9,10,11
来自人的细胞

图3

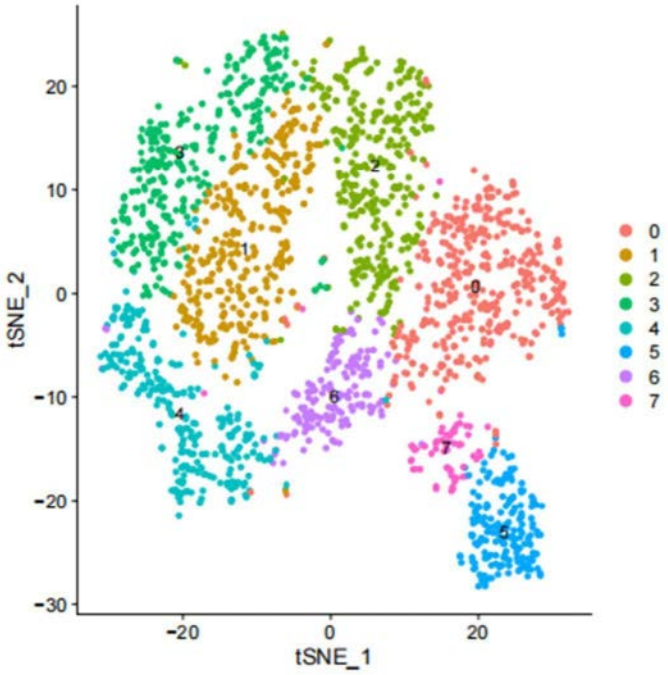


图4

	orig.ident	nCount_RNA	nFeature_RNA
AAACCCAGTAGTATAG	hgmm_5k	41037	7339
AAACGAAGTCAAAGAT	hgmm_5k	52028	8158
AAACGAATCAAGTTGC	hgmm_5k	24063	5134
AAACGAATCGTGCATA	hgmm_5k	23337	5547
AAACGCTAGATAACGT	hgmm_5k	61716	8395
AAACGCTCAGACCCGT	hgmm_5k	16304	4657

	orig.ident	nCount_RNA	nFeature_RNA
AAACCCAGTAGTATAG	hg_hgmm_5k_2	40579	7006
AAACGAAGTCAAAGAT	hg_hgmm_5k_2	51371	7737
AAACGAATCAAGTTGC	hg_hgmm_5k_2	23812	4989
AAACGAATCGTGCATA	hg_hgmm_5k_2	23126	5383
AAACGCTAGATAACGT	hg_hgmm_5k_2	61117	7935
AAACGCTCAGACCCGT	hg_hgmm_5k_2	16082	4519

图5

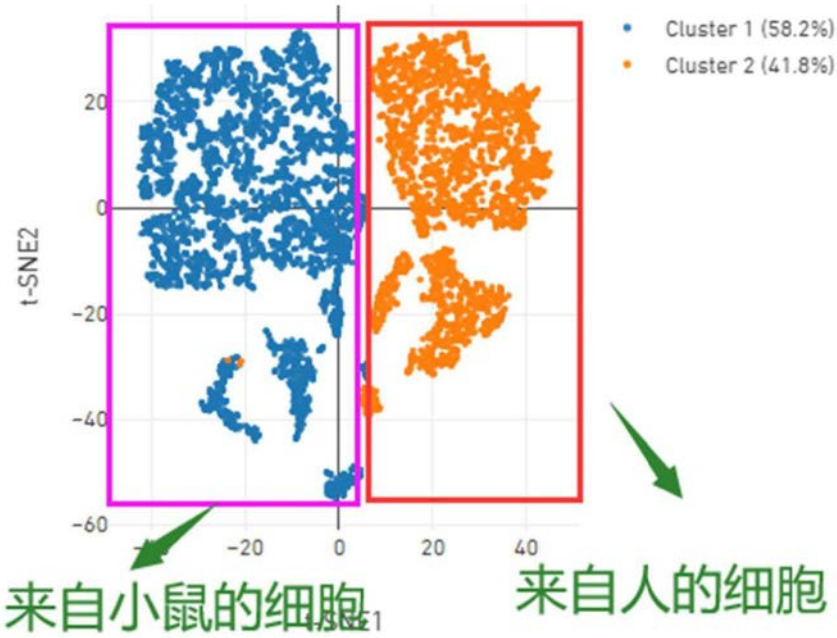


图6

Gene.stable.ID	Gene.name	Gene.stable.ID.1	Gene.name.1
ENSG00000278803	PWWP4	ENSMUSG00000071745	Pwwp4b
ENSG00000198464	ZNF480	ENSMUSG00000053211	Zfy1
ENSG00000198464	ZNF480	ENSMUSG00000000103	Zfy2
ENSG00000183833	CFAP91	ENSMUSG00000022805	Cfap91
ENSG00000163430	FSTL1	ENSMUSG00000022816	Fstl1
ENSG00000151320	AKAP6	ENSMUSG00000061603	Akap6
ENSG00000179304	FAM156B	ENSMUSG00000041353	Tmem29
ENSG00000171466	ZNF562	ENSMUSG00000053211	Zfy1
ENSG00000171466	ZNF562	ENSMUSG00000000103	Zfy2

图7

```
>ENSMUSG00000032094  Cd3d chromosome:GRCm39:9:44893084:44898637:1
ATGGAACACAGCGGGATTCTGGCTAGTCTGATACTGATTGCTGTTCTCCCCCAAGGGAGC
CCCTTCAAGATACAAGTGACCGAATATGAGGACAAAGTATTTGTGACCTGCAATACCAGC
GTCATGCATCTAGATGGAACGGTGGAAGGATGGTTTGCAAAGAATAAAACACTCAACTTG
GGCAAAGGCGTTCTGGACCCACGAGGGATATATCTGTGTAATGGGACAGAGCAGCTGGCA
AAGGTGGTGTCTTCTGTGCAAGTCCATTACCGAATGTGCCAGAACTGTGTGGAGCTAGAC
TCGGGCACCATGGCTGGTGTCATCTTCATTGACCTCATCGCAACTCTGCTCCTGGCTTTG
GGCGTCTACTGCTTTGCAGGACATGAGACCGGAAGGCCTTCTGGGGCTGCTGAGGTTCAA
GCACTGCTGAAGAATGAGCAGCTGTATCAGCCTCTTCGAGATCGTGAAGATACCCAGTAC
AGCCGTCTTGGAGGGAACTGGCCCCGGAACAAGAAATCTTAA
```

图8

```
>merge0000001 ENSG00000167286+ENSMUSG00000032094
ATGGAACATAGCACGTTTCTCTCTGGCCTGGTACTGGCTACCCTTCTCTCGCAAGTGAGC
CCCTTCAAGATACCTATAGAGGAACTTGAGGACAGAGTGTTTGTGAATTGCAATACCAGC
ATCACATGGGTAGAGGGAACGGTGGGAACACTGCTCTCAGACATTACAAGACTGGACCTG
GGAAAACGCATCCTGGACCCACGAGGAATATATAGGTGTAATGGGACAGATATATACAAG
GACAAAGAATCTACCGTGCAAGTTCATTATCGAATGTGCCAGAGCTGTGTGGAGCTGGAT
CCAGCCACCGTGGCTGGCATCATTGTCACTGATGTCATTGCCACTCTGCTCCTTGCTTTG
GGAGTCTTCTGCTTTGCTGGACATGAGACTGGAAGGCTGTCTGGGGCTGCCGACACACAA
GCTCTGTTGAGGAATGACCAGGTCTATCAGCCCCTCCGAGATCGAGATGATGCTCAGTAC
AGCCACCTTGGAGGAAACTGGGCTCGGAACAAGTGANNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATGG
AACACAGCGGGATTCTGGCTAGTCTGATACTGATTGCTGTTCTCCCCCAAGGGAGCCCCT
TCAAGATACAAGTGACCGAATATGAGGACAAAGTATTTGTGACCTGCAATACCAGCGTCA
TGCATCTAGATGGAACGGTGGAAGGATGGTTTGCAAAGAATAAAACACTCAACTTGGGCA
AAGGCGTTCTGGACCCACGAGGGATATATCTGTGTAATGGGACAGAGCAGCTGGCAAAGG
TGGTGTCTTCTGTGCAAGTCCATTACCGAATGTGCCAGAACTGTGTGGAGCTAGACTCGG
GCACCATGGCTGGTGTCATCTTCATTGACCTCATCGCAACTCTGCTCCTGGCTTTGGGCG
TCTACTGCTTTGCAGGACATGAGACCGGAAGGCCTTCTGGGGCTGCTGAGGTTCAAGCAC
TGCTGAAGAATGAGCAGCTGTATCAGCCTCTTCGAGATCGTGAAGATACCCAGTACAGCC
GTCTTGGAGGGAACTGGCCCCGGAACAAGAAATCTTAA
```

图9

```
merge0000001    customize   gene    1   1118    .   +   .   gene_id "merge0000001.1";
gene_name "merge0000001.1"; gene_biotype "protein_coding";
merge0000001    customize   transcript  1   516 .   +   .   gene_id
"merge0000001.1";transcript_id "merge0000001-trans1"; gene_name "merge0000001.1";
gene_biotype "protein_coding";
merge0000001    customize   exon    1   516 .   +   .   gene_id
"merge0000001.1";transcript_id "merge0000001-trans1"; gene_name "merge0000001.1";
gene_biotype "protein_coding";
merge0000001    customize   transcript  597 1118    .   +   .   gene_id
"merge0000001.1";transcript_id "merge0000001-trans2"; gene_name "merge0000001.1";
gene_biotype "protein_coding";
merge0000001    customize   exon    597 1118    .   +   .   gene_id
"merge0000001.1";transcript_id "merge0000001-trans2"; gene_name "merge0000001.1";
gene_biotype "protein_coding";
```

图10

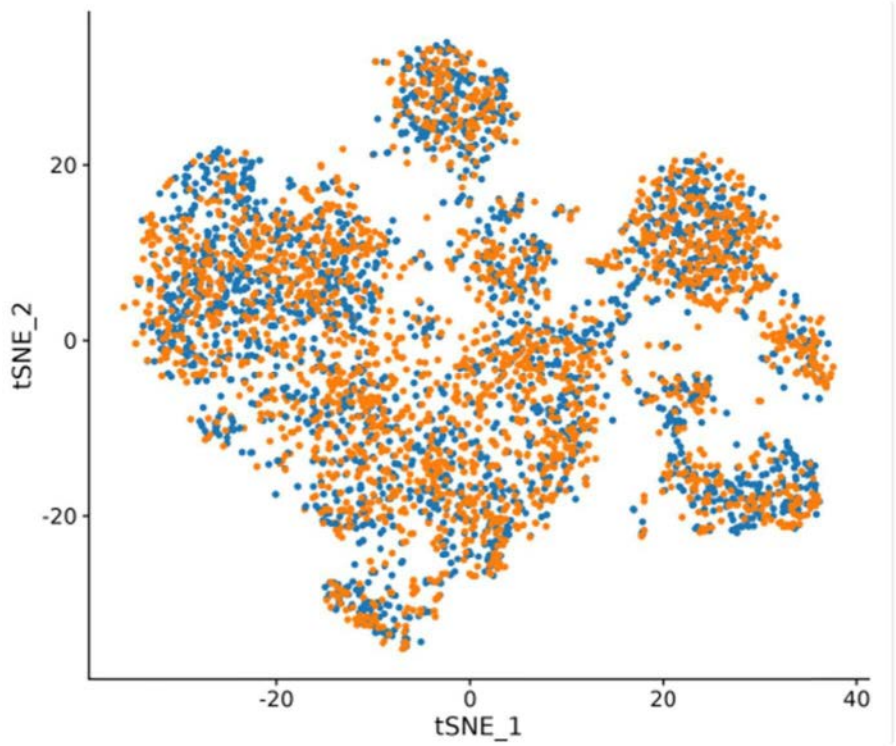

20
0
-20
tSNE_2
-20
0
20
40
tSNE_1

图11

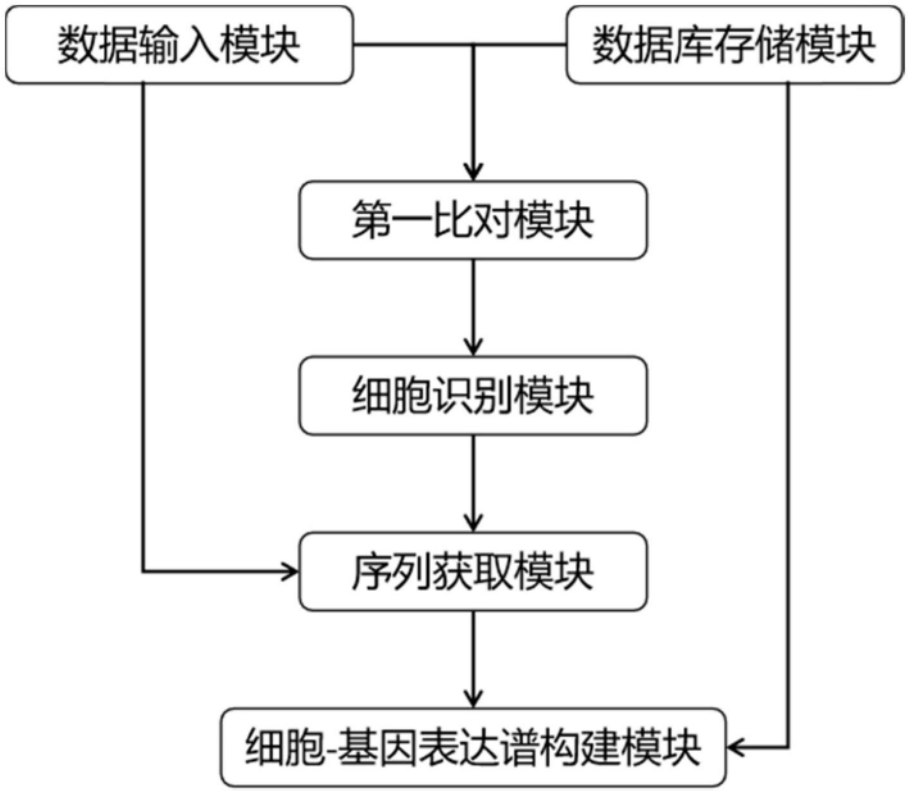

数据输入模块
数据库存储模块
第一比对模块
细胞识别模块
序列获取模块
细胞-基因表达谱构建模块

图12

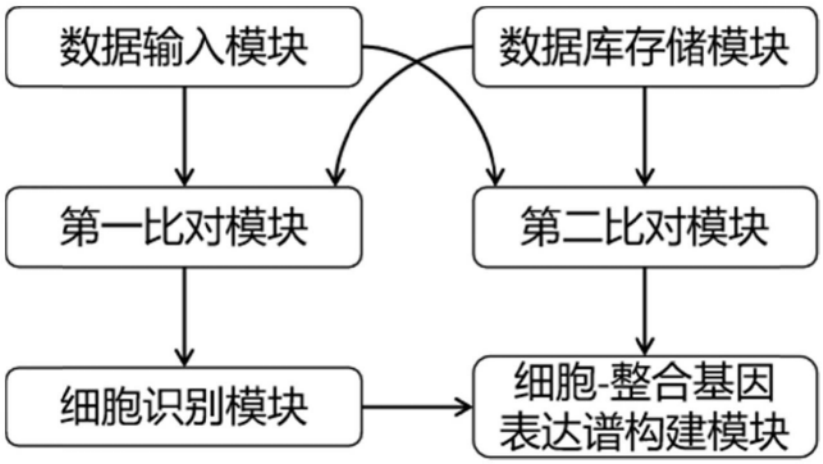

数据输入模块
数据库存储模块
第一比对模块
第二比对模块
细胞识别模块
细胞-整合基因
表达谱构建模块

图13

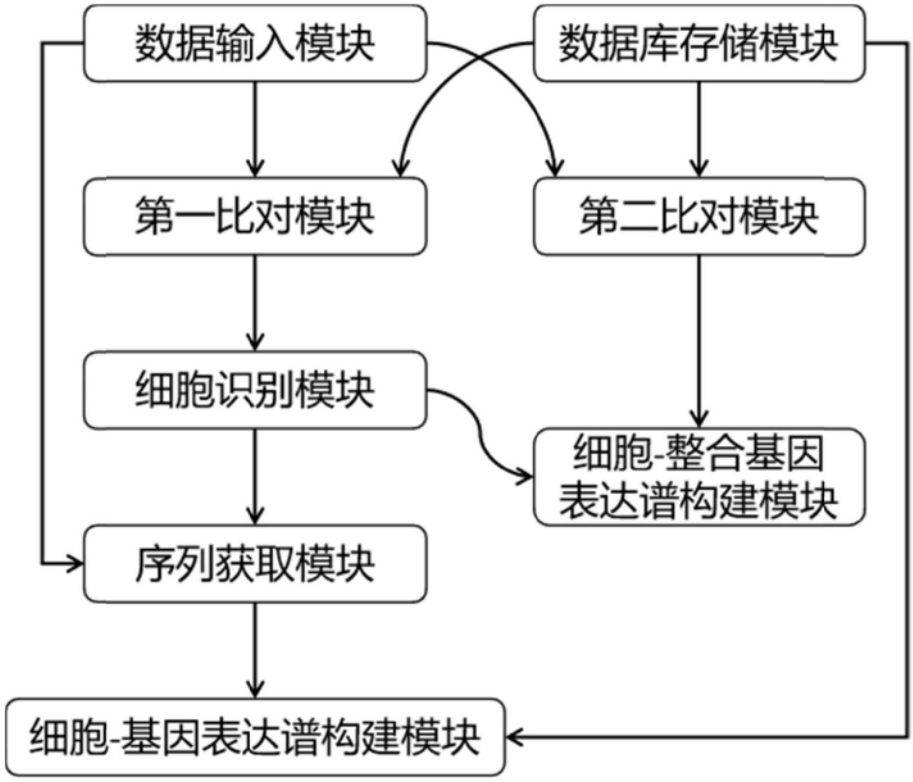

数据输入模块
数据库存储模块
第一比对模块
第二比对模块
细胞识别模块
细胞-整合基因
表达谱构建模块
序列获取模块
细胞-基因表达谱构建模块

图14