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

肖旭，苏乃斐，杨文娴，俞容山

# 摘要

本文提出了一种基于细胞内图像强度统计分布的方法对IMC多通道图像进行降噪的方法。本方法可以与现有的IMC图像单细胞分割方法联合使用，对采用不同蛋白标志物、不同实验室的不同仪器所产生的IMC数据集均适用，有强泛化能力。经过降噪处理的IMC图像在进行下游分析后，能提供更加精准的单细胞分析结果，帮助IMC图像成为临床研究的利器。

# 关键词

质谱成像细胞分析技术，降噪，归一化，中值滤波

# 背景知识

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。目前的IMC图像处理流程中，往往缺乏统一的质量控制体系以及去噪算法。这些图像噪声直接影响到单细胞的分割、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

本发明提出了一种IMC图像去噪及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。包括如下步骤：

S01：使用已有的图像单细胞分割方法，识别图像上单个细胞。

S02：对IMC图像进行预处理，包括像素值Arcsinh转换，并去除“热像素点”。

S03：将细胞蛋白表达值量化为该细胞所有像素值的平均值。

S04：在图像上按一定的规则选取一定数量的随机伪细胞，按照S03的方法对这些随机伪细胞蛋白表达值定量，形成随机伪细胞的蛋白表达值的分布。

S05：将图像内细胞的蛋白表达值分布与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。

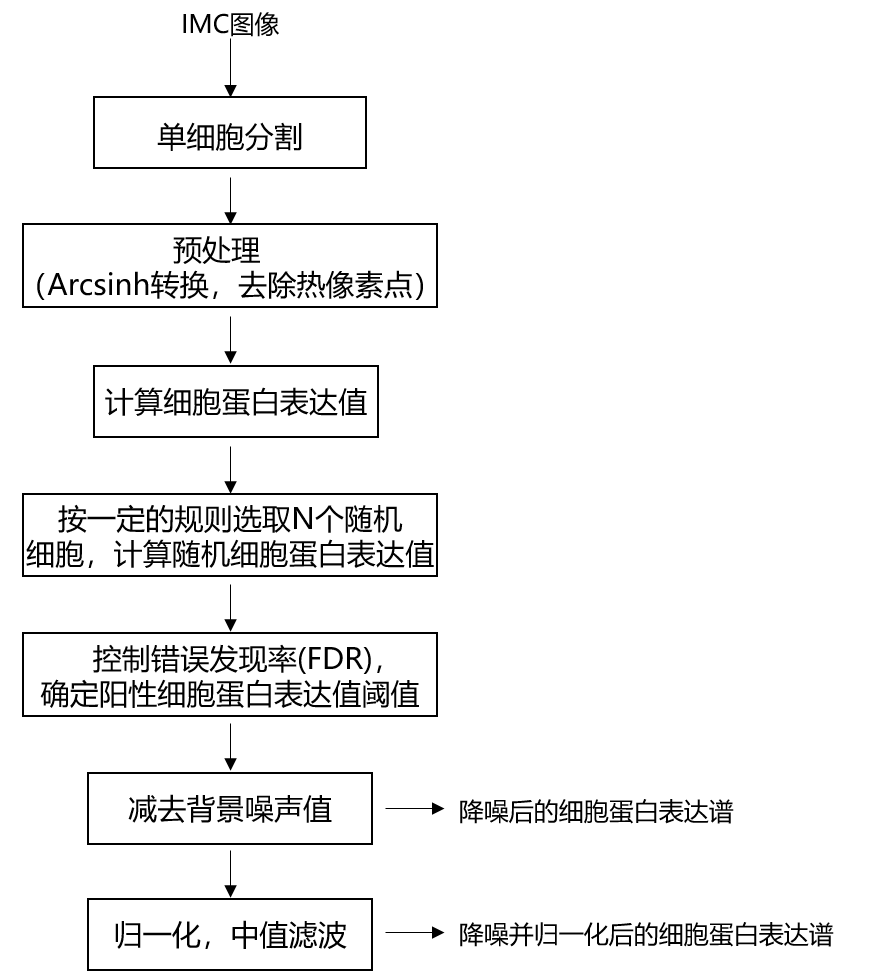
S06：消除图像中的背景噪声影响，得到降噪后细胞的蛋白表达值。

S07：对细胞的蛋白表达值归一化，并将图像通过中值滤波器，最后输出降噪并归一化后的IMC图像和细胞的蛋白表达值。

S08：用户交互式调参和去噪结果的可视化（可选）。

# 具体实施方式

下面结合本发明的流程图，对本发明的技术方案进行清楚、完整地描述。



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。在本实例中，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. 使用中值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；

* 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，是经过步骤2预处理后的像素值，是设定的最小表达量阈值，，是调整后的像素值。

经过上述处理步骤得到图像中，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。接着，调整后的图像使用中值滤波器。

* 1. 提取满足规则的随机伪细胞。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平均噪声水平。

步骤5：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N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。

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6：消除背景噪声影响；

为了消除图像中的背景噪声影响，将步骤3中降噪前细胞蛋白表达值减去步骤4中的图像平均噪声水平，即降噪后细胞蛋白表达值为：

其中，为降噪后的细胞蛋白表达值，为降噪前的细胞蛋白表达值，为随机伪细胞的细胞蛋白表达值。

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5所确定的阳性细胞的降噪后蛋白表达值的均值，并得出所有样本中最大的蛋白表达值均值；

（2）将每个样本中的细胞蛋白表达值归一化到最大均值水平上，即每个样本以一定的比例放大其细胞蛋白表达值，每个样本的比例是上述最大的蛋白表达值均值和该样本的蛋白表达量均值：

其中，是步骤5定义的阳性细胞，是当前样本阳性细胞降噪后的细胞蛋白表达值，是处理的所有样本中样本阳性细胞降噪后的细胞蛋白表达值。

则按比例放大后细胞的蛋白表达值为：

（3）为了去除椒盐噪点，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）。

# 本方法的一个应用实例

我们将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列举如下：

图1：步骤1使用Dice-XMBD得到的单细胞分割结果。

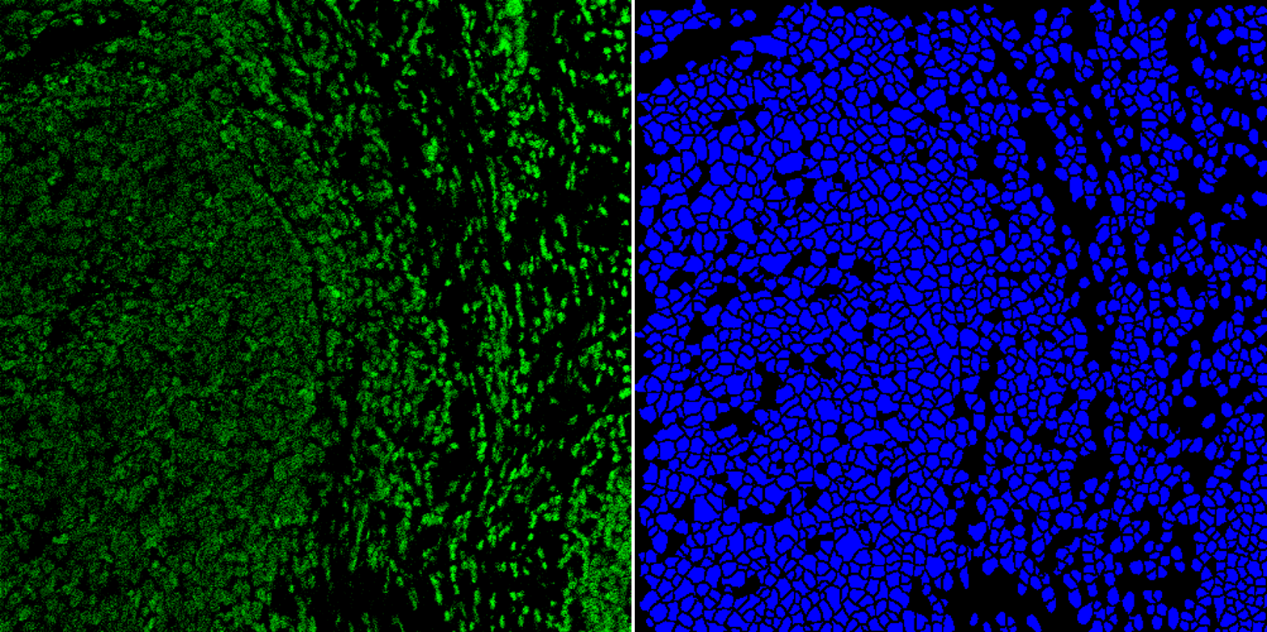
图2：步骤2预处理后的效果图。

图1：（左）原始图像的细胞核通道；（右）单细胞分割图像。

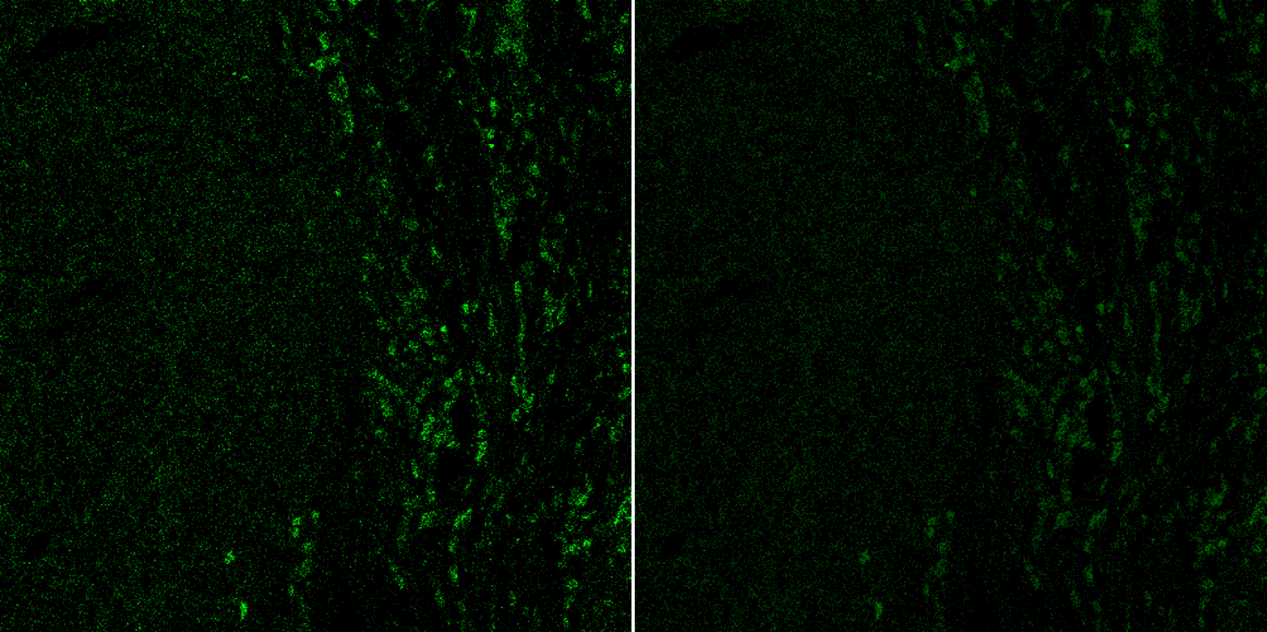


图2a：（左）CD3蛋白通道的原始图像；（右）经过Arcsinh转换后图像。

图2a：（左）CD3通道原始图像，（右）经过Arcsinh转换后图像

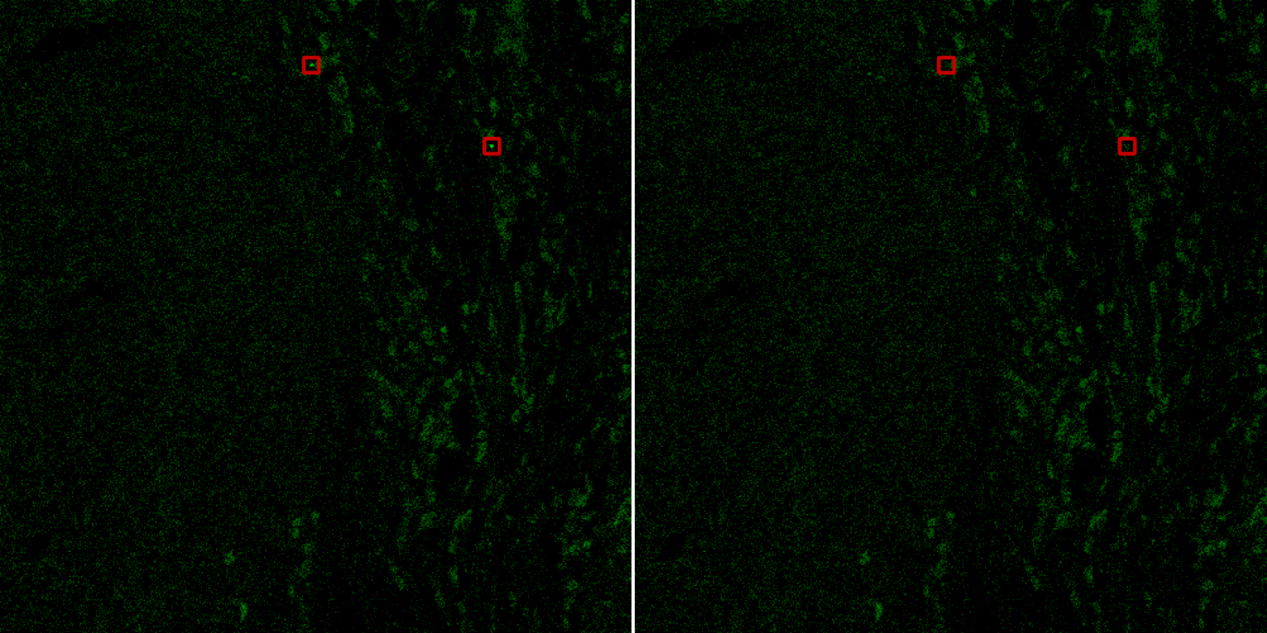


图2b：（左）经过Arcsinh转换后图像；（右）去除“热像素点”图像。

图3：步骤4设定最小表达量阈值参数为步骤2预处理后图像的0.05倍99分位像素值，区分出的真实细胞表达区域和噪声区域，及提取随机伪细胞的示意图。

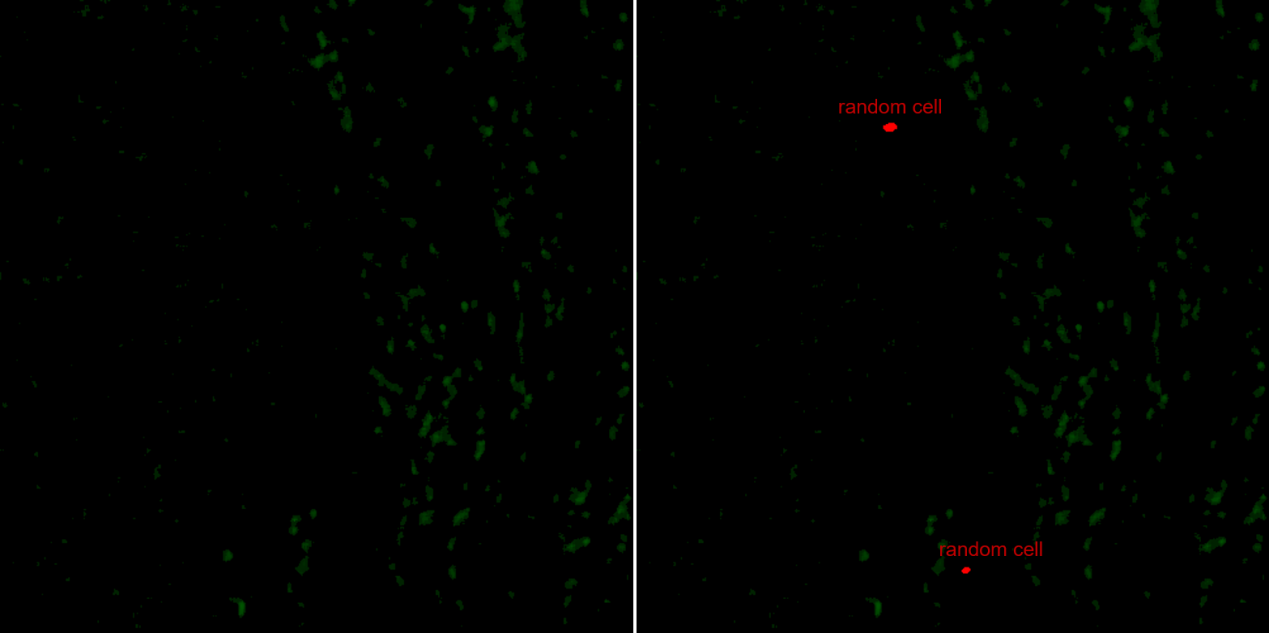


图3：（左）真实细胞表达区域和噪声区域区分图，绿色为细胞蛋白表达区域，黑色为噪声区域；（右）提取随机伪细胞示意图。

图4：步骤5计算FDR并确定阳性细胞的蛋白表达值阈值示意图。

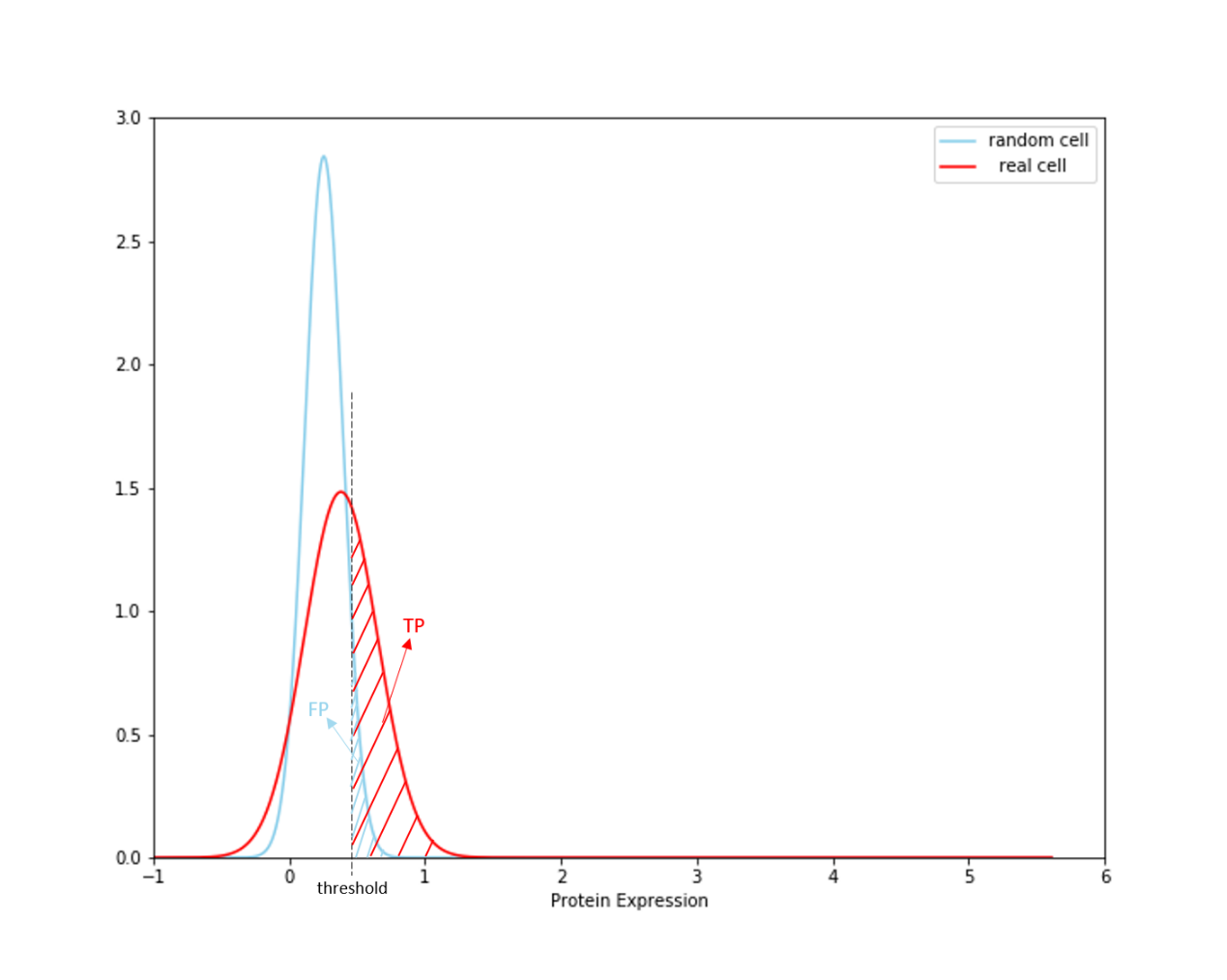


图4：蓝色曲线为随机伪细胞蛋白表达量分布曲线，蓝色阴影区域即假阳细胞数；

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5：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。

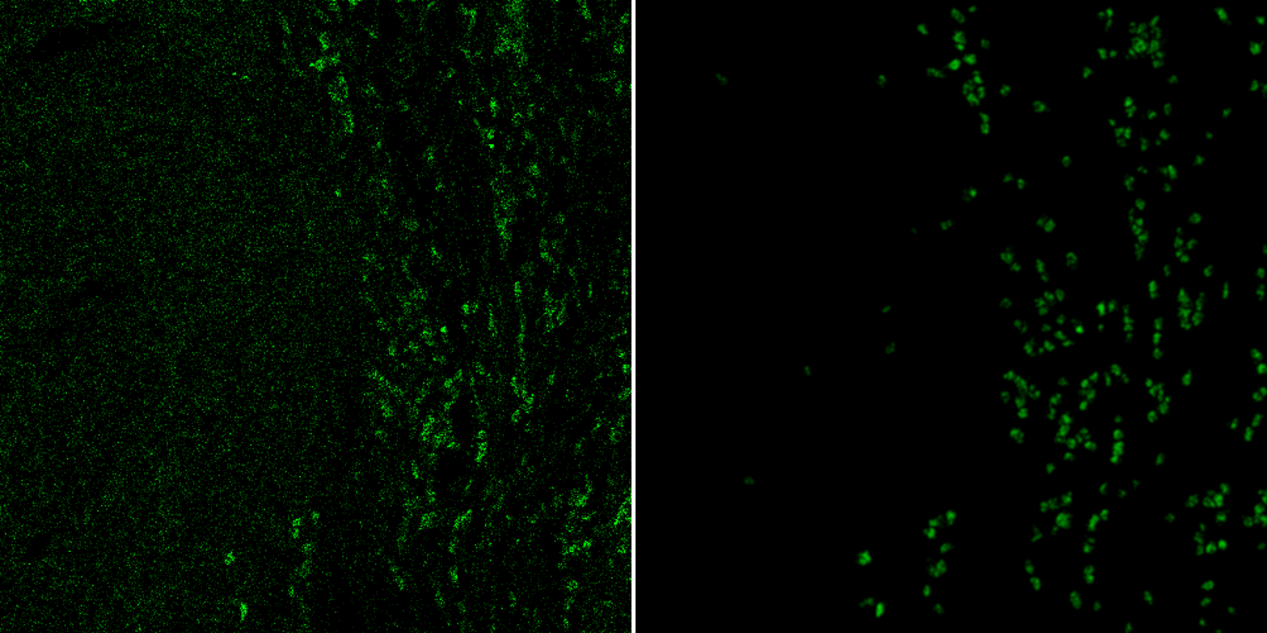


图5：（左）降噪前图像，（右）降噪后图像