
全球最大结直肠癌多组学研究：如何让疾病分类更准确

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/29353.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

全球最大结直肠癌多组学研究：如何让疾病分类更准确

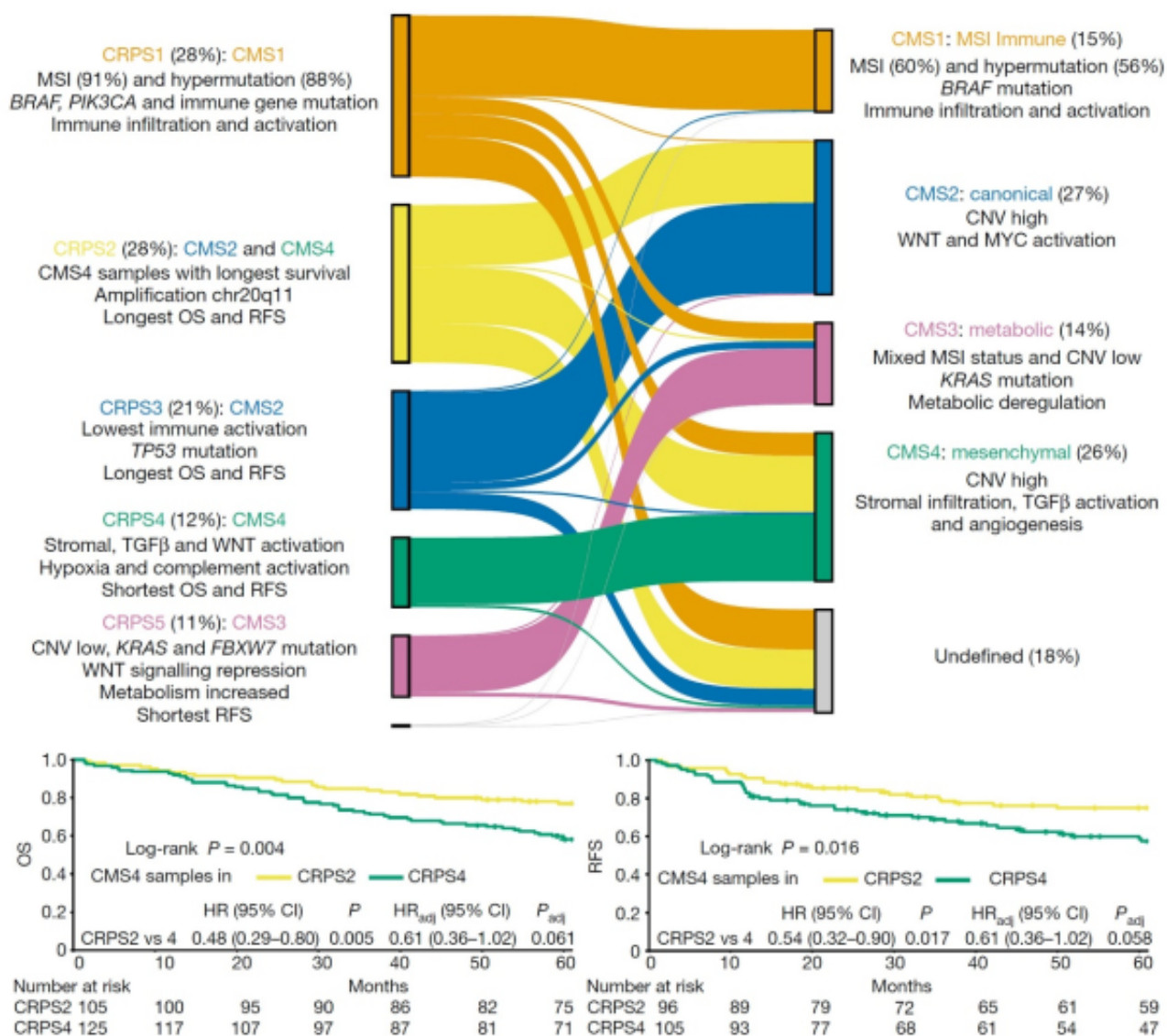
。· “在AI多组学模型建立的过程中通过深度学习可以判断出一个符合预期的结果，但这个结果具体是由什么机制推导出来的AI并不能给出特别具体的答案，也就是所谓的‘黑箱’，如果没有科学性的预测依据，可能结果不能完全令人信服。”

准确的疾病分类能够帮助医生对病人的病情、治疗手段以及预后（即预估治疗结果）进行快速和有效的判断。疾病分类的基础是致病机制，而在以前，分类往往是通过症状、病灶形态等表面现象进行的。现在，随着组学技术的发展，科学家能够深入基因，追本溯源，对疾病进行更为精准的分类，为诊疗作出贡献。

近日，一项研究通过全球最大规模的结直肠癌多组学数据分析，发现了多个与该疾病相关的关键基因，并在此基础上重构了结直肠癌类型学。该研究于2024年8月7日发表在《自然》（Nature）杂志上，作者是来自瑞典乌普萨拉大学、华大生命科学研究院以及华大基因智慧医学研究院的Tobias Sj?blom、林从、李甫强等多名研究者组成的团队。

结直肠癌发病率居所有癌症第三位，约20%的患者在确诊时已经发生转移，死亡率高居各类癌症第二位。该研究对瑞典知名癌症样本数据库U-CAN队列中的1063例结直肠癌样本进行全基因组和转录组测序分析，发现了一系列与癌症不同阶段相关的驱动基因，并识别了关键预后因子。

基于肿瘤基因表达差异，研究者们将结直肠癌预后分为了包含5个预后亚型的类型系统CRPS（CRC Prognostic Subtypes，结直肠癌预后分型）。与经典的结直肠癌分型CMS（Consensus Molecular Subtypes，共识分子分型）相比，该分类能够更加准确地预测预后。



通过对多组学数据的分析，研究团队对经典结直肠癌分类CMS（图片右侧）进行重构。新分型系统CRPS（图片左侧）对预后反应的预测效果更好。图片来源：《自然》杂志

基因与癌症是什么关系？什么是多组学数据？基因组与转录组有什么区别？AI能够帮助我们更好地对疾病进行分类吗？为了回答这些问题，澎湃科技采访了该研究共同通讯作者、华大基因智慧医学研究院专项科学家林从。

【对话】

基因：癌症的“密码”

澎湃科技：能否请你介绍一下结直肠癌的特点？该研究为什么选择这种癌症？

林从（论文共同通讯作者、华大基因智慧医学研究院专项科学家）：结直肠癌是全球第三大高发癌症，每年新增病例数约190万例，死亡人数接近90万，严重威胁人们的身体健康。尤其在发达国家，发病率和死亡率都较高。近年来，随着生活方式和饮食习惯的改变，许多发展中国家的发病率也在上升。

结直肠癌的病程相对缓慢，但发病隐匿，虽然有筛查和早期干预的机会但发现率有待提高。同时，结直肠癌的发生涉及遗传和环境等多种因素，肿瘤具有很高的异质性，大规模的研究可以帮助优化筛查方法，改善早期发现率，并帮助理解癌症发生的机制，在临床上也有助于研究新型治疗方案、标志物或药物靶点，可以改善患者生存率。

本研究使用的UCAN队列，是Uppsala University联合Umeå University、Stockholm university、Kungliga Tekniska Högskolan (KTH) 等研究团队于2010年启动建设的，截至2023年底，已经采集了27,476名癌症患者的样本、影像及临床数据，该队列对患者每年定期随访，近20%的患者随访超过10年，75%以上的患者随访超过5年。其中结直肠癌患者占比很高，为大规模研究提供了很好的基础。

澎湃科技：基因与癌症的关系是怎样的？前者决定后者的产生和发展吗？在很多人眼中，基因是先天决定的，因此癌症的风险也是先天决定的吗？

林从：基因与癌症的关系非常密切，癌症的产生发展通常涉及基因的突变、异常表达或调控失常。虽然基因确实在癌症风险中起到重要作用，但不是所有癌症的风险都是先天决定的。大多数癌症（约90-95%）与后天环境因素和生活方式密切相关（如吸烟、辐射暴露、饮食、运动不足、肥胖等），这些因素可能引发基因突变，从而导致癌症。

澎湃科技：该研究发现了多种新的癌症驱动基因。如何理解“驱动基因”？这次发现的这些基因有什么特点？它们是结直肠癌特有的吗？可以根据这些基因来对尚无症状的人进行癌症诊断吗？

林从：驱动基因一般是癌症发生发展过程中高度频发突变的基因，这种突变一般会为癌细胞提供一定生长优势，比如使它们能够不受控地增殖、转移和存活，从而推动肿瘤的形成和进展。

我们发现的这些基因不一定是结直肠癌特异的，而是与癌症普遍有关，但有些驱动基因是在结直肠癌更为常见。理论上，识别这些驱动基因的突变可以帮助早期诊断癌症，特别是在尚无症状的高危人群中，但需要后续在临床应用研究中再进一步严谨地验证，也需要考虑敏感性和特异性等因素。

组学分析：疾病溯源的利器

澎湃科技：该研究对1000多例结直肠癌样本进行了全基因组和转录组测序分析。什么是组学分析？转录组与基因组有什么区别，各能得到什么信息？

林从：组学分析是从不同层面和角度对生物系统中大量分子数据进行研究的方法。一般通过高通量技术（如基因测序、质谱等）得到不同类型的数据，如基因组、转录组、蛋白质组和代谢组等。

基因组是指人体内所有DNA的总和，包括所有的基因以及非编码区域，线粒体基因组等。转录组是指所有RNA分子的集合，转录组分析提供了基因表达的动态信息，揭示了在特定条件下哪些基因可被转录为RNA，以及它们的表达水平如何变化，对于理解基因功能和调控机制非常重要。

澎湃科技：该研究中的基因组学分析具体是如何开展的？文章提到该研究还结合了样本临床随访的数据，大致包括哪些方面的数据？

林从：大致流程就是收集病人样本并提取组织中DNA、RNA，然后利用我们自主研发的测序仪进行高通量测序，并将得到的测序数据与参考基因组进行比对，分析带有突变的基因以及被影响的基因表达等。

临床数据大概包含病人的年龄、性别等基本信息，以及肿瘤相关的信息，例如肿瘤分期、分级等，以及病人生存期的统计。

数据越多，分类越准？

澎湃科技：结直肠癌的分类体系是如何发展的？CRPSs有什么特点和优势？它为结直肠癌诊疗具体带来了哪些帮助？

林从：简单来说，最基础的是传统的形态学分类，根据肿瘤细胞形态和组织结构将结直肠癌分为腺癌、黏液腺癌等等，但这些亚型在预测患者预后和指导治疗方面的作用有限。

随着基因组学技术的发展，带有特定基因的突变，微卫星不稳定性（MSI）和染色体不稳定性（CIN）的肿瘤被逐一鉴别，这些分子特征的发现进一步细化了结直肠癌的分类。2015年，研究者提出了CMS分型（Consensus Molecular Subtypes），将结直肠癌系统的分为四个亚型：CMS1-4，每个亚型具有不同的基因表达特征、预后和治疗反应，这极大地完善了结直肠癌分子分型系统。

而在我们的研究里，进一步基于肿瘤基因表达差异谱解析出5个具有不同分子特征的预后亚型（Colorectal Cancer Prognostic Subtypes, CRPS），相较于CMS，我们利用独立队列转录组数据构建的CRPS能更加准确地预测预后。在经典的CMS分型中，CMS4间质型肿瘤具有基质细胞浸润较高，转化生长因子 β （TGF- β ）通路激活等特征，且被普遍认为是预后较差的肿瘤，但在CRPS分型系统中，部分CMS4型肿瘤实际上被判定为预后较好的CRPS2型，而CRPS2型肿瘤具有上皮细胞特征，且在基因组20q11区域具有较多的扩增，这些结果提示整合基因组和转录组数据的分子分型，能得到更精细准确的患者预后分层，对优化临床肿瘤分型，指导结直肠癌精准治疗具有重要意义。

澎湃科技：有人认为在AI和大数据时代，只要我们整合规模更大、模态更多的数据，就可以对疾病进行更加细致和本质的分类，甚至突破原先对疾病的理解。在你看来这一发展过程中还有哪些挑战？

林从：我认为挑战可能有以下几点：首先，数据质量参差不齐，会对AI产生噪音；且不同研究机构在收集样本/数据的过程中数据采集条件、测量标准、分析方法等方面存在差异，导致数据整合不一致性，影响AI模型泛化能力。

其次，不同模态的数据（如多组学数据、影像数据、临床数据等）有很高异质性，如何有效整合这些异构数据是一大挑战。

最后，在AI模型建立的过程中通过深度学习可以判断出一个符合预期的结果，但这个结果具体是由什么机制推导出来的AI并不能给出特别具体的答案，也就是所谓的“黑箱”，如果没有科学性的预测依据，可能结果不能完全令人信服。

（原标题：全球最大结直肠癌多组学研究：科学如何让疾病分类更准确？）

作者：季敬杰 来源：澎湃新闻

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发