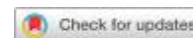


物品打开



一种用于分子标签转移的深度学习模型 从组织病理学图像中识别癌细胞

苏安德^{1,6}Lee先生^{2,6}，肖谭^{1,6}卡洛斯J. 苏亚雷斯³，Noemi安多尔^{2,5}、全阮¹✉和汉利P. 吉^{2,4}✉

深度学习分类系统有改善癌症诊断的潜力。然而，到目前为止，这些计算方法的发展依赖于先前的病理注释和大型训练数据集。手动注释是低分辨率，耗时，高度可变和受观察者方差。为了解决这个问题，我们开发了一种方法，H&E分子神经网络（HEMnet）。HEMnet利用免疫组化作为H&E图像上癌细胞的初始分子标记，并在重叠的临床组织病理学图像上训练癌症分类器。利用这种分子转移方法，HEMnet成功地从10张全幻灯片图像中生成并标记了21,939个肿瘤和8782个正常瓷砖，用于模型训练。建立模型后，HEMnet准确识别了结直肠癌区域，与p53染色和病理注释相比，ROC AUC值分别为0.84和0.73。我们的验证研究使用来自TCGA样本的组织病理学图像准确地估计了肿瘤纯度，这显示出与基于基因组测序数据的估计有显著的相关性（回归系数为0.8）。因此，HEMnet有助于解决癌症深度学习分析中的两个主要挑战，即需要有大量的图像进行训练和对病理学家手工标记的依赖。与人工组织病理学评估相比，HEMnet还能预测更高的分辨率的癌细胞。总的来说，我们的方法提供了一种全自动描述任何类型的肿瘤的路径，只要有一个癌症导向的分子染色可用于后续学习。软件，教程和交互式工具是可用的at: <https://github.com/>

npj精密肿瘤学（2022年）6: 14; <https://doi.org/10.1038/s41698-022-00252-0>

生物医学学习/HEMnet

背景

组织的组织病理学检查对于准确诊断和治疗癌症是必不可少的¹⁻³。通常，癌症的病理诊断和不同的亚型决定了特殊的治疗方案的使用⁴。目前的癌症诊断标准之一是用苏木精和伊红（H&E）染料联合染色的肿瘤组织切片的显微镜检查^{2,3}。基于活检切片的h&e染色图像，病理学家可以定性地评估癌症类型、分期和估计肿瘤纯度³。此外，组织病理学检查经常报告不同类型的细胞、器质状态和/或复杂组织内的细胞定位⁵。尽管病理学家之间的诊断一致性仍然很低⁶。对活检的组织病理学切片的目视检查是一项耗时的任务，并且缺乏对细胞特征的定量测量⁴。

近年来，数字病理学的新兴领域已经发展成为一种数字化、存储和分发癌症健康图像（WSIs）的方式。这种方法显著提高了癌症解剖病理的速度和途径。随着WSIs产量的增加，需要开发先进的计算方法，以快速、健壮和准确的方式分析这些医学图像，最终导致在自动癌症诊断中的应用⁷⁻¹⁰。

深度学习是数字组织学图像分析的首选方法，目前已经发展了许多方法用于肿瘤分类⁸。然而，深度学习的一个关键挑战是需要大量准确标记的数据¹¹。

对于这种方法，许多方法都需要由病理学家手动注释的wsi¹²。因此，生成训练数据集成为一个耗时的手工过程，这仍然有一个在病理学家之间的高差异的局限性¹³。这增加了成本，并使获取这些训练数据集的成本更加昂贵¹⁴。另一个挑战是这些幻灯片图像很大：一个×为10放大的图像可以包含数百万像素。然而，病理学家的注释通常不是在像素级，而是依赖于许多更粗糙的划分方法。因此，训练发生在一个较低的图像分辨率，缺乏细胞粒度¹⁵。我们的目标是解决三个关键的挑战，即依赖于模型训练的可变病理学家注释，需要大量的图像进行训练，以及实现高分辨率和定量预测癌细胞的需求。

在这里，我们描述了一种新的自动化方法，其中我们使用预先染色，以更高的图像分辨率将肿瘤与正常细胞区分开来。免疫组化（IHC）已经成为研究和临床诊断中的一个有用的工具——经典的组织病理学方法基于抗原-抗体的结合来定位和可视化特定的细胞或抗原。重要的是，IHC被广泛应用于福尔马林蜡副石蜡包埋（FFPE）组织，这是最常见的组织存档方法¹⁶。将H&E和分子标记物染色图像的手工耦合来检测（通过H&E）和进一步的鉴定（通过IHC）正越来越多地应用于组织病理学诊断⁶。这也为数字数据集创造了宝贵的机会。

¹昆士兰大学分子生物科学研究所，布里斯班，QLD 4072，澳大利亚。²美国斯坦福大学医学院医学系肿瘤科，斯坦福大学，加州94305。³斯坦福大学医学院病理学系，美国斯坦福大学，加州94305。⁴斯坦福基因组技术中心，斯坦福大学，帕洛阿尔托，美国加州94304。⁵现任地址：美国佛罗里达州坦帕市木兰花路12902号，莫菲特癌症中心综合数学肿瘤学系。⁶这些作者的贡献相同：苏安德鲁、李华俊、肖Tan✉. email: quan.nguyen@uq.edu.au; genomics_ji@stanford.edu

与明尼苏达大学荷美尔研究所合作出版

tu

组织形态和分子膜，尚未被利用^{2, 17, 18}。

我们开发了一种方法，称为H&E分子神经网络（HEMnet），该方法可以自动将IHC图像中的每个像素与H&E图像上相同位置的相应像素进行对齐。我们的方法将每个H&E像素标记为生物阳性或阴性标志物。在这项概念验证研究中，我们使用了一种针对癌症的IHC标记物来描述肿瘤细胞。我们使用了p53染色，这是一种重要的肿瘤抑制基因（TP53），它在许多不同的恶性肿瘤中容易发生高频率的基因改变^{19, 20}。大多数TP53突变是改变p53蛋白结构的错义类，使它们比野生型更稳定，有更长的半衰期。TP53突变导致p53在恶性细胞中的稳定并随后积累²¹，使它很容易被IHC检测到。野生型p53并不稳定，半衰期较短，因此正常细胞中的p53通常无法被IHC检测到²²。高达74%的结直肠癌样本显示p53异常高阳性染色（即棕色），这为结直肠癌的癌细胞提供了特殊的IHC标记^{19, 20, 23}。通过将p53 IHC图像映射/配准到H&E图像上，我们改进了模型训练和测试数据集，如下所述。

我们的研究利用创新的分子标签转移，从WSIs中提取了数万块的H&E瓷砖，而不需要人工检查或以最小的努力来协调自动标签。在这里，HEMnet对一组来自结直肠癌的p53染色和H&E WSI图像进行了训练。我们使用异常的p53染色模式，通过对齐这些图像来注释H&E切片中的结直肠癌细胞。基于内部的结直肠癌数据集，训练了数千个标记块的卷积神经网络分类器。通过这种训练和测试方法，我们在独立的一组组织病理学切片和图像上取得了高性能。HEMnet被扩展到测试癌症图像档案（TCIA），该档案有一个广泛的结直肠癌组织病理学成像数据存储库。通过与其他基于基因组学的方法进行比较，我们证明了一个具有显著正相关的高性能^{24, 25}。概括推广，只要分子标记对肿瘤细胞相对特殊，这一过程就能够对癌症和正常细胞进行简化和高分辨率的分子注释。HEMnet方法可以很容易地与其他有趣的生物标志物一起实现，如HER2和其他类型的癌症。多重标记物分析的最新发展，如大量细胞术成像，将使多个标记物的标记转移到H&E图像，从而在更大程度上分析癌症复杂性。鉴于其所取得的成功，该方法具有潜在的临床应用潜力。我们可以使用常见的组织病理学图像来发现组织内的癌症细胞几何模式，我们的软件能够自动检测这些模式，作为开发计算机辅助诊断工具的一部分。

结果

用于H&E图像注释的分子信息

我们开发了一种利用分子注释和深度学习方法来改进癌细胞识别的方法（图. 1）。HEMnet的开发管道包括四个主要步骤：（1）成对的P53和H&E图像的数据生成，（2）图像的预处理和分子标签的转移，（3）训练中性网络，（4）评估HEMnet的性能（图. 1）。HEMnet管道的设计适用于任何染色类型或癌症类型。

在这项研究中，我们开发了HEMnet来识别结直肠癌H&E图像中的肿瘤细胞。在第一步中，我们获得了32张高分辨率的H&E图像和相应的p53 IHC图像

27个癌症样本和5个非癌症样本。这是通过用H&E和p53对相邻的组织切片进行染色，为每个组织块生成匹配的配对wsi来实现的。步骤2是HEMnet将分子标记转移到H&E图像上的新贡献。HEMnet利用了分子信息，而不是手动的病理学家注释。我们通过将p53分子染色图像与相应的H&E图像对齐来实现这一点。3）。因此，p53染色模式被用于自动标记成对的H&E图像上的癌症区域，而不需要病理学家的干预。在步骤3中，每个标记的H&E图像被分割成数千个小块 $224 \times 224 \text{px}$ ，这样我们就可以从10个WSIs的小样本中生成成千上万的训练样本（图. 3d）。我们使用这些图像块来训练一个深度转移学习分类器，仅使用组织形态学特征来识别临床H&E图像中的癌症区域。步骤4提供了具有独立数据集的严格验证标准，将HEMnet与病理注释和7种计算基因组学方法进行了比较。

H&E染色的标准化减少了颜色的变化

除了实现使用分子标签的概念在挖掘模型，HEMnet管道的技术贡献在于无缝管道，包括一个步骤将多个图像纳入模型训练和测试数据集通过规范化不同的图像，其次是快速和准确的标签映射，之前训练一个神经网络。最初，具有相似组织结构的WSIs由于载玻片处理的不同（如染色时间、显微镜曝光）而染色不同的颜色。我们用染色归一化来解决这个问题，这导致这些WSIs采用了模板载玻片的染色着色剂，并增加了亮度以产生白色背景（图. 2a-c和图. S2）。该方法改变了归一化载玻片的平均R、G和B通道强度，使其与模板载玻片非常相似，同时保留了图像中R、G和B的颜色分布。在32个H&E WSIs中，染色归一化降低了平均R、G和B通道强度的变化（图. 2d）。此外，它还调整了中值通道强度的中值，使其更接近模板图像的平均通道强度。通过在输入模型之前对所有图像进行归一化，我们确保模型可以推广到与训练幻灯片不同的新幻灯片。

将p53分子标记转移到相应的H&E上 图像

对应的p53和h&e染色切片的WSIs经常错位（图. 3a）。为了使p53阳性细胞在H&E图像上准确地映射到癌细胞上，我们通过HEMnet自动图像配准将p53图像重新对齐到相应的H&E图像上（图. 3c）。我们基于强度的互信息优化配准方法是快速和准确的。3b, c）。接下来，我们根据p53染色模式标记H&E图像，其中p53阳性区域被标记为癌症，反之亦然。为了抵消p53染色的限制，标记癌细胞，只有p53阳性的瓷砖来自癌症切片和只有来自非癌症切片的p53阴性瓷砖被用于训练。所有其他的贴图都被标记为不确定的，并被排除在任何其他处理之外。在 $\times 10$ 放大时，一个WSI可以生成数千块瓷砖用于训练（图. 3c）。我们从分子标记的H&E图像中生成了 224×224 像素的瓷砖，以训练VGG16深度学习模型（图. 3d）。

分子注释质量控制产生了一个高的 Condians数据集

TP53肿瘤抑制基因是人类癌症中最常见的突变基因（50%），并且不成比例地具有这些突变基因

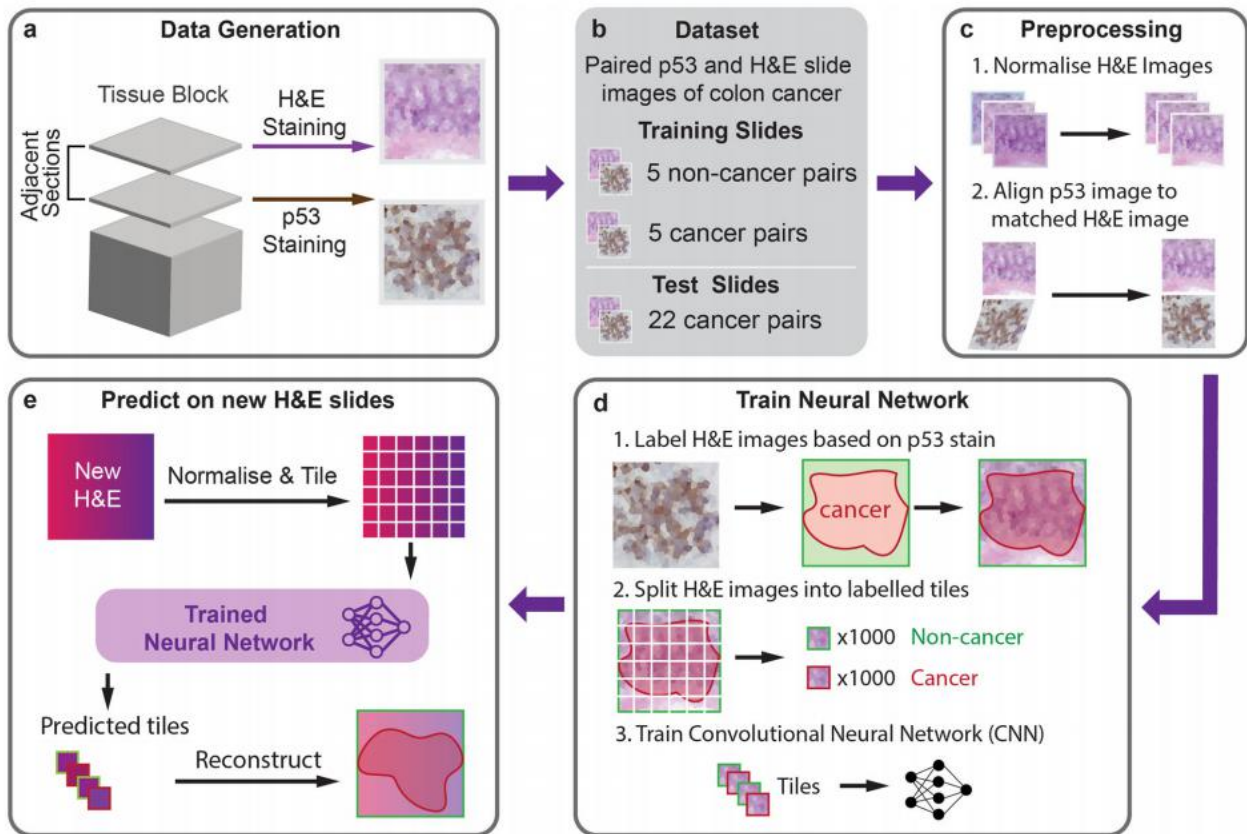


图1 H&E分子神经网络 (HEMnet) 工作效率概述。a匹配的p53 IHC染色和h&e染色的WSI来自于两个相邻的组织切片。b对配对的正常和癌症载玻片 (ive对) 进行训练。测试幻灯片被搁置起来, 在模型训练中看不到。c预处理, 通过染色归一化和图像配准来解释载玻片制备的技术变化。d分子标记从p53转移到H&E图像上。标签转移后, 每张图像被平铺生成数千个小样本 (224×224 像素) 来训练一个CNN。e应用HEMnet从新的临床H&E图像中预测癌症。

高达70%–80%的结肠癌患者的突变和其他基因改变^{26, 27}。由于其普遍流行, 它提供了一种高度可推广的方法来分子注释广泛的癌症。与其他IHC标记物类似, p53染色也有其局限性, 即在一幅图像内或图像之间, 该标记物并不总是指示癌症, 反之亦然。例如, p53的过表达和阳性染色可能发生在响应DNA损伤的正常细胞中。此外, p53在TP53缺失的p53基因缺失的癌细胞中可能缺失²²。为了克服这些限制, 在训练我们的模型时, 我们只考虑p53阳性细胞来自癌症玻片, 而p53阴性细胞来自癌症玻片, 其中细胞形态正常 (图. 3d)。通过这种方式, 我们一致认为细胞被正确标记了, 有8782个非癌症瓷砖和21, 939个癌症瓷砖。我们删除了23, 275个有一定程度的不确定性的瓷砖 (图. 3d)。

对癌细胞的高性能自动评估 丰度和空间分布

我们将训练过的HEMnet应用于看不见的WSIs来预测癌症区域。在测试数据集中17张未见的H&E切片中, 所有切片都有相应的p53染色切片, 13张有癌症区域的额外病理学注释。我们发现HEMnet可以准确预测p53染色模式 (ROC AUC = 0.73) 和病理学家注释的癌症区域 (ROC AUC = 0.84)。4a, b)。这些结果表明, 使用分子标记H&E图像开发的分类器, 可以从其一般形态学中预测特定组织样本的p53阳性癌症区域。

将p53标记的瓷砖与来自同一位置的病理学家标记的瓷砖进行比较, 我们发现瓷砖标签总体一致 (ROCAUC=0.67) (补充图. 6)。然而, 这项协议并不是绝对完美的。为了评估任何差异, 我们对每张幻灯片测量了p53染色注释癌症的能力。该分析涉及计算每个病理学家的p53染色和地面真实标签之间的ROC AUC。我们发现HEMnet p53表现 (ROC AUC) 更高, 其中p53更准确地标记癌症 (p53 vs病理学家标记ROC AUC) 显著, Pearson系数为1.02, 而 $R^2_{.94}=0$ (图. 4c)。这一结果表明, 该模型学会了识别癌细胞的特殊形态特征, 并不严格局限于识别具有高水平p53的细胞。这可能是由于癌细胞在形态上与正常细胞不同, 而p53阳性细胞和阴性细胞在形态上的差异更为微妙。我们注意到, 有一些例子表明, HEMnet可以识别由病理学家标记的癌症, 即使是在p53染色不能识别癌症的地方 (图. 4d, e)。综上所述, HEMnet能够准确识别癌症的组织形态特征。

外部验证和应用到TCGA建议 广泛适用性

作为一个使用外部数据集的独立验证, 我们将HEMnet应用于TCGA结肠癌样本中的结肠腺癌样本。我们利用这些CRC来研究该方法的推广性和临床应用

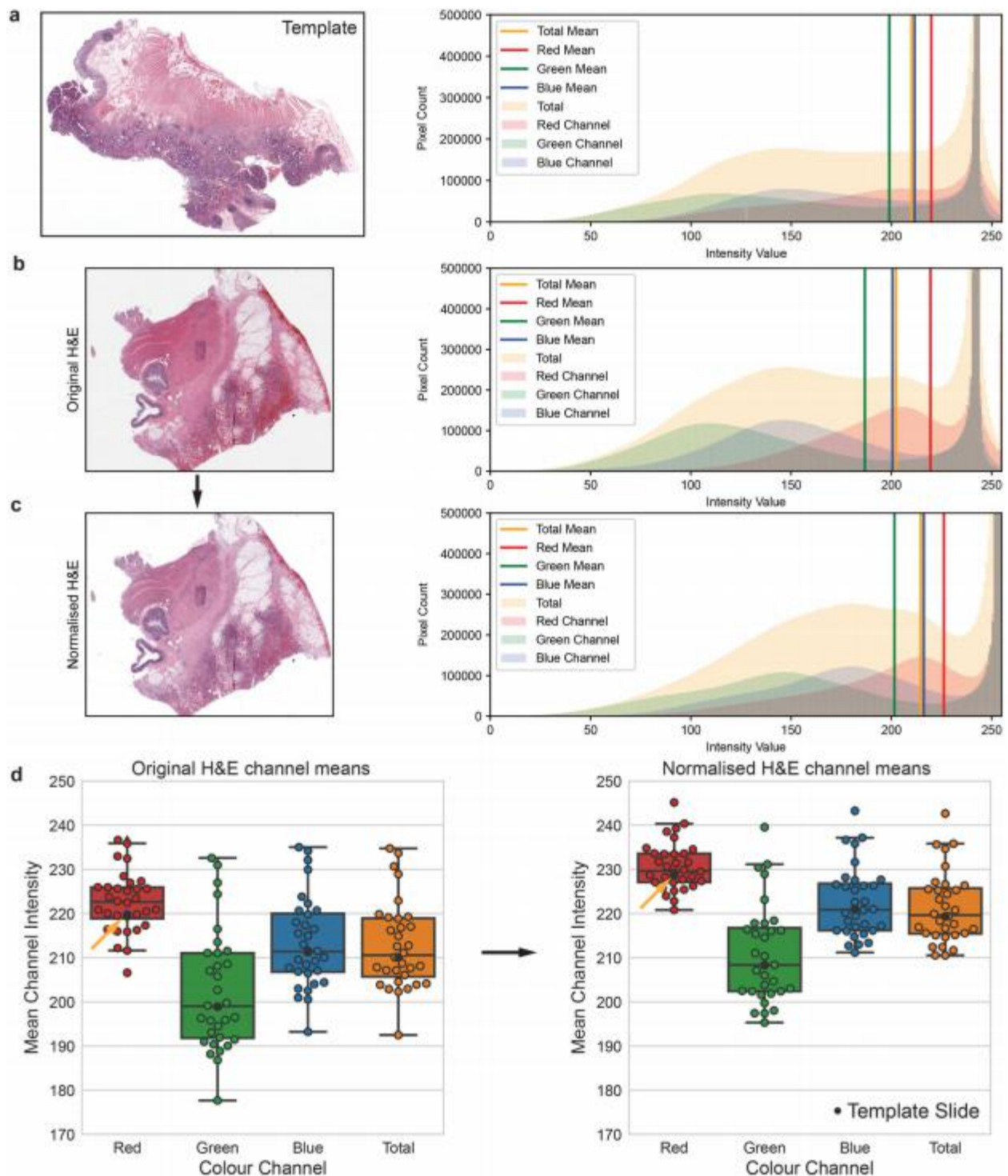


图2 H&E染色正常化。模板玻片-癌症玻片的平均R、G和B通道强度与所有图像的平均通道强度的中位数最相似。x2放大图像的直方图（a、b、c）。b归一化前的H&E图像。归一化后的C-H&E图像更接近于模板图像。增加图像亮度，保留像素强度分布。d所有幻灯片的归一化（ $n = 32$ ）。经过归一化后，平均通道强度的变化减少。模板幻灯片表示经过归一化（用箭头表示）后的通道强度更接近中线（箱线图中心线），并缩小了四分位数范围（箱线图边界框）。箱线图须表示数据范围，不包括异常值。

（补充表2）。通过本研究中描述的内部数据集，对未经修改的HEMnet模型进行训练，以预测结肠腺癌的H&E WSIs。通过将瓷砖水平的预测与每个瓷砖的细胞含量相结合，我们计算了每张载玻片的癌症组织占总组织的比例

（补充表2和图. 5a）。这是肿瘤纯度的近似值，我们与测序方法估计的匹配的基因组数据进行了比较。在我们的结肠癌数据和TCGA数据之间有几个差异。最重要的是，测序并没有在相同的平台上进行

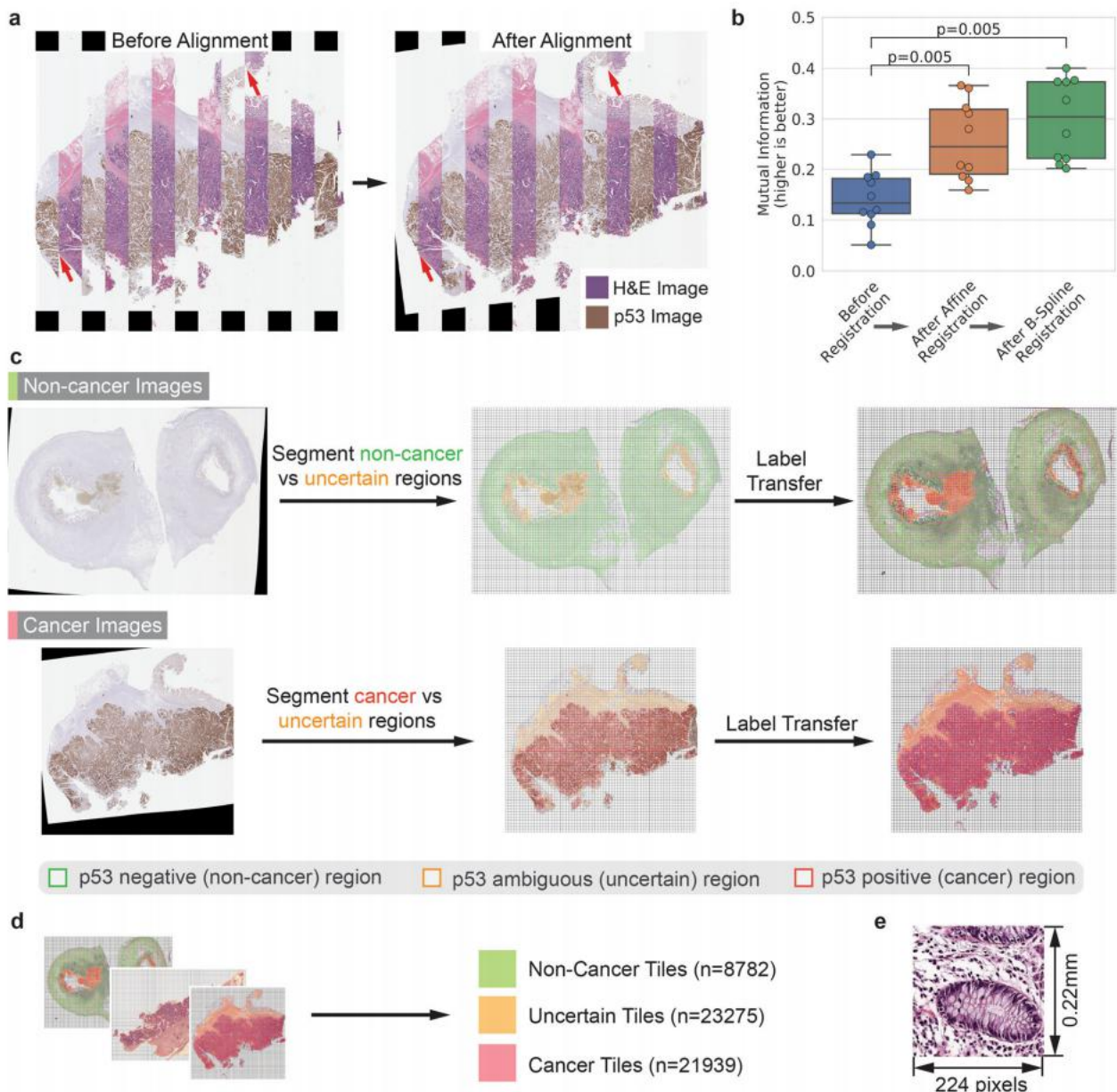


图3用H&E图像的分子标记来训练神经网络。aH&E和匹配的p53图像叠加显示配准后的对齐，用红色箭头突出显示。bp53图像与相应的H&E图像的准确对齐。连续的afine和b样条配准增加了互信息，一种图像相似性的度量。具有t检验的标志性试验。箱线图中心线表示中值，边界框表示四分位数范围，须表示数据范围。cp53图像的分割以标记匹配的H&E图像，其中只有非癌症幻灯片来自非癌症幻灯片，反之亦然。d 10训练H&E图像生成了数以万计的瓷砖，增加了样本量。e在 $\times 10$ 放大下生成并用于训练神经网络的癌症瓷砖的例子。

用于诊断成像的组织。尽管存在这些挑战，我们发现我们的方法与绝对估计的肿瘤纯度之间存在显著的相关性，回归系数为0.8，如图5所示。此外，我们还研究了HEMnet的表现是否受到以下因素的影响：(i) TP53突变状态，(ii) 临床分期，(iii) MSI状态，以及(iv) CMS-RF分类器。我们发现，无论TP53突变背景如何，HEMnet都表现良好。5a和补充图。8)。其他因素对HEMnet的性能没有显著影响。这些结果表明，HEMnet可以推广到新的结直肠癌临床数据，并能够可靠地预测TCGA图像。正如我们所观察到的，我们的预测在检测真阳性（癌细胞）和真阴性（正常细胞）方面一般都是准确的，但它也有很小的组织比例和假阳性

（预测正常的上皮细胞为癌细胞，通常被发现为模糊的区域，HEMnet概率得分低于癌症区域）。然而，我们相信，用我们的预测评分标注的瓷砖可以帮助病理学家快速检查幻灯片，并验证模糊的区域（补充图。9）。

讨论

H&E图像的组织病理学检查已经是几乎所有疑似癌症患者的病理诊断的金标准^{3,28}。机器学习工具分析H&E图像的现代应用越来越多^{7,29}，已经有了一些计算机辅助的图像诊断工具

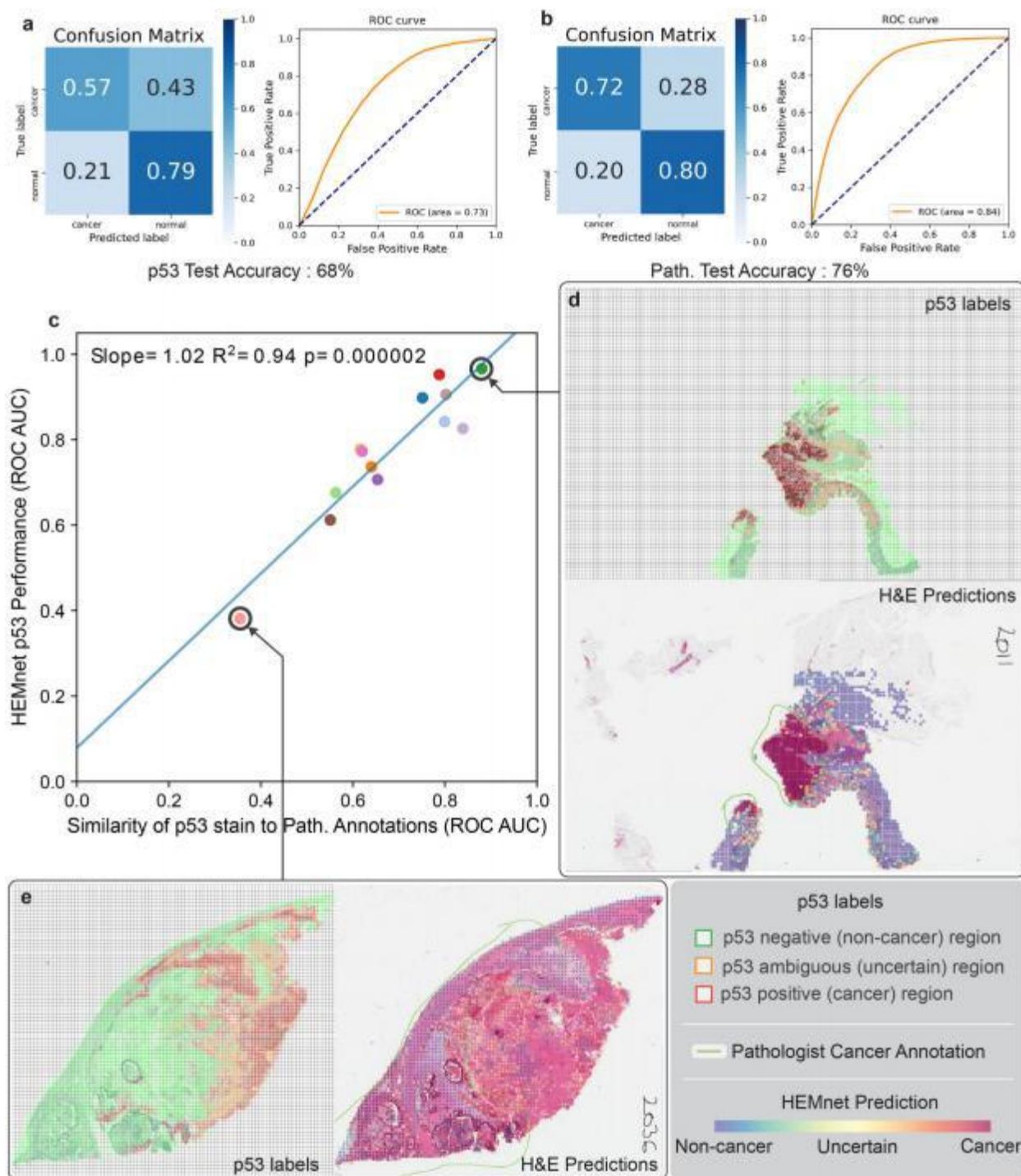


图4. HEMnet在看不见的H&E幻灯片上的表现。a对17张看不见的H&E切片上的p53染色模式的预测。b与病理学家的注释相比，对13张看不见的H&E切片上的癌症区域的预测。c如病理学家所注释的那样，p53染色模式的预测性能与p53标记组织上肿瘤区域的能力呈正相关。当p53染色与病理学家注释一致时，d HEMnet可以准确预测病理学家（下）和p53染色（上）注释的肿瘤区域。e HEMnet即使在p53染色模式（左）与病理学家的基础真相注释（右）不一致时，也能预测癌症区域。

经美国食品和药物管理局（FDA）批准的³⁰。数以百出的深度学习方法已经被开发出来，仅使用H&E图像来检测和诊断癌症⁷。虽然其中一些方法已经取得了很高的性能，但它们都依赖于病理注释将图像标记/分割为多个组织区域类别^{7, 31}。值得注意的是，病理学家的金标准注释并不总是基本的事实和

病理学家之间的注释存在着固有的差异。例如，在黑色素瘤的病例中，II类（35.2%）、III类（59.5%）和IV类（63.2%）的观察者内重现性较低³²。大多数方法还需要大量的注释图像来进行模型训练和评估^{33, 34}而缺乏大型标注数据集是深度学习图像分析面临的主要挑战⁷。我们开发了HEMnet作为一种癌症诊断框架

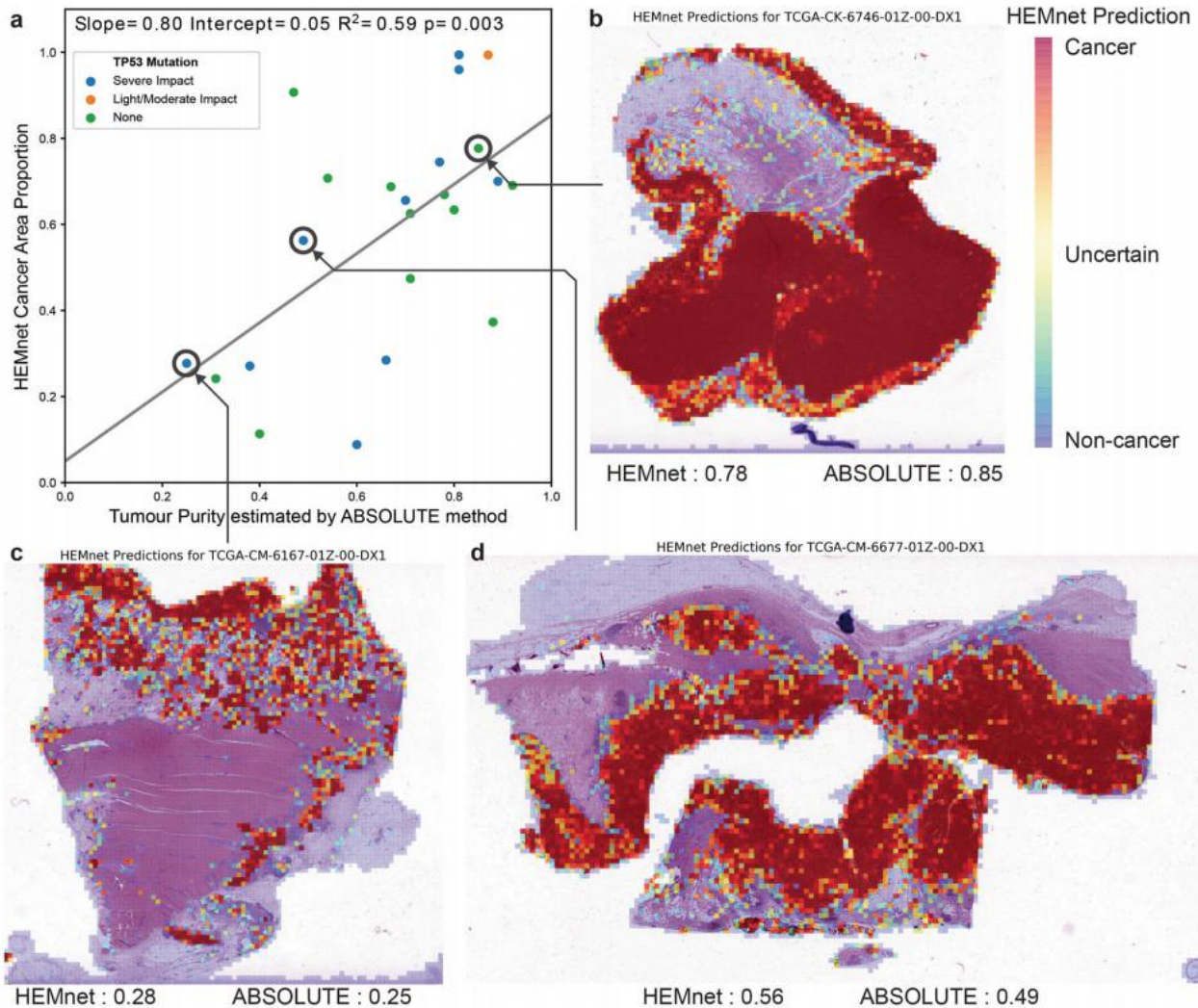


图5. 癌症基因组图谱 (TCGA) 的外部验证。a将HEMnet估计的肿瘤纯度——近似于肿瘤组织面积与总组织面积的比例——与使用绝对方法测序的肿瘤纯度估计值的比较 ($n = 24$)。这些点的颜色代表了从TCGA数据中获得的三类TP53突变。b, c, d HEMnet对福尔马林混合中低(c)、中(b)和高(d)肿瘤纯度的结肠腺癌的TCGA切片的癌症预测。

它使用数字标记和神经网络来解决这些挑战。

HEMnet结合了两种常见的组织病理学WSI数据，即H&E染色和免疫组化染色图像。HEMnet管道的新颖之处在于分子标记转移，它允许使用像素级的分子信息癌细胞（如P53阳性/负像素），其分辨率比人工病理分割高数十倍。在HEMnet中，我们解决了几个关键的技术挑战，以允许准确、快速和泛化的标签转移，最终目标是HEMnet可以实现到不同的数据集，包括那些具有高水平技术变化的数据集。遗憾的是，技术变异是由组织切片、安装、染色和成像过程引入的。

很少有研究调查其内在的技术变化，如对比度、亮度或信噪比⁷。与大多数方法不同的是，HEMnet实现了一个优化的预处理管道，允许去除图像之间的技术差异。HEMnet包括彻底执行背景校正、标准化、对齐、配准和标签转移的功能。在归一化之前，进行了光度标准化，以校正图像的亮度。我们比较了三种归一化方法，瓦哈达内³⁵莱因哈德³⁶和

马申科³⁷，并认为他们的表现更好。

Vahadane方法(设置为默认值)。图像配准实现了一种具有互信息最大化的概率方法。我们比较了多个选项，发现基于强度的配准，以及Affine的顺序组合和b样条配准³⁸，使用基于梯度下降的优化器来最小化互信息损失，注册H&E图像数据表现良好。我们还评估了计算和运行时间，因为配准是一个密集的过程。缩小规模是一种切实可行的解决方案。最后，为了对注册图像进行标记，我们开发了一个瓷砖级阈值策略，为 224×224 px的每个瓷砖区分癌症、非癌症和不确定标签。带有阈值、分类和过滤步骤的瓷砖标记允许我们为神经网络创建一个高质量的训练(和评估)数据集，最大限度地减少配准错误和不确定标记的技术噪声。总的来说，在HEMnet中实施的标签转移解决方案代表了一个重大的技术进步，是越来越重要的数字组织病理学分析所必需的。标签转移对模型训练产生了三个关键的有益影响。首先，像素级的标签允许我们将一幅图像分割成数百到数千个更小的、高分辨率的分子标记瓷砖，从而增加了模型的样本大小。

培训和测试。这使得开发只有少量幻灯片的精确模型，不像现有的方法需要数千张 ws_i ^{7,33}。一般来说，WSI的平铺为训练神经网络产生了大量的数据，因此能够克服图像分配中的差距。HEMnet成功鉴定了一些非p53染色的细胞为癌细胞，这证明了这一特征。⁴⁾通过像素级标记，癌细胞的分类的分辨率比病理学家的宏观图像高数百到数千倍。此外，分子标记是自动化的，使输出不那么依赖于由训练有素的病理学家进行的费力、手工和可变的注释。

HEMnet，与它的标签转移方法，可以有利于广泛的应用。当处理原始学习过程中未使用的独立验证集时，HEMnet预测了通过病理注释描述的相同的重叠区域（ROC AUC = 0.84）。我们通过系统地比较HEMnet与其他方法和地面真实病理注释来验证HEMnet。我们发现使用TCGA数据集与其他独立方法高度相关的结果（在预测癌症纯度=0.8方面相关性很强）³⁹。

我们选择了p53染色，一种已确定的癌细胞标记物，来开发HEMnet标记转移，正如我们所预期的那样，这个研究充分的问题允许我们评估我们的算法的性能。在非癌细胞中，p53蛋白通常无法被IHC检测到²²，而高达74%的结直肠癌细胞中p53呈棕色阳性^{19,20,23}。推广到其他类型的标记物和癌症，例如乳腺癌的HER2，可以进一步验证。研究了SOX10染色患者通过深度神经网络将H&E图像与IHC图像进行关联性的可行性⁴⁰和荧光癌症标记物图像，如全细胞角蛋白（panCK），或 α -平滑肌肌动蛋白（ α -SMA）⁴¹。

HEMnet是使用p53 IHC染色作为一种合适的结肠直肠癌标记物开发的，在70%-80%的结肠癌中表达¹⁹。我们期望HEMnet标记转移和阈值方法的待因酸性阳性癌症标记可以推广到其他癌症类型和免疫组化标记物。HEMnet可以很容易地用于新数据的培训——分析框架考虑了上面讨论的技术变化和可伸缩性。我们接受了对TCGA数据集鲁棒性性能的测试。标签转移管道可以扩展到许多其他应用程序，以整合来自邻近组织切片的成像数据。我们通过交互式谷歌协作工作区，使HEMnet成为一个易于适应大多数用户的工具，它允许用户上传他们的数据，并使用我们预先训练好的模型进行神经网络预测。

总之，HEMnet是目前一种独特的分子建模方法，它同时利用H&E和IHC图像来定量分类组织切片内的癌细胞。我们希望HEMnet有潜力作为一种计算机辅助工具，通过帮助病理学家提示重要的区域，如组织中的癌症部分^{29,42}。HEMnet不需要人类的病理注释，可以自动在像素分辨率下标记图像。HEMnet等软件的应用可以通过前所未有的分辨率、效率、再现性、准确性、速度、降低成本和增加病理服务的获取来支持癌症诊断。在一个老龄化的社会，有更多的活性和缺乏专业的解剖病理学家⁴³，这种计算创新越来越重要。我们相信HEMnet可以进一步加速计算病理学的应用和集成到病理工作常规中，协助疾病诊断，最终消除漏诊，改善患者预后。我们提供HEMnet作为一个开源软件，也作为一个可访问的基于云的工具，允许用户在不需要进一步编程的情况下分析他们的图像。

方法

H&E和IHC图像数据集的生成

我们从斯坦福医院收集了30名患者的癌症组织样本。所有患者均根据斯坦福大学医学院机构审查委员会（IRB11886）批准的研究方案入组。所有参与者均提供了参与研究的书面知情同意书。组织来自斯坦福癌症研究所组织库。此外，我们从15名患者中获得了匹配的正常、非癌症组织。每个样本被福尔马林混合和副素包埋（FFPE）作为组织块，从每个块取两个相邻的切片，确保这些切片接近相同。一个切片用H&E染色，另一个切片用斯坦福医学的D0-7单克隆抗体（罗氏，Cat#790-291212，预稀释）对p53进行IHC染色。所有的数字幻灯片图像都是由加州大学洛杉矶分校的转化病理学核心实验室以Aperio SVS格式生成的。这项研究是根据《赫尔辛基宣言》进行的。每个组织切片在 $\times 20$ 放大下扫描，共生成35对p53和H&E高分辨率WSIs。

训练、验证和测试数据集的生成

我们在机器学习中使用了一种常见的做法，即将我们的 ws_i 数据集分割成训练、验证和测试集。这些数据集之间不存在重叠，以确保测试数据和验证数据是完全独立。我们将正常的WSI对和癌症的WSI对分配到训练数据集。为了确保准确的训练数据集，我们还确认了病理学家在这些幻灯片中发现的大多数p53染色区域都是癌症。总之，这为该模型提供了区分癌症和非癌症组织的最佳学习程度（补充图。1a）。 ws_i 是以10亿像素的尺度捕获的（补充图。1b）允许我们使用一种平铺策略，将每个WSI分割成数千个更小的 224×224 px图像平铺，用于神经网络训练。我们留出了癌症WSI对作为验证数据集来优化我们的模型的超参数。剩下的17个癌症WSIs被分配到一个独立的测试数据集，以评估我们的模型在看不见的幻灯片上的表现。

H&E染色颜色归一化

由于不同的免疫组化试剂、治疗方案和载玻片扫描仪，在H&E染色和成像中会出现不理想的颜色变化³⁵。因此，组织中相同的细胞结构可能会根据组织染色和成像的方式而出现不同。为了确保我们的模型推广到不同设施的H&E幻灯片的图像，我们修正了染色和成像过程中的技术变化。首先，我们对成像亮度进行了校正，并通过光度标准化确保幻灯片背景为白色（补充图。2）。接下来，我们使用Vahadane等人将每个H&E WSI归一化为来自模板WSI的参考染色色剂³⁵。在染色工具中实现的染色标准化方法⁴⁴，Eq. (1)。

$od_{公寓} \% C * S$

(1)

OD_{1at}是由RGB WSI衍生的光密度（OD）阵列。染色基质（S）编码H&E染色的染色颜色，并使用Vahadane方法进行估计。该染色矩阵用于识别像素染色浓度矩阵（C）。为了将源WSI归一化为模板WSI，根据公式计算两幅图像的染色和浓度矩阵。(2)和(3)。

$od_{根源} \% C_{根源} * S_{根源} (2)$

$od_{样板} \% C_{样板} * S_{样板} (3)$

C_{根源}基质描述了在每个像素处的苏木精和伊红染色的浓度。使用模板图像中的染色矩阵（S_{样板}），我们对源浓度矩阵中的每个像素进行着色，以生成一个图像（Eq. (4)），就像源图像以与模板图像相同的方式被染色和捕获一样。

$od_{范数} \% C_{根源} * S_{样板} (4)$

通过将所有的 ws_i 归一化到模板图像上，我们确保了相似的细胞结构具有相似的外观，无论它们如何染色和进行图像扫描。

为了选择一个合适的模板WSI，我们将平均R、G、B通道强度最接近不同通道平均值的中位数的癌症载玻片进行分类

所有图像的通道 (R、G和B) 强度(补充图。1c)。此外，我们还实现了由Reinhard等人提出的两种用户可选择的、流行但不那么先进的图像归一化方法。³⁶和Macko等人。³⁷

IHC图像与H&E图像的配准

对于用于精确标记H&E图像的IHC图像，每个IHC图像与对应的H&E图像进行对齐。尽管来自同一组织块的相邻切片，但在切片、安装和成像方面的技术差异导致了IHC图像与H&E对应图像之间的错位。我们通过SimpleITK包实现图像配准来对齐这些图像³⁸。

在配准过程中，IHC图像被扭曲，使其与H&E图像对齐。通过只转换IHC图像，我们确保了H&E图像保持不变。H&E图像之间的技术变化，例如由于显微镜曝光时间和/或染色时间引起的亮度或颜色强度的变化，被归一化了(补充图。2和尤花果。2)。因此，在这些H&E图像上训练的神经网络可以应用于新的归一化，但在其他方面未经修改的H&E图像。

我们通过目测和定量的互信息度量来验证准确的配准。我们将配准的p53覆盖在相应的H&E图像上，以通过视觉检查是否正确对齐。此外，我们通过计算配准前、配准前后这些图像之间的互信息，来比较p53图像与H&E图像的对齐情况。互信息是一个信息论概念，可用于测量图像配准性能(补充图。3)。配准后互信息的增加表明了更好的图像对齐。IHC和H&E图像之间的互信息可以用等式计算出来(5)。

$$I(IHC; H\&E) = \sum_{ihc,h\&e} p(ihc; h\&e) \log \frac{p(ihc, h\&e)}{p(ihc)p(h\&e)} \quad (5)$$

其中， $p(ihc)$ 和 $p(h\&e)$ 分别为IHC和H&E图像中灰度像素强度的边际概率分布。 $p(ihc, h\&e)$ 是图像的灰度像素强度的联合分布。

配准策略可以广泛地分为基于特征的方法和基于强度的方法。基于特征的方法提取特征(e.g., 并对源图像进行变换，使源图像中的特征与目标图像中的匹配特征处于相同的位置。另一方面，基于强度的方法考虑了像素的强度或强度分布。这些方法还对源图像进行变换，使其与目标图像的像素强度或强度分布最密切相关。在初步测试中，我们发现基于强度的方法对H&E图像是有效的。

对于我们的基于强度的配准方法，我们选择了一个互信息代价函数来量化源图像和目标图像的配准程度。该代价函数度量源图像和目标图像的像素强度分布之间的互信息。配准的目标是对源图像进行变换，从而使源图像和目标图像之间的互信息最大化——这将意味着一个配准良好的图像。互信息是由灰度像素强度计算出来的，因此首先将IHC和h&e染色的图像转换为灰度。配准后，对RGB IHC图像的每个通道进行灰度IHC图像的最优变换，生成配准彩色图像。

为了实现准确的配准并达到全局最优，而不是局部最优，我们进行了精确配准，然后进行了b-样条配准。初始的线性仿形配准仅限于平移、尺度、剪切和旋转变换，而随后的b样条配准则是一个非线性变换。最初的基本步骤确保了在b样条注册iner细胞特征之前，图像中的大型架构特征被注册。仿射变换和b样条变换都通过基于梯度下降的优化器进行调整，以最小化互信息代价函数。

每个afine和b-样条配准步骤都包含了一个多分辨率的方法。这里的概念是相似的：通过在小特性之前注册大特性来实现更好的注册。在图像和b样条步骤的开始，一个低分辨率的图像被用来鼓励图像中的大特征的配准。逐渐使用更高和更高的分辨率来注册每一个so iner特征，直到达到所需的最终分辨率。由于配准是一个计算密集型的过程，特别是对于十亿像素的wsi，我们注册了更小版本的IHC和H&E图像，它们被缩小了5倍

因素是用户可调的。配准的主要输出是彩色5倍缩小的IHC图像精确配准到相应的相同大小的H&E图像。由于H&E图像可能捕获了与IHC图像不同的视图，任何以外的图像像素在IHC图像用白色填充。

基于p53染色的图像自动标记

配准将p53图像转换为与相应的H&E图像相同的坐标系。因此，对齐的p53图像中的每个像素都指向对应的H&E图像上相同位置的一个像素。这种对齐对于p53染色准确标记H&E图像至关重要。

为了将每个像素标记为与癌症和正常组织重叠的像素，我们对p53图像应用阈值分割。这个过程决定了哪些像素是正的(癌症的)或负的(正常的)染色。p53 IHC通过DAB(3,3'-二氨基联苯胺)沉积在组织上观察p53 IHC染色，阳性染色组织呈棕色。我们通过将RGB图像反卷积为苏木精、伊红和DAB通道，将DAB阳性像素和p53阳性像素与图像的其他部分区分出来。这个过程是基于瑞弗罗克和约翰斯顿开发的一种方法⁴⁵。通过这种方式，我们可以将阈值集中在DAB染色上，它可以改变每个像素上的p53蛋白水平。

我们观察到DAB通道内的像素可分为三类：p53阳性像素；微弱的组织背景染色；我们解释为p53阴性染色；幻灯片背景的像素，没有组织和p53染色。为了将其简化为两类阈值处理问题，我们使用苏木精通道将组织从幻灯片背景中分离出来——我们对DAB通道的组织区域应用了单独的阈值处理。在这两种情况下，我们使用Otsu阈值，使两类之间的类间方差最大化。通过用苏木精通道分割该组织，我们通过其低的，但明显高于幻灯片背景的染色水平来区分该组织。此外，它确保了我们保留了高水平苏木精的核，是p53蛋白的位置。在组织阈值分割后，我们将Otsu阈值仅应用于DAB通道的组织区域，并将每个像素分为两类：p53阳性的高强度像素；p53阴性的低强度背景染色像素。这一过程被自动且独立地应用于每张p53玻片，因此不会因为p53玻片之间染色的细微差异而发生像素误分类。

我们将每个H&E图像分成224张×224px瓷砖，用于模型训练和测试。随后，我们将p53像素级分类转化为倾斜癌/正常分类。配准的p53图像被5倍向下采样以方便配准，我们在这张图像上确定了像素和贴图标签，因为它与H&E对齐。因此，我们分析并标记了45倍px×45px的5倍下来样块，与原始图像的视图范围相同。这些瓷砖包含多个细胞——在肿瘤的组织区域，不是所有这些细胞都会是癌症。为了确保我们没有遗漏癌细胞，同时最小化错误染色的水平，如果瓷砖超过2%的像素是p53阳性，我们标记了一个瓷砖癌症。剩下的组织瓷砖被标记为正常或“非癌症”。

其他策略，以确保准确的瓷砖标签

病理回顾提供了这些组织的癌症与正常细胞状态。3个样本p53染色阳性，尽管没有肿瘤细胞的组织病理学指征，这将导致不准确的标记和模型错误分类。为了保证准确的模型训练和测试，这些样本中的p53和H&E WSIs被排除在分析中。总的来说，总共剩下32对H&E和p53 WSIs，27对癌症和live正常组织。

在某些情况下，p53染色还不够清晰，不足以提供指示性的标签，所以我们将不明确的瓷砖标记为不确定的，并丢弃它们。这些模糊的瓷砖可能会给训练数据增加噪声，并妨碍对模型性能的准确评估。我们通过设置一个用户可选择的DAB强度阈值来解决这个问题，以便将瓷砖标记为不确定。这些阈值应用于每个瓷砖的平均DAB强度。落在这些阈值之间的瓷砖被标记为不确定的，不用于训练或测试模型。剩余的癌症和非肿瘤贴标签从注册的p53图像转移到用于模型训练的H&E贴标签。

为了防止任何注册错误和确保准确的标签转移，如果一对p53/H&E瓷砖只有一个包含组织的瓷砖，则该H&E瓷砖将被丢弃。为了评估一个瓷砖，我们使用Rother等人的GrabCut算法从p53和H&E图像中的背景中分割组织。⁴⁶此外，为了确保一个干净的训练数据集，我们只使用了来自癌症样本的癌症阳性瓷砖，而只使用了来自非癌症样本的癌症阴性瓷砖。

训练一个卷积神经网络（CNN）

我们用10个224×WSIs的224×224px瓷砖训练模型。由于我们的平铺策略，我们可以从每个WSI中生成数千个样本，我们将它们汇集在一起训练模型。我们使用迁移学习开发了一个基于vgg16的CNN，用于将瓷砖分类为癌症或非癌症。我们的模型使用了VGG16架构，并对来自ImageNet的130万张图像进行了预训练⁴⁷，用于特征提取。HEMnet在图像训练过程中有多种选项来实现CNN模型，包括ResNet50、VGG16、VGG19、InceptionV3和Xception。我们比较了这些模型，发现了相似的性能，VGG16的运行速度略快，并产生了更高的精度（补充表1）。事实上，我们的HEMnet-VGG16模型比原始的VGG16模型少得多（>1000倍）（补充图。4），因为我们只使用了VGG16特征提取器和迁移学习方法，其中CNN基础模型中的参数没有被训练。此外，利用该预训练模型输出的最大池化层（1,1,512）作为输入，训练256个神经元的全连接层，输出一个具有类概率的s型神经元作为TP53二进制标签。通过在大量图像上使用预训练的权重，我们可以训练我们的模型成为一个相对较小的数据集，并且仍然可以在不被遗漏的情况下实现准确的预测。每个224×224px瓷砖的特征被输入一个完全连接的神经网络，以预测瓷砖的癌症状态。

完整的CNN是在10个训练WSI在×10放大下生成的10个瓷砖，持续100个时代。我们采用数据增强的方法来克服过拟合问题，提高了模型的通用性。由于给定的组织中肿瘤细胞内滤的程度无论视角或方向如何都保持不变，我们随机旋转和翻转瓷砖。在验证集上表现最好的超参数被用于训练在本工作中用于所有不可见的幻灯片测试的模型。我们使用张力洛作为深度学习框架，用Python实现了这个系统。

绩效评估

我们在H&E测试玻片上测试了我们的模型，评估了其p53染色模式和病理学家注释相比的性能。我们通过计算精度、混淆矩阵和接收机工作曲线（ROC）来测量模型的性能。为了评估针对p53注释的性能，我们使用描述为训练数据集的相同方法生成了一个测试数据集。考虑到这些切片上有细胞混合物，我们生成了仅代表癌症和正常组织的瓷砖。对于17张幻灯片中的13张，我们获得了WSIs上的病理学家癌症注释图。我们提取癌症注释包含的注释和标记贴标记为癌症，并将剩余的组织贴标记为非癌症（补充图。5）。

主要的性能指标是准确性和ROC AUC。这些数据是通过比较p53和病理学家测试数据集瓷砖标签与我们的模型预测的标签来计算的（图4、5和补充图。6）。由于癌症和非癌症贴片在这些数据集中并不均匀分布，我们通过对主导类进行子抽样来平衡每个类的贴片数量。

TCGA验证

我们用H&E图像对24例结直肠癌患者进行了验证，我们的模型。WSIs来自于TCIA，匹配的基因组数据来源于癌症基因组图谱（TCGA）。TCIA和TCGA分别是癌症医学成像数据（包括数字组织病理学数据）和癌症基因组数据的公共存储库。我们使用我们的模型预测来估计肿瘤纯度，并将其与从基因组测序研究获得的肿瘤纯度估计值进行了比较。对于这种基于图像的分析，我们通过将每个瓷砖内的组织面积加权来计算癌症组织面积占总组织面积的比例。这比使用癌症瓷砖与所有瓷砖的比例作为一些瓷砖更准确，特别是在组织的边缘。例如，一个只有一半背景和一半组织的瓷砖只会

贡献一半瓷砖的面积。我们将我们的估计与五种确定肿瘤纯度的方法进行了比较。这个比较包括了绝对的程序⁴⁸，扩展⁴⁹，估计⁵⁰，CPE⁵¹，IniniumPurify⁵²，和LUMP（白细胞未甲基化）（补充图。7）。

报告总结

关于研究设计的进一步信息可在链接到本文的《自然》研究报告摘要》中获得。

数据可用性

在当前的研究中使用和/或分析的数据集（所有高分辨率的H&E和TP53图像）可以从<https://dna发现中心免费获得>。斯坦福.edu/research/web-resources/HEMnet. 用于与HEMnet比较的绝对、估计、CPE、ininum-纯化、LUMP的结果来自<https://doi.org/10.1038/ncomms9971>提供的“补充数据1”。

代码可用性

源代码、教程和交互式分析工具可以在<https://github.com/BiomedicalMachineLearning/HEMnet>上找到。我们还提供了基于云的HEMnet实现（补充图。10），有谷歌Colab笔记本和一个ImJoy应用程序（这些应用程序的链接在HEMnet github页面上）。HEMnet也可以作为一个开源的PyPI python包使用（<https://pypi.org/project/hemnet>）。使用了HEMnet软件1.0.0版本，软件依赖项的版本信息列在环境中的HEMnetgitheeb站点中。ymile。我们使用SimpleITK版本1.2.3进行图像配准，使用Staintools版本2.1.2进行归一化。使用扩展2.0.0，使用默认设置估计肿瘤纯度。

收到日期：2021年3月27日；接受日期：2021年12月16日；

Published online: 02 March 2022

参考文献

1. Griffin, J. & Treanor, D. 临床应用中的数字病理学：我们现在在哪里，是什么阻碍了我们？组织病理学70、134–145（2017）。
2. 罗德尼格斯-卡纳莱斯, J., 埃伯利, F. C., 雅菲, E. S. 和 Emmert-Buck M. R. 为什么将病理学重新融入癌症研究中至关重要？Bioessays 33, 490–498（2011）。
3. Rosai, J. 为什么显微镜检查仍然是外科病理学的基石。实验室投资。87, 403–408（2007）。
4. Raab, S. S. 以及其他人在癌症诊断中解剖病理错误的临床影响和频率。癌症104, 2205–2213（2005）。
5. 阿斯瓦西, M. A. & Jagannath, M. 用数字组织病理学图像检测乳腺癌：现状和未来的可能性。通知。医学解锁874–79（2017）。
6. 范德拉克, 利金斯, G. & Ciompi, F. 组织病理学中的深度学习：通往诊所的途径。Nat. 医学27, 775–784（2021）。
7. 胡, Z. 以及其他深度学习用于基于图像的癌症检测和诊断-A项调查。模式识别。83, 134–149（2018）。
8. 贝拉, K., 沙尔珀, K. A., Rimm, D. L., Velcheti, V. 和马达布希。数字病理学的人工智能-诊断和精确肿瘤学的新工具。Nat. 发动机的旋转Clin. Oncol. 16, 703–715（2019）。
9. Acs, B. & Rimm, D. L. 不仅仅是数字病理，智能数字病理。JAMA协议。4, 403–404（2018）。
10. Huss, R. & Coupland, S. E. 数字组织病理学中的软件辅助决策支持。J. 帕索尔。250, 685–692（2020）。
11. 鲁, M. Y. 以及其他基于人工智能的病理学预测了未知原发性癌症的起源。《自然》杂志594、106–110页（2021年）。
12. Thakur, N., Yoon, H. & Chong Y. 结直肠癌病理图像分析的人工智能发展趋势：系统综述。癌症（巴塞尔）12, <https://doi.org/10.3390/cancers12071884>（2020）。
13. 埃尔摩, J. G. 以及其他病理学家对侵袭性黑色素瘤和黑素细胞增生的诊断：观察者的准确性和可重复性研究。BMJ 358, j3798（2017）。
14. Swiderska-Chadaj, Z. 以及其他学习通过深度学习通过免疫组化来检测淋巴细胞。医学图像肛门。58, 101547（2019）。
15. 坎帕内拉, G. 以及其他临床级计算病理学使用弱监督深度学习的整个幻灯片图像。Nat. 医学25, 1301–1309（2019）。
16. Magaki, S., Hojat, S. A., 魏, B., 如此 A. & Yong, W. H. 免疫组化性能的介绍。方法莫尔。比奥尔。1897, 289–298（2019）。

17. Himmell, E. 以及其他人在除了H&E: 对原位组织生物标志物成像的先进技术。ILAR J. 59, 51–65 (2018).
18. 费舍尔, H. 肿瘤生物学的进化: 寻求在基因表达促进和形态学研究之间的平衡。J. 摩尔诊断. 4, 65 (2002).
19. Kaserer, 以及其他人在结直肠癌中p53免疫组化的染色模式及其生物学意义。J. 帕索尔. 190, 450–456 (2000).
20. 中山, M. 和大岛, M. 结肠癌中的突变体p53. J. 摩尔细胞生物体. 11, 267–276 (2018).
21. Finlay, C. A. 以及其他人在激活p53转化的突变产生一个基因产物, 形成一个半衰期发生改变的hsc70-p53复合物。摩尔细胞生物体. 8, 531–539 (1988).
22. 宋楚瑜, R. 以及其他人在一系列常见的人类癌中, p53蛋白过表达与基因突变之间的一致性。哼唱帕索尔. 27, 1050–1055 (1996).
23. Murnyak, B. & 霍托巴吉, T. 癌症中TP53体细胞突变的免疫组化相关性。本体目标 764910–64920 (2016年)。
24. 克拉克, K. 以及其他人在癌症成像档案 (TCIA): 维护和操作一个公共信息存储库。J. 数字. 影像学检查26, 1045–1057 (2013年)。
25. 之前, F. 以及其他人在开放存取图像存储库: 高质量的数据, 使机器学习研究。Clin. 无线广播75, 7–12 (2020)。
26. 库马尔, 阿巴斯, A. K., 阿斯特, J. C. & 帕金斯, J. A. 罗宾斯基本病理学。(爱思唯尔, 2017)。
27. Andrysiak, Z. 以及其他人在识别一个具有高度分布的肿瘤抑制活性的核心TP53转录程序。基因组re. 27, 1645–1657 (2017)。
28. Rosai, J. 病理学: 一个历史上的机遇。是J. 帕索尔. 151, 3–6 (1997)。
29. Topol, E. J. 高性能医学: 人类智能与人工智能的融合。Nat. 医学25, 44–56 (2019)。
30. 埃文斯, J. 以及其他人在美国食品和药物管理局批准全幻灯片成像用于初步诊断: 达到了一个关键的里程碑, 并提出了新的问题。拱门。帕索尔. 实验室医学142, 1383–1387 (2018)。
31. 托马斯, M., 左派, J. G., 巴克斯特, G. 和汉密尔顿, N. A. 可解释的用于非黑色素瘤皮肤癌的多类分割和分类的深度系统。医学图像分析, 101915, <https://doi.org/10.1016/j.media.2020.101915> (2020)。
32. 埃尔摩, J. G. 以及其他人在病理学家对侵袭性黑色素瘤和黑素细胞增生的诊断: 观察者的准确性和可重复性研究。BMJ 357, j2813 (2017)。
33. 埃斯特瓦。以及其他人在利用深度神经网络进行皮肤癌的皮肤科医生级别的分类。自然542, 115 (2017)。
34. 傅, Y. 以及其他人在泛癌计算组织病理学显示突变, 肿瘤组成和预后。Nat. 癌症 1800–810 (2020年)。
35. 瓦哈丹。以及其他人在组织学图像的结构彩色归一化和稀疏染色分离。IEEE跨. 医学成像35, 1962–1971 (2016)。
36. 莱因哈德, E., 阿迪赫明, M., 古奇, B. & 雪莉, P. 图像之间的颜色转移。IEEE 计算机图. 应用程序. 21, 34–41 (2001)。
37. Macenko, M. 以及其他人在一种组织切片的归一化方法。2009年, IEEE第6届IEEE Int. 模拟。比奥梅德。成像, 1107–1110 (2009)。
38. Lowekamp, B. C., 陈, D. T., Ibanez, L. & Blezek, D. 简单ITK的设计。前面 Neuroinform 7, 45 (2013)。
39. 安多, N. 以及其他人在肿瘤内异质性的范围和后果的泛癌分析。Nat. 医学22, 105–113 (2016)。
40. 杰克逊, C. R., Sriharan, & Vaickus, L. J. 一种模拟免疫组化的机器学习算法: SOX10虚拟IHC的开发和对主要的黑素细胞肿瘤的评估。模块。帕索尔. 33, 1638–1648 (2020)。
41. 伯林盖姆, E. A., 马戈林 A. A., 灰色, J. W. 和张, Y. H. 转移: 使用条件生成对抗网络对整个幻灯片图像进行快速的组织病理学到免疫荧光翻译。过程SPIE国际。社会选择雕刻10581, <https://doi.org/10.1117/12.2293249> (2018)。
42. 郑, J. Z. 以及其他人在具有深度学习架构的计算机辅助诊断: 在美国图像中的乳腺病变和CT扫描中的肺结节中的应用。科学。校纹平布6, 24454 (2016)。
43. 威尔逊, M. L. 以及其他人在获得病理学和实验室医学服务: 一个关键的差距。《柳叶刀》第391页, 1927–1938年 (2018年)。
44. StainTools v. v2.1.3 (Zenodo, 2019)。
45. 瑞弗洛克, A. C. 和约翰斯顿, D. A. 用颜色反褶积法定量分析组织化学染色。肛门。细胞质。希斯托尔. 23, 291–299 (2001)。
46. 罗瑟, 科尔莫戈罗夫, V. 和布拉克。在ACM签名图2004论文309–314 (计算机协会, 洛杉矶, 加利福尼亚, 2004)。
47. 俄勒冈州罗萨科夫斯基。以及其他人大规模视觉识别的挑战。Int. J. 计算机Vis. 115, 211–252 (2015)。
48. 卡特, S. L. 以及其他人在人类癌症中体细胞DNA改变的绝对定量。Nat. 生物技术. 30, 413–421 (2012)。
49. 安多, N., 哈里斯, J. V., Muller, S., Mewes, H. W. 和彼得里奇, C. 扩展: 在嵌套的亚群体上扩展倍性和等位基因频率。生物信息学30, 50–60 (2014)。
50. 吉原, K. 以及其他人在从表达数据推断肿瘤纯度、间质和免疫细胞混合物。Nat. 通勤. 4, 2612 (2013)。
51. 阿兰, D., 西罗塔, M. 和Butte, J. 肿瘤纯度的系统泛癌分析。Nat. 通勤. 6, 8971 (2015)。
52. 郑十, 张, 吴, 何。和吴, H. 在癌症研究的DNA甲基化数据分析中估计和解释肿瘤纯度。基因组生物体. 18, 17 (2017)。

致谢

我们感谢阮的基因组学和机器学习实验室的所有成员和智研究小组的有益讨论。我们感谢Aykan Ozturk在H&E和IHC之间对齐方面的初步工作。这项工作得到了以下资助的支持: 澳大利亚研究委员会 (ARC DECRA DE190100116); 国家卫生研究委员会 (APP2001514); 昆士兰大学早期职业研究奖; 基因组创新中心外部项目拨款; 美国国立卫生研究院资助R01HG006137和U01CA217875。对H的额外支持。P. J. 和HJ. L. 来自克塞尔基金会。

作者贡献

Q.N., A.S., N. A., HJ.L., X.T., 和H.P.J. 构思了实验并开发了算法。A.S., X.T. 编写了软件。A.S., HJ.L., X.T., 和Q.N. 进行实验并分析数据。A.S., HJ.L., Q.N. 和H.P.J. 写的手稿。A.S., HJ.L. 和X.T. 对这项工作的贡献相同。所有作者均已对稿件进行了审核和审稿。

竞争利益

作者声明没有任何相互竞争的利益。


附加信息

补充信息在线版本包含补充材料, 可在<https://doi.org/10.1038/s41698-022-00252-0>获得。

材料的信件和要求应寄给阮或韩礼。吉。

重印和许可信息可在<http://www.nature.com/reprints>上获得。自然com/重印

出版商的说明《施普林格自然》对已出版的地图和机构机构中的管辖权主张保持中立。

 本文是在知识共享协议下获得授权的

4.0国际许可证, 允许使用、共享、以任何媒介或格式进行改编、分发和复制, 只要您给予原作者和资料来源适当的荣誉, 提供与知识共享许可相关的链接, 并表明是否进行了更改。本文中的图片或其他第三方材料都包含在文章的知识共享许可中, 除非在材料的信用额度中另有说明。如果材料没有包含在文章的知识共享许可中, 并且您的预期使用不被法律法规允许或超过了允许的使用, 您可能需要直接获得版权所有者的许可。要查看此许可证的副本, 请访问<http://creativecommons.org/licenses/by/4.0/>。

©作者(s) 2022