(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109937358 B
(45)授权公告日 2024.03.12

(21)申请号 201780069990.6

(22)申请日 2017.12.13

(65)同一申请的已公布的文献号
申请公布号 CN 109937358 A

(43)申请公布日 2019.06.25

(30)优先权数据
62/433,930 2016.12.14 US

(85)PCT国际申请进入国家阶段日
2019.05.13

(86)PCT国际申请的申请数据
PCT/US2017/065987 2017.12.13

(87)PCT国际申请的公布数据
WO2018/111982 EN 2018.06.21

(73)专利权人 佛罗乔有限责任公司
地址 美国俄勒冈州

(72)发明人 詹姆斯·阿尔玛罗德
约瑟夫·斯皮德伦
迈克尔·大卫·斯塔德尼斯凯

(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 陆建萍 杨明钊

(51)Int.Cl.
G16B 45/00(2019.01)
G16B 25/00(2019.01)
G16B 50/30(2019.01)
G06F 3/0481(2022.01)
G06F 3/0482(2013.01)

(56)对比文件
JP 2005352771 A,2005.12.22
JP 2001511546 A,2001.08.14
US 2010070904 A1,2010.03.18
US 2016130574 A1,2016.05.12
CN 103764848 A,2014.04.30
Tim Van den Bulcke et al..SynTReN:a generator of synthetic gene expression data for design and analysis of structure learning algorithms.《BMC Bioinformatics》.2006,第7卷
David DeTomaso et al..FastProject: a tool for low dimensional analysis of single-cell RNA-Seq data.《BMC Bioinformatics》.2016,第17卷

审查员 王晓钰

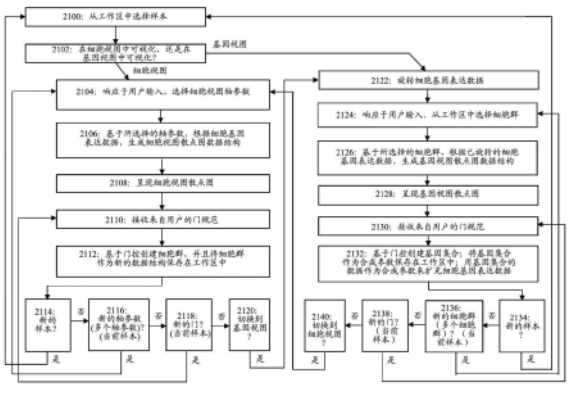
权利要求书5页 说明书12页 附图28页

(54)发明名称

应用计算机技术管理、合成、可视化和探索大型多参数数据集的参数

(57)摘要

公开了计算机技术,其将创新的数据处理和可视化技术应用于诸如细胞基因表达数据的大型多参数数据集,以发现诸如细胞和基因之间的关系的新关系,并在代表这些关系的数据集内创建新的关联数据结构。例如,基因表达数据的散点图可以在细胞视图和基因视图之间迭代地旋转,以找到用户关注的细胞群和基因集合。



CN 109937358 B

1.一种可视化多参数数据集的方法,包括:

处理器跨第一轴和第二轴生成所述多参数数据集的散点图,所述数据集包括多个数据项,每个数据项与多个参数相关联,并且显示每个相关联的参数的数据值,并且其中,所述第一轴和所述第二轴对应于在所述数据集内用户选择的参数,所述散点图包括多个点,每个点对应于所述数据集中的数据项,并且在所述散点图上沿着所述第一轴和所述第二轴的位置进行定位,位于与对应于所述轴的所述参数的所述数据值相对应的位置,

其中,所述散点图包括细胞视图散点图,

其中,所述方法还包括:处理器从所述细胞视图散点图旋转到基因视图散点图,所述基因视图散点图绘制了跨多个细胞群的多个不同基因,

其中,如果经由所述细胞视图散点图中的门控将多个细胞群的子集分组为两个细胞群,则所述两个细胞群在所述基因视图散点图中被旋转,以在所述基因视图散点图中可视化在所述两个细胞群之间的差异基因表达,并且

其中,所述基因视图散点图中的行显示了基因,所述基因在所述多参数数据集中是列,并且所述基因视图散点图中的列显示两个细胞群。

2.根据权利要求1所述的方法,其中,所述多参数数据集包括细胞基因表达数据,所述数据项包括多个细胞,其中,所述参数包括多个基因,并且其中,所述数据值包括指示在每个细胞中对应于基因的表达的计数的数据。

3.根据权利要求2所述的方法,其中,生成步骤包括处理器基于作为所述细胞视图散点图的轴参数的至少一个基因集合的用户规范,根据所述细胞基因表达数据,生成所述细胞视图散点图。

4.根据权利要求3所述的方法,还包括:

处理器门控所述细胞视图散点图中的一组点,以定义细胞群;以及

处理器在工作区中创建数据对象,所述数据对象代表所定义的细胞群。

5.根据权利要求4所述的方法,还包括处理器重复门控步骤和创建步骤,以在所述工作区中创建代表多个细胞群的多个数据对象。

6.根据权利要求4所述的方法,还包括:

处理器根据布尔运算将多个细胞群组合,以生成新的细胞群,以供作为数据对象被包含在所述工作区中。

7.根据权利要求3所述的方法,还包括:

改变所述细胞视图散点图,以在所述散点图的定义的零空间中展开对应于零的点。

8.根据权利要求3所述的方法,还包括:

处理器门控在所述基因视图散点图中的一组点,以定义基因集合;以及

处理器在工作区中创建合成参数数据对象,所述合成参数数据对象代表在所述工作区中的定义的基因集合。

9.根据权利要求8所述的方法,还包括:

处理器使用作为与所述细胞基因表达数据中的细胞相关联的合成参数的、所述定义的基因集合来扩充所述细胞基因表达数据。

10.根据权利要求8-9中任一项所述的方法,还包括:

处理器生成细胞视图散点图,其中,所述定义的基因集合被选择作为所述轴参数之一。

11.根据权利要求10所述的方法,还包括:

处理器基于所述定义的基因集合来门控在所述细胞视图散点图中的一组点,以定义另一个细胞群;以及

处理器在所述工作区中创建代表所定义的另一个细胞群的数据对象。

12.根据权利要求8所述的方法,还包括:

处理器根据布尔运算将多个基因集合组合,以生成新的基因集合,以供作为数据对象被包含在所述工作区中。

13.根据权利要求2所述的方法,其中,所述散点图包括根据所述细胞基因表达数据的基因视图散点图,并且其中,生成步骤包括处理器基于多个细胞群的用户规范来生成所述基因视图散点图。

14.根据权利要求13所述的方法,还包括:

处理器响应于所述多个细胞群的用户规范,从所述细胞基因表达数据的细胞视图散点图旋转到所述基因视图散点图。

15.根据权利要求13所述的方法,还包括:

处理器通过门控所述基因视图散点图中的多个基因来生成基因集合。

16.根据权利要求14所述的方法,还包括:

处理器基于用作所述细胞视图散点图中的轴参数的基因集合的用户规范,从所述基因视图散点图旋转到细胞视图散点图。

17.根据权利要求2所述的方法,其中,所述细胞基因表达数据包括从样本的随机标记中导出的多个随机标记。

18.根据权利要求1所述的方法,还包括:

处理器响应于用户输入,在所述散点图上叠加参数数据值的第三维度。

19.根据权利要求1所述的方法,还包括:

其中,所述处理器基于用户输入来调整怎样显示所述散点图。

20.一种可视化多参数数据集的方法,包括:

处理器门控细胞基因表达数据的细胞视图散点图的区域,以创建细胞群,其中,所述细胞基因表达数据包括针对多个参数的多个数据值,其中,所述细胞视图散点图包括对应于所述细胞基因表达数据的第一参数和第二参数的第一轴和第二轴;

处理器旋转到所述细胞基因表达数据的基因视图散点图,其中,所述基因视图散点图包括对应于所创建的细胞群的至少一个轴;以及

处理器门控所述细胞基因表达数据的所述基因视图散点图的区域,以创建基因集合;

其中,所述方法还包括:处理器用所创建的基因集合来扩充所述细胞基因表达数据,使得所述创建的基因集合能够用作所述细胞基因表达数据的合成参数;

其中,如果所述细胞群的子集经由所述细胞视图散点图中的门控分为两个细胞群,则所述两个细胞群在所述基因视图散点图中被旋转,以在所述基因视图散点图中可视化在所述两个细胞群之间的差异基因表达,并且

其中,所述基因视图散点图中的行显示了基因,所述基因在所述细胞基因表达数据中是列,并且所述基因视图散点图中的列显示两个细胞群。

21.根据权利要求20所述的方法,还包括:

处理器旋转到所述细胞基因表达数据的另一个细胞视图散点图,其中,所述另一个细胞视图散点图包括对应于所述创建的基因集合的至少一个轴。

22.根据权利要求20所述的方法,还包括:

处理器响应于用户输入,选择所述细胞基因表达数据的参数以用作所述散点图的轴。

23.根据权利要求20所述的方法,还包括:

分析所述两个细胞群以识别第一差异性地表达的基因集合,所述第一差异性地表达的基因集合与关于癌症类型的相对较好的存活机会相关;以及

分析所述两个细胞群以识别第二差异性地表达的基因集合,所述第二差异性地表达的基因集合与关于所述癌症类型的相对较差的存活机会相关。

24.根据权利要求23所述的方法,还包括:

处理器旋转到所述细胞基因表达数据的另一个细胞视图散点图,其中,所述另一个细胞视图散点图包括对应于所述第一差异性地表达的基因集合的第一轴和对应于所述第二差异性地表达的基因集合的第二轴。

25.根据权利要求24所述的方法,还包括:

处理器门控所述细胞基因表达数据的所述另一个细胞视图散点图的区域,以创建与关于所述癌症类型的相对较好的存活机会相关的细胞群。

26.根据权利要求25所述的方法,还包括:

处理器用与关于所述癌症类型的相对较好的存活机会相关的所述细胞群来扩充所述细胞基因表达数据,使得与关于所述癌症类型的相对较好的存活机会相关的所述细胞群能够用作所述细胞基因表达数据的合成参数。

27.根据权利要求24-26中任一项所述的方法,还包括:

处理器门控所述细胞基因表达数据的所述另一个细胞视图散点图的区域,以创建与关于所述癌症类型的相对较差的存活机会相关的细胞群。

28.根据权利要求27所述的方法,还包括:

处理器用与关于所述癌症类型的相对较差的存活机会相关的所述细胞群来扩充所述细胞基因表达数据,使得与关于所述癌症类型的相对较差的存活机会相关的所述细胞群能够用作所述细胞基因表达数据的合成参数。

29.根据权利要求23所述的方法,还包括:

分析所述两个细胞群以识别第三差异性地表达的基因集合,所述第三差异性地表达的基因集合与关于所述癌症类型的相对较好的治疗反应性相关。

30.根据权利要求29所述的方法,还包括:

处理器旋转到所述细胞基因表达数据的另一个细胞视图散点图,其中,所述另一个细胞视图散点图包括对应于所述第一差异性地表达的基因集合的第一轴和对应于所述第二差异性地表达的基因集合的第二轴;以及

处理器将所述细胞基因表达数据的第三维度叠加在所述另一个细胞视图散点图上,其中,所述第三维度对应于所述第三差异性地表达的基因集合。

31.根据权利要求30所述的方法,还包括:

处理器门控叠加了所述第三差异性地表达的基因集合的所述细胞基因表达数据的所述另一个细胞视图散点图的区域,以创建与关于所述癌症类型的相对较好的存活机会和关

于所述癌症类型的相对较好的治疗反应性均相关的细胞群。

32.根据权利要求31所述的方法,还包括:

处理器用与关于所述癌症类型的相对较好的存活机会和关于所述癌症类型的相对较好的治疗反应性均相关的所述细胞群来扩充所述细胞基因表达数据,使得与关于所述癌症类型的相对较好的存活机会和关于所述癌症类型的相对较好的治疗反应性均相关的所述细胞群能够用作所述细胞基因表达数据的合成参数。

33.根据权利要求23所述的方法,还包括:

分析所述两个细胞群以识别第四差异性地表达的基因集合,所述第四差异性地表达的基因集合与关于所述癌症类型的相对较差的治疗反应性相关。

34.根据权利要求33所述的方法,还包括:

处理器旋转到所述细胞基因表达数据的另一个细胞视图散点图,其中,所述另一个细胞视图散点图包括对应于所述第一差异性地表达的基因集合的第一轴和对应于所述第二差异性地表达的基因集合的第二轴;以及

处理器将所述细胞基因表达数据的第三维度叠加在所述另一个细胞视图散点图上,其中,所述第三维度对应于所述第四差异性地表达的基因集合。

35.根据权利要求34所述的方法,还包括:

处理器门控叠加了所述第四差异性地表达的基因集合的所述细胞基因表达数据的所述另一个细胞视图散点图的区域,以创建与关于所述癌症类型的相对较好的存活机会但关于所述癌症类型的相对较差的治疗反应性相关的细胞群。

36.根据权利要求35所述的方法,还包括:

处理器用与关于所述癌症类型的相对较好的存活机会但是关于所述癌症类型的相对较差的治疗反应性相关的所述细胞群来扩充所述细胞基因表达数据,使得与关于所述癌症类型的相对较好的存活机会但是关于所述癌症类型的相对较差的治疗反应性相关的所述细胞群能够用作所述细胞基因表达数据的合成参数。

37.根据权利要求20所述的方法,还包括:

处理器响应于用作所述细胞基因表达数据的第三维度的所述细胞基因表达数据的参数的用户选择,在至少一个所述散点图上叠加所述第三维度。

38.根据权利要求20所述的方法,还包括:

处理器用所述创建的细胞群来扩充所述细胞基因表达数据,使得所述创建的细胞群能够用作所述细胞基因表达数据的合成参数。

39.根据权利要求20所述的方法,其中,旋转步骤包括:

处理器对所述细胞基因表达数据中的多个参数数据值执行统计计算,以创建多个已旋转的数据值;以及

处理器用所述已旋转的数据值来填充所述细胞基因表达数据的已旋转的视图。

40.一种可视化多参数数据集的装置,包括:

存储器,其被配置为存储多参数数据集;以及

处理器,其用于与所述存储器协作,所述处理器被配置为执行权利要求1-39中任一项所述的方法。

41.根据权利要求40所述的装置,还包括:

显示器,其与所述处理器协作,其中,所述显示器被配置为以图形方式向用户呈现所述散点图。

42.一种非暂态计算机可读存储介质,包括:

多个处理器可执行指令,其驻留在所述非暂态计算机可读存储介质上,其中,所述指令被配置为在被处理器执行时,使所述处理器执行权利要求1-39中任一项所述的方法。

43.根据权利要求42所述的非暂态计算机可读存储介质,其中,所述指令在被所述处理器执行时,还被配置成指示显示器以图形方式向用户呈现所述散点图。

应用计算机技术管理、合成、可视化和探索大型多参数数据集的参数

[0001] 相关专利申请的交叉引用和优先权要求

[0002] 本专利申请要求于2016年12月14日提交的序列号为62/433,930并且标题为“Applied Computer Technology for Management,Synthesis,Visualization,and Exploration of Parameters in Large Multi-Parameter Data Sets”的美国临时专利申请的优先权,其全部公开内容通过引用并入本文。

技术领域

[0003] 本申请涉及计算机技术的应用,其使用遍及细胞表达数据的各个维度的创新的散点图(scatterplot)显示。

背景技术

[0004] 可用于大量人群和各个细胞的大量的遗传和基因表达信息已经增长到了对于调查人员来说变得难以控制(unwieldy)的程度。例如,细胞基因表达(gene expression)数据可以包括成千上万个基因(例如,10,000-30,000个或更多个基因)的基因表达数据,现在可以针对各个细胞进行测量,并且每个样本可以测量成千上万个细胞。这在细胞基因表达数据的可视化、分析、探索和理解领域提出了一个巨大的技术问题。

[0005] 例如,对于使用计算机来促进细胞基因表达数据可视化的传统方法,可视化是最终的终点,并且作为用户使用R编程语言手动编写脚本的结果,得出可视化,这要求用户具有不同库的知识,以便执行数据输入、重新格式化、操作、计算和绘图。这些脚本通常必须针对特定的数据集进行定制,并且它们的创建需要对编程语言、现有库以及用于产生结果所需的输入的专门知识。此外,这种常规方法阻碍了对异质细胞群(cell population)的深入探索。

发明内容

[0006] 作为该技术问题的解决方案,发明人公开了计算机技术的应用,其使用遍及细胞表达数据的各个维度的创新的散点图(scatterplot)显示,包括细胞(或细胞群)视图散点图,其中细胞被可视化为各个数据点(例如,细胞的基因相对于基因散点图的关系),以及基因视图散点图,其中,基因被可视化为各个数据点(例如,基因的细胞群相对于细胞群散点图的关系)。可以在这些散点图中执行门控(gate),以分别创建细胞群和基因集合,这些细胞群和基因集合可以充当生物学相关的维度,而被添加到工作区中作为新的数据对象,以用于扩充(augment)细胞基因表达数据,并且为有意义的调查开辟新的途径。作为对比,基于各个基因,以隔离的、孤立的方式进行这种分析很快变得难以控制,而在细胞视图散点图和基因视图散点图之间旋转(pivot)的能力允许用户找到生物学相关的基因分组,然后可以将其作为细胞视图散点图的合成(synthetic)参数进行进一步研究。

[0007] 如上所述,对于本领域的传统的可视化系统,可视化充当过程中的终点,而不能充

当进一步创建用于进一步研究的进一步可视化细化(refinements)的起点。作为示例,来自转移性黑色素瘤患者的免疫细胞样本可以包含T细胞,并且本领域的传统可视化系统将只能识别免疫细胞内的该亚群(subset)。然而,本文描述的创新计算机系统允许对T细胞亚群进行深入的探索和分析,以识别这些T细胞内的多个亚群,例如“耗尽(exhausted)”的T细胞,跟踪这种状态到各个基因,然后这些基因可以被靶向以逆转这种耗尽,激活T细胞,从而可能刺激免疫反应以根除转移,如下面参考示例实施例更详细解释的。

[0008] 因此,通过本文描述的创新可视化技术,计算机技术可以应用于细胞基因表达数据,以发现细胞和基因之间的新关系,并在代表这些关系的细胞基因表达数据内创建新的关联数据结构。

[0009] 通过这些和其他特征,本发明的示例实施例在应用生物信息学领域提供了显著的技术进步。

附图说明

[0010] 图1公开了一个示例计算机系统,其可以用于支持本文描述的创新数据处理和可视化技术。

[0011] 图2A描绘了细胞基因表达数据集的示例。

[0012] 图2B描绘了示例细胞基因表达数据的表格视图。

[0013] 图3描绘了示例细胞视图图形窗口散点图。

[0014] 图4描绘了执行以创建细胞视图散点图的示例过程流程。

[0015] 图5示出了包括参数选择菜单的示例细胞视图散点图用户界面。

[0016] 图6示出了如何在细胞视图散点图用户界面内执行门控以创建细胞群的示例。

[0017] 图7示出了允许用户扩展细胞视图散点图的零点的示例用户界面。

[0018] 图8示出了示例细胞视图散点图,其中并非基因的参数被选作轴参数。

[0019] 图9示出了可以如何在图8的细胞视图散点图内执行门控以创建细胞群的示例。

[0020] 图10示出了用于调整散点图表示中的显示设置的示例用户界面。

[0021] 图11描绘了包括细胞群选择菜单的示例基因视图图形窗口散点图。

[0022] 图12描绘了可以如何旋转细胞基因表达数据以创建基因视图散点图的细胞群数据的示例。

[0023] 图13A和图13B描绘了可以如何在工作区中创建并定义互补细胞群以用于基因视图散点图的示例。

[0024] 图14示出了可以如何在基因视图散点图用户界面内执行门控以创建基因集合的示例。

[0025] 图15示出了可以如何用基因集合作为合成参数来扩充细胞基因表达数据的示例。

[0026] 图16示出了细胞视图散点图的示例,其中基因集合作为选项被呈现在参数选择菜单中。

[0027] 图17A和图17B示出了用于从工作区中的其他基因集合中查看、编辑和创建新基因集合的示例用户界面。

[0028] 图18A-18D示出了用户选择的第三维度可以如何叠加在散点图上的示例。

[0029] 图19示出了在图18A-18D的3D散点图内可以如何执行门控的示例。

[0030] 图20A-20D示出了可以通过系统创建的示例报告。

[0031] 图21示出了系统执行细胞视图模式和基因视图模式之间切换的示例过程流程。

[0032] 图22示出了可以如何操作系统在细胞视图模式和基因视图模式之间切换以支持对细胞数据的调查和研究的示例。

具体实施方式

[0033] 图1公开了一个示例计算机系统100,其可以用于支持本文描述的创新数据处理和可视化技术。示例计算机系统100包括处理器102、存储器104、数据库106和显示器108,它们可以通过互连技术(诸如总线110)彼此通信。

[0034] 处理器102可以采取适于执行本文描述的操作的任何处理器的形式。例如,膝上型电脑或工作站的CPU适合用作处理器102。应当理解,处理器102可以包括多个处理器,包括通过网络彼此通信以执行本文描述的任务的分布式处理器(例如,云计算处理资源)。存储器104可以采取适于在执行本文描述的任务时与处理器102协作的任何计算机存储器的形式。应当理解,存储器104可以采取多个存储器设备的形式,包括遍布网络分布的存储器。类似地,数据库106可以采取处理器102可访问的任何数据存储库(例如,计算机上的文件系统、关系数据库等)的形式,并且应当理解,数据库106可以采取多个分布式数据库(例如,云存储)的形式。显示器108可以采取能够生成本文描述的可视化的计算机监视器或屏幕的形式。

[0035] 可以对细胞基因表达数据112执行本文描述的创新数据分析和可视化技术。可以通过下一代测序(例如,用于测量RNA测序(RNASeq)和单细胞RNA测序(scRNA-Seq)等等的测序方法)来生成细胞基因表达数据112。然而,这仅仅是示例,并且可以使用用于生成细胞基因表达数据112的其他技术。其他示例包括聚合酶链式反应方法,其包括数字液滴(digital droplet)和逆转录酶。还有更多的示例,包括通过流式细胞术(flow cytometry)和微阵列等等(其产生包含DNA和/或RNA的定量的数据文件)进行的RNA测量,或者通过处理原始读取数据(初级和次级分析)以生成基因表达数据文件的软件程序进行的RNA测量。再一个示例是从样本的随机标记(stochastic labeling)中推导出的基因表达数据。随机标记的基因表达数据的示例可以在于2017年9月25日提交的并且标题为“Measurement of Protein Expression Using Reagents with Barcoded Oligonucleotide Sequences”的专利申请15/715,028、以及第9,567,645号和第8,835,358号美国专利中找到,其中的每个的全部公开内容通过引用并入。

[0036] 此外,本文描述的创新数据分析和可视化技术可以通过各种手段应用于从单个细胞生成的数据。单细胞分析可以包括核酸或蛋白质或蛋白质和核酸的任意组合的随机标记。作为示例,本文描述的创新数据分析和可视化技术可以用于分析蛋白质密度或基因表达或其任意组合的定量特征。本文描述的创新数据分析和可视化技术还可以提供根据在单细胞中的各种蛋白质或核酸的随机标记生成的定量数据的改进的可视化。在各个细胞中的生物种群(核酸、蛋白质等)的定量值可以与其他的各个细胞进行比较,或者在细胞类型之间、或者甚至在生成数据的方法之间进行比较。例如,基因表达值可以被可视化为用于生成数据集的方法的函数。本文描述的创新数据分析和可视化技术也可以用于比较由在此描述的在方式方面独立于生成这种数据的方法的、提供用于定量生物种群数据的可视化的各种

手段生成的定量数据。

[0037] 该数据112可以被表征为大型多参数数据集,其在创建有意义的可视化(特别是当考虑到基础生物学时,使得生物学相关信息以视觉方式被有意义地呈现给用户)方面带来了特殊的技术挑战。例如,细胞基因表达数据可以包括大量的单个细胞和细胞群的数据,并且每个细胞或细胞群的参数可以扩展到10,000-30,000个或更多个参数。细胞基因表达数据112可以从数据库106中的文件被读出,并在分析和可视化程序114的执行期间,作为待处理器102操纵的多个数据结构116,被加载到存储器104中。程序114可以包括多个处理器可执行指令形式的处理器可执行计算机代码,这些指令驻留在非暂态计算机可读存储介质(诸如存储器104)上。

[0038] 图2A描绘了细胞基因表达数据集的示例,其中每个细胞(或细胞群)通过细胞ID被识别,并且与多个参数相关联,每个参数具有ID和与细胞ID相关的值。如所指示,细胞的基因表达数据是高维度的,并且无论是每个细胞还是每个细胞群,每个细胞的参数数量可以达到10,000-30,000个或更多个。细胞数据中参数的示例包括受试(subject)细胞中大量基因的基因表达的计数。因此,细胞1的参数1可以对应于基因1,并且其值可以是细胞1中的基因1的表达的计数。类似地,细胞1的参数2可以对应于基因2,并且其值可以是细胞1中的基因2的表达的计数。图2B描绘了示例细胞基因表达数据112的表格视图。表格200中的每一行对应于不同的细胞(参见细胞列),并且标记为基因1、基因2等等的各列对应于不同的基因,并且表格细胞识别每个受试细胞中相应基因的基因表达的计数。该表格还可以包括除了基因之外的参数。例如,细胞基因表达数据112可以包括每个表格细胞中的参数(诸如t分布随机邻域嵌入(tSNE)、主成分分析(PCA)、线性判别分析(LDA)等等)的数据值,其中这些数据值代表分析计算,此分析计算的值记录(capture)各个细胞在n个参数中的差异。可以以多种格式(例如,作为CSV文件、数据库表格(例如,作为关系数据库中的关系数据)、备用数据表示、二进制格式等等)中的任何一种来存储细胞基因表达数据112。

[0039] 图3描绘了可以通过执行程序114产生的示例图形窗口(GW),用于经由显示器108呈现给用户,其中图形窗口可视化了关于细胞群的基因1相对于基因2的关系。这种可视化可以被称为细胞基因表达数据112的细胞视图。如以下所解释的,对细胞视图中的各个细胞的选择进行门控,创建了细胞群。该图形窗口呈现了关于用户选择的文件306中的细胞群的用户选择的两个基因302和304(在该示例中分别是TMEM216和MMP2)的基因表达的散点图300。每个文件可以对应于细胞群的单个样本、单个细胞群(例如,已经通过流式细胞术活性细胞分类和分析进行分类的一个种群)、或者可能是已经联系在一起的多个样本(其来自于不同患者,或者来自于一个患者的不同时间点)。散点图中的每个点308代表受试文件306的细胞群中的细胞。在本示例中,X轴和Y轴的标度指示了相应基因的计数。因此,点308在水平X轴上的位置指示了对应于受试的点308的细胞中存在多少基因MMP2的计数,并且点308在垂直Y轴上的位置指示了对应于受试的点308的细胞中存在多少基因TMEM216的计数。因此,散点图300右上象限中的点308对应于TMEM216和MMP2都被高效表达(highly express)的细胞,而散点图300左下象限中的点对应于TMEM216和MMP2都被低水平地表达的细胞。同样,左上象限对应于其中TMEM216被高效表达而MMP2没有被高效表达的细胞,而右下象限对应于其中MMP2被高效表达而TMEM216没有被高效表达的细胞。如果这些细胞同样地表达两个所选择的基因,则在其中y＝x的对角线处定位对应于细胞的点308。因此,点308在任一方向上

远离该对角线的距离指示给定细胞中所选择的基因的差异表达的程度。预计对于许多细胞群来说,将会有大量细胞,其中这些细胞的所选择的基因的表达是处于零水平的。这导致点308分别在X轴和Y轴上的零水平310和312处大规模聚集。在310和312处,可以使用颜色编码来指示具有这种零水平基因表达的细胞的密度。

[0040] 图4描绘了作为程序114的一部分执行的示例过程流程,其描述了可以如何生成散点图300。在步骤400,处理器在存储器工作区中创建数据结构(参见图1中的116)。该数据结构可以用于保存细胞视图数据。

[0041] 对于第一轴,处理器基于用户输入,在细胞基因表达数据112中选择基因(步骤402)。例如,处理器可以响应于用户输入,在表格200中选择基因列。图5示出了用户界面可以如何向用户呈现每个轴的基因列表。为了访问选择菜单,用户可以分别为Y轴和X轴选择基因选择器302或304。如果我们假设该示例中的第一轴是Y轴,在选择302时,示出了参数选择菜单500,其呈现了可用于相对于Y轴进行选择的参数列表。可以用来自细胞基因表达数据112的参数对该列表进行填充(populate)。如图5所示,针对Y轴基因的用户选择502是TMEM216。在步骤404,处理器用来自细胞基因表达数据112的细胞列表(例如,表格200的细胞列中的细胞)填充数据结构。对于所选择的第一轴基因,每个列出的细胞与其计数值相关联。

[0042] 对于第二轴,处理器基于用户输入,在细胞基因表达数据112中选择另一基因(步骤406)。例如,处理器可以响应于用户输入,在表格200中选择另一基因列。图5示出了用户界面的示例,其中选择304来访问X轴参数选择菜单506(导致对MMP2的选择508)。在步骤408,处理器扩充细胞列表,以将所选第二轴基因的相关联计数值添加到每个细胞。因此,此时,数据结构包括与第一轴和第二轴的所选择的基因的计数对相关联的细胞列表;例如,列表可以包括细胞基因表达数据112中每个细胞的向量集{细胞ID 1,基因1计数,基因2计数}。

[0043] 在步骤410,过程解析该列表,以找到每个所选择的基因的最大值(最高计数)。然后,处理器使用这些最大值来定义散点图中的X轴和Y轴的适当标度(步骤412)。例如,如果X轴基因的最大值是10,则X轴标度可以是0-10。在步骤414,处理器使用细胞的相关计数值作为散点图中的X、Y坐标,基于细胞列表和所定义的标度来绘制散点图。结果是如图3所示的散点图300。

[0044] 回到图3,用户可以访问用户界面中的门创建工具(gate creation tool)320,以创建门,通过该门创建子细胞群。例如,用户可以访问工具320,以在散点图300中绘制包含对应于细胞的点308的子集的形状。图6描绘了示例门600,其被绘制以捕获具有两个选定基因的非零表达的细胞。落在门600的绘制形状内的点308被门控到它们自己的子细胞群内,并且该细胞群可以作为不同的对象602而被添加到工作区。为了创建细胞群,处理器可以将门600转换成门控细胞群的多个边界条件。例如,关于如门600所示的矩形形状,边界条件可以是所有细胞具有(1)在1和10之间的X轴值,和(2)在1和8之间的Y轴值。可以遍历细胞列表数据结构(例如,向量集{细胞ID 1,基因1计数,基因2计数}),以找到满足这些标准的所有细胞ID,并且这些细胞ID的数据可以在工作区中填充新的子细胞群数据结构。此外,随后可以呈现子细胞群(其中该细胞群包括一个或更多个对应于细胞的点308)的基因表达的相应散点图300。该门控可以允许用户聚焦于散点图300中的在生物学方面所关注的成群的对应

于细胞的点308。

[0045] 此外,图3的细胞视图模式允许用户使用导航工具322可视地比较不同的样本,包括子细胞群和父细胞群。例如,通过工具322中的后退(back)和下一步按钮,用户可以导航到工作区中的下一个和前一个样本的散点图300。通过工具322中的向下按钮,用户可以导航到分析层级中的子细胞群(例如,导航到经由图6的门600创建的子细胞群)。此外,通过向上按钮(未示出),用户可以导航到分析层次中的父细胞群。

[0046] 此外,当处于细胞视图模式时,预计对于许多细胞群来说,将会有大量细胞,其中这些细胞的所选择的基因的表达是处于零水平的。这导致分别在图3示出的散点图300的X轴和Y轴上的零水平310和312处大规模聚集点308。在310和312处可以使用颜色编码,以用这种零水平基因表达来指示细胞的密度。然而,发明人相信,通过扩展两个所选择的基因(或如以下所解释的基因集合)的零水平的可视化,可以增强显示,以在某些情况下向用户提供生物学相关的信息。为了实现这一点,用户界面可以包括如图7所示的“展开零点(spread zeros)”用户控件700(诸如,复选框(check box)、按钮等)。该用户控件700可以被设置在用于选择轴参数的菜单上,然而如果专业人员需要,用户控件700可以被定位在其他地方,诸如图3所示的用户界面上的某个地方。图7的上半部分示出了零点未展开时的散点图。图7的下半部分示出了当经由用户控件700选择展开零点选项时的散点图。

[0047] 如图7中的底部散点图所示,框702提供了细胞的扩展视图,其呈现了对于X轴基因(在本示例中为MMP2)的表达的零值。框704提供了细胞的扩展视图,其呈现了对于Y轴基因(在本示例中为TMEM216)的表达的零值。沿着X轴和Y轴的零空间的深度可以被定义为每个基因在零水平处的点/细胞的数量的函数,并且点/细胞可以作为在每个零位置处的细胞密度的函数而分布在由框702和704定义的零空间上,然而也可以使用其他分布技术。应当理解,图7的底部散点图的左下象限(由框702和704的叠加来定义)包括对X轴和Y轴基因均具有零表达的细胞。该散点图的右下象限包括这样的细胞:其具有Y轴基因的零表达,但具有X轴基因的正表达(positive expression),并且该散点图的左上象限包括这样的细胞:其具有X轴基因的零表达,但具有Y轴基因的正表达。该散点图的右上象限包括对两个基因都具有正表达的细胞(实际上是在图7的顶部部分示出的散点图)。从图7的底部散点图可以看出,右上象限的细胞群比其他象限稀疏得多。发明人相信,以这种展开方式使零水平可视化的能力可以为用户提供生物学相关的信息(诸如,在出现未预计到的分布的情况下(例如,右上象限不像通常预计的那样稀疏)评估基因组合的能力)。

[0048] 细胞视图散点图还可以显示细胞基因表达数据112中除基因之外的选定参数的细胞信息。如以上所指示的,细胞基因表达数据112可以包括来自维度降低的参数(诸如,根据tSNE、LDA、PCA等得出的参数)和质量控制参数(高于阈值、与核糖体RNA(rRNA)丰度(abundance)相关的参数等)。这些参数可以被呈现为参数选择菜单500和506上的选项。图8示出了一个示例,其中Y轴的选择800是参数tSNE axis 2,并且其中X轴的选择802是参数tSNE axis 1。可以经由图4的过程流程来创建图8的最终散点图。如以上参考图6所解释的,用户也可以在图8的散点图中门控所期望的细胞群(参见图9中的门900)。

[0049] 图3的细胞视图用户界面还可以包括用户控件330(例如,图3的“T”按钮),用户通过该控件可以改变散点图300中的信息显示。当选择T按钮330时,可以呈现图10的用户界面。通过该界面,用户可以交互式地且实时地改变散点图数据的舍弃(binning)/显示。图10

右侧的列表示出了针对散点图的X轴的可用参数和选定的参数。直方图1000示出了关于受试细胞群的所选择的参数的值的舍弃。给定零的高度普遍性,直方图1000在零水平处显示大尖峰(spike),并且该示例的标度模糊了直方图中的非零水平。通过示出的图10顶部的朝左/朝右箭头,用户可以快速预览不同样本的显示。通过图10中部的+/-控件,用户可以轻松更改直方图缩放级别。

[0050] 通过标度控件1002,用户可以以多种方式调整散点图的X轴。例如,可以将X轴标度定义为展示线性标度或诸如log2标度的一些其他标度(参见控件1004)。此外,通过最小/最大控件1006,用户可以定义在X轴上的最小和最大边界。例如,通过这些控件,最小值可以被定义为大于零的值,这将从直方图中去除零尖峰,并重新呈现新缩放的直方图,其中可以更清楚地看到非零值在X轴上的分布。滑块1008可以为用户提供对转换变量的简单控制。此外,应当理解,可以提供附加的转换选项,包括用户提供的转换,可以通过用户输入来调整其变量(例如,参见标题为“Plugin Interface and Framework for Integrating External Algorithms with Sample Data Analysis Software”的美国专利申请公开2016/0328249,其全部公开内容通过引用并入本文)。

[0051] 本文公开的本发明系统的一个特别创新且强大的方面是将散点图显示从细胞视图模式旋转到基因视图模式的能力。图11示出了基因视图模式下的散点图1100的示例。如本文中所使用的,“基因视图”可视化是指这样的可视化:其中基因是相对于轴维度测量的各个数据点(例如,基因是相对于两个细胞群测量出的散点图中的点)。在散点图1100中,轴参数是细胞群(在本示例中,参见指示B细胞群的X轴参数1102和指示“非B”细胞群的Y轴参数1104)。散点图1100中的点1106代表特定基因(而不是图3的散点图300中的各个细胞)。种群选择菜单1110提供可用于选择将要在X轴上使用的细胞群的列表,并且种群选择菜单1112提供可用于选择将要在Y轴上使用的细胞群的列表。如上所述,在该示例中,选择1114和1116分别对应于B细胞群和“非B”细胞群。

[0052] 凭借这种旋转,参考细胞基因表达数据112(诸如图2中的表格200),可以旋转表格200,使得基因变成表格中的行,并且细胞的子集被分组为两个细胞群,这两个细胞群变成表格中的列。此外,可以对表格200中的两个细胞群的细胞的基因计数进行计算,以确定将填充已旋转的表格的细胞的值。通过图11的控件1120,用户可以定义将对基因数据执行的计算,以计算在已旋转的表格中的细胞的值。在图11的示例中,选择了归一化的平均值计算。然而,应该理解,其他计算选项也是可用的,例如归一化中值、归一化模式、直接平均值、直接中值、直接模式等等。因此,应当理解,对已旋转的表格数据执行的计算可以是被认为与用户生物学相关的任何由用户定义的函数。发明人注意到,可能希望计算是基于考虑到两个细胞群的细胞计数中的潜在差异的某种类型的平均和/或考虑到在样本之间的可变性的归一化(例如,外部RNA质控联盟(ERCC)质控品(controls)的刺入(spike-in)的归一化,该ERCC质控品包含已知量的将要被测量的RNA)。例如,如果已旋转的表格值是两个细胞群中的每个基因的直接计数,并且如果两个细胞群中的细胞计数存在有意义的差异,则两个细胞群中的累计基因计数在比较意义上不会非常有益(informative)。然而,如果细胞群在大小上大致相似,则两个细胞群中的每个基因的直接基因计数可能仍然是有益的。

[0053] 图12描绘了从细胞视图到基因视图的示例旋转。图12中的表格200示出了样本中许多细胞的每个细胞的基因表达。每行对应于不同的细胞,并且列对应于相关联细胞中的

不同基因的表达计数。如果这些细胞的子集如图12所示被分组成两个细胞群1210和1212(例如,经由细胞视图散点图300中的门控),则这两个细胞群可以如图所示被旋转以创建已旋转的表格1200。在已旋转的表格1200中,行是基因,该基因在表格200中是列。表格1200中的列是两个细胞群1210和1212。每个表格细胞都凭借每个细胞群细胞中的相关联基因的基因计数的联系(concatenation)而进行填充。在图12的示例中,对这些值执行的计算是直接平均,以计算平均值(如已旋转的表格1200a所示)。再次,如上所述,应当理解,如果专业人员需要,可以执行其他计算。然后,可以使用与图4中描述的基本技术相同的技术(尽管使用已旋转的表格1200a而不是表格200),根据已旋转的表格1200a来创建图11中的散点图1100所示类型的基因视图散点图。

[0054] 回到图11,基因视图散点图1100因此以一种强大的方式向用户提供了在尺寸可能非常大的基因列表上可视化在两个细胞群之间的差异基因表达。发明人相信散点图1100代表了可视化细胞基因表达数据的开拓性新方法,其向专业人员打开了一系列新的广泛的调查选项,下面描述了其中的一些示例。图11的示例基因视图散点图1100为用户提供了在指定x/y轴种群和随着种群变化动态更新基因表达图方面的灵活性。示例基因视图散点图1100还向用户提供了大量的图形选项和分辨率。重要的是,分析不止于此图;它继续有能力创建新的基因集合,并且如下所述更深入地挖掘数据。没有这种方法,就不可能通过特定的新的单细胞方法来揭示检查细胞异质性和在样本/患者之间的差异。

[0055] 如果根据经由1120定义的度量,在两个细胞群中同样地表达基因1106,则将基因1106定位到y＝x的对角线。因此,基因1106在任一方向上与该对角线相距的距离指示了受试基因1106在两个所选择的细胞群之间的差异表达的程度。假设预计大多数基因1106将不会在许多细胞群中被表达(或者仅会被轻微表达),则预计在散点图1100的左下象限中通常将会有大群的基因1106。

[0056] 此外,如结合图6所述,由于用户可以创建多个分层相关的细胞群,这些细胞群作为数据对象存在于工作区中,所以用户可以在基因视图模式下选择这些细胞群中的任何一个。此外,应当理解,工作区可以用于创建细胞群的布尔表示式(例如,所有“非B细胞”的细胞群的或(OR)组合,以创建互补细胞群),用于在基因视图模式下的探索。图13A和13B示出了工作区中这种互补布尔细胞群的示例。因此,用户可以创建散点图1100,散点图1100示出了在B细胞群和“非B”细胞群之间的差异基因表达。

[0057] 本文描述的基因视图模式的另一个强大且创新的方面是用户在基因视图模式下门控基因从而创建基因集合的能力。图14示出了这样的示例:其中,在基因视图散点图中创建了门1400,以捕获用户认为值得进一步调查的基因集合。可以使用以上结合图6描述的技术在散点图上绘制门1400。该门控将在工作区中创建对应于由门1400定义的基因集合(其中基因集合包括一个或更多个基因,这取决于用户定义的门1400包含多少基因1106)的数据结构。然后,这样的基因集合可以用作与细胞基因表达数据112相关的合成参数,其可以在细胞视图模式下被选择用于可视化。在图14的示例中,门1400是梯形形状,其捕获具有在B细胞和非B细胞中的正差异表达的基因1106。正差异表达的程度由用户通过梯形的对角线与散点图的y＝x对角线的距离来控制。

[0058] 图15示出了可以如何用作为合成参数的基因集合来扩充细胞基因表达数据112的示例(以示例扩充的表格200的形式)。该表格200已经用两个基因集合作为参数进行了扩充

(参见基因集合1和基因集合2列)。在本示例中,基因集合1由{基因2,基因3}组成,并且基因集合2由{基因1,基因4}组成。这些基因集合(各自对应于表格行中的不同细胞)的表格细胞然后可以用受试基因集合的组成基因的基因计数的总和来填充。因此,用户可以返回到细胞视图模式以创建细胞视图散点图,其中,在X轴参数和Y轴参数中的一个或两个是基因集合。图16中示出这个的实例。图16中每个轴的参数选择菜单包括列出可用基因集合选项的部分1600。因此,图16的细胞视图散点图显示了对应于细胞的点308的散点图,其中,X轴参数1602是标记为"B_GeneTable"的基因集合,而Y轴参数1604是标记为"HighTHighB"的基因集合。因此,应当理解,用于创建基因集合的在基因视图散点图内的门控能力与将所创建的基因集合用作细胞视图散点图中所使用的合成参数的能力相结合,为用户提供了用于智能地减少细胞基因表达数据的多参数数据空间的前所未有的能力。这种结合使得用户能够克服识别在种群中的共同基因和识别在种群中的差异基因的极其困难的问题。换句话说,无论用户是在寻找哪些基因是共同的,还是寻找哪些基因是不同的,创建新的基因集合并将它们作为单个参数进行查看/分析的能力允许用户关注在细胞群和基因集合之间的重要关系。此外,当用户基于细胞群的各种比较来创建在生物学方面关注的基因集合时,可以通过将这些基因集合(作为被定义的所包含的基因的列表)传递给其他用户来与其他用户共享这些基因集合,使得这些其他用户可以用他们的细胞数据样本来评估基因集合。基于对不同细胞数据集的这种共享和独立研究,预计可以获得对关于细胞的基因行为的更深入了解。

[0059] 图17A示出了工作区的示例视图,其将工作区分解成存在于工作区内的各种基因集合1700、样本1702和细胞群1704。如图所示,列出基因集合1700的部分可以包括提供关于每个受试基因集合的元数据的各种显示字段(例如,基因集合内的基因的名称、计数和对基因集合的描述)。用户可以编辑名称和描述,这将有助于向用户通知基因集合的相关特征。此外,可以以指示基因集合内存在的任何层次关系的方式列出基因集合。

[0060] 基因集合控件1710(其在图17B中被更详细地示出)向用户提供了经由布尔运算根据其他基因集合创建基因集合的能力。例如,经由并集(union)按钮,用户可以创建新的基因集合,该基因集合是基因集合1700的列表内的两个或更多个选定基因集合的并集。经由交集(intersection)按钮,用户可以创建新的基因集合,该基因集合由基因集合1700的列表内两个或更多个所选择的基因集合中的两个/全部集合中所存在的基因组成。经由互补(complement)按钮,用户可以根据基因集合1700的列表内的两个基因集合(基因集合1和基因集合2)创建互补基因集合,其中,第一互补的新基因集合由在基因集合1(但不是基因集合2)中的所有基因组成,并且其中第二互补的新基因集合由在基因集合2(但不是基因集合1)中的所有基因组成。经由"全比较(all comparisons)"按钮,用户可以经由上述技术中的每种技术来同时创建几个新的基因集合(经由并集运算、交集运算和互补运算操作来创建新的基因集合)。这种根据现有基因集合创建新基因集合的灵活能力允许在治疗条件、患者、实验等之间比较基因集合。因此,通过使用控件1710,用户可以创建包括基因集合的组合的基因集。

[0061] 由程序114提供的可视化的另一个强大且创新的方面包括在细胞视图和/或基因视图散点图上叠加第三维度的能力。例如,可以将颜色编码(例如,热图)应用于散点图中的对应于细胞的点308或基因1106,以向数据呈现提供另一个维度。图18A-18D示出了这种情

况的示例。图18A示出了可以如何使用第三维度控件1800来定义如何在示例细胞视图散点图显示中呈现第三维度(例如,作为热图统计量)。通过控件1804,用户可以定义用于热图的统计量。图18B示出了在示例实施例中可用于该统计量1802的选项1806(例如,中值、平均值、几何平均值、变异系数(CV)、稳健CV、标准差(SD)、稳健SD等等)。热图统计量可以计算如下:图中的每个点代表一个或更多个细胞(通常是多个细胞)。对于每个点,为细胞(多个细胞)计算所选择的统计量(例如,平均值、中值等)。因此,如果图中有400个点,则计算400个统计值。对于400个统计值,确定最小值和最大值,并被用作索引到颜色图(其梯度从一种颜色变化到另一种颜色的颜色值的阵列)中的下限和上限,即最小值被映射到颜色阵列中的索引0,并且最大值被映射到颜色图中的最后一个索引。然后,颜色图中的颜色可以以某种方式被应用于在最小值和最大值之间的值(诸如,颜色对数值的线性分布)。

[0062] 图18C示出了用户如何从细胞基因表达数据中选择参数1808,以供用作第三维度。可以用细胞基因表达数据的参数的列表(例如,表格200中的列)对参数列表1810进行填充。在该示例中,用户已选择参数HighTHighB作为第三维度,如图18D所示,其经由细胞视图散点图上的颜色编码进行叠加。该第三维度叠加为用户提供了进一步有意义地解释散点图的能力。例如,用户可以使用颜色编码来指示感兴趣的细胞群,并门控这样的细胞群(参见图19中的门1900),从而在工作区中创建作为细胞群的另一个新对象,以供进一步调查。

[0063] 此外,根据所公开的系统的另一方面,可以通过将细胞群从工作区拖到报告编辑器中,来在报告编辑器中创建报告(参见图20A)。作为另一个示例,可以通过将一个细胞群拖到另一个细胞群的上面来叠加细胞群,从而创建2D叠加,并且按照基因集合来创建热图。图20A-20D中以两种方式示出了细胞群的叠加。在散点图中,不同的种群在同一个图中被显示为不同颜色的点图(并且您可以选择哪个种群位于另一个种群之上)。在热图中,不同的种群被水平地组织和附加,以作为热图中的附加列,即,种群1的所有细胞被一起渲染和标记,随后是种群2的所有细胞,等等。图20B示出了按基因集合和细胞群划分的示例热图。图20C示出了图20B最左边部分的放大视图,并且图20D示出了图20C最右边部分的放大视图。

示例用例:

[0064] 如以上所指示,所公开的系统通过在细胞视图模式和基因视图模式之间切换(反之亦然),同时在这两种视图模式中执行门控以聚焦于感兴趣的数据,为调查细胞基因表达数据112提供了强大的机制。

[0065] 图21描绘了作为程序114的一部分执行的示例过程流程,该示例过程流程描述了用户可以怎样在细胞视图模式和基因视图模式之间切换,以支持对细胞基因表达数据的深入调查。在步骤2100,处理器响应于用户输入,从工作区中选择样本。在步骤2102,做出关于是在细胞视图模式下还是在基因视图模式下操作的决定。可以响应于用户输入而做出该决定。如果选择了细胞视图模式,则过程流程前进到步骤2104。如果选择了基因视图模式,则过程流程前进到步骤2122。步骤2104-2120对应于在细胞视图模式下的操作,而步骤2122-2140对应于在基因视图模式下的操作。

[0066] 在步骤2104,处理器响应于用户输入而选择细胞视图轴参数(例如,诸如基因或基因集合的参数的选择)。在步骤2106,处理器基于在步骤2104所选择的轴参数,根据细胞基因表达数据112,生成细胞视图散点图数据结构。该步骤可以包括选择对应于表格200中所选择的参数的列,以获得细胞列表以及该细胞列表针对每个所选择的参数的相关联的值。

在步骤2108,处理器根据在步骤2106创建的细胞视图散点图数据结构内的数据生成细胞视图散点图300,以用于呈现给用户。

[0067] 在步骤2110,处理器响应于来自用户的输入,接收关于细胞视图散点图的门规范。该门控创建了细胞群(步骤2112),其中所创建的细胞群被保存为工作区中的新数据结构。此时,用户可以选择是(1)处理新样本(参见步骤2114,并且行进回到步骤2100),是(2)在细胞视图模式下相对于当前样本定义一个或更多个新轴参数(参见步骤2116,行进回步骤2104),是(3)在细胞视图模式下定义关于当前样本的新门(参见步骤2118,并且行进回到步骤2110),还是(4)切换到基因视图模式(参见步骤2120,并且进行到步骤2122)。

[0068] 在步骤2122,如以上讨论地,处理器旋转细胞基因表达数据112。然后,在步骤2124,处理器响应于用户输入,从工作区中选择细胞群。在步骤2126,处理器基于在步骤2124选择的细胞群,根据已旋转的细胞基因表达数据,生成基因视图散点图数据结构。该步骤可以包括在表格200的已旋转版本中选择对应于所选择的细胞群的已旋转的列,以获得每个所选择的细胞群的基因列表及其相关联的度量。在步骤2128,处理器根据在步骤2126创建的基因视图散点图数据结构内的数据生成基因视图散点图1100,以用于呈现给用户。

[0069] 在步骤2130,处理器响应于来自用户的输入,接收关于基因视图散点图的门规范。该门控创建基因集合(步骤2132),其中,所创建的基因集合被保存为工作区中的新的合成参数,以用于与细胞基因表达数据相关联。因此,可以用对应于在步骤2132创建的基因集合的新数据值来扩充细胞基因表达数据。此时,用户可以选择是(1)处理新样本(参见步骤2134,并且行进回到步骤2100),是(2)在基因视图模式下相对于当前样本定义一个或更多个新细胞群(参见步骤2136,行进回步骤2124),是(3)在基因视图模式下定义关于当前样本的新门(参见步骤2138,并且行进回到步骤2130),还是(4)切换到细胞视图模式(参见步骤2140,并且进行到步骤2104)。

[0070] 因此,图21示出了用户可以如何使用该系统在细胞视图模式和基因视图模式之间快速转换,同时经由可用于在各模式之间切换之后帮助可视化的门控来分别创建细胞群和基因集合。

[0071] 作为示例,图22的示例过程流程示出了一种强大且创新的操作模式,其中,用户经由图21的过程流程与程序114交互,以(1)在显示的细胞视图模式内创建一个或更多个细胞群(步骤2200),(2)旋转到基因视图模式,该模式比较性地显示了跨多个细胞群的基因表达(步骤2202),(3)在显示的基因视图模式内创建一个或更多个基因集合(步骤2204),该基因视图模式用新的基因集合(多个基因集合)作为合成参数来扩充细胞基因表达数据(步骤2206),(4)使用这些基因集合中的一个或更多个作为细胞视图散点图的轴参数(多个轴参数),旋转回到细胞视图散点图(步骤2208),以及(5)根据需要,迭代地重复这些操作,以深入到可能存在于细胞基因表达数据112内的生物学相关的关系中(并且其中可以通过在细胞视图或基因视图数据显示上叠加用户定义的第三维度来辅助操作(1)和(3))。例如,所公开的系统可以用作研究精确/个性化药物的强大工具,以执行工作(诸如,评估来自众多患者的细胞数据,以找到具有与各种癌症或其他疾病/病理的存活和治疗反应密切相关的基因的患者)。

[0072] 例如,通过本文提供的工具,用户可能能够分析细胞群,以识别与关于特定癌症X的更好存活机会相关的、差异性地表达的基因集合(我们可以将其标记为“存活基因集

合”)。与此同时,用户可能能够分析细胞群,以识别与关于癌症X低存活机会相关的、差异性地表达的基因集合(我们可以将其标记为“非存活基因集合”)。此外,用户可能能够分析细胞群,以识别与针对癌症X的治疗Y的反应良好相关的、差异性地表达的基因集合(我们可以将其标记为“治疗反应基因集合”)。然后,这些基因集合可以在系统的细胞视图模式中被用作合成参数,以在基因上倾向于对治疗Y的疗法(treatment)有良好反应的患者中找到细胞群,从而幸免于癌症X。例如,细胞视图散点图可以将存活基因集合用作一个轴参数(例如,X轴参数),并且将非存活基因集合用作另一个轴参数(例如,Y轴参数),同时将治疗反应基因集合用作第三维度叠加。所得出的散点图可以显示与存活和治疗反应性良好相关的细胞群(以及与存活或治疗反应性不良好相关的细胞群)。

[0073] 虽然上面已经结合本发明的示例实施例描述了本发明,但是可以对其进行各种修改,这些修改仍然落在本发明的范围内。例如,虽然本文所示的示例散点图将X轴表示为水平轴,并且将Y轴表示为垂直轴,但是应当理解,一些专业人员可能发现倾斜散点图是合乎需要的。一个示例是对角线y＝x被认为具有生物学重要性的场景。在这种情况下,散点图可能会倾斜,使得y＝x对角线呈现为水平线或垂直线,而不是45度线,从而帮助用户关注数据可能离y＝x线有多远。因此,应当理解,通过阅读本文的教导,对本发明的这些和其他修改将是可识别的。

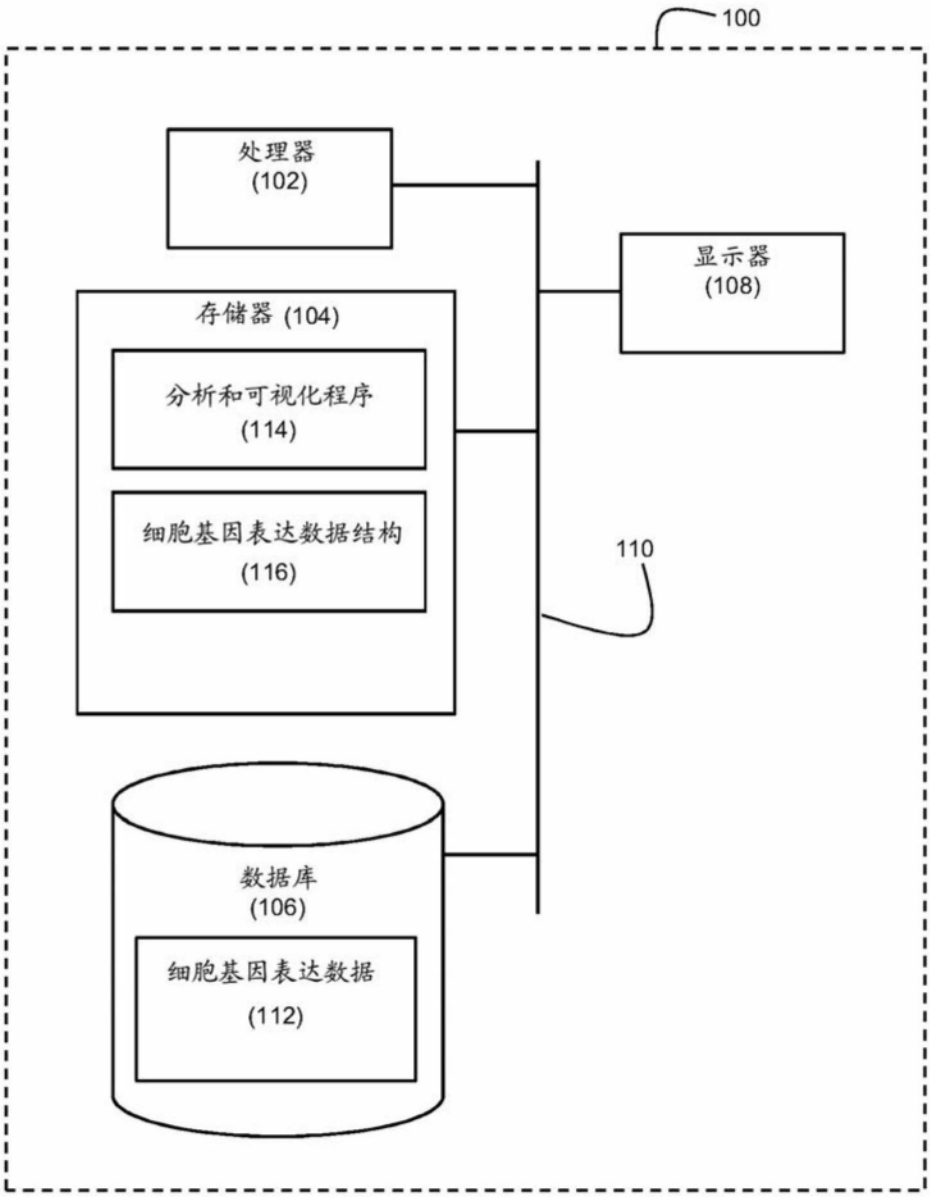

100
处理器
(102)
显示器
(108)
存储器 (104)
分析和可视化程序
(114)
细胞基因表达数据结构
(116)
110
数据库
(106)
细胞基因表达数据
(112)

图1

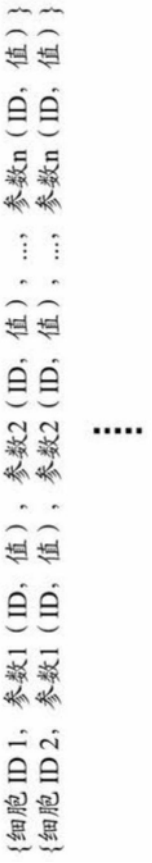


图2A

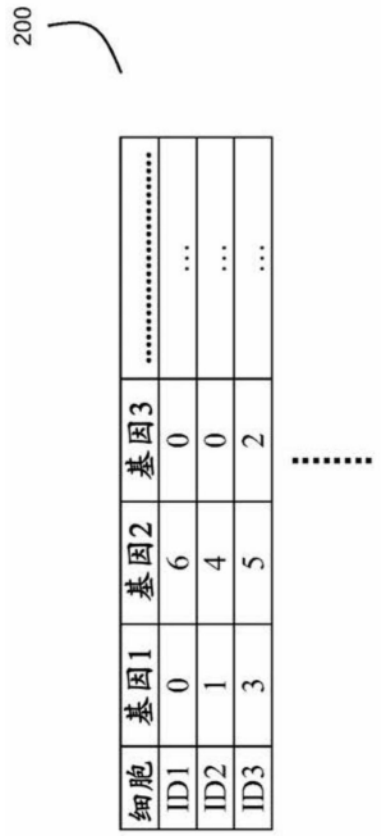


图2B

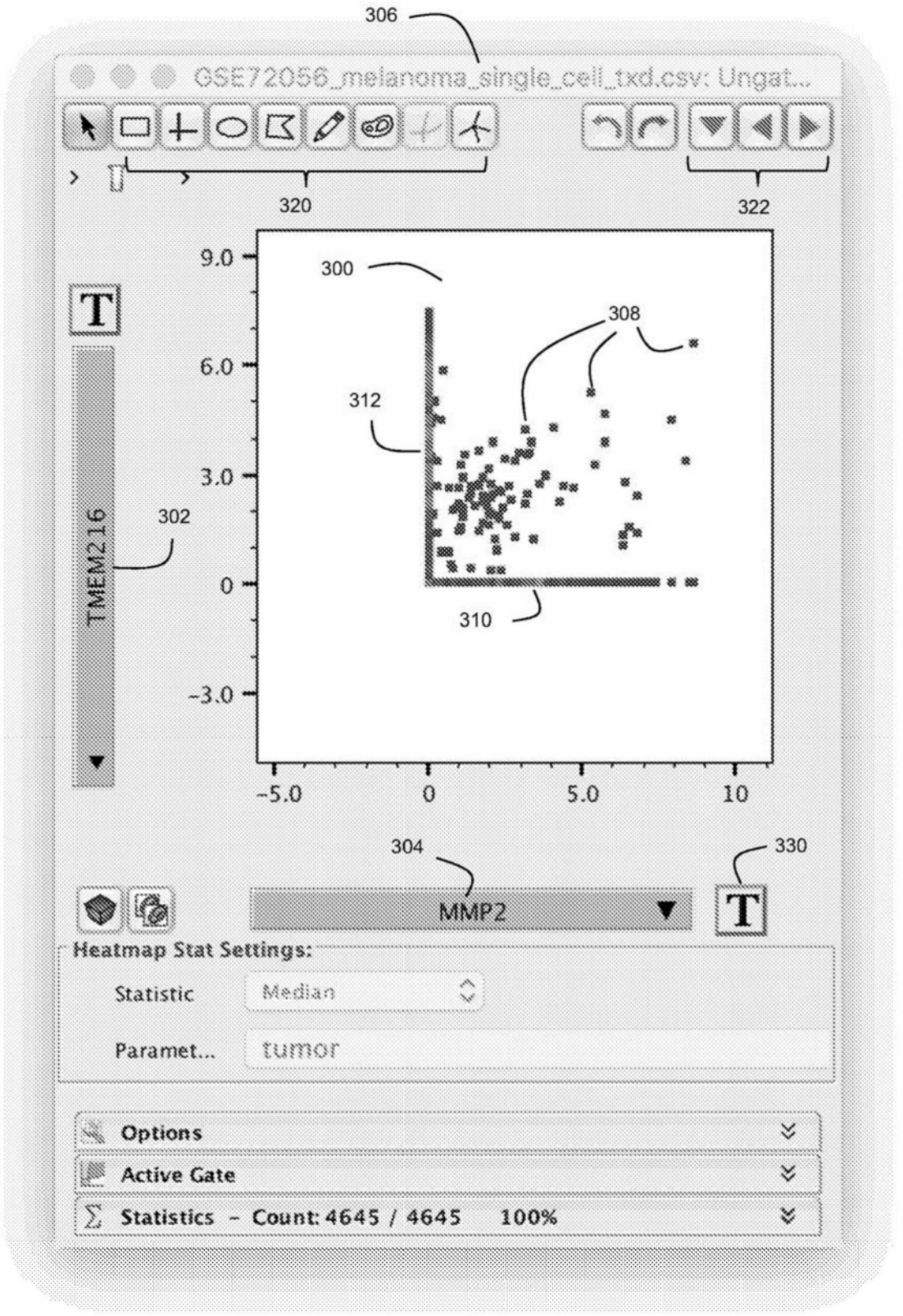

306
GSE72056_melanoma_single_cell_txd.csv: Ungat...
320
322
300
308
312
302
TMEM216
9.0
6.0
3.0
0
-3.0
310
-5.0
0
5.0
10
304
330
MMP2
Heatmap Stat Settings:
Statistic
Median
Paramet...
tumor
Options
Active Gate
Statistics - Count: 4645 / 4645 100%

图3

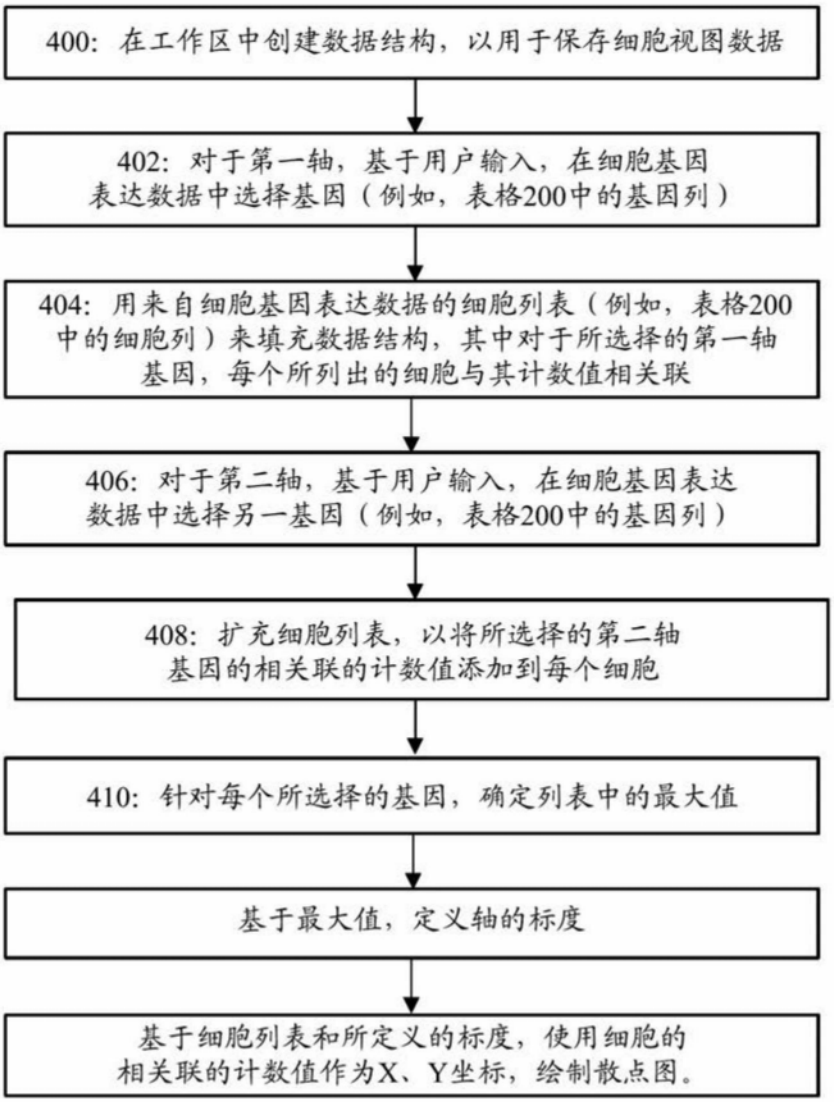

400：在工作区中创建数据结构，以用于保存细胞视图数据
402：对于第一轴，基于用户输入，在细胞基因表达数据中选择基因（例如，表格200中的基因列）
404：用来自细胞基因表达数据的细胞列表（例如，表格200中的细胞列）来填充数据结构，其中对于所选择的第一轴基因，每个所列出的细胞与其计数值相关联
406：对于第二轴，基于用户输入，在细胞基因表达数据中选择另一基因（例如，表格200中的基因列）
408：扩充细胞列表，以将所选择的第二轴基因的相关联的计数值添加到每个细胞
410：针对每个所选择的基因，确定列表中的最大值
基于最大值，定义轴的标度
基于细胞列表和所定义的标度，使用细胞的相关联的计数值作为X、Y坐标，绘制散点图。

图4


500
Y Axis Gene
hg19.Genes
hg19.NovelIntersectionGenes
hiMachiNK
hiThiB
hiThiB2
mm10.Genes
------ Genes ------
tumor
malignant(1=no,2=yes,0=unresoh
non-malignant cell type (1=T,2=B
C9orf152
RPS11
ELMO2
CREB3L1
PNMA1
MMP2
TMEM216
TRAF3IP2-AS1
LRRC37A5P
LOC653712
C10orf90
ZHX3
502
TMEM216
302
MMP2
304
Heatmap Stat Settings
Statistic
Parameter
Options
Active Gate
Statistics - Count 4645 / 4645 100%
X Axis Gene
Spread Zeroes
506
hg19.CellCycle.S.Genes
hg19.Genes
hg19.NovelIntersectionGenes
hiMachiNK
hiThiB
hiThiB2
mm10.Genes
------ Genes ------
tumor
malignant(1=no,2=yes,0=unresoh
non-malignant cell type (1=T,2=B
C9orf152
RPS11
ELMO2
CREB3L1
PNMA1
MMP2
TMEM216
TRAF3IP2-AS1
LRRC37A5P
LOC653712
C10orf90
508

图5

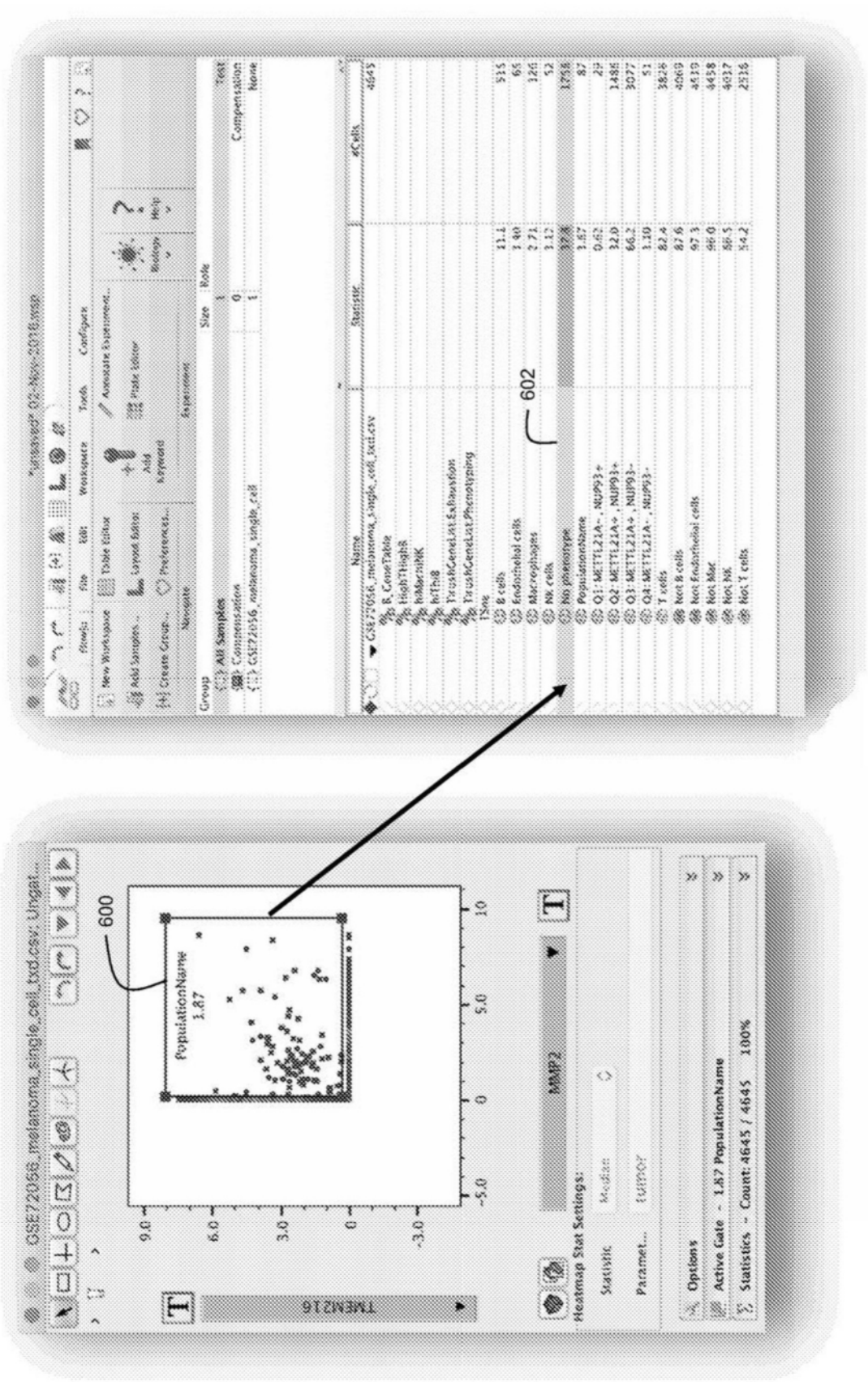

600
602
PopulationName
1.87
TMEM216
MMP2
Heatmap Stat Settings:
Statistic
Median
Paramet...
Tumor
Options
Active Gate - 1.87 PopulationName
Statistics - Count: 4645 / 4645 100%

图6

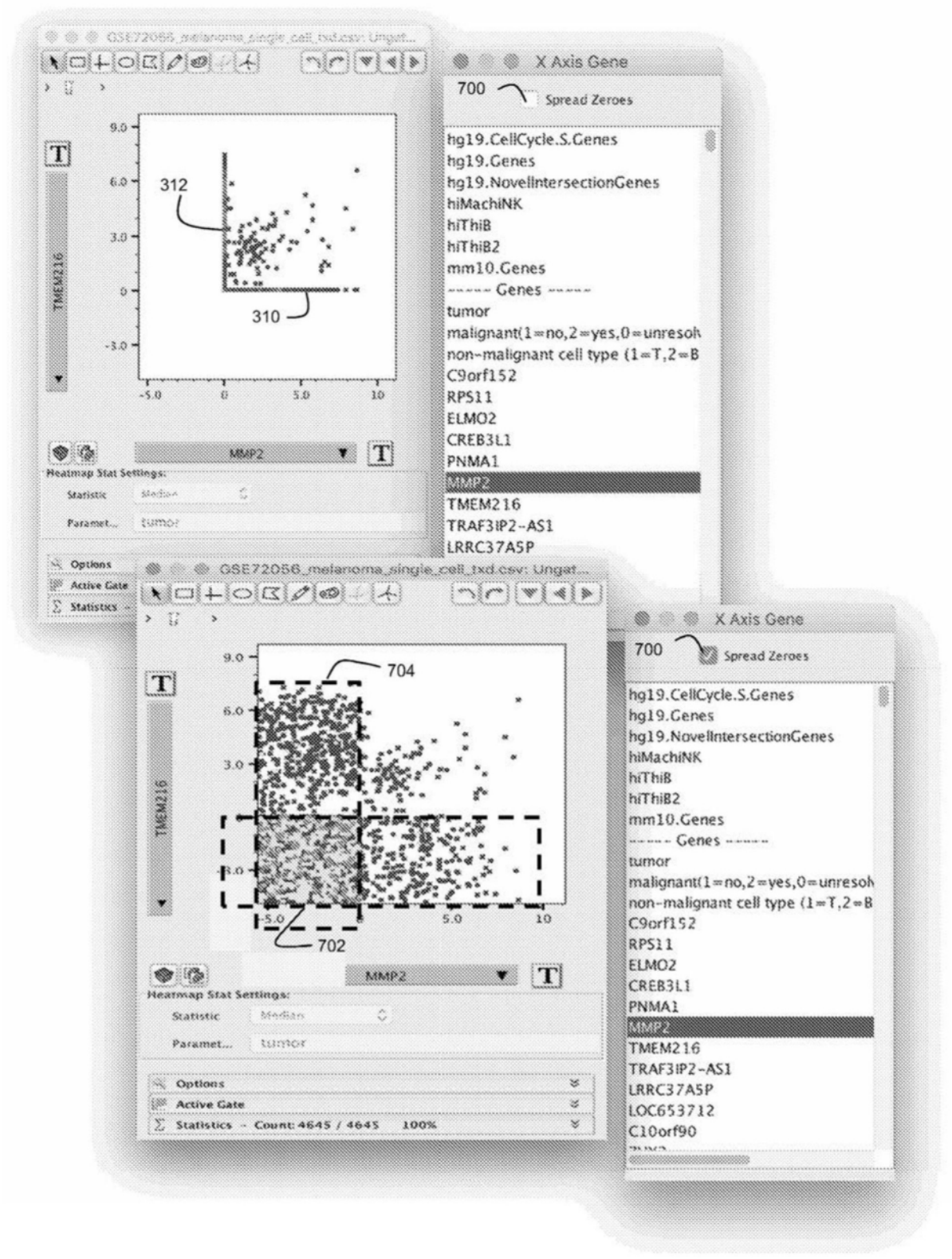

X Axis Gene
700
Spread Zeroes
hg19.CellCycle.S.Genes
hg19.Genes
hg19.NovelIntersectionGenes
hiMachiNK
hiThiB
hiThiB2
mm10.Genes
------ Genes ------
tumor
malignant(1=no,2=yes,0=unresoh
non-malignant cell type (1=T,2=B
C9orf152
RPS11
ELMO2
CREB3L1
PNMA1
MMP2
TMEM216
TRAF3IP2-AS1
LRRC37A5P
LOC653712
C10orf90
312
310
704
702
TMEM216
MMP2
Heatmap Stat Settings:
Statistic
Paramet...
Options
Active Gate
Statistics - Count: 4645 / 4645 100%

图7

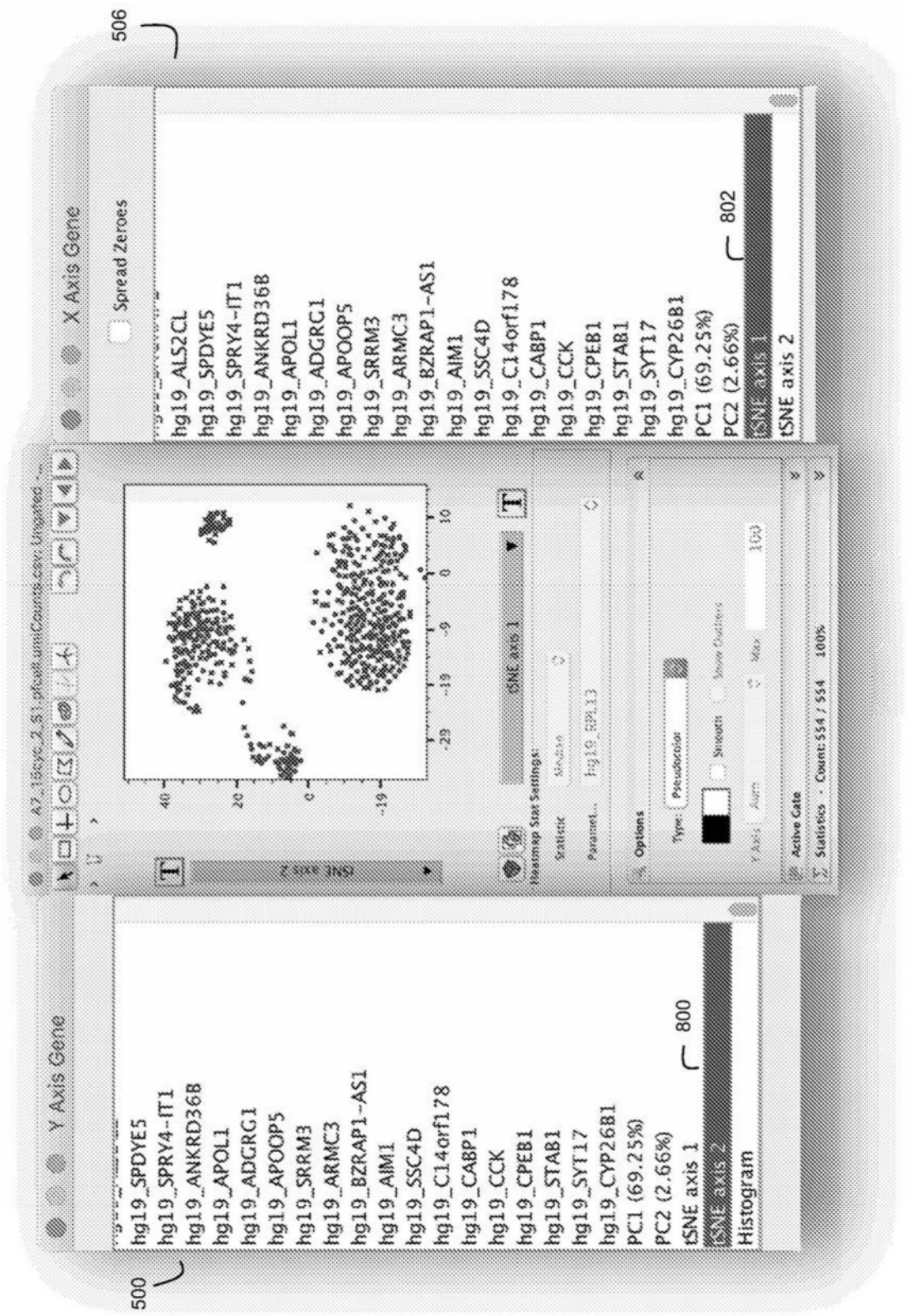

500
Y Axis Gene
hg19_SPDYE5
hg19_SPRY4-IT1
hg19_ANKRD36B
hg19_APOL1
hg19_ADGRG1
hg19_APOOP5
hg19_SRRM3
hg19_ARMC3
hg19_BZRAP1-AS1
hg19_AIM1
hg19_SSC4D
hg19_C14orf178
hg19_CABP1
hg19_CCK
hg19_CPEB1
hg19_STAB1
hg19_SYT17
hg19_CYP26B1
PC1 (69.25%)
PC2 (2.66%)
tSNE axis 1
tSNE axis 2
800
Histogram
506
X Axis Gene
Spread Zeroes
hg19_ALS2CL
hg19_SPDYE5
hg19_SPRY4-IT1
hg19_ANKRD36B
hg19_APOL1
hg19_ADGRG1
hg19_APOOP5
hg19_SRRM3
hg19_ARMC3
hg19_BZRAP1-AS1
hg19_AIM1
hg19_SSC4D
hg19_C14orf178
hg19_CABP1
hg19_CCK
hg19_CPEB1
hg19_STAB1
hg19_SYT17
hg19_CYP26B1
PC1 (69.25%)
PC2 (2.66%)
802
tSNE axis 1
tSNE axis 2
Heatmap Stat Settings
Statistic
Parameter...
Options
Type:
Pseudocolor
Smooth
Show Outliers
Active Cate
Statistics - Count:554 / 554 100%

图8

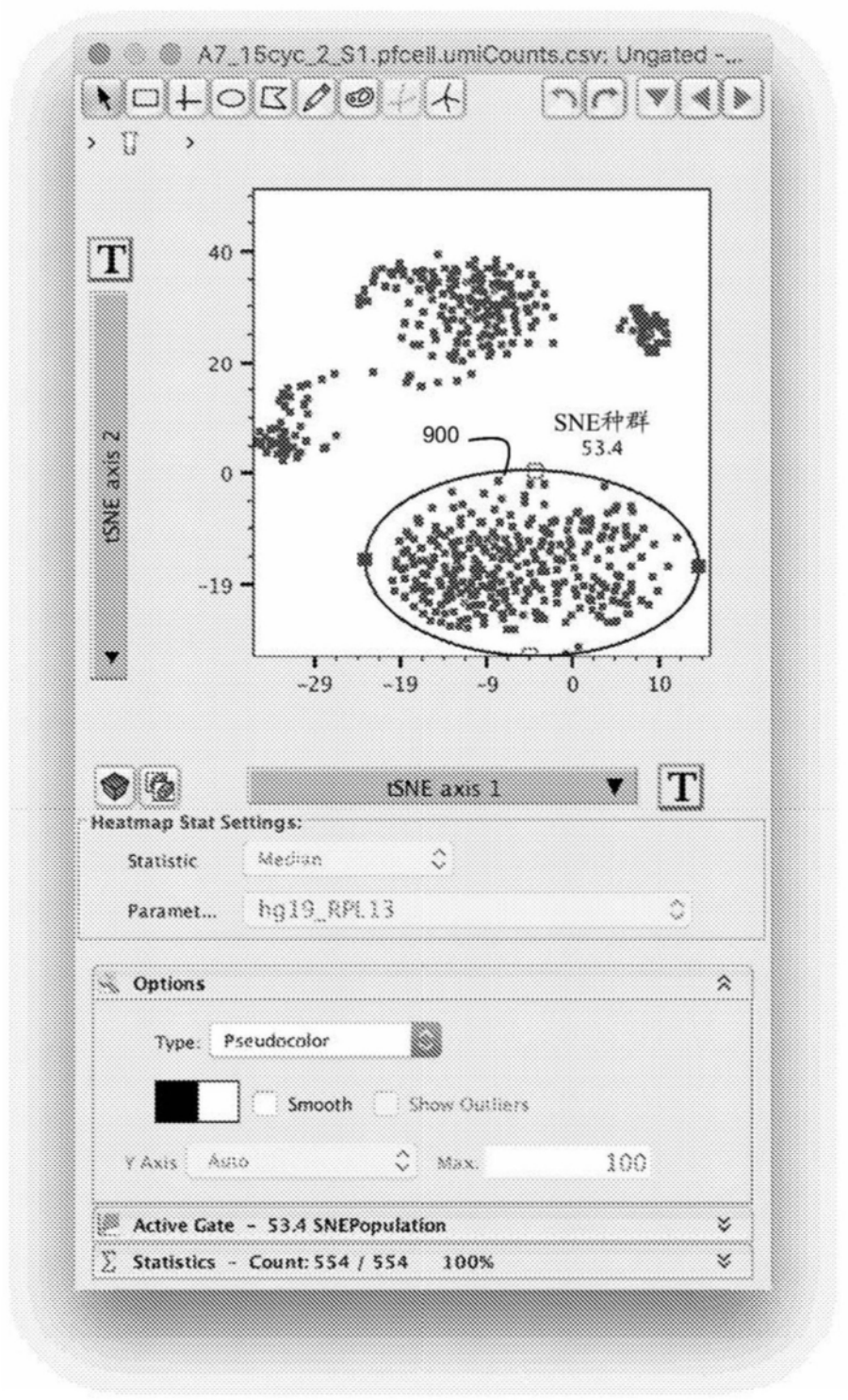

A7_15cyc_2_S1.pfcell.umiCounts.csv: Ungated -...
40
20
0
-19
tSNE axis 2
900
SNE种群
53.4
-29
-19
-9
0
10
tSNE axis 1
Heatmap Stat Settings:
Statistic
Median
Paramet...
hg19_RPL13
Options
Type:
Pseudocolor
Smooth
Show Outliers
Y Axis
Auto
Max:
100
Active Gate - 53.4 SNEPopulation
Statistics - Count: 554 / 554 100%

图9

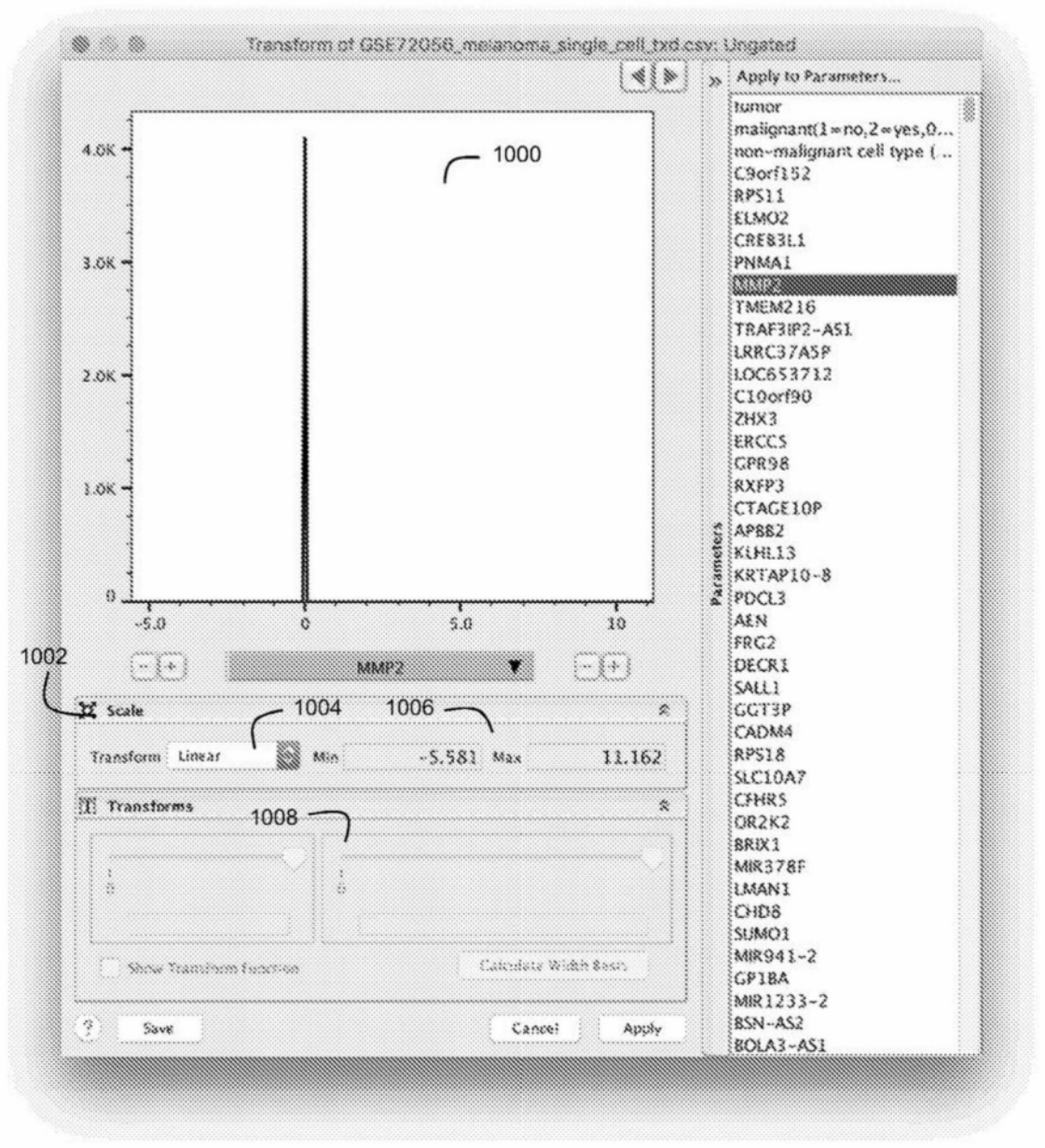

Transform of GSE72056_melanoma_single_cell_txd.csv: Ungated
Apply to Parameters...
tumor
malignant(1=no,2=yes,0...
non-malignant cell type (...
C9orf152
RPS11
ELMO2
CRE83L1
PNMA1
MMP2
TMEM216
TRAF3IP2-AS1
LRRC37A5P
LOC653712
C10orf90
ZHX3
ERCC5
GPR98
RXFP3
CTAGE10P
APBB2
KIHL13
KRTAP10-8
PDCL3
AEN
FRG2
DECR1
SALL1
GGT3P
CADM4
RPS18
SLC10A7
CFHR5
OR2K2
BRIX1
MIR378F
LMAN1
CHD8
SUMO1
MIR941-2
GP1BA
MIR1233-2
BSN-AS2
BOLA3-AS1
Parameters
4.0K
3.0K
2.0K
1.0K
0
-5.0
0
5.0
10
1000
1002
MMP2
Scale
1004
1006
Transform
Linear
Min
-5.581
Max
11.162
Transforms
1008
Show Transform function
Calculate Width Basis
Save
Cancel
Apply

图10

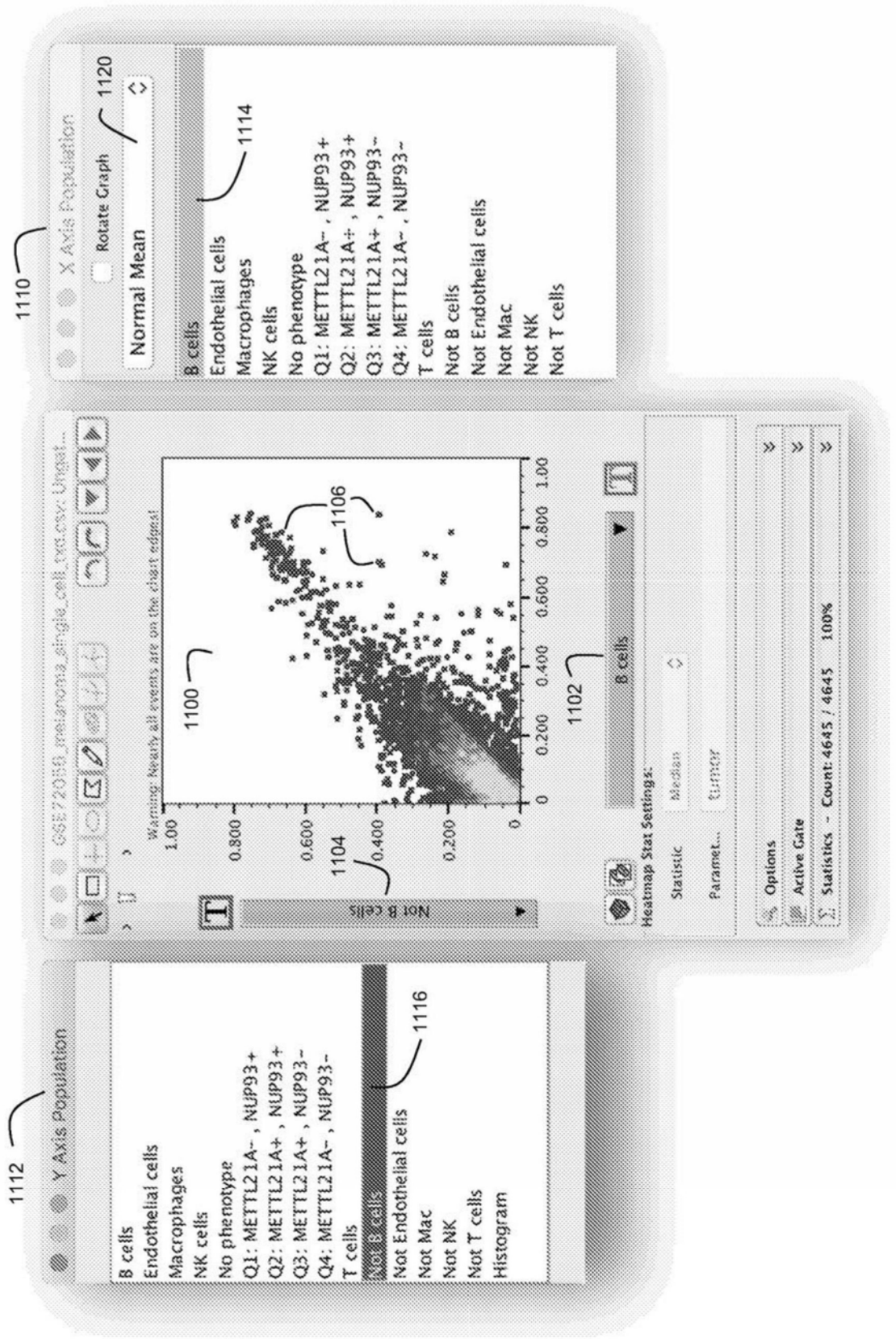
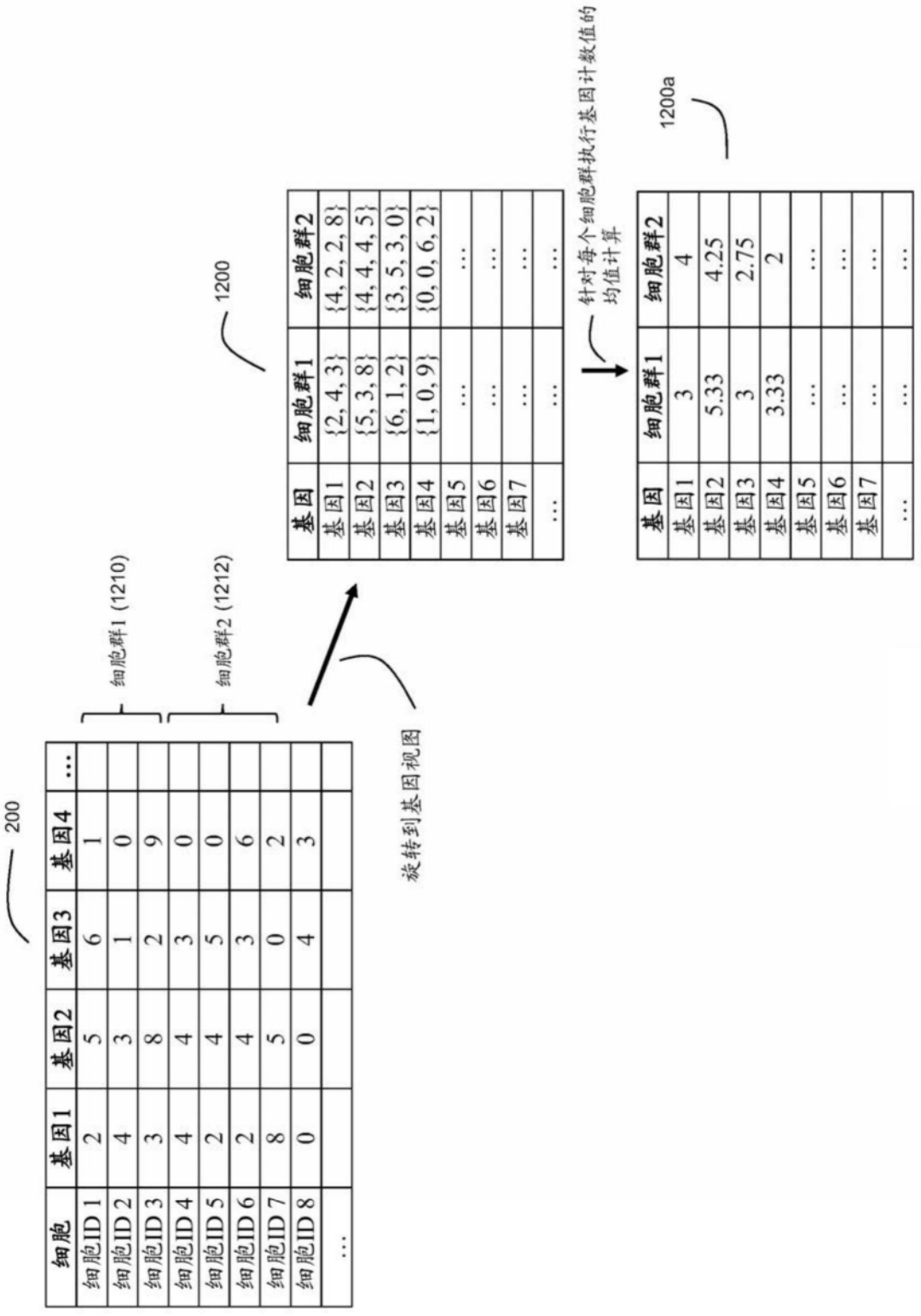

1110
X Axis Population
Rotate Graph
1120
Normal Mean
B cells
Endothelial cells
Macrophages
NK cells
No phenotype
Q1: METTL21A- , NUP93 +
Q2: METTL21A+ , NUP93 +
Q3: METTL21A+ , NUP93 -
Q4: METTL21A- , NUP93 -
T cells
Not B cells
Not Endothelial cells
Not Mac
Not NK
Not T cells
1114
1100
1106
1104
1102
Not B cells
B cells
Heatmap Stat Settings:
Statistic
Median
Paramet...
Tumor
Options
Active Gate
Statistics - Count: 4645 / 4645 100%
1112
Y Axis Population
Histogram
1116

图11



细胞	基因1	基因2	基因3	基因4	…
细胞ID 1	2	5	6	1	
细胞ID 2	4	3	1	0	
细胞ID 3	3	8	2	9	
细胞ID 4	4	4	3	0	
细胞ID 5	2	4	5	0	
细胞ID 6	2	4	3	6	
细胞ID 7	8	5	0	2	
细胞ID 8	0	0	4	3	
…					

基因	细胞群1	细胞群2
基因1	{2, 4, 3}	{4, 2, 2, 8}
基因2	{5, 3, 8}	{4, 4, 4, 5}
基因3	{6, 1, 2}	{3, 5, 3, 0}
基因4	{1, 0, 9}	{0, 0, 6, 2}
基因5	…	…
基因6	…	…
基因7	…	…
…	…	…

基因	细胞群1	细胞群2
基因1	3	4
基因2	5.33	4.25
基因3	3	2.75
基因4	3.33	2
基因5	…	…
基因6	…	…
基因7	…	…
…	…	…

图12

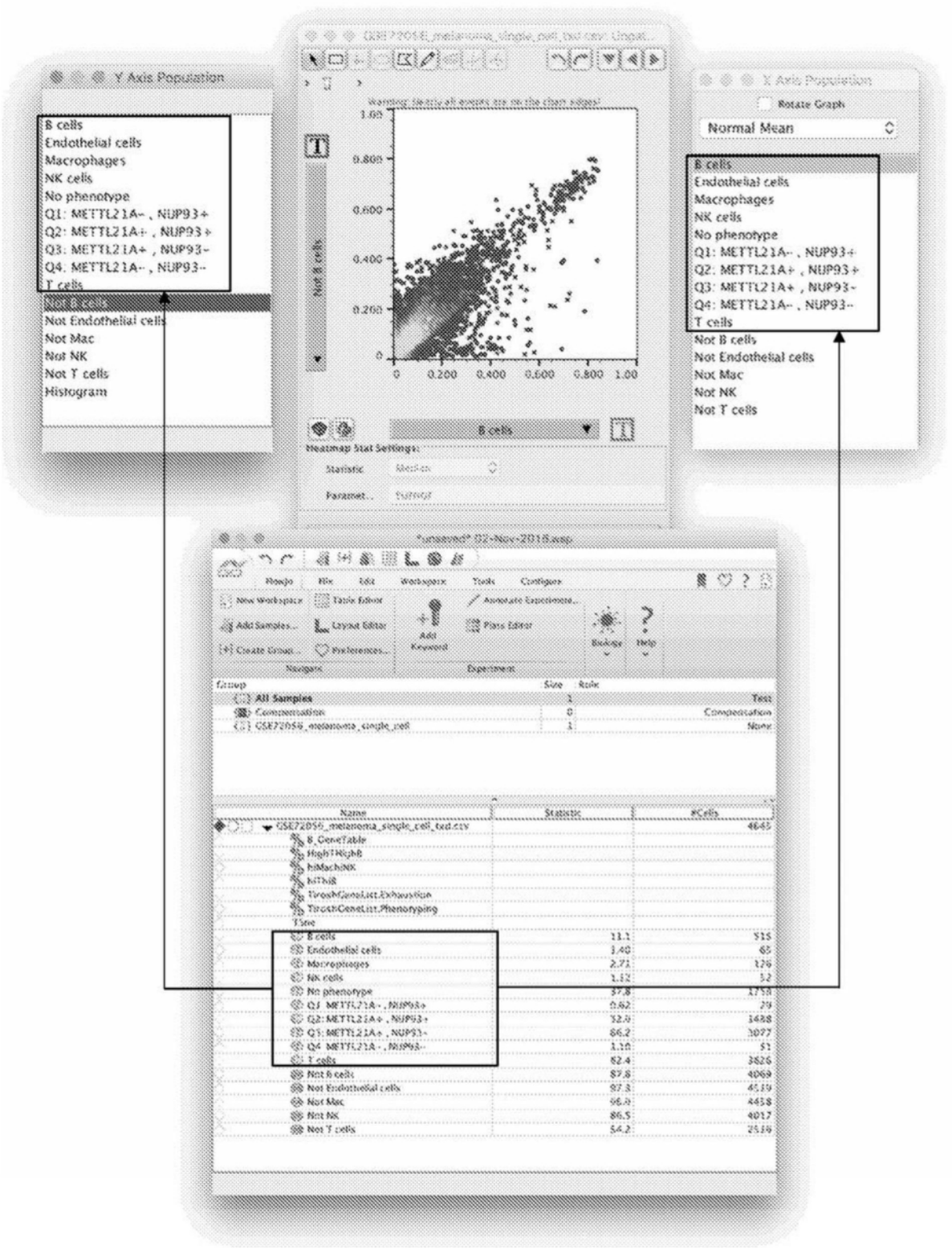

Y Axis Population
B cells
Endothelial cells
Macrophages
NK cells
No phenotype
Q1: METTL21A- , NUP93+
Q2: METTL21A+ , NUP93+
Q3: METTL21A+ , NUP93-
Q4: METTL21A- , NUP93-
T cells
Not B cells
Not Endothelial cells
Not Mac
Not NK
Not T cells
Histogram
X Axis Population
Rotate Graph
Normal Mean
Heatmap Stat Settings:
Statistic
Parameter

图13A

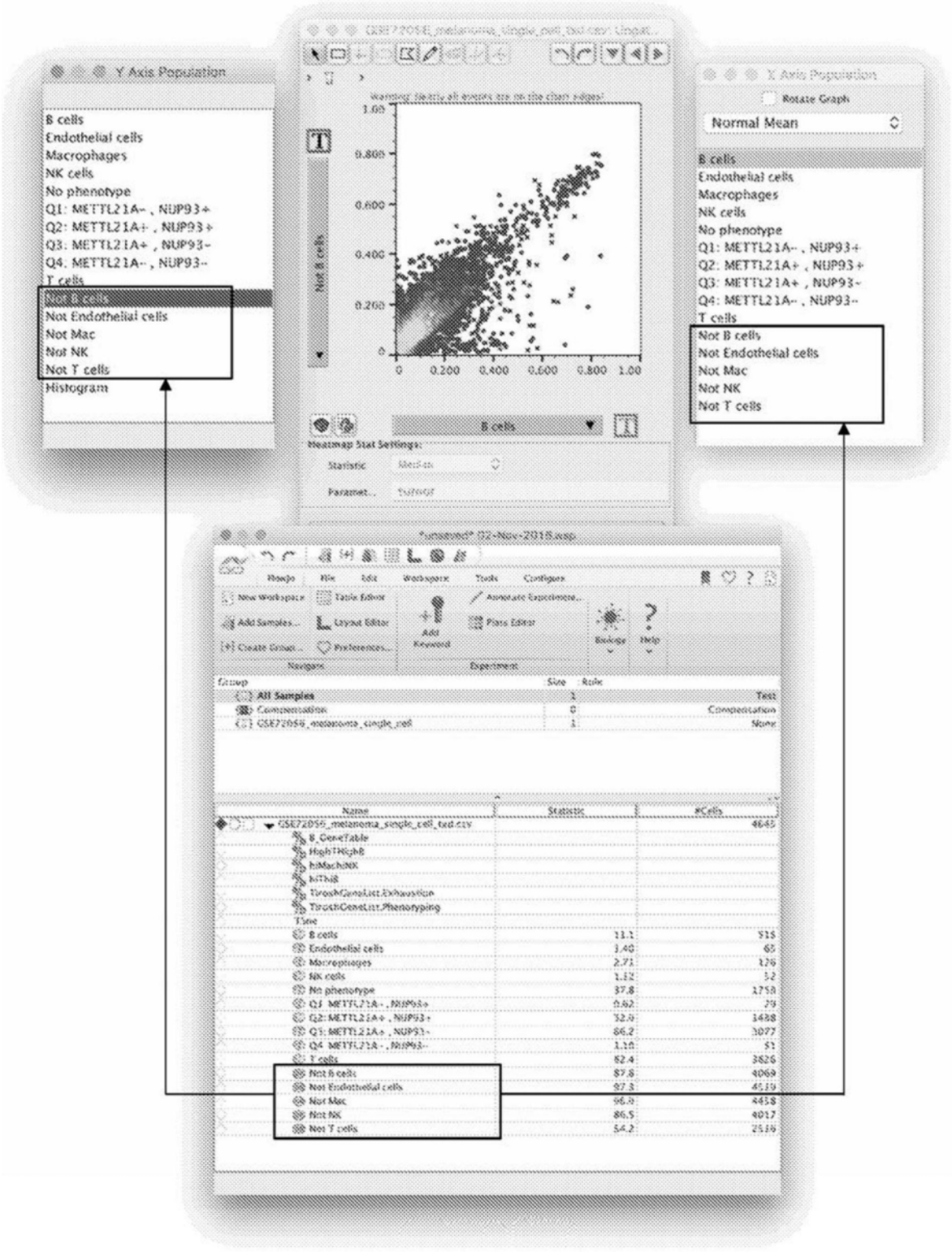

Y Axis Population
B cells
Endothelial cells
Macrophages
NK cells
No phenotype
Q1: METTL21A- , NUP93+
Q2: METTL21A+ , NUP93+
Q3: METTL21A+ , NUP93-
Q4: METTL21A- , NUP93-
T cells
Not B cells
Not Endothelial cells
Not Mac
Not NK
Not T cells
Histogram
X Axis Population
Rotate Graph
Normal Mean
B cells
Endothelial cells
Macrophages
NK cells
No phenotype
Q1: METTL21A- , NUP93+
Q2: METTL21A+ , NUP93+
Q3: METTL21A+ , NUP93-
Q4: METTL21A- , NUP93-
T cells
Not B cells
Not Endothelial cells
Not Mac
Not NK
Not T cells
Not B cells
0.800
0.600
0.400
0.200
0
0.200
0.400
0.600
0.800
1.00
B cells
Heatmap Stat Settings:
Statistic
Parameter...
Add Samples...
Layout Editor
Add Keyword
Plots Editor
Navigate
Experiment
All Samples
Name
Statistic
#Cells
B cells
Endothelial cells
Macrophages
NK cells
No phenotype
Q1: METTL21A- , NUP93+
Q2: METTL21A+ , NUP93+
Q3: METTL21A+ , NUP93-
Q4: METTL21A- , NUP93-
T cells
Not B cells
Not Endothelial cells
Not Mac
Not NK
Not T cells

图13B

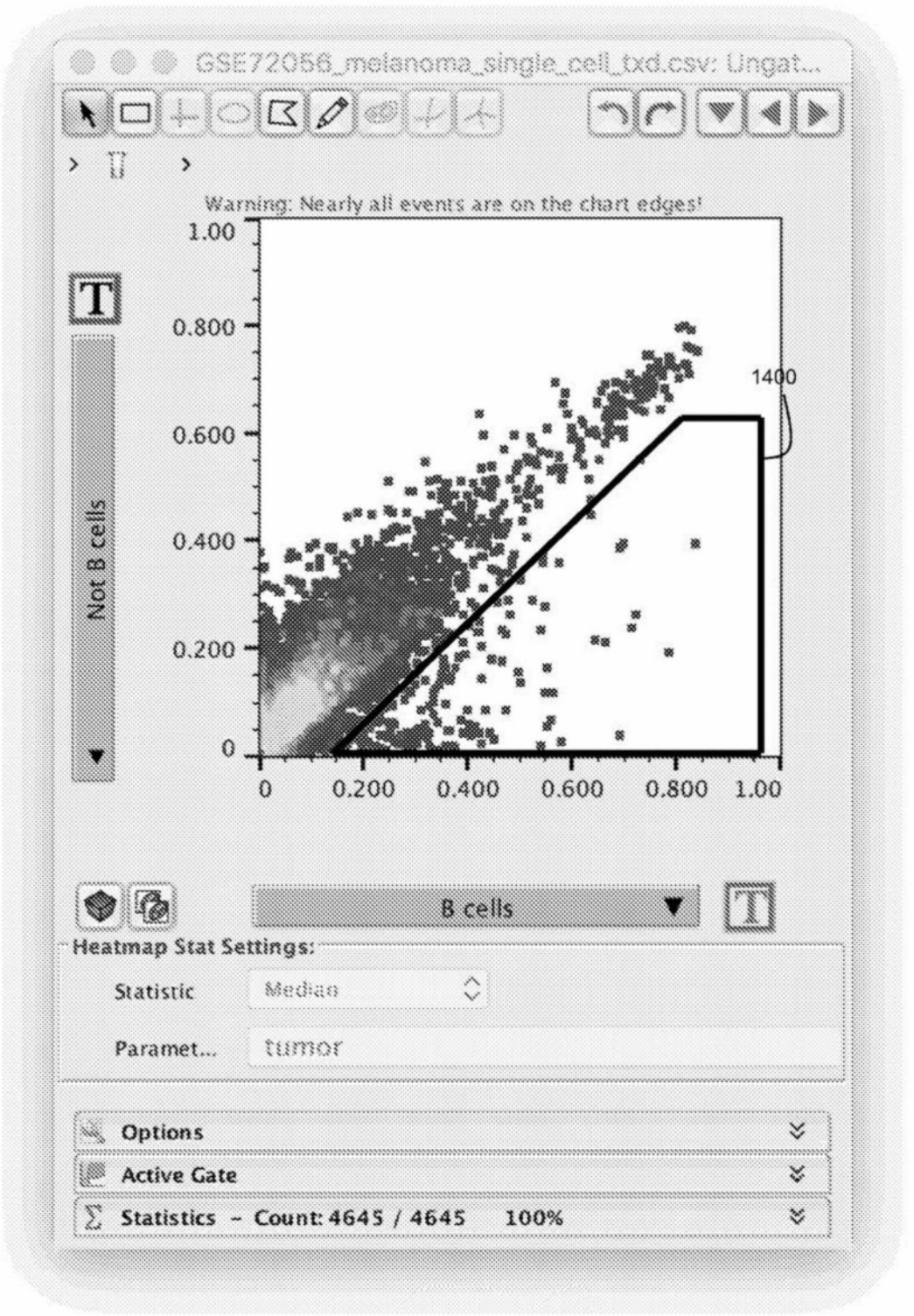

GSE72056_melanoma_single_cell_txd.csv: Ungat...
Warning: Nearly all events are on the chart edges!
1.00
0.800
0.600
0.400
0.200
0
Not B cells
1400
0
0.200
0.400
0.600
0.800
1.00
B cells
Heatmap Stat Settings:
Statistic
Median
Paramet...
tumor
Options
Active Gate
Statistics - Count: 4645 / 4645 100%

图14

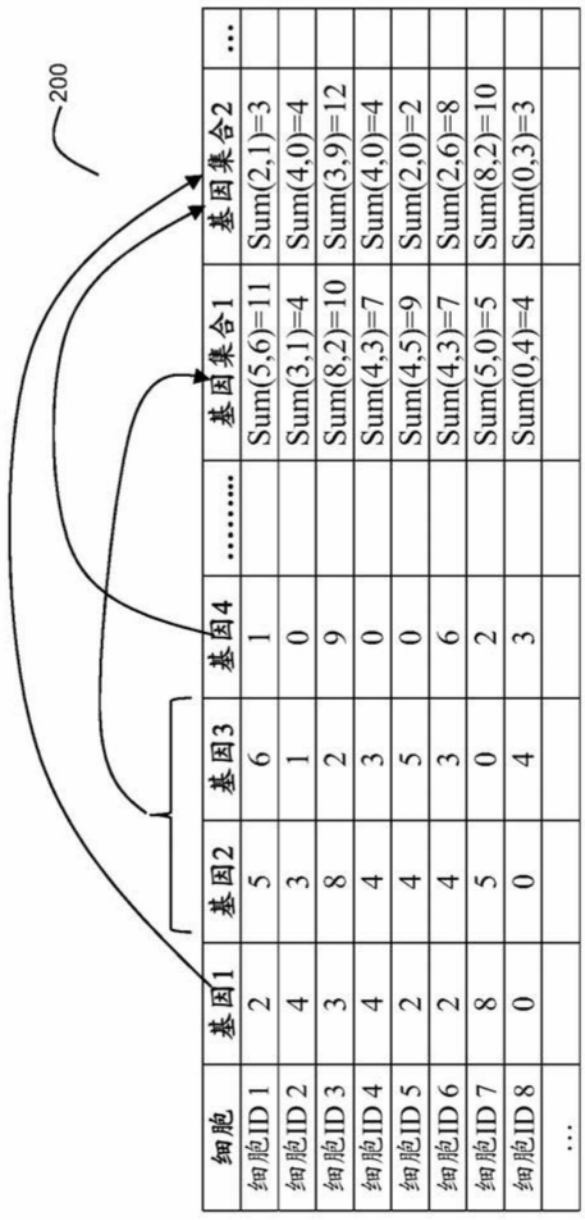

200
细胞
基因1
基因2
基因3
基因4
.....
基因集合1
基因集合2
...
细胞ID 1
2
5
6
1
Sum(5,6)=11
Sum(2,1)=3
细胞ID 2
4
3
1
0
Sum(3,1)=4
Sum(4,0)=4
细胞ID 3
3
8
2
9
Sum(8,2)=10
Sum(3,9)=12
细胞ID 4
4
4
3
0
Sum(4,3)=7
Sum(4,0)=4
细胞ID 5
2
4
5
0
Sum(4,5)=9
Sum(2,0)=2
细胞ID 6
2
4
3
6
Sum(4,3)=7
Sum(2,6)=8
细胞ID 7
8
5
0
2
Sum(5,0)=5
Sum(8,2)=10
细胞ID 8
0
0
4
3
Sum(0,4)=4
Sum(0,3)=3
...

图15

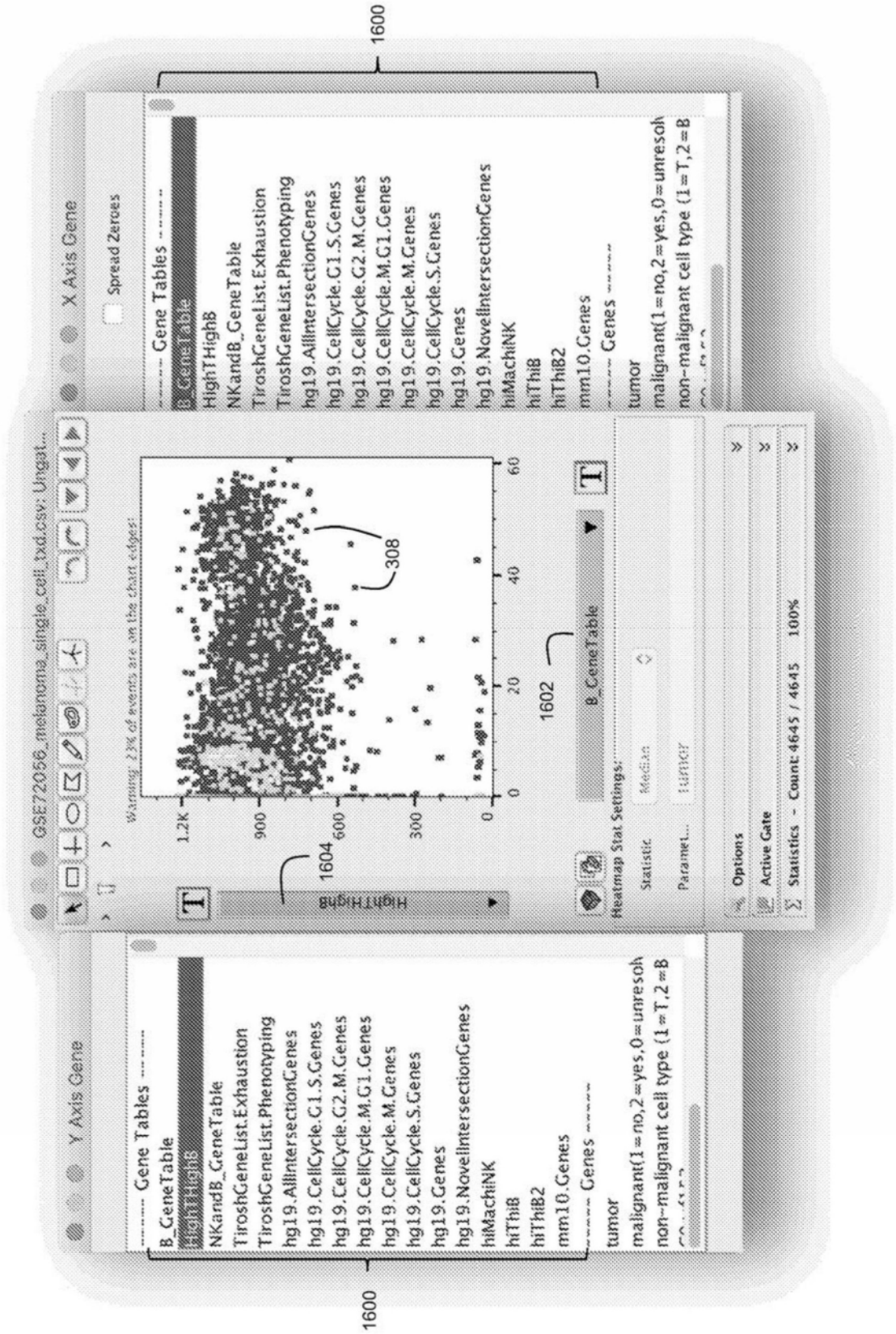

1600
X Axis Gene
Spread Zeroes
Gene Tables
B_GeneTable
HighTHighB
NKandB_GeneTable
TiroshGeneList.Exhaustion
TiroshGeneList.Phenotyping
hg19.AllIntersectionGenes
hg19.CellCycle.G1.S.Genes
hg19.CellCycle.G2.M.Genes
hg19.CellCycle.M.G1.Genes
hg19.CellCycle.M.Genes
hg19.CellCycle.S.Genes
hg19.Genes
hg19.NovelIntersectionGenes
hiMachiNK
hiThiB
hiThiB2
mm10.Genes
tumor
malignant(1=no,2=yes,0=unresoh
non-malignant cell type (1=T,2=B
GSE72056_melanoma_single_cell_txd.csv: Ungat...
Warning: 23% of events are on the chart edges!
1.2K
900
600
300
0
0
20
40
60
308
1604
1602
Heatmap Stat Settings:
Statistic
Median
Paramet...
Options
Active Gate
Statistics - Count: 4645 / 4645 100%
Y Axis Gene
1600

图16

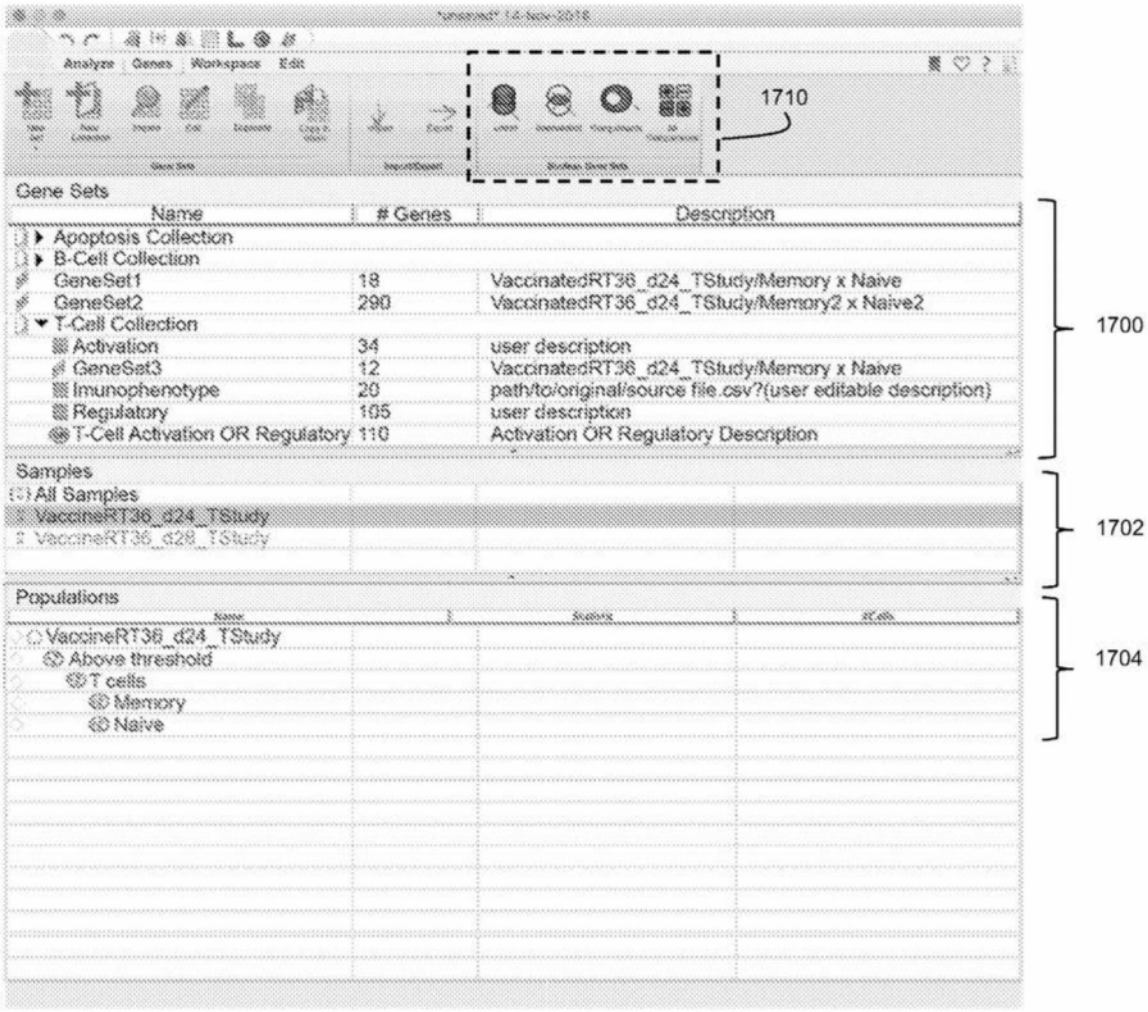

Analyze
Genes
Workspace
Edit
1710
Gene Sets
Name
Genes
Description
Apoptosis Collection
B-Cell Collection
GeneSet1
18
VaccinatedRT36_d24_TStudy/Memory x Naive
GeneSet2
290
VaccinatedRT36_d24_TStudy/Memory2 x Naive2
T-Cell Collection
Activation
34
user description
GeneSet3
12
VaccinatedRT36_d24_TStudy/Memory x Naive
Imunophenotype
20
path/to/original/source file.csv?(user editable description)
Regulatory
105
user description
T-Cell Activation OR Regulatory
110
Activation OR Regulatory Description
1700
Samples
All Samples
VaccineRT36_d24_TStudy
VaccineRT36_d28_TStudy
1702
Populations
VaccineRT36_d24_TStudy
Above threshold
T cells
Memory
Naive
1704

图17A

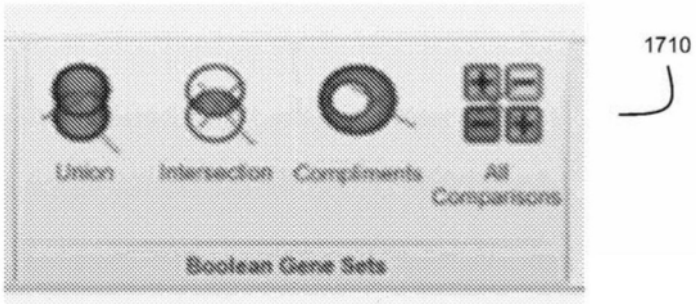

Union
Intersection
Compliments
All Comparisons
Boolean Gene Sets
1710

图17B

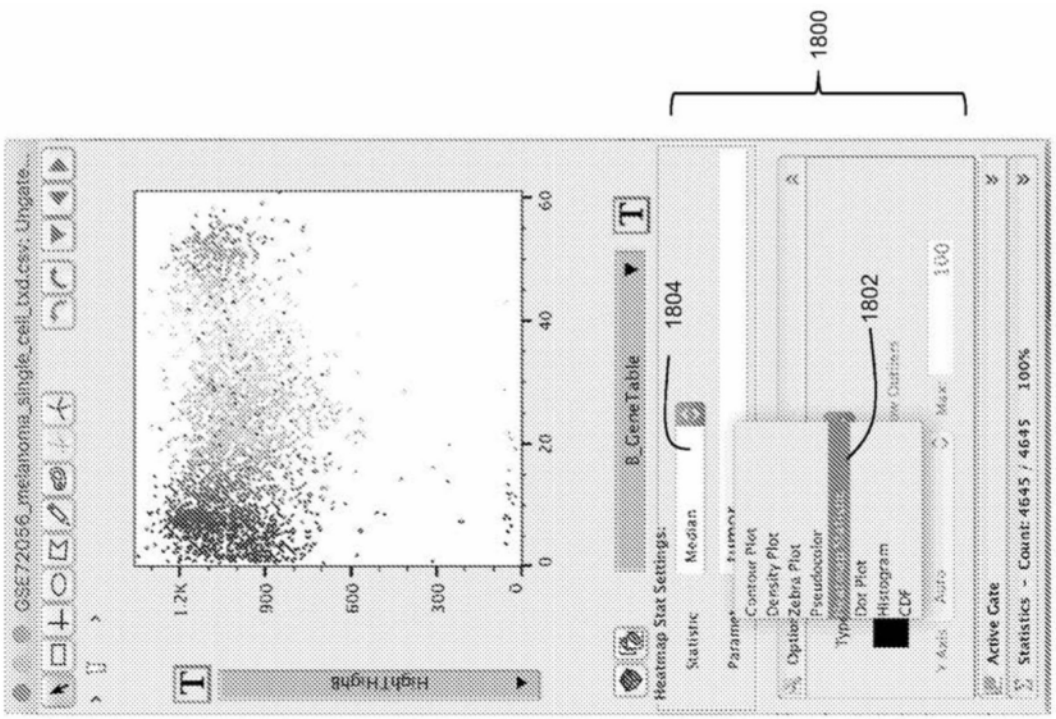

1800
1804
1802
B GeneTable
Heatmap Stat Settings:
Statistic
Median
Contour Plot
Density Plot
Zebra Plot
Pseudocolor
Dot Plot
Histogram
CDF
Active Gate
Statistics - Count: 4645 / 4645 100%
GSE72056_melanoma_single_cell_txd.csv: Ungate...
HighThighb

图18A

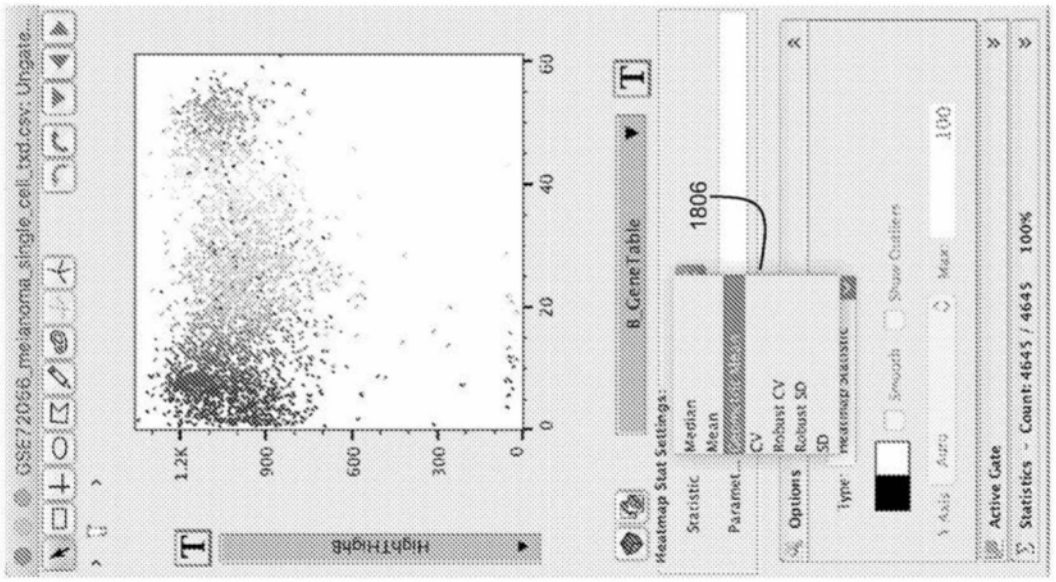

1806
B GeneTable
Heatmap Stat Settings:
Statistic
Median
Mean
CV
Robust CV
Robust SD
SD
Options
Active Gate
Statistics - Count: 4645 / 4645 100%
GSE72056_melanoma_single_cell_txd.csv: Ungate...
HighThighb

图18B

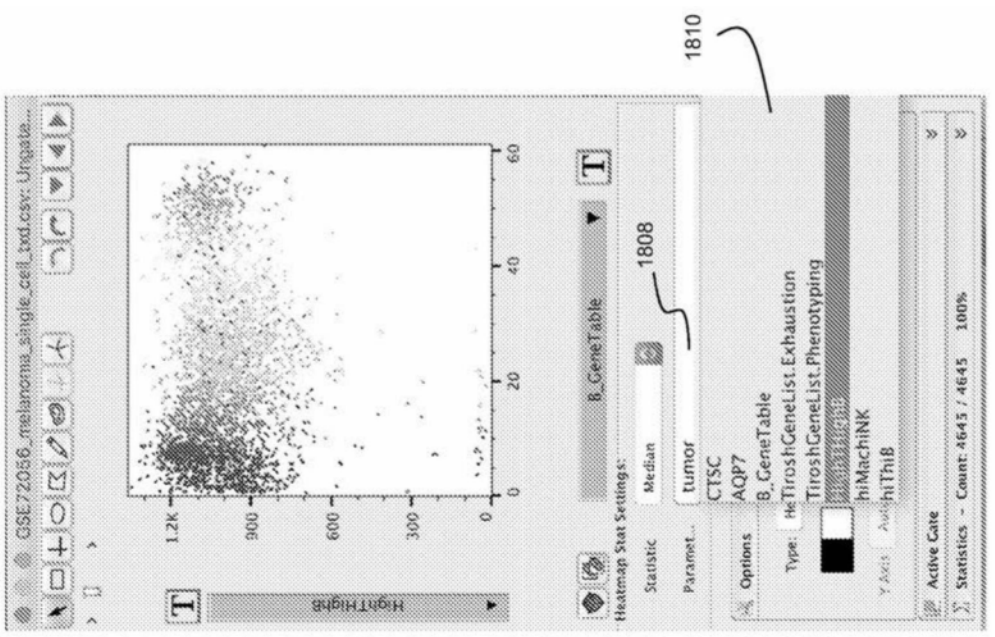

1810
1808
GSE72056_melanoma_single_cell_txd.csv: Ungate...
HighTHighB
B_GeneTable
Heatmap Stat Settings:
Statistic
Median
Paramet...
tumor
CTSC
AQP7
B_GeneTable
Options
Type:
HeTiroshGeneList.Exhaustion
TiroshGeneList.Phenotyping
hiMachiNK
hiThiB
Active Gate
Statistics - Count: 4645 / 4645 100%

图18C

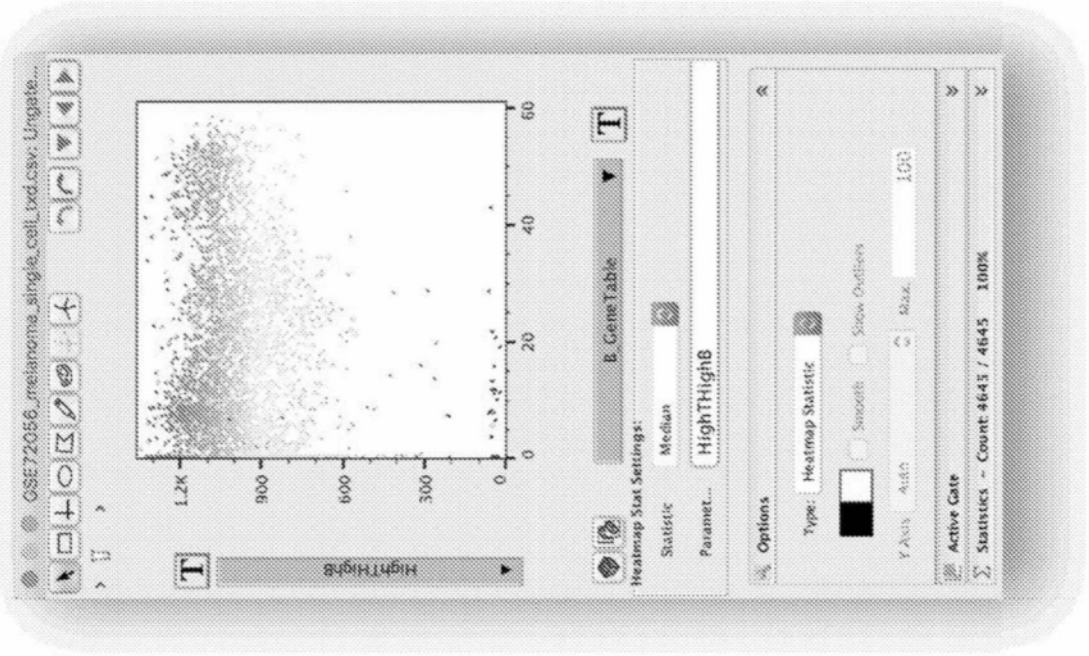

GSE72056_melanoma_single_cell_txd.csv: Ungate...
HighTHighB
B_GeneTable
Heatmap Stat Settings:
Statistic
Median
Paramet...
HighTHighB
Options
Type:
Heatmap Statistic
Active Gate
Statistics - Count: 4645 / 4645 100%

图18D

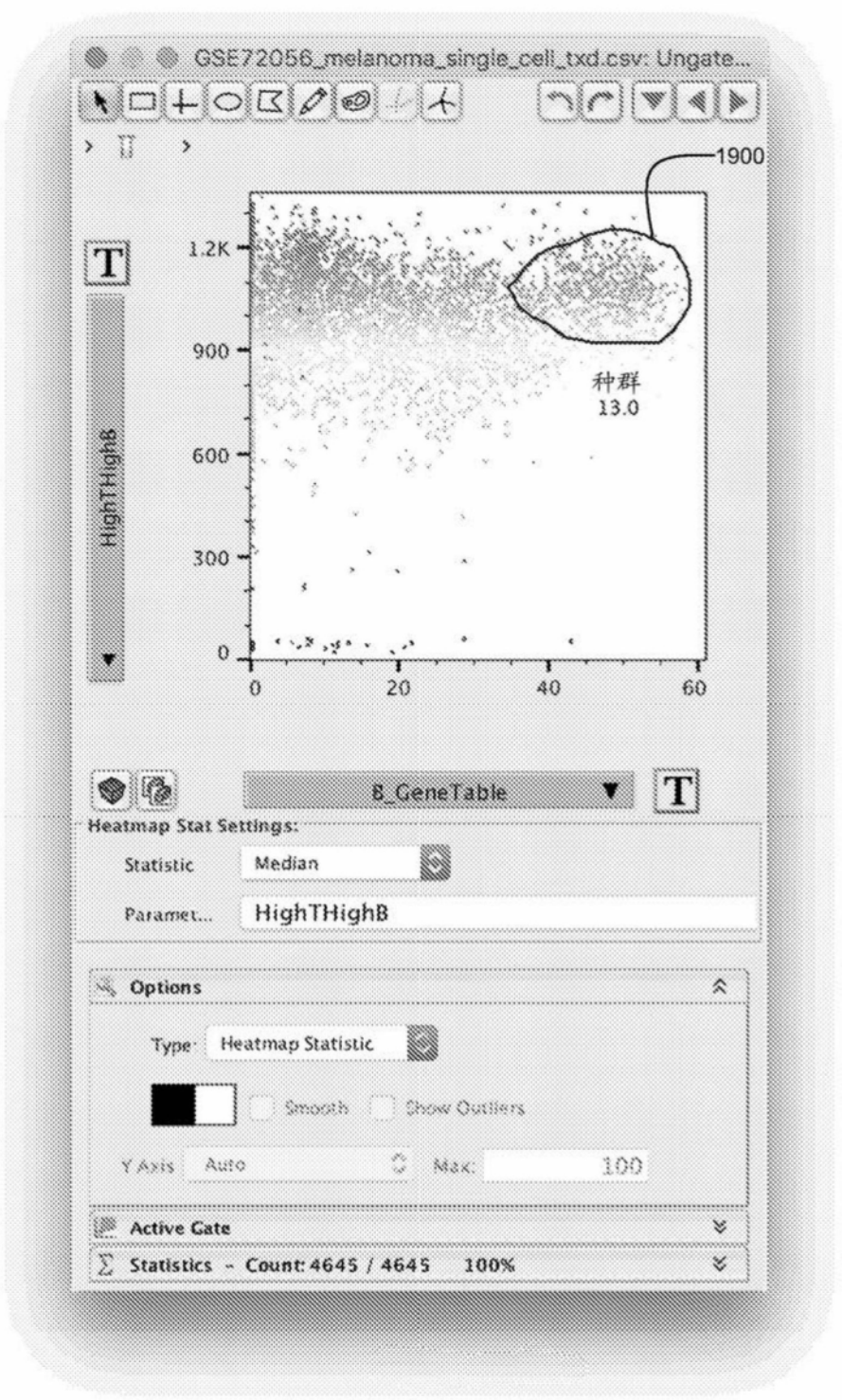

GSE72056_melanoma_single_cell_txd.csv: Ungate...
1900
1.2K
900
600
300
0
HighTHighB
种群
13.0
0
20
40
60
B_GeneTable
Heatmap Stat Settings:
Statistic
Median
Paramet...
HighTHighB
Options
Type:
Heatmap Statistic
Smooth
Show Outliers
Y Axis
Auto
Max:
100
Active Gate
Statistics - Count: 4645 / 4645 100%

图19

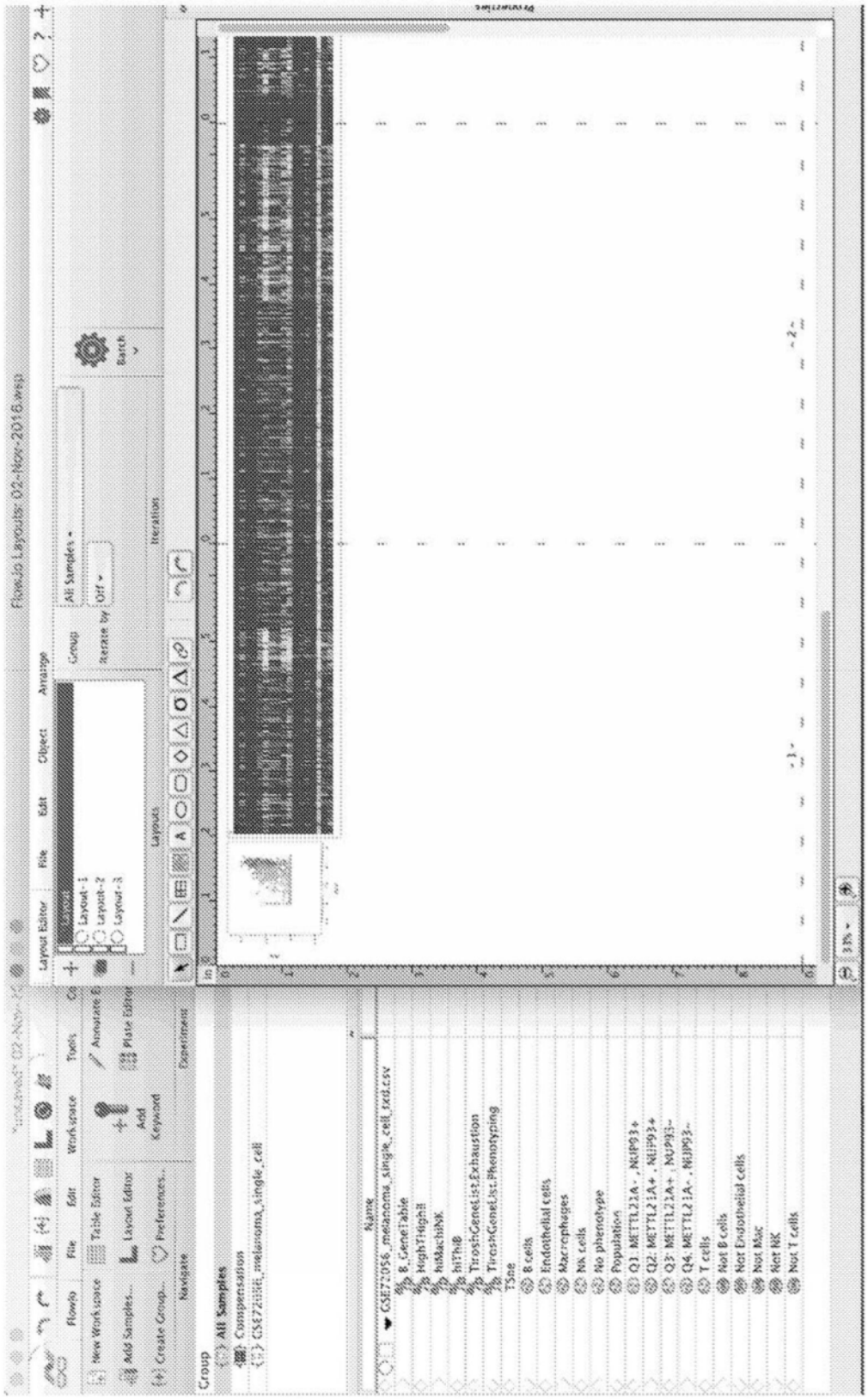


图20A

图20B

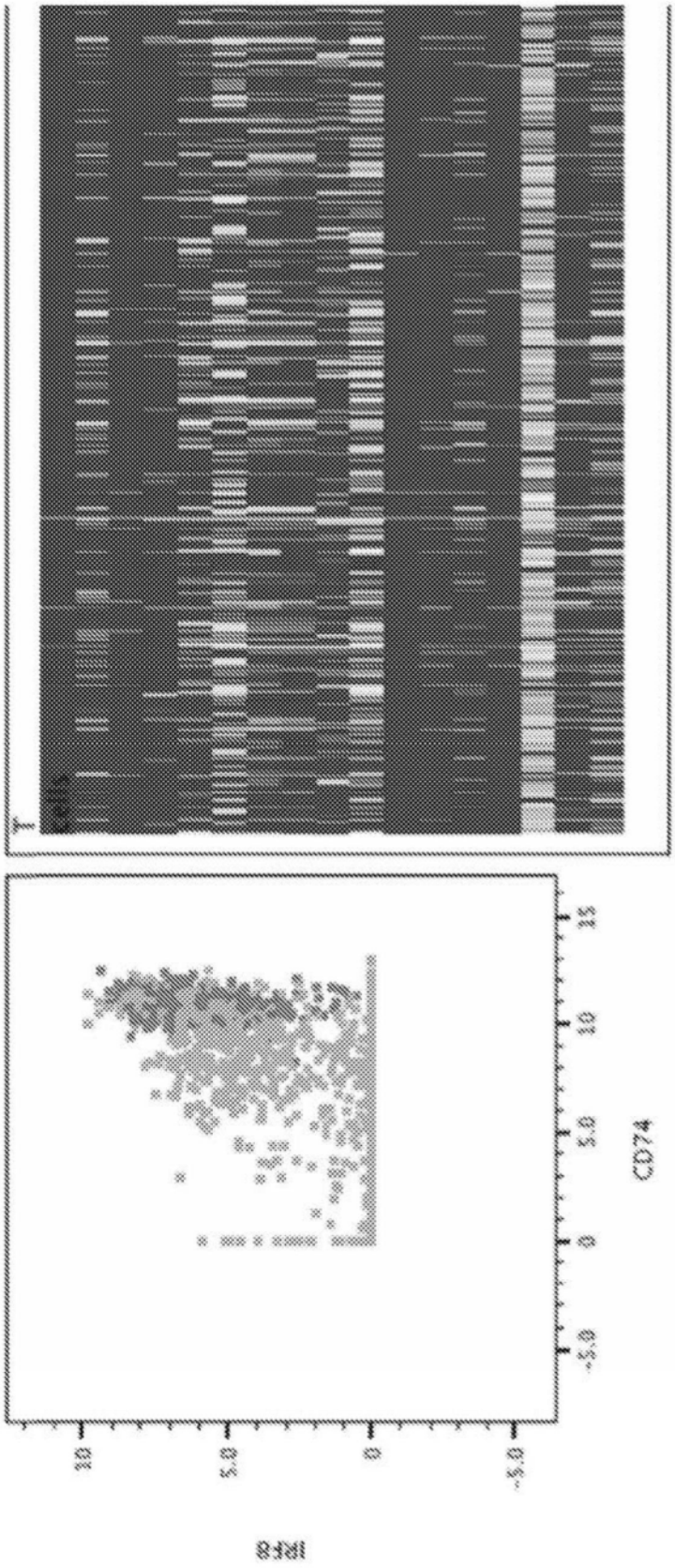


图20C

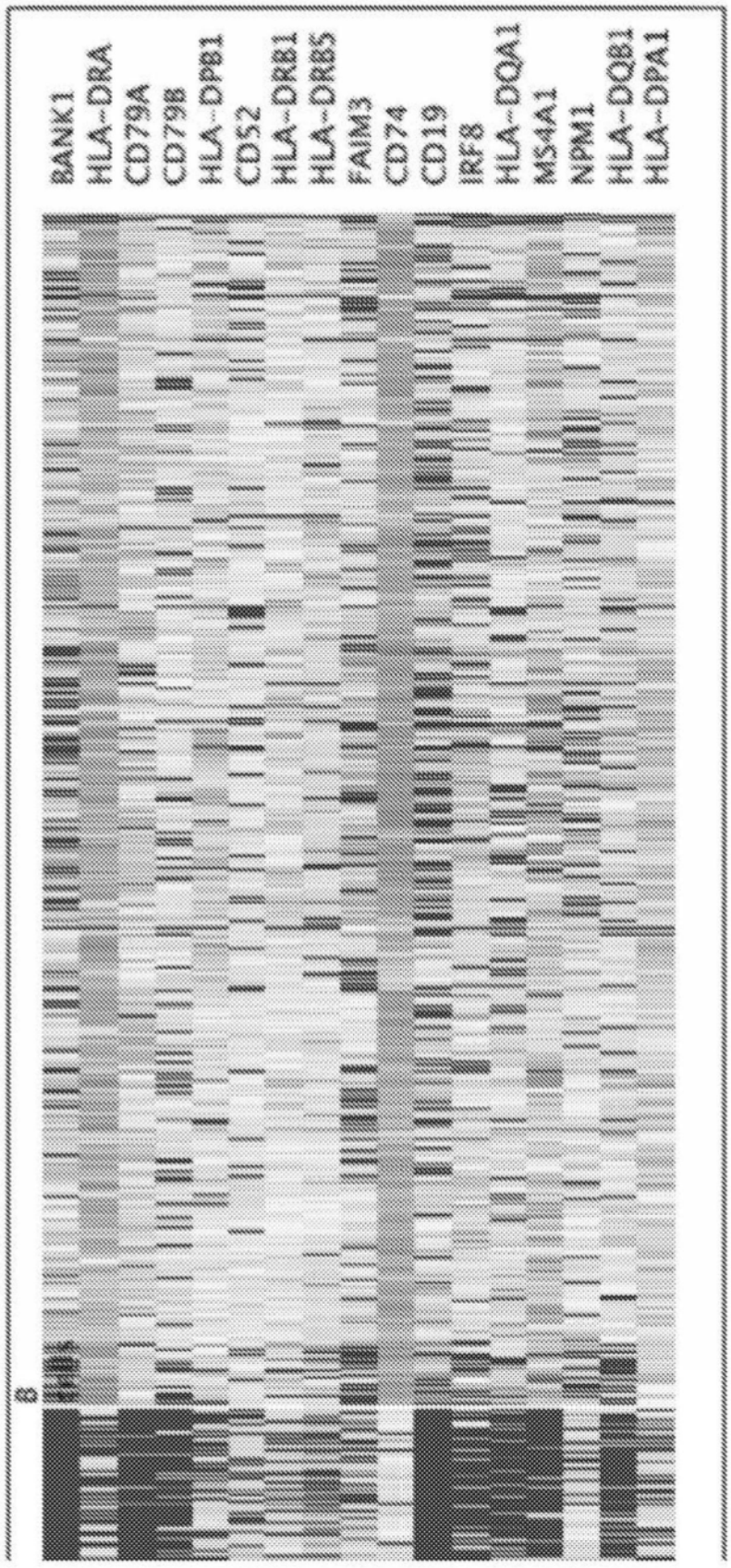

BANK1
HLA-DRA
CD79A
CD79B
HLA-DPB1
CD52
HLA-DRB1
HLA-DRB5
FAIM3
CD74
CD19
IRF8
HLA-DQA1
MS4A1
NPM1
HLA-DQB1
HLA-DPA1
PAX5
B

图20D

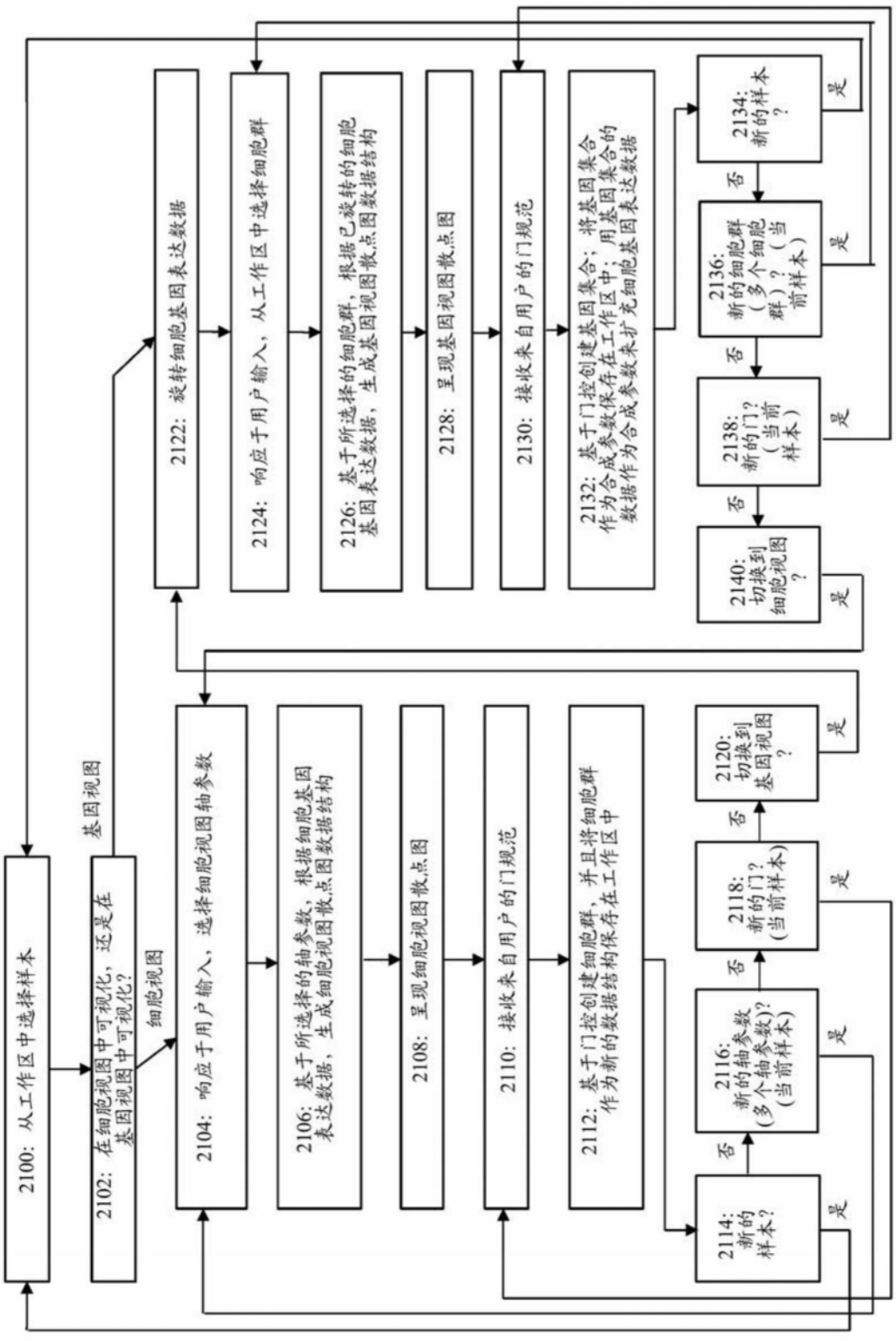

2100: 从工作区中选择样本
2102: 在细胞视图中可视化，还是在基因视图中可视化?
细胞视图
基因视图
2104: 响应于用户输入，选择细胞视图轴参数
2106: 基于所选择的轴参数，根据细胞基因表达数据，生成细胞视图散点图数据结构
2108: 呈现细胞视图散点图
2110: 接收来自用户的门规范
2112: 基于门控创建细胞群，并且将细胞群作为新的数据结构保存在工作区中
2114: 新的样本?
2116: 新的轴参数(多个轴参数)?(当前样本)
2118: 新的门?(当前样本)
2120: 切换到基因视图?
2122: 旋转细胞基因表达数据
2124: 响应于用户输入，从工作区中选择细胞群
2126: 基于所选择的细胞群，根据已旋转的细胞基因表达数据，生成基因视图散点图数据结构
2128: 呈现基因视图散点图
2130: 接收来自用户的门规范
2132: 基于门控创建基因集合；将基因集合作为合成参数保存在工作区中；用基因集合的数据作为合成参数来扩充细胞基因表达数据
2134: 新的样本?
2136: 新的细胞群(多个细胞群)?(当前样本)
2138: 新的门?(当前样本)
2140: 切换到细胞视图?
否
是

图21

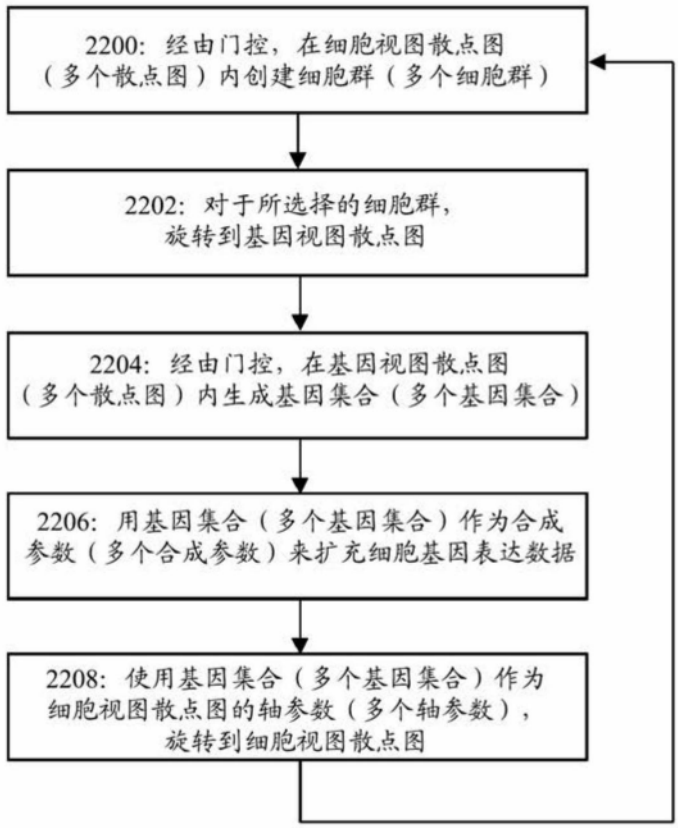

2200: 经由门控，在细胞视图散点图（多个散点图）内创建细胞群（多个细胞群）
2202: 对于所选择的细胞群，旋转到基因视图散点图
2204: 经由门控，在基因视图散点图（多个散点图）内生成基因集合（多个基因集合）
2206: 用基因集合（多个基因集合）作为合成参数（多个合成参数）来扩充细胞基因表达数据
2208: 使用基因集合（多个基因集合）作为细胞视图散点图的轴参数（多个轴参数），旋转到细胞视图散点图

图22