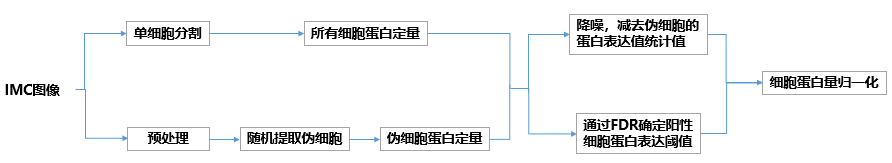
**说 明 书 摘 要**

本申请公开了一种对质谱成像细胞分析技术（IMC）细胞蛋白表达去噪及样本间蛋白量归一化的方法，该方法能够有效地降低图像噪声，提高细胞蛋白表达定量的准确性，并减少由于样本制备和保存带来的样本偏差。

**摘 要 附 图**



**权 利 要 求 书**

（已修改，见说明书）

**说 明 书**

一种基于统计学方法对多通道质谱图像进行降噪和归一化的方法

技术领域

本申请涉及图像处理领域，尤其涉及一种单细胞蛋白质谱成像图像上进行降噪和归一化的方法。

背景技术

质谱成像细胞分析技术（Imaging Mass Cytometry, IMC）是一种新型的多重分子成像方法，它是一种结合质谱分析和影像可视化的分子成像技术，使用激光消融来产生粒子后，使粒子通过惰性气流携带到质谱仪进行检测，可以在一个组织切片中同时测量多个蛋白标记物，提供组织切片的分子结构及其空间域信息。质谱成像技术作为分子成像和质谱领域的热点技术，近年来受到高度关注，并迅速发展成用于分析组织切片的重要工具。它能够检测基因、蛋白质以及药物等小分子在生物体内的分布特征及其含量的变化信息，提供生物体不同生理及病理过程中分子的变化，在临床医学、分子生物学和药学等领域有重大的应用前景。

随着这项技术的不断发展，也出现了许多亟待解决的技术问题。当前IMC图像的规模与信息量愈发庞大，急需发展相应的高效的数据解析方法。IMC图像处理流程主要包括用户交互式的单细胞分割，单细胞聚类和细胞类型的标注。然而，在IMC的成像过程中，由于采用的各种蛋白标记物的不同特性，以及实验室质谱仪、成像设备、或成像设备相应的图像处理软件所带来的各种系统误差和随机误差，给IMC图像引入了噪声和批次效应。目前的IMC图像处理流程中，往往缺乏有效的去噪算法和批次效应消除算法。这些问题直接影响到单细胞的表达定量、聚类和标注等下游分析的结果的准确性和精确度，对该技术的临床应用形成了障碍。

发明内容

在IMC图像处理流程中，本领域中通常使用的去噪方法是选取整张图像作为噪音区域进行中值滤波，但是，在真实细胞存在的区域，中值滤波器会导致去噪后去除了某些表达较弱的蛋白质信号，去噪后所得图像准确性和精确度较低。同时，对很多聚集成团的噪音点，单纯使用中值滤波器无法去除那些图像噪点，降噪结果不佳。

本发明提出了一种质谱成像细胞分析技术（IMC）图像去噪及样本间蛋白量归一化的方法，该方法能够有效地降低图像噪声，并保留图像的生物信息。具体而言，本发明通过如下技术方案，解决了本发明的技术问题。

1. 一种对质谱成像细胞分析技术（IMC）图像进行降噪及蛋白量归一化的方法，包括以下步骤对图像进行降噪：

1) 用图像单细胞分割方法识别所述IMC图像上的单个细胞；

2) 对所述IMC图像进行预处理；

3）将细胞的蛋白表达值量化为该细胞分割区域中所有像素值的统计值；

4）在图像上选取一定数量的随机伪细胞，按照步骤3）的方法对随机伪细胞进行蛋白表达值量化，得到随机伪细胞的蛋白表达值；

5）通过随机伪细胞的蛋白表达值，计算得到图像的背景噪声水平。消除图像中的背景噪声影响后，得到降噪后 IMC 图像；

2. 项目1所述的方法，进一步包括如下步骤对图像进行归一化：

1）将项目1步骤3）得到的降噪前的细胞的蛋白表达值分布与项目1步骤4）得到的随机伪细胞的蛋白表达值分布进行比较，通过控制错误发现率（FDR）来确定阳性细胞的蛋白表达值阈值；

2）利用多个样本上面的阳性细胞的蛋白表达值，计算每个样本的归一化缩放参数，对每个样本的 IMC 图像进行归一化。

3. 项目1所述的方法，其中所述预处理包括用Arcsinh函数对图像像素值进行转换，压缩像素值的尺度，使像素值更加平稳；和用滤波器去除图像上的“热像素点”，从而减小离群“热像素点”的噪声影响。

4. 项目1所述的方法，其中步骤4）随机伪细胞的选择，还包括在图像上选择随机伪细胞。选择随机伪细胞的办法如下：

1）计算项目1的步骤1）中通过细胞分割获得的所有细胞的长轴、短轴、方向和面积大小，并使用单独的高斯模型分别对长轴、短轴和方向进行拟合；

2）从上述的三个分布中抽取随机数，并在图像上随机选择一个坐标点，以该点为中心在图像上形成一个椭圆的随机伪细胞，该随机伪细胞需要满足以下两个条件：

条件1：随机伪细胞的面积大小在所有细胞的面积大小范围之内；

条件2：当随机伪细胞的部分位置超出图像边界时，该随机伪细胞在图像内的面积要大于5个像素点；

5. 项目4所述的方法，其中随机伪细胞的所有像素点都在图像的噪声区域内。噪声区域定义办法如下：

1. 将的像素置为0。其中，是原始IMC图像的像素值，是设定的最小表达量阈值；
2. 将经过上一步操作后的图像通过中值滤波器。

然后，定义如上第二步获得的图像上面像素值为0 的区域为噪声区域。

6. 项目1所述的方法，其中，图像的背景噪音水平 按照如下方式计算：

其中，为 IMC 图像上，随机伪细胞 上面的像素的平均蛋白表达值，即

其中， 为随机伪细胞 的蛋白表达量， 为随机伪细胞 的像素总数。

7. 项目6所述的方法，其中，降噪后的 IMC 图像的像素值 按照如下方式计算：

其中， 为降噪后的像素值， 为项目1步骤2）后得到的降噪前的像素值， 为图像的背景噪音水平。

8. 项目2所述的方法，其中错误发现率（FDR）的计算方法为：

其中，是假阳细胞数，即项目1的步骤4）中所提取的随机伪细胞中蛋白表达值大于阈值的细胞个数，是真阳细胞数，即项目1的步骤3）中降噪前细胞蛋白表达值大于阈值的细胞个数。

9. 项目2所述的方法，其中细胞的蛋白表达值归一化包括：

对每个蛋白标志物，在降噪后的 IMC 图像上，计算每个样本上的所有阳性细胞的蛋白表达值，然后，利用每个样本上的所有阳性细胞蛋白表达值，按照如下方法求得每个样本的归一化缩放参数 ：

其中， 为样本 上所有阳性细胞蛋白表达值的平均值。得到 后，按照如下方式对 IMC 图像进行归一化

其中， 为样本 归一化后的像素值， 为样本 降噪后的像素值， 为样本总数。

10. 项目2所述办法，最后将归一化的图像通过滤波器，如中值滤波器，优选中值滤波器，最后输出降噪并归一化后的IMC图像，并在降噪并归一化后的IMC图像上按照项目1步骤3）的方法进行定量，获得细胞的蛋白表达值；

11. 项目1-10任一项所述的方法，进一步包含去噪结果的可视化，将阳性细胞范围外的全部像素置0，然后，对产生的图像进行高斯滤波，并将滤波结果的图像可视化，将去噪后的图像呈现在研究者面前。

附图说明

图1是对IMC图像进行图像的降噪处理和样本间蛋白表达值归一化的流程图。

图2是项目1的步骤1）得到的单细胞分割结果（在本实例中我们使用Dice-XMBD方法进行图像的单细胞分割）。左图为原始图像的细胞核通道，右图为单细胞分割图像。

图3是项目1的步骤2）预处理后的效果图。图3a左图为CD3蛋白通道的原始图像，右图为经过Arcsinh转换后图像。图3b左图为经过Arcsinh转换后图像，右图为去除“热像素点”图像。

图4是项目5中，设定最小表达量阈值参数为原始IMC图像的0.05倍99分位像素值（缺省参数值）时，区分出的噪声区域，及项目4中提取随机伪细胞的示意图。左图为真实细胞表达区域和噪声区域区分图，绿色为细胞蛋白表达区域，黑色为噪声区域；右图为提取随机伪细胞示意图。

图5是项目2步骤1）中，计算FDR并确定阳性细胞的蛋白表达值阈值示意图。蓝色曲线为随机伪细胞蛋白表达量分布曲线，蓝色阴影区域即假阳细胞数；红色曲线为图像内细胞蛋白表达量分布曲线，红色阴影区域即真阳细胞数。

图6是设置随机伪细胞数量为1000，错误发现率FDR为0.01，将阳性细胞范围外的全部像素置0，采用高斯滤波后的降噪效果图。左图为降噪前图像，右图为降噪后图像。

图7是对CD3蛋白通道图像的手工降噪方法得到的效果图。

图8是对CD3蛋白通道图像应用中值滤波降噪方法得到的效果图。

具体实施方式

为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明作进一步的详细说明。

可以对IMC图像进行以下步骤的处理，进行图像的降噪及样本间蛋白量归一化处理（其流程图如图1所示）。

步骤1：细胞分割；

使用已有的单细胞分割方法对图像进行单细胞分割。在一个实例中，我们采用预训练的Dice-XMBD模型进行单细胞分割。细胞分割后得到与原图大小一致、像素级精度的细胞掩码。舍弃细胞面积小于5个像素点的分割区域，得到单细胞分割结果。

步骤2：图像预处理；

1. 使用Arcsinh函数对图像像素值进行转换，压缩像素值的尺度，使像素值更加平稳；
2. 使用图像滤波器去除图像上的“热像素点”。当中值滤波器窗口的中心像素值位于所有像素表达值前 2%，且比窗口中所有像素值中值高至少 4 倍时，该中心像素点为判定成“热像素点”，并将其像素值调整为窗口中所有像素点的中值。这一步减小了离群“热像素点”的噪声影响。

步骤3：定量单细胞蛋白表达值；

经过细胞分割（步骤1）和图像预处理（步骤2）后，将降噪前的细胞蛋白表达值量化为预处理后图像上该细胞内所有像素点值的统计值，在一个实例中，我们计算细胞内所有像素点值的平均值来定量降噪前的细胞蛋白表达值。

步骤4：提取一定数量的随机伪细胞并计算其蛋白表达值；

* 1. 在提取随机伪细胞前，先区分图像上概率较高的真实细胞表达区域和噪声区域。图像上像素值表达量高的区域定义为置信度高的真实细胞表达区域，相反地，像素值很低的区域定义为噪声区域。将噪声区域的像素值置零，即图像的像素值被调整为：

其中，是原始IMC图像的像素值，是设定的最小表达量阈值，是调整后的像素值。

接着，在调整后的图像上使用中值滤波器。

经过上述处理步骤得到的图像中，噪声区域像素值为0，置信度高的真实细胞表达区域像素值大于0，即完成了置信度高的真实细胞表达区域和噪声区域的区分。

* 1. 提取满足规则的随机伪细胞。在一个实例中，我们提取1000个随机伪细胞。计算步骤1细胞分割获得的所有细胞的长轴、短轴、方向和面积大小，并使用单独的高斯模型分别对长轴、短轴和方向进行拟合。从上述的三个分布中抽取随机数，并在图像上随机选择一个坐标点，以该点为中心在图像上形成一个椭圆的随机伪细胞。该随机伪细胞需要满足以下三个条件：

1. 条件1：随机伪细胞的面积大小在所有细胞的面积大小范围之内；
2. 条件2：当随机伪细胞的部分位置超出图像边界时，保证该随机伪细胞在图像内的面积大于5个像素点；
3. 条件3：随机伪细胞的所有像素点都在上述噪声区域内。
   1. 使用步骤3的方法量化随机伪细胞的蛋白表达值。
   2. 将提取的随机伪细胞的平均蛋白表达值定义为图像的背景噪声水平，其计算方式如下：

其中， 为 IMC 图像上，随机伪细胞 上面的像素的平均蛋白表达值，即

其中， 为随机伪细胞 的蛋白表达量， 为随机伪细胞 的像素总数。

步骤5：消除背景噪声影响；

为了消除图像中的背景噪声影响，将降噪前步骤2得到的图像的像素值减去步骤4（步骤4.4）中的图像背景噪声水平，即降噪后图像的像素值为：

其中，为降噪后的像素值，为降噪前的像素值，为图像的背景噪音水平。

步骤6：确定阳性细胞的蛋白表达值阈值；

在本发明中，错误发现率（FDR）的计算方法为：

其中， (False Positive)是假阳细胞数，即步骤4中所提取的1000个随机伪细胞中蛋白表达值大于阈值的细胞个数， (True Positive)是真阳细胞数，即步骤3中降噪前细胞蛋白表达值大于阈值的细胞个数。

在一个实例中，我们将错误发现率FDR设为0.01，即控制假阳细胞数在阳性细胞中的比例为0.01，从而确定阳性细胞的蛋白表达值阈值，并将步骤3中降噪前细胞蛋白表达值大于等于阈值的细胞定义为图像上的阳性细胞，小于阈值的细胞确定为阴性细胞。

步骤7：细胞蛋白表达值归一化；

不同样本间的同一蛋白标志物存在表达强度不同的情况，所以需要对同一蛋白标志物中所有样本的细胞蛋白表达值进行归一化处理，以便进行合理的后续生物数据分析，例如细胞无监督聚类。

（1）在每个蛋白标志物中，计算每个样本中步骤6所确定的阳性细胞的降噪后蛋白表达值，然后，利用每个样本上的阳性细胞蛋白表达值，按照如下方法求得每个样本的归一化缩放参数 ：

其中， 为样本 上阳性细胞蛋白表达值的平均值，为样本总数。（2）

得到归一化缩放参数 后，按照如下方式对 IMC 图像进行归一化

其中， 为样本 归一化后的像素值， 为样本 降噪后的像素值。

（3）为了进一步去除椒盐噪点，将经过背景噪音消除和归一化后的图像经过滤波器，优选中值滤波器，输出最终降噪并归一化后的IMC图像，并使用步骤3的方法对单细胞进行蛋白定量，输出最终降噪并归一化后的细胞蛋白表达值。

步骤 8：交互式调参和去噪结果可视化（可选）。

在步骤6的基础上，将阳性细胞范围外的全部像素置0，然后，对产生的图像进行高斯滤波，并可视化，将去噪后的图像呈现在研究者面前，使他们能够观察、判断，和调整参数。用户可调整的参数包括：最小表达量阈值（缺省值为预处理后图像0.05倍99分位像素值），伪细胞数量（缺省值为1000），以及错误发现率FDR（缺省值为0.01）。

实施例

将发明本方法应用于黑色素瘤患者的IMC图像中。该数据集的采集方法如下：

首先筛选靶标蛋白，得到抗体panel（其中包含25个预标记抗体、10个第三方抗体和DNA），设计不同的抗体浓度梯度预实验，获得最佳的抗体浓度配比，然后对样本组织石蜡切片进行染色。抗体孵育完成后使用Hyperion组织质谱成像系统（Fluidigm）对组织切片进行全景图扫描，并挑选合适的感兴趣区域（ROI）进行成像。

我们对原始的IMC图像使用本方法进行降噪和归一化，各步骤的实验参数及实验结果如下：

步骤1中使用Dice-XMBD（代码地址为<https://github.com/xmuyulab/Dice-XMBD）得到的单细胞分割结果（见图2>）；

步骤2进行图像预处理，包括像素值做Arcsinh函数转换（代码为Python标准库math模块的asinh函数），运用中值滤波器将位于前 2%且比4 倍窗口内像素值中值高的离群热像素点调整为窗口内像素值的中值，所得结果见图3；

步骤3：经过细胞分割（步骤1）和图像预处理（步骤2）后，将降噪前的细胞蛋白表达值量化为预处理后图像上该细胞内所有像素点值的平均值；

步骤4：提取一定数量的随机伪细胞并计算其蛋白表达值，首先设定最小表达量阈值参数为原始IMC图像的0.05倍99分位像素值，将小于最小表达量阈值的像素值置为0后所得到的图像再通过中值滤波器，从而区分出置信度高的真实细胞表达区域和噪声区域，接着，根据步骤1所得的单细胞分割结果得到IMC图像上所有细胞的面积、长轴、短轴及方向，并使用单独的高斯模型分别对长轴、短轴和方向进行拟合，最后，在上述三个分布中提取长轴、短轴和方向随机数并在噪声区域内形成面积在所有细胞面积范围内的伪随机细胞（示意图见图4），设定随机伪细胞数量为1000，完成1000个伪细胞提取后使用步骤3的方法量化随机伪细胞的蛋白表达值，并将提取的随机伪细胞的平均蛋白表达值定义为图像的背景噪声水平；

步骤5：将步骤2预处理后降噪前图像的像素值减去步骤4中的图像背景噪声水平，从而消除图像中的背景噪声影响；

步骤6：形成步骤4提取的1000个随机伪细胞的蛋白表达值和步骤3得到的降噪前的细胞蛋白表达值的分布，假阳细胞数定义为1000个随机伪细胞中蛋白表达值大于阈值的细胞个数，真阳细胞数定义为降噪前细胞蛋白表达值大于阈值的细胞个数，计算FDR即假阳细胞数在阳性细胞中的比例并确定阳性细胞的蛋白表达值阈值，区分阳性（真实）细胞和阴性细胞，具体过程见图5；

步骤7：在每种蛋白标志物中，计算每个样本中步骤6所确定的阳性细胞的降噪后蛋白表达值，并求出每个样本阳性细胞蛋白表达值的均值，每个样本的归一化缩放参数定义为所有样本中最大的阳性细胞蛋白表达值均值和本样本中阳性细胞蛋白表达值均值的比值，将降噪后图像的像素值乘以归一化缩放参数并通过中值滤波器获得降噪并归一化后的IMC图像，在降噪并归一化后的IMC图像上按照步骤3的方法进行定量，获得降噪并归一化后的细胞蛋白表达值；

步骤8：设置随机伪细胞数量为1000，错误发现率FDR为0.01，将阳性细胞范围外的全部像素置0，采用高斯滤波，降噪后效果图见图6。

对比例

手动确定信号阈值降噪

通过图像可视化软件（Fluidigm公司的MCD Viewer软件，版本号为v1.0.560.6），观察IMC图像的每个通道，在通道设置中可以人为设置不同的最小和最大蛋白表达量阈值。软件将低于蛋白表达量最小阈值的信号确定为噪声并将其值置为零，将高于蛋白表达量最大阈值的信号置为最大阈值，以消除热像素值对后续单细胞蛋白定量分析的影响。如本发明中的示例中，在CD3蛋白通道图像中，在软件中人为设置最小蛋白表达量阈值为3，最大蛋白表达阈值为4.8，得到软件降噪后的效果图（图7）。

由于不同样本间和不同蛋白存在差异性，因此样本的不同蛋白通道的噪音阈值大小不一致，通过手动确定阈值来降噪的方法需要耗费大量的人力和时间，降噪结果也存在强主观性。

应用中值滤波降噪

使用中值滤波器对IMC图像进行滤波，以消除类似椒盐噪点的噪声。在本发明的示例中，使用OpenCV-python软件库（版本号为4.2.0.34）自带的函数medianBlur，窗口大小参数设为5，对CD3蛋白通道图像进行中值滤波操作。中值滤波是将图像上每一像素点值设置为该点邻域窗口内所有像素值的中值，从而对图像进行局部平均(见图8)。

通过图6-8的比较，可以看出，相比手动确定阈值降噪方法，中值滤波降噪方法速度较快，但可能出现聚集成团的噪声消除不干净和去除了某些表达较弱的蛋白质信号的情况，降噪结果不佳。而本发明的降噪方法效果最佳（见图6）。

以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

**说 明 书 附 图**

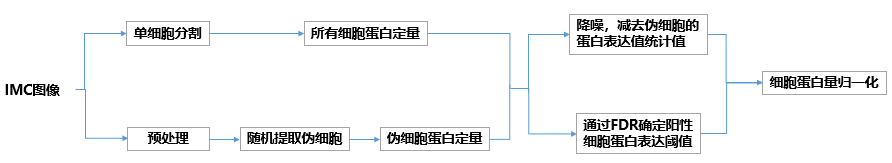


图1

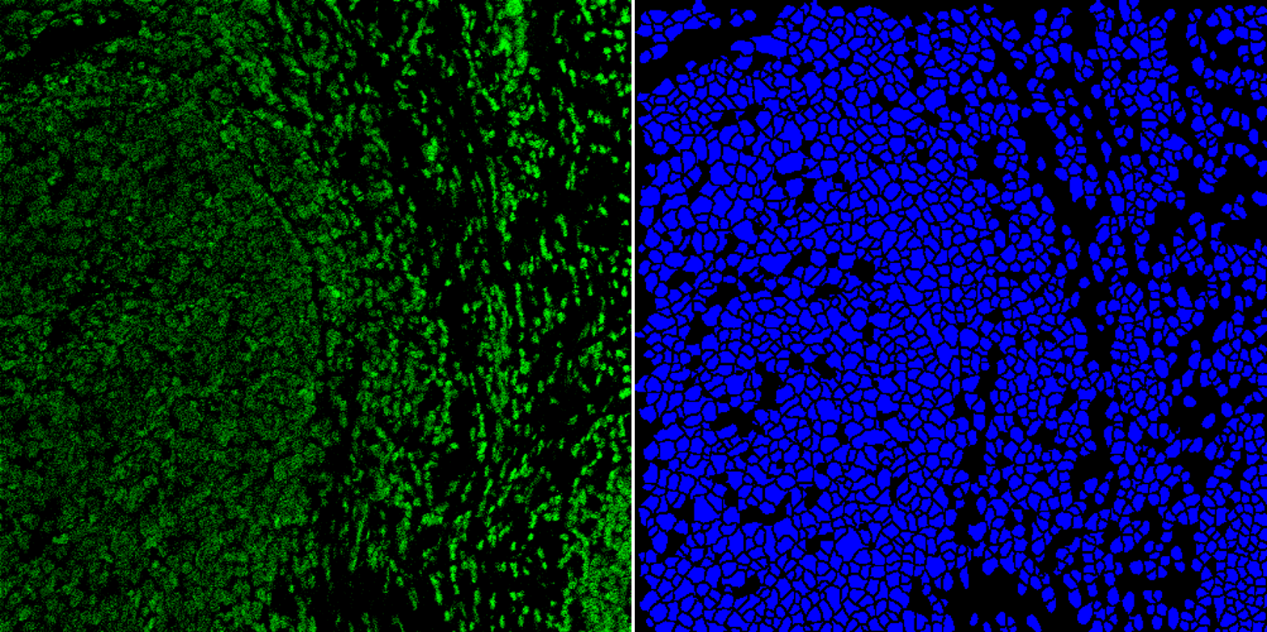
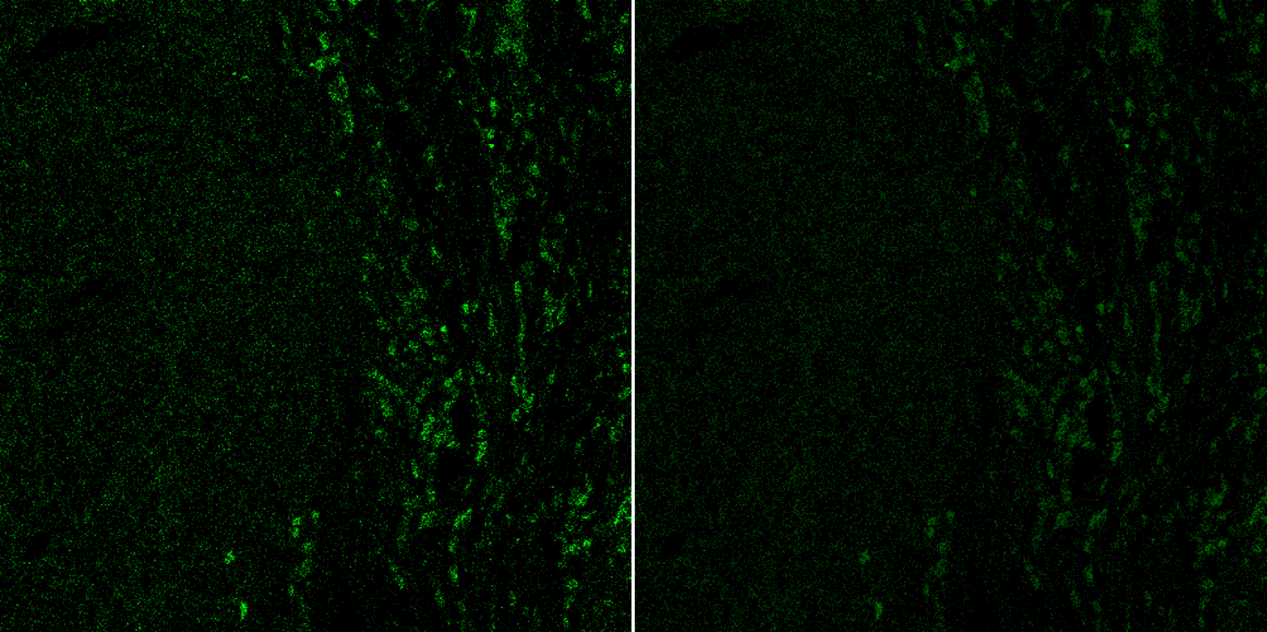
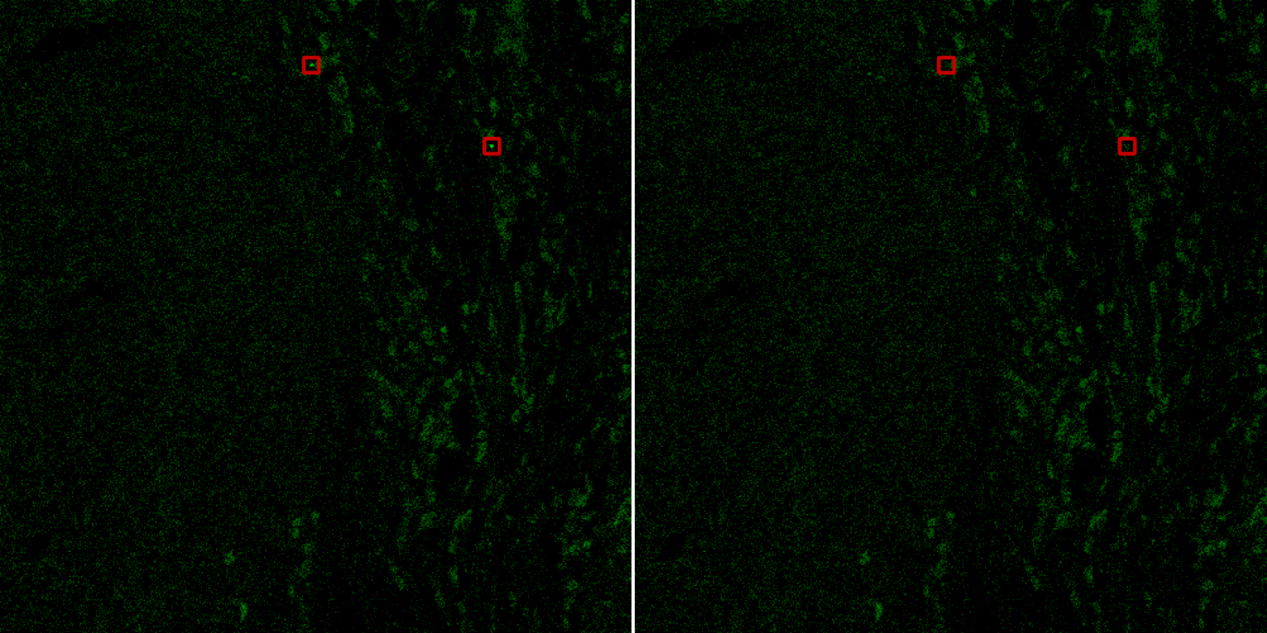


图2



a.



b.

图3

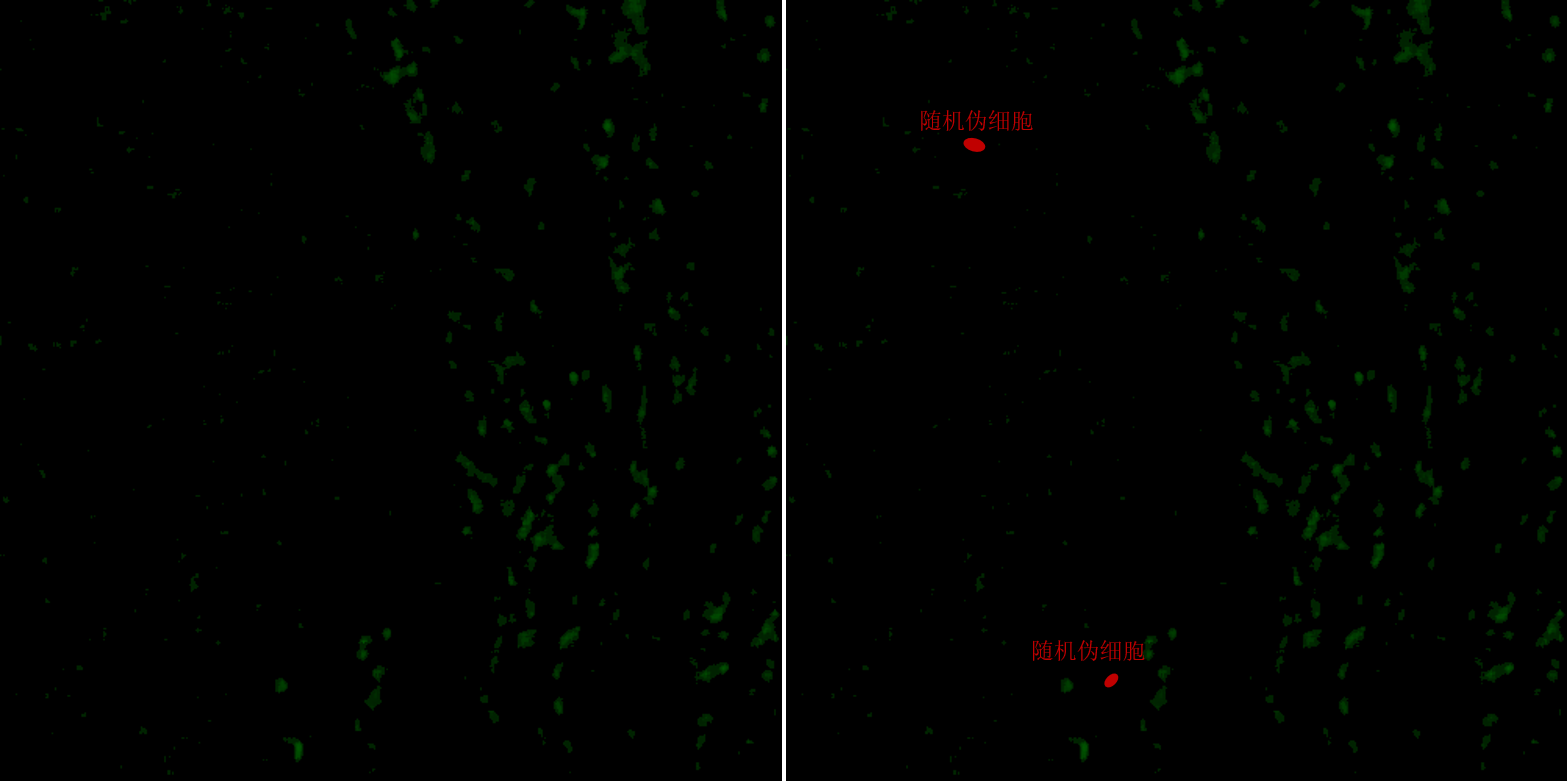


图4

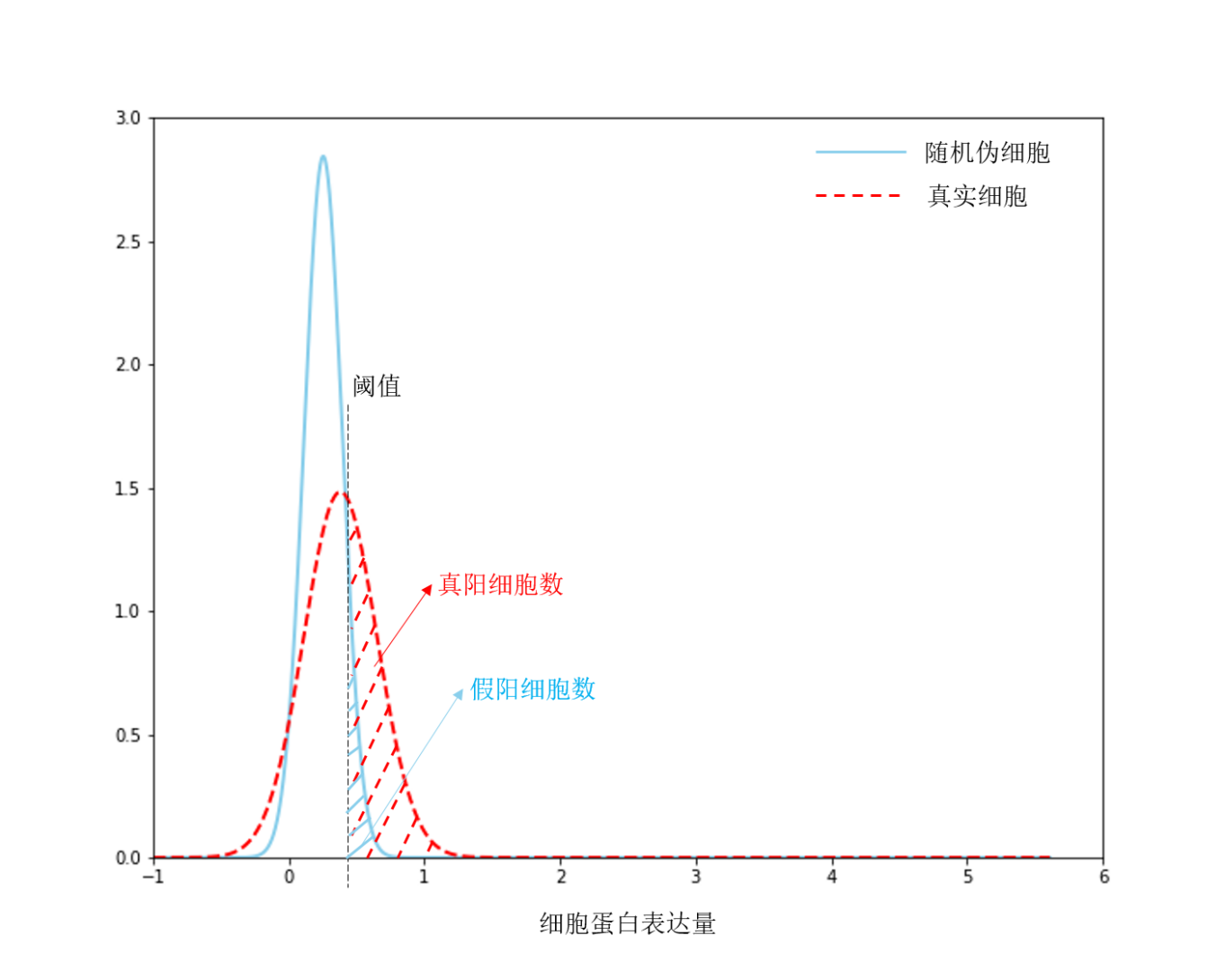


图5

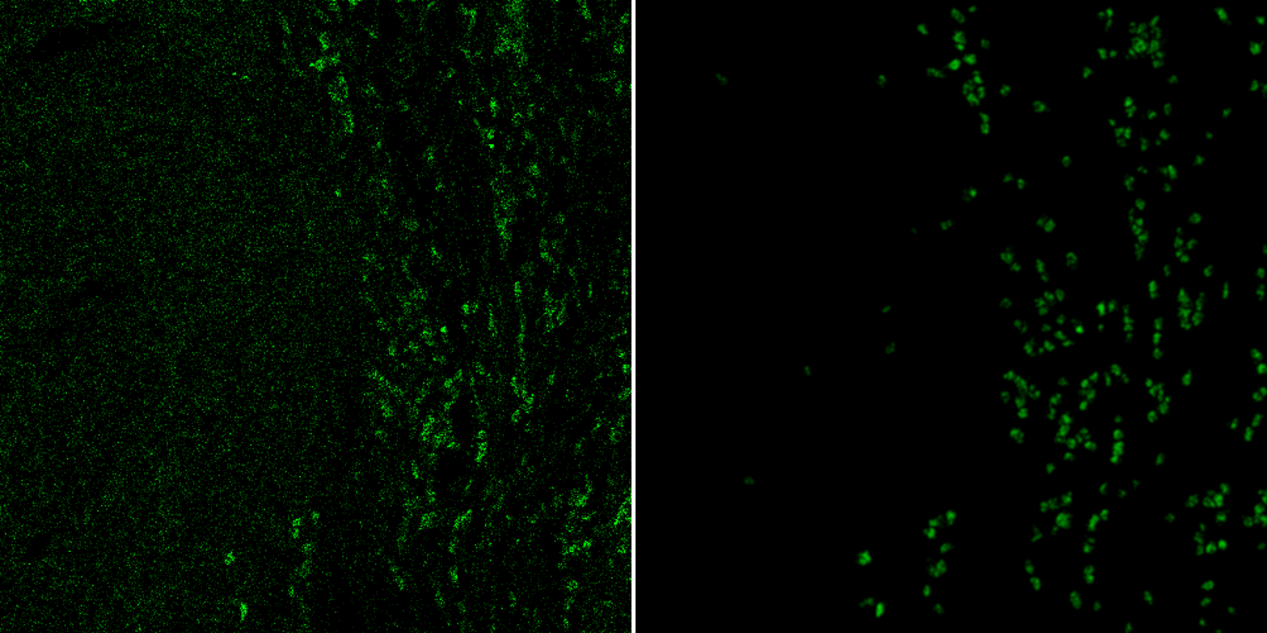


图6

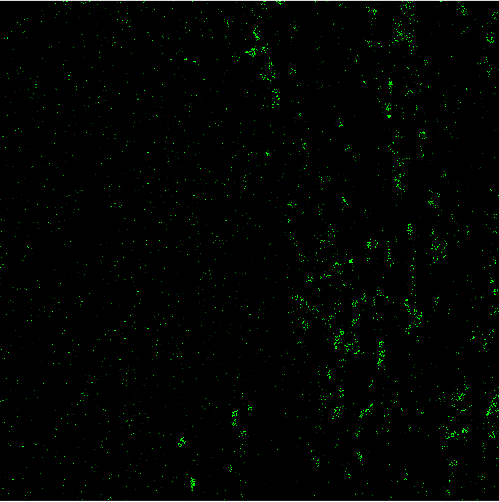


图7

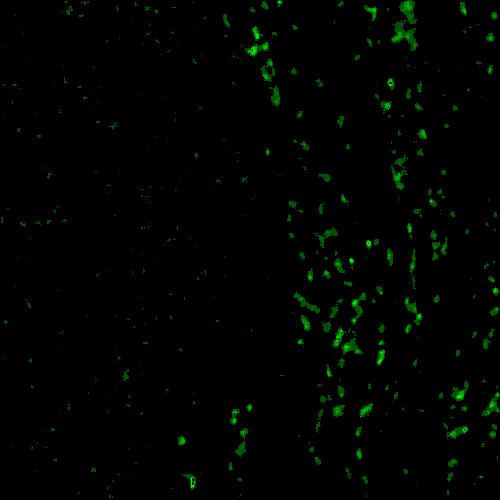


图8