(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109952612 B
(45)授权公告日 2023.12.01

(21)申请号 201780068294.3

(22)申请日 2017.11.07

(65)同一申请的已公布的文献号
申请公布号 CN 109952612 A

(43)申请公布日 2019.06.28

(30)优先权数据
62/419,291 2016.11.08 US
62/446,227 2017.01.13 US

(85)PCT国际申请进入国家阶段日
2019.05.05

(86)PCT国际申请的申请数据
PCT/US2017/060451 2017.11.07

(87)PCT国际申请的公布数据
WO2018/089378 EN 2018.05.17

(73)专利权人 贝克顿迪金森公司
地址 美国新泽西州

(72)发明人 范珏 杰西·张 胡静

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 张瑞 郑霞

(51)Int.Cl.
G16B 40/00(2019.01)
G16B 99/00(2019.01)
G16B 25/00(2019.01)

(56)对比文件
US 2010145893 A1,2010.06.10
WO 2016162309 A1,2016.10.13
CN 1878873 A,2006.12.13
CN 103003444 A,2013.03.27
CN 102194133 A,2011.09.21
WO 03053215 A2,2003.07.03
WO 03093445 A2,2003.11.13
Therese Sørlie等.“Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”.《PROCEEDINGS NATIONAL ACADEMY OF SCIENCES PNAS,NATIONAL ACADEMY OF SCIENCES》.2001,
patrik D’haeseleer.“How does gene expression clustering work?”.《Nature Biotechnology》.2005,

审查员 江梓琴

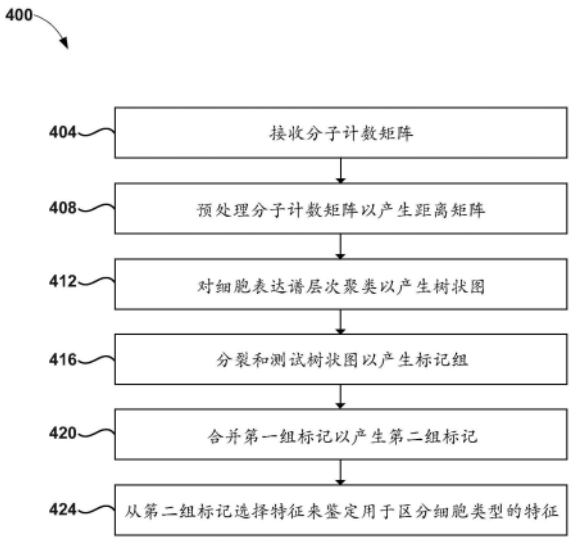
权利要求书8页 说明书69页 附图60页

(54)发明名称

用于表达谱分类的方法

(57)摘要

本文中公开了鉴定用于区分细胞类型的靶的方法和系统。在一些实施例中,该方法包括:将细胞的表达谱层次聚类以产生树状图,其中每个叶代表不同个体细胞的表达谱。可以通过消除无效分裂(及其子)来修剪该树状图。可以将该树状图中的剩余叶独立于其在该树状图中的位置基于其彼此的距离进行合并以产生表达谱聚类。该方法可以基于在聚类中差异表达的靶来鉴定用于区分细胞类型的靶。



CN 109952612 B

1.一种用于鉴定区分细胞类型的靶的方法,该方法包括:

(a)接收靶计数数据结构,其中该靶计数数据结构包含多个细胞的表达谱,并且其中该多个细胞的表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目;

(b)基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类,以产生代表该多个细胞的表达谱的树状图,其中该多个细胞的表达谱之间的距离包含该多个细胞的表达谱之间的成对相关距离,其中该树状图包含多个节点,其中该多个节点包括根节点、多个叶节点和多个非根非叶节点,其中该多个叶节点中的每个叶节点代表该多个细胞中不同细胞的表达谱,其中该根节点代表该多个细胞的表达谱,其中该多个叶节点和该多个非根非叶节点中的每一个与母节点相关联,并且

其中该根节点和该多个非根非叶节点中的每一个与左子节点和右子节点相关联,并代表由该节点的该左子节点和该右子节点代表的表达谱;

(c)在从该树状图的根节点到该树状图的多个叶节点遍历该树状图的多个节点的每个节点时:

(1)通过以下确定该节点分裂成该节点的子节点是有效还是无效的:

在该左子节点和该右子节点之间的距离高于分裂阈值,确定分裂有效;并且

在该左子节点和该右子节点之间的距离不高于分裂阈值,确定分裂无效,

其中该左子节点和该右子节点之间的距离基于对该左子节点和该右子节点代表的表达谱之间多个靶中每个靶实施的统计学检验来确定;和

(2)在该节点分裂成该节点的子节点是无效时,将该节点添加到合并聚类集中;

(d)迭代地,对于该合并聚类集中的每个第一节点,

在该合并聚类集中的该第一节点与该合并聚类集中最接近该第一节点的第二节点之间的距离在合并距离阈值内时,将该第一节点与该第二节点合并以生成合并的节点,该合并的节点包含由该第一节点和该第二节点代表的表达谱;和

(e)对于该合并聚类集中的每个节点,基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶,其中鉴定用于区分细胞类型的靶包括:

确定由该节点代表的表达谱和由该合并聚类集中另一节点代表的表达谱之间在与每个靶相关联的具有不同序列的分子标记的数目方面的差;以及

如果所确定的与靶相关联的具有不同序列的分子标记的数目的差大于显著性阈值,则鉴定用于区分不同细胞类型的靶。

2.如权利要求1所述的方法,其中该靶计数数据结构包括靶计数矩阵。

3.如权利要求2所述的方法,其中该靶计数矩阵的每行或每列包含针对该多个细胞中不同个体细胞的多个靶中每个靶的数目。

4.如权利要求1所述的方法,该方法包括:在(a)中接收靶计数数据结构之前:

(f)使用多个条形码对该多个细胞中的该多个靶进行条形码化以创建多个经条形码化的靶,其中该多个条形码中的每一个包括细胞标记和分子标记,其中从不同细胞的靶创建的经条形码化的靶具有不同的细胞标记,并且其中从该多个细胞中一个细胞的靶创建的经条形码化的靶具有不同的分子标记;

(g)获得该多个经条形码化的靶的测序数据;和

(h)对于该多个细胞的每个:

(1)针对该细胞,对该测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数;和

(2)基于(h)(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目,估算该细胞的多个靶中每个靶的数目。

5.如权利要求4所述的方法,其中接收该靶计数数据结构包括:

从(h)(2)中估算的该细胞的多个靶中每个靶的数目生成靶计数数据结构,其中该多个细胞中该细胞的表达谱包含(h)(2)中估算的该细胞的多个靶中每个靶的数目。

6.如权利要求1所述的方法,该方法包括:在(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类以产生代表该多个细胞的表达谱的树状图之前:

(i)确定该靶计数数据结构的元素的距离数据结构,其中该距离数据结构包含该多个细胞的表达谱之间的距离。

7.如权利要求6所述的方法,其中该距离数据结构包括距离矩阵。

8.如权利要求7所述的方法,其中该距离矩阵的每个对角元素具有零值。

9.如权利要求6所述的方法,其中在(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类以产生代表该多个细胞的表达谱的树状图包括:基于该靶计数数据结构和该距离数据结构对该多个细胞的表达谱进行层次聚类。

10.如权利要求6所述的方法,其中该多个细胞的表达谱之间的距离是该多个细胞的表达谱之间的成对相关距离。

11.如权利要求6所述的方法,该方法包括:在(i)中确定该靶计数数据结构的元素的距离数据结构之前,将该靶计数数据结构对数转换为经对数转换的靶计数数据结构,

其中确定该靶计数数据结构的元素的距离数据结构包括确定该经对数转换的靶计数数据结构的距离数据结构,并且

其中(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类包括:基于该经对数转换的靶计数数据结构和该距离数据结构对该多个细胞的表达谱进行层次聚类以产生树状图。

12.如权利要求11所述的方法,其中将该靶计数数据结构对数转换为经对数转换的靶计数数据结构包括通过增量增加该靶计数数据结构的每个元素的值。

13.如权利要求12所述的方法,其中该增量为1。

14.如权利要求1所述的方法,其中(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类包括:

将该多个细胞的每个表达谱分配给不同的叶节点;并且

当该第二节点是该多个节点中最接近该第一节点的节点时,迭代地将该多个节点的第一节点和第二节点组合以生成该第一节点和该第二节点的母节点。

15.如权利要求14所述的方法,其中该第一节点和该第二节点之间的距离是具有由该第一节点代表的表达谱的任何细胞与具有由该第二节点代表的表达谱的任何细胞之间的最大距离。

16.如权利要求14所述的方法,其中该第一节点的节点内相关性和该第二节点的节点

内相关性中的至少一个高于该第一节点和该第二节点的节点间相关性。

17.如权利要求14所述的方法,其中该第一节点的节点内相关性和该第二节点的节点内相关性的指标高于该第一节点和该第二节点的节点间相关性。

18.如权利要求17所述的方法,其中该第一节点的节点内相关性和该第二节点的节点内相关性的指标基于以下中的至少一种:

该第一节点和该第二节点的节点内最大相关性,

该第一节点和该第二节点的节点内平均相关性,

该第一节点和该第二节点的节点内中值相关性,

该第一节点和该第二节点的节点内最小相关性,以及

其任何组合。

19.如权利要求16所述的方法,其中该第一节点的节点内相关性基于以下中的至少一种:

该第一节点的节点内最大相关性,

该第一节点的节点内平均相关性,

该第一节点的节点内中值相关性,

该第一节点的节点内最小相关性,以及

其任何组合。

20.如权利要求16所述的方法,其中该第二节点的节点内相关性基于以下中的至少一种:

该第二节点的节点内最大相关性,

该第二节点的节点内平均相关性,

该第二节点的节点内中值相关性,

该第二节点的节点内最小相关性,以及

其任何组合。

21.如权利要求16所述的方法,其中该第一节点和该第二节点的节点间相关性基于以下中的至少一种:

该第一节点和该第二节点的节点间最大相关性,

该第一节点和该第二节点的节点间平均相关性,

该第一节点和该第二节点的节点间中值相关性,

该第一节点和该第二节点的节点间最小相关性,以及

其任何组合。

22.如权利要求1所述的方法,其中遍历该树状图的多个节点包括:

从该节点遍历到该节点的左子节点和右子节点;和

当分裂无效时,停止从该节点遍历到该节点的左子节点和右子节点。

23.如权利要求1所述的方法,其中该统计学检验包括Welch t检验。

24.如权利要求1所述的方法,其中该左子节点和该右子节点之间的距离基于对该左子节点代表的每个表达谱和该右子节点代表的每个表达谱之间多个靶中每个靶实施的统计学检验的最大p值来确定。

25.如权利要求1-24中任一项所述的方法,该方法包括:在遍历该树状图的多个节点

时,在每个节点处:(3)当该节点代表单细胞的表达谱时,将该节点添加到合并聚类集。

26.如权利要求1-24中任一项所述的方法,该方法包括:在遍历该树状图的多个节点时,在每个节点处:为该节点分配节点标记。

27.如权利要求26所述的方法,

其中当该节点代表单细胞的表达谱时,则该节点的节点标记包括单细胞标识,

否则当该节点是该母节点的左子节点时,则该节点的节点标记包括该母节点的节点标记和左标识,以及

否则,该节点的节点标记包括该母节点的节点标记和右标识。

28.如权利要求1-24中任一项所述的方法,该方法包括:在(d)中将该第一节点与该第二节点合并以生成合并的节点之前:

当该合并聚类集中代表单细胞的表达谱的第三节点和该合并聚类集中的第四节点之间的距离在节点距离阈值内时,将该第三节点与该第四节点合并。

29.如权利要求1-24中任一项所述的方法,该方法包括:基于合并聚类集中代表细胞表达谱的节点对该多个细胞进行分类。

30.如权利要求1-24中任一项所述的方法,该方法包括:基于所鉴定的用于区分细胞类型的靶设计全转录组测定。

31.如权利要求1-24中任一项所述的方法,该方法包括:基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

32.一种用于鉴定区分细胞类型的靶的方法,该方法包括:

(a)接收多个细胞的表达谱,其中这些表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目;

(b)基于该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行聚类以产生多个表达谱聚类,其中每个聚类与(1)母聚类和(2)两个或更多个子聚类中的一者或两者具有一个或多个关联,其中该母聚类代表由该聚类代表的该多个细胞的一个或多个细胞的表达谱,且其中该聚类代表由该两个或更多个子聚类代表的表达谱,其中该多个细胞的表达谱之间的距离包含该多个细胞的表达谱之间的成对相关距离;

(c)对于具有两个或更多个子聚类的每个聚类,当该聚类与该两个或更多个子聚类之间的关联无效时将该聚类添加到合并聚类集,其中当左子聚类和右子聚类之间的距离不高于关联阈值时,确定该聚类与该两个或更多个子聚类之间的关联无效,

其中当左子聚类和右子聚类之间的距离高于关联阈值时,确定该聚类与该两个或更多个子聚类之间的关联有效,

其中该左子聚类和该右子聚类之间的距离基于对该左子聚类和该右子聚类代表的表达谱之间多个靶中每个靶实施的统计学检验来确定;

(d)迭代地,对于该合并聚类集中的每个第一聚类,当该合并聚类集中的该第一聚类与该合并聚类集中最接近该第一聚类的第二聚类之间的距离在合并距离阈值内时,将该第一聚类和该第二聚类合并以生成合并的聚类,其中该合并的聚类包含该第一聚类和该第二聚类的表达谱;和

(e)对于该合并聚类集中的每个聚类,基于该聚类所代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶,其中鉴定用于区分细胞类型的靶包括:

确定由该聚类代表的表达谱和由该合并聚类集中另一聚类代表的表达谱之间在与每个靶相关联的具有不同序列的分子标记的数目方面的差;以及

如果所确定的与靶相关联的具有不同序列的分子标记的数目的差大于显著性阈值,则鉴定用于区分不同细胞类型的靶。

33.如权利要求32所述的方法,其中接收该多个细胞的表达谱包括接收靶计数数据结构。

34.如权利要求33所述的方法,其中该靶计数数据结构包括靶计数矩阵。

35.如权利要求34所述的方法,其中该靶计数矩阵的每行或每列包含该多个细胞中不同个体细胞的表达谱。

36.如权利要求32所述的方法,其中基于该多个细胞的表达谱之间的距离将该多个细胞的表达谱聚类成多个表达谱聚类包括:

基于该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类,以产生代表该多个细胞的表达谱的树状图,其中该树状图包含多个聚类,其中该多个聚类包括根聚类、多个叶聚类和多个非根非叶聚类。

37.如权利要求36所述的方法,

其中该多个叶聚类和该多个非根非叶聚类中的每个与母聚类关联,

其中该根聚类和该多个非根非叶聚类中的每个与左子聚类和右子聚类关联,且代表由该聚类的该左子聚类和该右子聚类所代表的表达谱,并且

其中该根聚类代表该多个细胞的表达谱。

38.如权利要求36所述的方法,其中将该聚类添加到合并聚类集中包括:

当从该树状图的根聚类到该树状图的多个叶聚类遍历该树状图的每个聚类时:

(1)确定该聚类与该聚类的子聚类的关联是否是无效的;和

(2)如果关联无效,则将该聚类添加到合并聚类集。

39.如权利要求32所述的方法,该方法包括:在(a)中接收该多个细胞的表达谱之前:

(f)使用多个条形码对该多个细胞中的该多个靶进行条形码化以创建多个经条形码化的靶,其中该多个条形码中的每一个包括细胞标记和分子标记,其中从不同细胞的靶创建的经条形码化的靶具有不同的细胞标记,并且其中从该多个细胞中一个细胞的靶创建的经条形码化的靶具有不同的分子标记;

(g)获得该多个经条形码化的靶的测序数据;和

(h)对于该多个细胞的每个:

(1)针对该细胞,对该测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数;和

(2)基于(h)(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目,估算该细胞的多个靶中每个靶的数目。

40.如权利要求39所述的方法,其中该多个细胞中该细胞的表达谱包含(h)(2)中估算的该细胞的多个靶中每个靶的数目。

41.如权利要求32所述的方法,该方法包括:在(b)中基于该多个细胞的表达谱之间的距离对该多个细胞的表达谱聚类以产生多个表达谱聚类之前:

(i)确定该多个细胞的表达谱的距离数据结构。

42.如权利要求41所述的方法,其中该距离数据结构包含该多个细胞的表达谱的距离矩阵。

43.如权利要求42所述的方法,其中该距离矩阵的每个对角元素具有零值。

44.如权利要求42所述的方法,其中在(b)中基于该多个细胞的表达谱之间的距离将该多个细胞的表达谱聚类以产生多个表达谱聚类包括:基于该距离矩阵将该多个细胞的表达谱聚类以产生多个表达谱聚类。

45.如权利要求41所述的方法,其中该多个细胞的表达谱之间的距离是该多个细胞的表达谱之间的成对相关距离。

46.如权利要求41所述的方法,该方法包括:在(i)中确定该距离数据结构之前,将靶计数数据结构对数转换为经对数转换的靶计数数据结构,

其中确定该靶计数数据结构的元素的距离数据结构包括确定该经对数转换的靶计数数据结构的距离数据结构,并且

其中在(b)中基于该多个细胞的表达谱之间的距离将该多个细胞的表达谱聚类以产生多个表达谱聚类包括:基于该经对数转换的靶计数数据结构和该距离数据结构对该多个细胞的表达谱进行聚类以产生多个聚类。

47.如权利要求46所述的方法,其中将该靶计数数据结构对数转换为经对数转换的靶计数数据结构包括通过增量增加该靶计数数据结构的每个元素的值。

48.如权利要求47所述的方法,其中该增量为1。

49.如权利要求48所述的方法,其中在(b)中基于该多个细胞的表达谱之间的距离将该多个细胞的表达谱聚类包括:

将该多个细胞的每个表达谱分配给该多个聚类中不同的叶聚类;和

当该第二聚类是该多个聚类中最接近该第一聚类的聚类时,迭代地将该多个聚类的第一聚类和第二聚类组合以生成该第一聚类和该第二聚类的母聚类。

50.如权利要求49所述的方法,其中该第一聚类和该第二聚类之间的距离是由该第一聚类代表的任意表达谱与由该第二聚类代表的任意表达谱之间的最大距离。

51.如权利要求49所述的方法,其中该第一聚类的聚类内相关性和该第二聚类的聚类内相关性高于该第一聚类和该第二聚类的聚类间相关性。

52.如权利要求49所述的方法,其中该第一聚类的聚类内相关性和该第二聚类的聚类内相关性的指标高于该第一聚类和该第二聚类的聚类间相关性。

53.如权利要求52所述的方法,其中该第一聚类的聚类内相关性和该第二聚类的聚类内相关性的指标基于以下中的至少一种:

该第一聚类和该第二聚类的聚类内最大相关性,

该第一聚类和该第二聚类的聚类内平均相关性,

该第一聚类和该第二聚类的聚类内中值相关性,

该第一聚类和该第二聚类的聚类内最小相关性,以及

其任何组合。

54.如权利要求51所述的方法,其中该第一聚类的聚类内相关性基于以下中的至少一种:

该第一聚类的聚类内最大相关性,

该第一聚类的聚类内平均相关性，

该第一聚类的聚类内中值相关性，

该第一聚类的聚类内最小相关性，以及

其任何组合。

55.如权利要求51所述的方法，其中该第二聚类的聚类内相关性基于以下中的至少一种：

该第二聚类的聚类内最大相关性，

该第二聚类的聚类内平均相关性，

该第二聚类的聚类内中值相关性，

该第二聚类的聚类内最小相关性，以及

其任何组合。

56.如权利要求51所述的方法，其中该第一聚类和该第二聚类的聚类间相关性基于以下中的至少一种：

该第一聚类和该第二聚类的聚类间最大相关性，

该第一聚类和该第二聚类的聚类间平均相关性，

该第一聚类和该第二聚类的聚类间中值相关性，

该第一聚类和该第二聚类的聚类间最小相关性，以及

其任何组合。

57.如权利要求38所述的方法，其中遍历该树状图的该多个聚类包括：

从该聚类遍历到该聚类的左子聚类和右子聚类；和

当关联无效时，则停止从该聚类遍历到该聚类的左子聚类和右子聚类。

58.如权利要求32所述的方法，其中该统计学检验包括Welch t检验。

59.如权利要求32所述的方法，其中该左子聚类和该右子聚类之间的距离基于对该左子聚类代表的表达谱和该右子聚类代表的每个表达谱之间多个靶中每个靶实施的统计学检验的最大p值来确定。

60.如权利要求38所述的方法，该方法包括：在遍历该树状图的该多个聚类时，在每个聚类处：(3)当该聚类代表单细胞的表达谱时，将该聚类添加到合并聚类集。

61.如权利要求38所述的方法，该方法包括：在遍历该树状图的该多个聚类时，在每个聚类处：为该聚类分配聚类标记。

62.如权利要求61所述的方法，

其中当该聚类代表单细胞的表达谱时，该聚类的聚类标记包括单细胞标识，

否则，当该聚类是母聚类的左子聚类时，则该聚类的聚类标记包括母聚类的聚类标记和左标识，以及

否则，该聚类的聚类标记包括母聚类的聚类标记和右标识。

63.如权利要求32-62中任一项所述的方法，该方法包括：在(d)中将该第一聚类与该第二聚类合并以生成合并的聚类之前：

如果该合并聚类集中代表单细胞的表达谱的第三聚类和该合并聚类集中的第四聚类之间的距离在聚类距离阈值内，将该第三聚类和该第四聚类合并。

64.如权利要求32-62中任一项所述的方法，该方法包括：基于合并聚类集中代表细胞

表达谱的聚类对该多个细胞进行分类。

65.如权利要求32-62中任一项所述的方法，该方法包括：基于所鉴定的用于区分细胞类型的靶设计全转录组测定。

66.如权利要求32-62中任一项所述的方法，该方法包括：基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

67.一种用于确定靶的数目的计算机系统，该计算机系统包含：

硬件处理器；和

具有存储在其上的指令的非暂时性存储器，这些指令当通过该硬件处理器执行时使得该处理器执行如权利要求1-24和32-62中任一项所述的方法。

68.一种计算机可读介质，其包含用于执行如权利要求1-24和32-62中任一项所述的方法的代码。

用于表达谱分类的方法

[0001] 相关申请的交叉引用

[0002] 本申请要求于2016年11月8日提交的美国临时专利申请号62/419291、和2017年1月13日提交的美国临时专利申请号62/446227的优先权。这些相关申请中每个申请的内容在此通过引用以其全文明确地并入本文中。

[0003] 版权和商标声明

[0004] 本专利文件的公开内容的一部分含有受到版权保护的材料。版权所有者不反对任何人对专利文件或专利披露进行传真复制，因为它出现在专利商标局的专利文件或记录中，但无论如何在其他方面保留所有版权权利。

[0005] 发明背景

技术领域

[0006] 本公开一般涉及对表达谱进行分类的领域，更具体地涉及鉴定靶来区分细胞类型的领域。

背景技术

[0007] 方法和技术如条形码化(例如，随机条形码化)可用于细胞分析。例如，条形码化可用于破译单细胞的基因表达谱以确定其状态，使用例如逆转录、聚合酶链反应(PCR)扩增和下一代测序(NGS)。然而，这些方法和技术产生的大量数据需要进一步分析，以鉴定用于区分细胞类型的标志物和确定所分析细胞的类型。

发明内容

[0008] 本文中公开了用于鉴定靶以区分细胞类型的方法。在一些实施例中，该方法包括：(a)接收靶计数数据结构，其中该靶计数数据结构包含多个细胞的表达谱，并且其中该多个细胞的表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目；(b)基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类，以产生代表该多个细胞的表达谱的树状图，其中该树状图包含多个节点，其中该多个节点包括根节点、多个叶节点和多个非根非叶节点，其中该多个叶节点中的每个叶节点代表该多个细胞中不同细胞的表达谱，且其中该根节点代表该多个细胞的表达谱；(c)在从该树状图的根节点到该树状图的多个叶节点遍历该树状图的多个节点的每个节点时：(1)确定该节点分裂成该节点的子节点是有效还是无效的(例如，子节点之间的差异不显著)；和(2)如果该节点分裂成该节点的子节点是无效的，则将该节点添加到合并聚类集中；(d)迭代地，对于该合并聚类集中的每个第一节点，如果该合并聚类集中的该第一节点与该合并聚类集中最接近该第一节点的第二节点之间的距离在合并距离阈值内，则将该第一节点与该第二节点合并以生成合并的节点，该合并的节点包含由该第一节点和该第二节点代表的表达谱；和(e)对于该合并聚类集中的每个节点，基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0009] 在一些实施例中，该靶计数数据结构包含靶计数矩阵。该靶计数矩阵的每行或每列可包含针对该多个细胞中不同个体细胞的多个靶中每个靶的数目。

[0010] 在一些实施例中，多个叶节点和多个非根非叶节点中的每一个可以与母节点相关联，并且根节点和多个非根非叶节点中的每一个可以与左子节点和右子节点相关联，并代表由该节点的左子节点和右子节点代表的表达谱。

[0011] 在一些实施例中，该方法包括在(a)中接收靶计数数据结构之前：(f)使用多个条形码对该多个细胞中的该多个靶进行条形码化以创建多个经条形码化的靶，其中该多个条形码中的每一个包括细胞标记和分子标记，其中从不同细胞的靶创建的经条形码化的靶具有不同的细胞标记，并且其中从该多个细胞中一个细胞的靶创建的经条形码化的靶具有不同的分子标记；(g)获得该多个经条形码化的靶的测序数据；以及(h)对于该多个细胞的每个：(1)针对该细胞，对测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数；以及(2)基于(h)(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目，估算该细胞的多个靶中每个靶的数目。例如，该方法可以包括，在(a)中接收靶计数数据结构之前：步骤(s)(f)使用多个随机条形码随机地对该多个细胞的多个靶进行条形码编码以创建多个经随机经条形码化的靶，其中该多个经随机条形码中的每一个包含细胞标记和分子标记，其中从不同细胞的靶创建的随机经条形码化的靶具有不同的细胞标记，且其中从该个多个细胞中的一细胞的靶创建的经随机经条形码化的靶具有不同的分子标记；和/或(g)获得该多个经随机经条形码化的靶的测序数据。接收靶计数数据结构可以包括：从(h)(2)中估算的该细胞的多个靶中每个靶的数目生成靶计数数据结构，其中该多个细胞中该细胞的表达谱包含(h)(2)中估算的该细胞的多个靶中每个靶的数目。

[0012] 在一些实施例中，该方法包括在(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类以产生代表该多个细胞的表达谱的树状图之前：(i)确定该靶计数矩阵的元素的距离数据结构，其中该距离数据结构包含该多个细胞的表达谱之间的距离。该距离数据结构可以包含距离矩阵。该距离矩阵的每个对角元素可以具有零值。(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类以产生代表多个细胞的表达谱的树状图可以包括，基于该靶计数数据结构和该距离数据结构对该多个细胞的表达谱进行层次聚类。该多个细胞的表达谱之间的距离可以包含该多个细胞的表达谱之间的成对相关距离。

[0013] 在一些实施例中，在(i)中确定靶计数数据结构的元素的距离数据结构之前，将靶计数数据结构对数转换为经对数转换的靶计数数据结构，其中确定靶计数数据结构的元素的距离数据结构包括确定该经对数转换的靶计数数据结构的距离数据结构，并且其中(b)中基于靶计数数据结构和多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类包括：基于该经对数转换的靶计数数据结构和该距离数据结构对该多个细胞的表达谱进行层次聚类以产生树状图。将靶计数数据结构对数转换为经对数转换的靶计数数据结构可以包括通过增量(例如1)增加靶计数数据结构的每个元素的值。

[0014] 在一些实施例中，(b)中基于靶计数数据结构和多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类包括：将该多个细胞的每个表达谱分配给不同的叶节点；且迭代地将该多个节点的第一节点和第二节点组合以生成该第一节点和该第二节点的母节点，前提是如果该第二节点是该多个节点中最接近该第一节点的节点的话。该第一节点

和该第二节点之间的距离是具有由该第一节点代表的表达谱的任何细胞与具有由该第二节点代表的表达谱的任何细胞之间的最大距离。

[0015] 在一些实施例中，该方法包括：在遍历树状图的多个节点时，在每个节点处：如果分裂有效，则继续从该节点遍历到该节点的左子节点和右子节点；如果分裂无效，则停止从该节点遍历到该节点的左子节点和右子节点。该第一节点的节点内相关性和该第二节点的节点内相关性中的至少一个可以高于该第一节点和该第二节点的节点间相关性。该第一节点的节点内相关性和该第二节点的节点内相关性的量度或指标可以高于该第一节点和该第二节点的节点间相关性。第一节点的节点内相关性和第二节点的节点内相关性的量度可以基于以下中的至少一种：第一节点和第二节点的节点内最大相关性，第一节点和第二节点的节点内平均相关性，第一节点和第二节点的节点内中值相关性，第一节点和第二节点的节点内最小相关性，以及其任何组合。第一节点的节点内相关性可以基于以下中的至少一种：第一节点的节点内最大相关性，第一节点的节点内平均相关性，第一节点的节点内中值相关性，第一节点的节点内最小相关性，以及其任何组合。第二节点的节点内相关性可以基于以下中的至少一种：第二节点的节点内最大相关性，第二节点的节点内平均相关性，第二节点的节点内中值相关性，第二节点的节点内最小相关性，以及其任何组合。第一节点和第二节点的节点间相关性可以基于以下中的至少一种：第一节点和第二节点的节点间最大相关性，第一节点和第二节点的节点间平均相关性，第一节点和第二节点的节点间中值相关性，第一节点和第二节点的节点间最小相关性，以及其任何组合。

[0016] 在一些实施例中，确定节点分裂成该节点的子节点是有效还是无效的包括：如果左子节点和右子节点之间的距离高于分裂阈值，则确定分裂有效，否则无效。左子节点和右子节点之间的距离可以基于对左子节点和右子节点代表的表达谱之间多个靶中每个靶实施的统计学检验来确定。所述统计学检验可以包括Welch t检验。左子节点和右子节点之间的距离可以基于对左子节点代表的每个表达谱和右子节点代表的每个表达谱之间多个靶中每个靶实施的统计学检验的最大p值来确定。

[0017] 在一些实施例中，确定节点分裂成该节点的子节点是有效还是无效的包括：如果该第一节点的节点内相关性和该第二节点的节点内相关性中的至少一个高于该该第一节点和该第二节点的节点间相关性，则确定分裂有效，否则为无效的。在一些实施例中，确定节点分裂成该节点的子节点是有效还是无效的包括：如果该第一节点的节点内相关性和该第二节点的节点内相关性的量度或指标高于该该第一节点和该第二节点的节点间相关性，则确定分裂有效，否则为无效的。第一节点的节点内相关性和第二节点的节点内相关性的量度可以基于以下中的至少一种：第一节点和第二节点的节点内最大相关性，第一节点和第二节点的节点内平均相关性，第一节点和第二节点的节点内中值相关性，第一节点和第二节点的节点内最小相关性，以及其任何组合。所述第一节点的节点内相关性可以基于以下的至少一种：第一节点的节点内最大相关性，第一节点的节点内平均相关性，第一节点的节点内中值相关性，第一节点的节点内最小相关性，以及其任何组合。所述第二节点的节点内相关性可以基于以下的至少一种：第二节点的节点内最大相关性，第二节点的节点内平均相关性，第二节点的节点内中值相关性，第二节点的节点内最小相关性，以及其任何组合。第一节点和第二节点的节点间相关性可以基于以下中的至少一种：第一节点和第二节点的节点间最大相关性，第一节点和第二节点的节点间平均相关性，第一节点和第二节点

的节点间中值相关性，第一节点和第二节点的节点间最小相关性，以及其任何组合。

[0018] 在一些实施例中，该方法包括：在遍历树状图的多个节点时，在每个节点处：(3)如果该节点代表单细胞的表达谱，则将该节点添加到合并聚类集。在一些实施例中，该方法可以包括在遍历树状图的多个节点时在每个节点处：为该节点分配节点标记。如果该节点表示单细胞的表达谱，则该节点的节点标记包括单细胞标识，否则如果该节点是母节点的左子节点，则该节点的节点标记包括母节点的节点标记和左标识，否则该节点的节点标记包括母节点的节点标记和右标识。

[0019] 在一些实施例中，对于合并聚类集中的每个节点，基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶包括：确定由该节点代表的表达谱和由合并聚类集中的另一节点代表的表达谱之间在与用于区分细胞类型的靶相关联的具有不同序列的分子标记的数目方面的差大于显著性阈值。

[0020] 在一些实施例中，该方法包括在(d)中在将第一节点与第二节点合并以生成合并的节点之前：将合并聚类集中代表单细胞的表达谱的每个第三节点与合并聚类集中的第四节点合并，前提是如果该第三节点和该第四节点之间的距离在节点距离阈值内的话。在一些实施例中，该方法包括基于合并聚类集中代表细胞表达谱的节点对多个细胞进行分类。该方法可以包括基于所鉴定的用于区分细胞类型的靶设计全转录组测定。在一些实施例中，该方法可包括基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

[0021] 本文中公开了鉴定用于区分细胞类型的靶的方法。在一些实施例中，该方法包括：(a)接收多个细胞的表达谱，其中所述表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目；(b)基于靶计数矩阵和该多个细胞的表达谱之间的距离对该该多个细胞的表达谱进行层次聚类，以产生代表该多个细胞的表达谱的树状图，其中该树状图包含多个节点，其中该多个节点包括根节点、多个叶节点和多个非根非叶节点，其中该多个叶节点中的每个叶节点代表该多个细胞中不同细胞的表达谱，且其中该根节点代表所述多个细胞的表达谱；(c)在从该树状图的根节点到该树状图的多个叶节点遍历该树状图的多个节点的每个节点时：(1)确定该节点的两个子分支(例如，由该节点的子节点代表)是否显著不同；和(2)如果该节点的两个子分支显著不同，则将该节点分成两个聚类集(例如，通过遍历至该节点的两个子分支)。在一些实施例中，该方法包括，(3)如果该节点分裂成该节点的子节点是无效的，则将该节点添加到合并聚类集中。在一些实施例中，该方法包括：(d)迭代地，对于该合并聚类集中的每个第一节点，如果合并聚类集中的第一节点与最接近该第一节点的合并聚类集中的第二节点之间的距离在合并距离阈值内，则将该第一节点与第二节点合并以生成该合并聚类集中的合并的节点；和(e)对于该合并聚类集中的每个节点，基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0022] 本文中公开了鉴定用于区分细胞类型的靶的方法。在一些实施例中，该方法包括：(a)接收多个细胞的表达谱，其中所述表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目；(b)基于该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行聚类以产生多个表达谱聚类，其中每个聚类与(1)母聚类和(2)两个或更多个子聚类中的一者或两者具有一个或多个关联，其中该母聚类代表由该聚类代表的该多个细胞的一个或多个细胞的表达谱，且其中该聚类代表由两个或更多个子聚类代表的表达谱；(c)对于具有两个或更多个子聚类的每个聚类，如果该聚类与该两个或更多个子聚类之间的关联无效(例如，该

两个或更多个子聚类之间的差异不显著)，则将该聚类添加到合并聚类集中；(d)迭代地，对于该合并聚类集中的每个第一聚类，如果该合并聚类集中的该第一聚类与该合并聚类集中最接近该第一聚类的第二聚类之间的距离在合并距离阈值内，则将该第一聚类和该第二聚类合并以生成合并的聚类，其中该合并的聚类包含该第一聚类和该第二聚类的表达谱；和(e)对于该合并聚类集中的每个聚类，基于该聚类所代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0023] 在一些实施例中，接收该多个细胞的表达谱包括接收靶计数数据结构。该靶计数数据结构可以包含靶计数矩阵。该靶计数矩阵的每行或每列可包含多个细胞中不同个体细胞的表达谱。基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类成多个表达谱聚类可以包括：基于多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类，以产生代表多个细胞的表达谱的树状图。该树状图可以包含多个聚类。该多个聚类可以包含根聚类、多个叶聚类、和多个非根非叶聚类。叶聚类可代表细胞的表达谱。非根非叶聚类可代表由该非根非叶聚类的子聚类所代表的细胞的表达谱。根聚类可代表其子聚类的表达谱。多个叶聚类和多个非根非叶聚类中的每个可与母聚类关联。根聚类和多个非根非叶聚类中的每个可与左子聚类和右子聚类关联，且代表由聚类的左子聚类和右子聚类所代表的表达谱。根聚类可代表多个细胞的表达谱。

[0024] 在一些实施例中，对于具有两个或更多个子聚类的每个聚类，如果该聚类与该两个或更多个子聚类之间的关联无效则将该聚类添加到合并聚类集包括：当从树状图的根聚类到树状图的多个叶聚类遍历树状图的每个聚类时：(1)确定该聚类与该聚类的子聚类的关联是有效还是无效的；和(2)如果关联无效，则将该聚类添加到合并聚类集。

[0025] 在一些实施例中，该方法包括：在(a)中接收多个细胞的表达谱之前：(f)使用多个条形码对该多个细胞中的该多个靶进行条形码化以创建多个经条形码化的靶，其中该多个条形码中的每一个包括细胞标记和分子标记，其中从不同细胞的靶创建的经条形码化的靶具有不同的细胞标记，并且其中从该多个细胞中一个细胞的靶创建的经条形码化的靶具有不同的分子标记；(g)获得该多个经条形码化的靶的测序数据；以及(h)对于该多个细胞的每个：(1)针对该细胞，对测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数；以及(2)基于(h)(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目，估算该细胞的多个靶中每个靶的数目。例如，该方法可包括：在(a)中接收多个细胞的表达谱之前：步骤(s)(f)使用多个随机条形码随机地对该多个细胞的多个靶进行条形码编码以创建多个经随机经条形码化的靶，其中该多个经随机条形码中的每一个包含细胞标记和分子标记，其中从不同细胞的靶创建的随机经条形码化的靶具有不同的细胞标记，且其中从该个多个细胞中的一细胞的靶创建的经随机经条形码化的靶具有不同的分子标记；和/或(g)获得该多个经随机经条形码化的靶的测序数据。

[0026] 在一些实施例中，多个细胞中该细胞的表达谱包含(h)(2)中估算的该细胞的多个靶中每个靶的数目。在一些实施例中，该方法包括，在(b)中基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类之前：(i)确定多个细胞的表达谱的距离数据结构。该距离数据结构可包含多个细胞的表达谱的距离矩阵。该距离矩阵的每个对角元素具有零值。在(b)中基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类可包括：基于距离矩阵将多个细胞的表达谱聚类以产生多个表达谱

聚类。多个细胞的表达谱之间的距离可以是多个细胞的表达谱之间的成对相关距离。

[0027] 在一些实施例中，所述方法包括，在(i)中确定距离数据结构之前，将靶计数数据结构对数转换为经对数转换的靶计数数据结构，其中确定靶计数数据结构的元素的距离数据结构包括确定该经对数转换的靶计数数据结构的距离数据结构，并且其中(b)中基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类包括基于该经对数转换的靶计数数据结构和该距离数据结构将多个细胞的表达谱聚类以产生多个聚类。将靶计数数据结构对数转换为经对数转换的靶计数数据结构可以包括通过增量增加靶计数数据结构的每个元素的值。该增量可以是1。

[0028] 在一些实施例中，在(b)中基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类包括：将多个细胞的每个表达谱分配给多个聚类中不同的叶聚类；且迭代地将多个聚类的第一聚类和第二聚类组合以生成该第一聚类和该第二聚类的母聚类，前提是如果该第二聚类是多个聚类中最接近第一聚类的聚类的话。该第一聚类和该第二聚类之间的距离可以是由该第一聚类代表的任意表达谱与由该第二聚类代表的任意表达谱之间的最大距离。

[0029] 在一些实施例中，第一聚类的聚类内相关性和第二聚类的聚类内相关性高于第一聚类和第二聚类的聚类间相关性。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度或指标高于第一聚类和第二聚类的聚类间相关性。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度可以基于以下的至少一种：第一聚类和第二聚类的聚类内最大相关性，第一聚类和第二聚类的聚类内平均相关性，第一聚类和第二聚类的聚类内中值相关性，第一聚类和第二聚类的聚类内最小相关性，以及其任何组合。第一聚类的聚类内相关性可以基于以下的至少一种：第一聚类的聚类内最大相关性，第一聚类的聚类内平均相关性，第一聚类的聚类内中值相关性，第一聚类的聚类内最小相关性，以及其任何组合。第二聚类的聚类内相关性可以基于以下的至少一种：第二聚类的聚类内最大相关性，第二聚类的聚类内平均相关性，第二聚类的聚类内中值相关性，第二聚类的聚类内最小相关性，以及其任何组合。第一聚类和第二聚类的聚类间相关性可以基于以下的至少一种：第一聚类和第二聚类的聚类间最大相关性，第一聚类和第二聚类的聚类间平均相关性，第一聚类和第二聚类的聚类间中值相关性，第一聚类和第二聚类的聚类间最小相关性，以及其任何组合。

[0030] 在一些实施例中，在遍历树状图的多个聚类时该方法在每个聚类处包括：如果关联有效，则继续从该聚类遍历到该聚类的左子聚类和右子聚类；且如果关联无效，则停止从该聚类遍历到该聚类的左子聚类和右子聚类。确定该聚类与该聚类的子聚类的关联是有效还是无效可包括：如果左子聚类和右子聚类之间的距离高于关联阈值，则确定该关联有效，否则为无效。

[0031] 一些实施例中，左子聚类和右子聚类之间的距离可以基于左子聚类和右子聚类代表的表达谱之间对多个靶的每个靶实施的统计学检验来确定。所述统计学检验可以包括Welch t检验。左子聚类和右子聚类之间的距离可以基于在由左子聚类代表的表达谱和右子聚类代表的每种表达谱之间对多个靶的每个靶实施的统计学检验的最大p值来确定。

[0032] 在一些实施例中，确定该聚类与该聚类的子聚类的关联是有效还是无效可包括：如果第一聚类的聚类内相关性和第二聚类的聚类内相关性中的至少一个高于第一聚类和第二聚类的聚类间相关性，则确定该关联有效，否则为无效。在一些实施例中，确定该聚类

与该聚类的子聚类的关联是有效还是无效可包括:如果第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度或指标高于第一聚类和第二聚类的聚类间相关性,则确定该关联有效。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度可以基于以下的至少一种:第一聚类和第二聚类的聚类内最大相关性,第一聚类和第二聚类的聚类内平均相关性,第一聚类和第二聚类的聚类内中值相关性,第一聚类和第二聚类的聚类内最小相关性,以及其任何组合。第一聚类的聚类内相关性可以基于以下的至少一种:第一聚类的聚类内最大相关性,第一聚类的聚类内平均相关性,第一聚类的聚类内中值相关性,第一聚类的聚类内最小相关性,以及其任何组合。第二聚类的聚类内相关性可以基于以下的至少一种:第二聚类的聚类内最大相关性,第二聚类的聚类内平均相关性,第二聚类的聚类内中值相关性,第二聚类的聚类内最小相关性,以及其任何组合。第一聚类和第二聚类的聚类间相关性可以基于以下的至少一种:第一聚类和第二聚类的聚类间最大相关性,第一聚类和第二聚类的聚类间平均相关性,第一聚类和第二聚类的聚类间中值相关性,第一聚类和第二聚类的聚类间最小相关性,以及其任何组合。

[0033] 在一些实施例中,在遍历树状图的多个聚类时该方法在每个聚类处包括:(3)如果该聚类代表单细胞的表达谱,则将该聚类添加到合并聚类集。在遍历树状图的多个聚类时该方法在每个聚类处包括:为该聚类分配聚类标记。在一些实施例中,如果该聚类代表单细胞的表达谱,则该聚类的聚类标记包括单细胞标识,否则如果该聚类是母聚类的左子聚类,则该聚类的聚类标记包括母聚类的聚类标记和左标识,否则该聚类的聚类标记包括母聚类的聚类标记和右标识。

[0034] 在一些实施例中,对于合并聚类集中的每个聚类,基于由该聚类代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶包括:确定由该聚类代表的表达谱和由合并聚类集中的另一聚类代表的表达谱之间在与用于区分细胞类型的靶相关联的具有不同序列的分子标记的数目方面的差大于显著性阈值。

[0035] 在一些实施例中,该方法包括,在(d)中在将第一聚类与第二聚类合并以生成合并的聚类之前:将合并聚类集中代表单细胞的表达谱的每个第三聚类与合并聚类集中的第四聚类合并,前提是如果该第三聚类和该第四聚类之间的距离在聚类距离阈值内的话。该方法包括基于代表细胞表达谱的合并聚类集中的聚类对多个细胞进行分类。该方法可以包括基于所鉴定的用于区分细胞类型的靶设计全转录组测定,或基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

[0036] 本文中公开了鉴定用于区分细胞类型的靶的系统。在一些实施例中,该系统包含:硬件处理器;和具有存储在其上的指令的非暂时性存储器,这些指令当通过该硬件处理器执行时使得该处理器执行本文公开的任何方法。本文中公开了鉴定用于区分细胞类型的靶的计算机可读介质。在一些实施例中,该计算机可读介质包含用于实施本文中公开的任何方法的代码。

[0037] 本文中公开了鉴定用于区分细胞类型的靶的系统的实施例。在一些实施例中,该系统包含:配置为存储可执行指令的非暂时性存储器,和与该非暂时性存储器通讯的硬件处理器,该硬件处理器由可执行指令程序化设定为:(a)接收靶计数数据结构,其中该靶计数数据结构包含多个细胞的表达谱,并且其中该多个细胞的表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目;(b)基于靶计数数据结构和该多个细胞的表达谱之间

的距离对该多个细胞的表达谱进行层次聚类，以产生代表该多个细胞的表达谱的树状图，其中该树状图包含多个节点，其中该多个节点包括根节点、多个叶节点和多个非根非叶节点，其中该多个叶节点中的每个叶节点代表该多个细胞中不同细胞的表达谱，且其中该根节点代表该多个细胞的表达谱；(c)在从该树状图的根节点到该树状图的多个叶节点遍历该树状图的多个节点的每个节点时：(1)确定该节点分裂成该节点的子节点是有效还是无效的；和(2)如果将节点分裂成该节点的子节点是无效的，则将该节点添加到合并聚类集中；(d)迭代地，对于该合并聚类集中的每个第一节点，如果合并聚类集中的第一节点与最接近该第一节点的合并聚类集中的第二节点之间的距离在合并距离阈值内，则将该第一节点与第二节点合并以生成合并的节点，其包含由该第一节点和该第二节点代表的表达谱；和(e)对于合并聚类集中的每个节点，基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0038] 在一些实施例中，该靶计数数据结构包含靶计数矩阵。该靶计数矩阵的每行或每列可包含针对该多个细胞中不同个体细胞的多个靶中每个靶的数目。多个叶节点和多个非根非叶节点中的每一个可以与母节点相关联，并且根节点和多个非根非叶节点中的每一个可以与左子节点和右子节点相关联，并代表由该节点的左子节点和右子节点代表的表达谱。

[0039] 在一些实施例中，该硬件处理器可程序化设定为在(a)中接收靶计数数据结构之前：(f)使用多个条形码对该多个细胞中的多个靶进行条形码编码以创建多个经条形码化的靶，其中该多个条形码中的每一个包括细胞标记和分子标记，其中从不同细胞的靶创建的经条形码化的靶具有不同的细胞标记，并且其中从该多个细胞中一个细胞的靶创建的经条形码化的靶具有不同的分子标记；(g)获得该多个经条形码化的靶的测序数据；以及(h)对于该多个细胞的每个：(1)针对该细胞，对测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数；以及(2)基于(h)(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目，估算该细胞的多个靶中每个靶的数目。为了接收靶计数数据结构，硬件处理器可以程序化设定为：对于，从(h)(2)中估算的该细胞的多个靶中每个靶的数目生成靶计数数据结构，其中该多个细胞中该细胞的表达谱包含(h)(2)中估算的该细胞的多个靶中每个靶的数目。

[0040] 在一些实施例中，硬件处理器可以程序化设定为，在(b)中基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类以产生代表该多个细胞的表达谱的树状图之前：(i)确定该靶计数数据结构的元素的距离数据结构，其中该距离数据结构包含该多个细胞的表达谱之间的距离。该距离数据结构包含距离矩阵。该距离矩阵的每个对角元素具有零值。

[0041] 在一些实施例中，为了在(b)中基于靶计数数据结构和多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类以产生代表多个细胞的表达谱的树状图，硬件处理器可以程序化设定为：基于靶计数数据结构和距离数据结构对多个细胞的表达谱进行层次聚类。该多个细胞的表达谱之间的距离可以包含该多个细胞的表达谱之间的成对相关距离。

[0042] 在一些实施例中，硬件处理器可以程序化设定为，在(i)中确定靶计数数据结构的元素的距离数据结构之前，将靶计数数据结构对数转换为经对数转换的靶计数数据结构。

为了确定靶计数数据结构的元素的距离数据结构，硬件处理器可以程序化设定为：确定经对数转换的靶计数数据结构的距离数据结构。为了在(b)中基于靶计数数据结构和多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类，硬件处理器可以程序化设定为：基于该经对数转换的靶计数数据结构和距离数据结构对多个细胞的表达谱进行层次聚类以产生树状图。为了将靶计数数据结构对数转换为经对数转换的靶计数数据结构，硬件处理器可以程序化设定为：通过增量增加靶计数数据结构的每个元素的值。该增量可以是1。

[0043] 为了在(b)中基于靶计数数据结构和多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类，硬件处理器可以程序化设定为：将该多个细胞的每个表达谱分配给不同的叶节点；且迭代地将多个节点的第一节点和第二节点组合以生成该第一节点和该第二节点的母节点，如果该第二节点是多个节点中最接近第一节点的节点的话。第一节点和第二节点之间的距离可以是具有由第一节点代表的表达谱的任何细胞与具有由第二节点代表的表达谱的任何细胞之间的最大距离。

[0044] 在一些实施例中，该第一节点的节点内相关性和该第二节点的节点内相关性中的至少一个可以高于该第一节点和该第二节点的节点间相关性。该第一节点的节点内相关性和该第二节点的节点内相关性的量度或指标可以高于该第一节点和该第二节点的节点间相关性。第一节点的节点内相关性和第二节点的节点内相关性的量度可以基于以下中的至少一种：第一节点和第二节点的节点内最大相关性，第一节点和第二节点的节点内平均相关性，第一节点和第二节点的节点内中值相关性，第一节点和第二节点的节点内最小相关性，以及其任何组合。第一节点的节点内相关性可以基于以下中的至少一种：第一节点的节点内最大相关性，第一节点的节点内平均相关性，第一节点的节点内中值相关性，第一节点的节点内最小相关性，以及其任何组合。第二节点的节点内相关性可以基于以下中的至少一种：第二节点的节点内最大相关性，第二节点的节点内平均相关性，第二节点的节点内中值相关性，第二节点的节点内最小相关性，以及其任何组合。第一节点和第二节点的节点间相关性可以基于以下中的至少一种：第一节点和第二节点的节点间最大相关性，第一节点和第二节点的节点间平均相关性，第一节点和第二节点的节点间中值相关性，第一节点和第二节点的节点间最小相关性，以及其任何组合。

[0045] 在一些实施例中，硬件处理器可以程序化设定为在遍历树状图的多个节点时在每个节点处：如果分裂有效，则继续从该节点遍历到该节点的左子节点和右子节点；如果分裂无效，则停止从该节点遍历到该节点的左子节点和右子节点。为了确定节点分裂成该节点的子节点是有效还是无效的，硬件处理器可以程序化设定为：如果左子节点和右子节点之间的距离高于分裂阈值，则确定分裂有效，否则无效。左子节点和右子节点之间的距离可以基于对左子节点和右子节点代表的表达谱之间多个靶中每个靶实施的统计学检验来确定。所述统计学检验可以包括Welch t检验。左子节点和右子节点之间的距离可以基于对左子节点代表的每个表达谱和右子节点代表的每个表达谱之间多个靶中每个靶实施的统计学检验的最大p值来确定。

[0046] 在一些实施例中，硬件处理器可以程序化设定为在遍历树状图的多个节点时在每个节点处：(3)如果该节点代表单细胞的表达谱，则将该节点添加到合并聚类集。在一些实施例中，在遍历树状图的多个节点时在每个节点处，硬件处理器可以程序化设定为：为该节

点分配节点标记。如果该节点表示单细胞的表达谱，则该节点的节点标记可以包括单细胞标识，否则如果该节点是母节点的左子节点，则该节点的节点标记可以包括母节点的节点标记和左标识，否则该节点的节点标记可以包括母节点的节点标记和右标识。

[0047] 在一些实施例中，对于合并聚类集中的每个节点，基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶，硬件处理器可以程序化设定为：确定由该节点代表的表达谱和由合并聚类集中的另一节点代表的表达谱之间在与用于区分细胞类型的靶相关联的具有不同序列的分子标记的数目方面的差大于显著性阈值。

[0048] 在一些实施例中，硬件处理器可以程序化设定为：在(d)中将第一节点与第二节点合并以生成合并的节点之前：将合并聚类集中代表单细胞的表达谱的每个第三节点与合并聚类集中的第四节点合并，前提是如果该第三节点和该第四节点之间的距离在节点距离阈值内的话。硬件处理器可以程序化设定为：基于合并聚类集中代表细胞表达谱的节点对多个细胞进行分类。硬件处理器可以程序化设定为：基于所鉴定的用于区分细胞类型的靶设计全转录组测定。硬件处理器可以程序化设定为：基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

[0049] 本文中公开了鉴定用于区分细胞类型的靶的系统的实施例。在一些实施例中，该系统包含：配置为存储可执行指令的非暂时性存储器，和与该非暂时性存储器通讯的硬件处理器，该硬件处理器由可执行指令程序化设定为：(a)接收多个细胞的表达谱，其中所述表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目；(b)基于该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行聚类以产生多个表达谱聚类，其中每个聚类与(1)母聚类和(2)两个或更多个子聚类中的一者或两者具有一个或多个关联，其中该母聚类代表由该聚类代表的该多个细胞的一个或多个细胞的表达谱，且其中该聚类代表由两个或更多个子聚类代表的表达谱；(c)对于具有两个或更多个子聚类的每个聚类，如果该聚类与该两个或更多个子聚类之间的关联无效则将该聚类添加到合并聚类集；(d)迭代地，对于合并聚类集中的每个第一聚类，如果合并聚类集中的第一聚类与最接近该第一聚类的合并聚类集中的第二聚类之间的距离在合并距离阈值内，则将该第一聚类和该第二聚类合并以生成合并的聚类，其中该合并聚类包含该第一聚类和该第二聚类的表达谱；和(e)对于合并聚类集中的每个聚类，基于该聚类所代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0050] 在一些实施例中，硬件处理器可以程序化设定为：接收该多个细胞的表达谱包括接收靶计数数据结构。该靶计数数据结构可以包含靶计数矩阵。该靶计数矩阵的每行或每列可包含多个细胞中不同个体细胞的表达谱。

[0051] 在一些实施例中，为了基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类成多个表达谱聚类，硬件处理器可程序化设定为：基于多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类，以产生代表多个细胞的表达谱的树状图，其中该树状图包含多个聚类，其中该多个聚类包括根聚类、多个叶聚类和多个非根非叶聚类。多个叶聚类和多个非根非叶聚类中的每个可与母聚类关联。根聚类和多个非根非叶聚类中的每个可与左子聚类和右子聚类关联，且代表由聚类的左子聚类和右子聚类所代表的表达谱。根聚类可代表多个细胞的表达谱。对于具有两个或更多个子聚类的每个聚类，如果该聚类与该两个或更多个子聚类之间的关联无效则将该聚类添加到合并聚类集，当从树状图的根聚类到

树状图的多个叶聚类遍历树状图的每个聚类时硬件处理器可程序化设定为:(1)确定该聚类与该聚类的子聚类的关联是有效还是无效的;和(2)如果关联无效,则将该聚类添加到合并聚类集。

[0052] 在一些实施例中,硬件处理器可程序化设定为,在(a)中接收多个细胞的表达谱之前:(f)使用多个条形码对该多个细胞中的多个靶进行条形码编码以创建多个经条形码化的靶,其中该多个条形码中的每一个包括细胞标记和分子标记,其中从不同细胞的靶创建的经条形码化的靶具有不同的细胞标记,并且其中从该多个细胞中一个细胞的靶创建的经条形码化的靶具有不同的分子标记;(g)获得该多个经条形码化的靶的测序数据;以及(h)对于该多个细胞的每个:(1)针对该细胞,对测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数;以及(2)基于(h)(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目,估算该细胞的多个靶中每个靶的数目。多个细胞中该细胞的表达谱可以包含(h)(2)估算的该细胞的多个靶中每个靶的数目。

[0053] 在一些实施例中,硬件处理器可程序化设定为,在(b)中基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类之前:(i)确定多个细胞的表达谱的距离数据结构。该距离数据结构可包含多个细胞的表达谱的距离矩阵。该距离矩阵的每个对角元素可以具有零值。为了在(b)中基于多个细胞的表达谱之间的距离对多个细胞的表达谱进行聚类以产生多个表达谱聚类,硬件处理器可程序化设定为:基于距离矩阵将多个细胞的表达谱聚类以产生多个表达谱聚类。多个细胞的表达谱之间的距离可以是多个细胞的表达谱之间的成对相关距离。

[0054] 在一些实施例中,硬件处理器可程序化设定为,在(i)中确定距离数据结构之前,将靶计数数据结对数转换成经对数转换的靶计数数据结构。为了确定靶计数数据结构的元素的距离数据结构,硬件处理器可以程序化设定为:确定经对数转换的靶计数数据结构的距离数据结构。为了在(b)中基于多个细胞的表达谱之间的距离对多个细胞的表达谱进行聚类以产生多个表达谱聚类,硬件处理器可程序化设定为:基于该经对数转换的靶计数数据结构和距离数据结构对多个细胞的表达谱进行聚类以产生多个聚类。为了将靶计数数据结构对数转换为经对数转换的靶计数数据结构,硬件处理器可以程序化设定为:通过增量增加靶计数数据结构的每个元素的值。该增量可以是1。

[0055] 在一些实施例中,为了在(b)中基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类,硬件处理器可程序化设定为:将多个细胞的每个表达谱分配给多个聚类中不同的叶聚类;且迭代地将多个聚类的第一聚类和第二聚类组合以生成该第一聚类和该第二聚类的母聚类,如果该第二聚类是多个聚类中最接近第一聚类的聚类的话。该第一聚类和该第二聚类之间的距离可以是由该第一聚类代表的任意表达谱与由该第二聚类代表的任意表达谱之间的最大距离。

[0056] 在一些实施例中,第一聚类的聚类内相关性和第二聚类的聚类内相关性高于第一聚类和第二聚类的聚类间相关性。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度或指标可以高于第一聚类和第二聚类的聚类间相关性。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度可以基于以下的至少一种:第一聚类和第二聚类的聚类内最大相关性,第一聚类和第二聚类的聚类内平均相关性,第一聚类和第二聚类的聚类内中值相关性,第一聚类和第二聚类的聚类内最小相关性,以及其任何组合。第一聚类的聚类内相

关性可以基于以下的至少一种:第一聚类的聚类内最大相关性,第一聚类的聚类内平均相关性,第一聚类的聚类内中值相关性,第一聚类的聚类内最小相关性,以及其任何组合。第二聚类的聚类内相关性可以基于以下的至少一种:第二聚类的聚类内最大相关性,第二聚类的聚类内平均相关性,第二聚类的聚类内中值相关性,第二聚类的聚类内最小相关性,以及其任何组合。第一聚类和第二聚类的聚类间相关性可以基于以下的至少一种:第一聚类和第二聚类的聚类间最大相关性,第一聚类和第二聚类的聚类间平均相关性,第一聚类和第二聚类的聚类间中值相关性,第一聚类和第二聚类的聚类间最小相关性,以及其任何组合。

[0057] 在一些实施例中,硬件处理器可以程序化设定为在遍历树状图的多个聚类时在每个聚类处:如果关联有效,则继续从该聚类遍历到该聚类的左子聚类和右子聚类;且如果关联无效,则停止从该聚类遍历到该聚类的左子聚类和右子聚类。为了确定聚类与该聚类的子聚类的关联是有效还是无效,硬件处理器可程序化设定为:如果左子聚类和右子聚类之间的距离高于关联阈值,则确定该关联有效,否则为无效。左子聚类和右子聚类之间的距离可以基于左子聚类和右子聚类代表的表达谱之间对多个靶的每个靶实施的统计学检验来确定。所述统计学检验可以包括Welch t检验。左子聚类和右子聚类之间的距离可以基于在由左子聚类代表的表达谱和右子聚类代表的每种表达谱之间对多个靶的每个靶实施的统计学检验的最大p值来确定。

[0058] 在一些实施例中,硬件处理器可以程序化设定为在遍历树状图的多个聚类时在每个聚类处:(3)如果该聚类代表单细胞的表达谱,则将该聚类添加到合并聚类集。硬件处理器可以程序化设定为:在遍历树状图的多个聚类时,在每个聚类处:为该聚类分配聚类标记。如果该聚类代表单细胞的表达谱,则该聚类的聚类标记包括单细胞标识,否则如果该聚类是母聚类的左子聚类,则该聚类的聚类标记包括母聚类的聚类标记和左标识,否则该聚类的聚类标记包括母聚类的聚类标记和右标识。

[0059] 在一些实施例中,对于合并聚类集中的每个聚类,基于由该聚类代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶,硬件处理器可以程序化设定为:确定由该聚类代表的表达谱和由合并聚类集中的另一聚类代表的表达谱之间在与用于区分细胞类型的靶相关联的具有不同序列的分子标记的数目方面的差大于显著性阈值。硬件处理器可以程序化设定为,在(d)中将第一聚类与第二聚类合并以生成合并的聚类之前:将合并聚类集中代表单细胞的表达谱的每个第三聚类与合并聚类集中的第四聚类合并,前提是如果该第三聚类和该第四聚类之间的距离在聚类距离阈值内的话。

[0060] 在一些实施例中,硬件处理器可以程序化设定为:基于代表细胞表达谱的合并聚类集中的聚类对多个细胞进行分类。硬件处理器可以程序化设定为:基于所鉴定的用于区分细胞类型的靶设计全转录组测定。硬件处理器可以程序化设定为:基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

[0061] 附图简要说明

[0062] 图1说明了非限制性示例性条形码(例如随机条形码)。

[0063] 图2显示了进行条形码化和数字计数(例如进行随机条形码化和数字计数)的非限制性示例性工作流程。

[0064] 图3是显示用于从多个靶产生经(例如经条形码化的靶随机经条形码化的靶)的索

引文库的非限制性示例性过程的示意图。

[0065] 图4是显示通过使用树状图对细胞的表达谱聚类来鉴定用于区分细胞类型的靶的非限制性示例性方法的流程图。

[0066] 图5是示例性树状图的示意图。

[0067] 图6是显示通过对细胞的表达谱聚类来鉴定用于区分细胞类型的靶的非限制性示例性方法的流程图。

[0068] 图7是配置为执行本公开的方法的例示性计算系统的方框图。

[0069] 图8小图(a)-(d)显示在分裂和合并单细胞的表达谱之后二维空间中表达谱的非限制性示例性图。

[0070] 图9小图(a)-(x)是二维空间中表达谱的非限制性示例性图,显示了怎样能决定分裂。

[0071] 图10显示了在第五次分裂循环后二维空间中表达谱的非限制性示例性图。

[0072] 图11小图(a)-(l)是二维空间中表达谱的非限制性示例性图,显示了为何树状图中的某些分裂被保留用于图10中所示的第五次分裂循环。

[0073] 图12小图(a)-(i)是二维空间中表达谱的非限制性示例性图,显示了怎样能决定合并。

[0074] 图13显示了第二次合并循环后二维空间中表达谱的非限制性示例性图。

[0075] 图14小图(a)-(d)是二维空间中表达谱的非限制性示例性图,显示了图13中所示的第二次合并循环是怎样决定的。

[0076] 图15小图(a)-(f)是显示差异表达分析的一个非限制性示例类型的图。

[0077] 图16小图(a)-(o)是显示差异表达分析的另一个非限制性示例类型的图。

[0078] 图17小图(a)-(g)是使聚类之间距离可视化的非限制性示例图。

[0079] 图18小图(a)-(e)显示了非限制性示例性树状图。

[0080] 图19小图(a)-(s)是显示参数扫描的非限制性示例图。

[0081] 图20是显示怎样使用参数扫描来鉴定阈值的非限制性示例图。

[0082] 图21小图(a)-(j)是显示第一次分裂结果的非限制性示例图。

[0083] 图22是例示二维空间中表达谱的分裂结果的非限制性示例图。

[0084] 图23小图(a)-(d)显示了非限制性示例性树状图,其显示了被分为两个聚类的表达谱。

[0085] 图24是显示参数扫描的另一个非限制性示例图。

具体实施方式

[0086] 在以下详细说明中参考了形成本文的一部分的附图。在附图中,除非上下文另有指示,否则相似的符号通常标识相似的组件。在具体实施例、附图和权利要求中描述的说明性实施例不意味着是限制性的。在不脱离本文提出的主题的精神或范围的情况下,可以利用其他实施例,并且可以做出其他改变。容易理解的是,如本文一般描述的以及图中说明的本披露的方面能以各种不同的配置来布置、替换、组合、分离和设计,所有这些都在本文中明确考虑并且构成本披露内容的一部分。

[0087] 来自GenBank的所有专利、公开的专利申请、其他出版物、和序列,以及本文提及的

其他数据库关于相关技术通过引用以其整体并入。

[0088] 对少量核酸或靶(例如信使核糖核苷酸(mRNA)分子)进行量化对于确定例如在不同发育阶段或在不同环境条件下在细胞中表达的基因是临床上重要的。然而,确定核酸分子(例如,mRNA分子)的绝对数目也是非常具有挑战性的,尤其是当分子数目非常小时。确定样品中分子的绝对数目的一种方法是数字聚合酶链式反应(PCR)。可以将具有独特分子标记(ML,也称为分子指数(MI))的条形码(例如随机条形码)用于计数分子数目。对每个细胞标记独特的具有分子标记的条形码可用于计数每个细胞中的分子数目。条形码化(例如随机条形码化)的非限制性示例测定法包括Precise™测定法(Cellular Research,Inc.(Palo Alto,CA))、Resolve™测定法(Cellular Research,Inc.(Palo Alto,CA))、或Rhapsody™测定法(Cellular Research,Inc.(Palo Alto,CA))。

[0089] Rhapsody™测定法可利用具有大量的(例如6561至65536个)条形码(例如随机条形码)的非耗尽性池、聚(T)寡核苷酸上的独特分子标记,以在RT步骤期间与样品中的所有聚(A)-mRNA杂交。除了分子标记,可以使用条形码的细胞标记来鉴定微孔板的每个孔中的每个单细胞。条形码(例如随机条形码)可包括通用PCR引发位点。在RT期间,靶基因分子与条形码随机地反应。每个靶分子可以与条形码杂交,从而产生经条形码化的互补核糖核苷酸(cDNA)分子(例如经随机条形码化的cDNA分子)。在标记后,可将来自微孔板微孔的经条形码化的cDNA分子合并到单个管中用于PCR扩增和测序。可以分析原始测序数据来得到具有独特分子标记的条形码(例如随机条形码)的数目。

[0090] 本文中公开了用于鉴定靶以区分细胞类型的方法。在一些实施例中,该方法包括:(a)接收靶计数数据结构,其中该靶计数数据结构包含多个细胞的表达谱,并且其中该多个细胞的表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目;(b)基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类,以产生代表该多个细胞的表达谱的树状图,其中该树状图包含多个节点,其中该多个节点包括根节点、多个叶节点和多个非根非叶节点,其中该多个叶节点中的每个叶节点代表该多个细胞中不同细胞的表达谱,且其中该根节点代表该多个细胞的表达谱;(c)在从该树状图的根节点到该树状图的多个叶节点遍历该树状图的多个节点的每个节点时:(1)确定节点分裂成该节点的子节点是有效还是无效的(例如,子节点之间的差异不显著);和(2)如果将该节点分裂成该节点的子节点是无效的,则将该节点添加到合并聚类集中;(d)迭代地,对于该合并聚类集中的每个第一节点,如果该合并聚类集中的该第一节点与该合并聚类集中最接近该第一节点的第二节点之间的距离在合并距离阈值内,则将该第一节点与该第二节点合并以生成合并的节点,该合并的节点包含由该第一节点和该第二节点代表的表达谱;和(e)对于该合并聚类集中的每个节点,基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0091] 本文中公开了用于鉴定靶以区分细胞类型的系统。在一些实施例中,该系统包含:硬件处理器;和具有存储在其上的指令的非暂时性存储器,这些指令当通过该硬件处理器执行时使得该处理器执行本文公开的任何方法。本文中公开了鉴定用于区分细胞类型的靶的计算机可读介质。在一些实施例中,该计算机可读介质包含用于实施本文中公开的任何方法的代码。

[0092] 定义

[0093] 除非另外定义,本文所用的技术术语和科学术语具有与本披露所属领域的普通技术人员通常所理解的相同意义。参见,例如,Singleton等人,Dictionary of Microbiology and Molecular Biology[微生物学和分子生物学词典],第2版,约翰·威利父子出版公司(J.Wiley&Sons),(纽约,纽约州1994);Sambrook等人,Molecular Cloning,A Laboratory Manual[分子克隆:实验室手册],冷泉港实验室出版社(冷泉港,纽约州1989)。出于本披露的目的,以下术语定义如下。

[0094] 如本文所用的,术语"衔接子"可以意指促进相关联的核酸的扩增或测序的序列。相关联的核酸可包括靶核酸。相关联的核酸可包含空间标记、靶标记、样品标记、索引标记、条形码、随机条形码、或分子标记中的一个或多个。衔接子可以是线性的。衔接子可以是预腺苷酸化的衔接子。衔接子可以是双链或单链的。一个或多个衔接子可以位于核酸的5'或3'端。当衔接子在5'和3'端包括已知序列时,已知序列可以是相同或不同的序列。位于多核苷酸的5'和/或3'端的衔接子能够与固定在表面上的一个或多个寡核苷酸杂交。在一些实施例中,衔接子可包括通用序列。通用序列可以是两个或更多个核酸分子共有的核苷酸序列的区域。两个或更多个核酸分子可具有不同序列的区域。因此,例如,5'衔接子可包括相同和/或通用核酸序列,且3'衔接子可包括相同和/或通用序列。可存在于多个核酸分子的不同成员中的通用序列可允许使用与通用序列互补的单个通用引物复制或扩增多个不同序列。相似地,可以存在于核酸分子的集合中的不同成员中的至少一个、两个(例如,一对)或更多个通用序列可以允许使用与通用序列互补的至少一个、两个(例如,一对)或更多个单个通用引物复制或扩增多个不同序列。因此,通用引物包括可与此类通用序列杂交的序列。可以修饰携带靶核酸序列的分子以将通用衔接子(例如,非靶核酸序列)附接至不同靶核酸序列的一端或两端。与靶核酸附接的一个或多个通用引物可以提供通用引物杂交的位点。与靶核酸附接的一个或多个通用引物可以彼此相同或不同。

[0095] 如本文使用的,术语"关联"或"与……相关联"可意指两个或更多个种类可以被鉴定为在某个时间点处共定位。关联可意指两个或更多个种类在或曾经在相似的容器内。关联可以是信息学关联,其中例如关于两个或更多个种类的数字信息被存储并且可以用于确定所述种类中的一个或多个在某个时间点处共定位。关联可以是物理关联。在一些实施例中,两个或更多个相关联的种类彼此之间或与共同的固体或半固体表面是"连接的"、"附接的"或"固定的"。关联可以指用于将标记附接到固体或半固体支持物(如珠)上的共价或非共价方式。关联可以是靶与标记之间的共价键。

[0096] 如本文使用的,术语"互补性"可以指两个核苷酸之间精确配对的能力。例如,如果核酸的在给定位置的核苷酸能够与另一个核酸的核苷酸以氢键结合,则两个核酸被认为在所述位置处是彼此互补的。两单链核酸分子之间的互补性可以是"部分的",其中该核苷酸中仅一些结合,或者当该单链分子之间存在完全互补性时,这种互补性可以是完全的。如果第一核苷酸序列与第二核苷酸序列互补,则可以认为第一核苷酸序列是第二序列的"互补体"。如果第一核苷酸序列互补于和第二序列相反的序列(即,核苷酸顺序相反),则可以认为第一核苷酸序列是第二序列的"反向互补体"。如本文使用的,术语"互补体"、"互补"和"反向互补体"可以互换使用。从本披露可以理解,如果一个分子可以与另一个分子杂交,则其可以是杂交的分子的互补体。

[0097] 如本文使用的,术语"数字计数"可以指用于估计样品中靶分子数目的方法。数字

计数可以包括确定已经与样品中的靶相关联的独特标记的数目的步骤。这种随机方法将计数分子的问题从相同分子的定位和鉴定之一转化为有关检测到一组预定义标记的一系列是/否数字问题。

[0098] 如本文使用的，术语(多个)“标记”可以指与样品中的靶相关联的核酸代码。标记可以是例如核酸标记。标记可以是完全或部分可扩增的标记。标记可以是完全或部分可测序的标记。标记可以是可鉴定为有区别的天然核酸的一部分。标记可以是已知的序列。标记可以包括核酸序列的接点，例如天然和非天然序列的接点。如本文使用的，术语“标记”可以与术语“索引”、“标签”或“标记-标签”互换使用。标记可以传达信息。例如，在各种实施例中，可以使用标记来确定样品的身份、样品的来源、细胞的身份和/或靶。

[0099] 如本文使用的，术语“非耗尽性储库(non-depleting reservoir)”可以指由许多不同标记组成的随机条形码池。非耗尽性储库可以包括大量不同的随机条形码，使得当非耗尽性储库与靶池相关联时，每个靶可能与独特的随机条形码相关联。每个经标记的靶分子的独特性可以通过随机选择的统计来确定，并且取决于与多样的标记相比在集合中相同的靶分子的拷贝数。所得的经标记的靶分子集合的大小可以通过条形码化处理的随机性质来确定，然后对检测到的随机条形码的数目的分析允许计算原始集合或样品中存在的靶分子的数目。当存在的靶分子的拷贝数与独特的随机条形码的数目的比率低时，经标记的靶分子是高度独特的(即，用给定的标记来标记多于一个靶分子的概率非常低)。

[0100] 如本文所用的，术语“核酸”是指多核苷酸序列、或其片段。核酸可包括核苷酸。核酸对于细胞可以是外源的或内源的。核酸可以存在于无细胞环境中。核酸可以是基因或其片段。核酸可以是DNA。核酸可以是RNA。核酸可以包括一种或多种类似物(例如改变的骨架、糖或核碱基)。类似物的一些非限制性实例包括：5-溴尿嘧啶、肽核酸、外来核酸、吗啉代、锁核酸、二醇核酸、苏糖核酸、二脱氧核苷酸、虫草菌素、7-脱氮-GTP、荧光团(例如，罗丹明或与糖连接的荧光黄素)、含有核苷酸的硫醇、生物素连接的核苷酸、荧光基类似物、CpG岛、甲基-7-鸟苷、甲基化的核苷酸、肌苷、硫代尿苷、假尿苷、二氢尿苷、辫苷、以及怀俄苷。“核酸”、“多核苷酸”、“靶多核苷酸”和“靶核酸”可以互换使用。

[0101] 核酸可以包括一种或多种修饰(例如，碱基修饰、骨架修饰)，以为核酸提供新的或增强的特征(例如，改进的稳定性)。核酸可以包括核酸亲和标签。核苷可以是碱基-糖组合。核苷的碱基部分可以是杂环碱基。此类杂环碱基的两个最常见的类别是嘌呤和嘧啶。核苷酸可以是还包括与核苷的糖部分共价连接的磷酸基团的核苷。对于包括呋喃戊糖的那些核苷，磷酸基团可以连接到糖的2’、3’或5’羟基部分。在形成核酸中，磷酸基团可以将相邻的核苷彼此共价连接以形成线性高分子化合物。转而此线性高分子化合物的各自端可以进一步接合而形成环状化合物；然而，线性化合物通常是合适的。此外，线性化合物可以具有内部核苷酸碱基互补性，并且因此可以按产生完全或部分双链化合物的方式折叠。在核酸中，该磷酸基团通常可以被称为形成核酸的核苷间骨架。连键或骨架可以是3’到5’磷酸二酯键。

[0102] 核酸可以包括修饰的骨架和/或修饰的核苷间键。修饰的骨架可以包括在骨架中保留磷原子和在骨架中不具有磷原子的那些。其中含有磷原子的合适修饰的核酸骨架可以含有例如硫代磷酸酯；手性硫代磷酸酯；二硫代磷酸酯；磷酸三酯；氨基烷基磷酸三酯；甲基膦酸酯和其他烷基膦酸酯，如3’-亚烷基膦酸酯、5’-亚烷基膦酸酯；手性膦酸酯；亚磷酸酯；

包括3'-氨基磷酰胺酯和氨基烷基磷酰胺酯的磷酰胺酯;磷二酰胺酯;硫代羰基磷酰胺酯;硫代羰基烷基膦酸酯;硫代羰基烷基磷酸三酯;硒代磷酸酯;以及具有正常3'-5'键的硼烷磷酸酯,2'-5'连接的类似物和具有反向极性的那些,其中一个或多个核苷酸间键是3'至3'、5'至5'或2'至2'键。

[0103] 核酸可以包括由短链烷基或环烷基核苷间键、混合杂原子、和烷基或环烷基核苷间键或者一个或多个短链杂原子的或杂环的核苷间键形成的多核苷酸骨架。这些可包括具有以下结构的那些:吗啉代键(从核苷的糖部分部分地形成);硅氧烷骨架;硫化物、亚砜和砜骨架;甲酰乙酰基和硫代甲酰乙酰基骨架;亚甲基甲酰乙酰基和硫代甲酰乙酰基骨架;核糖乙酰基骨架;含烯的骨架;氨基磺酸盐骨架;亚甲亚氨基和亚甲肼基骨架;磺酸酯和磺酰胺骨架;酰胺骨架;和具有混合的N、O、S和CH2组分部分的其他骨架。

[0104] 核酸可以包括核酸模拟物。术语"模拟物"可以旨在包括其中只有呋喃糖环或呋喃糖环和核苷酸间键两者被非呋喃糖基团替代的多核苷酸,仅替代呋喃糖环可以称为糖替代物。可以保持杂环碱基部分或修饰的杂环碱基部分以便与适当的靶核酸杂交。一种这样的核酸可以是肽核酸(PNA)。在PNA中,多核苷酸的糖骨架可以被含酰胺的骨架(特别是氨基乙基甘氨酸骨架)替代。核苷酸可以被保持并且直接或间接地结合至骨架的酰胺部分的氮杂氮原子上。PNA化合物中的骨架可以包括两个或更多个连接的氨基乙基甘氨酸单元,其给予PNA含酰胺的骨架。杂环碱基部分可以直接或间接地结合到骨架的酰胺部分的氮杂氮原子上。

[0105] 核酸可以包括吗啉代骨架结构。例如,核酸可以包括代替核糖环的6元吗啉代环。在这些实施例的一些中,磷二酰胺酯或其他非磷酸二酯核苷间键可替代磷酸二酯键。

[0106] 核酸可以包括具有附接到吗啉代环上的杂环碱基的连接的吗啉代单元(即吗啉代核酸)。连接基团可以连接吗啉代核酸中的吗啉代单体单元。非离子型基于吗啉代的寡聚化合物可以与细胞蛋白具有较少的不希望的相互作用。基于吗啉代的多核苷酸可以是核酸的非离子模拟物。吗啉代类别中的多种化合物可以使用不同的连接基团连接。另一类多核苷酸模拟物可称为环己烯基核酸(CeNA)。通常存在于核酸分子中的呋喃糖环可以被环己烯基环替代。可以制备CeNA DMT保护的亚磷酰胺单体,并用于使用亚磷酰胺化学的寡聚化合物合成。将CeNA单体并入核酸链可以增加DNA/RNA杂交体的稳定性。CeNA寡聚腺苷酸可以与具有和天然复合物相似的稳定性的核酸互补体形成复合物。另外的修饰可以包括锁核酸(LNA),其中2'-羟基基团连接到糖环的4'碳原子,从而形成2'-C,4'-C-氧亚甲基键,由此形成双环糖部分。该键可以是桥连2'氧原子和4'碳原子的基团亚甲基(-CH2-),其中n是1或2。LNA和LNA类似物可以表现出与互补性核酸非常高的双链体热稳定性(Tm=+3℃至+10℃)、对3′-外切核苷酸降解的稳定性以及良好的溶解度特性。

[0107] 核酸还可以包括核碱基(通常简称为"碱基")修饰或取代。如本文使用的,"未修饰的"或"天然的"核碱基可以包括嘌呤碱基(例如腺嘌呤(A)和鸟嘌呤(G))、以及嘧啶碱基(例如胸腺嘧啶(T)、胞嘧啶(C)和尿嘧啶(U))。经修饰的核碱基可以包括其他合成以及天然的核碱基,如5-甲基胞嘧啶(5-me-C),5-羟甲基胞嘧啶,黄嘌呤,次黄嘌呤,2-氨基腺嘌呤,腺嘌呤和鸟嘌呤的6-甲基以及其他烷基衍生物,腺嘌呤和鸟嘌呤的2-丙基以及其他烷基衍生物,2-硫尿嘧啶,2-硫胸腺嘧啶以及2-硫胞嘧啶,5-卤代尿嘧啶以及胞嘧啶,5-丙炔基(-C=C-CH3)尿嘧啶及胞嘧啶以及嘧啶碱基的其他炔基衍生物,6-偶氮基尿嘧啶,胞嘧啶以及胸

腺嘧啶,5-尿嘧啶(假尿嘧啶),4-硫尿嘧啶,8-卤基、8-氨基、8-巯基、8-硫烷基、8-羟基以及其他8-取代的腺嘌呤和鸟嘌呤,5-卤基特别是5-溴、5-三氟甲基以及其他5-取代的尿嘧啶和胞嘧啶、7-甲基鸟嘌呤和7-甲基腺嘌呤、2-F-腺嘌呤、2-氨基腺嘌呤、8-氮杂鸟嘌呤和8-氮杂腺嘌呤、7-脱氮杂鸟嘌呤和7-脱氮杂腺嘌呤、以及3-脱氮杂鸟嘌呤和3-脱氮杂腺嘌呤。经修饰的核碱基可以包括三环嘧啶如吩噁嗪胞苷(1H-嘧啶并(5,4-b)(1,4)苯并噁嗪-2(3H)-酮)、吩噻嗪胞苷(1H-嘧啶并(5,4-b)(1,4)苯并噻嗪-2(3H)-酮),G-夹(clamp)如取代的吩噁嗪胞苷(例如9-(2-氨基乙氧基)-H-嘧啶并(5,4-(b)(1,4)苯并噁嗪-2(3H)-酮)、吩噻嗪胞苷(1H-嘧啶并(5,4-b)(1,4)苯并噻嗪-2(3H)-酮),G-夹(clamp)如取代的吩噁嗪胞苷(例如9-(2-氨基乙氧基)-H-嘧啶并(5,4-(b)(1,4)苯并噁嗪-2(3H)-酮)、咔唑胞苷(2H-嘧啶并(4,5-b)吲哚-2-酮)、吡啶并吲哚胞苷(H吡啶并(3',2':4,5)吡咯并[2,3-d]嘧啶-2-酮)。

[0108] 如本文使用的,术语“样品”可以指包括靶的组合物。用于通过披露的方法、装置、和系统进行分析的合适样品包括细胞、组织、器官、或生物体。

[0109] 如本文使用的,术语“采样装置”或“装置”可以指可以取一部分样品和/或将所述部分放置在基底上的装置。采样装置可以指例如荧光激活细胞分选(FACS)机、细胞分选机、活检针、活检装置、组织切片装置、微流体装置、叶栅和/或超薄切片机。

[0110] 如本文使用的,术语“固体支持物”可以指可以附接多个随机条形码的离散固体或半固体表面。固体支持物可以包括任何类型的实心的、多孔的或空心的球体、球、承座、圆柱体或其他类似配置,其由塑料、陶瓷、金属或高分子材料(例如,水凝胶)构成,其上可以固定核酸(例如,共价地或非共价地)。固体支持物可以包括可以是球形的(例如,微球)或具有非球形或不规则形状的离散颗粒,所述形状是如立方形、长方形、锥形、圆柱形、圆锥形、椭圆形或圆盘形等。以阵列间隔开的多个固体支持物可以不包括基底。固体支持物可以与术语“珠”互换使用。

[0111] 固体支持物可以指“基底”。基底可以是一种固体支持物。基底可以指可以在其上进行本披露的方法的连续的固体或半固体表面。例如,基底可以指阵列、盒、芯片、装置和载玻片。

[0112] 如本文使用的,术语“空间标记”可以指其可以在空间中与位置相关联的标记。

[0113] 如本文使用的,术语“随机条形码”可以指包含标记的多核苷酸序列。随机条形码可以是可用于随机条形码化的多核苷酸序列。随机条形码可用于对样品中的靶定量。随机条形码可用于控制标记与靶相关联后可能发生的错误。例如,随机条形码可用于评估扩增或测序错误。与靶相关联的随机条形码可以称为随机条形码-靶或随机条形码-标签-靶。

[0114] 如本文所用,术语“基因特异性随机条形码”可以指包含标记和基因特异性的靶结合区的多核苷酸序列。随机条形码可以是可用于随机条形码化的多核苷酸序列。随机条形码可用于对样品中的靶定量。随机条形码可用于控制标记与靶相关联后可能发生的错误。例如,随机条形码可用于评估扩增或测序错误。与靶相关联的随机条形码可以称为随机条形码-靶或随机条形码-标签-靶。

[0115] 如本文使用的,术语“随机条形码化”可以指核酸的随机标记(例如,条形码化)。随机条形码化可以利用递归泊松策略来关联并对与靶相关联的标记进行定量。如本文所用的,术语“随机条形码化”可以与“基因特异性随机条形码化”互换地使用。

[0116] 如本文使用的，术语“靶”可以指可与随机条形码相关联的组合物。用于通过披露的方法、装置和系统进行分析的示例性合适的靶包括寡核苷酸、DNA、RNA、mRNA、微小RNA、tRNA等。靶可以是单链的或双链的。在一些实施例中，靶可以是蛋白质。在一些实施例中，靶是脂质。

[0117] 如本文所用的，术语“逆转录酶”可以指具有逆转录酶活性(即，催化从RNA模板合成DNA)的一组酶。通常，这样的酶包括但不限于逆转录病毒逆转录酶、逆转录转座子逆转录酶、逆转录质粒逆转录酶、逆转录子逆转录酶、细菌逆转录酶、II型内含子衍生的逆转录酶，及其突变体、变体或衍生物。非逆转录病毒逆转录酶包括非LTR逆转录转座子逆转录酶、逆转录质粒逆转录酶、逆转录子逆转录酶和II型内含子逆转录酶。II型内含子逆转录酶的实例包括乳酸乳球菌LI.LtrB内含子逆转录酶、细长嗜热聚球藻(Thermosynechococcus elongatus)TeI4c内含子逆转录酶或嗜热脂肪土芽孢杆菌GsI-IIC内含子逆转录酶。其他类别的逆转录酶可以包括许多类型的非逆转录病毒逆转录酶(即，逆转录子、II型内含子、以及多样性产生型逆转录元件等等)。

[0118] 本文中公开了鉴定用于区分细胞类型的靶的系统和方法。在一些实施例中，该方法包括：(a)接收包含表达谱的靶计数数据结构(例如靶计数矩阵)；(b)对多个细胞的表达谱进行层次聚类以产生代表该表达谱的树状图；(c)在从该树状图的根节点到该树状图的叶节点遍历该树状图的每个节点时：(1)确定该节点分裂成该节点的子节点是有效还是无效的(例如，子节点之间的差异不显著)；和(2)如果该节点分裂成该节点的子节点是无效的，则将该节点添加到合并聚类集中；(d)迭代地，对于该合并聚类集中的每个第一节点，如果该合并聚类集中的该第一节点与该合并聚类集中最接近该第一节点的第二节点之间的距离在合并距离阈值内，则将该第一节点与该第二节点合并以生成合并的节点，该合并的节点包含由该第一节点和该第二节点代表的表达谱；和(e)对于该合并聚类集中的每个节点，基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0119] 条形码

[0120] 条形码化(如，随机条形码化)已描述于例如US 20150299784、WO 2015031691、以及Fu等人，Proc Natl Acad Sci[美国国家图书馆院刊]U.S.A.2011年5月31日；108(22)：9026-31和Fan等人，Science[科学](2015)347(6222)：1258367中；这些出版物的内容通过引用以其整体结合在此。在一些实施例中，本文披露的条形码可以是随机条形码，该随机条形码可以是可用于对靶进行随机标记(例如，条形码，标签)的多核苷酸序列。如果随机条形码的不同的条形码序列的数目与待标记的任何靶的出现次数的比率可以是、或约1:1、2:1、3:1、4:1、5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、16:1、17:1、18:1、19:1、20:1、30:1、40:1、50:1、60:1、70:1、80:1、90:1、100:1、或在这些值的任何两个之间的数字或范围，则条形码可以称为随机条形码。靶可以是，例如包括具有相同或几乎相同序列的mRNA分子的mRNA种类。如果随机条形码的不同的条形码序列的数目与待标记的任何靶的出现次数的比率是至少、或至多1:1、2:1、3:1、4:1、5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、16:1、17:1、18:1、19:1、20:1、30:1、40:1、50:1、60:1、70:1、80:1、90:1、或100:1，则条形码可以称为随机条形码。随机条形码的条形码序列可以称为分子标记。

[0121] 条形码(例如随机条形码)可以包括一个或多个标记。示例性标记可包括通用标记、细胞标记、条形码序列(例如，分子标记)、样品标记、板标记、空间标记、和/或前空间

(pre-spatial)标记。图1说明了具有空间标记的示例性条形码104。条形码104可包括可将条形码与固体支持物105连接的5'胺。条形码可包括通用标记、维度标记、空间标记、细胞标记、和/或分子标记。条形码中不同标记(包括但不限于通用标记、维度标记、空间标记、细胞标记、和分子标记)的顺序可以改变。例如,如图1中显示,通用标记可以是5'-末端标记,且分子标记可以是3'-末端标记。空间标记、维度标记、和细胞标记能以任何顺序。在一些实施例中,通用标记、空间标记、维度标记、细胞标记、和分子标记是以任何顺序的。条形码可以包括靶结合区。靶结合区可以与样品中的靶(例如,靶核酸、RNA、mRNA、DNA)相互作用。例如,靶结合区可以包括可以与mRNA的聚(A)尾相互作用的寡聚(dT)序列。在一些情况下,条形码的标记(例如,通用标记、维度标记、空间标记、细胞标记和条形码序列)可以由1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、或20个或更多个核苷酸分开。

[0122] 标记(例如细胞标记)可包括一组独特的定义长度的核酸子序列,例如每个七个核苷酸(相当于一些汉明错误校正代码中使用的比特数目),其可以设计为提供错误校正能力。可以设计包括七个核苷酸序列的错误校正子序列组,使得所述组中的序列的任何成对组合展现出定义的"遗传距离"(或错配碱基数),例如一组纠错子序列能被设计为展现三个核苷酸的遗传距离。在这种情况下,对于经标记的靶核酸分子的序列数据组中的错误校正序列的审查(在下面更全面地描述)能允许检测或校正扩增或测序误差。在一些实施例中,用于产生错误校正代码的核酸子序列的长度可以变化,例如,它们可以是、或是约1、2、3、4、5、6、7、8、9、10、15、20、30、31、40、50个、或在这些值的任何两个之间的数字或范围的核苷酸长度。在一些实施例中,其他长度的核酸子序列可以用来产生错误校正代码。

[0123] 条形码可以包括靶结合区。靶结合区可以与样品中的靶相互作用。该靶可以是、或包括核糖核酸(RNA)、信使RNA(mRNA)、微小RNA、小干扰RNA(siRNA)、RNA降解产物、各自含有聚(A)尾的RNA、或其任何组合。在一些实施例中,多个靶可包括脱氧核糖核酸(DNA)。

[0124] 在一些实施例中,靶结合区可以包括可以与mRNA的聚(A)尾相互作用的寡聚(dT)序列。条形码的一个或多个标记(例如,通用标记、维度标记、空间标记、细胞标记、和条形码序列(例如,分子标记))可以通过间隔物与条形码的剩余标记的另一个或两个分开。间隔物可以是例如,1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、或20个或更多个核苷酸。在一些实施例中,条形码的标记中没有标记被间隔物分开。

[0125] 通用标记

[0126] 条形码可以包括一个或多个通用标记。在一些实施例中,对于条形码组中的所有条形码(附接到给定的固体支持物上的),一个或多个通用标记可以是相同的。在一些实施例中,对于附接到多个珠上的所有条形码,一个或多个通用标记可以是相同的。在一些实施例中,通用标记可以包括能够与测序引物杂交的核酸序列。测序引物可以用于对包括通用标记的条形码进行测序。测序引物(例如,通用测序引物)可以包括与高通量测序平台相联系的测序引物。在一些实施例中,通用标记可以包括能够与PCR引物杂交的核酸序列。在一些实施例中,通用标记可以包括能够与测序引物和PCR引物杂交的核酸序列。能够与测序或PCR引物杂交的通用标记的核酸序列可以被称为引物结合位点。通用标记可以包括可用于引发条形码转录的序列。通用标记可以包括可用于延伸条形码或条形码内的区域的序列。通用标记的长度可以是或是约1、2、3、4、5、10、15、20、25、30、35、40、45、50个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。例如,通用标记可包括至少约10个核苷酸。通

用标记的长度可以是至少、或至多1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、或300个核苷酸。在一些实施例中，可切割接头或修饰的核苷酸可以是通用标记序列的一部分，以使条形码能够从支持物上被切割下来。

[0127] 维度标记

[0128] 条形码可以包括一个或多个维度标记。在一些实施例中，维度标记可以包括提供关于标记(例如，随机标记)发生的维度的信息的核酸序列。例如，维度标记可以提供关于对靶进行随机条形码化的时间的信息。维度标记可以与样品中条形码化(例如，随机条形码化)的时间相关联。维度标记可以在标记的时间处被激活。不同的维度标记可以在不同的时间被激活。该维度标记提供关于靶、靶组和/或样品被随机条形码化的顺序的信息。例如，在细胞周期的G0期可以对细胞群进行随机条形码化。在细胞周期的G1期，可以用条形码(例如，随机条形码)对这些细胞再次进行脉冲处理。在细胞周期的S期，可以用条形码对所述细胞再次进行脉冲处理，等等。每个脉冲(例如，细胞周期的每个阶段)处的条形码可以包括不同的维度标记。以这种方式，该维度标记提供关于哪些靶在细胞周期的哪个时期被标记的信息。维度标记可以探询许多不同的生物阶段。示例性的生物学时间可以包括但不限于细胞周期、转录(例如，转录起始)和转录物降解。在另一个实例中，样品(例如，细胞、细胞群)可以在用药物和/或疗法治疗之前和/或之后随机标记。不同靶的拷贝数的变化可以指示样品对药物和/或疗法的反应。

[0129] 维度标记可以是可激活的。可以在特定时间点激活可激活的维度标记。可激活的标记可以被例如组成性地激活(例如，不关闭)。该可激活的维度标记可以被例如可逆地激活(例如，该可激活的维度标记可以打开和关闭)。该维度标记可以被例如可逆地激活至少1、2、3、4、5、6、7、8、9或10次或更多次。该维度标记可以被可逆地激活例如至少1、2、3、4、5、6、7、8、9或10次或更多次。在一些实施例中，可以用荧光；光；化学事件(例如，切割，另一种分子的连接，修饰的添加(例如，聚乙二醇化、sumo化、乙酰化、甲基化、去乙酰化、去甲基化)；光化学事件(例如，光锁定)；以及引入非天然的核苷酸将该维度标记激活。

[0130] 在一些实施例中，该维度标记对于附接到给定的固体支持物(例如，珠)上的所有条形码(例如，随机条形码)可以是相同的，但对于不同的固体支持物(例如，珠)是不同的。在一些实施例中，相同固体支持物上的至少60％、70％、80％、85％、90％、95％、97％、99％或100％的条形码可以包括相同的维度标记。在一些实施例中，相同固体支持物上的至少60％的条形码可以包括相同的维度标记。在一些实施例中，相同固体支持物上的至少95％的条形码可以包括相同的维度标记。

[0131] 多个固体支持物(例如，珠)可以表现多达10^6个或更多个独特维度标记序列。维度标记的长度可以是或是约1、2、3、4、5、10、15、20、25、30、35、40、45、50个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。维度标记的长度可以是至少、或至多1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、或300个核苷酸。维度标记可包括在约5至约200个之间的核苷酸。维度标记可包括在约10至约150个之间的核苷酸。维度标记可包括长度在约20至约125个之间的核苷酸。

[0132] 空间标记

[0133] 条形码可以包括一个或多个空间标记。在一些实施例中，空间标记可以包括提供与条形码相关联的靶分子的空间取向的信息的核酸序列。空间标记可以与样品中的坐标相

关联。该坐标可以是固定的坐标。例如可以参考基底固定坐标。空间标记可以参考二维或三维网格。可以参考界标固定坐标。在空间中界标是可被鉴定的。界标可以是可被成像的结构。界标可以是生物学结构，例如解剖学界标。界标可以是细胞界标，例如细胞器。界标可以是非天然界标，如具有可鉴定标识(如色码、条形码、磁性、荧光、放射性或独特尺寸或形状)的结构。空间标记可以与物理分区(例如，孔、容器或液滴)相关联。在一些实施例中，将多个空间标记一起用于编码在空间中的一个或多个位置。

[0134] 所述空间标记对于附接到给定的固体支持物(例如，珠)上的所有条形码可以是相同的，但对于不同的固体支持物(例如，珠)是不同的。在一些实施例中，包括相同空间标记的相同固体支持物上的条形码的百分比可以是、或是约60％、70％、80％、85％、90％、95％、97％、99％、100％、或在这些值的任何两个之间的数字或范围。在一些实施例中，包括相同空间标记的相同固体支持物上的条形码的百分比可以是至少、或至多60％、70％、80％、85％、90％、95％、97％、99％、或100％。在一些实施例中，相同固体支持物上的至少60％的条形码可以包括相同的空间标记。在一些实施例中，相同固体支持物上的至少95％的条形码可以包括相同的空间标记。

[0135] 多个固体支持物(例如，珠)可以表现多达10^6个或更多个独特空间标记序列。空间标记的长度可以是或是约1、2、3、4、5、10、15、20、25、30、35、40、45、50个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。空间标记的长度可以是至少、或至多1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、或300个核苷酸。空间标记可包括在约5至约200个之间的核苷酸。空间标记可包括在约10至约150个之间的核苷酸。空间标记可包括长度在约20至约125个之间的核苷酸。

[0136] 细胞标记

[0137] 条形码可以包括一个或多个细胞标记。在一些实施例中，细胞标记可以包括提供用于确定哪个靶核酸来自哪个细胞的信息的核酸序列。在一些实施例中，该细胞标记对于附接到给定的固体支持物(例如，珠)上的所有条形码是相同的，但对于不同的固体支持物(例如，珠)是不同的。在一些实施例中，包括相同细胞标记的相同固体支持物上的条形码的百分比可以是、或是约60％、70％、80％、85％、90％、95％、97％、99％、100％、或在这些值的任何两个之间的数字或范围。在一些实施例中，包括相同细胞标记的相同固体支持物上的条形码的百分比可以是、或是约60％、70％、80％、85％、90％、95％、97％、99％、或100％。例如，相同固体支持物上的至少60％的条形码可以包括相同的细胞标记。作为另一个实例，相同固体支持物上的至少95％的条形码可以包括相同的细胞标记。

[0138] 多个固体支持物(例如，珠)可以表现多达10^6个或更多个独特细胞标记序列。细胞标记的长度可以是或是约1、2、3、4、5、10、15、20、25、30、35、40、45、50个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。细胞标记的长度可以是至少、或至多1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、或300个核苷酸。例如，细胞标记可以包括约5至约200个之间的核苷酸。作为另一个实例，细胞标记可以包括约10至约150个之间的核苷酸。还作为另一个实例，细胞标记可包括长度在约20至约125个之间的核苷酸。

[0139] 条形码序列

[0140] 条形码可以包括一个或多个条形码序列。在一些实施例中，条形码序列可以包括为与条形码杂交的特定类型的靶核酸种类提供鉴定信息的核酸序列。条形码序列可以包括

如下核酸序列，该核酸序列为与条形码(例如，靶结合区)杂交的靶核酸种类的特定出现提供计数器(例如，提供粗略近似)。

[0141] 在一些实施例中，将一组不同的条形码序列附接到给定的固体支持物(例如，珠)上。在一些实施例中，可以有、或约有10^2、10^3、10^4、10^5、10^6、10^7、10^8、10^9个、或在这些值的任何两个之间的数字或范围的独特分子标记序列。例如，多个条形码可以包括具有不同序列的约6561个条形码序列。作为另一个实例，多个条形码可以包括具有不同序列的约65536个条形码序列。在一些实施例中，可以有至少、或至多10^2、10^3、10^4、10^5、10^6、10^7、10^8、或10^9个独特条形码序列。独特分子标记序列可以附接至给定的固体支持物(例如，珠)上。

[0142] 条形码的长度可以是、或是约1、2、3、4、5、10、15、20、25、30、35、40、45、50个、或在这些值的任何两个之间的数字或范围的核苷酸。条形码的长度可以是至少、或至多1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、或300个核苷酸。

[0143] 分子标记

[0144] 条形码可以包括一个或多个分子标记。分子标记可以包括条形码序列。在一些实施例中，分子标记可以包括为与随机条形码杂交的特定类型的靶核酸种类提供鉴定信息的核酸序列。分子标记可以包括如下核酸序列，该核酸序列为与随机条形码(例如，靶结合区)杂交的靶核酸种类的特定出现提供计数器。

[0145] 在一些实施例中，将一组不同的分子标记附接到给定的固体支持物(例如，珠)上。在一些实施例中，可以有、或约有10^2、10^3、10^4、10^5、10^6、10^7、10^8、10^9个、或许多或一定范围的独特分子标记序列。例如，多个随机条形码可包括具有不同序列的约6561个分子标记。作为另一个实例，多个随机条形码可包括具有不同序列的约65536个分子标记。在一些实施例中，可以有至少、或至多10^2、10^3、10^4、10^5、10^6、10^7、10^8、或10^9个独特分子标记序列。具有独特分子标记序列的随机条形码可以附接至给定固体支持物(例如，珠)上。

[0146] 对于使用多个随机条形码的随机条形码化，不同分子标记序列的数目与任何靶的出现次数的比率可以是、或约1:1、2:1、3:1、4:1、5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、16:1、17:1、18:1、19:1、20:1、30:1、40:1、50:1、60:1、70:1、80:1、90:1、100:1、或在这些值的任何两个之间的数字或范围。靶可以是包括具有相同或几乎相同序列的mRNA分子的mRNA种类。在一些实施例中，不同分子标记序列的数目与任何靶的出现次数的比率是至少、或至多1:1、2:1、3:1、4:1、5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、16:1、17:1、18:1、19:1、20:1、30:1、40:1、50:1、60:1、70:1、80:1、90:1、或100:1。

[0147] 分子标记的长度可以是或是约1、2、3、4、5、10、15、20、25、30、35、40、45、50个、或在这些值的任何两个之间的数字或范围的核苷酸。分子标记的长度可以是至少、或至多1、2、3、4、5、10、15、20、25、30、35、40、45、50、100、200、或300个核苷酸。

[0148] 靶结合区

[0149] 条形码可以包括一个或多个靶结合区，如捕获探针。在一些实施例中，靶结合区可以与目的靶杂交。在一些实施例中，该靶结合区可包括与靶(例如，靶核酸、靶分子，例如待分析的细胞核酸)特异性杂交(例如与特定基因序列杂交)的核酸序列。在一些实施例中，靶结合区可以包括可附接(例如，杂交)至特定靶核酸的特定位置的核酸序列。在一些实施例中，靶结合区可以包括能够与限制性酶位点突出端(例如EcoRI粘性末端突出端)进行特异

性杂交的核酸序列。然后条形码可以连接到包括与限制性位点突出端互补的序列的任何核酸分子。

[0150] 在一些实施例中，靶结合区可以包括非特异性靶核酸序列。非特异性靶核酸序列可以指独立于靶核酸的特定序列可与多个靶核酸结合的序列。例如，靶结合区可以包括与mRNA分子上的聚(A)尾杂交的随机多聚体序列或寡聚(dT)序列。随机多聚体序列可以是例如随机二聚体、三聚体、四聚体、五聚体、六聚体、七聚体、八聚体、九聚体、十聚体或任何长度的更高多聚体序列。在一些实施例中，对于附接至给定珠的所有条形码，所述靶结合区是相同的。在一些实施例中，对于附接到给定珠上的多个条形码，靶结合区可以包括两个或更多个不同的靶结合序列。靶结合区的长度可以是或是约5、10、15、20、25、30、35、40、45、50个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。靶结合区的长度可以是至多约5、10、15、20、25、30、35、40、45、50个或更多个核苷酸。

[0151] 在一些实施例中，靶结合区可以包括寡聚(dT)，该寡聚(dT)可以与包括聚腺苷酸化端的mRNA杂交。靶结合区可以是基因特异性的。例如，可以将靶结合区配置为与靶的特定区域杂交。靶结合区的长度可以是或是约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。靶结合区的长度可以是至少、或至多1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、或30个核苷酸。靶结合区的长度可以是约5-30个核苷酸。当条形码包括基因特异性靶结合区时，该条形码可以称为基因特异性条形码。

[0152] 定向特性

[0153] 条形码可以包括一种或多种可用于定向(例如，比对)条形码的定向特性。条形码可以包括用于等电聚焦的部分。不同的条形码可以包括不同的等电聚焦点。当将这些条形码被引入样品中时，该样品可以经历等电聚焦，以便于将所述条形码定位成已知的方式。以这种方式，该定向特性可以用于开发样品中条形码的已知的映射。示例性定向特性可以包括电泳迁移率(例如，基于条形码的尺寸)、等电点、自旋、电导率和/或自组装。例如，条形码具有自组装的定向特性，当激活时可以自组装成特定定向(例如，核酸纳米结构)。

[0154] 亲和特性

[0155] 条形码可以包括一种或多种亲和特性。例如，空间标记可以包括亲和特性。亲和力特性可包括在化学和/或生物部分中，该特性可以促进该条形码与另一种实体(例如，细胞受体)的结合。例如，亲和特性可包括抗体，例如，对于样品上的具体部分(例如，受体)特异性的抗体。在一些实施例中，抗体可以将条形码引导到特定细胞类型或分子上。在特定细胞类型或分子处的和/或附近的靶可以被随机标记。在一些实施例中，除了空间标记的核苷酸序列，亲和力特性可以提供空间信息，因为该抗体可以将该条形码引导至特定位置。抗体可以是治疗性抗体，例如单克隆抗体或多克隆抗体。抗体可以是人源化的或嵌合的。抗体可以是裸抗体或融合抗体。

[0156] 抗体可以是全长(即，天然存在的或通过正常免疫球蛋白基因片段重组过程形成的)免疫球蛋白分子(例如，IgG抗体)或免疫球蛋白分子的免疫活性(即，特异性结合)部分(像抗体片段)。

[0157] 抗体片段可以是例如抗体的一部分，如F(ab')2、Fab'、Fab、Fv、sFv等。在一些实施

例中，抗体片段可以与由全长抗体识别的相同的抗原结合。抗体片段可以包括由抗体的可变区组成的分离的片段，如由重链和轻链的可变区组成的“Fv”片段和其中轻链和重链可变区通过肽接头连接的重组单链多肽分子(“scFv蛋白”)。示例性抗体可以包括但不限于癌细胞抗体、病毒抗体、结合至细胞表面受体(CD8、CD34、CD45)的抗体、和治疗性抗体。

[0158] 通用衔接子引物

[0159] 条形码可包括一个或多个通用衔接子引物。例如，基因特异性条形码(如，基因特异性随机条形码)可以包括通用衔接子引物。通用衔接子引物可以指在所有条形码上通用的核苷酸序列。通用衔接子引物可以用于构建基因特异性条形码。通用衔接子引物的长度可以是、或是约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30个核苷酸、或在这些值的任何两个之间的数字或范围的核苷酸。通用衔接子引物的长度可以是至少、或至多1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、或30个核苷酸。通用衔接子引物的长度可以是从5-30个核苷酸。

[0160] 接头

[0161] 当条形码包括多于一个类型的标记(例如，多于一个细胞标记或多于一个条形码序列，如一个分子标记)时，这些标记可以穿插着接头标记序列。接头标记序列的长度可以是至少约5、10、15、20、25、30、35、40、45、50个或更多个核苷酸。接头标记序列的长度可以是至多约5、10、15、20、25、30、35、40、45、50个或更多个核苷酸。在一些情况下，接头标记序列的长度是12个核苷酸。可以将接头标记序列用于促进条形码的合成。该接头标记可以包括错误校正(例如，汉明)代码。

[0162] 固体支持物

[0163] 在一些实施例中，本文披露的条形码(如随机条形码)可以与固体支持物关联。例如，固体支持物可以是合成颗粒。在一些实施例中，固体支持物上的多个条形码(例如，第一多个条形码)的一些或所有条形码序列(如，随机条形码(例如，第一条形码序列)的分子标记)具有至少一个核苷酸的差异。相同固体支持物上的条形码的细胞标记可以是相同的。不同的固体支持物上的条形码的细胞标记可以具有至少一个核苷酸的差异。例如，第一固体支持物上的第一多个条形码的第一细胞标记可以具有相同的序列，且第二固体支持物上的第二多个条形码的第二细胞标记可以具有相同的序列。第一固体支持物上的第一多个条形码的第一细胞标记和第二固体支持物上的第二多个条形码的第二细胞标记可以具有至少一个核苷酸的差异。细胞标记例如可以约5-20个核苷酸长。条形码序列例如可以约5-20个核苷酸长。合成颗粒例如可以是珠。

[0164] 珠可以例如是硅胶珠、可控孔径玻璃珠、磁珠、Dynabead、交联葡聚糖/琼脂糖珠、珠状纤维素、聚苯乙烯珠、或其任何组合。珠可包括如聚二甲基硅氧烷(PDMS)、聚苯乙烯、玻璃、聚丙烯、琼脂糖、明胶、水凝胶、顺磁物质、陶瓷、塑料、玻璃、甲基苯乙烯、丙烯酸聚合物、钛、乳胶、琼脂糖凝胶、纤维素、尼龙、硅酮、或其任何组合的材料。

[0165] 在一些实施例中，珠可以是高分子微球(例如可变形的珠或凝胶珠)，其用条形码或随机条形码功能化(如来自10X基因组公司(10X Genomics)(旧金山，加利福尼亚)的凝胶珠)。在一些实现方式中，凝胶珠可包括基于聚合物的凝胶。例如，可以通过将一种或多种聚合物前体包封进液滴来产生凝胶珠。在将聚合物前体暴露于促进剂(例如，四甲基乙二胺

(TEMED))后,可以产生凝胶珠。

[0166] 在一些实施例中,颗粒可以是可降解的。例如,高分子微球可以例如在所希望的条件下溶解、熔化或降解。所希望的条件可包括环境条件。所希望的条件可导致高分子微球以受控方式溶解、熔化或降解。由于化学刺激、物理刺激、生物刺激、热刺激、磁刺激、电刺激、光刺激或其任何组合,凝胶珠可以溶解、融化或降解。

[0167] 分析物和/或试剂(如寡核苷酸条形码)例如可以偶联/固定到凝胶珠的内表面(寡核苷酸条形码和/或用于产生寡核苷酸条形码的材料的扩散的可接近内部)和/或凝胶珠的外表面或本文描述的任何其他微胶囊。偶联/固定可以经由任何形式的化学键(例如,共价键、离子键)或物理现象(例如,范德华力、偶极-偶极相互作用等)。在一些实施例中,试剂与凝胶珠或本文描述的任何其他微胶囊的偶联/固定可以是可逆的,例如经由不稳定部分(例如,经由化学交联剂,包括本文描述的化学交联剂)。在施加刺激后,不稳定部分可以被切割并释放固定化的试剂。在一些实施例中,不稳定部分是二硫键。例如,在经由二硫键将寡核苷酸条形码固定到凝胶珠上的情况下,将二硫键暴露于还原剂可以切割二硫键并从珠释放寡核苷酸条形码。不稳定部分可以作为凝胶珠或微胶囊的一部分、作为将试剂或分析物与凝胶珠或微胶囊连接的化学接头的一部分、和/或作为试剂或分析物的一部分包括在内。在一些实施例中,多个条形码的至少一个条形码可固定在颗粒上、部分固定在颗粒上、包封在颗粒中、部分包封在颗粒中、或其任何组合。

[0168] 在一些实施例中,凝胶珠可包括广泛多种不同的聚合物,包括但不限于:聚合物、热敏聚合物、光敏聚合物、磁性聚合物、pH敏感聚合物、盐敏感聚合物、化学敏感聚合物、聚电解质、多糖、肽、蛋白质和/或塑料。聚合物可包括但不限于以下材料:如聚(N-异丙基丙烯酰胺)(PNIPAAm)、聚(苯乙烯磺酸酯)(PSS)、聚(烯丙基胺)(PAAm)、聚(丙烯酸)(PAA)、聚(乙烯亚胺)(PEI)、聚(双烯丙基二甲基-氯化铵)(PDADMAC)、聚(吡咯)(PPy)、聚(乙烯基吡咯烷酮)(PVPON)、聚(乙烯基吡啶)(PVP)、聚(甲基丙烯酸甲酯)(PMAA)、聚(甲基丙烯酸甲酯)(PMMA)、聚苯乙烯(PS)、聚(四氢呋喃)(PTHF)、聚(邻苯二甲醛)(PTHF)、聚(己基紫精)(PHV)、聚(L-赖氨酸)(PLL)、聚(L-精氨酸)(PARG)、聚(乳酸-聚羟基乙酸)(PLGA)。

[0169] 许多化学刺激可用于触发珠的破坏、溶解、或降解。这些化学改变的实例可包括但不限于pH介导的珠壁改变、经由交联键的化学裂解使珠壁分解、珠壁的触发解聚、和珠壁转换反应。批量改变也可用于触发珠的破坏。

[0170] 通过各种刺激对微胶囊的批量或物理变化在设计胶囊以释放试剂方面也提供了许多优点。在宏观尺度上发生批量或物理变化,其中珠破裂是由刺激引起的机械-物理力的结果。这些过程可包括但不限于压力引起的破裂、珠壁熔化、或珠壁的孔隙率的改变。

[0171] 生物刺激也可用于触发珠的破坏、溶解、或降解。通常,生物触发剂类似于化学触发剂,但是许多实例使用生物分子、或生命系统中常见的分子,如酶、肽、糖、脂肪酸、核酸等。例如,珠可包括具有肽交联的聚合物,该肽交联通过特定蛋白酶对切割敏感。更具体地,一个实例可包括含有GFLGK肽交联的微胶囊。在添加生物触发物(如蛋白酶组织蛋白酶B)后,壳孔的肽交联被切割且珠的内容物被释放。在其他情况下,蛋白酶可以是热激活的。在另一个实例中,珠包括含有纤维素的壳壁。水解酶壳聚糖的添加用作纤维素键裂解、壳壁解聚、和内部内容物释放的生物触发剂。

[0172] 还可以在施加热刺激后诱导珠释放其内容物。温度的变化可导致珠的各种变化。

热量的变化可能导致珠熔化，使得珠壁崩解。在其他情况下，热量可能增加珠内部组分的内部压力，使得珠破裂或爆炸。在仍其他情况下，热量可以使珠变成收缩的脱水状态。热量还可以作用于珠壁内的热敏聚合物，从而引起珠的破坏。

[0173] 将磁性纳米颗粒包括在微胶囊的珠壁中可以允许珠的触发破裂以及将珠引导成阵列。本披露的装置可包括用于任一感兴趣的磁珠。在一个实例中，将Fe_3O_4纳米颗粒并入含聚电解质的珠中在振荡磁场刺激的存在下触发破裂。

[0174] 由于电刺激的结果，珠也可能被破坏、溶解、或降解。与先前部分中描述的磁性颗粒相似，电敏珠可以允许珠的触发破裂以及其他功能，如电场中的对准、电导率或氧化还原反应。在一个实例中，含有电敏材料的珠在电场中排列，从而可以控制内部试剂的释放。在其他实例中，电场可以在珠壁本身内引起氧化还原反应，这可以增加孔隙率。

[0175] 也可用光刺激来破坏珠。许多光触发是可能的，并可以包括使用各种分子(如能够吸收特定波长范围的光子的纳米颗粒和发色团)的系统。例如，金属氧化物涂层可用作胶囊触发剂。涂覆有SiO_2的聚电解质胶囊的UV照射可导致珠壁的崩解。在又另一个实例中，可以将可光切换材料(如偶氮苯基团)并入珠壁中。在施加UV或可见光后，如这些的化学物质在吸收光子后经历可逆的顺式-反式异构化。在此方面，光子切换的并入导致珠壁在施加光触发剂后可崩解或变得更多孔。

[0176] 例如，在图2中说明的条形码化(随机条形码化)的非限制性实例中，在框208处将细胞(如单个细胞)引入微孔阵列的多个微孔之后，在框212处可以将珠引入微孔阵列的多个微孔上。每个微孔可包括一个珠。珠可包括多个条形码。条形码可包括附接至珠的5’胺区域。条形码可以包括通用标记、条形码序列(例如，分子标记)、靶结合区、或其任何组合。

[0177] 本文披露的条形码可以与固体支持物(例如，珠)关联(例如，附接)。与固体支持物关联的条形码每个可包括选自下组的条形码序列，该组由以下组成：具有独特序列的至少100或1000个条形码序列。在一些实施例中，与固体支持物关联的不同条形码可包括不同序列的条形码序列。在一些实施例中，与固体支持物关联的条形码的百分比包括相同的细胞标记。例如，该百分比可以是、或是约60％、70％、80％、85％、90％、95％、97％、99％、100％、或在这些值的任何两个之间的数字或范围。作为另一个实例，该百分比可以是至少、或至多60％、70％、80％、85％、90％、95％、97％、99％、或100％。在一些实施例中，与固体支持物关联的条形码可以具有相同的细胞标记。与不同固体支持物关联的条形码可以具有选自下组的不同的细胞标记，该组由以下组成：具有独特序列的至少100或1000个细胞标记。

[0178] 本文披露的条形码可以与固体支持物(例如，珠)关联(例如，附接)。在一些实施例中，可以用包括与多个条形码关联的多个合成的颗粒的固体支持物对样品中的多个靶进行随机地条形码化。在一些实施例中，固体支持物可包括与多个条形码关联的多个合成的颗粒。不同固体支持物上的多个条形码的空间标记可以具有至少一个核苷酸的差异。固体支持物例如可以包括处于二维或三维的多个条形码。合成的颗粒可以是珠。珠可以是硅胶珠、可控孔径玻璃珠、磁珠、Dynabead、交联葡聚糖/琼脂糖珠、珠状纤维素、聚苯乙烯珠、或其任何组合。固体支持物可包括聚合物、基质、水凝胶、针阵列装置、抗体、或其任何组合。在一些实施例中，固体支持物可以自由浮动。在一些实施例中，固体支持物可嵌入半固体或固体阵列中。条形码可以不与固体支持物关联。条形码可以是单独的核苷酸。条形码可与基底相关联。

[0179] 如本文使用的，术语“拴系”、“附接”和“固定”可互换使用，并且可以指用于将条形码附接到固体支持物上的共价或非共价方式。可以将多种不同的固体支持物中的任何一种用作固体支持物，以用于附接预先合成的条形码或用于条形码的原位固相合成。

[0180] 在一些实施例中，固体支持物是珠。珠可以包括一种或多种类型的实心的、多孔的或空心的球体、球、承座、圆柱体或其他相似配置，其上可以固定核酸(例如，共价地或非共价地)。珠可以例如由塑料、陶瓷、金属、聚合物材料、或其任何组合构成。珠可以是、或包括球形的(例如，微球)或具有非球形或不规则形状的离散颗粒，该形状是如立方形、长方形、锥形、圆柱形、圆锥形、椭圆形或圆盘形等。在一些实施例中，珠的形状可以是非球形的。

[0181] 珠可以包含多种材料，包括但不限于顺磁性材料(例如镁、钼、锂和钽)、超顺磁性材料(例如铁氧体(Fe_3O_4；磁铁矿)纳米颗粒)、铁磁材料(例如，铁、镍、钴，其一些合金，以及一些稀土金属化合物)、陶瓷、塑料、玻璃、聚苯乙烯、二氧化硅、甲基苯乙烯、丙烯酸聚合物、钛、胶乳、交联琼脂糖、琼脂糖、水凝胶、聚合物、纤维素、尼龙、或其任何组合。

[0182] 在一些实施例中，珠(例如，标记所附接的珠)是水凝胶珠。在一些实施例中，珠包括水凝胶。

[0183] 本文披露的一些实施例包括一个或多个颗粒(例如珠)。颗粒的每个可包括多个寡核苷酸(例如，条形码)。多个寡核苷酸的每个可包括条形码序列(例如，分子标记)、细胞标记、和靶结合区(例如，寡聚(dT)序列、基因特异性序列、随机多聚体、或其组合)。该多个寡核苷酸的每个的细胞标记序列可以是相同的。不同颗粒上的寡核苷酸的细胞标记序列可以是不同的，使得可以鉴定不同颗粒上的寡核苷酸。在不同实现方式中，不同细胞标记序列的数目可以是不同的。在一些实施例中，细胞标记序列的数目可以是、或是约10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、10^6、10^7、10^8、10^9、在这些值的任何两个之间的数字或范围、或更多。在一些实施例中，细胞标记序列的数目可以是至少、或至多10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、10^6、10^7、10^8、或10^9。在一些实施例中，多个颗粒中不超过1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、或更多个包括具有相同细胞序列的寡核苷酸。在一些实施例中，包括具有相同细胞序列的寡核苷酸的多个颗粒可以是至多0.1％、0.2％、0.3％、0.4％、0.5％、0.6％、0.7％、0.8％、0.9％、1％、2％、3％、4％、5％、6％、7％、8％、9％、10％或更多。在一些实施例中，多个颗粒中没有颗粒具有相同的细胞标记序列。

[0184] 在每个颗粒上的多个寡核苷酸可以包括不同的条形码序列(例如，分子标记)。在一些实施例中，条形码序列的数目可以是、或约10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、10^6、10^7、10^8、10^9、或在这些值的任何两个之间的数字或范围。在一些实施例中，条形码序列的数目可以是至少、或至多10、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000、10^6、10^7、10^8、或10^9。例如，多个寡核苷酸中的至少100个包括不同的条形码序列。作为另一个实例，在单个颗粒中，多

个寡核苷酸中的至少100、500、1000、5000、10000、15000、20000、50000个、这些值的任何两个之间的数字或范围、或更多个包括不同的条形码序列。一些实施例提供了包括条形码的多个颗粒。在一些实施例中，待标记的靶和不同条形码序列的出现(或拷贝或数目)的比率可以是至少1:1、1:2、1:3、1:4、1:5、1:6、1:7、1:8、1:9、1:10、1:11、1:12、1:13、1:14、1:15、1:16、1:17、1:18、1:19、1:20、1:30、1:40、1:50、1:60、1:70、1:80、1:90、或更高。在一些实施例中，多个寡核苷酸的每个进一步包括样品标记、通用标记、或两者。颗粒例如可以是纳米颗粒或微颗粒。

[0185] 珠的尺寸可以改变。例如，珠的直径范围可以从0.1微米至50微米。在一些实施例中，珠的直径可以是或是约0.1、0.5、1、2、3、4、5、6、7、8、9、10、20、30、40、50微米、或在这些值的任何两个之间或数字或范围。

[0186] 珠的直径可以与基底的孔的直径相关。在一些实施例中，相比孔的直径，珠的直径可以更长或更短、或约10％、20％、30％、40％、50％、60％、70％、80％、90％、100％、或在这些值的任何两个之间的数字或范围。珠的直径可以与细胞(例如，由基底的孔截留的单个细胞)的直径有关。在一些实施例中，相比孔的直径，珠的直径可以是至少、或至多10％、20％、30％、40％、50％、60％、70％、80％、90％、或100％更长或更短。珠的直径可以与细胞(例如，由基底的孔截留的单个细胞)的直径有关。在一些实施例中，相比细胞的直径，珠的直径可以更长或更短、或约10％、20％、30％、40％、50％、60％、70％、80％、90％、100％、150％、200％、250％、300％、或在这些值的任何两个之间的数字或范围。在一些实施例中，相比孔的直径，珠的直径可以是至少、或至多10％、20％、30％、40％、50％、60％、70％、80％、90％、100％、150％、200％、250％、或300％更长或更短。

[0187] 珠可以附接到和/或包埋在基底中。可以将珠附接和/或嵌入凝胶、水凝胶、聚合物，和/或基质中。使用存在于珠中可以充当位置地址的条形码上的空间标记，可以鉴定在基质(例如，凝胶、基质、支架，或聚合物)中珠的空间位置。

[0188] 珠的实例可以包括但不限于链霉亲和素珠、琼脂糖珠、磁珠、Dynabeads®、MACS®微珠、缀合抗体的珠(例如，抗免疫球蛋白微珠)、缀合A蛋白的珠、缀合G蛋白的珠、缀合A/G蛋白的珠、缀合L蛋白的珠、缀合寡聚(dT)的珠、二氧化硅珠、二氧化硅样珠、抗生物素微珠、抗荧光染料微珠以及BcMag™羧基封端的磁珠。

[0189] 珠可以关联有(例如浸渍有)量子点或荧光染料，以使其在一个荧光光通道或多个光通道中是荧光的。珠可以关联有氧化铁或氧化铬，以使其具有顺磁性或铁磁性。珠是可被鉴定的。例如，使用照相机可以将珠成像。珠可以具有与所述珠相关联的可检测的代码。例如，珠可包括条形码。珠的尺寸可以变化，例如由于在有机或无机溶液中的溶胀。珠可以是疏水的。珠可以是亲水的。珠可以是生物相容的。

[0190] 可以使固体支持物(例如，珠)可视化。固体支持物可以包括可视化标签(例如，荧光染料)。可以用标识(例如，数字)将固体支持物(例如，珠)蚀刻。通过对珠成像可以将所述标识可视化。

[0191] 固体支持物可以包括不溶性、半溶性或不溶性材料。当固体支持物包括接头、支架、结构单元或附接至其上的其他反应性部分时，它可以被称为“官能化的”，而当固体支持物缺少附接至其上的这样一个反应性部分时，它可以被称为“非官能化的”。固体支持物可以在溶液中不受约束地利用，如以微量滴定孔形式；以流通形式，如在柱中；或在试纸条

(dipstick)中。

[0192] 固体支持物可以包括膜、纸、塑料、涂覆的表面、平表面、玻璃、载玻片、芯片、或其任何组合。固体支持物可以采用树脂、凝胶、微球或其他几何构型的形式。固体支持物可以包括二氧化硅芯片;微颗粒;纳米颗粒;平板;阵列;毛细管;平支持物,如玻璃纤维过滤器,玻璃表面,金属表面(钢、金、银、铝、硅以及铜),玻璃支持物,塑料支持物,硅支持物,芯片,过滤器,膜,微孔板,载玻片;塑料材料包括多孔板或膜(例如,由聚乙烯、聚丙烯、聚酰胺、聚偏二氟乙烯形成);和/或晶片;梳状物(comb);针或针头(例如,适于组合合成或分析的针阵列);或平表面(如晶片(例如,硅晶片)、带有具有或不具有滤底的凹陷的晶片)的凹陷或纳升孔的阵列中的珠。

[0193] 所述固体支持物可以包括聚合物基质(例如,凝胶、水凝胶)。该聚合物基质可能能够渗透细胞内间隙(例如,细胞器周围)。该聚合物基质可能能够贯穿循环系统进行泵送。

[0194] 固体支持物可以是生物分子。例如,固体支持物可以是核酸、蛋白质、抗体、组蛋白、细胞区室、脂质、碳水化合物等。作为生物分子的固体支持物可被扩增、翻译、转录、降解和/或修饰(例如,聚乙二醇化、sumo化、乙酰化、甲基化)。除了附接至生物分子的空间标记之外,作为生物分子的固体支持物可以提供空间和时间信息。例如,生物分子可以在未修饰时包括第一构象,但是在修饰时可以改变为第二构象。这些不同的构象可以将本披露的条形码(例如,随机条形码)暴露给靶。例如,生物分子可以包括由于生物分子的折叠而不可接近的条形码。在修饰生物分子(例如,乙酰化)时,这些生物分子可以改变构象以暴露这些条形码。修饰的时间设置可以为本披露的条形码化的方法提供另一时间维度。

[0195] 在一些实施例中,包括本披露的条形码试剂的生物分子可以位于细胞的细胞质中。激活时,该生物分子可以移动到细胞核,于此可以进行条形码化。以这种方式,所述生物分子的修饰可以编码由条形码鉴别的靶的另外的空间-时间信息。

[0196] 基底和微孔阵列

[0197] 如本文所用,基底可以指一种固体支持物。基底可以指可包括本披露的条形码和随机条形码的固体支持物。例如,基底可以包括多个微孔。例如,基底可以是包括两个或更多个微孔的孔阵列。在一些实施例中,微孔可以包括具有确定体积的小反应室。在一些实施例中,微孔可以截留一个或多个细胞。在一些实施例中,微孔只能截留一个细胞。在一些实施例中,微孔可以截留一个或多个固体支持物。在一些实施例中,微孔只能截留一个固体支持物。在一些实施例中,微孔截留单个细胞和单个固体支持物(例如,珠)。微孔可以包括本披露的组合条形码试剂。

[0198] 条形码化的方法

[0199] 本披露提供了用于估计身体样品(例如,组织、器官、肿瘤、细胞)中的不同位置处的不同靶的数目的方法。这些方法可以包括将条形码(例如,随机条形码)靠近样品放置,裂解样品,将不同靶与条形码相关联,对这些靶进行扩增和/或对靶进行数字计数。该方法可以进一步包括对获得自条形码上的空间标记的信息进行分析和/或可视化。在一些实施例中,该方法包括使样品中的多个靶可视化。将多个靶映射到样品的映射图上可以包括产生样品的二维映射图或三维映射图。可以在对样品中的多个靶进行条形码化(例如,随机条形码化)之前或之后产生二维映射图和三维映射图。将样品中的多个靶可视化包括将多个靶映射到样品的映射图上。将多个靶映射到样品的映射图上可以包括产生样品的二维映射图

或三维映射图。可以在对样品中的多个靶进行条形码化之前或之后产生二维映射图和三维映射图。在一些实施例中，可以在裂解样品之前或之后产生二维映射图和三维映射图。在产生二维映射图或三维映射图之前或之后裂解样品可包括加热样品、使样品与洗涤剂接触、改变样品的pH、或其任何组合。

[0200] 在一些实施例中，对多个靶进行条形码化包括将多个条形码与多个靶杂交以创建经条形码化的靶(例如，经随机条形码化的靶)。对多个靶进行条形码化可包括产生经条形码化的靶的索引文库。产生经条形码化的靶的索引文库可以用包括多个条形码(例如，随机条形码)的固体支持物进行。

[0201] 使样品和条形码接触

[0202] 本披露提供了用于使样品(例如，细胞)与本披露的基底接触的方法。可以使包括例如细胞、器官或组织薄片的样品与条形码(例如，随机条形码)接触。例如，通过重力流可以使这些细胞接触，其中可以使这些细胞沉淀并且产生单层细胞。该样品可以是组织薄切片。可以将薄切片置于基底上。该样品可以是一维的(例如，形成平面)。可以将该样品(例如，细胞)涂布于基底上，例如，通过在基底上生长/培养这些细胞。

[0203] 当条形码靠近靶时，靶可以与条形码进行杂交。条形码可以按不可耗尽的比率接触，使得每个不同的靶可以与本披露的不同条形码相关联。为了确保靶与条形码之间的有效关联，可以将靶与条形码交联。

[0204] 细胞裂解

[0205] 在细胞和条形码的分布之后，可以裂解细胞以释放靶分子。细胞裂解可以通过多种手段中的任何一种来完成，例如通过化学或生化手段，通过渗透冲击，或通过热裂解、机械裂解或光学裂解。可以通过添加包括洗涤剂(例如SDS、十二烷基硫酸锂、Triton X-100、Tween-20或NP-40)的细胞裂解缓冲液、有机溶剂(例如甲醇或丙酮)或消化酶(例如蛋白酶K、胃蛋白酶或胰蛋白酶)或其任何组合来裂解细胞。为了增加靶和条形码的关联，可通过例如降低裂解物的温度和/或增加裂解物的粘度来改变靶分子的扩散速率。

[0206] 在一些实施例中，可以使用滤纸将样品裂解。可以用在滤纸上部的裂解缓冲液浸渍该滤纸。在压力下可以将滤纸应用于样品，该压力可以促进样品的裂解，以及样品的靶与基底的杂交。

[0207] 在一些实施例中，裂解可以通过机械裂解、热裂解、光学裂解、和/或化学裂解来进行。化学裂解可以包括使用消化酶类，如蛋白酶K、胃蛋白酶、以及胰蛋白酶。可以通过将裂解缓冲液添加到基底中进行裂解。裂解缓冲液可以包括Tris HCl。裂解缓冲液可包括至少约0.01M、0.05M、0.1M、0.5M、或1M或更多Tris HCl。裂解缓冲液可包括至多约0.01M、0.05M、0.1M、0.5M、或1M或更多Tris HCL。裂解缓冲液可以包括约0.1M Tris HCl。裂解缓冲液的pH可以是至少约1、2、3、4、5、6、7、8、9、或10，或更高。裂解缓冲液的pH可以是至多约1、2、3、4、5、6、7、8、9、或10，或更高。在一些实施例中，该裂解缓冲液的pH是约7.5。该裂解缓冲液可以包括盐(例如，LiCl)。在裂解缓冲液中盐的浓度可以是至少约0.1M、0.5M、或1M、或更高。在裂解缓冲液中盐的浓度可以是至多约0.1M、0.5M、或1M、或更高。在一些实施例中，在裂解缓冲液中盐的浓度是约0.5M。裂解缓冲液可以包括洗涤剂(例如，SDS、十二烷基硫酸锂、曲通X、tween、NP-40)。在裂解缓冲液中洗涤剂的浓度可以是至少约0.0001％、0.0005％、0.001％、0.005％、0.01％、0.05％、0.1％、0.5％、1％、2％、3％、4％、5％、6％、或7％、或更

高。在裂解缓冲液中洗涤剂的浓度可以是至多约0.0001％、0.0005％、0.001％、0.005％、0.01％、0.05％、0.1％、0.5％、1％、2％、3％、4％、5％、6％、或7％、或更高。在一些实施例中，在裂解缓冲液中洗涤剂的浓度是约1％十二烷基硫酸锂。该裂解方法中所用时间可以依赖于所用洗涤剂的量。在一些实施例中，所用洗涤剂越多，裂解所需时间越短。裂解缓冲液可以包括螯合剂(例如，EDTA、EGTA)。在裂解缓冲液中螯合剂的浓度可以是至少约1mM、5mM、10mM、15mM、20mM、25mM、或30mM或更高。在裂解缓冲液中螯合剂的浓度可以是至多约1、5、10、15、20、25、或30mM或更高。在一些实施例中，在裂解缓冲液中的螯合剂的浓度是约10mM。裂解缓冲液可以包括还原剂(例如，β-巯基乙醇、DTT)。在裂解缓冲液中还原剂的浓度可以是至少约1、5、10、15、或20mM或更高。在裂解缓冲液中还原剂的浓度可以是至多约1、5、10、15、或20mM或更高。在一些实施例中，在裂解缓冲液中还原剂的浓度是约5mM。在一些实施例中，裂解缓冲液可以包括约0.1M Tris HCl、约pH 7.5、约0.5M LiCl、约1％十二烷基硫酸锂、约10mM EDTA，以及约5mM DTT。

[0208] 可以在约4℃、10℃、15℃、20℃、25℃、或30℃的温度进行裂解。裂解可以进行约1分钟、5分钟、10分钟、15分钟、或20分钟或更多分钟。裂解的细胞可以包括至少约100000、200000、300000、400000、500000、600000、或700000个靶核酸分子，或更多。裂解的细胞可以包括至多约100000、200000、300000、400000、500000、600000、或700000个靶核酸分子，或更多。

[0209] 将条形码附接至靶核酸分子

[0210] 在细胞裂解和核酸分子从释放之后，核酸分子可以随机地与共定位的固体支持物的条形码相关联。关联可以包括将条形码的靶识别区与靶核酸分子的互补部分的杂交(例如，条形码的寡聚(dT)可与靶的聚(A)尾相互作用)。可以选择用于杂交的测定条件(例如缓冲液pH、离子强度、温度等)以促进形成特定的稳定的杂交体。在一些实施例中，可以将从裂解的细胞释放出的核酸分子与基底上的多个探针(例如，与基底上的探针杂交)相关联。当该探针包括寡聚(dT)时，可以将mRNA分子与探针杂交，并且进行逆转录。可以将寡核苷酸的寡聚(dT)部分充当用于cDNA分子的第一链合成的引物。例如，图2中(在框216上)说明的条形码化的非限制性实例中，mRNA分子可以与珠上的条形码杂交。例如，单链的核苷酸片段可以与条形码的靶结合区杂交。

[0211] 附接可以进一步包括将条形码的靶识别区与靶核酸分子的一部分连接。例如，靶结合区可以包括可能够与限制性位点突出端(例如EcoRI粘性末端突出端)进行特异性杂交的核酸序列。测定程序还可以包括用限制性酶(例如EcoRI)处理靶核酸以产生限制性位点突出端。然后条形码可以连接到包括与限制性位点突出端互补的序列的任何核酸分子。连接酶(例如，T4DNA连接酶)可用于连接两个片段。

[0212] 例如，在图2(在框220处)中说明的条形码化的非限制性实例中，随后可以将来自多个细胞(或多个样品)的经标记的靶(例如，靶-条形码分子)例如池化至管中。经标记的靶可以通过例如回收条形码和/或附接靶-条形码分子的珠来池化。

[0213] 可以通过使用磁珠和外部施加的磁场来实现附接的靶-条形码分子的基于固体支持物的集合的检索。一旦该靶-条形码分子已经池化，所有进一步的处理可以在单个反应容器中进行。进一步的处理可以包括例如逆转录反应、扩增反应、切割反应、解离反应和/或核酸延伸反应。进一步的处理反应可以在微孔内进行，即，不首先池化来自多个细胞的经标记

的靶核酸分子。

[0214] 逆转录

[0215] 本披露提供了使用逆转录来产生靶-条形码缀合物的方法(在图2的框224中)。靶-条形码缀合物可以包括条形码以及靶核酸(即,条形码化的cDNA分子,如随机条形码化的cDNA分子)的全部或部分的互补性序列。关联的RNA分子的逆转录可以通过添加逆转录引物连同逆转录酶一起而发生。逆转录引物可以是寡聚dT引物、随机六核苷酸引物或靶特异性寡核苷酸引物。寡聚(dT)引物的长度可以是、或可以是约12-18个核苷酸,并与哺乳动物mRNA的3'端的内源性聚(A)尾结合。随机六核苷酸引物可在多个互补位点处结合至mRNA。靶特异性寡核苷酸引物通常选择性地引发目的mRNA。

[0216] 在一些实施例中,标记的RNA分子的逆转录可通过添加逆转录引物而进行。在一些实施例中,该逆转录引物是寡聚(dT)引物、随机六核苷酸引物或靶特异性寡核苷酸引物。通常,寡聚(dT)引物的长度为12-18个核苷酸,并结合至在哺乳动物mRNA的3'端的内源性聚(A)+尾。随机六核苷酸引物可在多个互补位点处结合至mRNA。靶特异性寡核苷酸引物通常选择性地引发目的mRNA。

[0217] 逆转录可以重复地发生以产生多个经标记的cDNA分子。本文披露的方法可包括进行至少约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20次逆转录反应。该方法可包括进行至少约25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100次逆转录反应。

[0218] 扩增

[0219] 可以进行一个或多个核酸扩增反应(例如,在图2的框228中)以产生经标记的靶核酸分子的多个拷贝。扩增能以多路方式进行,其中多个靶核酸序列同时进行扩增。扩增反应可用于向核酸分子添加测序衔接子。扩增反应可以包括扩增样品标记(如果存在)的至少一部分。扩增反应可以包括扩增细胞标记和/或条形码序列(例如,分子标记)的至少一部分。扩增反应可以包括扩增样品标签、细胞标记、空间标记、条形码(例如,分子标记)、靶核酸或其组合的至少一部分。扩增反应可包括扩增多个核酸的0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、97%、100%、或在这些值的任何两个之间的数字或范围。该方法可以进一步包括进行一个或多个cDNA合成反应以产生包括样品标记、细胞标记、空间标记和/或条形码序列(例如,分子标记)的靶-条形码分子的一个或多个cDNA拷贝。

[0220] 在一些实施例中,可以使用聚合酶链式反应(PCR)进行扩增。如本文使用的,PCR可以指用于通过DNA的互补链的同时引物延伸使特异性DNA序列体外扩增的反应。如本文使用的,PCR可包括所述反应的派生形式,包括但不限于RT-PCR、实时PCR、巢式PCR、定量PCR、多重PCR、数字PCR、和组装PCR。

[0221] 经标记的核酸的扩增可以包括基于非PCR的方法。非基于PCR的方法的实例包括但不限于多重置换扩增(MDA)、转录介导的扩增(TMA)、基于核酸序列的扩增(NASBA)、链置换扩增(SDA)、实时SDA、滚环扩增或环到环扩增(circle-to-circle amplification)。其他非基于PCR的扩增方法包括DNA依赖性RNA聚合酶驱动的RNA转录扩增或RNA指导的DNA合成和转录的多个循环以扩增DNA或RNA靶、连接酶链式反应(LCR)、和Qβ复制酶(Qβ)方法、回文探针的使用、链置换扩增、使用限制性内切核酸酶的寡核苷酸驱动的扩增、使引物与核酸序列

杂交并且将所得双链体在延伸反应和扩增之前切割的扩增方法、使用缺乏5'外切核酸酶活性的核酸聚合酶的链置换扩增、滚环扩增和分支延伸扩增(RAM)。在一些实施例中,扩增不产生环化转录物。

[0222] 在一些实施例中,本文披露的方法进一步包括对经标记的核酸(例如,经标记的RNA、经标记的DNA、经标记的cDNA)进行聚合酶链式反应,以产生经标记的扩增子(例如,经随机标记的扩增子)。经标记的扩增子可以是双链分子。双链分子可包括双链RNA分子、双链DNA分子或者与DNA分子杂交的RNA分子。双链分子的一条或两条链可以包括样品标记、空间标记、细胞标记、和/或条形码序列(例如,分子标记)。经标记的扩增子可以是单链分子。单链分子可包括DNA、RNA或其组合。本披露的核酸可以包括合成的或改变的核酸。

[0223] 扩增可以包括使用一个或多个非天然核苷酸。非天然核苷酸可包括光不稳定或可触发的核苷酸。非天然核苷酸的实例可以包括但不限于肽核酸(PNA)、吗啉代和锁核酸(LNA)、以及二醇核酸(GNA)与苏糖核酸(TNA)。可以将非天然核苷酸添加至扩增反应的一个或多个循环中。添加非天然核苷酸也可以用于鉴别扩增反应中特定循环或时间点的产物。

[0224] 进行一个或多个扩增反应可以包括使用一个或多个引物。一个或多个引物可以包括例如1、2、3、4、5、6、7、8、9、10、11、12、13、14、或15个或更多个核苷酸。一个或多个引物可以包括至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、或15个或更多个核苷酸。一个或多个引物可以包括少于12-15个核苷酸。一个或多个引物可以退火至多个经标记的靶(例如,随机地经经标记的靶)的至少一部分。一个或多个引物可以退火至多个经标记的靶的3'端或5'端。一个或多个引物可以退火至多个经标记的靶的内部区域。内部区可以是从该多个经标记的靶的3'端的至少约50、100、150、200、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、510、520、530、540、550、560、570、580、590、600、650、700、750、800、850、900或1000个核苷酸。一个或多个引物可以包括一组固定的引物。一个或多个引物可以包括至少一个或多个定制引物。一个或多个引物可以包括至少一个或多个对照引物。一个或多个引物可以包括至少一个或多个基因特异性引物。

[0225] 一个或多个引物可以包括通用引物。通用引物可以退火至通用引物结合位点。一个或多个定制引物可以退火至第一样品标记、第二样品标记、空间标记、细胞标记、条形码序列(例如,分子标记)、靶、或其任何组合。一个或多个引物可以包括通用引物和定制引物。定制引物可以设计用于扩增一个或多个靶。靶可以包括一个或多个样品中总核酸的子集。靶可以包括一个或多个样品中总的经标记的靶的子集。一个或多个引物可以包括至少96个或更多个定制引物。一个或多个引物可以包括至少960个或更多个定制引物。一个或多个引物可以包括至少9600个或更多个定制引物。一个或多个定制引物可以退火至两个或更多个不同的经标记的核酸。两个或更多个不同的经标记的核酸可以对应于一个或多个基因。

[0226] 可以在本披露的方法中使用任何扩增方案。例如,在一个方案中,第一轮PCR可以使用基因特异性引物和针对通用亿明达(Illumina)测序引物1序列的引物来扩增附接到珠上的分子。第二轮PCR可以使用侧翼于亿明达测序引物2序列的巢式基因特异性引物和针对通用亿明达测序引物1序列的引物扩增第一PCR产物。第三轮PCR添加P5和P7以及样品索引,以便使PCR产物进入亿明达测序文库。使用150bp x 2测序的测序可以揭示读数1上的细胞标记和条形码序列(例如,分子标记)、读数2上的基因、以及索引1读数上的样品索引。

[0227] 在一些实施例中，使用化学切割可以将核酸从基底中去除。例如，可以将存在于核酸中的化学基团或经修饰的碱基用于促进将其从固体支持物中去除。例如，酶可以用于从基底中去除核酸。例如，通过限制性内切核酸酶消化，可以将核酸从基底中去除。例如，使用尿嘧啶-d-糖基化酶(UDG)处理含有dUTP或ddUTP的核酸可以从基底中去除核酸。例如，可以使用用于核苷酸切除(例如，碱基切除修复酶(例如，脱嘌呤/脱嘧啶(AP)核酸内切酶))的酶将核酸从基底中去除。在一些实施例中，可以使用可光解(photocleavable)基团以及光将核酸从基底中去除。在一些实施例中，可以使用可切割接头从基底中去除核酸。例如，可切割接头可以包括以下中的至少一种：生物素/亲和素、生物素/链霉抗生物素蛋白、生物素/中性链亲和素、Ig蛋白A、光不稳定性接头、酸或碱不稳定性接头基团、或适配体。

[0228] 当探针是基因特异性时，可以将这些分子与探针杂交，并且进行逆转录和/或扩增。在一些实施例中，在核酸已经合成(例如，逆转录)之后，可以将其扩增。扩增能以多重方式进行，其中多个靶核酸序列同时进行扩增。扩增可以将测序衔接子添加至核酸。

[0229] 在一些实施例中，例如，用桥接扩增可以将扩增在基底上进行。cDNA可以是同聚物尾部，使用基底上的寡聚(dT)探针，以产生用于桥接扩增的相容端。在桥接扩增中，与模板核酸的3'端互补的引物可以是共价附接至固体颗粒的每对引物的第一引物。当含有模板核酸的样品与颗粒接触并进行单个热循环时，可以将模板分子退火至第一引物，并且第一引物通过添加核苷酸而向前延伸以形成双链体分子，该双链体分子由模板分子和与模板互补的新形成的DNA链构成。在下一循环的加热步骤中，双链体分子可以变性，从颗粒释放模板分子，并通过第一引物将互补性DNA链附接至颗粒。在随后的退火和延伸步骤的退火阶段中，互补链可以与第二引物杂交，该第二引物在从第一引物去除的位置处与互补链的片段互补。该杂交可导致互补链在通过共价键固定到第一引物的第一和第二引物之间形成桥接，并通过杂交形成第二引物。在延伸阶段，通过在相同的反应混合物中添加核苷酸，第二引物可以按相反方向延伸，从而将桥转化为双链桥。然后开始下一个循环，并且该双链桥可以变性以产生两个单链核酸分子，每个单链核酸分子的一端分别经第一和第二引物附接至颗粒表面，其中每个单链核酸分子的另一端是未附接的。在该第二个循环的退火和延伸步骤中，每条链可以与先前未使用的另外的互补引物杂交在相同的颗粒上，以形成新的单链桥。将现在杂交的两个先前未使用的引物延伸从而将两个新桥转换成双链桥。

[0230] 扩增反应可以包括扩增多个核酸的至少1％、2％、3％、4％、5％、6％、7％、8％、9％、10％、15％、20％、25％、30％、35％、40％、45％、50％、55％、60％、65％、70％、75％、80％、85％、90％、95％、97％或100％。

[0231] 对经标记的核酸的扩增可以包括基于PCR的方法或非基于PCR的方法。对经标记的核酸的扩增可以包括对经标记的核酸的指数式扩增。对经标记的核酸的扩增可以包括对经标记的核酸的线性扩增。扩增可以通过聚合酶链式反应(PCR)来进行。PCR可指用于通过DNA的互补链的同时引物延伸使特异性DNA序列体外扩增的反应。PCR可涵盖该反应的派生形式，包括但不限于，RT-PCR、实时PCR、巢式PCR、定量PCR、多重PCR、数字PCR、抑制PCR、半抑制PCR以及装配PCR。

[0232] 在一些实施例中，所述经标记的核酸的扩增包括非基于PCR的方法。非基于PCR的方法的实例包括但不限于多重置换扩增(MDA)、转录介导的扩增(TMA)、基于核酸序列的扩增(NASBA)、链置换扩增(SDA)、实时SDA、滚环扩增或环到环扩增(circle-to-circle

amplification)。其他非基于PCR的扩增方法包括DNA依赖性RNA聚合酶驱动的RNA转录扩增或RNA指导的DNA合成和转录的多个循环以扩增DNA或RNA靶、连接酶链式反应(LCR)、Qβ复制酶(Qβ)、回文探针的使用、链置换扩增、使用限制性内切核酸酶的寡核苷酸驱动的扩增、使引物与核酸序列杂交并且将所得双链体在延伸反应和扩增之前切割的扩增方法、使用缺乏5'外切核酸酶活性的核酸聚合酶的链置换扩增、滚环扩增和分支延伸扩增(RAM)。

[0233] 在一些实施例中,本文披露的这些方法进一步包括对扩增的扩增子(例如,靶)进行巢式聚合酶链式反应。扩增子可以是双链分子。双链分子可包括双链RNA分子、双链DNA分子或者与DNA分子杂交的RNA分子。双链分子的一条或两条链可包括样品标签或分子鉴定物标记。可替代地,该扩增子可以是单链分子。单链分子可包括DNA、RNA或其组合。本发明的核酸可以包括合成的或改变的核酸。

[0234] 在一些实施例中,该方法包括反复扩增经标记的核酸以产生多个扩增子。本文披露的方法可包括进行至少约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20次扩增反应。可替代地,该方法包括进行至少约25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100次扩增反应。

[0235] 扩增可进一步包括将一个或多个对照核酸添加至一个或多个包括多个核酸的样品中。扩增可进一步包括将一个或多个对照核酸添加至多个核酸中。对照核酸可以包括对照标记。

[0236] 扩增可以包括使用一个或多个非天然核苷酸。非天然核苷酸可以包括光不稳定和/或可触发的核苷酸。非天然核苷酸的实例包括但不限于肽核酸(PNA)、吗啉代和锁核酸(LNA)以及二醇核酸(GNA)与苏糖核酸(TNA)。可以将非天然核苷酸添加至扩增反应的一个或多个循环中。添加非天然核苷酸也可以用于鉴别扩增反应中特定循环或时间点的产物。

[0237] 进行一个或多个扩增反应可以包括使用一个或多个引物。一个或多个引物可以包括一个或多个寡核苷酸。一个或多个寡核苷酸可以包括至少约7至9个核苷酸。一个或多个寡核苷酸可包括少于12-15个核苷酸。一个或多个引物可以退火至多个经标记的核酸的至少一部分。一个或多个引物可以退火至多个经标记的核酸的3'端和/或5'端。一个或多个引物可以退火至多个经标记的核酸的内部区。内部区可以是从该多个经标记的核酸的3'端的至少约50、100、150、200、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、510、520、530、540、550、560、570、580、590、600、650、700、750、800、850、900或1000个核苷酸。一个或多个引物可以包括一组固定的引物。一个或多个引物可以包括至少一个或多个定制引物。一个或多个引物可以包括至少一个或多个对照引物。一个或多个引物可以包括至少一个或多个管家基因引物。一个或多个引物可以包括通用引物。通用引物可以退火至通用引物结合位点。一个或多个定制引物可以退火至第一样品标签、第二样品标签、分子鉴定物标记、核酸或它们的产物。一个或多个引物可以包括通用引物和定制引物。定制引物可以被设计成扩增一个或多个靶核酸。靶核酸可以包括一个或多个样品中总核酸的子集。在一些实施例中,这些引物是附接至本披露的阵列的探针。

[0238] 在一些实施例中,条形码化(例如,随机地条形码化)在样品中的多个靶进一步包括产生条形码化的片段的索引文库。不同的条形码的条形码序列(例如,不同的随机条形码的分子标记)可以彼此不同。产生经条形码化的靶(例如,经随机条形码化的靶)的索引文库

包括从样品中的多个靶产生多个索引多核苷酸。例如，对于包括第一索引靶和第二索引靶的经条形码化的靶的索引文库，第一索引多核苷酸的标记区与第二索引多核苷酸的标记区可以具有、具有约、具有至少、或具有至多1、2、3、4、5、6、7、8、9、10、20、30、40、50个核苷酸的差异、或在这些值的任何两个之间的数字或范围的核苷酸差异。在一些实施例中，产生经条形码化的靶的索引文库包括使多个靶(例如mRNA分子)与包括聚(T)区和标记区的多个寡核苷酸接触；以及使用逆转录酶进行第一链合成以产生单链标记的cDNA分子(每个包括cDNA区和标记区)，其中多个靶包括不同序列的至少两个mRNA分子，且多个寡核苷酸包括不同序列的至少两个寡核苷酸。产生经条形码化的靶的索引文库可进一步包括扩增单链标记的cDNA分子以产生双链标记的cDNA分子；以及在双链标记的cDNA分子上进行巢式PCR以产生标记的扩增子。在一些实施例中，该方法可包括产生衔接子标记的扩增子。

[0239] 随机条形码化可以使用核酸条形码或标签以标记单个核酸(例如，DNA或RNA)分子。在一些实施例中，其涉及将DNA条形码或标签添加至cDNA分子，因为它们是从mRNA产生的。可以进行巢式PCR以最小化PCR扩增偏差。可以使用例如下一代测序(NGS)添加衔接子用于测序。例如在图2的框232处，可以使用测序结果以确定靶的一个或多个拷贝的细胞标记、条形码序列(例如，分子标记)、和核苷酸片段的序列。

[0240] 图3是显示产生经条形码化的靶(例如，经随机条形码化的靶)例如mRNA的索引文库的非限制性示例性过程的示意图。如步骤1显示，逆转录过程可以编码具有独特条形码序列(例如，分子标记)、细胞标记和通用PCR位点的每个mRNA分子。例如，通过将一组条形码(例如随机条形码)310杂交(例如随机杂交)到RNA分子302的聚(A)尾区308，可以将RNA分子302逆转录以产生经标记的cDNA分子304(包括cDNA区306)。每个条形码310可包括靶结合区，例如聚(dT)区312、条形码序列或分子标记314、和通用PCR区316。

[0241] 在一些实施例中，细胞标记可包括3至20个核苷酸。在一些实施例中，条形码序列(例如，分子标记)可包括3至20个核苷酸。在一些实施例中，多个随机条形码的每个进一步包括通用标记和细胞标记的一个或多个，其中通用标记对于固体支持物上的多个随机条形码是相同的且细胞标记对于固体支持物上的多个随机条形码是相同的。在一些实施例中，通用标记可包括3至20个核苷酸。在一些实施例中，细胞标记包括3至20个核苷酸。

[0242] 在一些实施例中，标记区314可包括条形码序列或分子标记318和细胞标记320。在一些实施例中，标记区314可包括通用标记、维度标记、和细胞标记的一个或多个。条形码序列或分子标记318的长度可以是、可以是约、可以是至少、或可以是至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100个核苷酸、或在这些值的任何之间的数目或范围的核苷酸。细胞标记320的长度可以是、可以是约、可以是至少、或可以是至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100个核苷酸、或在这些值的任何之间的数字或范围的核苷酸。通用标记的长度可以是、可以是约、可以是至少、或可以是至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100个核苷酸、或在这些值的任何之间的数字或范围的核苷酸。对于固体支持物上的多个随机条形码，通用标记可以是相同的，且对于固体支持物上的多个随机条形码，细胞标记是相同的。维度标记的长度可以是、可以是约、可以是至少、或可以是至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100个核苷酸、或在这些值的任何之间的数字或范围的核苷酸。

[0243] 在一些实施例中，标记区314可包含、包含约、包含至少、或包含至多1、2、3、4、5、6、

7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000个不同标记、或在这些值的任何之间的数字或范围的不同标记，如条形码序列或分子标记318和细胞标记320。每个标记的长度可以是、可以是约、可以是至少、或可以是至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100个核苷酸、或在这些值的任何之间的数字或范围的核苷酸。一组条形码或随机条形码310可以含有、含有约、含有至少、或可以是至多10、20、40、50、70、80、90、10^2、10^3、10^4、10^5、10^6、10^7、10^8、10^9、10^{10}、10^{11}、10^{12}、10^{13}、10^{14}、10^{15}、10^{20}个条形码或随机条形码310、或在这些值的任何之间的数字或范围的条形码或随机条形码310。并且条形码或随机条形码310的组可以例如，各自含有独特标记区314。经标记的cDNA分子304可以进行纯化以去除过量条形码或随机条形码310。纯化可以包括Ampure珠纯化。

[0244] 如步骤2所示，来自逆转录过程的产物在步骤1中可以池化至1管中，且用第1PCR引物池和第1通用PCR引物进行PCR扩增。因为独特标记区314，池化是可能的。特别地，可以将标记的cDNA分子304扩增以产生巢式PCR标记的扩增子322。扩增可包括多重PCR扩增。扩增可以包括在单一反应体积中用96种多重引物进行的多重PCR扩增。在一些实施例中，在单一反应体积中，多重PCR扩增可以利用、利用约、利用至少、或利用至多10、20、40、50、70、80、90、10^2、10^3、10^4、10^5、10^6、10^7、10^8、10^9、10^{10}、10^{11}、10^{12}、10^{13}、10^{14}、10^{15}、10^{20}个多重引物、或在这些值的任何之间的数字或范围的多重引物。扩增可包括靶向特异性基因的定制引物326A-C和通用引物328的第1PCR引物池324。定制引物326可以与经标记的cDNA分子304的cDNA部分306'内的区域杂交。通用引物328可以与经标记的cDNA分子304的通用PCR区域316杂交。

[0245] 如图3的步骤3中显示，来自步骤2中的PCR扩增的产物可以用巢式PCR引物池和第2通用PCR引物进行扩增。巢式PCR可以最小化PCR扩增偏差。例如，巢式PCR标记的扩增子322可通过巢式PCR进行进一步扩增。巢式PCR可以包括在单个反应体积中用巢式PCR引物332a-c的巢式PCR引物池330和第2通用PCR引物328'的多重PCR。巢式PCR引物池328可含有、含有约、含有至少、或含有至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000个不同巢式PCR引物330、或在这些值的任何之间的数字或范围的不同巢式PCR引物330。巢式PCR引物332可含有衔接子334，并与经标记的扩增子322的cDNA部分306"内的区域杂交。通用引物328'可含有衔接子336，并与经标记的扩增子322的通用PCR区域316杂交。因此，步骤3产生衔接子标记的扩增子338。在一些实施例中，巢式PCR引物332和第2通用PCR引物328'可以不含有衔接子334和336。相反，衔接子334和336可以连接到巢式PCR的产物以产生衔接子标记的扩增子338。

[0246] 如步骤4中显示，可以使用文库扩增引物将来自步骤3的PCR产物进行PCR扩增用于测序。特别地，可以将衔接子334和336用于对衔接子标记的扩增子338执行一个或多个另外的测定。衔接子334和336可以与引物340和342杂交。一个或多个引物340和342可以是PCR扩增引物。一个或多个引物340和342可以是测序引物。一个或多个衔接子334和336可以用于衔接子标记的扩增子338的进一步扩增。一个或多个衔接子334和336可以用于对衔接子标记的扩增子338进行测序。引物342可含有板索引344，使得使用同一组条形码或随机条形码310产生的扩增子可以使用下一代测序(NGS)在一轮测序反应中测序。

[0247] 使用树状图对表达谱进行聚类

[0248] 本文中公开了用于鉴定靶以区分细胞类型的方法。在一些实施例中，该方法包括：

(a)接收靶计数数据结构,其中该靶计数数据结构包含多个细胞的表达谱,并且其中该多个细胞的表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目;(b)基于该靶计数数据结构和该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行层次聚类,以产生代表该多个细胞的表达谱的树状图,其中该树状图包含多个节点,其中该多个节点包括根节点、多个叶节点和多个非根非叶节点,其中该多个叶节点中的每个叶节点代表该多个细胞中不同细胞的表达谱,且其中该根节点代表该多个细胞的表达谱;(c)在从该树状图的根节点到该树状图的多个叶节点遍历该树状图的多个节点的每个节点时:(1)确定该节点分裂成该节点的子节点是有效还是无效的(例如,子节点之间的差异不显著);和(2)如果该节点分裂成该节点的子节点是无效的,则将该节点添加到合并聚类集中;(d)迭代地,对于该合并聚类集中的每个第一节点,如果该合并聚类集中的该第一节点与该合并聚类集中最接近该第一节点的第二节点之间的距离在合并距离阈值内,则将该第一节点与该第二节点合并以生成合并的节点,该合并的节点包含由该第一节点和该第二节点代表的表达谱;和(e)对于该合并聚类集中的每个节点,基于由该节点代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。

[0249] 图4是显示鉴定用于区分细胞类型的靶的非限制性示例性方法400的流程图。方法400将分子计数数据结构(例如,分子计数矩阵)映射到一组聚类标记和一组重要基因。在一些实施例中,输入可以是分子计数的N×M矩阵或阵列(其中第i,j个条目描述了使用来自细胞i的读数观察到的基因j的分子数目)。该算法可以生成两个输出。第一输出可以是一组N个标记,每个细胞一个(具有相同标记的细胞可以被推断为“相似”)。第二输出可以是一组可用于将聚类彼此区分的基因。

[0250] 方法400使用树状图分裂、测试和合并方法生成两个输出。在预处理数据并生成距离数据结构(例如,距离矩阵)D之后,算法可以将D层次聚类以产生树状图。该算法可包括两个阶段。在分裂和测试阶段期间,方法400从树状图的顶部开始(例如,图5中树或树状图500的根节点504)。树状图500包括根节点504,多个叶节点520a-520h,以及多个非根非叶节点508a-508b,512a-512b和516a-516b。在树状图的每个节点处(叶节点除外,例如叶节点520a-520h),树被分裂成两个子树(例如,根节点504被分裂成两个子树508a、508b)。该分裂对应于聚类(例如,包含两个或更多个细胞的表达谱)被分分裂两个候选的子聚类(例如,各包含至少一个细胞的表达谱)。可以对分裂的质量进行评分。如果认为子聚类足够不同,则继续在每个子树上运行算法。否则,对于树状图的这部分终止该算法。此阶段为数据集生成一组标记。在合并阶段期间,方法400使用在分裂和测试阶段期间生成的标记来确定是否应该组合这些聚类中的任何聚类以形成一个聚类。在一些实施例中,分裂和测试阶段倾向于产生几个样本各自的小聚类。合并阶段可以通过将较小的聚类与较大的聚类合并来“清理”较小的聚类。

[0251] 在方框404处,方法400可以包括接收分子计数数据结构(例如,分子计数矩阵)。该矩阵可以仅包括全部非负条目,并且趋向于大且稀疏的。在一些实施例中,输入可以是分子计数的N×M矩阵或阵列(其中第i,j个条目描述了使用来自细胞i的读数观察到的基因j的分子数目)。

[0252] 在方框408处,方法400可以包括预处理分子计数数据结构以生成距离数据结构(例如,距离矩阵)。在一些实施例中,输入数据结构是经对数转换的。在取自然对数之前,向

每个条目添加值1。可以使用相关距离来描述每对细胞之间的成对相异性d。对于细胞c_i和c_j,两种细胞之间的相关距离可使用等式[1]

[0253] $$d(\boldsymbol{c}_i,\boldsymbol{c}_j)=1-\frac{(\boldsymbol{c}_i-\bar{\boldsymbol{c}}_i)\cdot(\boldsymbol{c}_j-\bar{\boldsymbol{c}}_j)}{\|\boldsymbol{c}_i-\bar{\boldsymbol{c}}_i\|_2\|\boldsymbol{c}_j-\bar{\boldsymbol{c}}_j\|_2}$$确定, 等式[1]

[0254] 其中$\bar{\boldsymbol{c}}_i$指c_i的所有要素的均值。预处理步骤的输出可以是沿对角线具有0的距离的方形对称矩阵D。

[0255] 在方框412处,方法400可以包括对细胞的表达谱进行层次聚类以生成树状图。对细胞的表达谱进行层次聚类以生成树状图可以包括迭代地合并树状图的两个最接近的聚类。所有聚类可以以具有如上所述的成对距离的各个点启动。计算聚类之间的距离D是使用完全连接完成的。对于聚类A和B,两个聚类之间的距离可以使用等式[2]:

[0256] $D(A,B)=\max_{a\in A,b\in B}d(a,b)$确定。等式[2]

[0257] 在该方框处可获得全树状图。在一些实施例中,聚类A的聚类内相关性和聚类B的聚类内相关性高于聚类A和聚类B的聚类间相关性。聚类A的聚类内相关性和聚类B的聚类内相关性的量度或指标高于聚类A和聚类B的聚类间相关性。聚类A的聚类内相关性和聚类B的聚类内相关性的量度可以基于以下的至少一种:聚类A和聚类B的聚类内最大相关性,聚类A和聚类B的聚类内平均相关性,聚类A和聚类B的聚类内中值相关性,聚类A和聚类B的聚类内最小相关性,以及其任何组合。聚类A的聚类内相关性可以基于以下的至少一种:聚类A的聚类内最大相关性,聚类A的聚类内平均相关性,聚类A的聚类内中值相关性,聚类A的聚类内最小相关性,以及其任何组合。聚类B的聚类内相关性可以基于以下的至少一种:聚类B的聚类内最大相关性,聚类B的聚类内平均相关性,聚类B的聚类内中值相关性,聚类B的聚类内最小相关性,以及其任何组合。聚类A和聚类B的聚类间相关性可以基于以下的至少一种:聚类A和聚类B的聚类间最大相关性,聚类A和聚类B的聚类间平均相关性,聚类A和聚类B的聚类间中值相关性,聚类A和聚类B的聚类间最小相关性,以及其任何组合。例如,两个子聚类的聚类内中值相关性可以高于聚类间中值相关性。

[0258] 在方框416处,方法400可以包括分裂和测试树状图以生成一组标记。可以在树状图的顶部开始分裂和测试。给定树状图子树T0,树可以分裂成恰好两个子树TL和TR。可以执行统计学检验以确定左子树TL中的细胞是否与右子树TR中的细胞充分不同。在一些实施例中,统计学检验涉及对两个群体的每个基因进行Welch t检验。如果两个群体中的方差估算为0,则可能出现无穷大的t统计量;这些情况可以忽略。如果所有检验中的最小p值低于某个阈值(针对错误检测率经保守校正),则可以认为该分裂是有效的,并且在两个子树上再次执行该算法。如果最小p值不低于该阈值,则方法400对于子树T0终止。如果TL包含恰好1个样本(即,TL是单例(singleton)),则可以忽略TL并且算法对TR重复该过程。如果TR包含恰好1个样本,则可以忽略TR并且算法对TL重复该过程。如果TL和TR各自包含恰好1个样本,则算法对子树T0终止。

[0259] 在方框416处,方法400可以包括如下确定聚类标记。首先,所有子树都可以标记为“r”。每次发生分裂并且由于p值问题而未被拒绝时,TL中的所有细胞的标记都附加有“L”,并且TR中的所有细胞的标记可以附加“R”。这意味着当跳过单例时,标记仍然受到影响。单例自动获得不与任何其他数据点共享的独特标记。

[0260] 在方框416处,方法400可以包括确定每个最终聚类的内聚性。如果最终聚类内的

所有样本彼此相距很远(即,聚类内的成对距离不在底部,例如所有距离的50百分位),则可以解散聚类。然后可以将每个样本标记为单例。

[0261] 在方框420处,方法400可以包括合并在方框416处生成的标记组以生成另一组标记。在一些实施例中,合并可以是两阶段过程。在第一阶段,每个单例可以与最近邻居放在同一个聚类中,所述最近邻居是使用来自预处理方框408的距离数据结构(例如,距离矩阵)确定的。如果单例与其最近邻居的距离位于距离的前10%内(即它远离所有其他细胞),则该单例可以被标记为异常者并保留在其自己的聚类中。此第一步确保所有聚类包含至少两个非异常成员。在第二步,在去除异常者之后,使用若干统计学检验计算聚类之间的成对距离,得到成对聚类距离矩阵D_c。两个聚类之间的距离估算为通过对所有基因的独立Welch t检验获得的最小p值的负对数。从总体最小距离开始,合并两个相应的聚类。计算新聚类与所有现有聚类的距离,并重复该过程,直到所有成对距离都高于特定距离。软件包还通过在D_c上运行Louvain算法提供基于社区检测的方法来合并。

[0262] 在方框424处,方法400可以包括从在方框420处确定的标记组中选择特征以鉴定用于区分细胞类型的特征。在一些实施例中,方法400可以使用从合并方框420生成的标记来执行两种类型的特征选择。对于第一种类型的特征选择,在分裂和测试方框期间,每次保留分裂时,保存具有K个最小p值的K个基因。可以从更靠近树状图顶部的分裂保存更多的基因。最终,输出来自所有分裂的并集的独特基因的列表。对于第二种类型的特征选择,对于每个聚类,仅使用在感兴趣的聚类中具有更高均值的基因来完成若干个一对余检验。可以输出每个聚类的重要基因的表以及关于每个基因的一些额外信息(例如p值,倍数变化,聚类内的均值表达水平)。

[0263] 方法400可以包括实施探索性分析。在一些实施例中,方法400可以利用若干个函数来使分裂和合并阶段中的某些步骤可视化。例如,这些函数例示了分裂(或合并)中涉及的细胞,在每个子树(或组合的聚类)中结束的细胞,以及指示该分裂(或合并)的基因。再例如,方法400可以在所有聚类之间实施成对比较(例如,以确定哪些基因区分每对聚类)和实施函数用于绘制树状图。方法400可以基于聚类内成对距离的分布。在一些实施例中,方法400可以包括还提供实施参数扫描(parameter sweep)。

[0264] 对表达谱进行聚类

[0265] 本文中公开了用于鉴定靶以区分细胞类型的方法。在一些实施例中,该方法包括:(a)接收多个细胞的表达谱,其中所述表达谱包含针对该多个细胞中每个细胞的多个靶中每个靶的数目;(b)基于该多个细胞的表达谱之间的距离对该多个细胞的表达谱进行聚类以产生多个表达谱聚类,其中每个聚类与(1)母聚类和(2)两个或更多个子聚类中的一者或两者具有一个或多个关联,其中该母聚类代表由该聚类代表的该多个细胞的一个或多个细胞的表达谱,且其中该聚类代表由两个或更多个子聚类代表的表达谱;(c)对于具有两个或更多个子聚类的每个聚类,如果该聚类与该两个或更多个子聚类之间的关联无效(例如,该两个或更多个子聚类之间的差异不显著),则将该聚类添加到合并聚类集中;(d)迭代地,对于该合并聚类集中的每个第一聚类,如果该合并聚类集中的该第一聚类与该合并聚类集中最接近该第一聚类的第二聚类之间的距离在合并距离阈值内,则将该第一聚类和该第二聚类合并以生成合并的聚类,其中该合并的聚类包含该第一聚类和该第二聚类的表达谱;和(e)对于该合并聚类集中的每个聚类,基于该聚类所代表的细胞的多个靶的表达谱来鉴定

用于区分细胞类型的靶。

[0266] 图6是显示通过对细胞的表达谱聚类来鉴定用于区分细胞类型的靶的非限制性示例性方法的流程图。在方框604，方法600接收多个细胞的表达谱。每个表达谱可包含针对多个细胞中不同细胞的多个靶中每个靶的数目。在一些实施例中，接收该多个细胞的表达谱包括接收靶计数数据结构(例如，靶计数矩阵)。靶计数矩阵的每行可包含多个细胞中细胞的表达谱。

[0267] 在不同实现方式中，接收的表达谱的数目可以是不同的。在一些实施例中，接收的表达谱的数目可以是、或约10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、或在这些值的任何两个之间的数字或范围。在一些实施例中，接收的表达谱的数目可以是至少、或至多10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、或10000。

[0268] 在一些实施例中，方法600包括：在方框604处接收多个细胞的表达谱之前：(f)使用多个随机条形码对该多个细胞中的多个靶进行随机条形码编码以创建多个经随机经条形码化的靶，其中该多个随机条形码中的每一个包括细胞标记和分子标记，其中从不同细胞的靶创建的经随机经条形码化的靶具有不同的细胞标记，并且其中从该多个细胞中一个细胞的靶创建的经随机经条形码化的靶具有不同的分子标记；获得该多个经随机经条形码化的靶的测序数据；以及对于多个细胞中的每个：(1)针对该细胞，对测序数据中与该多个靶中每个靶相关联的具有不同序列的分子标记的数目进行计数；以及(2)基于(1)中计数的测序数据中与该靶相关联的具有不同序列的分子标记的数目，估算该细胞的多个靶中每个靶的数目。在一些实施例中，多个细胞中该细胞的表达谱包含(2)中估算的该细胞的多个靶中每个靶的数目。

[0269] 在方框608处，方法600可以包括将多个细胞的表达谱聚类以生成多个表达谱聚类。方法600可以基于多个细胞的表达谱之间的距离来生成表达谱的聚类。由每个聚类代表的表达谱的数目在不同的实现方式中可以是不同的。在一些实施例中，每个聚类代表的表达谱的数目可以是、或约1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、或在这些值的任何两个之间的数字或范围内。在一些实施例中，每个聚类代表的表达谱的数目可以是至少、或至多1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、或10000。在一些实施例中，每个聚类代表的表达谱的数目可以是、或约方框604处接收的表达谱的数目的1％，2％，3％，4％，5％，6％，7％，8％，9％，10％，20％，30％，40％，50％，60％，70％，80％，90％，100％、或在这些值的任何两个之间的数字或范围内。在一些实施例中，每个聚类代表的表达谱的数目可以是至少、或至多方框604处接收的表达谱的数目的1％，2％，3％，4％，5％，6％，7％，8％，9％，10％，20％，30％，40％，50％，60％，70％，80％，90％，或100％。

[0270] 每个聚类可以与(1)母聚类和(2)两个或更多个子聚类(如3、4、5、6、7、8、9、10或更多个子聚类)的一者或两者具有关联。母聚类代表由该聚类代表的多个细胞的一个或多个细胞的表达谱。聚类代表由其两个或更多个子聚类代表的表达谱。

[0271] 在一些实施例中，表达谱可以参照图4，如图4的方框412所描述的进行聚类。例如，方法600可以包括基于多个细胞的表达谱之间的距离对多个细胞的表达谱进行层次聚类，以产生代表多个细胞的表达谱的树状图。该树状图可以包含多个聚类。该多个聚类可以包含根聚类、多个叶聚类、和多个非根非叶聚类。叶聚类的数目可以例如与表达谱的数目n相同。非根非叶聚类的数目可以是，例如n-2。

[0272] 多个叶聚类和多个非根非叶聚类中的每个可与母聚类关联。根聚类和多个非根非叶聚类中的每个可与左子聚类和右子聚类关联，且代表由聚类的左子聚类和右子聚类所代表的表达谱。根聚类可代表多个细胞的表达谱。在一些实现方式中，叶聚类可代表细胞的表达谱。非根非叶聚类可代表由该非根非叶聚类的子聚类所代表的细胞的表达谱。根聚类可代表其子聚类的表达谱。

[0273] 在一些实施例中，在方框608处基于多个细胞的表达谱之间的距离对多个细胞的表达谱进行聚类包括：将多个细胞的每个表达谱分配给多个聚类中不同的叶聚类；且迭代地将多个聚类的第一聚类和第二聚类组合以生成该第一聚类和该第二聚类的母聚类，前提是如果该第二聚类是多个聚类中最接近第一聚类的聚类的话。第一聚类和第二聚类之间的距离可以是具有由第一聚类代表的表达谱的任何细胞与具有由第二聚类代表的表达谱的任何细胞之间的最大距离。

[0274] 在一些实施例中，第一聚类的聚类内相关性和第二聚类的聚类内相关性高于第一聚类和第二聚类的聚类间相关性。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度或指标高于第一聚类和第二聚类的聚类间相关性。第一聚类的聚类内相关性和第二聚类的聚类内相关性的量度可以基于以下的至少一种：第一聚类和第二聚类的聚类内最大相关性，第一聚类和第二聚类的聚类内平均相关性，第一聚类和第二聚类的聚类内中值相关性，第一聚类和第二聚类的聚类内最小相关性，以及其任何组合。第一聚类的聚类内相关性可以基于以下的至少一种：第一聚类的聚类内最大相关性，第一聚类的聚类内平均相关性，第一聚类的聚类内中值相关性，第一聚类的聚类内最小相关性，以及其任何组合。第二聚类的聚类内相关性可以基于以下的至少一种：第二聚类的聚类内最大相关性，第二聚类的聚类内平均相关性，第二聚类的聚类内中值相关性，第二聚类的聚类内最小相关性，以及其任何组合。第一聚类和第二聚类的聚类间相关性可以基于以下的至少一种：第一聚类和第二聚类的聚类间最大相关性，第一聚类和第二聚类的聚类间平均相关性，第一聚类和第二聚类的聚类间中值相关性，第一聚类和第二聚类的聚类间最小相关性，以及其任何组合。

[0275] 在一些实施例中，方法600可包括，在方框608处基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类之前：确定多个细胞的表达谱的距离数据结构(例如距离矩阵)。该矩阵可以具有n x n的维度，其中n表示在方框604处接收的表达谱的数目。该距离矩阵的每个对角元素具有零值。在方框608处基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类可包括：基于距离数据结构将多个细胞的表达谱聚类以产生多个表达谱聚类。多个细胞的表达谱之间的距离可以是多个细胞的表达谱之间的成对相关距离。

[0276] 在一些实施例中，方法600可包括，在(i)中确定距离数据结构之前，将靶计数数据结构对数转换为经对数转换的靶计数数据结构(例如经对数转换的靶计数矩阵)。确定靶计数数据结构的元素的距离数据结构包括确定该经对数转换的靶计数数据结构的距离数据

结构。在方框608处基于多个细胞的表达谱之间的距离将多个细胞的表达谱聚类以产生多个表达谱聚类包括基于该经对数转换的靶计数数据结构和该距离数据结构将多个细胞的表达谱聚类以产生多个聚类。将靶计数数据结构对数转换为经对数转换的靶计数数据结构可以包括通过增量增加靶计数数据结构的每个元素的值。增量可以是例如,1、2、3、4、5、6、7、8、9、10或更高。

[0277] 在方框612处,方法600可包括将具有两个或更多个子聚类的表达谱聚类的每个聚类添加到合并聚类集中,如果该聚类与其子聚类之间的关联无效的话(例如,子聚类之间的差异不显著)。在一些实施例中,如果表达谱已经在方框608处聚类以产生树状图,则方法600可通过参照图4如图4的方框416所描述的分裂和测试树状图以产生标记组,将具有两个或更多个子聚类的每个聚类添加到合并聚类集中。

[0278] 在一些实施例中,对于具有两个或更多个子聚类的每个聚类,如果该聚类与其两个或更多个子聚类之间的关联无效的话,则方法600可通过以下将该聚类添加到合并聚类集中:当从树状图的根聚类到树状图的多个叶聚类遍历树状图的每个聚类时:(1)确定该聚类与该聚类的子聚类的关联是有效还是无效的;和(2)如果关联无效,则将该聚类添加到合并聚类集。

[0279] 在方框616处,方法600可包括将合并聚类集中的每个聚类与合并聚类集中其最近的聚类合并,如果这两个聚类之间的距离在合并距离阈值内的话。合并过的聚类包含该第一聚类和第二聚类的表达谱。方法600可如参照图4如图4方框420所描述的将合并聚类集中的每个聚类与其最近的聚类合并。

[0280] 在一些实施例中,方法600可包括在遍历树状图的多个聚类时该方法在每个聚类处包括:如果关联有效,则继续从该聚类遍历到该聚类的左子聚类和右子聚类;且如果关联无效,则停止从该聚类遍历到该聚类的左子聚类和右子聚类。确定该聚类与该聚类的子聚类的关联是有效还是无效可包括:如果左子聚类和右子聚类之间的距离高于关联阈值,则确定该关联有效,否则为无效。

[0281] 在一些实施例中,该第一节点的节点内相关性和该第二节点的节点内相关性中的至少一个可以高于该第一节点和该第二节点的节点间相关性。该第一节点的节点内相关性和该第二节点的节点内相关性的量度或指标可以高于该第一节点和该第二节点的节点间相关性。第一节点的节点内相关性和第二节点的节点内相关性的量度可以基于以下中的至少一种:第一节点和第二节点的节点内最大相关性,第一节点和第二节点的节点内平均相关性,第一节点和第二节点的节点内中值相关性,第一节点和第二节点的节点内最小相关性,以及其任何组合。第一节点的节点内相关性可以基于以下中的至少一种:第一节点的节点内最大相关性,第一节点的节点内平均相关性,第一节点的节点内中值相关性,第一节点的节点内最小相关性,以及其任何组合。第二节点的节点内相关性可以基于以下中的至少一种:第二节点的节点内最大相关性,第二节点的节点内平均相关性,第二节点的节点内中值相关性,第二节点的节点内最小相关性,以及其任何组合。第一节点和第二节点的节点间相关性可以基于以下中的至少一种:第一节点和第二节点的节点间最大相关性,第一节点和第二节点的节点间平均相关性,第一节点和第二节点的节点间中值相关性,第一节点和第二节点的节点间最小相关性,以及其任何组合。

[0282] 一些实施例中,左子聚类和右子聚类之间的距离可以基于左子聚类和右子聚类代

表的表达谱之间对多个靶的每个靶实施的统计学检验来确定。统计学检验可以是例如Welch t检验。左子聚类和右子聚类之间的距离可以基于在由左子聚类代表的表达谱和右子聚类代表的每种表达谱之间对多个靶的每个靶实施的统计学检验的最大p值来确定。

[0283] 在一些实施例中，方法600包括在方框616处将第一聚类与第二聚类合并以产生合并聚类之前：将合并聚类集中代表单细胞的表达谱的每个第三聚类与合并聚类集中的第四聚类合并，前提如果该第三聚类和该第四聚类之间的距离在聚类距离阈值内的话。该方法包括基于代表细胞表达谱的合并聚类集中的聚类对多个细胞进行分类。该方法可以包括基于所鉴定的用于区分细胞类型的靶设计全转录组测定，或基于所鉴定的用于区分细胞类型的靶设计靶向性转录组测定。

[0284] 在一些实施例中，方法600包括在遍历树状图的多个聚类时该方法在每个聚类处包括：(3)如果该聚类代表单细胞的表达谱，则将该聚类添加到合并聚类集。在遍历树状图的多个聚类时该方法在每个聚类处包括：为该聚类分配聚类标记。在一些实施例中，如果该聚类代表单细胞的表达谱，则该聚类的聚类标记包括单细胞标识，否则如果该聚类是母聚类的左子聚类，则该聚类的聚类标记包括母聚类的聚类标记和左标识，否则该聚类的聚类标记包括母聚类的聚类标记和右标识。

[0285] 在方框620处，方法600可包括基于由合并聚类集中的每个聚类代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶。方法600可以参照图4，例如图4的方框424所描述的来鉴定用于区分细胞类型的靶。在一些实施例中，对于合并聚类集中的每个聚类，基于由该聚类代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶包括：确定由该聚类代表的表达谱和由合并聚类集中的另一聚类代表的表达谱之间在与用于区分细胞类型的靶相关联的具有不同序列的分子标记的数目方面的差大于显著性阈值。

[0286] 测序

[0287] 在一些实施例中，估算不同的经条形码化的靶(例如经随机条形码化的靶)的数目可以包括确定经标记的靶、空间标记、分子标记、样品标记、细胞标记、或其任何产物(例如经标记的扩增子、或经标记的cDNA分子)的序列。扩增的靶可以经受测序。确定经条形码化的靶(例如经随机条形码化的靶)或其任何产物的序列可以包括进行测序反应以确定样品标记、空间标记、细胞标记、分子标记的至少一部分的序列，经标记的靶(例如经随机标记的靶)、其互补体、其反向互补体的至少一部分的序列，或其任何组合的序列。

[0288] 可以使用多种测序方法确定经条形码化的靶或经随机条形码化的靶的序列(例如，扩增的核酸、标记的核酸、标记的核酸的cDNA拷贝等)，这些方法包括但不限于杂交测序(SBH)、连接法测序(SBL)、量化增量荧光核苷酸附加测序(quantitative incremental fluorescent核苷酸addition sequencing)(QIFNAS)、分段连接与断裂、荧光共振能量转移(FRET)、分子信标、TaqMan报告探针消化、焦磷酸测序、荧光原位测序(FISSEQ)、FISSEQ珠、摆动测序(wobble sequencing)、多重测序、聚合集群(polymerized colony)(POLONY)测序；纳米格滚环测序(nanogrid rolling circle sequencing，ROLONY)、等位基因特异性寡核苷酸连接检验(allele-specific oligo ligation assay)(例如，寡核苷酸连接检验(OLA)、使用连接的线性探针和滚环扩增(RCA)读出、连接的持锁探针的单模板分子(single template molecule)OLA、或使用连接的环形持锁探针和滚环扩增(RCA)读出的单模板分子OLA)等。

[0289] 在一些实施例中，确定经条形码化的靶(例如经随机条形码化的靶)或其任何产物的序列包括配对端测序、纳米孔测序、高通量测序、鸟枪法测序、染料终止剂测序、多重引物DNA测序、引物步移、桑格双脱氧测序法、马克西姆-吉尔伯特(Maxim-Gilbert)测序、焦磷酸测序、真正的单分子测序或其任何组合。可替代地，可以通过电子显微镜分析法或化学敏感场效应晶体管(chemFET)阵列来确定经条形码化的靶或其任何产物的序列。

[0290] 可以使用高通量测序方法，如使用平台(如Roche 454、Illumina Solexa、ABI-SOLiD、ION Torrent、Complete Genomics、Pacific Bioscience、Helicos、或Polonator平台)的循环阵列测序。在一些实施例中，测序可以包含MiSeq测序。在一些实施例中，测序可以包含HiSeq测序。

[0291] 经标记的靶(例如经随机标记的靶)可以包括代表来自生物体基因组基因的约0.01％至生物体基因组基因的约100％的核酸。例如，可以使用包括多个多聚体的靶标互补区域，通过从所述样品中捕获含有互补序列的基因，对约0.01％的生物体基因组基因至约100％的生物体基因组基因进行测序。在一些实施例中，经条形码化的靶包括代表从约0.01％的生物体转录组转录物至约100％的生物体转录组转录物的核酸。例如，可以使用包括聚(T)尾的靶标互补区域，通过从所述样品中捕获mRNA，对约0.501％的生物体转录组转录物至约100％的生物体转录组转录物进行测序。

[0292] 确定多个条形码(例如随机条形码)的空间标记和分子标记的序列可包括对多个条形码的0.00001％、0.0001％、0.001％、0.01％、0.1％、1％、2％、3％、4％、5％、6％、7％、8％、9％、10％、20％、30％、40％、50％、60％、70％、80％、90％、99％、100％、或在这些值的任何两个之间的数字或范围进行测序。确定多个条形码的标记(例如样品标记、空间标记、和分子标记)的序列可包括对多个条形码中的1、10、20、30、40、50、60、70、80、90、100、10^3、10^4、10^5、10^6、10^7、10^8、10^9、10^{10}、10^{11}、10^{12}、10^{13}、10^{14}、10^{15}、10^{16}、10^{17}、10^{18}、10^{19}、10^{20}个、或在这些值的任何两个之间的数字或范围进行测序。对多个条形码中的一些或所有进行测序可包括产生具有、具有约、具有至少、或具有至多10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000个、或在这些值的任何两个之间的数字或范围内的核苷酸或碱基的读数长度的序列。

[0293] 测序可包括对经条形码化的靶的至少或至少约10、20、30、40、50、60、70、80、90、100个或更多个核苷酸或碱基对进行测序。例如，测序可包括通过对多个经条形码化的靶进行聚合酶链式反应(PCR)扩增，产生测序数据，其中序列具有50、75、或100个、或更多个核苷酸的读数长度。测序可包括对经条形码化的靶的至少或至少约200、300、400、500、600、700、800、900、1,000个或更多个核苷酸或碱基对进行测序。测序可包括对经条形码化的靶的至少或至少约1500、2000、3000、4000、5000、6000、7000、8000、9000、或10000个或更多个核苷酸或碱基对进行测序。

[0294] 测序可以包括至少约200、300、400、500、600、700、800、900、1,000个或更多个测序读数/运行。在一些实施例中，测序包括每次运行对至少或至少约1500、2000、3000、4000、5000、6000、7000、8000、9000、或10000个或更多个测序读数进行测序。测序可以包括小于或等于约1,600,000,000个测序读数/运行。测序可以包括小于或等于约200,000,000个读数/运行。

[0295] 样品

[0296] 在一些实施例中，多个靶可以包含在一个或多个样品中。样品可包含一个或多个细胞、或者来自一个或多个细胞的核酸。样品可以是单细胞或来自单细胞的核酸。一个或多个细胞可以是一种或多种细胞类型。一种或多种细胞类型中的至少一种可以是脑细胞、心脏细胞、癌细胞、循环肿瘤细胞、器官细胞、上皮细胞、转移性细胞、良性细胞、原代细胞、循环细胞或其任何组合。

[0297] 用于在本披露的方法中使用的样品可以包括一个或多个细胞。样品可以指一个或多个细胞。在一些实施例中，多个细胞可包括一种或多种细胞类型。一种或多种细胞类型中的至少一种可以是脑细胞、心脏细胞、癌细胞、循环肿瘤细胞、器官细胞、上皮细胞、转移性细胞、良性细胞、原代细胞、循环细胞或其任何组合。在一些实施例中，该细胞是从癌组织切除的癌细胞，例如乳腺癌、肺癌、结肠癌、前列腺癌、卵巢癌、胰腺癌、脑癌、黑素瘤和非黑素瘤皮肤癌等。在一些实施例中，该细胞源自癌症，但是从体液收集(例如，循环肿瘤细胞)。癌症的非限制性实例可以包括腺瘤、腺癌、鳞状细胞癌、基底细胞癌、小细胞癌、大细胞未分化癌、软骨肉瘤、以及纤维肉瘤。样品可包括组织、单层细胞、固定的细胞、组织切片、或其任何组合。样品可包括生物学样品、临床样品、环境样品、生物学流体、来自受试者的组织或细胞。样品可以获得自人、哺乳动物、狗、大鼠、小鼠、鱼、蝇、蠕虫、植物、真菌、细菌、病毒、脊椎动物、或无脊椎动物。

[0298] 在一些实施例中，该细胞是已经被病毒感染并含有病毒寡核苷酸的细胞。在一些实施例中，病毒感染可以由如单链(+链或“有义”)DNA病毒(例如，细小病毒)、或双链RNA病毒(例如，呼吸道肠道病毒)的病毒引起。在一些实施例中，所述细胞是细菌。这些可以包括革兰氏阳性细菌或革兰氏阴性细菌。在一些实施例中，所述细胞是真菌。在一些实施例中，所述细胞是原生动物或其他寄生虫。

[0299] 如本文使用的，术语“细胞”可以指一个或多个细胞。在一些实施例中，细胞是正常细胞，例如，处于不同发育阶段的人细胞，或来自不同器官或组织类型的人细胞。在一些实施例中，该细胞是非人类细胞，例如其他类型的哺乳动物细胞(例如小鼠、大鼠、猪、狗、牛或马)。在一些实施例中，所述细胞是其他类型的动物或植物细胞。在其他实施例中，细胞可以是任何原核或真核细胞。

[0300] 在一些实施例中，在将细胞与珠关联之前对所述细胞进行分选。例如，所述细胞可以通过荧光活化细胞分选或磁活化细胞分选进行分选，或更普遍地通过流式细胞术进行分选。可以按尺寸过滤细胞。在一些实施例中，滞留物包含有待与珠相关联的细胞。在一些实施例中，流过物(flow through)包含有待与珠相关联的细胞。

[0301] 样品可以指多个细胞。样品可以指单层细胞。样品可以指薄切片(例如，组织薄切片)。样品可以指细胞的固体或半固体集合，其可以放置在阵列上的一个维度上。

[0302] 执行环境

[0303] 本公开提供了程序化设定为执行本公开方法(例如方法400或方法600)的计算机系统。图7显示了计算机系统700，其程序化设定为或另外配置为执行本文中公开的任何方法。计算机系统700可以是用户的电子设备或相对于电子设备远程定位的计算机系统。电子设备可以是移动电子设备。

[0304] 计算机系统700包括中央处理单元(CPU，本文也称为“处理器”和“计算机处理器”)705，其可以是单核或多核处理器，或者是用于并行处理的多个处理器。计算机系统700还包

括存储器或存储器位置710(例如,随机存取存储器、只读存储器、闪存),电子存储单元715(例如,硬盘),用于与一个或多个其他系统通信的通信接口720(例如,网络适配器),以及外围设备725,例如高速缓存、其他存储器、数据存储和/或电子显示适配器。存储器710,存储单元715,接口720和外围设备725通过通信总线(实线)如母板与CPU 705通信。存储单元715可以是用于存储数据的数据存储单元(或数据存储库)。计算机系统700可以借助于通信接口720可操作地耦合到计算机网络(“网络”)730。网络730可以是因特网、因特网和/或外联网、或者与因特网通信的内联网和/或外联网。在某些情况下,网络730是电信和/或数据网络。网络730可以包括一个或多个计算机服务器,其可以实现分布式计算,例如云计算。在一些情况下借助于计算机系统700,网络730可以实现对等网络(peer-to-peer network),其可以使耦合到计算机系统700的设备能够充当客户端或服务器。

[0305] CPU 705可以执行一系列机器可读指令,其可以具体化在程序或软件中。指令可以存储在存储器位置中,例如存储器710中。可以将指令引导到CPU 705,CPU 705随后可以程序化设定为或另外配置CPU705以执行本公开的方法。由CPU 705执行的操作的例子可以包括获取,解码,执行和回写。CPU 705可以是电路例如集成电路的一部分。系统700的一个或多个其他组件可以包括在电路中。在一些情况下,该电路是专用集成电路(ASIC)。

[0306] 存储单元715可以存储文件,例如驱动程序、库和保存的程序。存储单元715可以存储用户数据,例如用户偏好和用户程序。在一些情况下,计算机系统700可以包括在计算机系统700外部的一个或多个附加数据存储单元,例如位于通过内联网或因特网与计算机系统700通信的远程服务器上。

[0307] 计算机系统700可以通过网络730与一个或多个远程计算机系统通信。例如,计算机系统700可以与用户(例如,微生物学家)的远程计算机系统通信。远程计算机系统的例子包括个人计算机(例如便携式PC),触屏电脑或平板电脑(例如,Apple®iPad、Samsung® GalaxyTab),电话,智能电话(例如,Apple®iPhone、支持Android的设备、Blackberry®)),或个人数字助理。用户可以经由网络730访问计算机系统700。

[0308] 计算机系统700可包括电子显示器735或与电子显示器735通信,电子显示器735包括用户界面(UI)740,用于提供例如指示多个微生物类群的串共现(string co-occurrence)或相互作用的输出,如字符串表示的。UI的例子包括但不限于图形用户界面(GUI)和基于web的用户界面。

[0309] 本文中描述的方法可以通过存储在计算机系统700的电子存储位置上,如例如在存储器710或电子存储单元715上的机器(例如,计算机处理器)可执行代码来实现。机器可执行代码或机器可读代码可以以软件的形式提供。在使用期间,代码可以由处理器705执行。在一些情况下,可以从存储单元715取回代码并将其存储在存储器710上以供处理器705随时访问。在一些情况下,可以排除电子存储单元715,并且机器可执行指令存储在存储器710上。

[0310] 可以预编译和配置代码以用于具有适于执行代码的处理器的机器,或者可以在运行时期间编译。代码可以以编程语言提供,可以选择该编程语言使代码能够以预编译或如编译的方式执行。

[0311] 本文提供的系统和方法的各方面,例如计算机系统700,可以具体化在编程中。该技术的各个方面可以被认为是通常以一种类型的机器可读介质中承载或具体化的机器(或

处理器)可执行代码和/或相关数据的形式的“产品”或“制品”。机器可执行代码可以存储在电子存储单元,例如存储器(例如,只读存储器,随机存取存储器,闪存)或硬盘上。“存储”型介质可以包括计算机的有形存储器、处理器等、或其相关模块,例如各种半导体存储器、磁带驱动器、磁盘驱动器等的任何一个或全部,其可以随时提供非暂时性存储用于软件编程。软件的全部或部分有时可以通过因特网或各种其他电信网络进行通信。例如,这类通信可以使软件从一个计算机或处理器加载到另一个计算机或处理器,例如,从管理服务器或主计算机加载到应用服务器的计算机平台。因此,可以承载软件元件的另一种类型的介质包括光波、电波和电磁波,例如跨本地设备之间的物理接口,通过有线和光学陆线网络以及通过各种空中链路使用的。携带这类波的物理元件,例如有线或无线链路、光链路等,也可以被认为是承载软件的介质。如本文所使用的,除非限定为非暂时性的有形“存储”介质,诸如计算机或机器“可读介质”之类的术语是指参与向处理器提供指令以供执行的任何介质。

[0312] 因此,机器可读介质如计算机可执行代码可以采用许多形式,包括但不限于,有形存储介质、载波介质或物理传输介质。非易失性存储介质包括例如光盘或磁盘,如任何计算机等中的任何存储设备,如可用于实现附图中所示的数据库等。易失性存储介质包括动态存储器,例如这类计算机平台的主存储器。有形传输介质包括同轴电缆;铜线和光纤,包括在计算机系统内构成总线的线。载波传输介质可以采用电或电磁信号,或声波或光波的形式,例如在射频(RF)和红外(IR)数据通信期间产生的那些。因此,计算机可读介质的常见形式包括,例如:软盘(floppy disk),软磁盘(flexible disk),硬盘,磁带,任何其他磁介质,CD-ROM,DVD或DVD-ROM,任何其他光学介质,穿孔卡纸带,任何其他带孔图案的物理存储介质,RAM,ROM,PROM和EPROM,FLASH-EPROM,任何其他存储器芯片或盒式磁带,传输数据或指令的载波,传输此类载波的电缆或链路,或计算机可从中读取编程代码和/或数据的任何其他介质。许多这些形式的计算机可读介质可涉及将一个或多个指令的一个或多个序列传送到处理器以供执行。

[0313] 在一些实施例中,计算机系统700的一些或全部分析功能性可包装在单个软件包内。在一些实施例中,完整的数据分析能力集合可以包括一套软件包。在一些实施例中,数据分析软件可以是独立于测定仪器系统使用户可用的独立包。在一些实施例中,软件可以是基于网络的,并且可以允许用户共享数据。在一些实施例中,可商购的软件可用于执行数据分析的全部或一部分,例如,可以使用七桥(Seven Bridges,https://www.sbgenomics.com/)软件来编译在整个细胞集合中每个细胞中出现的一个或多个基因的拷贝数目表。

[0314] 可以通过一种或多种算法或方法来执行本公开的方法和系统。可以在由中央处理单元705执行时通过软件实现方法。通过软件执行的算法或方法的示例性应用包括,用于序列读取处理(例如,合并、过滤、修剪、聚类),对齐和调用,以及串数据和光密度数据的处理(例如,最可能的数目和可培养的丰度测定)的生物信息学方法。

[0315] 在示例性实施例中,计算机系统700可以在通过实施单细胞、随机条形码化测定而生成的序列数据集上执行数据分析。数据分析功能的实例包括但不限于(i)用于解码/多路解编通过对在运行测定中产生的随机条形码文库进行测序所提供的样品标记、细胞标记、空间标记和分子标记以及靶序列数据的算法,(ii)用于基于数据确定每个细胞每个基因的读段数目和每个细胞每个基因的独特转录物分子数目、并且创建汇总表的算法,(iii)对序

列数据的统计分析，例如用于通过基因表达数据对细胞聚类，或用于预测确定每个细胞每个基因的转录物分子数目的置信区间等，(iv)用于鉴定罕见细胞亚群的算法，例如使用主成分分析、分层聚类、k均值聚类、自组织映射、神经网络等，(v)基因序列数据与已知参考序列比对以及突变、多态性标记物和剪接变体的检测的序列比对能力，以及(vi)分子标记的自动聚类，以补偿扩增或测序错误。在一些实施例中，计算机系统700可以以有用的图形格式输出测序结果，例如，指示在细胞集合的每个细胞中存在的一个或多个基因的拷贝数的热图。在一些实施例中，计算机系统700可以执行用于从测序结果提取生物学意义的算法，例如通过将在细胞集合的每个细胞中存在的一个或多个基因的拷贝数与一种类型的细胞、一种类型的稀有细胞、或来源于具有特定疾病或病症的受试者的细胞相关联。在一些实施例中，计算机系统700可以执行用于比较跨不同生物样品的细胞群的算法。

[0316] 实例

[0317] 以上讨论的实施例的一些方面在以下实例中进一步详细披露，其不是旨在以任何方式限制本披露的范围。

[0318] 实例1

[0319] 通过递归树状图分裂和测试然后合并进行聚类

[0320] 本实例描述了通过递归分裂(例如，递归树状图分裂)和测试然后合并的聚类方法。

[0321] 注意

[0322] 在本实例所示的方法中，在树状图分裂步骤期间，如果算法可以找到至少一个达到足够低的p值(或足够高的-log10(p值))的基因，则分裂被认为(例如默认)是生物学相关联的。换言之，在一些实施例中，要细调的唯一超参数是得分阈值参数。较高的得分阈值(例如100)对应于较低的p值(10e-100)，这意味着必须找到更重要的基因以使分裂被认为是有效的。得分阈值越高，聚类的数目越少。

[0323] 如果在分裂步骤之后生成太多聚类，则用户可以尝试增加得分阈值。如果分裂步骤后生成的聚类太少，则用户可以尝试降低得分阈值。可以在同一距离矩阵上测试多个得分阈值。通过预先计算距离矩阵，可以节省大量的计算时间。

[0324] 如果扫描扫过不同的得分阈值持续产生无意义的结果，则问题可能在首先产生的树状图(即距离矩阵)上。如下面的单元[3]所示，算法的第一步需要从分子计数的矩阵运行到距离矩阵(预处理步骤)。可能需要尝试不同类型的预处理。用户也许可以尝试另一种距离量度，尝试不取对数，或预过滤细胞和/或基因，这可能会为您的应用生成更准确的距离量度。

[0325] 如果分裂步骤产生许多看起来不相关的小聚类，则可以降低解散(disband)百分位参数。该参数基于最后聚类的成对距离的多少位于总体成对距离的底部解散百分位内来决定是否保留该最后聚类。例如，运行具有解散百分位20的算法将仅在至少一个成对距离位于总体距离的底部20％内的情况下保留聚类。

[0326] 为了确定将聚类分成两个聚类的原因，鉴定那两个聚类并进行成对t检验分析。这可以在实例2中的单元[13]中对每对聚类完成。此功能将显示区分两个聚类的标志。另外，参阅如何决定分裂的探索功能以查看导致分裂的分裂算法中的确切步骤。

[0327] 要确定是否将不正确的聚类合并在一起，在合并步骤中降低得分阈值参数。此处

的得分阈值越大,则两个不同的聚类合并在一起的可能性就越大。另外,参阅如何决定合并的探索功能以查看导致合并的合并算法中的确切步骤。

[0328] 为了鉴定更多的异常者,尝试降低合并步骤中的异常者_阈值_百分位_参数(outlier_threshold_percentile_parameter)。

[0329] 依赖关系

[0330] 模块具有以下依赖关系:-numpy(1.10.4)-scipy(0.17.0)-matplotlib(1.5.1)-sklearn(0.17.1)-networkx(1.11)-community-rpy2(2.8.2)

[0331] networkx,community和rpy2默认不需要。networkx和community用于社群检测。networkx还用于最大化权重匹配(maxed-weight-matching)(作为两组标记如何接近的量度)。rpy2用于运行sigclust,针对两个群体是否实际上应该是一个群体的统计学检验。为了运行sigclust,用户可能还需要在安装sigclust包的同时安装R。

[0332] 在[1]中:#加载相关的模块和库

[0333] %load_ext autoreload

[0334] %autoreload 2

[0335] %matplotlib inline

[0336] from dendrosplit import split,merge import pickle

[0337] import numpy as np

[0338] import matplotlib.pyplot as plt np.set_printoptions(precision＝2,suppress＝True)

[0339] 运行流水线

[0340] 对流水线的输入是分子计数(自然数)的N×M矩阵称为'X'。'基因'是基因名称的长度M的列表。'x1'和'x2'代表数据的二维嵌入,其使用用户选择的任何方法。'x1'和'x2'仅用于使流水线以及中间步骤的结果可视化。该算法要求删除总和为0的所有'X'的列,并且由此代码单元负责处理。

[0341] 在[2]中:#加载数据

[0342] dataset＝'Resolve4'

[0343] pickledir＝'/Users/user1/Desktop/datasets/'

[0344] X,genes＝pickle.load(file(pickledir+dataset+'.pickle'))

[0345] x1,x2＝pickle.load(file(pickledir+dataset+'tsne.pickle'))

[0346] #删除总和为0的'X'的列

[0347] X,genes＝split.filter_genes(X,genes)

[0348] 对于跨所有单元具有>0计数,保持19307个基因

[0349] 可以首先从计数矩阵生成距离矩阵。下面的单元通过计算经对数转换的样本(log(X+1))之间的成对相关距离来实现这一点。虽然用户可以输入距离矩阵以及如下所示,但算法的分裂部分仅需要计数矩阵作为输入。算法的这部分为样本返回了一组长度为N的标记(字符串)以及'历史(history)',这是一种跟踪算法生成的所有中间信息的数据结构。'历史'对于后面用于剖析算法如何生成此类标记(以及哪些特征对于生成此类标记最重要)的函数非常有用。标记是根据使用距离矩阵生成的树状图指示聚类位于何处的字符串。例如,'rLLR'表示此点属于根的左子树的左子树的右子树。

[0350] 在[3]中:#获得第一组标记高度推荐在算法外计算距离矩阵

[0351] D＝split.log_correlation(X)

[0352] ys,shistory＝split.dendrosplit((D,X),preprocessing＝'precomputed',score_threshold＝10,verbose＝True,disband_percentile＝50)

[0353] 可能的分裂结果:883和3

[0354] dendrosplit/feature selection.py:106:RuntimeWarning:divide by zero encountered in log10

[0355] gene_scores＝np.nan_to_num(-np.log10(p[keep_inds]))

[0356] 分裂得分1.8E+308

[0357] 可能的分裂结果:1和882

[0358] 可能的分裂结果:484和398

[0359] /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/

[0360] _distn_infrastructure.py:1748:Runtime

[0361] cond1＝(scale>0)&(x>self.a)&(x<self.b)

[0362] /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/_distn_infrastructure.py:1748:Runtime

[0363] cond1＝(scale>0)&(x>self.a)&(x<self.b)

[0364] /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/_distn_infrastructure.py:1749:Runtime

[0365] cond2＝cond0&(x<＝self.a)

[0366] 分裂得分182.26

[0367] 可能的分裂结果:481和3分裂得分1.8E+308

[0368] 可能的分裂结果:1和480

[0369] 可能的分裂结果:1和479

[0370] 可能的分裂结果:195和284

[0371] 分裂得分125.49

[0372] 可能的分裂结果:177和18

[0373] 分裂得分15.35

[0374] 可能的分裂结果:1和176

[0375] 可能的分裂结果:1和175

[0376] 可能的分裂结果:1和174

[0377] 可能的分裂结果:12和162

[0378] 分裂得分18.88

[0379] 可能的分裂结果:1和11

[0380] 可能的分裂结果:1和10

[0381] 可能的分裂结果:2和8

[0382] 分裂得分6.11

[0383] 可能的分裂结果:1和161

[0384] 可能的分裂结果:1和160

[0385] 可能的分裂结果:28和132

[0386] 分裂得分12.32

[0387] 可能的分裂结果:25和3

[0388] 分裂得分13.94

[0389] 可能的分裂结果:1和24

[0390] 可能的分裂结果:11和13

[0391] 分裂得分4.77

[0392] 可能的分裂结果:1和2

[0393] 可能的分裂结果:1和1

[0394] 解散(聚类中的点彼此太远)可能的分裂结果:122和10

[0395] 分裂得分18.52

[0396] 可能的分裂结果:13和109

[0397] 分裂得分24.92

[0398] 可能的分裂结果:6和7

[0399] 分裂得分3.77

[0400] 可能的分裂结果:105和4

[0401] 分裂得分31.72

[0402] 可能的分裂结果:15和90

[0403] 分裂得分11.31

[0404] 可能的分裂结果:3和12

[0405] 分裂得分6.55

[0406] 可能的分裂结果:17和73

[0407] 分裂得分8.91

[0408] 可能的分裂结果:2和2

[0409] 分裂得分1.58

[0410] 解散(聚类中的点彼此太远)可能的分裂结果:2和8

[0411] 分裂得分5.79

[0412] 可能的分裂结果:1和17

[0413] 可能的分裂结果:1和16

[0414] 可能的分裂结果:1和15

[0415] 可能的分裂结果:4和11

[0416] 分裂得分4.57

[0417] 可能的分裂结果:1和283

[0418] 可能的分裂结果:1和282

[0419] 可能的分裂结果:1和281

[0420] 可能的分裂结果:271和10

[0421] 分裂得分38.04

[0422] 可能的分裂结果:2和269

[0423] 分裂得分233.23

[0424] 可能的分裂结果:1和1

[0425] 解散(聚类中的点彼此太远)可能的分裂结果:1和268

[0426] 可能的分裂结果:265和3

[0427] 分裂得分80.24

[0428] 可能的分裂结果:4和261

[0429] 分裂得分100.26

[0430] 可能的分裂结果:1和3

[0431] 可能的分裂结果:1和2

[0432] 可能的分裂结果:1和1

[0433] 解散(聚类中的点彼此太远)可能的分裂结果:192和69

[0434] 分裂得分9.66

[0435] 可能的分裂结果:1和2

[0436] 可能的分裂结果:1和1

[0437] 解散(聚类中的点彼此太远)可能的分裂结果:2和8

[0438] 分裂得分5.12

[0439] 可能的分裂结果:1和2

[0440] 可能的分裂结果:1和1

[0441] 解散(聚类中的点彼此太远)可能的分裂结果:1和397

[0442] 可能的分裂结果:1和396

[0443] 可能的分裂结果:1和395

[0444] 可能的分裂结果:392和3

[0445] 分裂得分228.58

[0446] 可能的分裂结果:1和391

[0447] 可能的分裂结果:1和390

[0448] 可能的分裂结果:1和389

[0449] 可能的分裂结果:1和388

[0450] 可能的分裂结果:1和387

[0451] 可能的分裂结果:1和386

[0452] 可能的分裂结果:32和354

[0453] 分裂得分33.24

[0454] 可能的分裂结果:1和31

[0455] 可能的分裂结果:1和30

[0456] 可能的分裂结果:21和9

[0457] 分裂得分7.20

[0458] 可能的分裂结果:1和353

[0459] 可能的分裂结果:1和352

[0460] 可能的分裂结果:1和351

[0461] 可能的分裂结果:19和332

[0462] 分裂得分32.86

[0463] 可能的分裂结果:1和18

[0464] 可能的分裂结果:3和15

[0465] 分裂得分8.90

[0466] 可能的分裂结果:6和326

[0467] 分裂得分83.57

[0468] 可能的分裂结果:1和5

[0469] 可能的分裂结果:1和4

[0470] 可能的分裂结果:2和2

[0471] 分裂得分1.32

[0472] 解散(聚类中的点彼此太远)可能的分裂结果:3和323

[0473] 分裂得分148.25

[0474] 可能的分裂结果:1和2

[0475] 可能的分裂结果:1和1

[0476] 解散(聚类中的点彼此太远)可能的分裂结果:314和9

[0477] 分裂得分71.43

[0478] 可能的分裂结果:221和93

[0479] 分裂得分48.70

[0480] 可能的分裂结果:1和220

[0481] 可能的分裂结果:1和219

[0482] 可能的分裂结果:1和218

[0483] 可能的分裂结果:1和217

[0484] 可能的分裂结果:215和2

[0485] 分裂得分133.42

[0486] 可能的分裂结果:166和49

[0487] 分裂得分7.64

[0488] 可能的分裂结果:1和1

[0489] 可能的分裂结果:40和53

[0490] 分裂得分9.31

[0491] 可能的分裂结果:4和5

[0492] 分裂得分3.20

[0493] 解散(聚类中的点彼此太远)可能的分裂结果:1和2

[0494] 可能的分裂结果:1和1

[0495] 解散(聚类中的点彼此太远)可能的分裂结果:1和2

[0496] 可能的分裂结果:1和1

[0497] 解散(聚类中的点彼此太远)

[0498] #调用得分函数的次数40

[0499] 总计算时间为9.532s

[0500] 合并步骤涉及对上面的分裂过程生成的所有聚类进行成对比较。从两个最相似的聚类开始,将差异不够大的聚类合并在一起。与分裂步骤一样,合并步骤返回标记(长度N)

和中间步骤的历史两者。标记是整数。异常者标记为“-1”。请参阅下文，了解基于社群检测的合并步骤的备选方法。

[0501] 在[4]中:#合并聚类标记

[0502] ym,mhistory＝merge.dendromerge((D,X),ys,score_threshold＝10,

[0503] preprocessing＝’precomputed’,verbose＝True,outlier_threshold_percentile＝90)

[0504] 886个样本中0个是单例

[0505] 异常者阈值是0.51

[0506] 821的最近邻居:72,聚类76中(D＝0.375)

[0507] 661的最近邻居:29,聚类76中(D＝0.379)

[0508] 729的最近邻居:281,聚类76中(D＝0.381)

[0509] 559的最近邻居:79,聚类76中(D＝0.381)

[0510] 690的最近邻居:171,聚类76中(D＝0.381)

[0511] 564的最近邻居:79,聚类76中(D＝0.381)

[0512] 776的最近邻居:474,聚类38中(D＝0.387)

[0513] 860的最近邻居:340,聚类38中(D＝0.390)

[0514] 816的最近邻居:379,聚类78中(D＝0.390)

[0515] 787的最近邻居:63,聚类38中(D＝0.391)

[0516] 737的最近邻居:72,聚类76中(D＝0.392)

[0517] 874的最近邻居:220,聚类76中(D＝0.392)

[0518] 743的最近邻居:72,聚类76中(D＝0.394)

[0519] 877的最近邻居:174,聚类76中(D＝0.394)

[0520] 753的最近邻居:190,聚类76中(D＝0.397)

[0521] 774的最近邻居:158,聚类38中(D＝0.398)

[0522] 565的最近邻居:190,聚类76中(D＝0.399)

[0523] 785的最近邻居:79,聚类76中(D＝0.401)

[0524] 706的最近邻居:101,聚类18中(D＝0.403)

[0525] 829的最近邻居:213,聚类38中(D＝0.404)

[0526] 701的最近邻居:179,聚类76中(D＝0.404)

[0527] 770的最近邻居:453,聚类38中(D＝0.404)

[0528] 630的最近邻居:79,聚类76中(D＝0.406)

[0529] 866的最近邻居:87,聚类38中(D＝0.407)

[0530] 795的最近邻居:159,聚类76中(D＝0.407)

[0531] 865的最近邻居:179,聚类76中(D＝0.407)

[0532] 869的最近邻居:101,聚类18中(D＝0.409)

[0533] 830的最近邻居:165,聚类38中(D＝0.412)

[0534] 851的最近邻居:29,聚类76中(D＝0.412)

[0535] 782的最近邻居:50,聚类76中(D＝0.412)

[0536] 627的最近邻居:72,聚类76中(D＝0.412)

[0537] 848的最近邻居:83,聚类76中(D＝0.413)

[0538] 883的最近邻居:687,聚类12中(D＝0.413)

[0539] 793的最近邻居:107,聚类76中(D＝0.414)

[0540] 631的最近邻居:101,聚类18中(D＝0.416)

[0541] 720的最近邻居:101,聚类18中(D＝0.418)

[0542] 885的最近邻居:101,聚类18中(D＝0.418)

[0543] 813的最近邻居:101,聚类18中(D＝0.419)

[0544] 788的最近邻居:278,聚类38中(D＝0.420)

[0545] 748的最近邻居:101,聚类18中(D＝0.422)

[0546] 762的最近邻居:158,聚类38中(D＝0.423)

[0547] 804的最近邻居:177,聚类18中(D＝0.425)

[0548] 854的最近邻居:101,聚类18中(D＝0.426)

[0549] 605的最近邻居:159,聚类76中(D＝0.437)

[0550] 849的最近邻居:101,聚类18中(D＝0.437)

[0551] 835的最近邻居:101,聚类18中(D＝0.438)

[0552] 790的最近邻居:32,聚类76中(D＝0.442)

[0553] 744的最近邻居:188,聚类38中(D＝0.448)

[0554] 822的最近邻居:282,聚类38中(D＝0.449)

[0555] 723的最近邻居:170,聚类76中(D＝0.456)

[0556] 884的最近邻居:101,聚类18中(D＝0.459)

[0557] 563的最近邻居:34,聚类76中(D＝0.463)

[0558] 867的最近邻居:160,聚类18中(D＝0.463)

[0559] 771的最近邻居:34,聚类76中(D＝0.473)

[0560] 826的最近邻居:165,聚类38中(D＝0.475)

[0561] 777的最近邻居:174,聚类76中(D＝0.478)

[0562] 759的最近邻居:101,聚类18中(D＝0.483)

[0563] 855的最近邻居:101,聚类18中(D＝0.485)

[0564] 702的最近邻居:160,聚类18中(D＝0.492)

[0565] 750的最近邻居:230,聚类76中(D＝0.495)

[0566] 704的最近邻居:216,聚类78中(D＝0.497)

[0567] 711的最近邻居:55,聚类76中(D＝0.502)

[0568] 708的最近邻居:537,聚类78中(D＝0.510)

[0569] 791的最近邻居:115,聚类76中(D＝0.534)

[0570] 722的最近邻居:15,聚类76中(D＝0.547)

[0571] 700的最近邻居:107,聚类76中(D＝0.549)

[0572] 846的最近邻居:72,聚类76中(D＝0.552)

[0573] 876的最近邻居:85,聚类76中(D＝0.560)

[0574] 868的最近邻居:740,聚类78中(D＝0.562)

[0575] 569的最近邻居:68,聚类76中(D＝0.572)

[0576] 817的最近邻居:56,聚类76中(D＝0.582)

[0577] 798的最近邻居:310,聚类38中(D＝0.585)

[0578] 717的最近邻居:216,聚类78中(D＝0.597)

[0579] 879的最近邻居:209,聚类76中(D＝0.612)

[0580] 727的最近邻居:96,聚类76中(D＝0.616)

[0581] 828的最近邻居:142,聚类38中(D＝0.618)

[0582] 840的最近邻居:632,聚类78中(D＝0.640)

[0583] 747的最近邻居:202,聚类76中(D＝0.698)

[0584] 842的最近邻居:797,聚类38中(D＝0.703)

[0585] 442的最近邻居:336,聚类78中(D＝0.735)

[0586] 异常者的总数目:18

[0587] 分配的单例(0.052s)

[0588] 产生的Dc(13.181s)

[0589] 合并前:14个聚类

[0590] 合并距离3.60的标记0(N＝10)和6(N＝15)

[0591] 合并前:13个聚类

[0592] 合并距离4.31的标记2(N＝15)和4(N＝10)合并前:12个聚类

[0593] 合并距离4.37的标记1(N＝13)和11(N＝25)合并前:11个聚类

[0594] 合并距离5.23的标记0(N＝25)和10(N＝38)合并前:10个聚类

[0595] 合并距离6.04的标记3(N＝30)和7(N＝95)合并前:9个聚类

[0596] 合并距离6.81的标记2(N＝10)和5(N＝2)合并前:8个聚类

[0597] 合并距离7.19的标记4(N＝25)和5(N＝63)合并前:7个聚类

[0598] 合并距离7.23的标记2(N＝18)和5(N＝12)合并前:6个聚类

[0599] 合并距离9.76的标记3(N＝125)和5(N＝30)合并聚类用时25.977s

[0600] 基于社群检测的合并使用了networkx和社群python模块。这里返回的历史数据结构仅含有输入标记和后单例处理(post-singleton-processing)标记。

[0601] 在[5]中:ym_community＝merge.dendromerge((D,X),ys,preprocessing＝'precomputed',verbose＝True,outlier_threshold_percentile＝90,perform_community_detection＝True)

[0602] 886个样本中80个是单例异常者阈值是0.51

[0603] 821的最近邻居:72,聚类76中(D＝0.375)

[0604] 661的最近邻居:29,聚类76中(D＝0.379)

[0605] 729的最近邻居:281,聚类76中(D＝0.381)

[0606] 559的最近邻居:79,聚类76中(D＝0.381)

[0607] 690的最近邻居:171,聚类76中(D＝0.381)

[0608] 564的最近邻居:79,聚类76中(D＝0.381)

[0609] 776的最近邻居:474,聚类38中(D＝0.387)

[0610] 860的最近邻居:340,聚类38中(D＝0.390)

[0611] 816的最近邻居:379,聚类78中(D＝0.390)

[0612] 787的最近邻居:63,聚类38中(D＝0.391)

[0613] 737的最近邻居:72,聚类76中(D＝0.392)

[0614] 874的最近邻居:220,聚类76中(D＝0.392)

[0615] 743的最近邻居:72,聚类76中(D＝0.394)

[0616] 877的最近邻居:174,聚类76中(D＝0.394)

[0617] 753的最近邻居:190,聚类76中(D＝0.397)

[0618] 774的最近邻居:158,聚类38中(D＝0.398)

[0619] 565的最近邻居:190,聚类76中(D＝0.399)

[0620] 785的最近邻居:79,聚类76中(D＝0.401)

[0621] 706的最近邻居:101,聚类18中(D＝0.403)

[0622] 829的最近邻居:213,聚类38中(D＝0.404)

[0623] 701的最近邻居:179,聚类76中(D＝0.404)

[0624] 770的最近邻居:453,聚类38中(D＝0.404)

[0625] 630的最近邻居:79,聚类76中(D＝0.406)

[0626] 866的最近邻居:87,聚类38中(D＝0.407)

[0627] 795的最近邻居:159,聚类76中(D＝0.407)

[0628] 865的最近邻居:179,聚类76中(D＝0.407)

[0629] 869的最近邻居:101,聚类18中(D＝0.409)

[0630] 830的最近邻居:165,聚类38中(D＝0.412)

[0631] 851的最近邻居:29,聚类76中(D＝0.412)

[0632] 782的最近邻居:50,聚类76中(D＝0.412)

[0633] 627的最近邻居:72,聚类76中(D＝0.412)

[0634] 848的最近邻居:83,聚类76中(D＝0.413)

[0635] 883的最近邻居:687,聚类12中(D＝0.413)

[0636] 793的最近邻居:107,聚类76中(D＝0.414)

[0637] 631的最近邻居:101,聚类18中(D＝0.416)

[0638] 720的最近邻居:101,聚类18中(D＝0.418)

[0639] 885的最近邻居:101,聚类18中(D＝0.418)

[0640] 813的最近邻居:101,聚类18中(D＝0.419)

[0641] 788的最近邻居:278,聚类38中(D＝0.420)

[0642] 748的最近邻居:101,聚类18中(D＝0.422)

[0643] 762的最近邻居:158,聚类38中(D＝0.423)

[0644] 804的最近邻居:177,聚类18中(D＝0.425)

[0645] 854的最近邻居:101,聚类18中(D＝0.426)

[0646] 605的最近邻居:159,聚类76中(D＝0.437)

[0647] 849的最近邻居:101,聚类18中(D＝0.437)

[0648] 835的最近邻居:101,聚类18中(D＝0.438)

[0649] 790的最近邻居:32,聚类76中(D＝0.442)

[0650] 744的最近邻居:188,聚类38中(D＝0.448)

[0651] 822的最近邻居:282,聚类38中(D＝0.449)

[0652] 723的最近邻居:170,聚类76中(D＝0.456)

[0653] 884的最近邻居:101,聚类18中(D＝0.459)

[0654] 563的最近邻居:34,聚类76中(D＝0.463)

[0655] 867的最近邻居:160,聚类18中(D＝0.463)

[0656] 771的最近邻居:34,聚类76中(D＝0.473)

[0657] 826的最近邻居:165,聚类38中(D＝0.475)

[0658] 777的最近邻居:174,聚类76中(D＝0.478)

[0659] 759的最近邻居:101,聚类18中(D＝0.483)

[0660] 855的最近邻居:101,聚类18中(D＝0.485)

[0661] 702的最近邻居:160,聚类18中(D＝0.492)

[0662] 750的最近邻居:230,聚类76中(D＝0.495)

[0663] 704的最近邻居:216,聚类78中(D＝0.497)

[0664] 711的最近邻居:55,聚类76中(D＝0.502)

[0665] 708的最近邻居:537,聚类78中(D＝0.510)

[0666] 791的最近邻居:115,聚类76中(D＝0.534)

[0667] 722的最近邻居:15,聚类76中(D＝0.547)

[0668] 700的最近邻居:107,聚类76中(D＝0.549)

[0669] 846的最近邻居:72,聚类76中(D＝0.552)

[0670] 876的最近邻居:85,聚类76中(D＝0.560)

[0671] 868的最近邻居:740,聚类78中(D＝0.562)

[0672] 569的最近邻居:68,聚类76中(D＝0.572)

[0673] 817的最近邻居:56,聚类76中(D＝0.582)

[0674] 798的最近邻居:310,聚类38中(D＝0.585)

[0675] 717的最近邻居:216,聚类78中(D＝0.597)

[0676] 879的最近邻居:209,聚类76中(D＝0.612)

[0677] 727的最近邻居:96,聚类76中(D＝0.616)

[0678] 828的最近邻居:142,聚类38中(D＝0.618)

[0679] 840的最近邻居:632,聚类78中(D＝0.640)

[0680] 747的最近邻居:202,聚类76中(D＝0.698)

[0681] 842的最近邻居:797,聚类38中(D＝0.703)

[0682] 442的最近邻居:336,聚类78中(D＝0.735)

[0683] 异常者的总数目:18

[0684] 分配的单例(0.054s)

[0685] 产生的Dc(12.773s)

[0686] 用14个节点和24个边构建图(12.774s)

[0687] 合并聚类用时12.775s

[0688] 总而言之,这些数据通过递归分裂和测试然后合并例示了进行聚类。

[0689] 实例2

[0690] 将通过递归树状图分裂和测试然后合并进行聚类的结果可视化

[0691] 本实例描述了将实例1中例示的通过递归树状图分裂和测试然后合并进行聚类的结果可视化。

[0692] 检查分裂和合并步骤后生成的聚类标记。

[0693] 在[6]中:plt.scatter(x1,x2,edgecolors＝'none')

[0694] _＝plt.axis('off')

[0695] plt.title('Pre-clustering')

[0696] #使用合并前的标记的聚类结果(标记单例)

[0697] plt.figure()

[0698] split.plot_labels_legend(x1,x2,split.str_labels_to_ints(ys))

[0699] plt.title('After splitting step')

[0700] #使用合并后的标记的聚类结果

[0701] plt.figure()

[0702] split.plot_labels_legend(x1,x2,ym)

[0703] plt.title('After merging step')

[0704] #使用合并后的标记的聚类结果

[0705] plt.figure()

[0706] split.plot_labels_legend(x1,x2,ym_community)

[0707] plt.title('After merging step using community detection')

[0708] Out[6]:<matplotlib.text.Text at 0x112674510>

[0709] 图8小图(a)-(d)显示在分裂和合并单细胞的表达谱之后二维空间中表达谱的非限制性示例性图。

[0710] 探索如何决定分裂

[0711] 使用“print_history”函数,“print_history”函数可用于探索该方法的分裂步骤如何处理数据集。第i行描述了第i个有效分裂。如果生成的两个聚类都高于“min_clust_size”并且从分裂生成的最低p值低于阈值,则认为分裂有效。

[0712] 在[7]中:split.print_history(genes,shistory)

[0713] 分裂前:886L:883R:3得分:1.8E+308最高基因:RPL31最高基因得分:1.8E+308

[0714] 分裂前:882L:484R:398得分:182.26最高基因:FTL最高基因得分:182.26

[0715] 分裂前:484L:481R:3得分:1.8E+308最高基因:RPL23最高基因得分:1.8E+308

[0716] 分裂前:479L:195R:284得分:125.49最高基因:IGHM最高基因得分:125.49

[0717] 分裂前:195L:177R:18得分:15.35最高基因:RRP7A最高基因得分:15.35

[0718] 分裂前:174L:12R:162得分:18.88最高基因:ANXA11最高基因得分:18.88

[0719] 分裂前:160L:28R:132得分:12.32最高基因:TTF1最高基因得分:12.32

[0720] 分裂前:28L:25R:3得分:13.94最高基因:SRPK1最高基因得分:13.94

[0721] 分裂前:132L:122R:10得分:18.52最高基因:TOP2A最高基因得分:18.52

[0722] 分裂前:122L:13R:109得分:24.92最高基因:CACYBP最高基因得分:24.92

[0723] 分裂前:109L:105R:4得分:31.72最高基因:RPSA最高基因得分:31.72

[0724] 分裂前:105L:15R:90得分:11.31最高基因:PSMD14最高基因得分:11.31

[0725] 分裂前:281L:271R:10得分:38.04最高基因:RNASEH2B最高基因得分:38.04

[0726] 分裂前:271L:2R:269得分:233.23最高基因:GAS8最高基因得分:233.23

[0727] 分裂前:268L:265R:3得分:80.24最高基因:CNPY3最高基因得分:80.24

[0728] 分裂前:265L:4R:261得分:100.26最高基因:MZB1最高基因得分:100.26

[0729] 分裂前:395L:392R:3得分:228.58最高基因:CREB3L1最高基因得分:228.58

[0730] 分裂前:386L:32R:354得分:33.24最高基因:VMP1最高基因得分:33.24

[0731] 分裂前:351L:19R:332得分:32.86最高基因:EIF2B1最高基因得分:32.86

[0732] 分裂前:332L:6R:326得分:83.57最高基因:NUDT5最高基因得分:83.57

[0733] 分裂前:326L:3R:323得分:148.25最高基因:TMSB4X最高基因得分:148.25

[0734] 分裂前:323L:314R:9得分:71.43最高基因:C12orf57最高基因得分:71.43

[0735] 分裂前:314L:221R:93得分:48.70最高基因:RPL27A最高基因得分:48.70

[0736] 分裂前:217L:215R:2得分:133.42最高基因:JUN最高基因得分:133.42

[0737] 使每个分裂中牵涉的点可视化。每行有2个图。行i描述了保存的第i个分裂。图9小图(a)-(x)是二维空间中表达谱的非限制性示例性图,显示了怎样决定分裂。在每个小图中,左图指示怎样实施分裂。蓝点完全未牵涉到分裂中。红点和绿点曾经在同一个聚类中,而之后分开了。左图的标题指示分裂数目和实现最大t统计量(在取绝对值后)的3个基因。与每个基因关联的数字是对应p值的-log10。每个基因旁边的括号中的‘0’或‘1’指示具有该基因的更高的均值表达的聚类。右图显示实现最大t统计量的基因的对数表达。

[0738] 在[8]中:split.visualize_history(np.log(1+X),x1,x2,genes,shistory)/Users/user1/anaconda2/lib/python2.7/site-packages/matplotlib/pyplot.py:516:RuntimeWarning:More max open warning,RuntimeWarning)

[0739] “analyze_split”函数可用于更仔细地查看指示为何保留特定分裂的基因。还使用关键词“show_background”来显示分裂中未牵涉到的细胞。使用“clust”仅去查看在特定聚类中更高表达的基因。“num_genes”可用于显示自定义数目的基因。

[0740] 在[9]中:#查看分裂5

[0741] split_num＝5

[0742] cluster_of_interest＝None

[0743] show_background＝False

[0744] split.analyze_split(X,x1,x2,genes,shistory,split_num,num_genes＝12,show_background＝show_background,clust＝cluster_of_interest)

[0745] 图10显示了在第五次分裂循环后二维空间中表达谱的非限制性示例性图。图11小图(a)-(l)是二维空间中表达谱的非限制性示例性图,显示了为何树状图中的某些分裂被保留用于图10中所示的第五次分裂循环。

[0746] 探索如何决定合并

[0747] 用于探索如何决定分裂的函数可用于探索如何实施合并。

[0748] 在[10]中:split.print_history(genes,mhistory)

[0749] split.visualize_history(np.log(1+X),x1,x2,genes,mhistory)

[0750] 886个样本中80个是单例

[0751] 单例442、569、700、708、717、722、727、747、791、798、817、828、840、842、846、868、

876、879m

[0752] 单例15与聚类12(N＝24)合并以形成聚类1(N＝25)

[0753] 单例3、4、6、7、9、10、11、13、14、19、20、21、22、24、25、26与聚类18(N＝90)合并

[0754] 单例5、29、30、31、32、33、34、35、36、37、39、40、41与聚类38(N＝261)合并以形成

[0755] 单例28、53、55、56、58、59、60、61、63、64、65、67、68、69、70、71、72、73、74、75、79、80、81

[0756] 单例54、66与聚类78(N＝93)合并以形成聚类13(N＝95)

[0757] 合并后:25L:10R:15得分:3.60最高基因:ENOSF1最高基因得分:3.60

[0758] 合并后:25L:15R:10得分:4.31最高基因:MAGED1最高基因得分:4.31

[0759] 合并后:38L:13R:25得分:4.37最高基因:PRPF40A最高基因得分:4.37

[0760] 合并后:63L:25R:38得分:5.23最高基因:ALDOC最高基因得分:5.23

[0761] 合并后:125L:30R:95得分:6.04最高基因:PARP1最高基因得分:6.04

[0762] 合并后:12L:10R:2得分:6.81最高基因:IGLC3最高基因得分:6.81

[0763] 合并后:88L:25R:63得分:7.19最高基因:HMGB2最高基因得分:7.19

[0764] 合并后:30L:18R:12得分:7.23最高基因:VIM最高基因得分:7.23

[0765] 合并后:155L:125R:30得分:9.76最高基因:HMGN5最高基因得分:9.76

[0766] 图12小图(a)-(i)是二维空间中表达谱的非限制性示例性图,显示了怎样决定合并。

[0767] 在[11]中:#查看合并2

[0768] merge_num＝2

[0769] cluster_of_interest＝None

[0770] show_background＝False

[0771] split.analyze_split(X,x1,x2,genes,mhistory,merge_num,num_genes＝4,show_background＝show_background,clust＝cluster_of_interest)

[0772] 图13显示了第二次合并循环后二维空间中表达谱的非限制性示例性图。图14小图(a)-(d)是二维空间中表达谱的非限制性示例性图,显示了图13中所示的第二次合并循环是怎样决定的。

[0773] 差异表达

[0774] 模块允许两种简单的差异表达分析。第一种对每个聚类进行一对余比较,根据每个基因的t检验使每个聚类的最重要基因可视化。第二种对每两对聚类进行成对比较。

[0775] 在[12]中:

[0776] split.save_more_highly_expressed_genes_in_one_clust(X,genes,ym,x1,x2,num_genes＝3,show_plots＝True)

[0777] 图15小图(a)-(f)是显示差异表达分析的一个非限制性示例类型的图。

[0778] 在[13]中:

[0779] split.pairwise_cluster_comparison(X,genes,ym,x1＝x1,x2＝x2,num_genes＝3,show_plots＝True,verbose＝F

[0780] dendrosplit/utils.py:39:FutureWarning:elementwise comparison failed;

[0781] returning scalar instead,but in

[0782] plt.plot(x1[y==i],x2[y==i],'.',c=RGBs[j],label=str(i)+'

[0783] ('+str(np.sum(y==i))+')')

[0784] dendrosplit/feature selection.py:221:RuntimeWarning:divide by zero encountered in double scalars fold=g mean j/g mean i

[0785] 图16小图(a)-(o)是显示差异表达分析的另一个非限制性示例类型的图。

[0786] 距离分布

[0787] 模块还允许用户可视化每个聚类内的距离分布。对于给定的聚类,此函数对成对距离(对于所有聚类中的所有点)的总集合的每个百分位仓(bin)绘制成对距离(聚类中的点之间)的比例。例如,在1情况下0.3指示成对距离的30%落在总距离的第5和第10百分位之间。使用此函数可以感知根据最初距离矩阵,聚类的内聚性如何。直观地说,一个好的聚类应该具有均彼此接近的点。例如,在底部10个仓(即底部50个百分位)中没有距离的聚类将被认为是较差的。注意,正如预期的那样,下面的聚类"-1"就是这种情况,其中包含异常者。

[0788] 在[14]中:

[0789] merge.visualize_within_cluster_distance_distributions(D,ym,show_D_dist=True)

[0790] 图17小图(a)-(g)是使聚类之间距离可视化的非限制性示例图。

[0791] 树状图

[0792] 模块还允许用户生成树状图并根据树状图获得细胞的次序。树状图可能难以在iPython笔记本中查看。在一些实施例中,可以保存树状图,如下所示。用户可以输入聚类标记("labels"关键字)。如果需要,该函数可以为聚类内所有样本的名称着上相同的颜色。

[0793] 在[15]中:cell_order=

[0794] split.plot_dendro(D,return_cell_order=True,labels=ym,save_name='/Users/user1/Desktop/dendrogram')

[0795] dendrosplit/split.py:233:FutureWarning:comparison to `None`will result in an elementwise object comp if labels!=None:

[0796] 图18显示了非限制性示例性树状图。

[0797] 总而言之,这些数据展示了本公开用于可视化递归分裂和测试然后合并的各种步骤和结果的各种工具。

[0798] 实例3

[0799] 用于通过递归树状图分裂和测试然后合并进行聚类的参数扫描

[0800] 本实例描述了用于优化参数来进行递归分裂和测试然后合并的参数扫描。

[0801] 在该方法的分割步骤期间,可以细调两个超参数:得分阈值和解散百分位。可以探索使用不同的超参数能生成怎样不同的聚类。通过利用以较小得分阈值(较小阈值导致更多聚类)生成的聚类分隔以较大得分阈值生成的聚类的事实,可以快速获得几种(分裂后合并前)聚类结果。首先,以非常低的阈值运行分裂步骤。其次,使用来自history()函数的获得聚类。

[0802] 下面显示了扫描扫过各种得分阈值的例子。可以对解散百分位值进行同样操作。

[0803] 在[16]中:ys,shistory=split.dendrosplit((D,X),preprocessing='

precomputed',score_threshold＝2,verbose＝False,disband_percentile＝50)

[0804] ys_sweep＝[ ]

[0805] thresholds＝range(5,100,5)

[0806] for threshold in thresholds:

[0807] ys_sweep.append(split.get_clusters_from_history(D,shistory,threshold,50))

[0808] plt.figure()

[0809] split.plot_labels_legend(x1,x2,split.str_labels_to_ints(ys_sweep[-1]))

[0810] plt.title('Clustering result using a threshold of%.3f'%(threshold))

[0811] 图19小图(a)-(s)是显示参数扫描的非限制性示例图。

[0812] 通过扫描参数,可以研究聚类的数目如何作为阈值的函数而变化。这可以为用户提供如何为特定应用选择最佳阈值方面的洞察。

[0813] 在[17]中:def count_nonsingleton_clusters(y):

[0814] return sum([1for i in np.unique(y)if np.sum(y＝＝i)!＝1])

[0815] plt.plot(thresholds,[count_nonsingleton_clusters(i)for i in ys_sweep])

[0816] plt.grid()

[0817] plt.xlabel('thresholds(-log10(p-value))')

[0818] plt.ylabel('number of nonsingletone clusters')

[0819] Out[17]:<matplotlib.text.Text at 0x117fb3290>

[0820] 图20是显示怎样使用参数扫描来鉴定阈值的非限制性示例图。由于大量的表达谱聚类以仅5的阈值鉴定,较少的表达谱聚类以更大的阈值来鉴定(例如图19小图(h)中显示的40)。

[0821] 总之,这些数据例示了通过用于递归分裂和测试然后合并的参数扫描来优化超参数。

[0822] 实例4

[0823] 通过递归树状图分裂和测试然后合并进行聚类

[0824] 本实例描述了通过递归分裂(例如,递归树状图分裂)和测试然后合并的聚类方法。在树状图的每个聚类或节点处(叶节点除外),在本实例中,两个子聚类的聚类内中值相关性高于聚类间中值相关性。

[0825] 在357个细胞的表达谱的分裂和测试阶段期间,从树状图的顶部开始,将树分成两个候选子树。在两个子聚类的聚类内中值相关性应该高于聚类间中值相关性的约束下,分裂对应于聚类被分成两个候选的子聚类。对分裂的质量打分。如果认为子聚类足够不同,则该阶段继续到每个子树。否则,该方法终止于树状图的这一部分。此阶段为数据集生成了一组标记。

[0826] 图21小图(a)-(j)是显示第一次分裂结果的非限制性示例图。在第一次分裂期间,二十个基因(表1中显示)被确定为在357个细胞中是差异表达的。

[0827] 表1.在第一次分裂期间二十个基因被确定为在357个细胞中是差异表达的。

[0828]

分裂	基因	p值	较大聚类
1	IGLC3|ENST00000390325.2|Reference_end	201.35	0
1	JCHAIN|NM_144646.3|Reference_end	105.57	0
1	ADA|NM_000022.3|Reference_end	89.27	1
1	TCL1A|NM_021966.2|Reference_end	81.19	0
1	CD74|NM_004355.3|Reference_end	62.65	0
1	CD3D|NM_000732.4|Reference_end	50.32	1
1	POU2AF1|NM_006235.2|Reference_end	39.94	0
1	CD52|NM_001803.2|Reference_end	39.1	0
1	QPCT|NM_012413.3|Reference_end	38.87	0
1	HLA-DRA|NM_019111.4|Reference_end	26.64	0
1	CD22|NM_001771.3|Reference_end	25.96	0
1	IRF8|NM_002163.2|Reference_end	21.25	0
1	MS4A1|NM_021950.3|PolyA_1	19.99	0
1	CD37|NM_001774.2|Reference_end	18.49	0
1	LEF1|NM_016269.4|Reference_end	17.63	1
1	MME|NM_000902.3|Reference_end	15.59	0
1	BCL6|NM_001706.4|Reference_end	13.39	0
1	CD27|NM_001242.4|Reference_end	11.02	0
1	IL32|NM_004221.4|Reference_end	10.86	1
1	CD38|NM_001775.3|Reference_end	10.65	0

[0829] 图22是非限制性示例性的t分布随机邻域嵌入(t-SNE)图，其例示了357个细胞的表达谱的分裂结果，显示357个细胞被以阈值10分类为两个聚类。图23显示了非限制性示例性树状图，其显示了基于表2中所示的特征分类为两个聚类的表达谱(表1中的聚类0对应于表2中的聚类1，且表2中的聚类1对应于表2中的聚类2)。图24是显示参数扫描的非限制性示例性图。因为两个表达谱聚类是用仅10的阈值鉴定的，所以用更大的阈值鉴定同样的两个表达谱聚类(比较图24和图20)。

[0830] 表2.按p值排序的两个聚类的成对聚类特征

[0831]

比较	基因	p 值	较大聚类	较大聚类的表达的倍数变化
聚类 1 相对于聚类 2	IGLC3|ENST00000390325.2|Reference_end	201.35	1	183.947
聚类 1 相对于聚类 2	JCHAIN|NM_144646.3|Reference_end	105.572	1	50.085
聚类 1 相对于聚类 2	ADA|NM_000022.3|Reference_end	89.274	2	11.82
聚类 1 相对于聚类 2	TCL1A|NM_021966.2|Reference_end	81.191	1	134.689
聚类 1 相对于聚类 2	CD74|NM_004355.3|Reference_end	62.653	1	173.268
聚类 1 相对于聚类 2	CD3D|NM_000732.4|Reference_end	50.32	2	109.703
聚类 1 相对于聚类 2	POU2AF1|NM_006235.2|Reference_end	39.943	1	19.778
聚类 1 相对于聚类 2	CD52|NM_001803.2|Reference_end	39.105	1	50.988
聚类 1 相对于聚类 2	QPCT|NM_012413.3|Reference_end	38.87	1	INF
聚类 1 相对于聚类 2	HLA-DRA|NM_019111.4|Reference_end	26.642	1	71.744
聚类 1 相对于聚类 2	CD22|NM_001771.3|Reference_end	25.96	1	64.976
聚类 1 相对于聚类 2	IRF8|NM_002163.2|Reference_end	21.245	1	INF

[0832]

聚类1相对于聚类2	MS4A1|NM_021950.3|PolyA_1	19.993	1	INF
聚类1相对于聚类2	CD37|NM_001774.2|Reference_end	18.492	1	INF
聚类1相对于聚类2	LEF1|NM_016269.4|Reference_end	17.63	2	4.807
聚类1相对于聚类2	MME|NM_000902.3|Reference_end	15.585	1	37.451
聚类1相对于聚类2	BCL6|NM_001706.4|Reference_end	13.393	1	INF
聚类1相对于聚类2	CD27|NM_001242.4|Reference_end	11.018	1	INF
聚类1相对于聚类2	IL32|NM_004221.4|Reference_end	10.862	2	INF
聚类1相对于聚类2	CD38|NM_001775.3|Reference_end	10.649	1	INF

[0833] 总而言之,这些数据通过递归分裂和测试然后合并例示了进行聚类。在本实例中,在树状图的每个聚类或节点处(叶节点除外),两个子聚类的聚类内中值相关性高于聚类间中值相关性。

[0834] 在至少一些先前描述的实施例中,在一个实施例中使用的一个或多个元素可以互换地用于另一个实施例中,除非这种替换在技术上不可行。本领域技术人员将理解,在不脱离所要求保护的主题的范围的情况下,可以对上述方法和结构进行各种其他的省略、添加和修改。所有此类修改和改变都旨在落在由所附权利要求书限定的主题的范围内。

[0835] 关于本文中使用基本上任何复数和/或单数术语,在对于背景和/或应用适当的情况下,本领域技术人员可以从复数转换为单数和/或从单数转换为复数。为了清楚起见,可以在本文明确阐述各种单数/复数排列。如本说明书和所附权利要求书中使用的,除非上下文另有明确指示,否则单数形式“一个/一种(a/an)”和“所述(the)”包括复数的提及物。除非另外说明,在本文中对“或”的任何提及旨在涵盖“和/或”。

[0836] 本领域技术人员将理解,一般来说,本文使用的术语,尤其是所附权利要求书(例如,所附权利要求书的主体)中的术语,通常旨在作为“开放性的”术语(例如,术语“包括(including)”应解释为“包括但不限于(including but not limited to)”,术语“具有(having)”应解释为“具有至少(having at least)”,术语“包括(includes)”应解释为“包括但不限于(includes but is not limited to)”等)。本领域技术人员将进一步理解,如果预期到所介绍的权利要求陈述的特定数目,这样的预期将明确地陈述于权利要求中,并且在不存在这种陈述的情况下没有这种意图存在。例如,作为对理解的帮助,以下所附权利要求书可以包含介绍性短语“至少一个”和“一个或多个”的使用,以介绍权利要求陈述。然

而,此类短语的使用不应解读为意味着由不定冠词“一个”或“一种”介绍权利要求陈述会将任何包含这种介绍的权利要求陈述的具体权利要求限制到包含仅一个这种陈述的实施例中,甚至当相同的权利要求包括介绍性短语“一个或多个”或“至少一个”以及不定冠词如“一个”或“一种”时也是如此(例如,“一个”和/或“一种”应解释为意指“至少一个”或“一个或多个”);这对于使用定冠词来介绍权利要求陈述同样适用。此外,即使明确地陈述了介绍的权利要求陈述的特定数目,本领域技术人员将认识到,这种陈述应解释为意指至少所陈述的数字(例如,仅陈述“两个陈述”而没有其他修饰词意指至少两个陈述、或两个或更多个陈述)。此外,在使用类似于“A、B和C等中的至少一个”的惯例的那些情况下,通常这种句法结构是在本领域技术人员将理解该惯例的意义上预期(例如,“具有A、B和C中的至少一个的系统”将包括但不限于仅具有A,仅具有B,仅具有C,A和B一起,A和C一起,B和C一起,和/或A、B、和C一起等的系统)。在使用类似于“A、B或C等中的至少一个”的惯例的那些情况下,通常这种句法结构是在本领域技术人员将理解该惯例的意义上预期(例如,“具有A、B或C中的至少一个的系统”将包括但不限于仅具有A,仅具有B,仅具有C,A和B一起,A和C一起,B和C一起,和/或A、B、和C一起等的系统)。本领域技术人员将进一步理解,实际上,无论在说明书、权利要求书还是在附图中,呈现两个或更多个替代术语的任何分离性词语和/或短语应被理解为考虑到包括术语之一、任一术语或两个术语的可能性。例如,短语“A或B”将被理解为包括“A”或“B”或“A和B”的可能性。

[0837] 此外,当本披露的特征或方面以马库什组(Markush group)描述时,本领域技术人员将意识到本披露还由此以马库什组的任何单独的成员或成员子组描述。

[0838] 如本领域技术人员将理解的,出于任何和所有目的,如在提供书面描述方面,本文披露的所有范围还包括任何和所有可能的它的子范围和子范围组合。任何列出的范围都可以很容易地被识别为充分描述并使相同的范围能被分解为至少相等的一半,三分之一,四分之一,五分之一,十分之一等。作为非限制性示例,这里讨论的每个范围可以容易地分解为下三分之一,中三分之一和上三分之一等。如本领域技术人员还将理解的,所有语言,如“多至”、“至少”、“大于”、“小于”等包括所陈述的数字,并且指代可以随后分解为如上讨论的子范围的范围。最后,如本领域技术人员将理解的,范围包括每个单独的成员。因此,例如,具有1-3个物品的组是指具有1、2或3个物品的组。类似地,具有1-5个物品的组指代具有1、2、3、4或5个物品的组,等等。

[0839] 尽管本文已经披露了各种方面和实施例,但其他方面和实施例对本领域技术人员将是明显的。本文披露的各种方面和实施例用于说明的目的而并不意于限制由以下权利要求所指出的真实范围和精神。

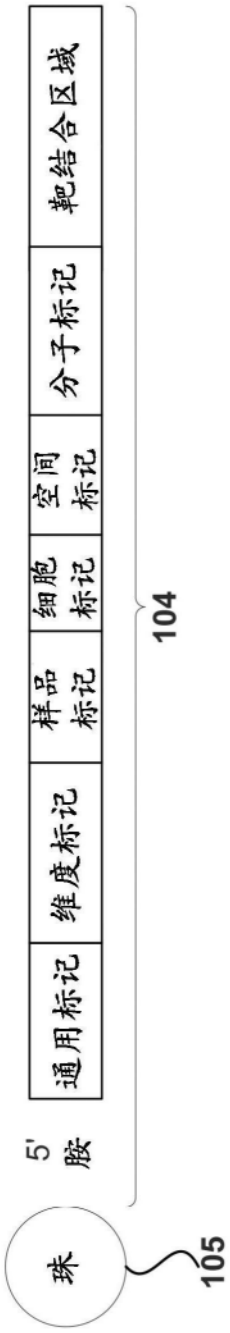


图1

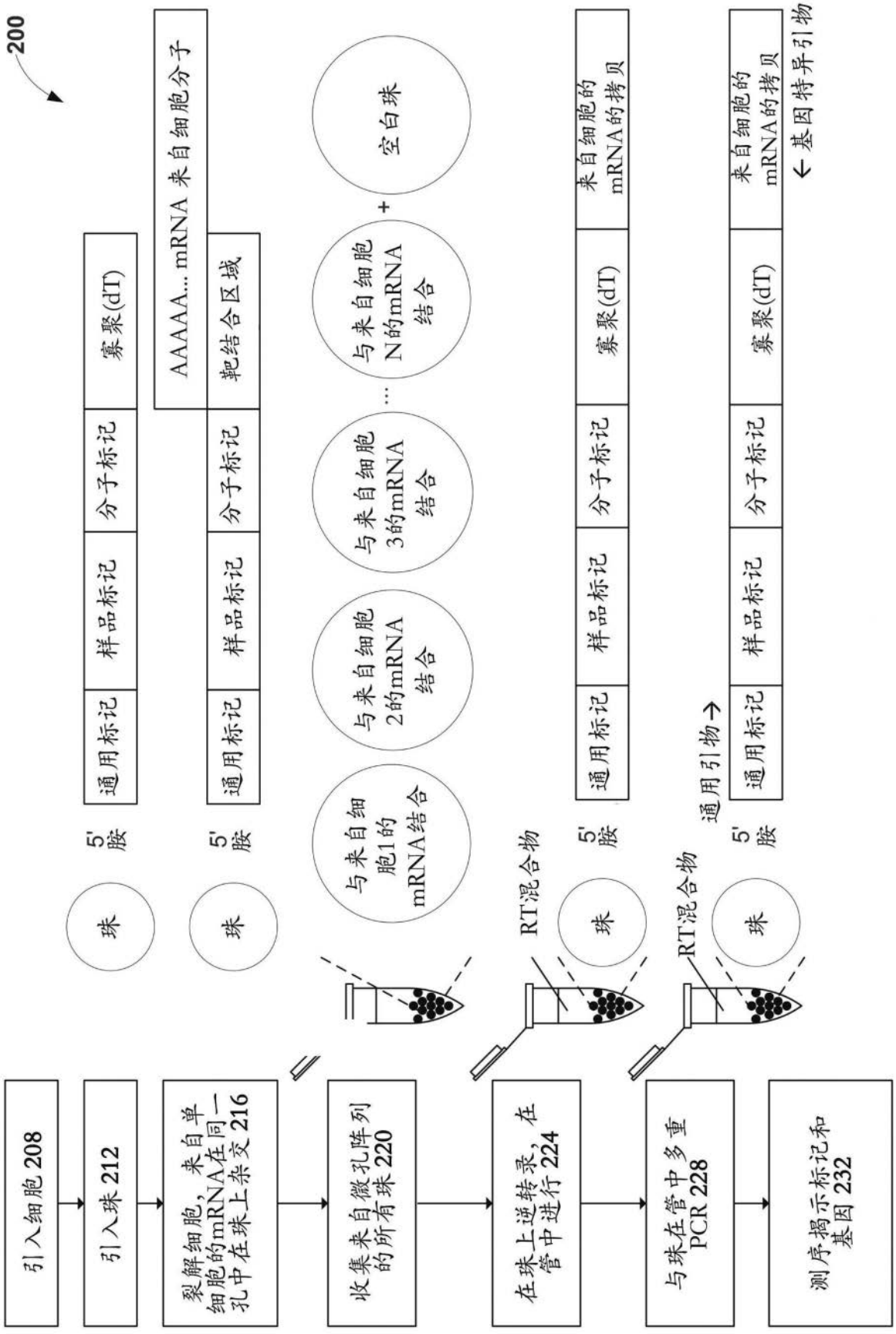

200
引入细胞 208
引入珠 212
裂解细胞，来自单细胞的mRNA在同一孔中在珠上杂交 216
收集来自微孔阵列的所有珠 220
在珠上逆转录，在管中进行 224
与珠在管中多重PCR 228
测序揭示标记和基因 232
珠
5' 胺
通用标记
样品标记
分子标记
寡聚(dT)
AAAAA... mRNA 来自细胞分子
靶结合区域
与来自细胞1的mRNA结合
与来自细胞2的mRNA结合
与来自细胞3的mRNA结合
与来自细胞N的mRNA结合
空白珠
RT混合物
通用引物→
来自细胞的mRNA的拷贝
← 基因特异引物

图2

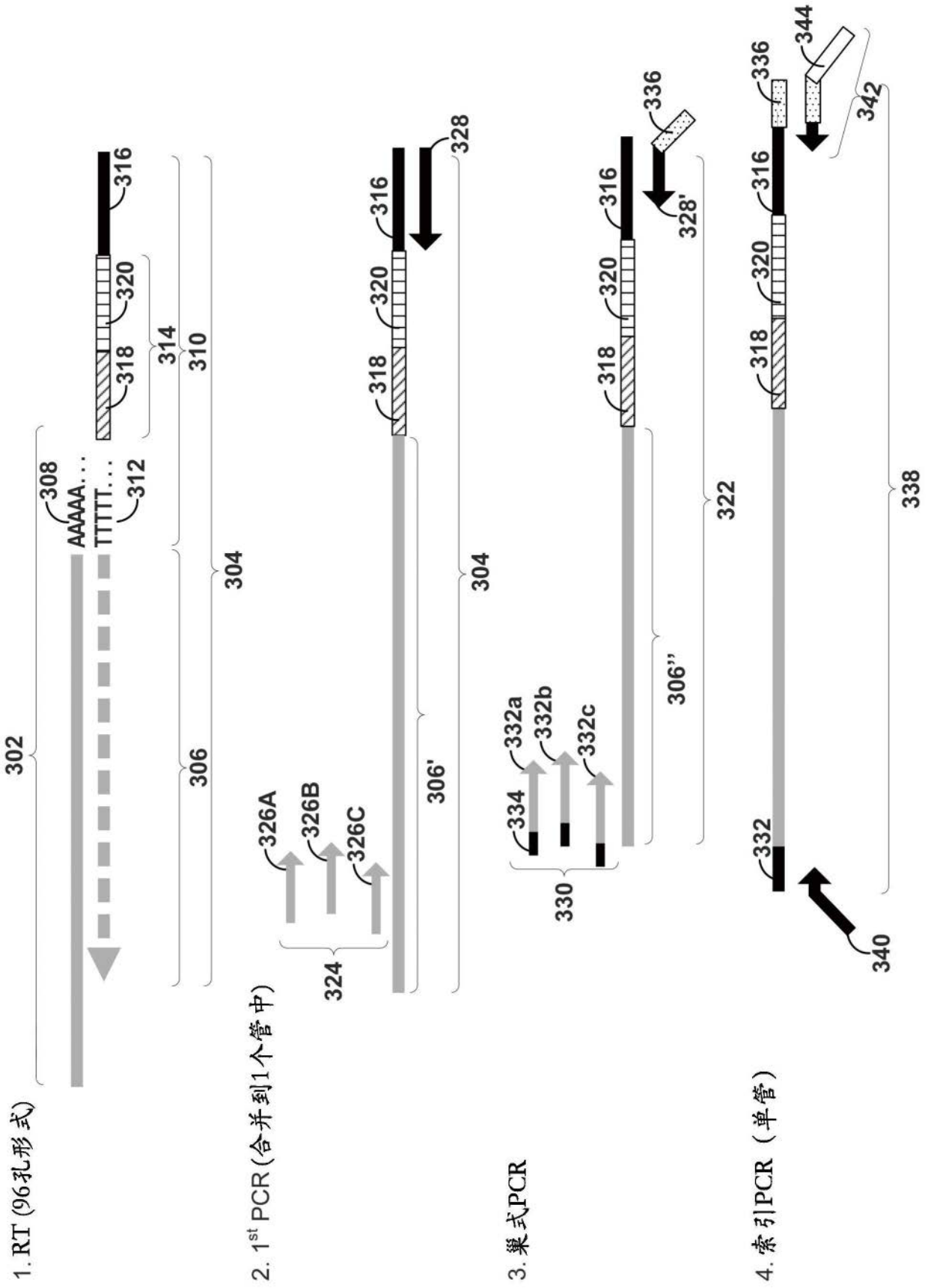

1. RT (96孔形式)
2. 1st PCR (合并到1个管中)
3. 巢式PCR
4. 索引PCR (单管)
302
308
AAAAA...
TTTTT...
312
316
320
318
314
310
304
306
306'
306''
322
324
326A
326B
326C
328
328'
330
332
332a
332b
332c
334
336
338
340
342
344

图3

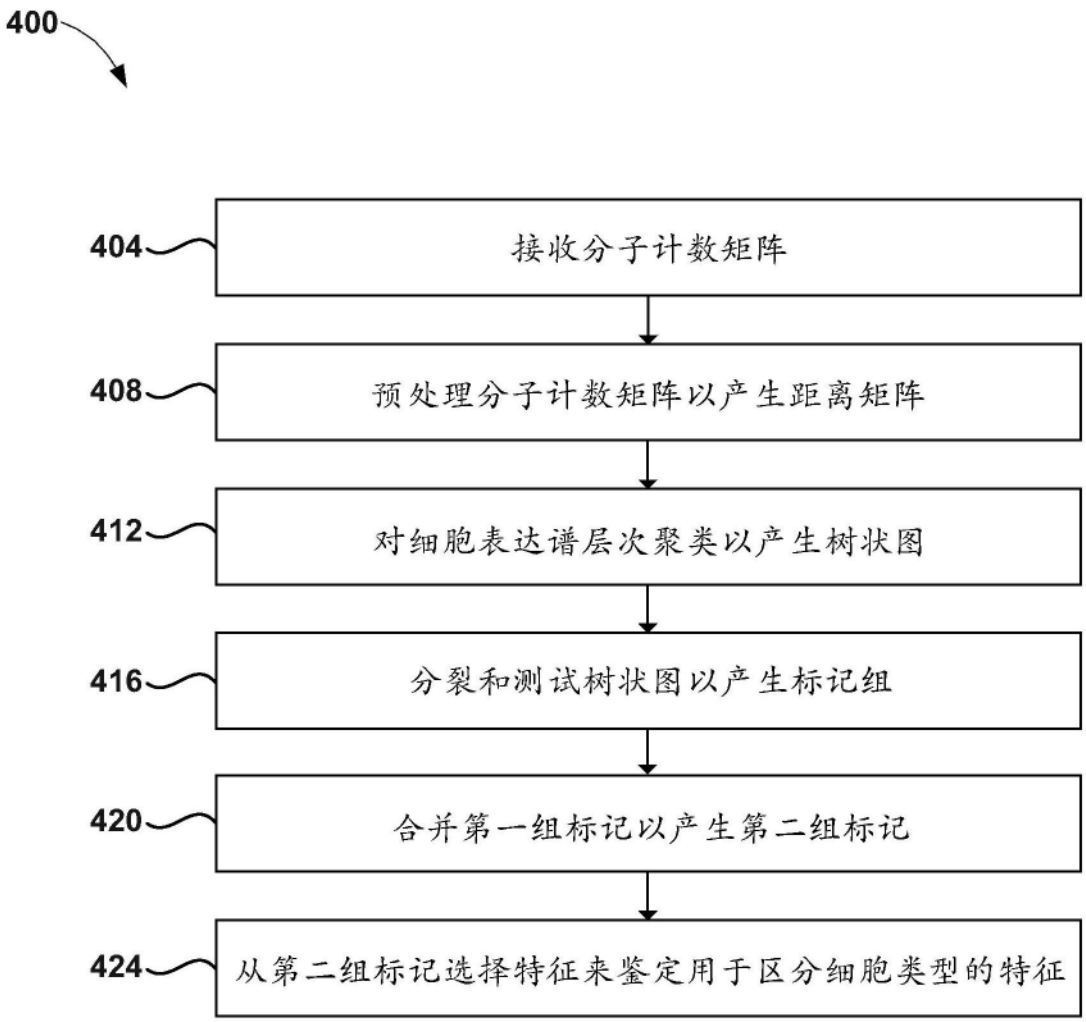

400
404
接收分子计数矩阵
408
预处理分子计数矩阵以产生距离矩阵
412
对细胞表达谱层次聚类以产生树状图
416
分裂和测试树状图以产生标记组
420
合并第一组标记以产生第二组标记
424
从第二组标记选择特征来鉴定用于区分细胞类型的特征

图4

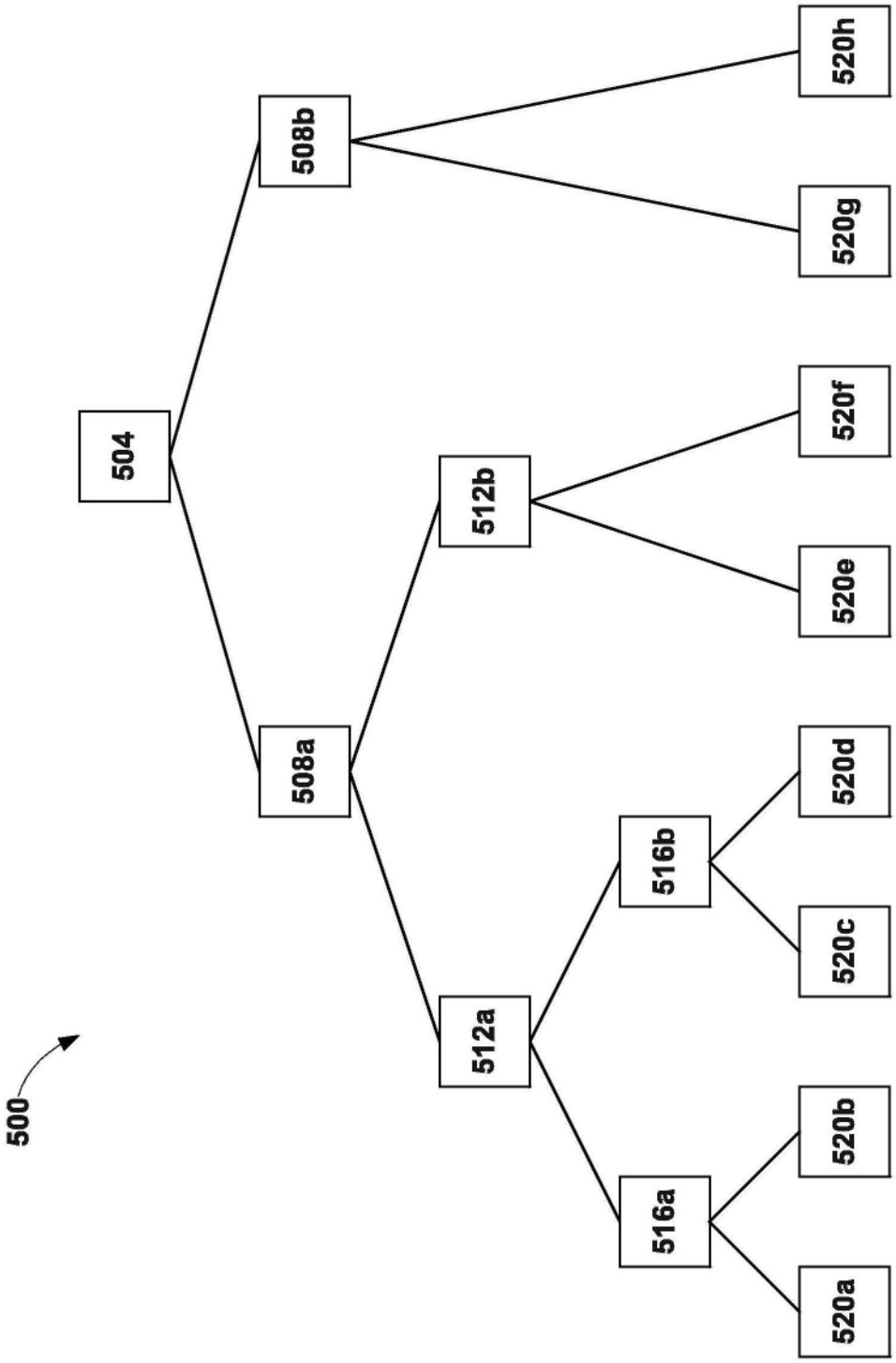

500
504
508b
508a
512b
512a
516b
516a
520h
520g
520f
520e
520d
520c
520b
520a

图5

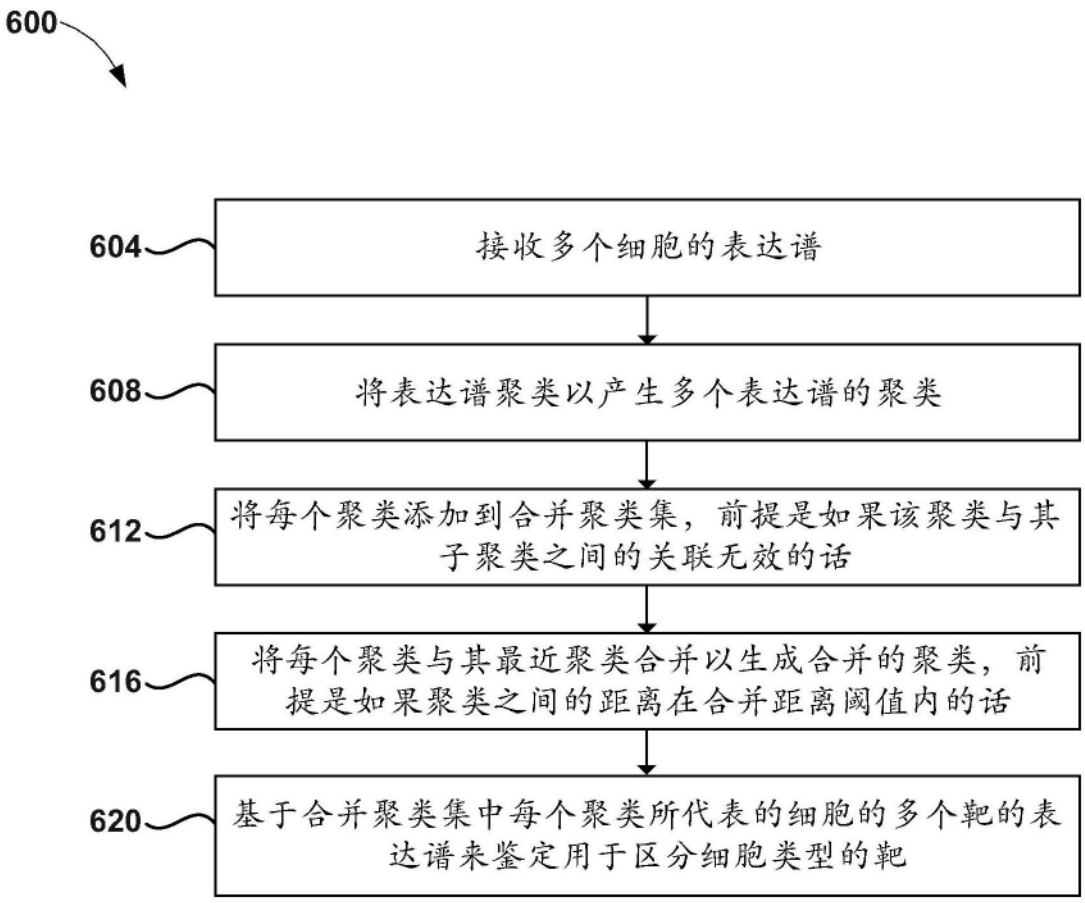

600
604
接收多个细胞的表达谱
608
将表达谱聚类以产生多个表达谱的聚类
612
将每个聚类添加到合并聚类集，前提是如果该聚类与其子聚类之间的关联无效的话
616
将每个聚类与其最近聚类合并以生成合并的聚类，前提是如果聚类之间的距离在合并距离阈值内的话
620
基于合并聚类集中每个聚类所代表的细胞的多个靶的表达谱来鉴定用于区分细胞类型的靶

图6

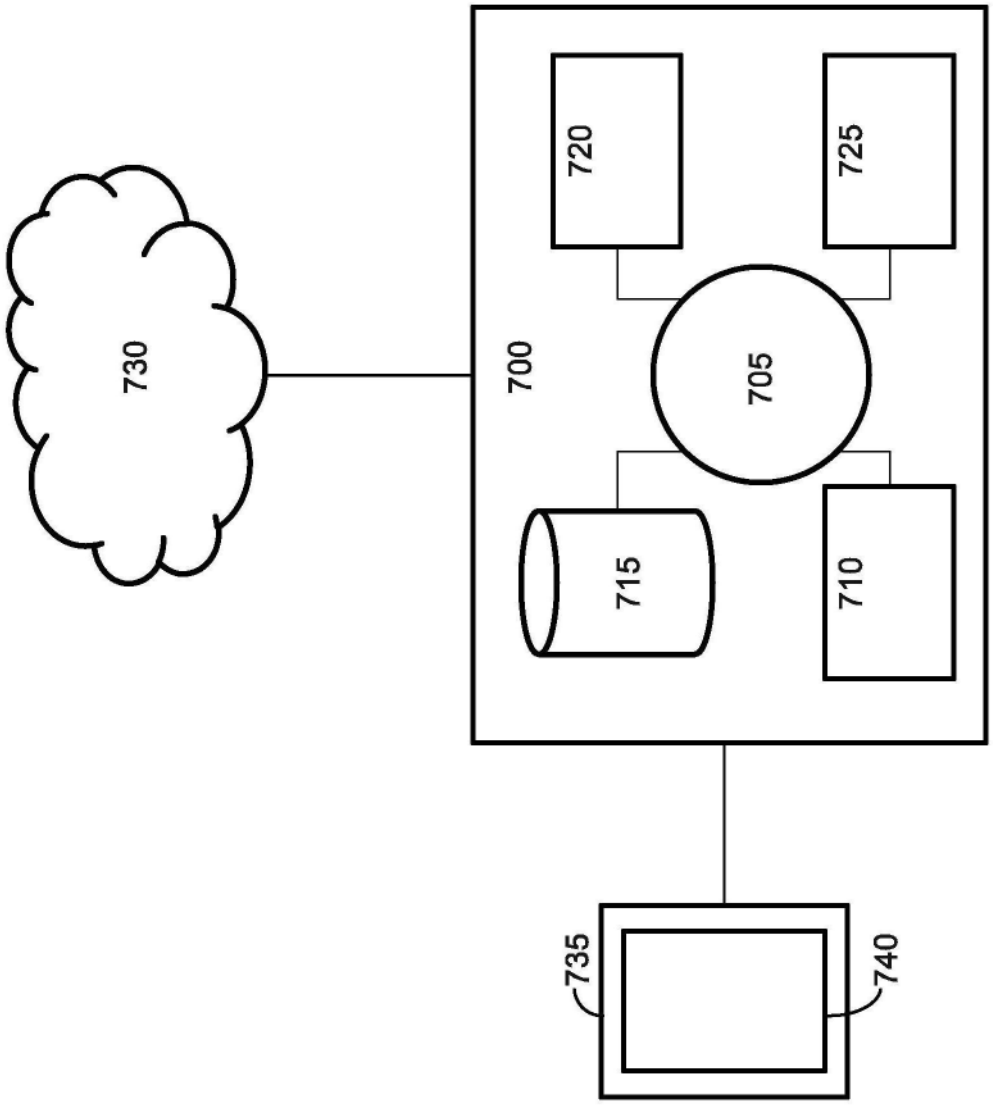


图7

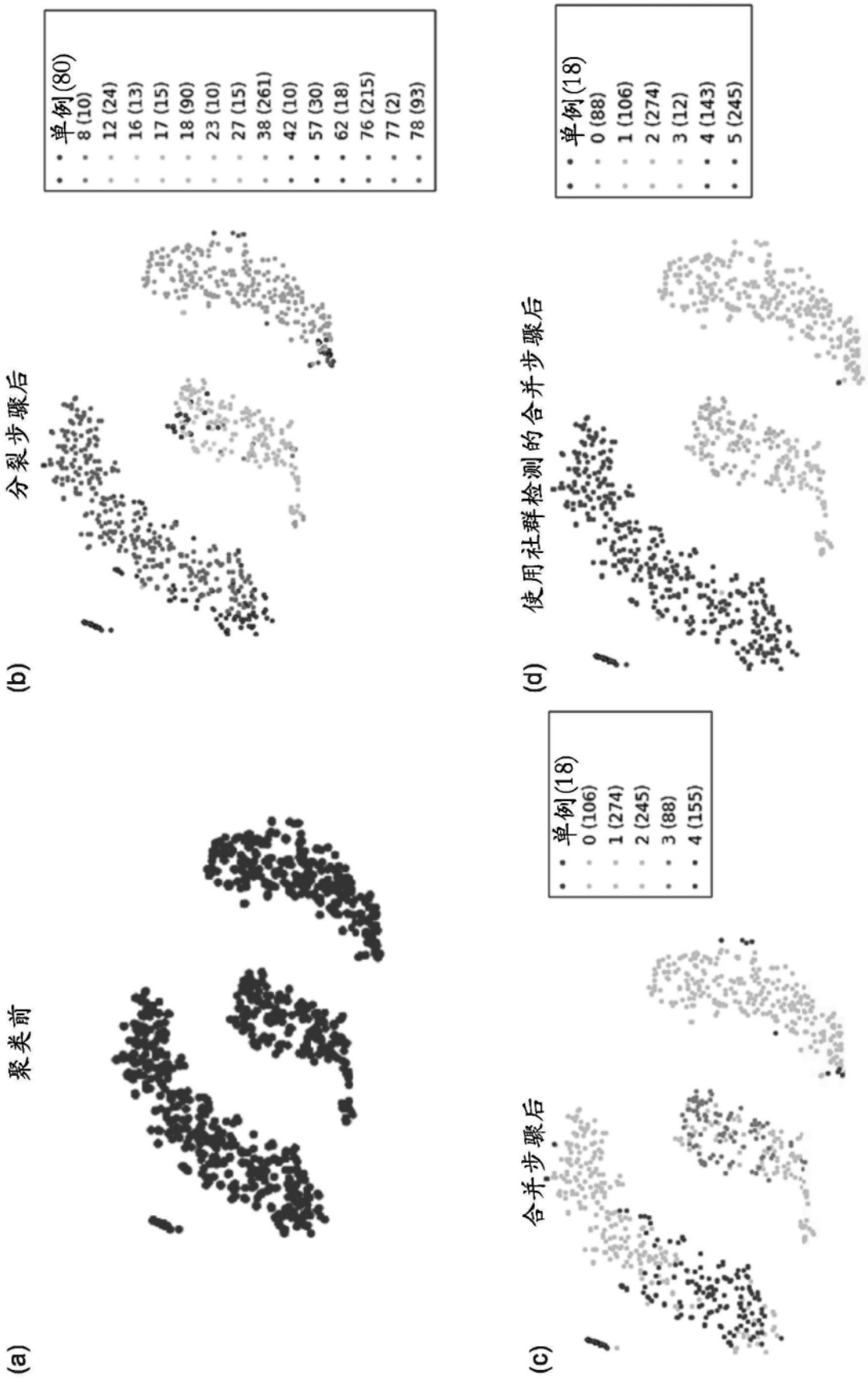

(a)
聚类前
(b)
分裂步骤后
单例(80)
8 (10)
12 (24)
16 (13)
17 (15)
18 (90)
23 (10)
27 (15)
38 (261)
42 (10)
57 (30)
62 (18)
76 (215)
77 (2)
78 (93)
(c)
合并步骤后
单例(18)
0 (106)
1 (274)
2 (245)
3 (88)
4 (155)
(d)
使用社群检测的合并步骤后
单例(18)
0 (88)
1 (106)
2 (274)
3 (12)
4 (143)
5 (245)

图8

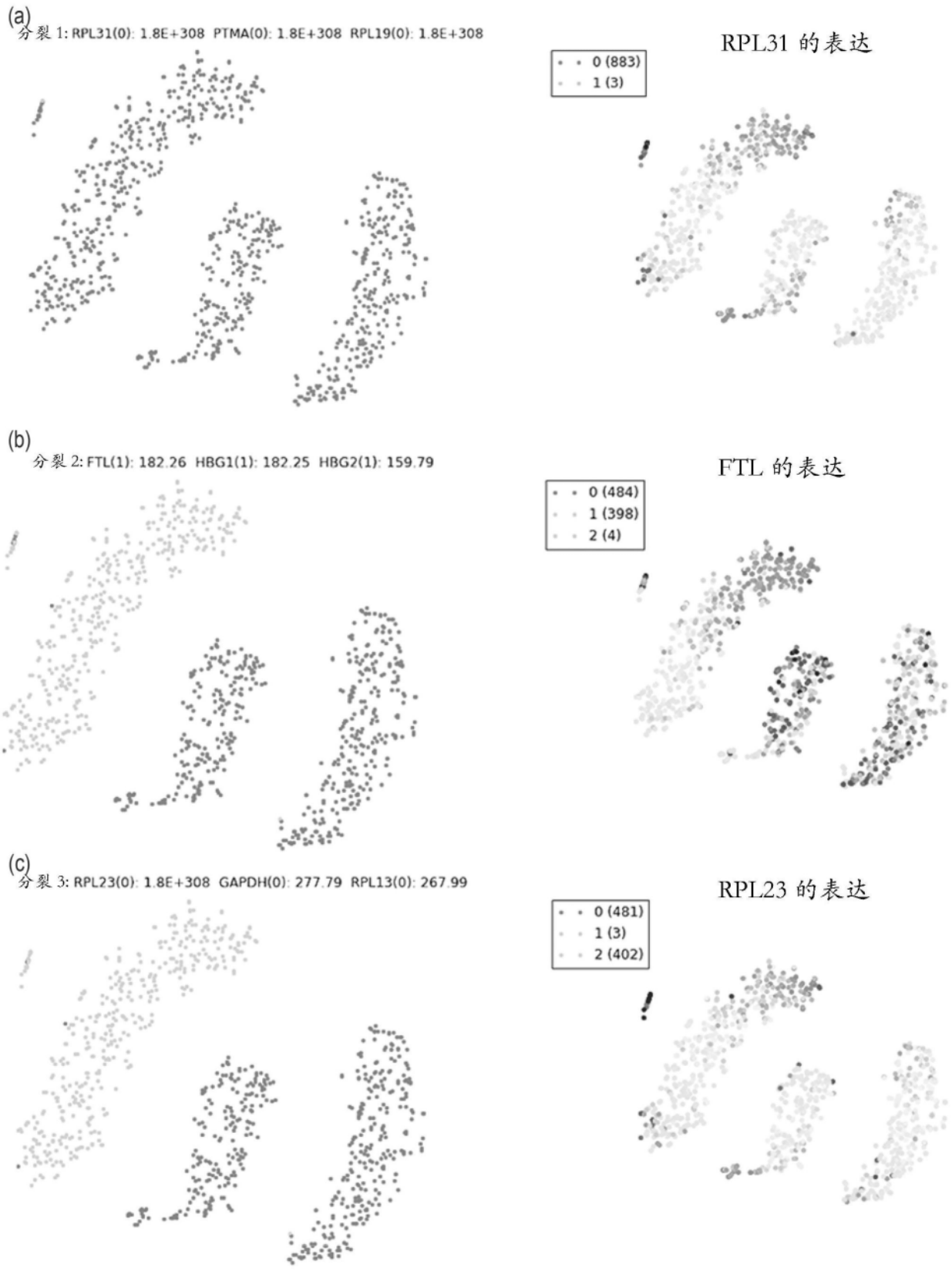

(a)
分裂 1: RPL31(0): 1.8E+308 PTMA(0): 1.8E+308 RPL19(0): 1.8E+308
RPL31 的表达
0 (883)
1 (3)
(b)
分裂 2: FTL(1): 182.26 HBG1(1): 182.25 HBG2(1): 159.79
FTL 的表达
0 (484)
1 (398)
2 (4)
(c)
分裂 3: RPL23(0): 1.8E+308 GAPDH(0): 277.79 RPL13(0): 267.99
RPL23 的表达
0 (481)
1 (3)
2 (402)

图9

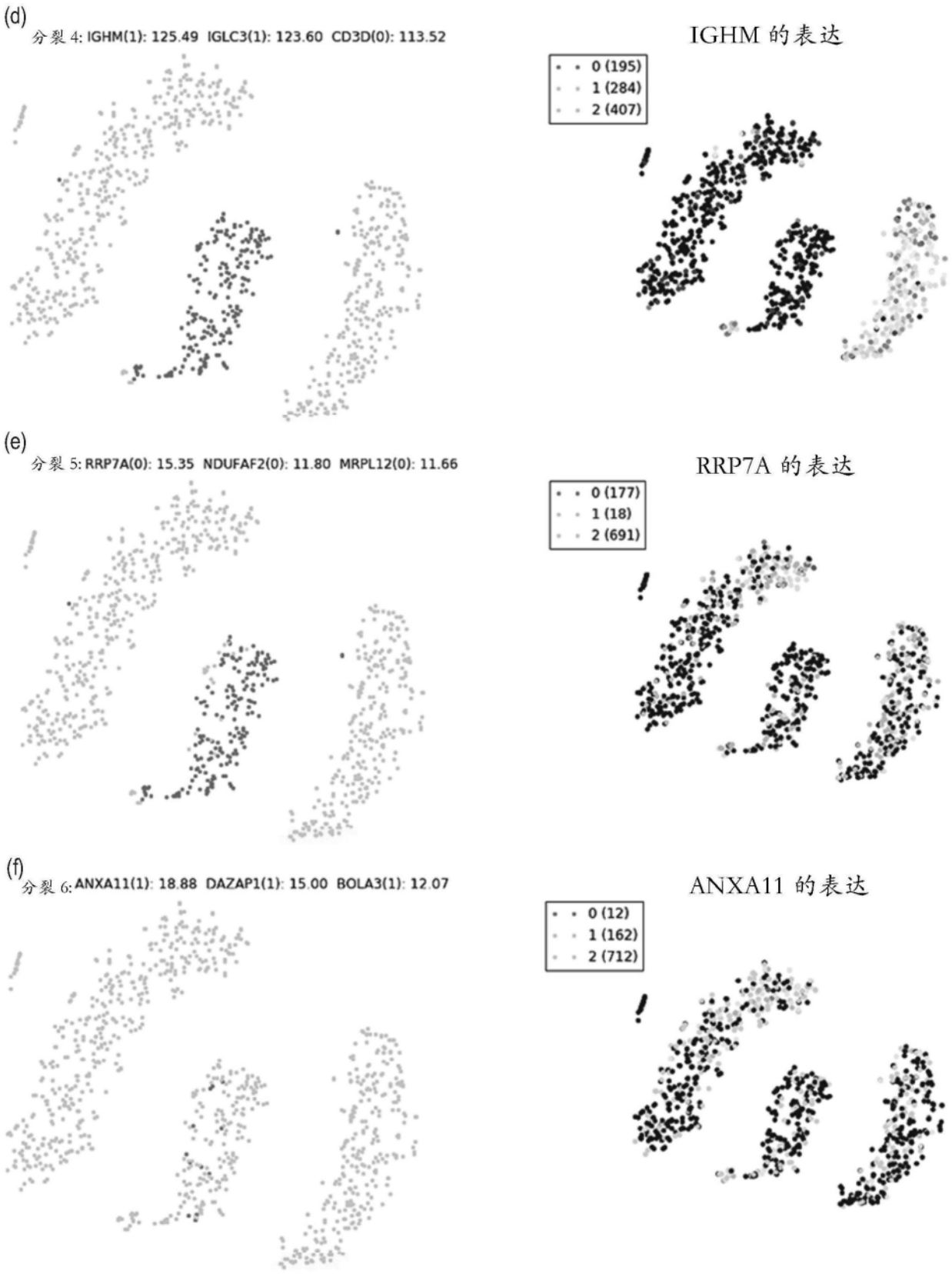

(d)
分裂 4: IGHM(1): 125.49 IGLC3(1): 123.60 CD3D(0): 113.52
IGHM 的表达
0 (195)
1 (284)
2 (407)
(e)
分裂 5: RRP7A(0): 15.35 NDUFAF2(0): 11.80 MRPL12(0): 11.66
RRP7A 的表达
0 (177)
1 (18)
2 (691)
(f)
分裂 6: ANXA11(1): 18.88 DAZAP1(1): 15.00 BOLA3(1): 12.07
ANXA11 的表达
0 (12)
1 (162)
2 (712)

图9(续)

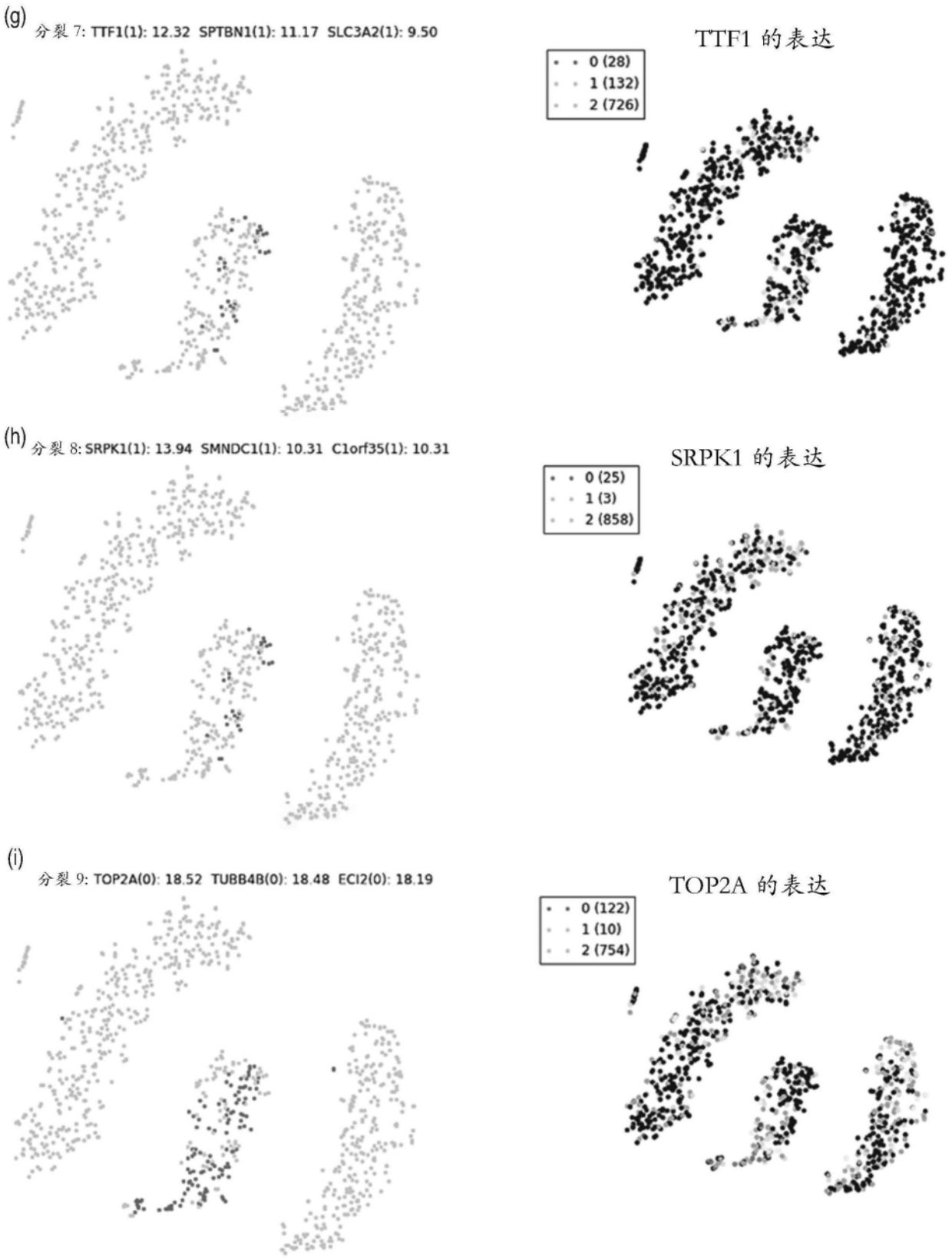

(g)
分裂 7: TTF1(1): 12.32 SPTBN1(1): 11.17 SLC3A2(1): 9.50
TTF1 的表达
0 (28)
1 (132)
2 (726)
(h)
分裂 8: SRPK1(1): 13.94 SMNDC1(1): 10.31 C1orf35(1): 10.31
SRPK1 的表达
0 (25)
1 (3)
2 (858)
(i)
分裂 9: TOP2A(0): 18.52 TUBB4B(0): 18.48 ECI2(0): 18.19
TOP2A 的表达
0 (122)
1 (10)
2 (754)

图9(续)

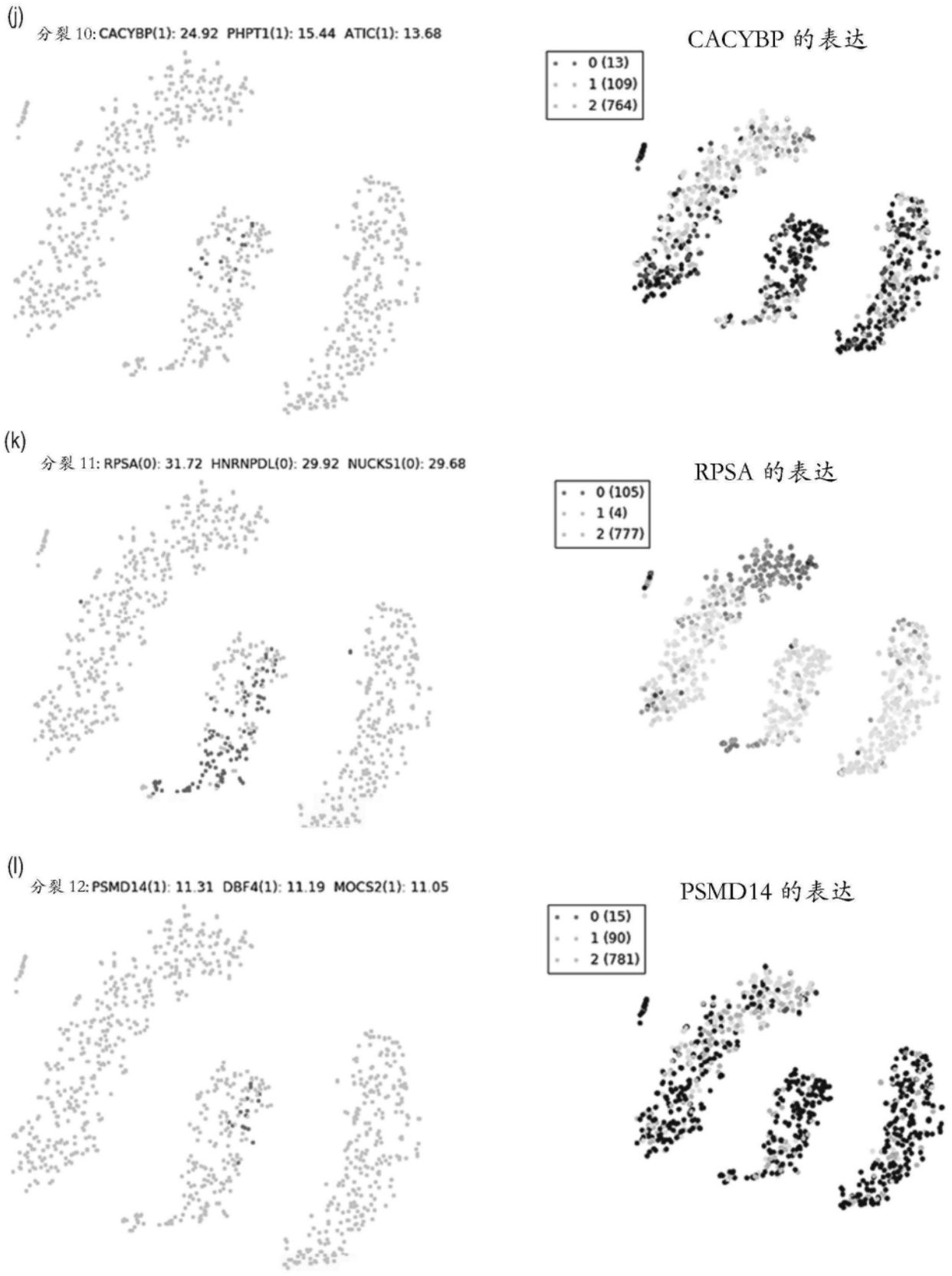

(j)
分裂 10: CACYBP(1): 24.92 PHPT1(1): 15.44 ATIC(1): 13.68
CACYBP 的表达
0 (13)
1 (109)
2 (764)
(k)
分裂 11: RPSA(0): 31.72 HNRNPDL(0): 29.92 NUCKS1(0): 29.68
RPSA 的表达
0 (105)
1 (4)
2 (777)
(l)
分裂 12: PSMD14(1): 11.31 DBF4(1): 11.19 MOCS2(1): 11.05
PSMD14 的表达
0 (15)
1 (90)
2 (781)

图9(续)

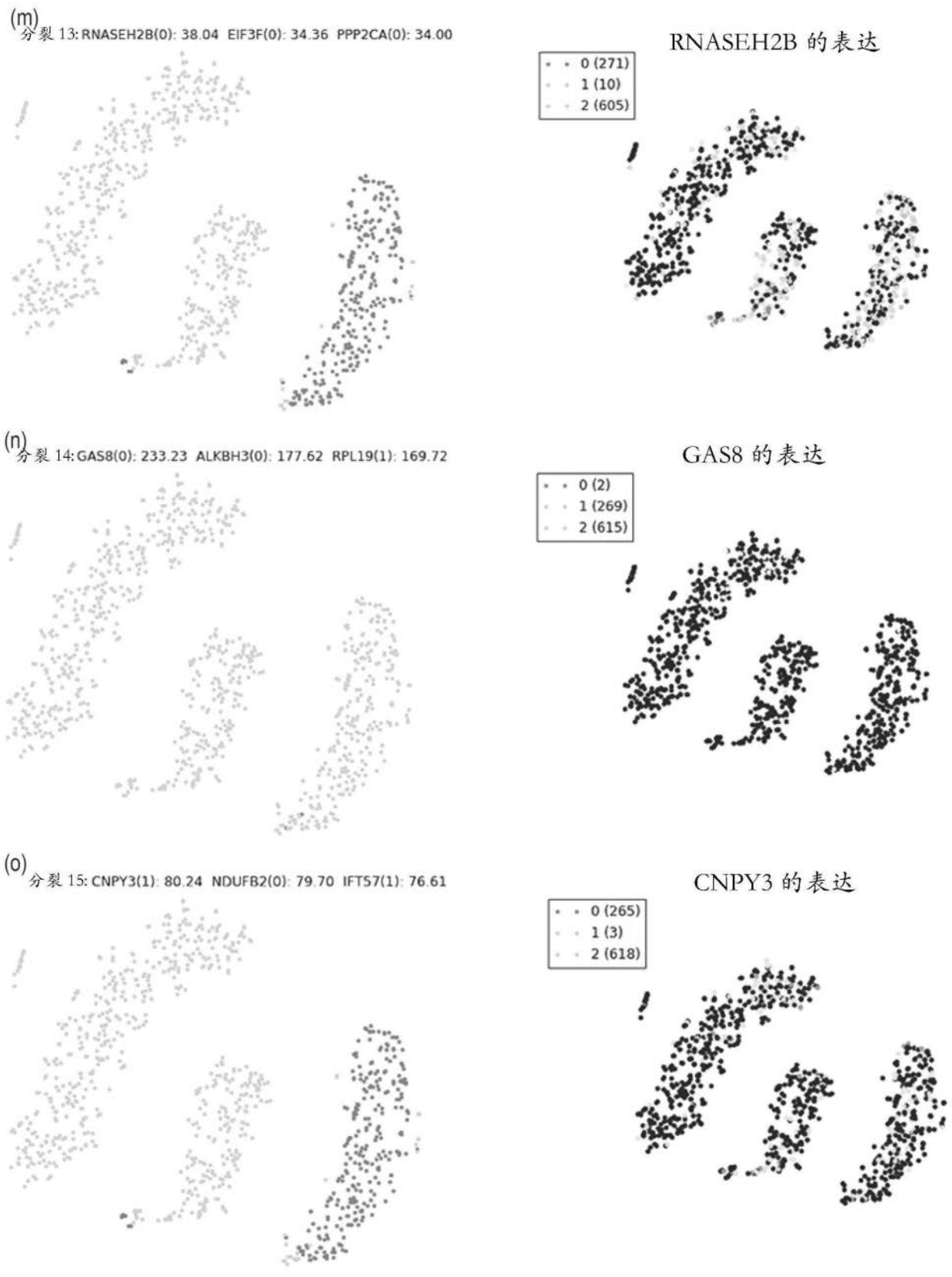

(m)
分裂 13: RNASEH2B(0): 38.04 EIF3F(0): 34.36 PPP2CA(0): 34.00
RNASEH2B 的表达
0 (271)
1 (10)
2 (605)
(n)
分裂 14: GAS8(0): 233.23 ALKBH3(0): 177.62 RPL19(1): 169.72
GAS8 的表达
0 (2)
1 (269)
2 (615)
(o)
分裂 15: CNPY3(1): 80.24 NDUFB2(0): 79.70 IFT57(1): 76.61
CNPY3 的表达
0 (265)
1 (3)
2 (618)

图9(续)

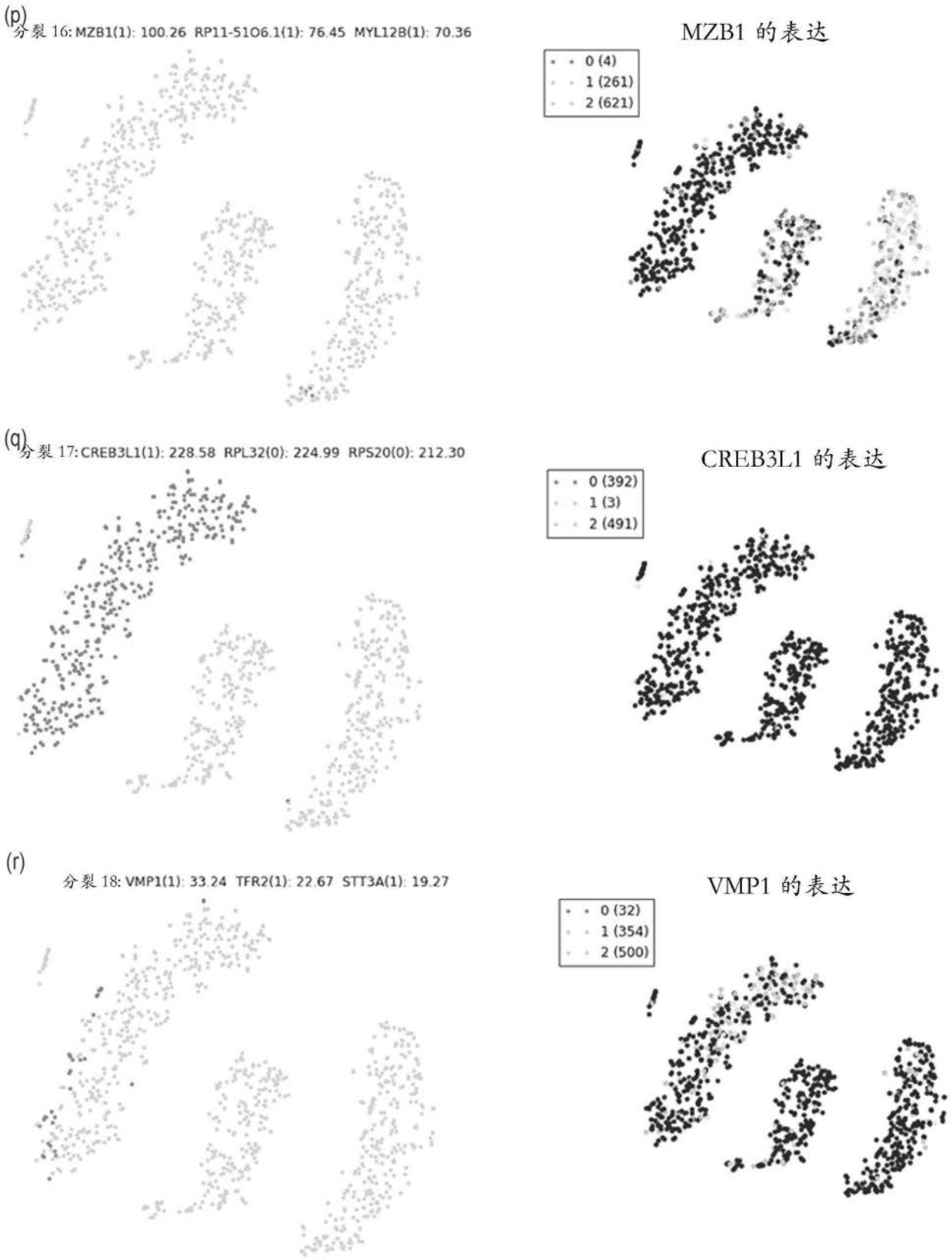

(p)
分裂 16: MZB1(1): 100.26 RP11-51O6.1(1): 76.45 MYL12B(1): 70.36
MZB1 的表达
0 (4)
1 (261)
2 (621)
(q)
分裂 17: CREB3L1(1): 228.58 RPL32(0): 224.99 RPS20(0): 212.30
CREB3L1 的表达
0 (392)
1 (3)
2 (491)
(r)
分裂 18: VMP1(1): 33.24 TFR2(1): 22.67 STT3A(1): 19.27
VMP1 的表达
0 (32)
1 (354)
2 (500)

图9(续)

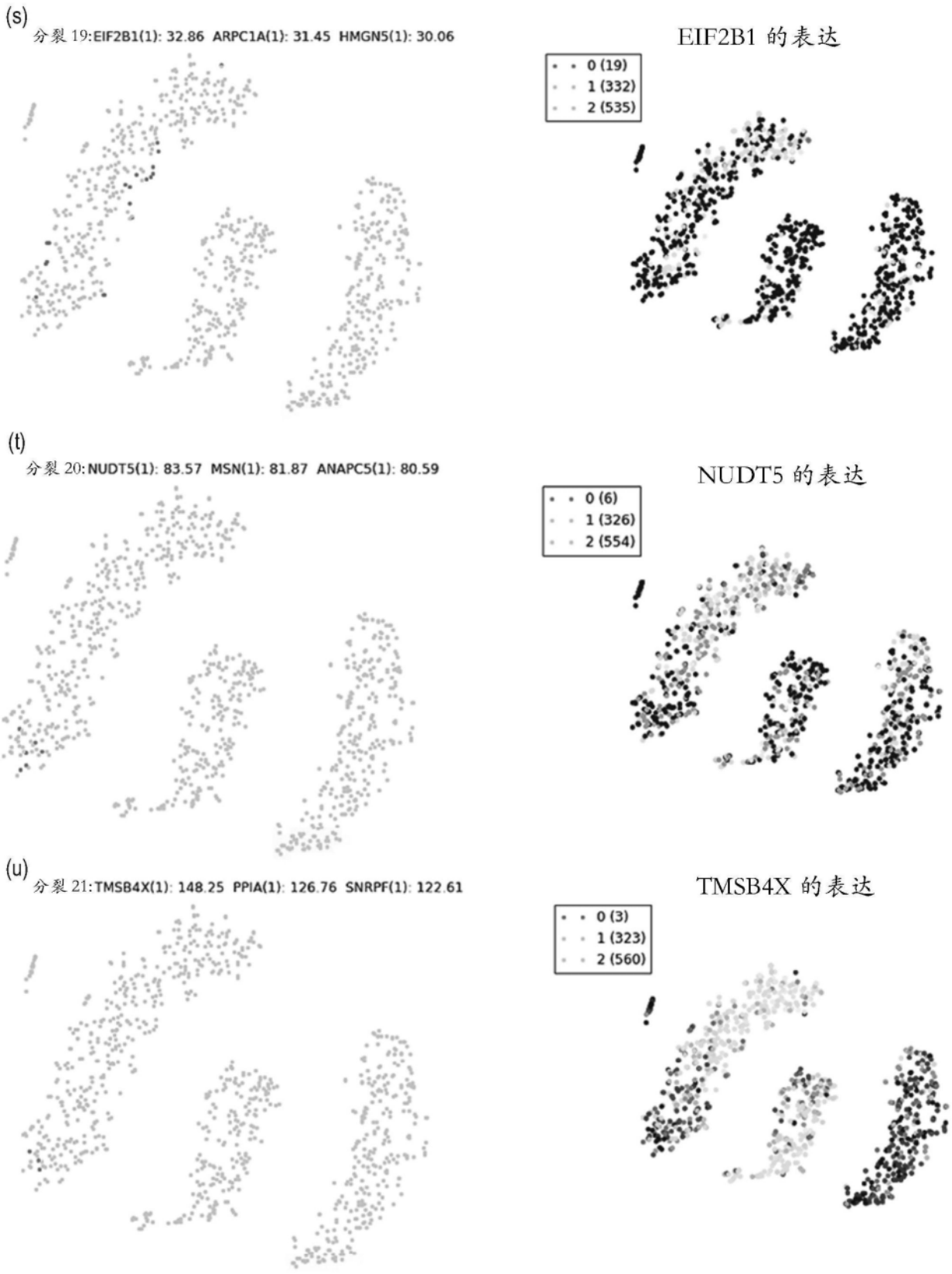

(s)
分裂 19:EIF2B1(1): 32.86 ARPC1A(1): 31.45 HMGN5(1): 30.06
EIF2B1 的表达
0 (19)
1 (332)
2 (535)
(t)
分裂 20:NUDT5(1): 83.57 MSN(1): 81.87 ANAPC5(1): 80.59
NUDT5 的表达
0 (6)
1 (326)
2 (554)
(u)
分裂 21:TMSB4X(1): 148.25 PPIA(1): 126.76 SNRPF(1): 122.61
TMSB4X 的表达
0 (3)
1 (323)
2 (560)

图9(续)

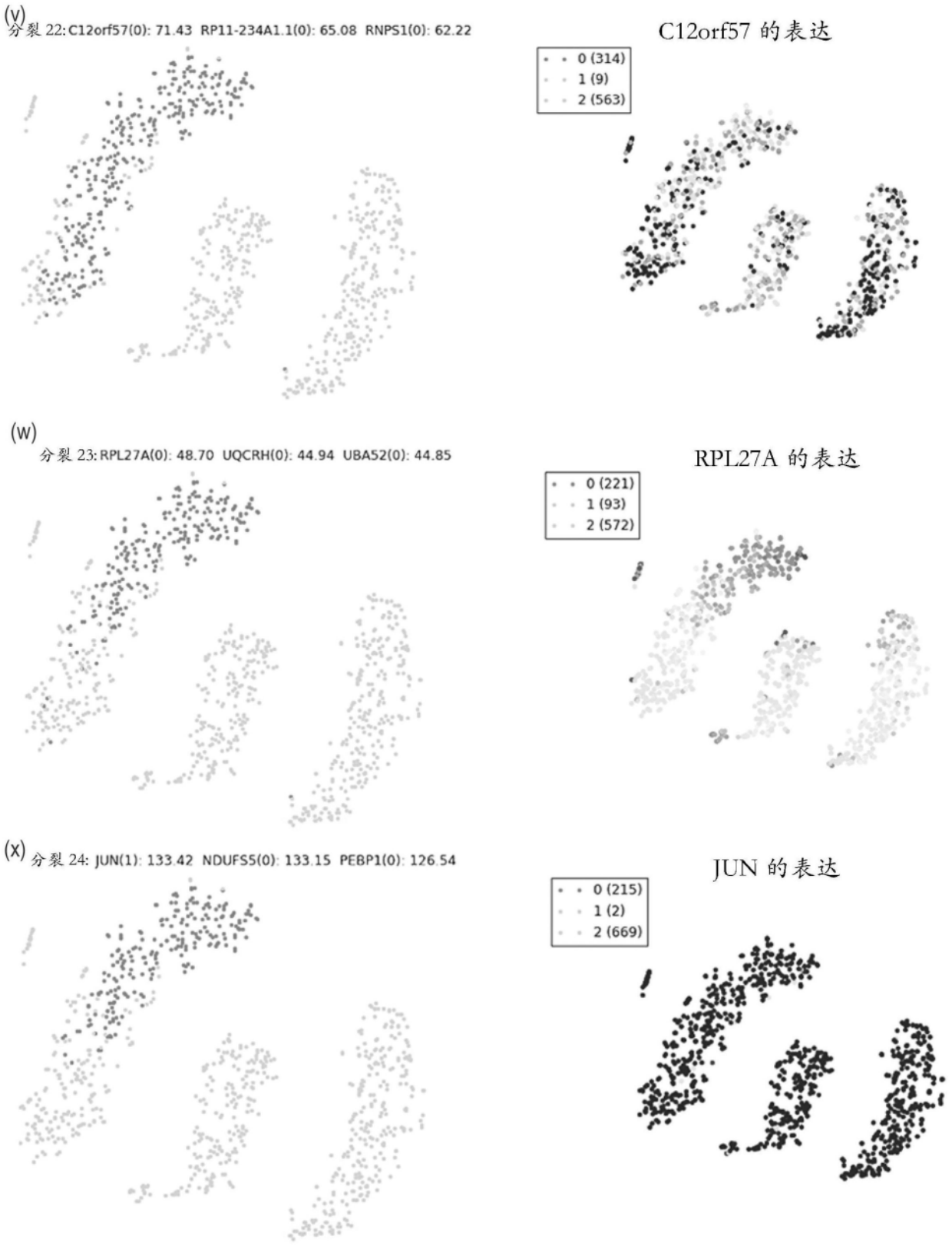

(V)
分裂 22: C12orf57(0): 71.43 RP11-234A1.1(0): 65.08 RNPS1(0): 62.22
C12orf57 的表达
0 (314)
1 (9)
2 (563)
(W)
分裂 23: RPL27A(0): 48.70 UQCRH(0): 44.94 UBA52(0): 44.85
RPL27A 的表达
0 (221)
1 (93)
2 (572)
(X)
分裂 24: JUN(1): 133.42 NDUFS5(0): 133.15 PEBP1(0): 126.54
JUN 的表达
0 (215)
1 (2)
2 (669)

图9(续)

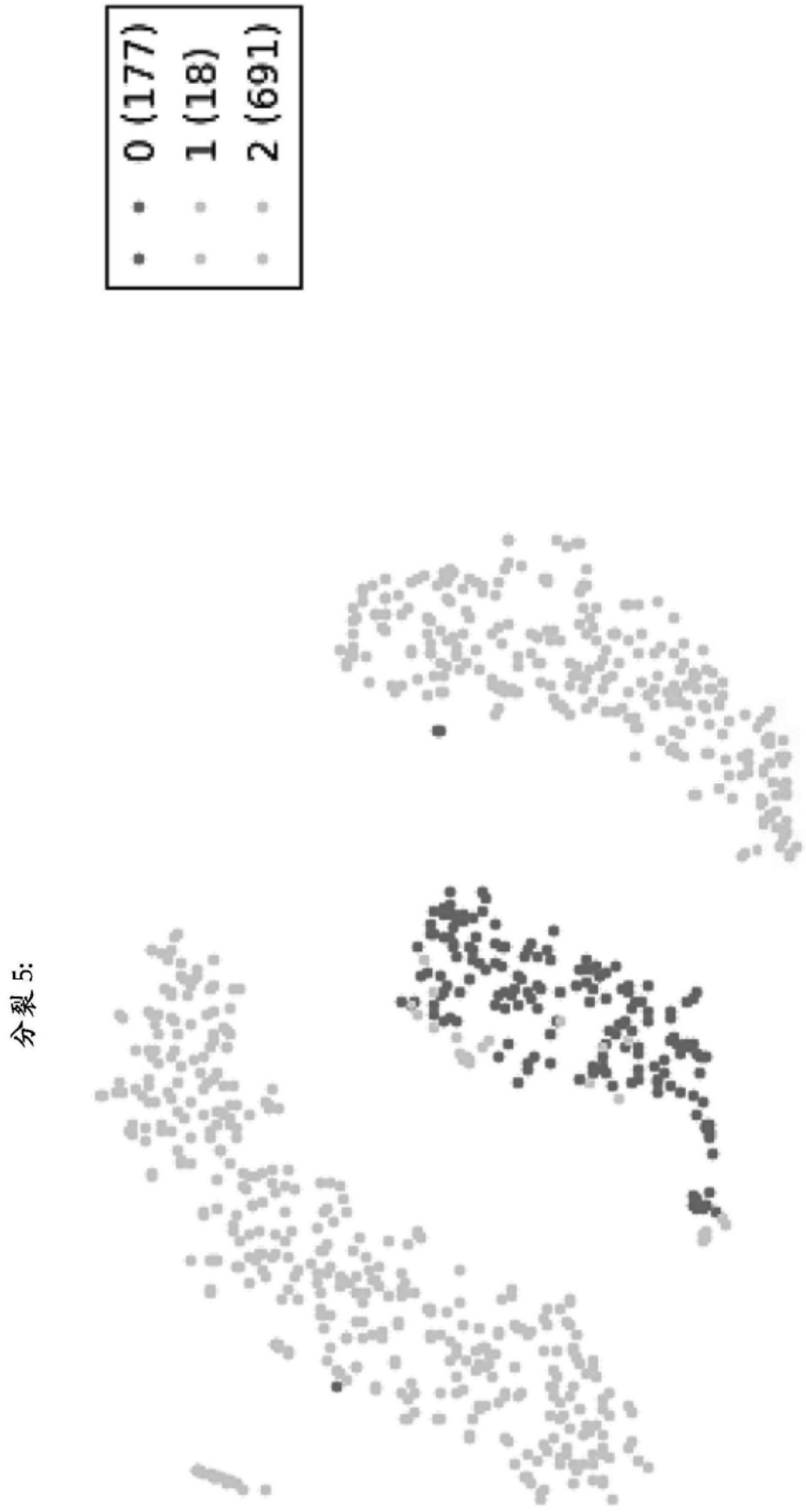


图10

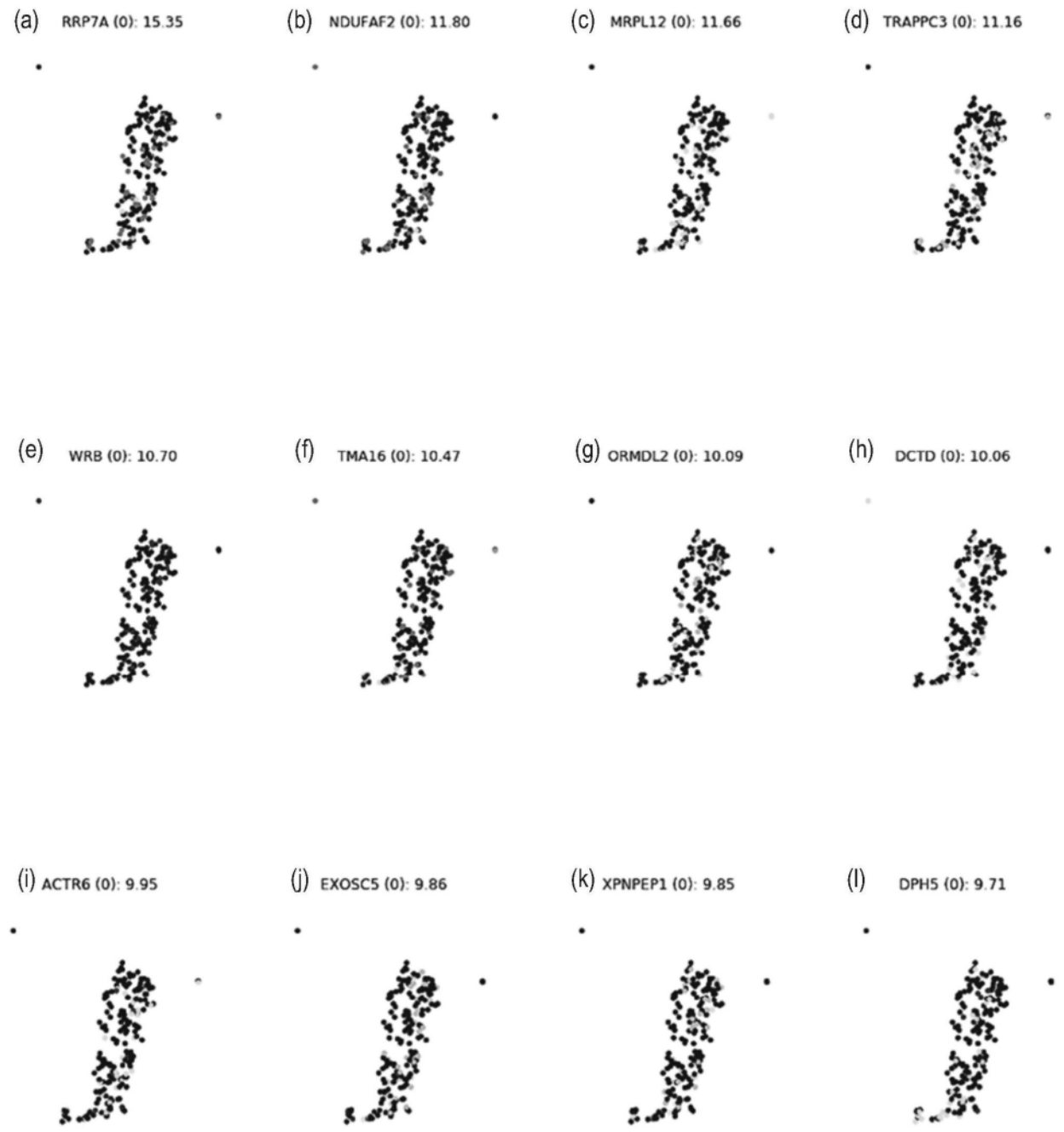

(a) RRP7A (0): 15.35
(b) NDUFAF2 (0): 11.80
(c) MRPL12 (0): 11.66
(d) TRAPPC3 (0): 11.16
(e) WRB (0): 10.70
(f) TMA16 (0): 10.47
(g) ORMDL2 (0): 10.09
(h) DCTD (0): 10.06
(i) ACTR6 (0): 9.95
(j) EXOSC5 (0): 9.86
(k) XPNPEP1 (0): 9.85
(l) DPH5 (0): 9.71

图11

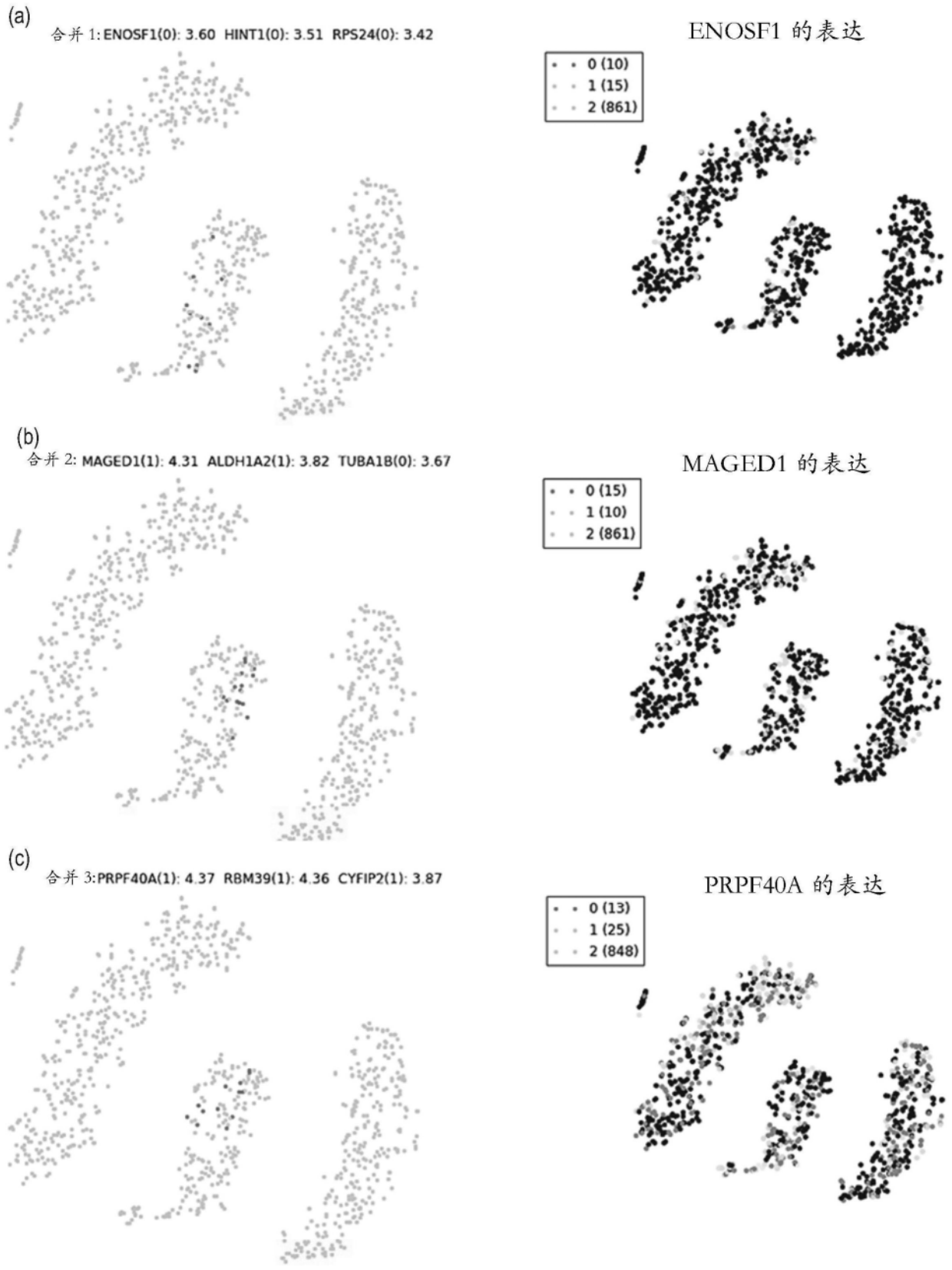

(a)
合并 1: ENOSF1(0): 3.60 HINT1(0): 3.51 RPS24(0): 3.42
ENOSF1 的表达
0 (10)
1 (15)
2 (861)
(b)
合并 2: MAGED1(1): 4.31 ALDH1A2(1): 3.82 TUBA1B(0): 3.67
MAGED1 的表达
0 (15)
1 (10)
2 (861)
(c)
合并 3: PRPF40A(1): 4.37 RBM39(1): 4.36 CYFIP2(1): 3.87
PRPF40A 的表达
0 (13)
1 (25)
2 (848)

图12

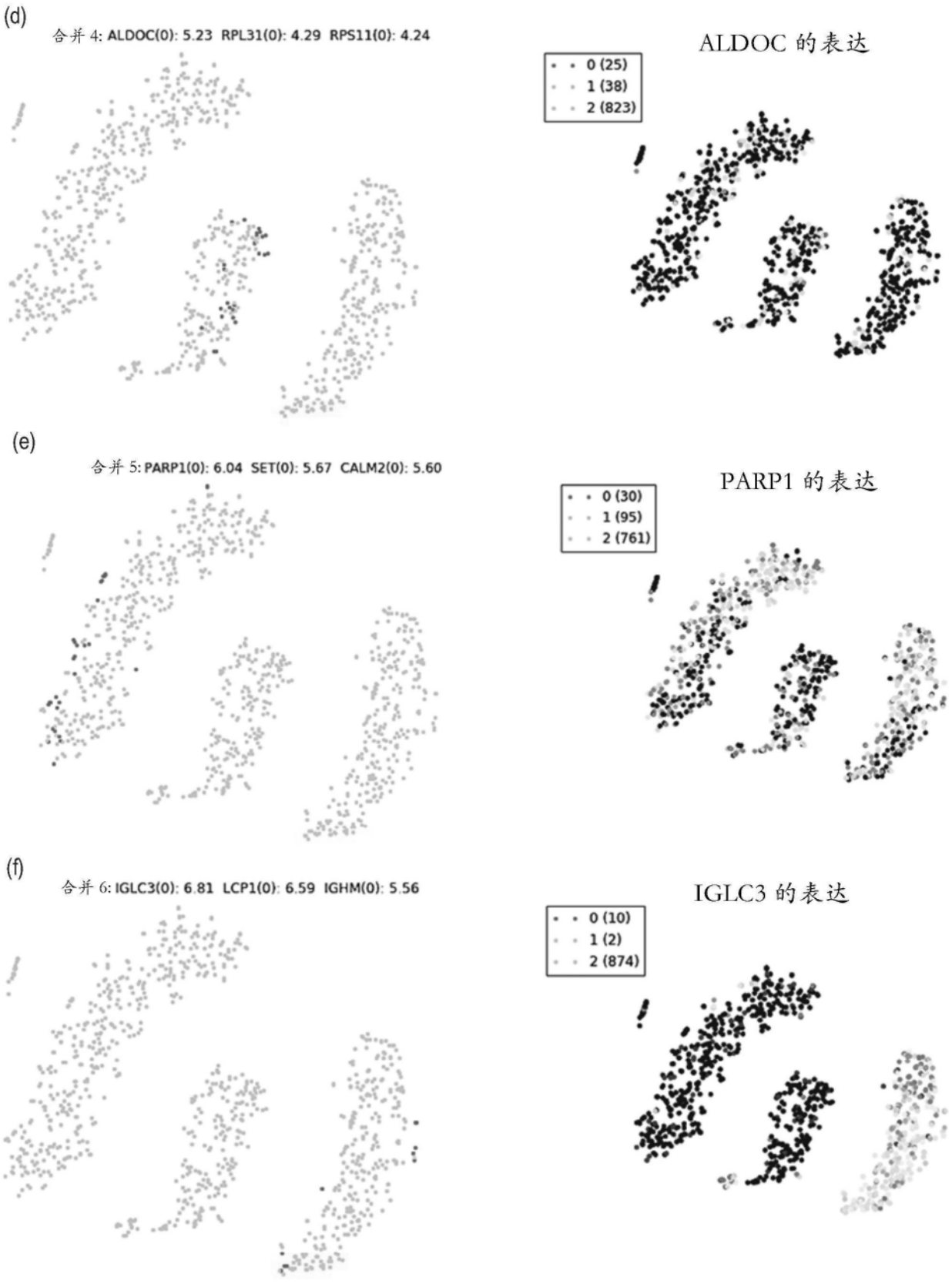

(d)
合并 4: ALDOC(0): 5.23 RPL31(0): 4.29 RPS11(0): 4.24
ALDOC 的表达
0 (25)
1 (38)
2 (823)
(e)
合并 5: PARP1(0): 6.04 SET(0): 5.67 CALM2(0): 5.60
PARP1 的表达
0 (30)
1 (95)
2 (761)
(f)
合并 6: IGLC3(0): 6.81 LCP1(0): 6.59 IGHM(0): 5.56
IGLC3 的表达
0 (10)
1 (2)
2 (874)

图12(续)

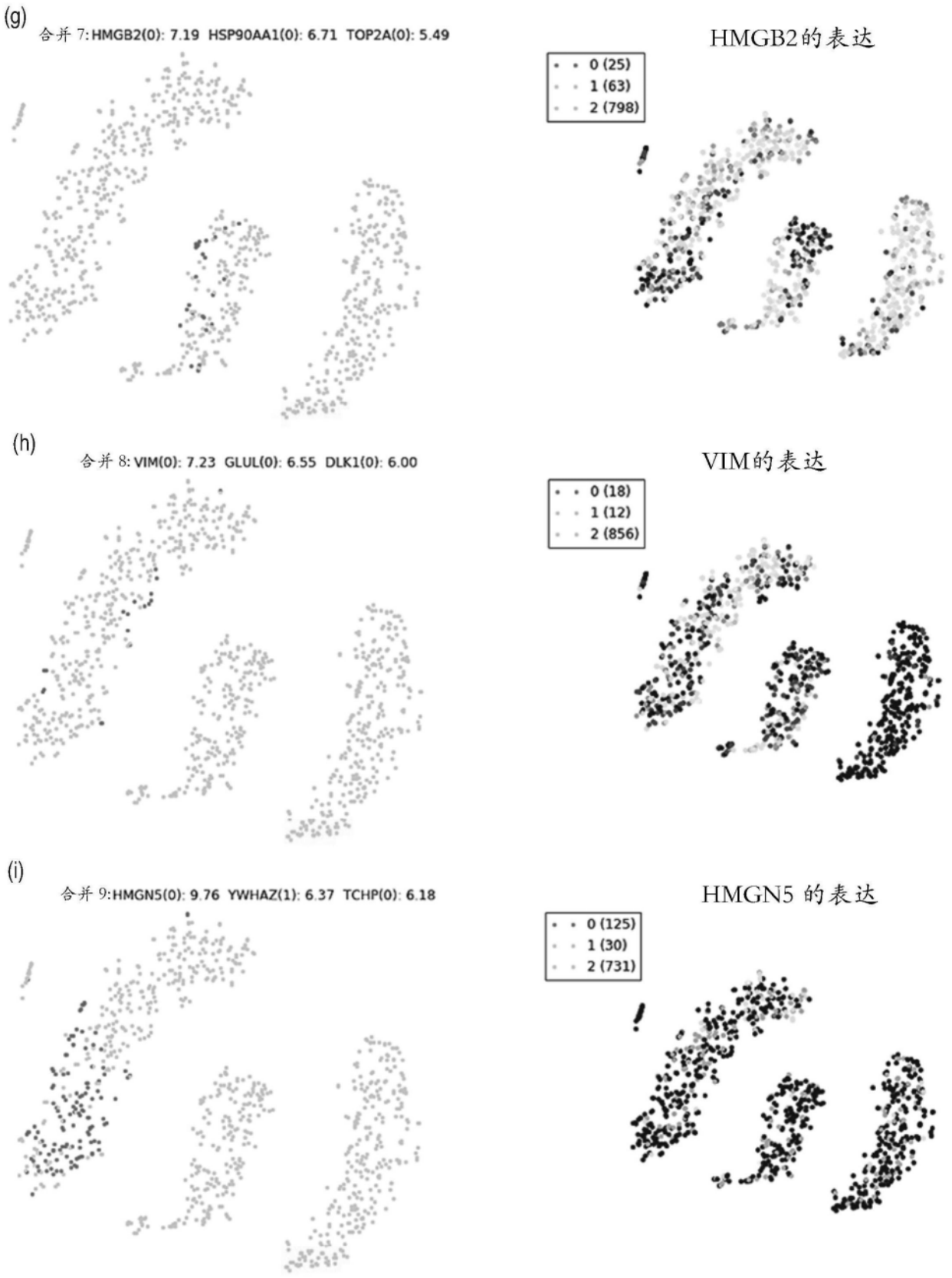

(g)
合并 7: HMGB2(0): 7.19 HSP90AA1(0): 6.71 TOP2A(0): 5.49
HMGB2的表达
0 (25)
1 (63)
2 (798)
(h)
合并 8: VIM(0): 7.23 GLUL(0): 6.55 DLK1(0): 6.00
VIM的表达
0 (18)
1 (12)
2 (856)
(i)
合并 9: HMGN5(0): 9.76 YWHAZ(1): 6.37 TCHP(0): 6.18
HMGN5 的表达
0 (125)
1 (30)
2 (731)

图12(续)

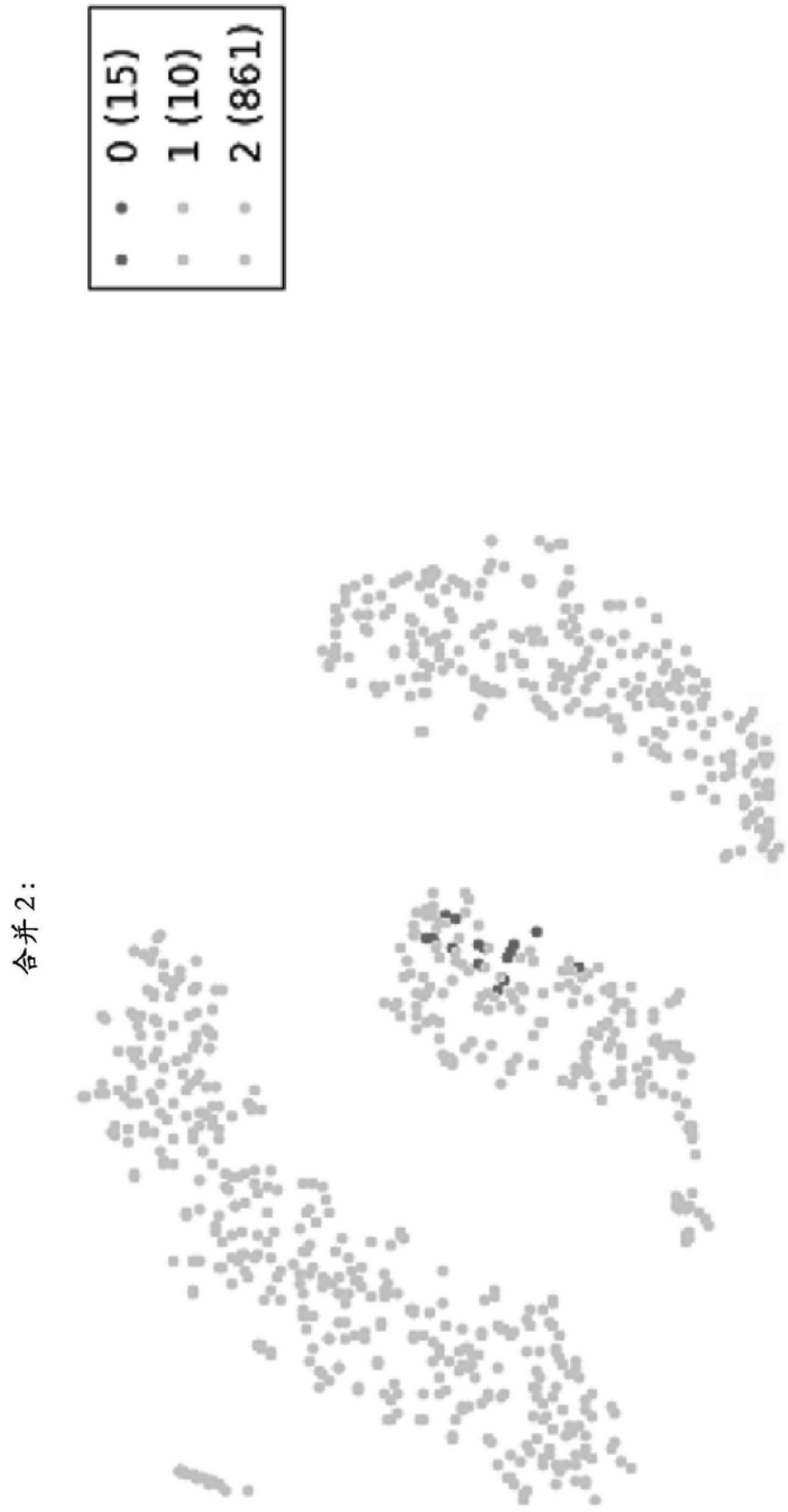


图13

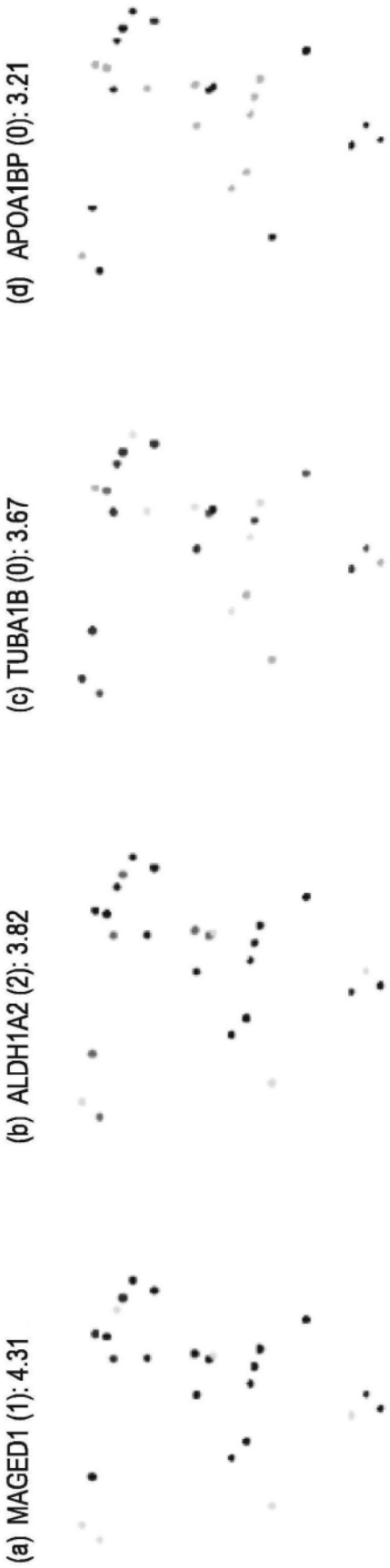


图14

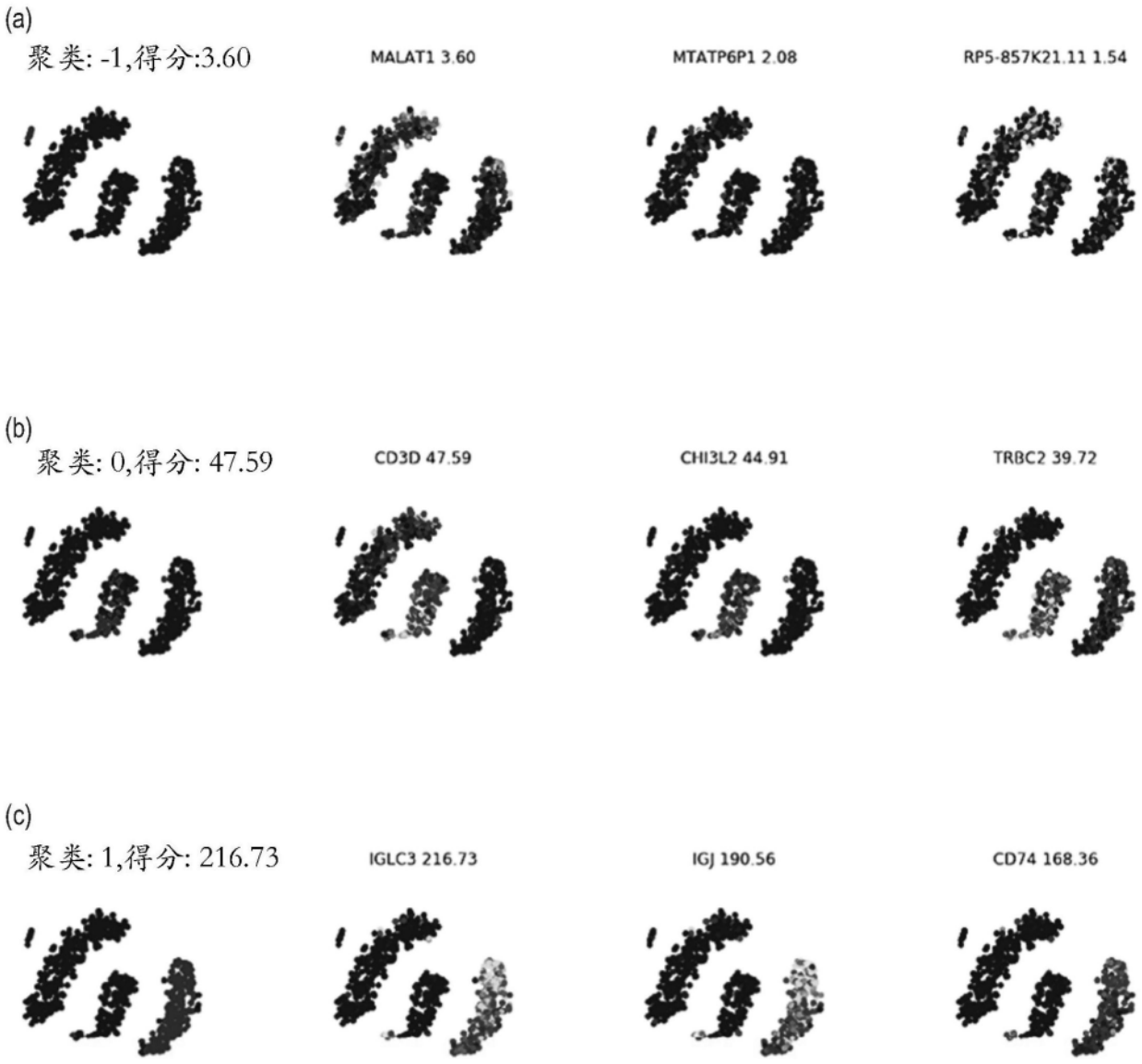

(a)
聚类: -1,得分:3.60
MALAT1 3.60
MTATP6P1 2.08
RP5-857K21.11 1.54
(b)
聚类: 0,得分: 47.59
CD3D 47.59
CHI3L2 44.91
TRBC2 39.72
(c)
聚类: 1,得分: 216.73
IGLC3 216.73
IGJ 190.56
CD74 168.36

图15


(d)
聚类: 2,得分: 137.51
FTL 137.51
HBG1 120.99
UQCRH 114.51
(e)
聚类: 3,得分: 32.53
CD3D 32.53
ADA 28.54
CHI3L2 28.50
(f)
聚类: 4,得分: 18.15
FTL 18.15
PRAME 14.63
KRT18 10.43

图15(续)

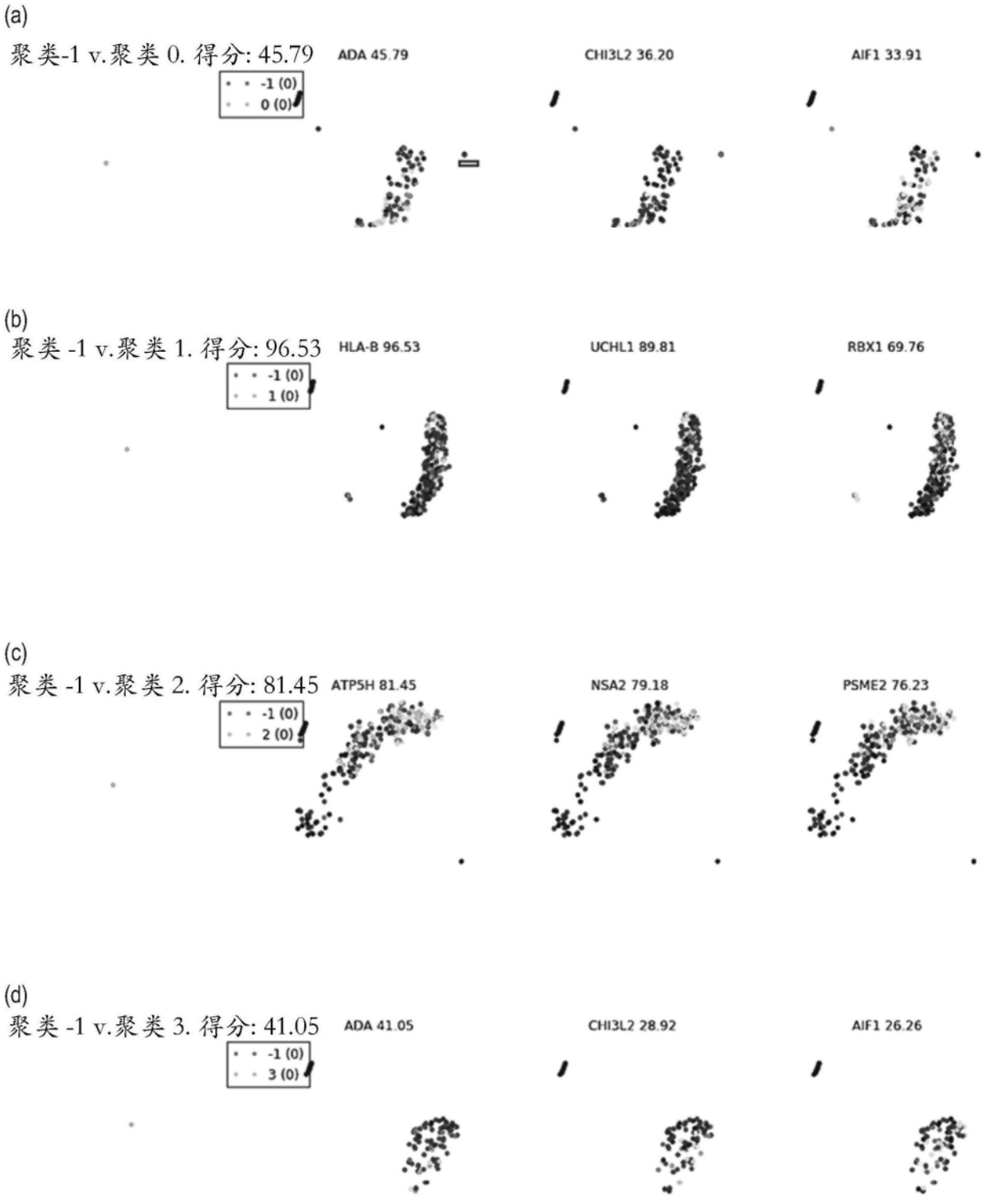

(a)
聚类-1 v.聚类 0. 得分: 45.79
-1 (0)
0 (0)
ADA 45.79
CHI3L2 36.20
AIF1 33.91
(b)
聚类 -1 v.聚类 1. 得分: 96.53
-1 (0)
1 (0)
HLA-B 96.53
UCHL1 89.81
RBX1 69.76
(c)
聚类 -1 v.聚类 2. 得分: 81.45
-1 (0)
2 (0)
ATP5H 81.45
NSA2 79.18
PSME2 76.23
(d)
聚类 -1 v.聚类 3. 得分: 41.05
-1 (0)
3 (0)
ADA 41.05
CHI3L2 28.92
AIF1 26.26

图16

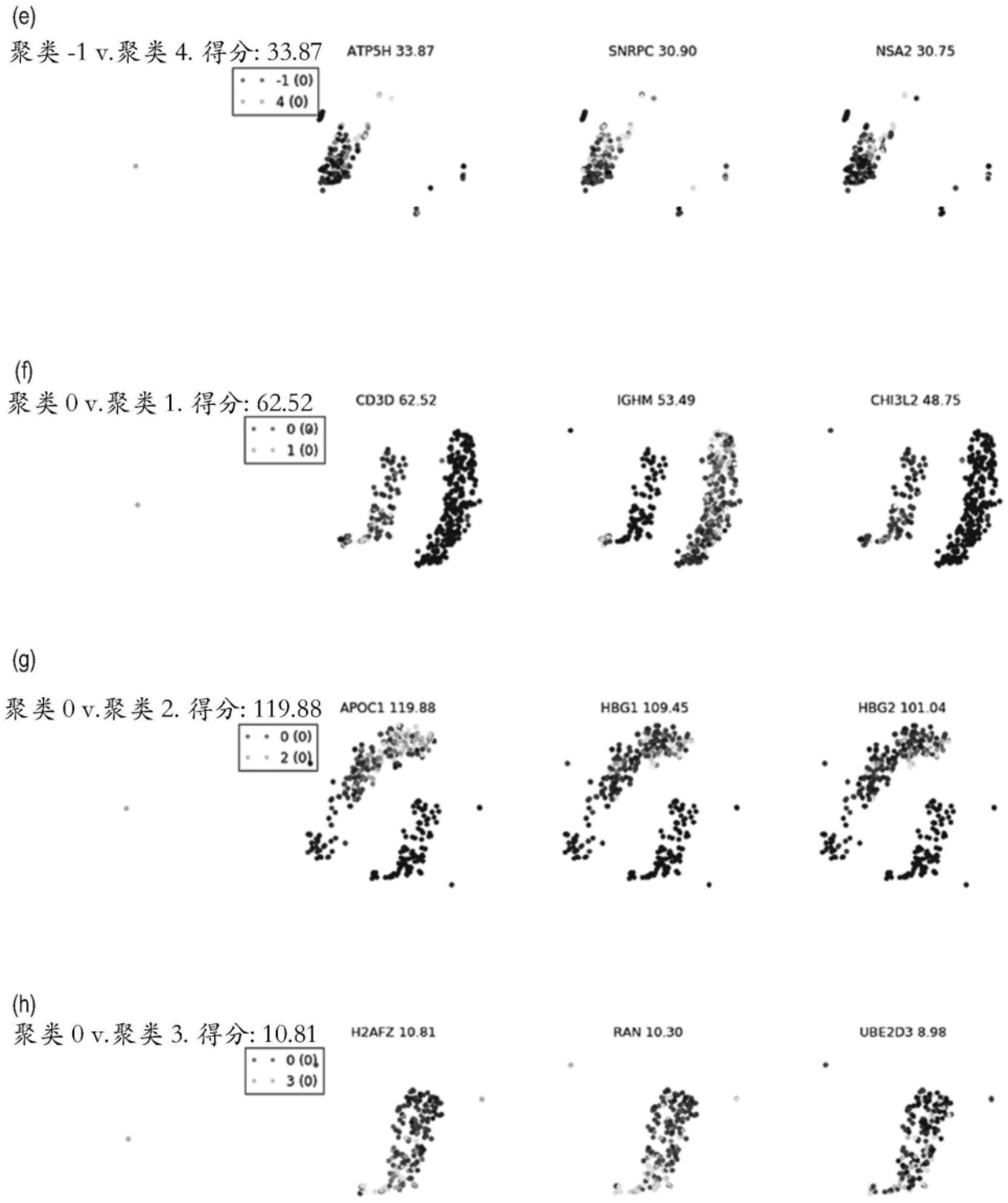

(e)
聚类 -1 v.聚类 4. 得分: 33.87
-1 (0)
4 (0)
ATP5H 33.87
SNRPC 30.90
NSA2 30.75
(f)
聚类 0 v.聚类 1. 得分: 62.52
0 (0)
1 (0)
CD3D 62.52
IGHM 53.49
CHI3L2 48.75
(g)
聚类 0 v.聚类 2. 得分: 119.88
0 (0)
2 (0)
APOC1 119.88
HBG1 109.45
HBG2 101.04
(h)
聚类 0 v.聚类 3. 得分: 10.81
0 (0)
3 (0)
H2AFZ 10.81
RAN 10.30
UBE2D3 8.98

图16(续)

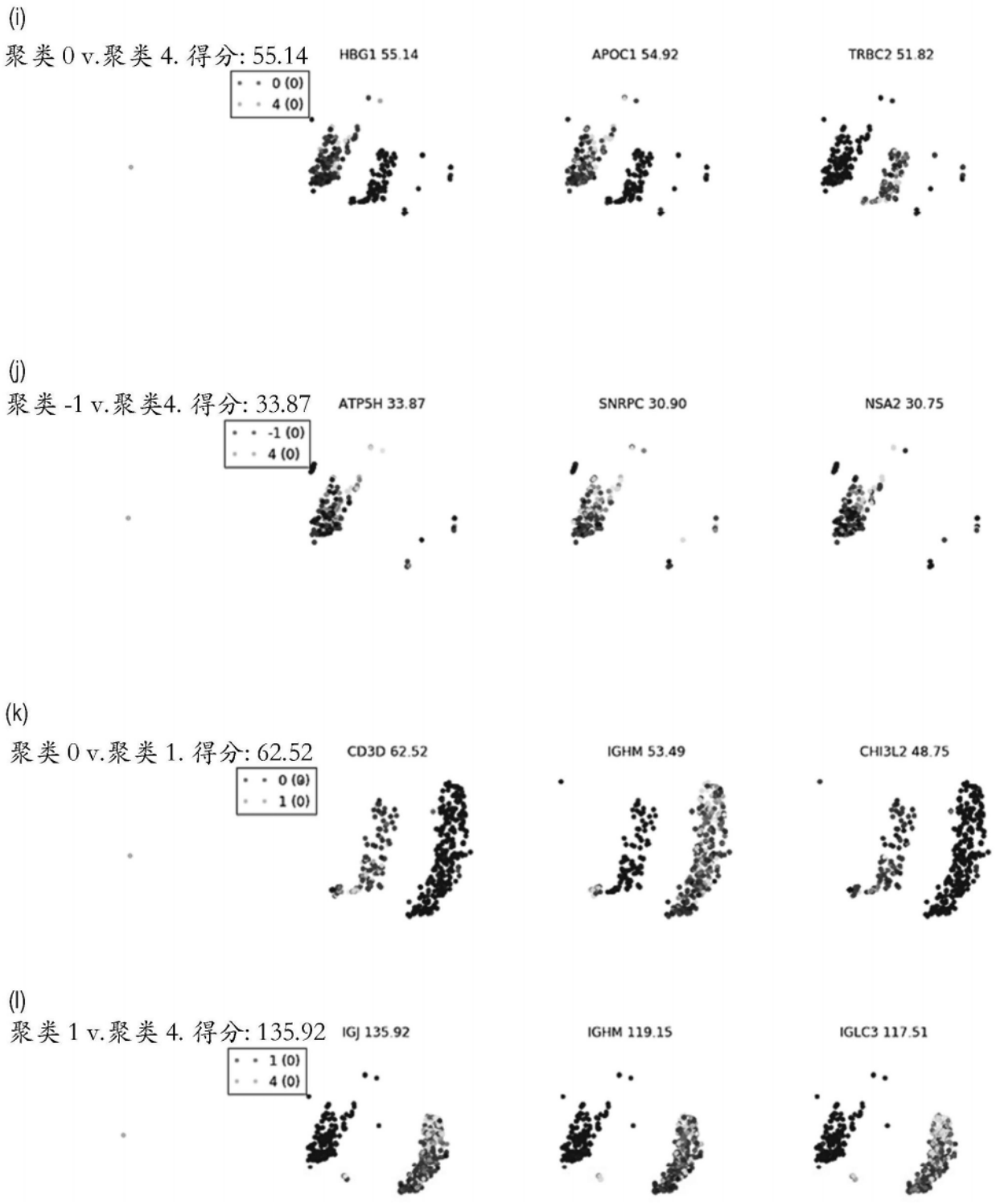

(i)
聚类 0 v.聚类 4. 得分: 55.14
HBG1 55.14
APOC1 54.92
TRBC2 51.82
0 (0)
4 (0)
(j)
聚类 -1 v.聚类4. 得分: 33.87
ATP5H 33.87
SNRPC 30.90
NSA2 30.75
-1 (0)
4 (0)
(k)
聚类 0 v.聚类 1. 得分: 62.52
CD3D 62.52
IGHM 53.49
CHI3L2 48.75
0 (0)
1 (0)
(l)
聚类 1 v.聚类 4. 得分: 135.92
IGJ 135.92
IGHM 119.15
IGLC3 117.51
1 (0)
4 (0)

图16(续)

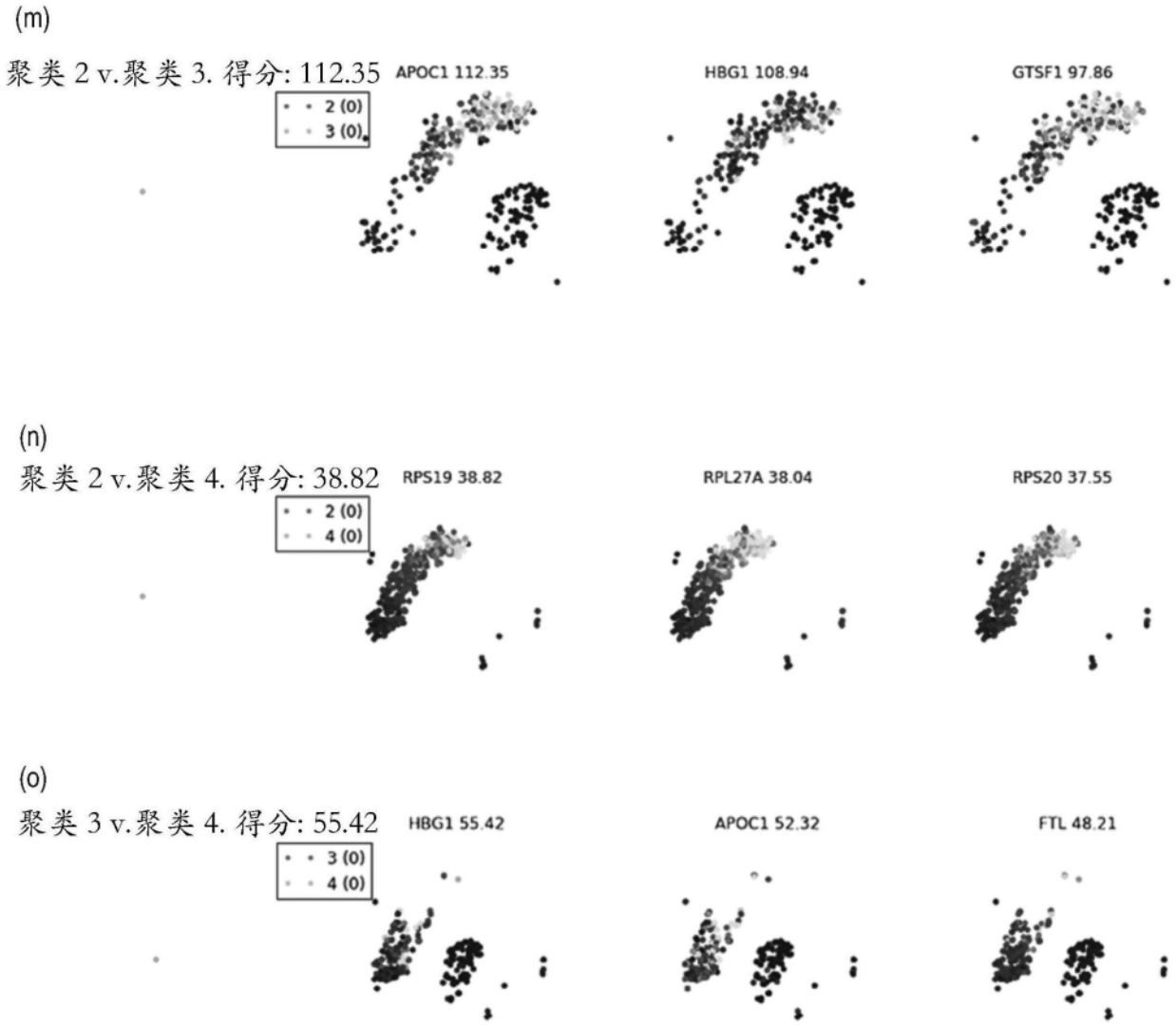

(m)
聚类 2 v.聚类 3. 得分: 112.35
APOC1 112.35
HBG1 108.94
GTSF1 97.86
2 (0)
3 (0)
(n)
聚类 2 v.聚类 4. 得分: 38.82
RPS19 38.82
RPL27A 38.04
RPS20 37.55
2 (0)
4 (0)
(o)
聚类 3 v.聚类 4. 得分: 55.42
HBG1 55.42
APOC1 52.32
FTL 48.21
3 (0)
4 (0)

图16(续)

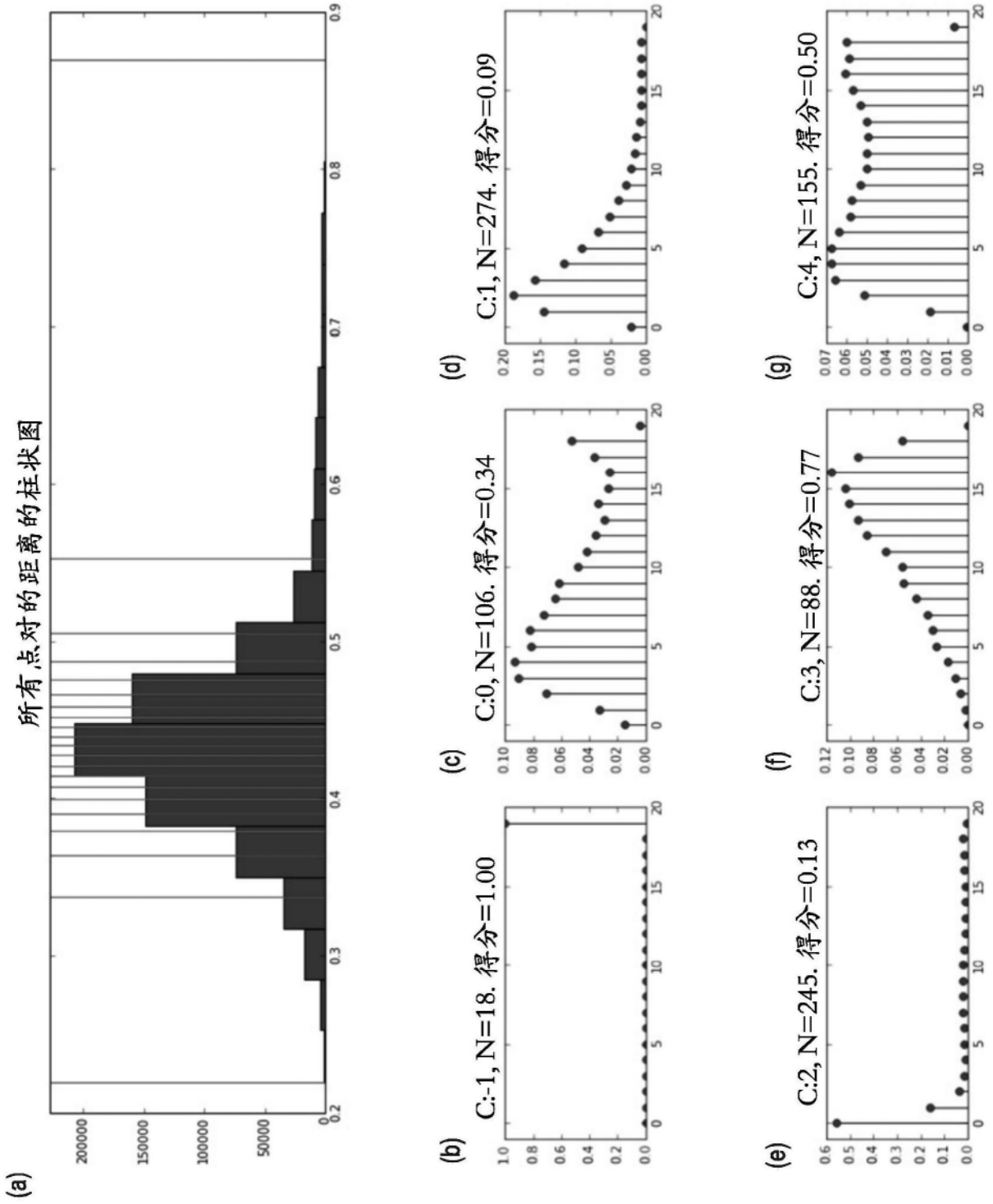

(a)
所有点对的距离的柱状图
(b)
C:-1, N=18. 得分=1.00
(c)
C:0, N=106. 得分=0.34
(d)
C:1, N=274. 得分=0.09
(e)
C:2, N=245. 得分=0.13
(f)
C:3, N=88. 得分=0.77
(g)
C:4, N=155. 得分=0.50

图17

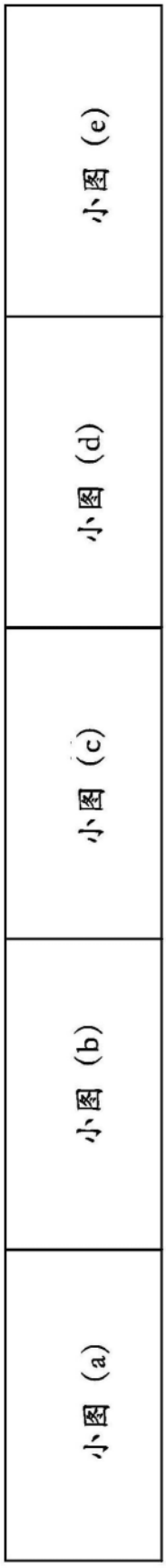


图18



图18(续)

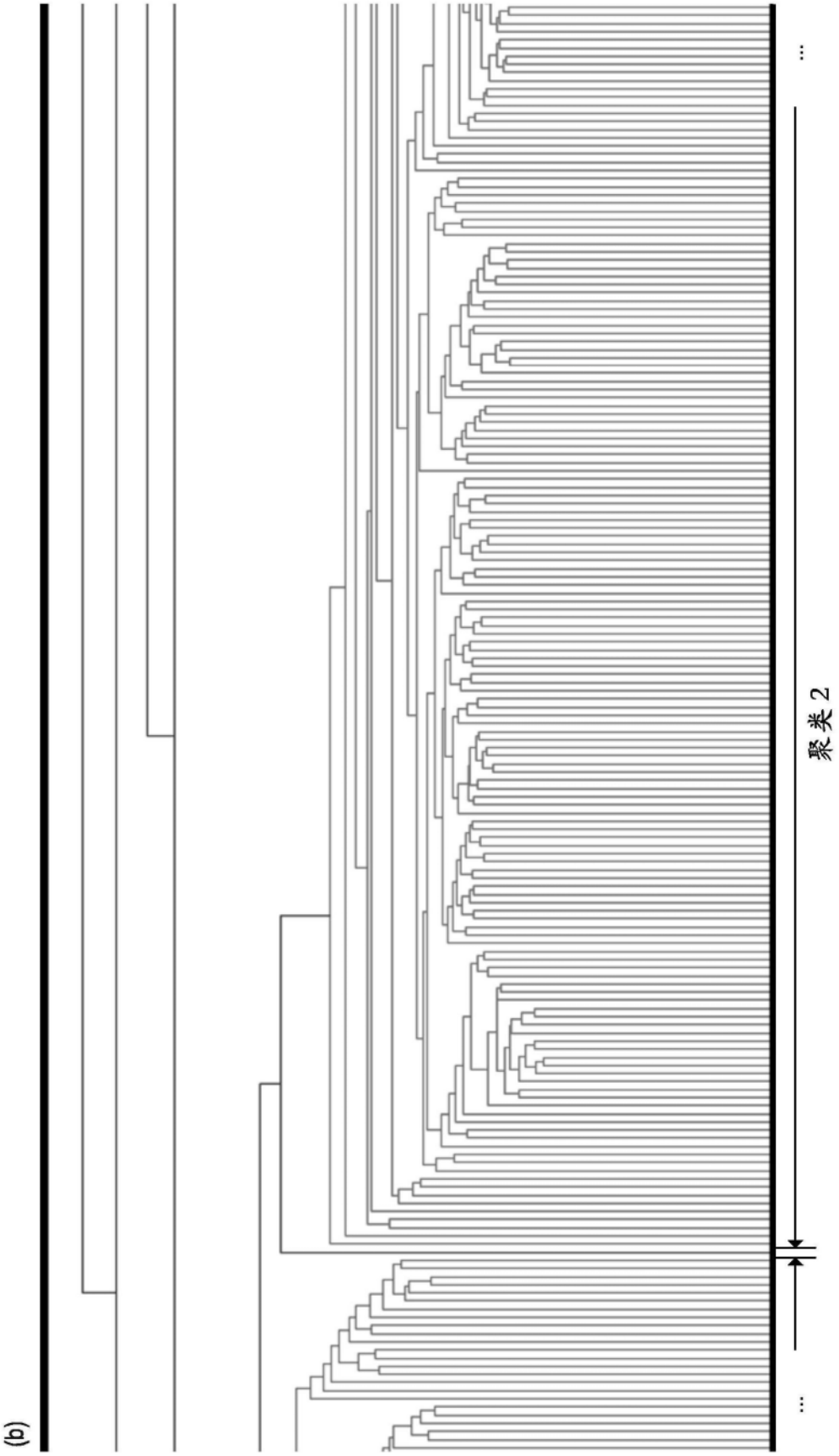


图18(续)

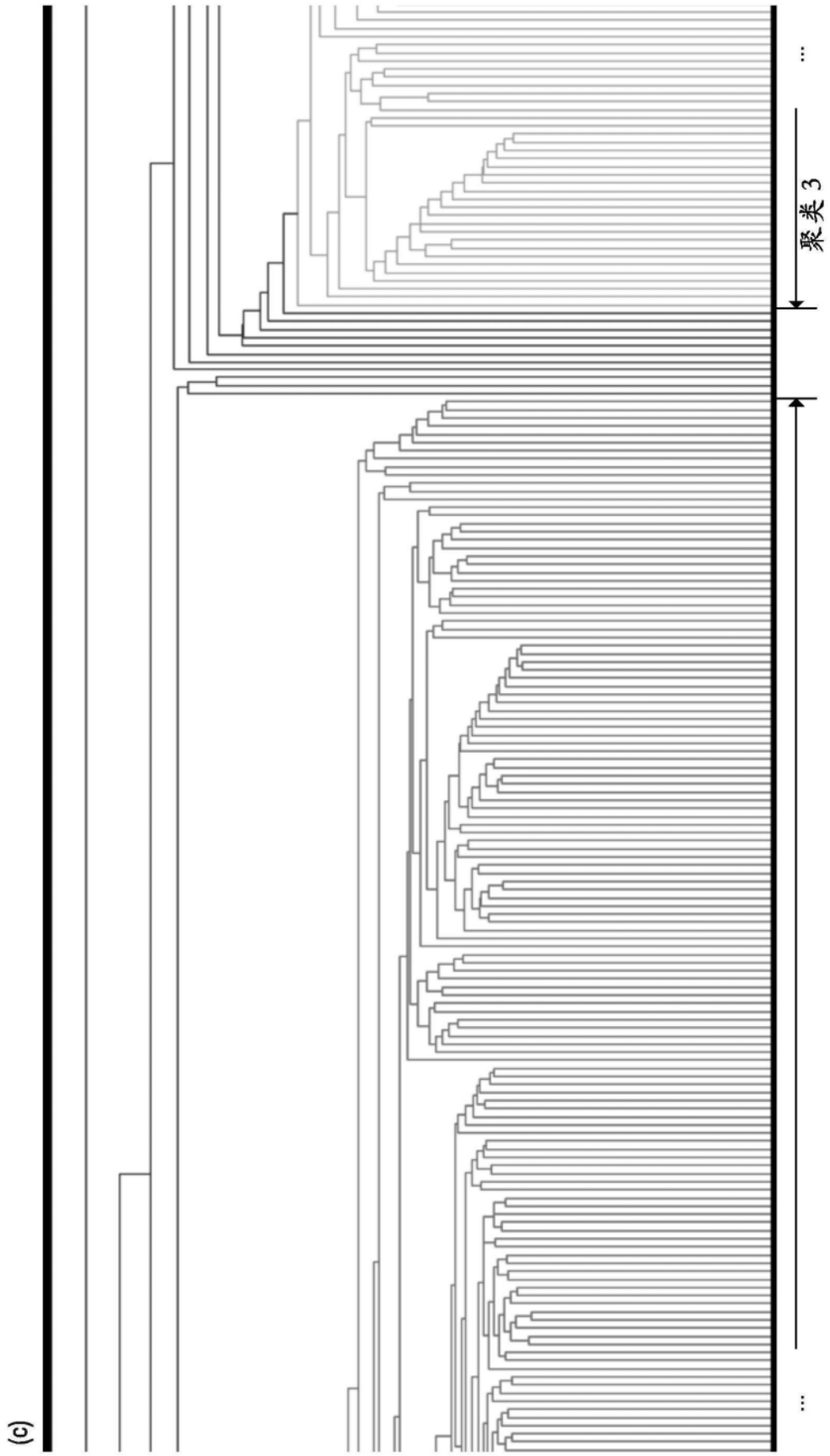


图18(续)



图18(续)

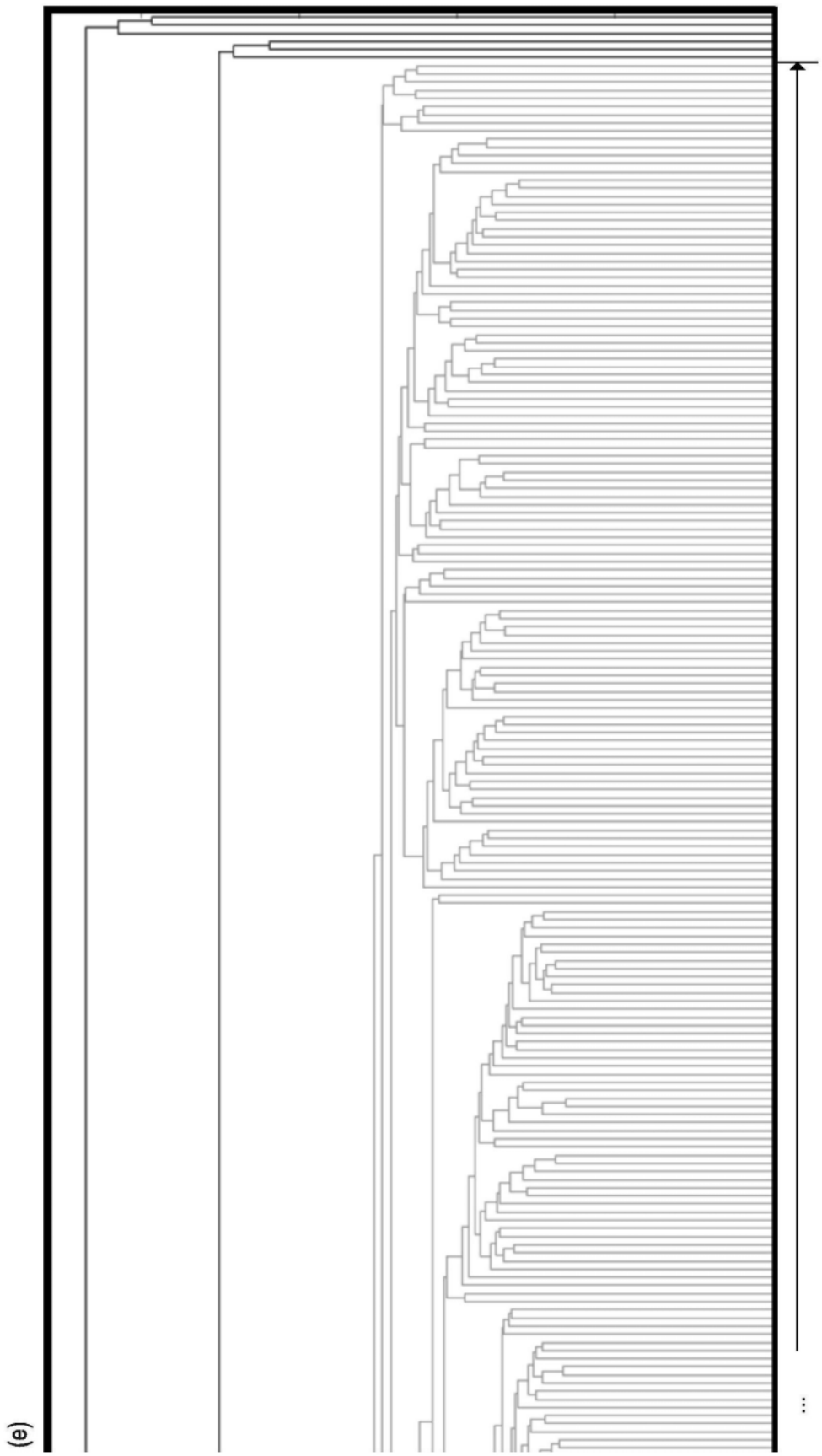


图18(续)

(a)

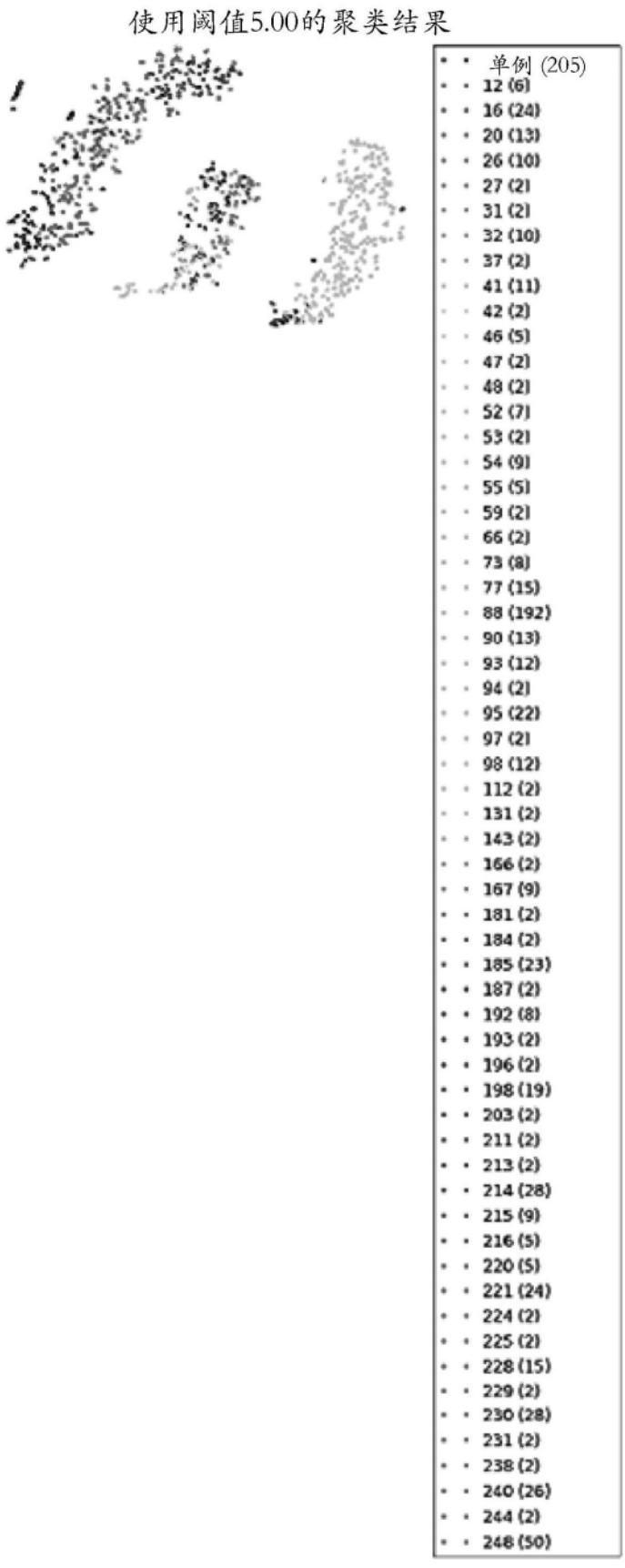
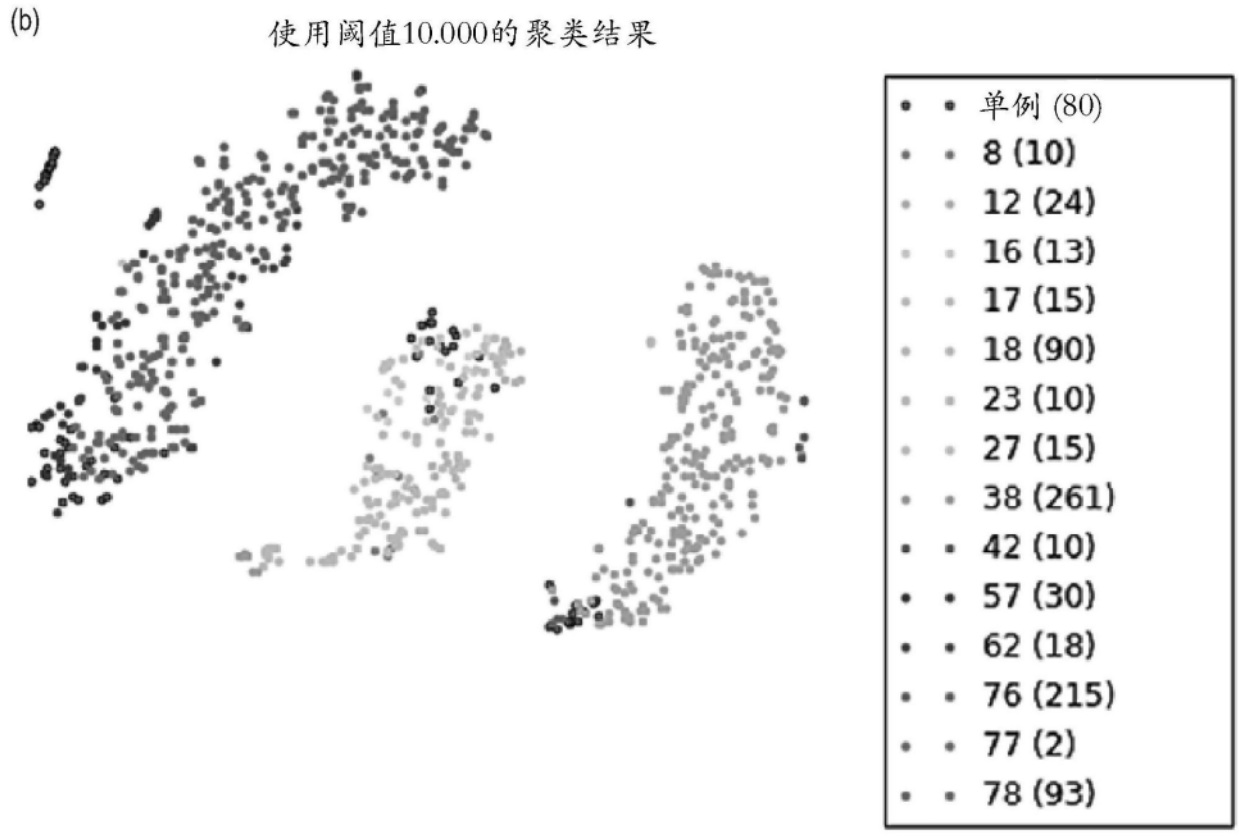

使用阈值5.00的聚类结果
单例 (205)
12 (6)
16 (24)
20 (13)
26 (10)
27 (2)
31 (2)
32 (10)
37 (2)
41 (11)
42 (2)
46 (5)
47 (2)
48 (2)
52 (7)
53 (2)
54 (9)
55 (5)
59 (2)
66 (2)
73 (8)
77 (15)
88 (192)
90 (13)
93 (12)
94 (2)
95 (22)
97 (2)
98 (12)
112 (2)
131 (2)
143 (2)
166 (2)
167 (9)
181 (2)
184 (2)
185 (23)
187 (2)
192 (8)
193 (2)
196 (2)
198 (19)
203 (2)
211 (2)
213 (2)
214 (28)
215 (9)
216 (5)
220 (5)
221 (24)
224 (2)
225 (2)
228 (15)
229 (2)
230 (28)
231 (2)
238 (2)
240 (26)
244 (2)
248 (50)

图19


(b)
使用阈值10.000的聚类结果
单例 (80)
8 (10)
12 (24)
16 (13)
17 (15)
18 (90)
23 (10)
27 (15)
38 (261)
42 (10)
57 (30)
62 (18)
76 (215)
77 (2)
78 (93)

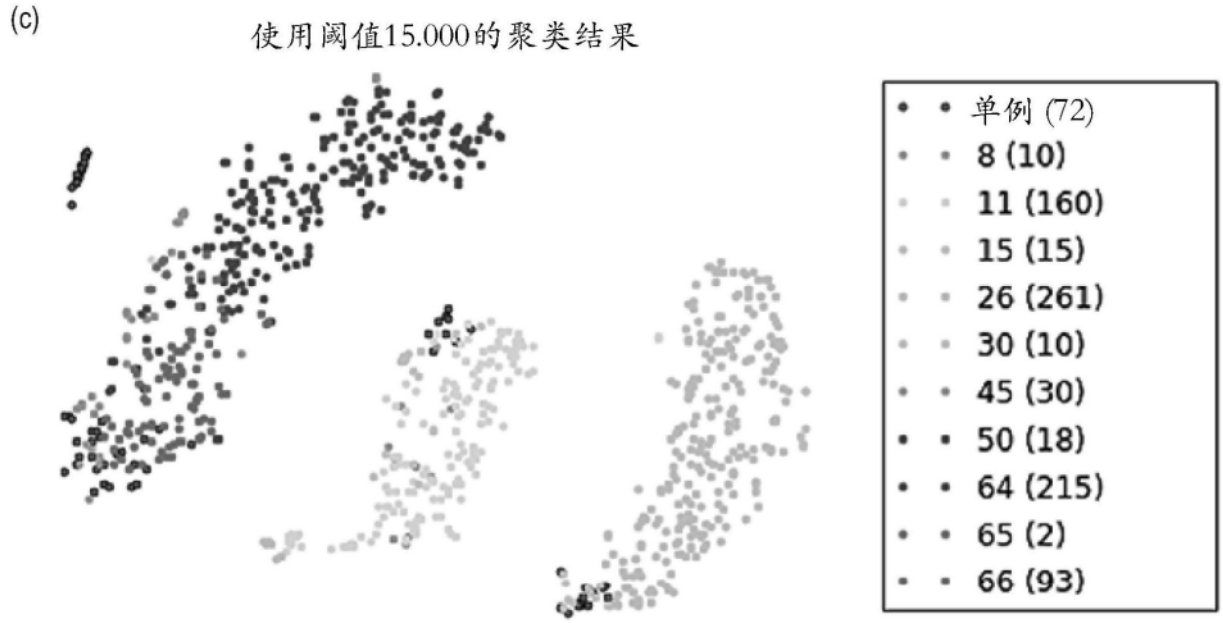
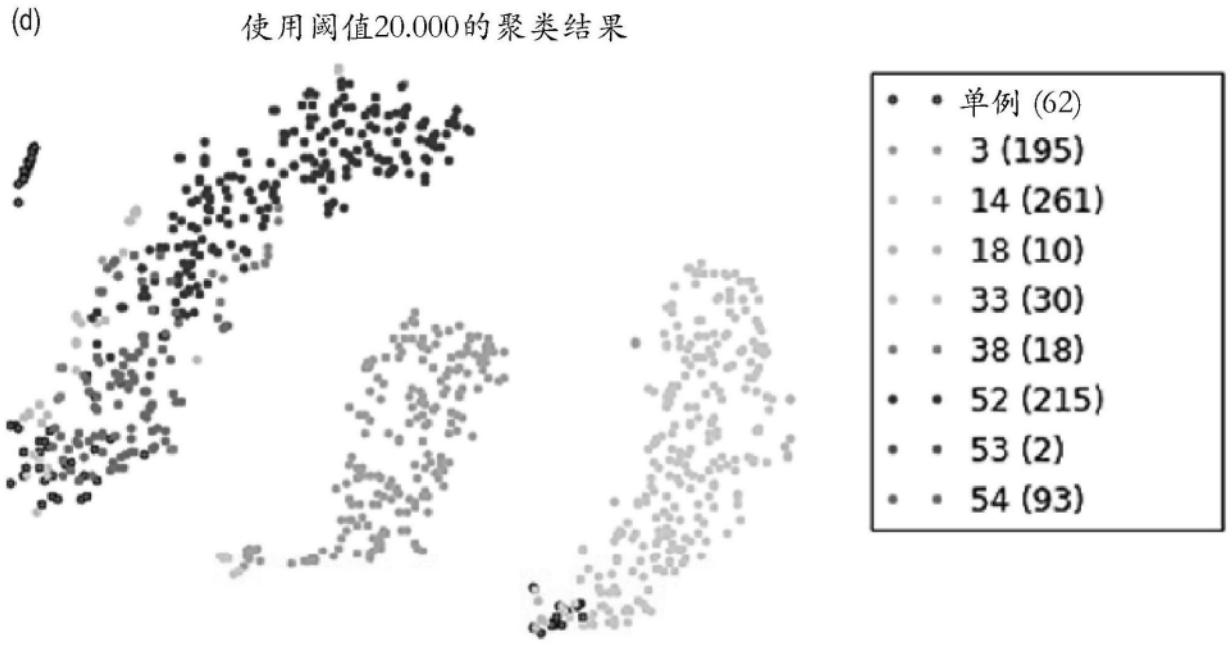

(c)
使用阈值15.000的聚类结果
单例 (72)
8 (10)
11 (160)
15 (15)
26 (261)
30 (10)
45 (30)
50 (18)
64 (215)
65 (2)
66 (93)

图19(续)


(d)
使用阈值20.000的聚类结果
单例 (62)
3 (195)
14 (261)
18 (10)
33 (30)
38 (18)
52 (215)
53 (2)
54 (93)

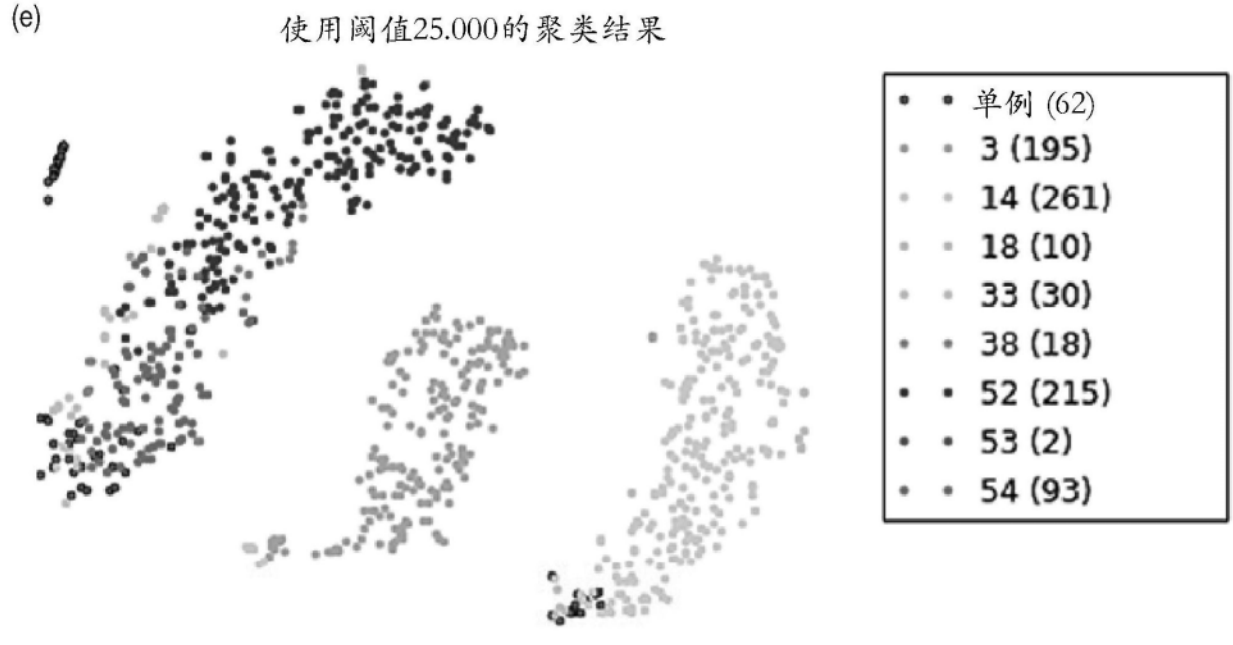
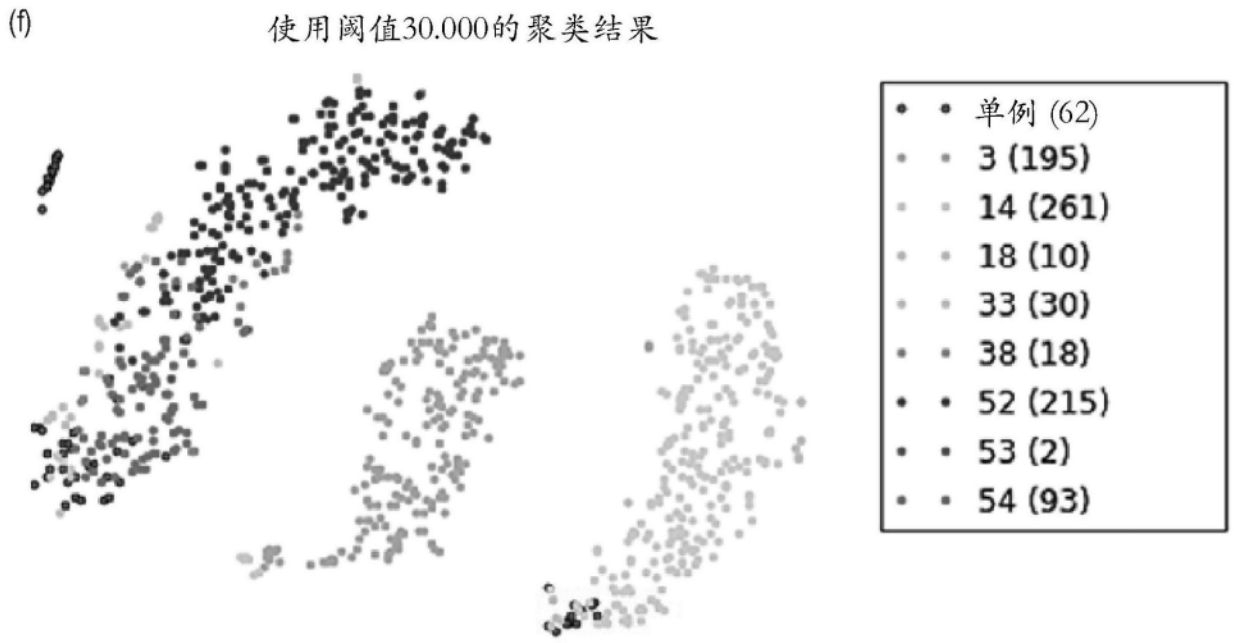

(e)
使用阈值25.000的聚类结果
单例 (62)
3 (195)
14 (261)
18 (10)
33 (30)
38 (18)
52 (215)
53 (2)
54 (93)

图19(续)


(f)
使用阈值30.000的聚类结果
单例 (62)
3 (195)
14 (261)
18 (10)
33 (30)
38 (18)
52 (215)
53 (2)
54 (93)

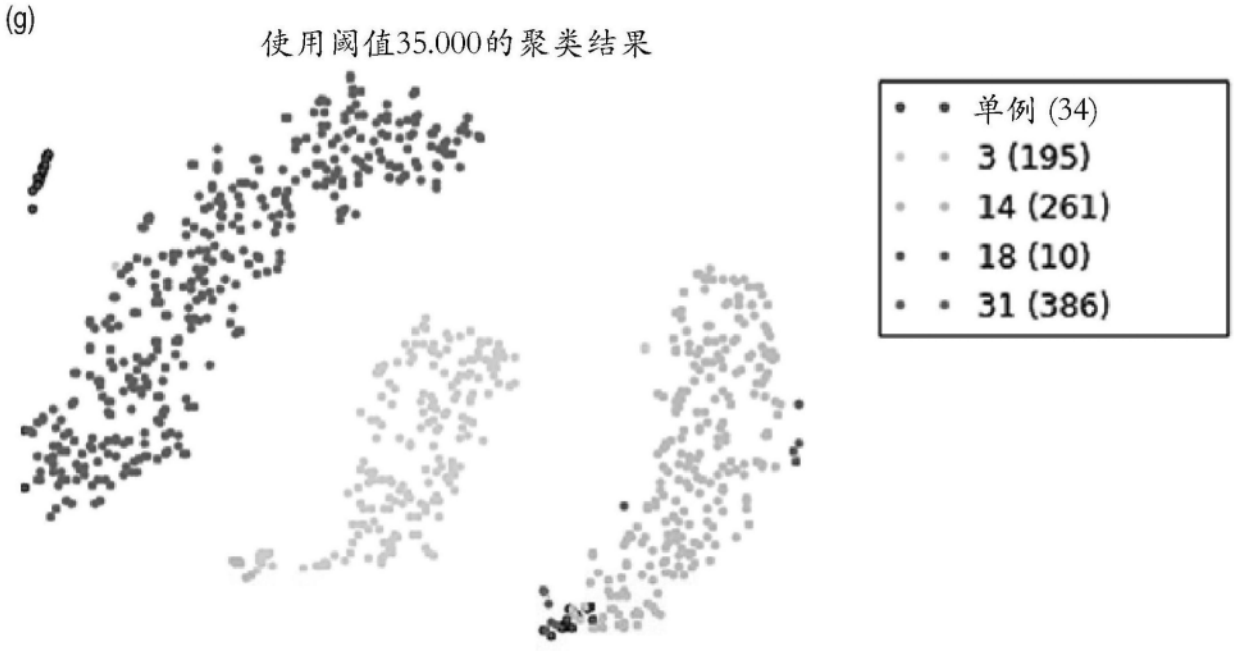
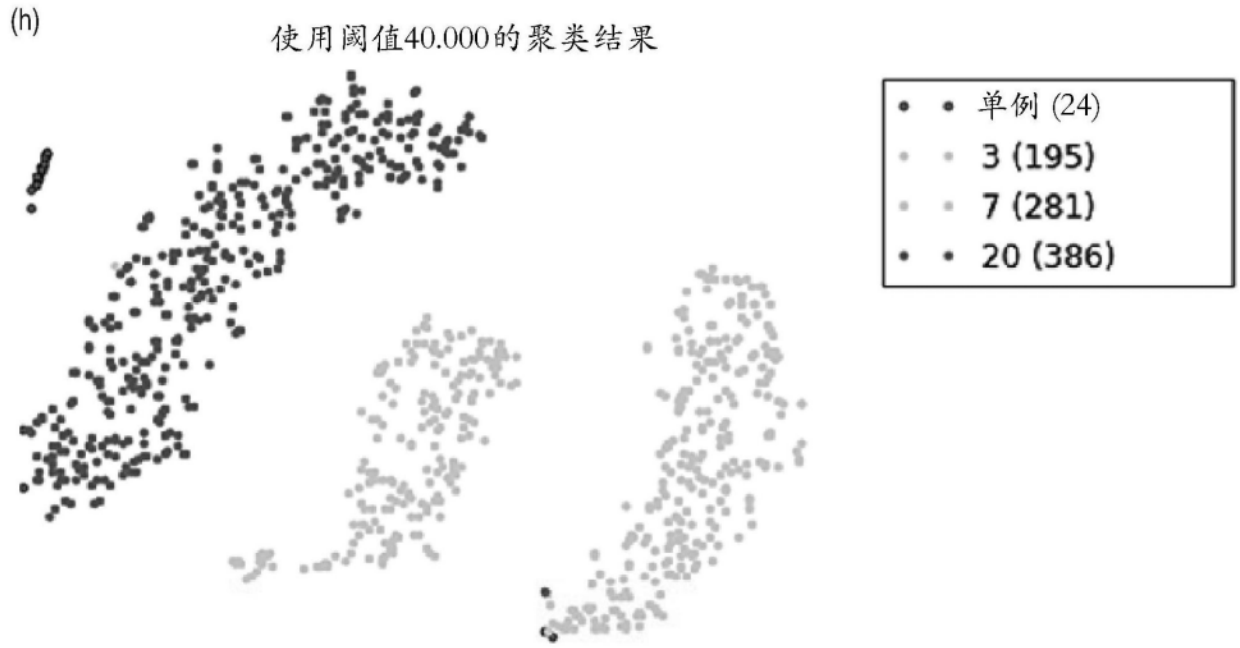

(g)
使用阈值35.000的聚类结果
单例 (34)
3 (195)
14 (261)
18 (10)
31 (386)

图19(续)


(h)
使用阈值40.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

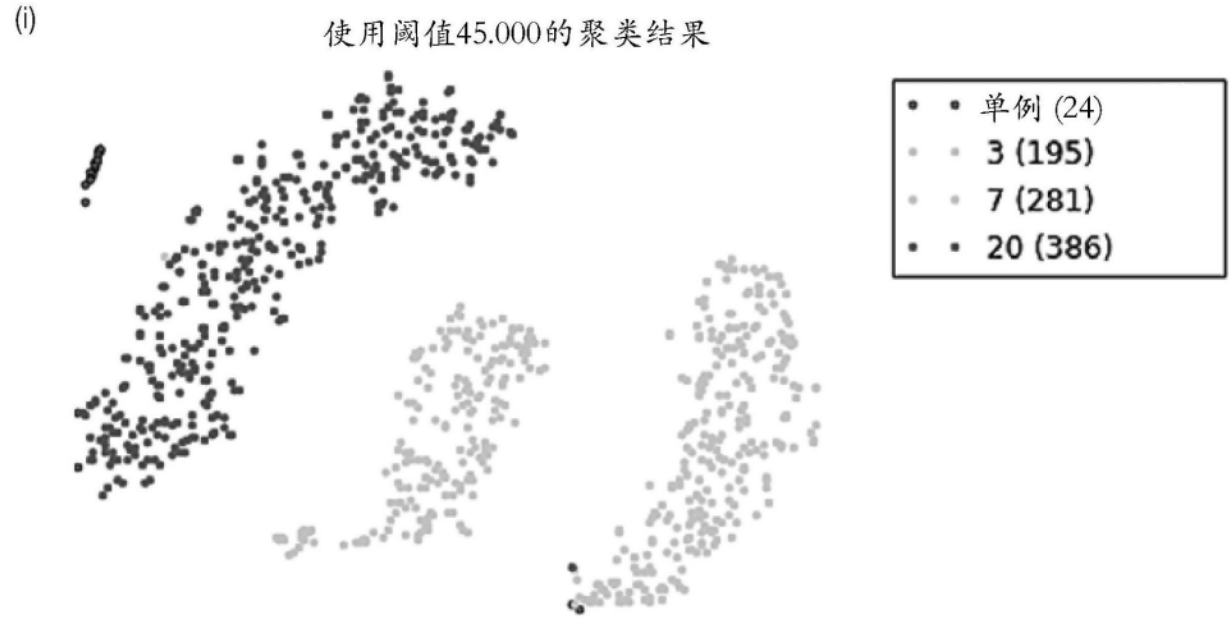
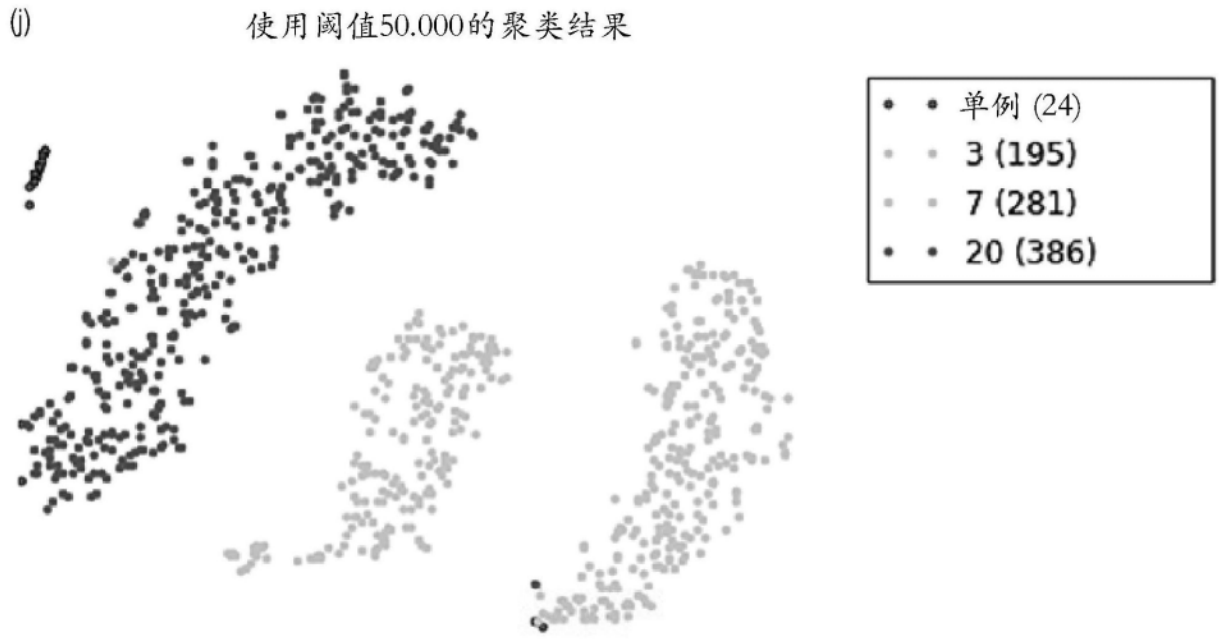

(i)
使用阈值45.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

图19(续)


(j)
使用阈值50.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

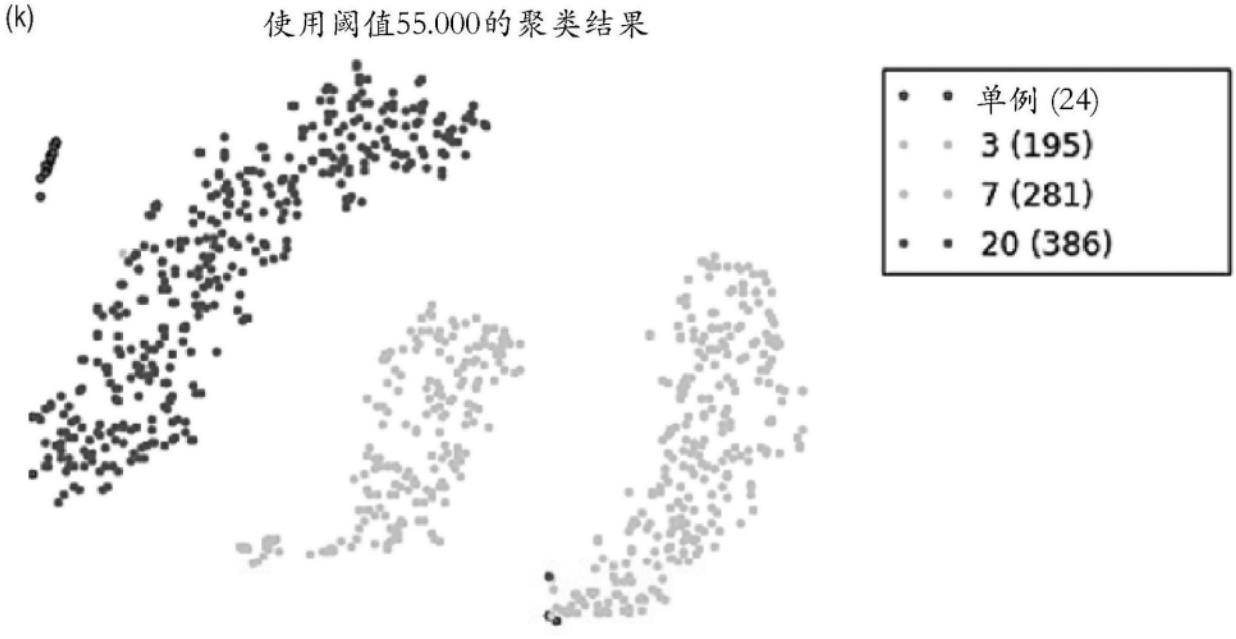
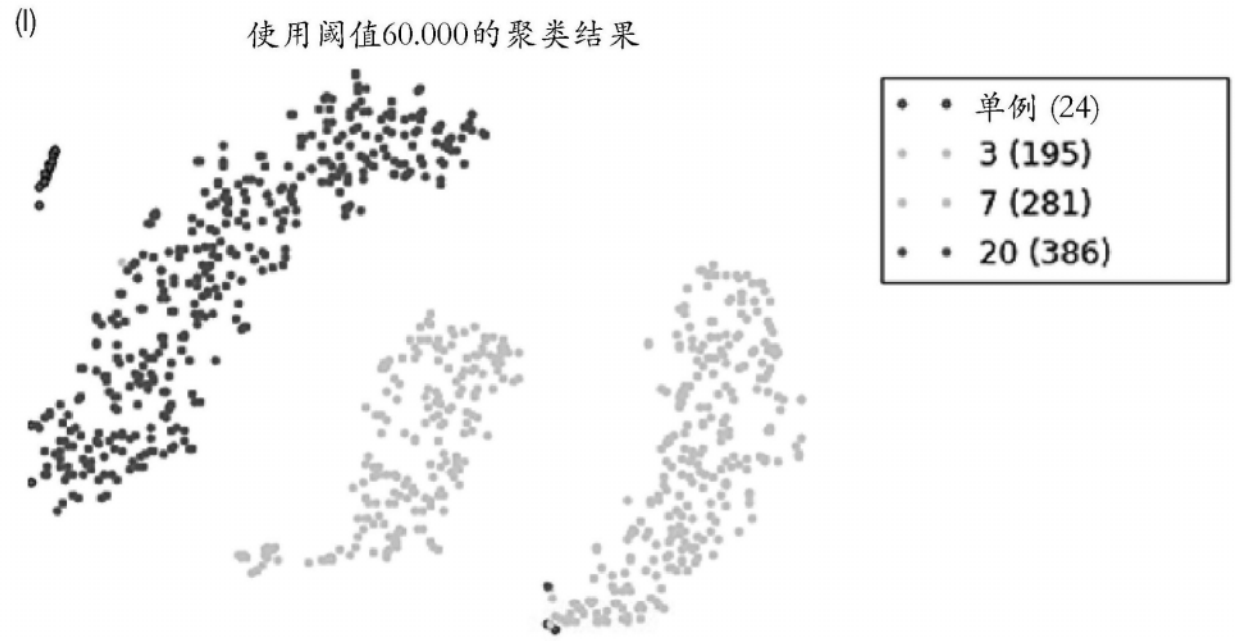

(k)
使用阈值55.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

图19(续)


(l)
使用阈值60.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

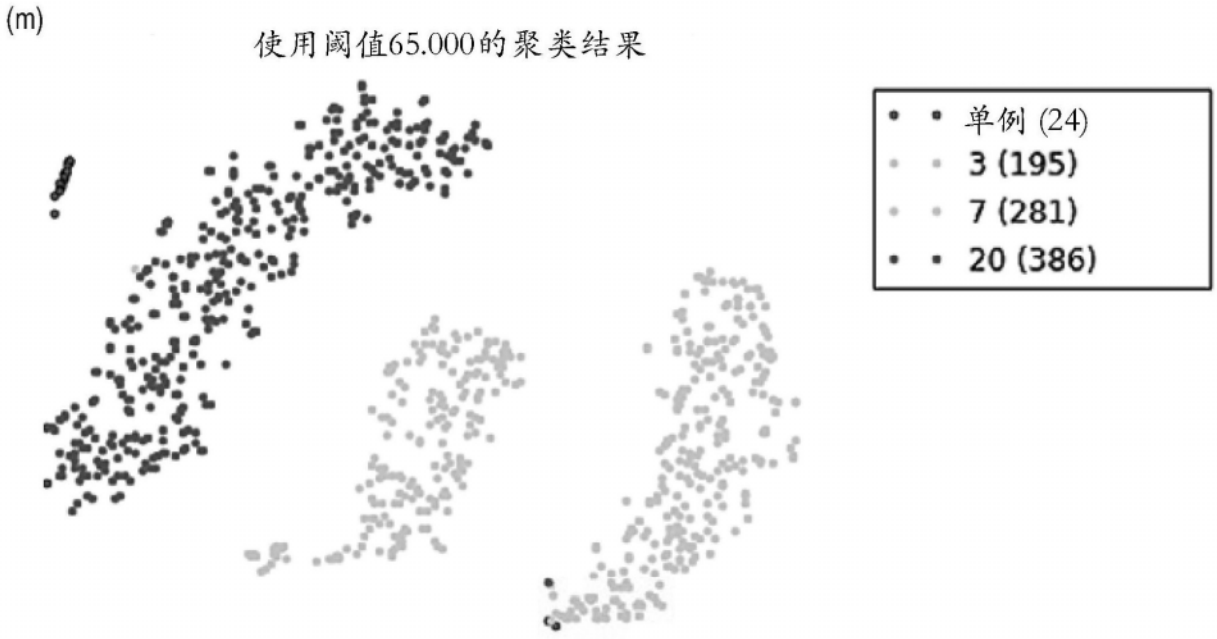

(m)
使用阈值65.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

图19(续)

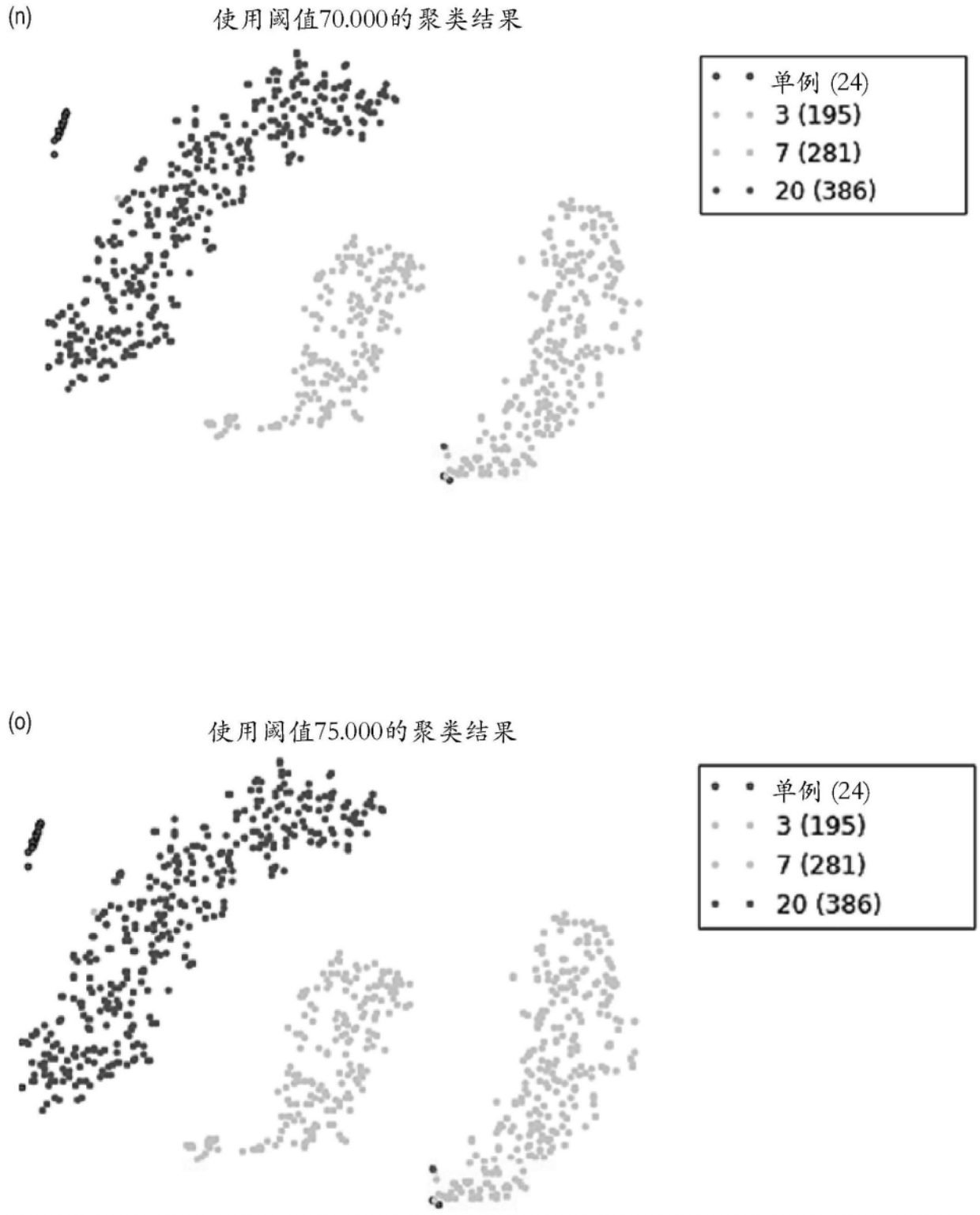
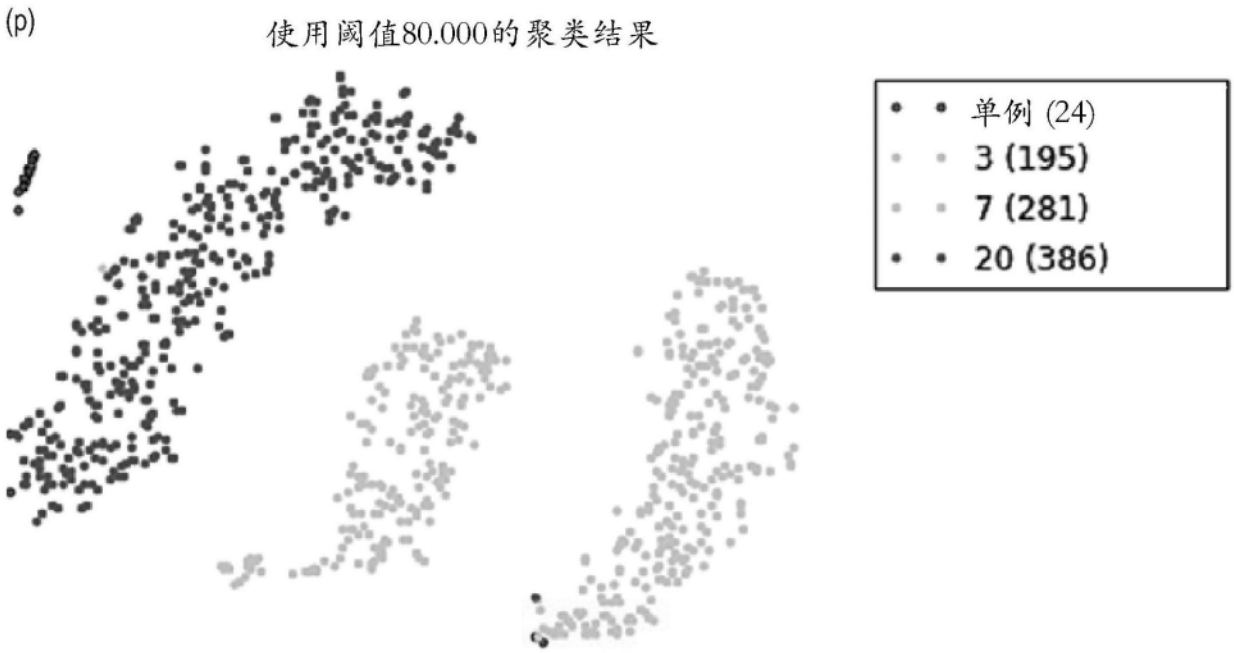

(n)
使用阈值70.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)
(o)
使用阈值75.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

图19(续)


(p)
使用阈值80.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

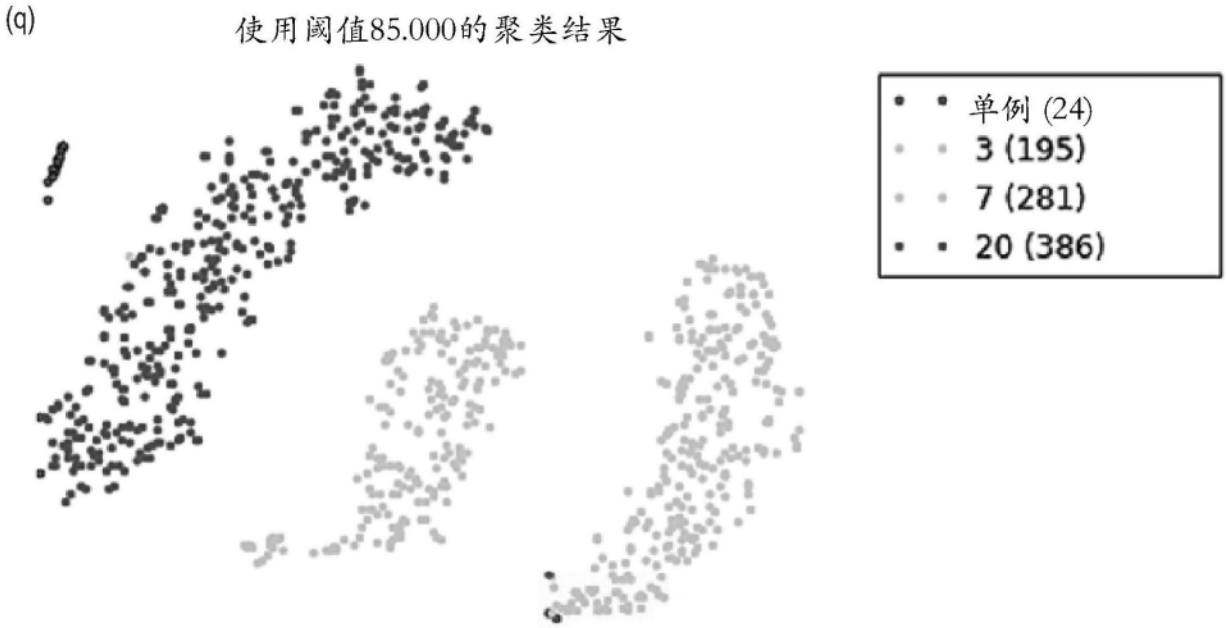
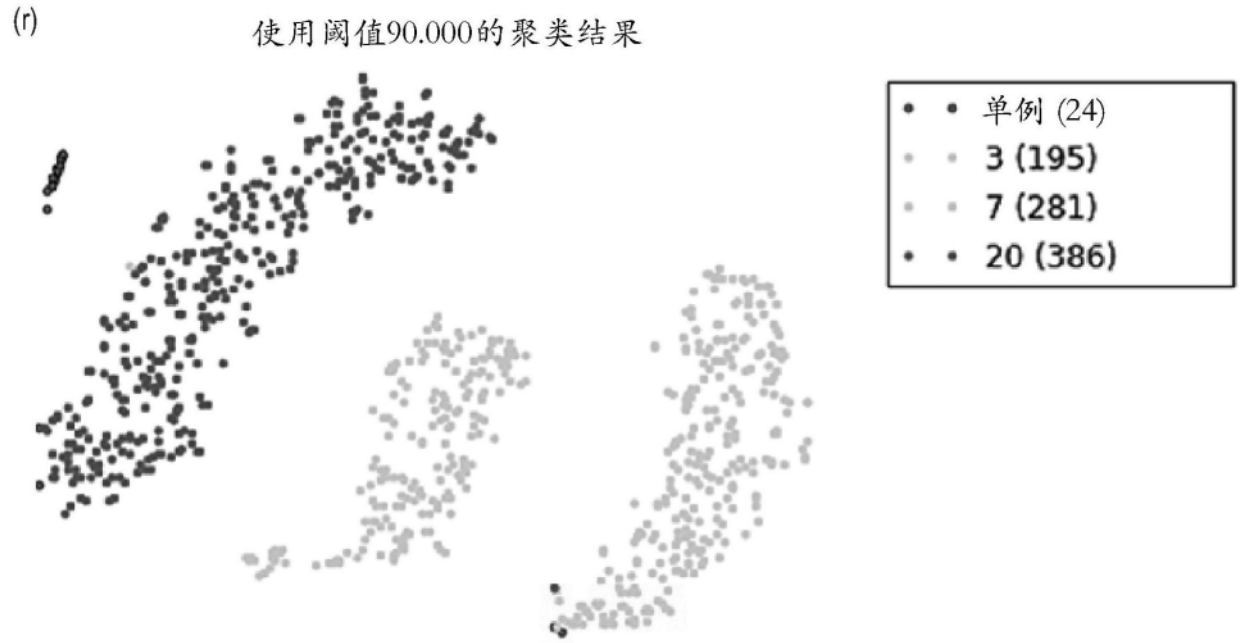

(q)
使用阈值85.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

图19(续)


(r)
使用阈值90.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

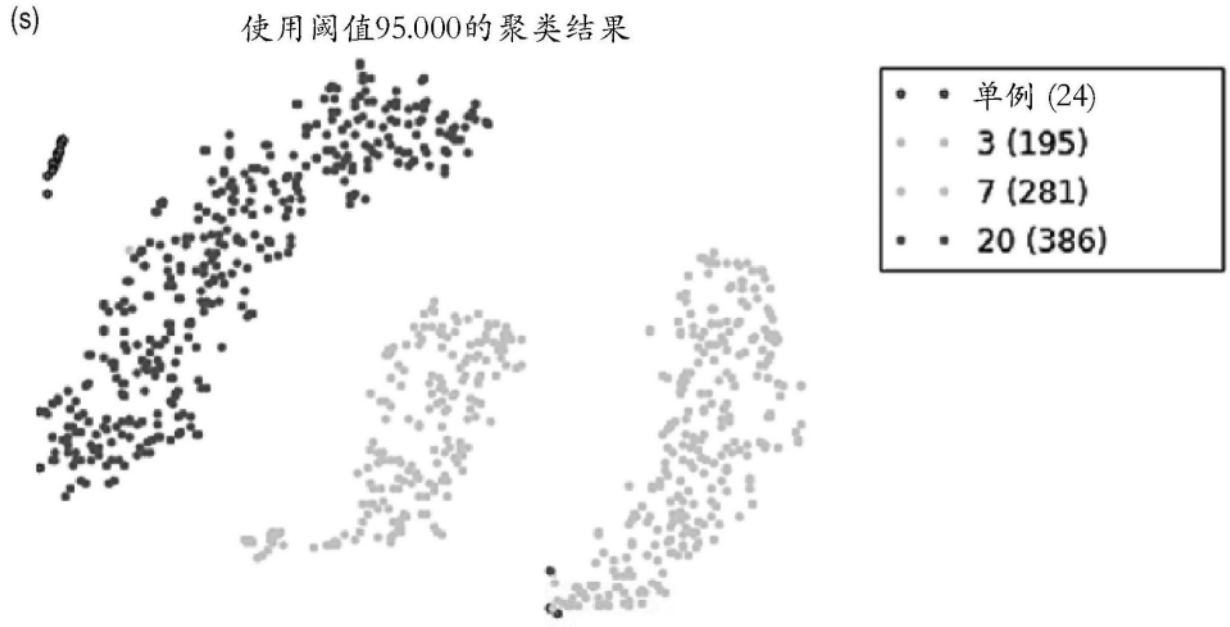

(s)
使用阈值95.000的聚类结果
单例 (24)
3 (195)
7 (281)
20 (386)

图19(续)

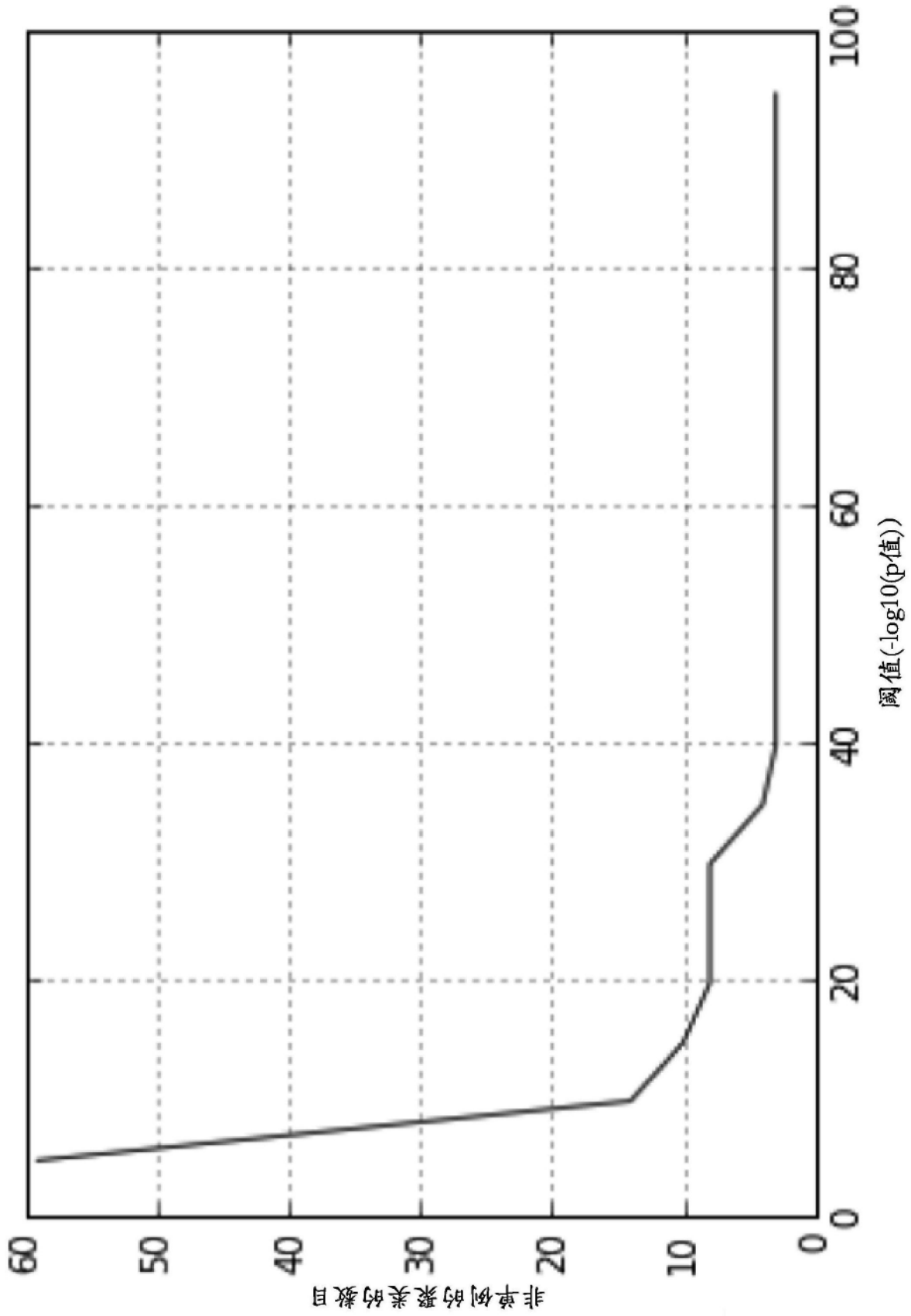


图20

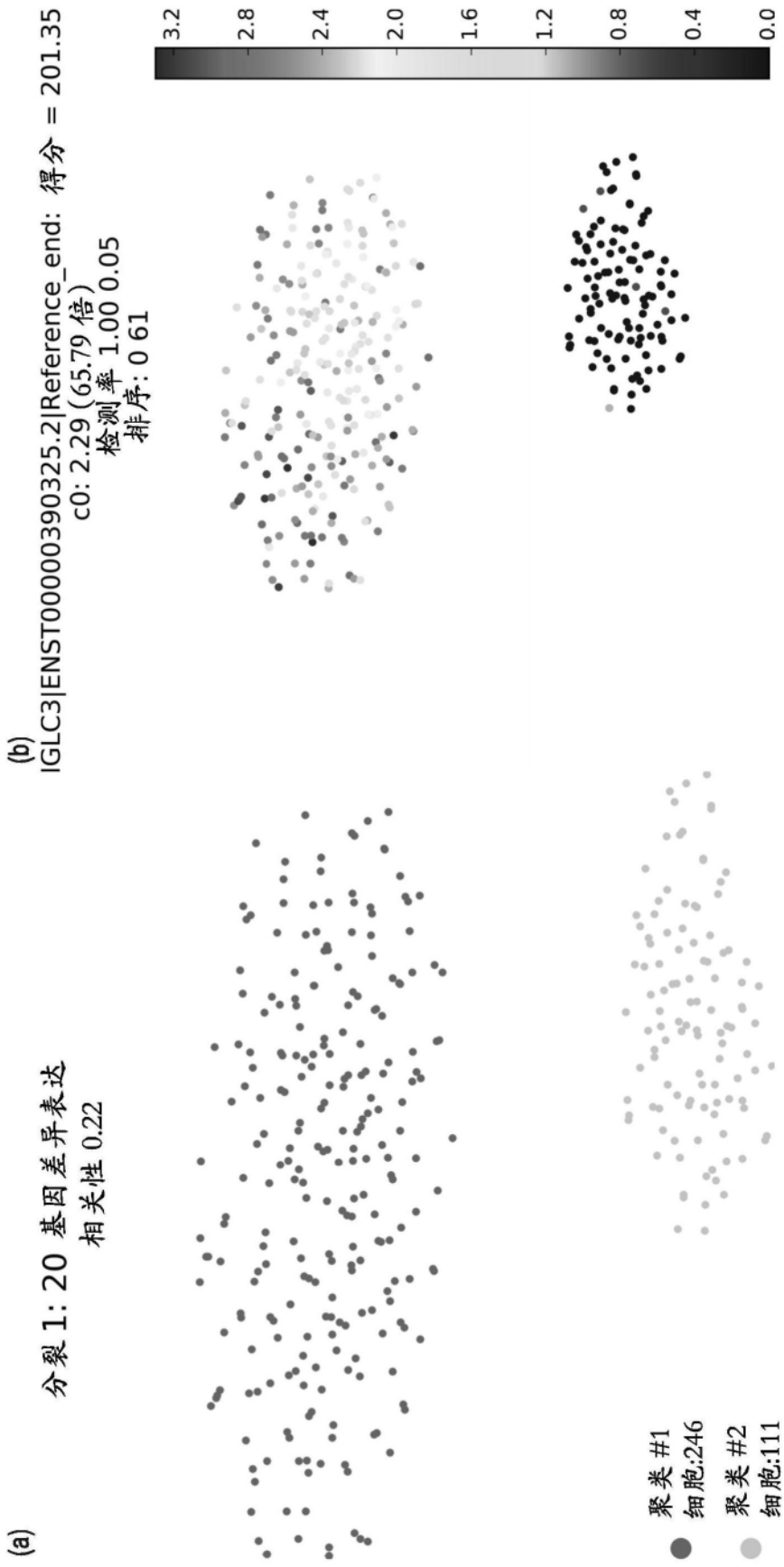

(a)
分裂 1: 20 基因差异表达
相关性 0.22
聚类 #1
细胞:246
聚类 #2
细胞:111
(b)
IGLC3|ENST00000390325.2|Reference_end: 得分 = 201.35
c0: 2.29 (65.79 倍)
检测率 1.00 0.05
排序: 0 61
3.2
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0.0

图21

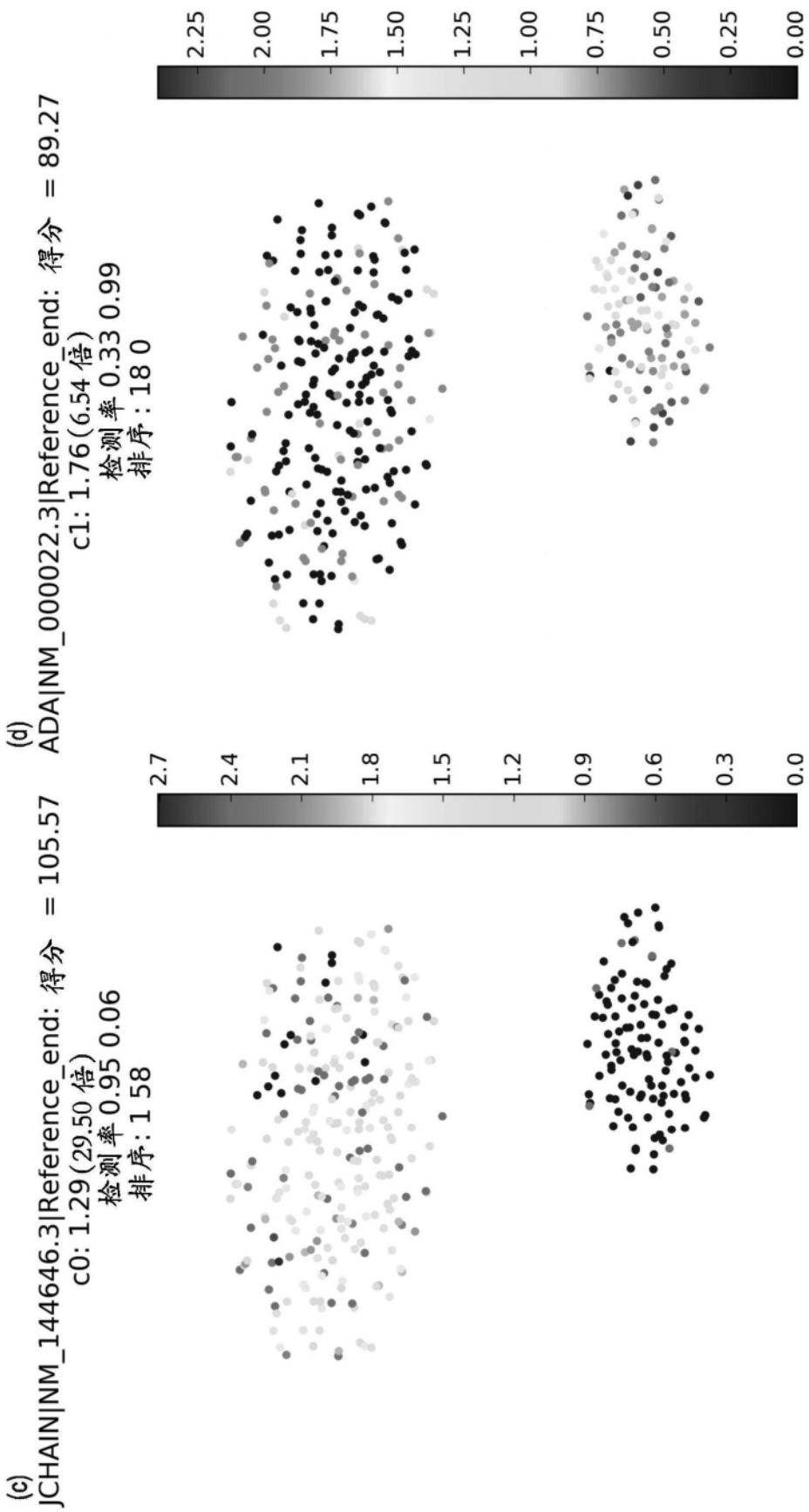

(c)
JCHAIN|NM_144646.3|Reference_end: 得分 = 105.57
c0: 1.29 (29.50 倍)
检测率 0.95 0.06
排序: 1 58
(d)
ADA|NM_000022.3|Reference_end: 得分 = 89.27
c1: 1.76 (6.54 倍)
检测率 0.33 0.99
排序: 18 0
2.7
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
0.0
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

图21(续)

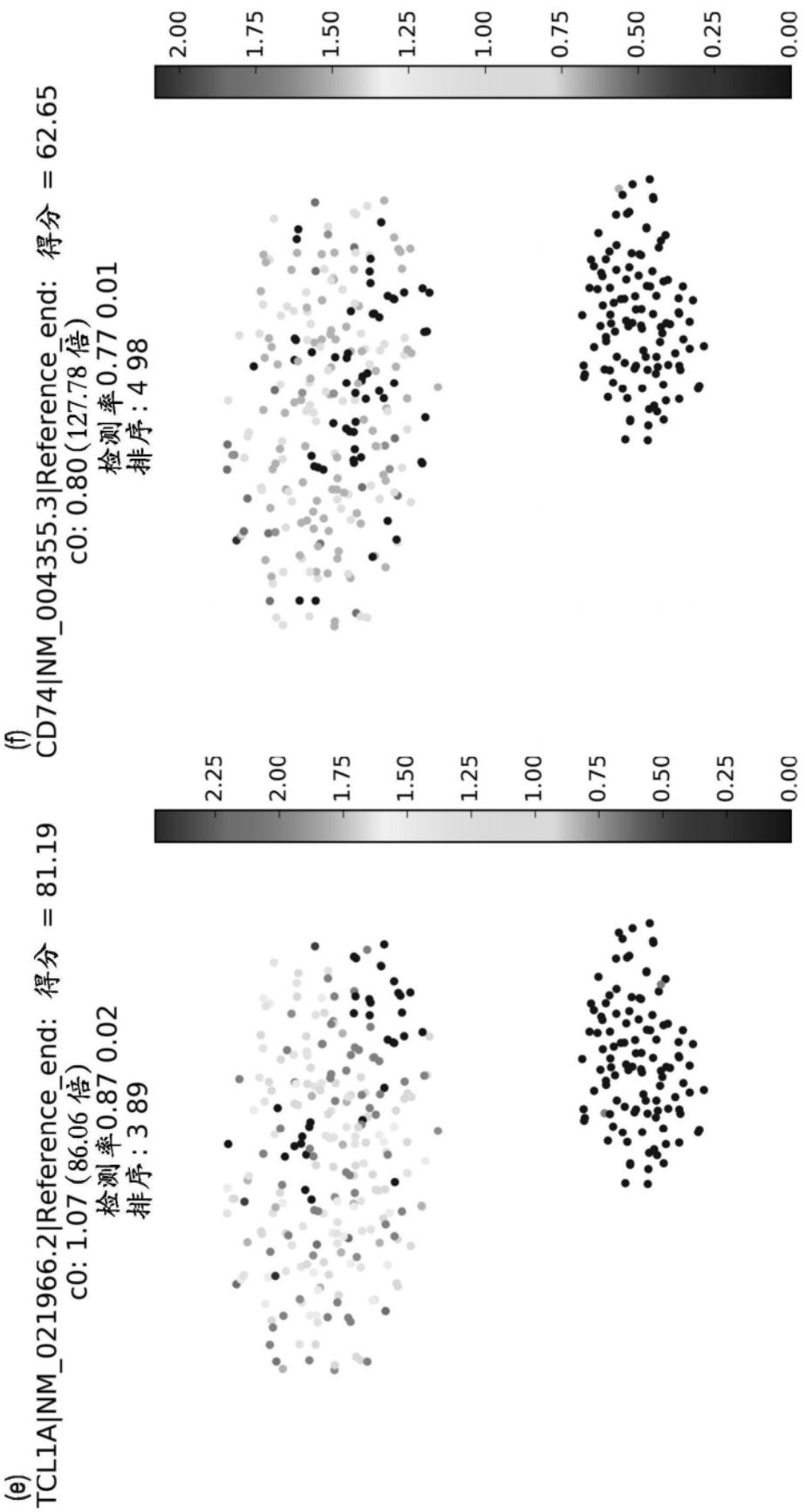

(e)
TCL1A|NM_021966.2|Reference_end: 得分 = 81.19
c0: 1.07 (86.06 倍)
检测率0.87 0.02
排序：3 89
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
(f)
CD74|NM_004355.3|Reference_end: 得分 = 62.65
c0: 0.80 (127.78 倍)
检测率0.77 0.01
排序：4 98
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

图21(续)

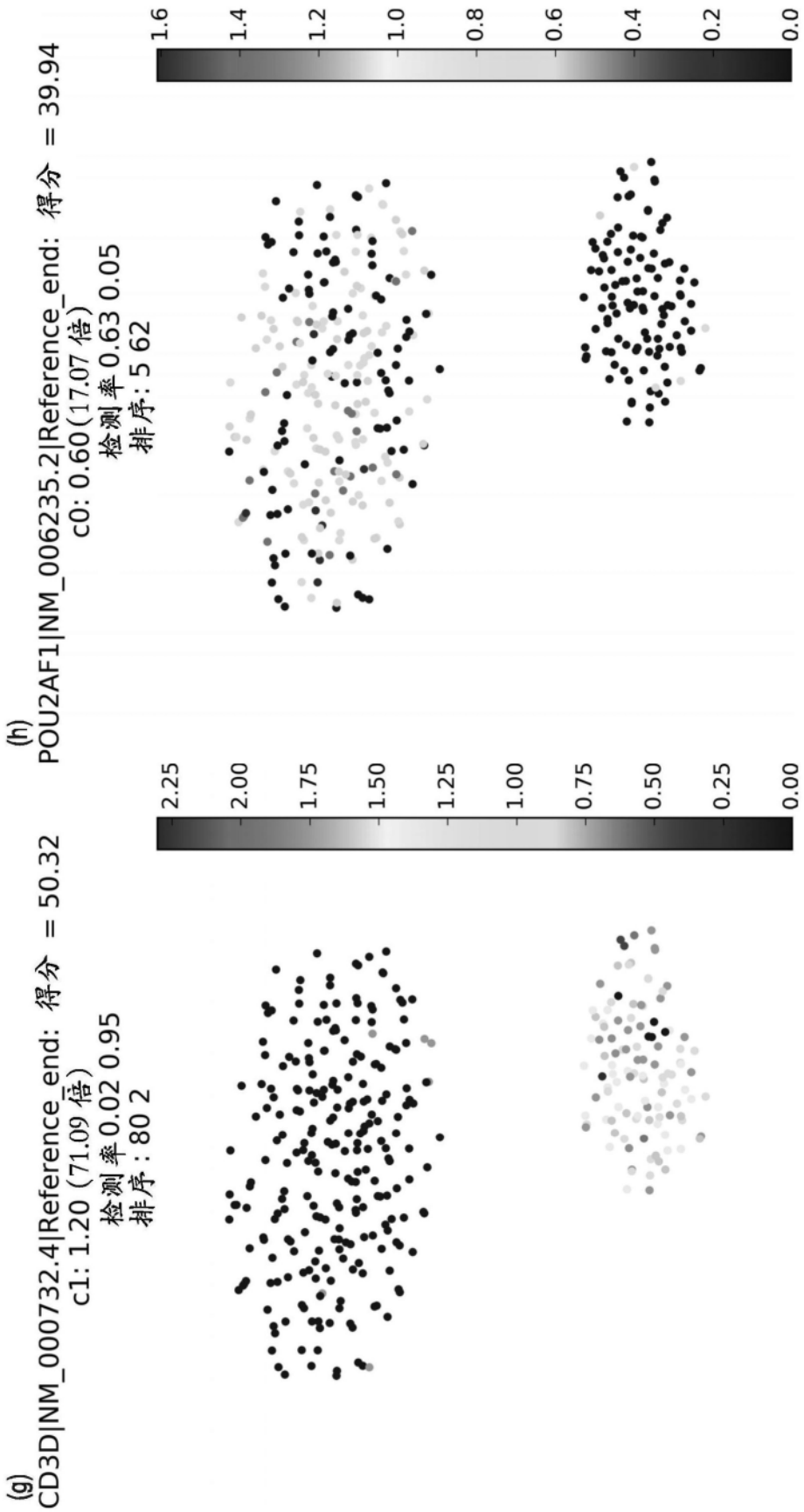

(g)
CD3D|NM_000732.4|Reference_end: 得分 = 50.32
c1: 1.20 (71.09 倍)
检测率0.02 0.95
排序：80 2
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
(h)
POU2AF1|NM_006235.2|Reference_end: 得分 = 39.94
c0: 0.60(17.07 倍)
检测率0.63 0.05
排序：5 62
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

图21(续)

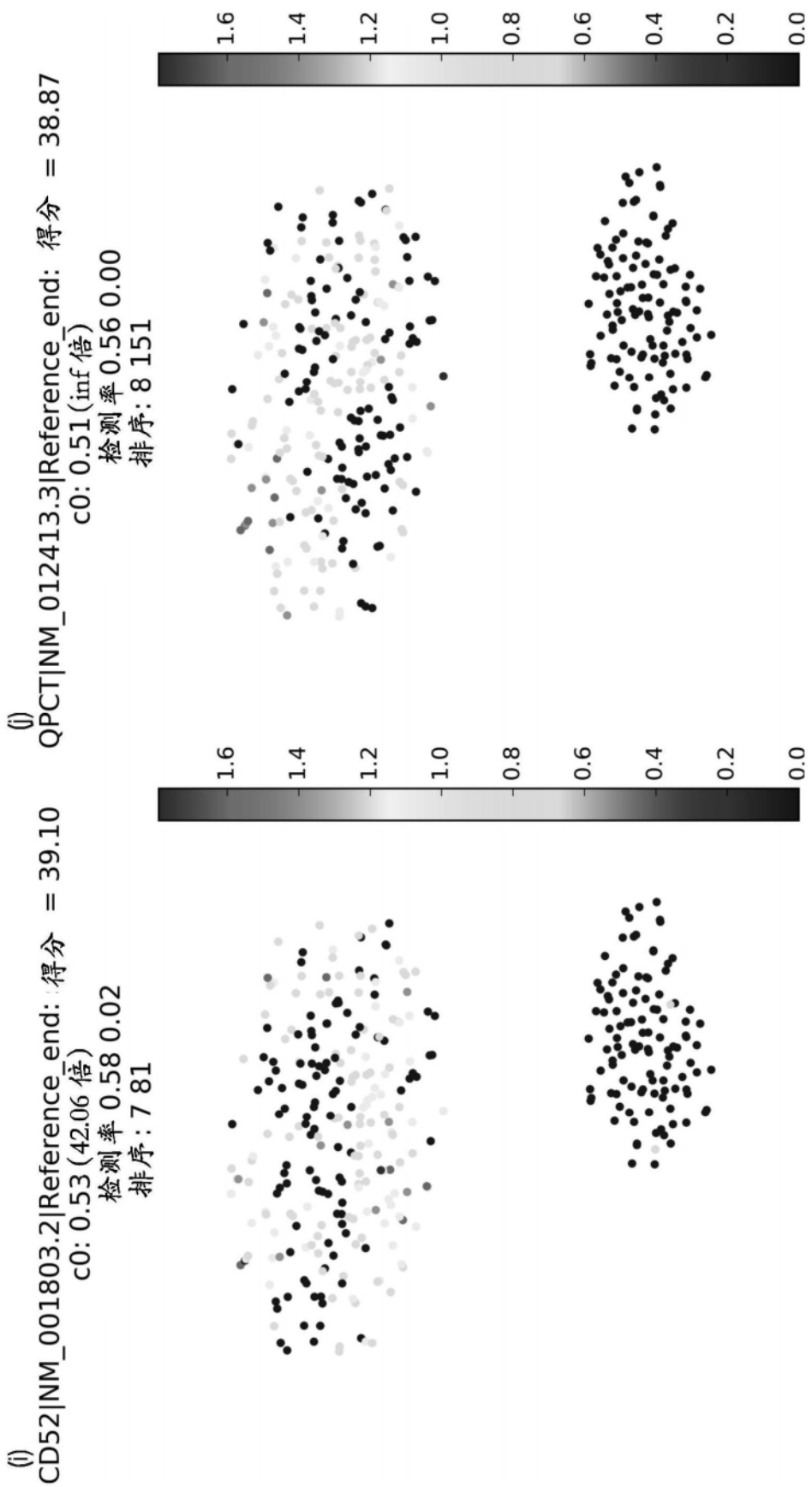

(j)
QPCT|NM_012413.3|Reference_end: 得分 = 38.87
c0: 0.51(inf倍)
检测率 0.56 0.00
排序: 8 151
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
(i)
CD52|NM_001803.2|Reference_end: 得分 = 39.10
c0: 0.53(42.06倍)
检测率 0.58 0.02
排序: 7 81

图21(续)

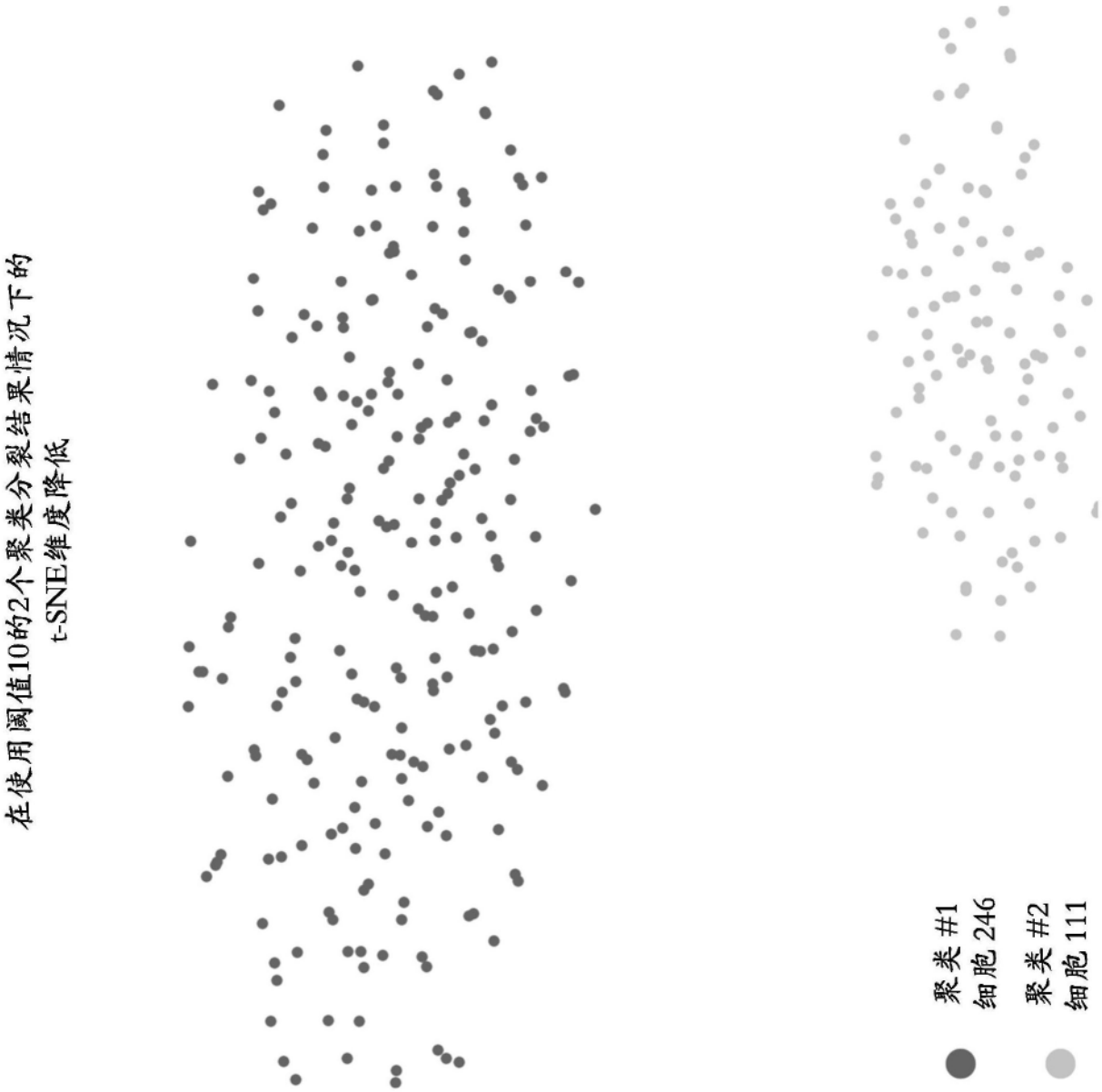


图22



图23

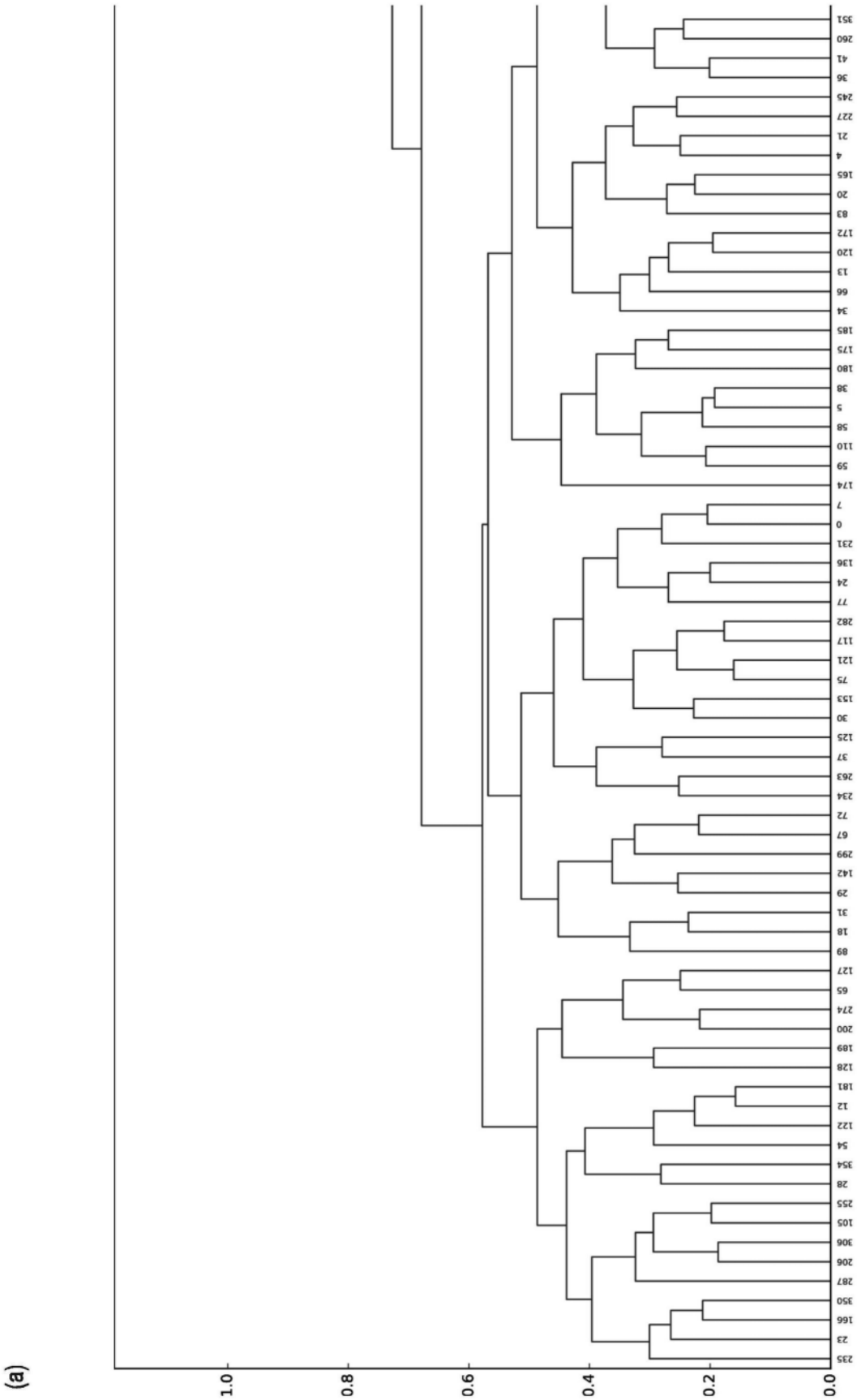

(a)
1.0
0.8
0.6
0.4
0.2
0.0
351
260
41
36
245
227
21
4
165
20
83
172
120
13
66
34
185
175
180
38
5
58
110
59
174
7
0
231
136
24
77
282
117
121
75
153
30
125
37
263
234
72
67
299
142
29
31
18
89
127
65
274
200
109
128
181
12
122
54
354
28
255
105
306
206
287
350
166
23
235

图23(续)

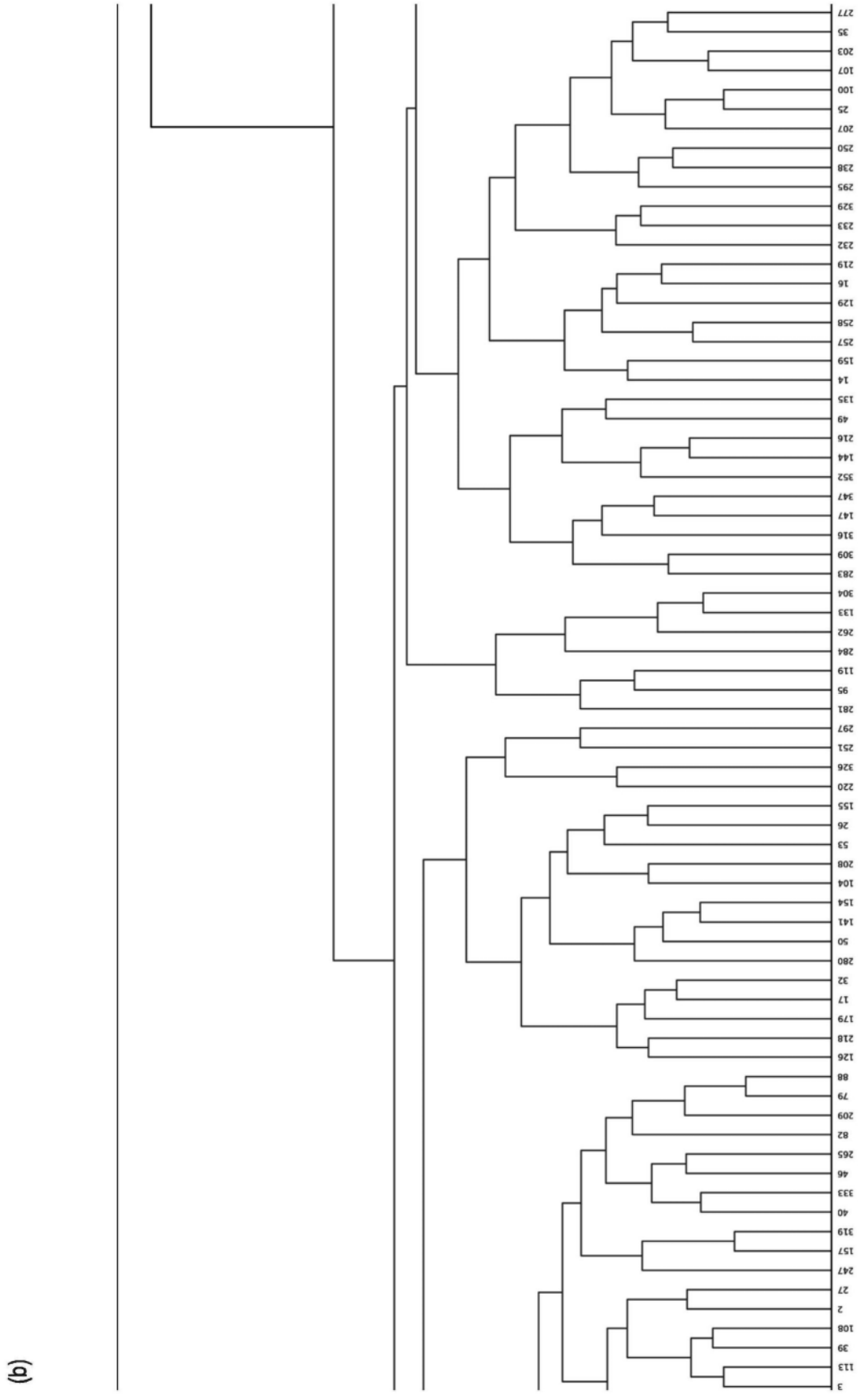

277
35
203
107
100
25
207
250
238
295
329
233
232
219
16
129
258
257
159
14
135
49
216
144
352
347
147
316
309
283
304
133
262
284
119
95
281
297
251
326
220
155
26
53
208
104
154
141
50
280
32
17
179
218
126
88
79
209
82
265
46
333
40
319
157
247
27
2
108
39
113
3
(b)

图23(续)

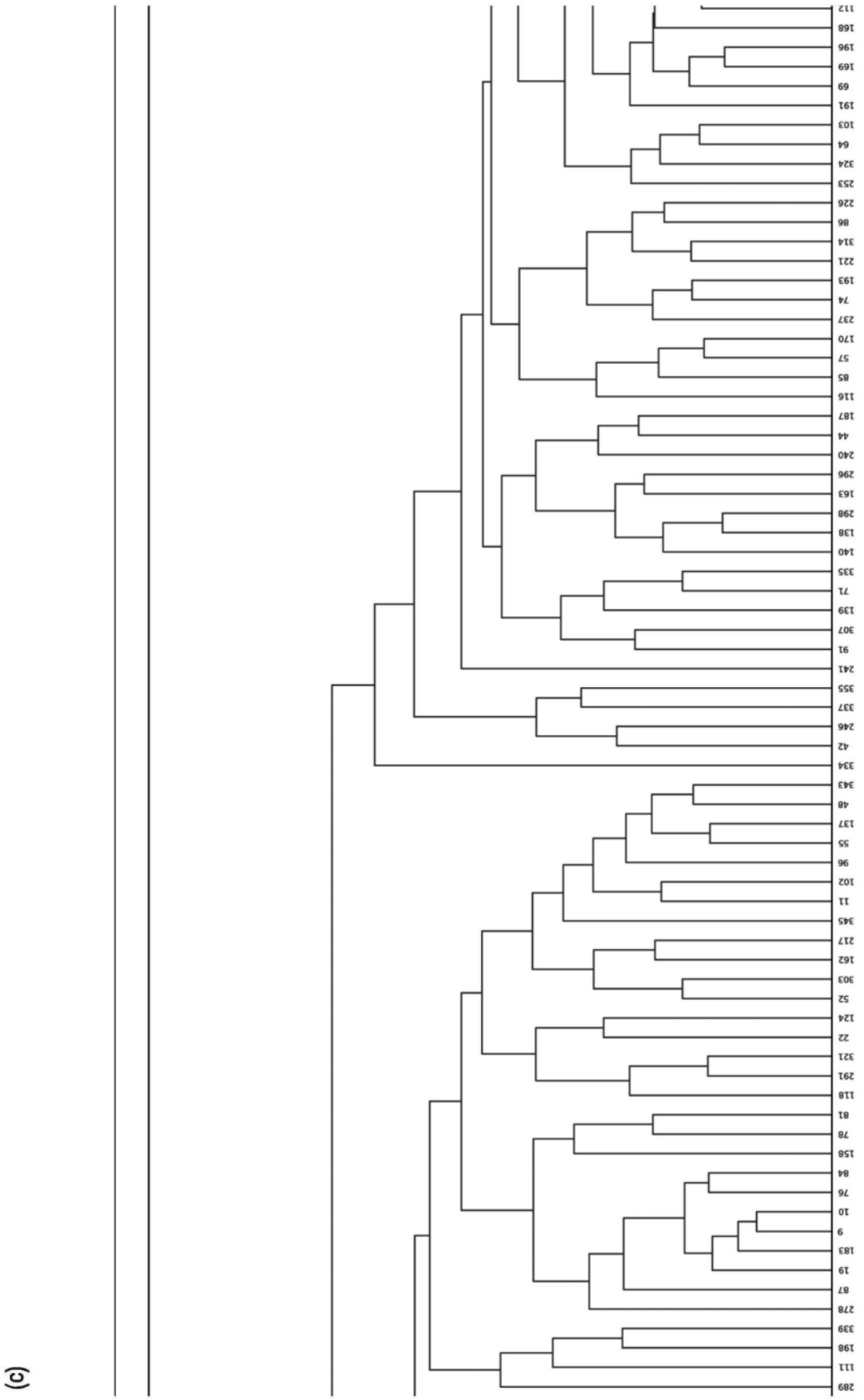

(c)
112
168
196
169
69
191
103
64
324
253
226
86
314
221
193
74
237
170
57
85
116
187
44
240
296
163
298
138
140
335
71
139
307
91
241
355
337
246
42
334
343
48
137
55
96
102
11
345
217
162
303
52
124
22
321
291
118
81
78
158
84
76
10
9
183
19
87
278
339
198
111
289

图23(续)

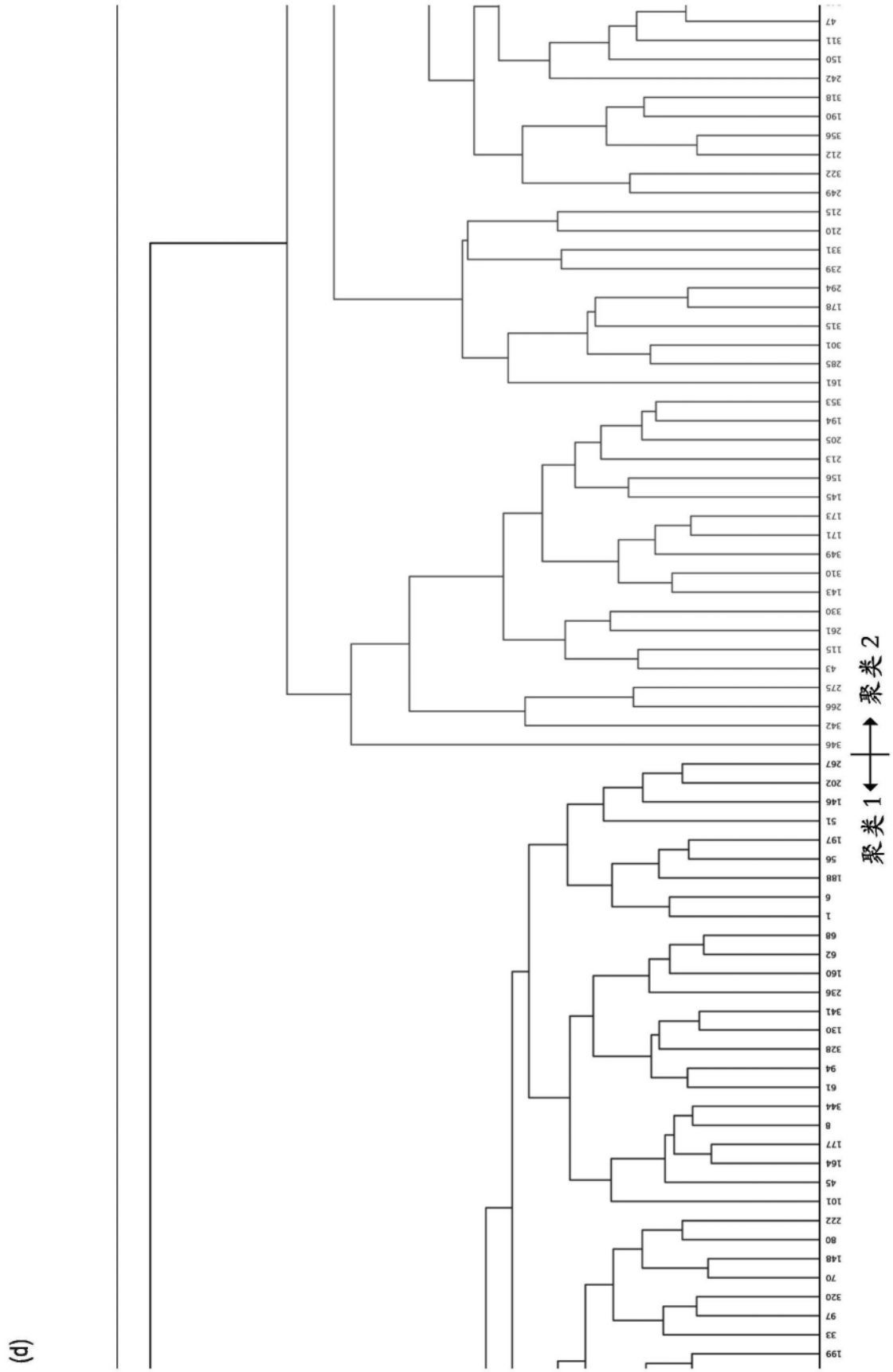


图23(续)

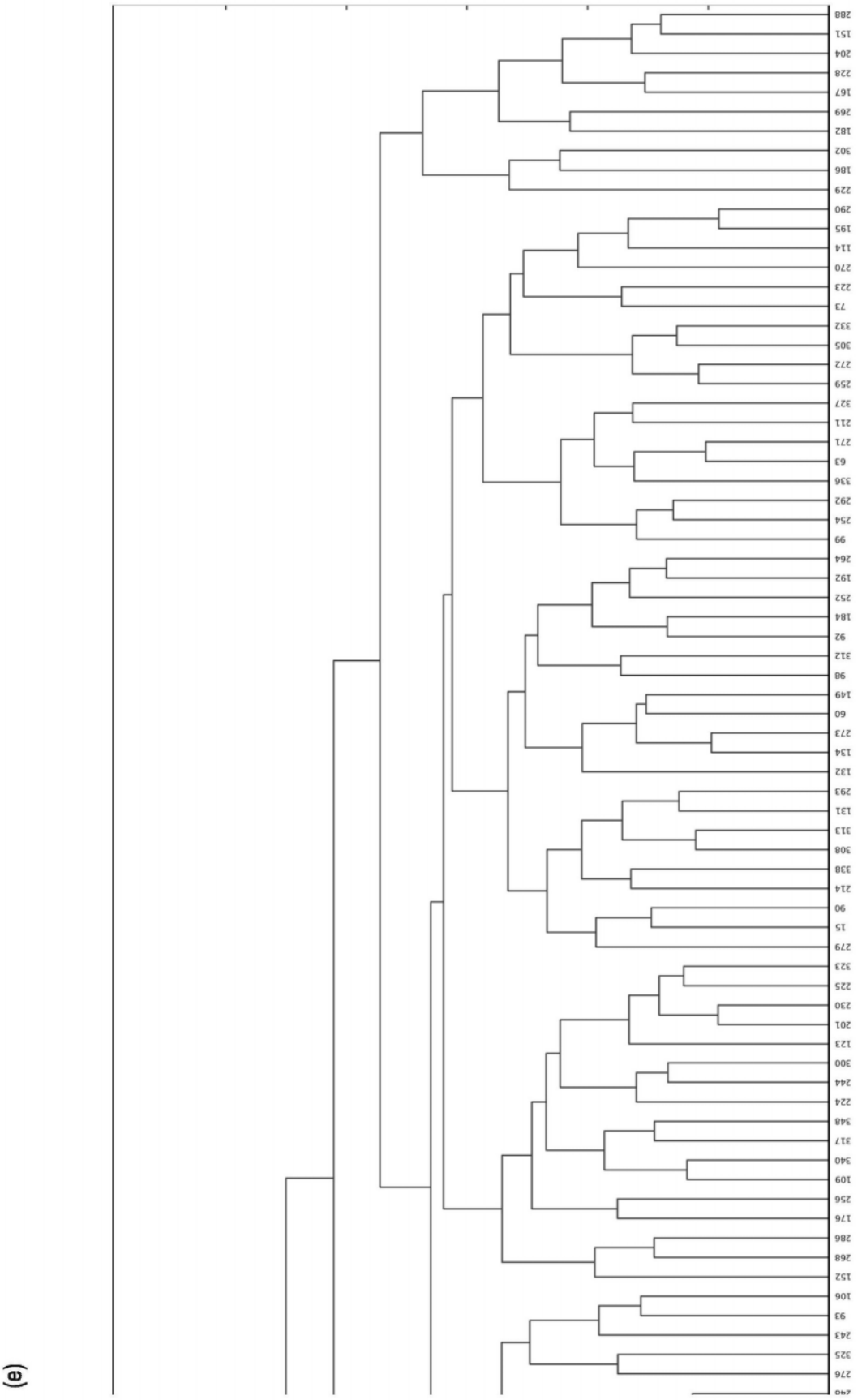


图23(续)

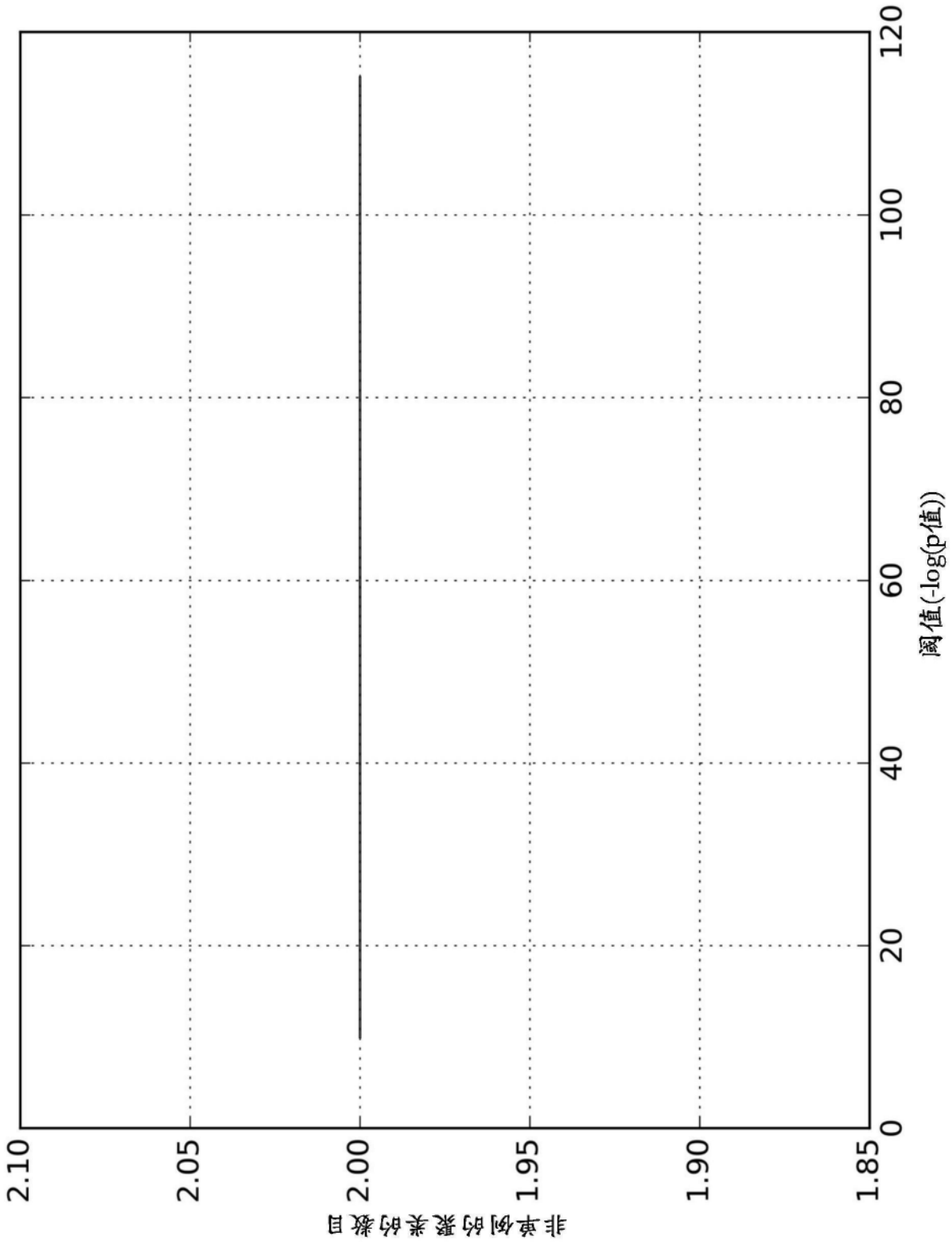


图24