Universiteit Leiden
The Netherlands

利用血小板 RNA 检测和定位早期和晚期癌症

In't Veld, S.G.J.G.; Arkani, M.; Post, E.; Antunes-Ferreira, M.; D'Ambrosi, S.; Vessies, D.C.L.
;
... ; Wurdinger, T.

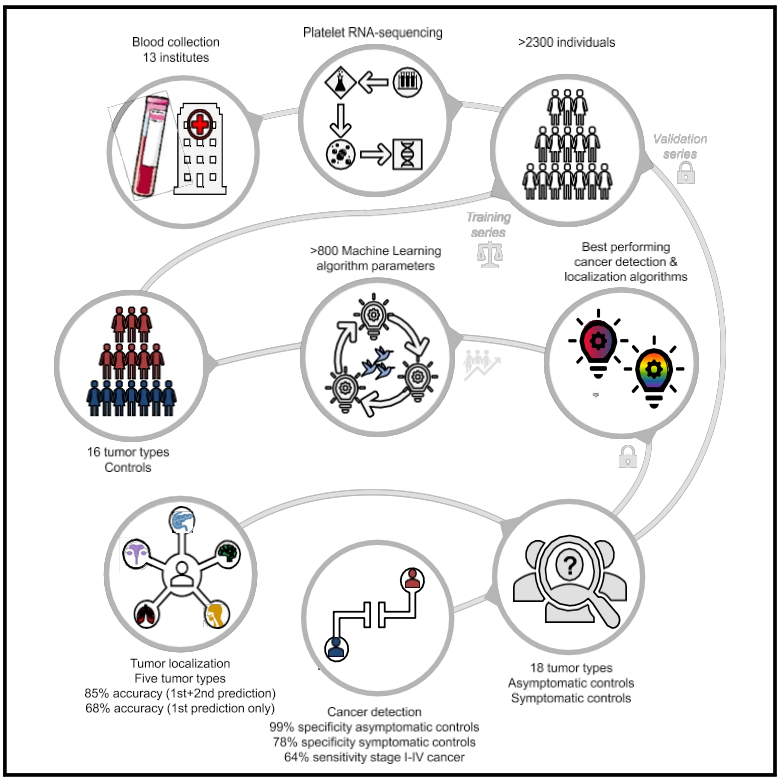
注：引用本出版物时，请使用最终出版版本（如适用）。

Cancer Cell

文章

利用血小板 RNA 检测和定位早期和晚期癌症

图形摘要



亮点

- 通过血小板 RNA 分析确定 18 种肿瘤类型，特异性极高
- 与肿瘤类型相关的血小板 RNA 图谱可用于肿瘤原发部位分析
- 多处肿瘤活动可能对血小板产生影响
- 血小板 RNA 可为液体活检领域提供补充

作者

Sjors G.J.G. In 't Veld、穆罕默德-阿尔卡尼、爱德华-波斯特、……、尼克-索尔、迈伦-G-贝斯特、托马斯-沃丁格

通信 m.g.best@amsterdamumc.nl (M.G.B.), t.wurdinger@amsterdamumc.nl (T.W.)

简而言之

In 't Veld 等人利用血小板 RNA 图谱开发了一种高度特异性的泛癌症血液检测方法，涵盖 18 种不同的肿瘤类型，并能对原发肿瘤进行定位。这项研究强调了血小板在早期癌症检测中的价值，并可作为 "液体活检 "的补充生物资源。

In 't Veld 等人，2022 年，《癌细胞》*40 期*，999-1009 页

September 12, 2022
https://doi.org/10.1016/j.ccell.2022.08.006

CellPress

文章

探测和定位利用血小板 RNA 检测早期和晚期癌症

Sjors G.J.G. In 't Veld,[1,2,3,4,5] Mohammad Arkani,[1,2,3,6,68] Edward Post,[1,2,3,68] Mafalda Antunes-Ferreira,[1,2,3,68] Silvia D'Ambrosi,[1,2,3,68] Daan C.L. Vessies,[7] Lisa Vermunt,[4,5] Adrienne Vancura,[1,2,3] Mirte Muller,[8] Anna-Larissa N.Niemeijer、[6] Jihane Tannous、[9、10] Laura L. Meijer、[2、11] Tessa Y.S. Le Large、[2、11] Giulia Mantini、[2、12] Niels E. Wondergem、[2、13] Kimberley M. Heinhuis、[14、15] Sandra van Wilpe、[16] A.Josien Smits,[6] Esther E.E. Drees,[2,17] Eva Roos,[11] Cyra E. Leurs,[5,18,19] Lee-Ann Tjon Kon Fat,[20] Ewoud J. van der Lelij,[1,2,3] Govert Dwarshuis,[1,2,3] Maarten J. Kamphuis,[1,2,3] Lisanne E. Visser,[1,2,3] Romee Harting,[1,2,3] Annemijn Gregory,[1,2,3] Markus W. Schweiger,[1,2,3,9,10] Laurine E. Wedekind,[1,2,3] Jip Ramaker,1[,2,3] Kenn Zwaan,[1,2,3] Heleen Verschueren,[1,2,3] Idris Bahce,[6]

(作者名单见下页)

1Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Neurosurgery, Boelelaan 1117, Amsterdam, the Netherlands阿姆斯特丹，阿姆斯特丹自由大学神经外科所在地

2 荷兰阿姆斯特丹癌症中心和液体活检中心

3 荷兰阿姆斯特丹脑肿瘤中心

4 阿姆斯特丹 UMC 所在阿姆斯特丹自由大学临床化学系神经化学实验室，荷兰阿姆斯特丹 Boelelaan 1117 号

5荷兰阿姆斯特丹神经科学校园

6Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Pulmonary Medicine, Boelelaan 1117, Amsterdam, Netherlands 7Department of Laboratory Medicine, the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, Netherlands 8Department of Thoracic Oncology, the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, Netherlands 9Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

10美国马萨诸塞州波士顿哈佛医学院神经科学项目

11Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Surgery, Boelelaan 1117, Amsterdam, the Netherlands 12Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Medical Oncology, Boelelaan 1117, Amsterdam, the Netherlands 13Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Otolaryngology and Head and Neck Surgery, Boelelaan 1117、

荷兰阿姆斯特丹

(附属机构见下页)

摘要

癌症患者可从早期肿瘤检测中获益，因为晚期癌症的治疗效果更佳。血小板参与了癌症的进展，并在局部和全身线索的作用下改变其 RNA 含量，因此被认为是一种很有前景的癌症检测生物资源。我们的研究表明，基于肿瘤教育血小板（TEP）RNA的血液检测能检测出18种癌症类型。ThromboSeq 对无症状对照组的特异性高达 99%，能正确检测出 1,096 份 I-IV 期癌症患者血液样本中三分之二的癌症存在，以及 352 份 I-III 期肿瘤样本中一半的癌症存在。无症状对照组（包括炎症和心血管疾病）和良性肿瘤的假阳性检测结果增加，平均特异性为 78%。此外，在五种不同类型的肿瘤中，血栓串联基因能正确确定80%以上癌症患者的肿瘤起源部位。这些结果凸显了 TEP 衍生 RNA 面板的潜在优势，可作为目前基于血液的癌症筛查方法的补

引言

有几种测序技术可对血液中循环的蛋白质和核酸进行深入分析，包括血浆衍生的无细胞（cf）DNA 和 RNA 分子，它们也可用于微创癌症检测。然而，在早期癌症患者血浆中的突变 cfDNA 水平相对较低，这取决于癌症类型，而且由于衰老相关过程中自然存在的非癌症 cfDNA 变异，其检测也变得复杂（Heitzer 等人，2019 年）。因此，我们需要补充性液体生物资源，以便能在以下情况下检测癌症

Cancer Cell *40*, 999-1009, September 12, 2022

Adrianus J. de Langen,[8] Egbert F. Smit,[8] Michel M. van den Heuvel,[8,21] Koen J. Hartemink,[22] Marijke J.E. Kuijpers,[23,24] Mirjam G.A. oude Egbrink,[25] Arjan W. Griffioen,[2,12] Rafael Rossel,[26,27,28,29] T. Jeroen N. Hiltermann,[30] Elizabeth Lee-Lewandrowski,[31] Kent B. Lewandrowski,[31] Philip C. De Witt Hamer,[1,2,3] Mathilde Kouwenhoven,[2,3,18] Jaap C. Reijneveld,2,3,18,32 William P. J. Leenders,33 Ann Hoeben,34 Irma M. Verdonck-de Leeuw,2,13,35Reijneveld,[2,3,18,32] William P.J. Leenders,[33] Ann Hoeben,[34] Irma M. Verdonck-de Leeuw,[2,13,35] C.Rene´ Leemans，[13] Robert J. Baatenburg de Jong，[36] Chris H.J. Terhaard，[37] Robert P. Takes，[38] Johannes A. Langendijk，[39] Saskia C. de Jager，[40] Adriaan O. Kraaijeveld，[41] Gerard Pasterkamp，[40] Minke Smits，[16] Jack A. Schalken，[42，43]

Sylwia qapin´ska-Szumczyk,[44] Anna qajkowska,[44] Anna J. Z_ aczek,[45] Henk Lokhorst,[2,46] Niels W.C.J. van de Donk,[2,46] Inger Nijhof,[2,46] Henk-Jan Prins,[2,46] Jose´ M. Zijlstra,[2,46] Sander Idema,[1,2,3] Johannes C. Baayen,[1,2,3] Charlotte E. Teunissen,[4,5] Joep Killestein,[5,18,19] Marc G. Besselink,[11] Lindsay Brammen,[47] Thomas Bachleitner-Hofmann,[48] Farrah Mateen,[9] John T.M. Plukker,[49] Michal Heger,[50,51] Quirijn de Mast,[52]

(作者名单见下页)

[14] 荷兰阿姆斯特丹安东尼-范-列文虎克医院荷兰癌症研究所肿瘤内科 15 荷兰阿姆斯特丹安东尼-范-列文虎克医院荷兰癌症研究所临床药理学部 [16] 荷兰奈梅亨拉德布德大学医学中心肿瘤内科

[17] 阿姆斯特丹 UMC 所在地：荷兰阿姆斯特丹 Boelelaan 1117 号阿姆斯特丹 Vrije 大学病理学系 [18] 阿姆斯特丹 UMC 所在地：荷兰阿姆斯特丹 Boelelaan 1117 号阿姆斯特丹 Vrije 大学神经病学系 [19] 阿姆斯特丹MS 中心，荷兰阿姆斯特丹

20瑞典于默奥大学肿瘤放射科学系

21荷兰奈梅亨拉德布德大学医学中心呼吸疾病部

22 荷兰阿姆斯特丹安东尼-范-列文虎克医院荷兰癌症研究所胸外科 [23] 荷兰马斯特里赫特马斯特里赫特大学马斯特里赫特心血管研究所生物化学系 [24] 荷兰马斯特里赫特马斯特里赫特大学医学中心心脏和血管中心血栓形成专家中心 [25] 荷兰马斯特里赫特马斯特里赫特大学马斯特里赫特心血管研究所生理学系 [26] 西班牙巴塞罗那奎罗德克塞斯大学医院罗塞尔肿瘤研究所翻译研究室博士。罗塞尔肿瘤研究所，西班牙巴塞罗那基罗恩德克斯大学医院

27Pangaea Biotech SL，西班牙巴塞罗那

28 西班牙巴塞罗那日耳曼特里亚斯普霍尔医院加泰罗尼亚肿瘤研究所

29 西班牙巴塞罗那分子肿瘤学研究（MORe）基金会

30 格罗宁根大学肺病系，格罗宁根大学医学中心，荷兰格罗宁根

[31] 美国马萨诸塞州波士顿哈佛医学院马萨诸塞州总医院病理学系 32 荷兰海姆斯特德 Stichting Epilepsie Instellingen Nederland (SEIN) 神经病学系 [33] 荷兰奈梅亨 Radboud 分子生命科学研究所生物化学系

[34]荷兰马斯特里赫特，马斯特里赫特大学医学中心肿瘤与发育生物学学院（GROW）肿瘤内科学系

35荷兰阿姆斯特丹，阿姆斯特丹自由大学行为与运动科学学院临床、神经与发展心理学系及阿姆斯特丹公共卫生研究所

36荷兰鹿特丹伊拉斯姆斯医学中心癌症研究所耳鼻咽喉头颈外科部

37荷兰乌得勒支乌得勒支大学医学中心放射治疗部

[38] 荷兰奈梅亨拉德布德大学医学中心耳鼻咽喉头颈外科 [39] 荷兰格罗宁根格罗宁根大学医学中心格罗宁根大学放射肿瘤学系 [40] 荷兰乌得勒支乌得勒支大学医学中心实验心脏病学系

41荷兰乌得勒支乌得勒支大学医学中心心肺科心脏病学系

[42] 荷兰奈梅亨拉德布德大学医学中心医学研究实验室

43荷兰奈梅亨拉德布德大学医学中心泌尿科

[44] 波兰格但斯克，格但斯克医科大学妇科、妇科肿瘤学和妇科内分泌学系 [45] 波兰格但斯克，格但斯克大学和格但斯克医科大学校际生物技术学院转化肿瘤学实验室

46 阿姆斯特丹 UMC 所在地：荷兰阿姆斯特丹 Boelelaan 1117 号阿姆斯特丹自由大学血液学系

47奥地利维也纳，维也纳医科大学外科，普通外科部

48 奥地利维也纳，维也纳医科大学实验室医学临床研究所

49荷兰格罗宁根大学格罗宁根大学医学中心外科系

[50]浙江省嘉兴市嘉兴大学医学院光子医学与实验治疗学嘉兴市重点实验室药学系

51荷兰鹿特丹伊拉斯姆斯医学中心病理学系，实验肿瘤学实验室

[52]荷兰奈梅亨拉德布德大学医学中心内科部

53 荷兰格罗宁根大学格罗宁根大学医学中心外科研究实验室

54波兰格但斯克医科大学肿瘤学和放射治疗系

55荷兰莱顿莱顿大学医学中心妇产科

56荷兰阿姆斯特丹安东尼-范-列文虎克癌症研究所妇科肿瘤部

57 荷兰阿姆斯特丹安东尼-范-列文虎克癌症研究所阿姆斯特丹妇科肿瘤中心

(附属机构见下页)

Ton Lisman,[49,53] D. Michiel Pegtel,[2,17] Harm-Jan Bogaard,[6] Jacek Jassem,[54] Anna Supernat,[45] Niven Mehra,[16] Winald Gerritsen,[16] Cornelis D. de Kroon,[55] Christianne A.R. Lok,[56,57] Jurgen M.J. Piek,[58] Neeltje Steeghs,[14,15] Winan J. van Houdt,59 Ruud H. Brakenhoff,2,13 Gabe S. Sonke,[14] Henk M. Verheul,16 Elisa Giovannetti,2,12,60van Houdt,[59] Ruud H. Brakenhoff,[2,13] Gabe S. Sonke,[14] Henk M. Verheul,[16] Elisa Giovannetti,[2,12,60] Geert Kazemier,[2,11] Siamack Sabrkhany,[61] Ed Schuuring,[62] Erik A. Sistermans,[63,64] Rob Wolthuis,[63] Hanne Meijers-Heijboer,[63] Josephine Dorsman,[63] Cees Oudejans,[65] Bauke Ylstra,[2,17] Bart A. Westerman,[1,2,3] Daan van den Broek,[7] Danijela Koppers-Lalic,[1,2,3] Pieter Wesseling,[2,3,17,66] R. Jonas A. Nilsson,[20] W. Peter Vandertop,[1,2,3] David P. Noske,[1,2,3] Bakhos A. Tannous,[9,10] Nik Sol,[2,3,18] Myron G. Best,[1,2,3,67,*] and Thomas Wurdinger[1,2,3,67,69,*]

[58] 荷兰埃因霍温 Catharina 医院妇产科和 Catharina 癌症研究所 [59] 荷兰阿姆斯特丹 Antoni van Leeuwenhoek 医院荷兰癌症研究所肿瘤外科

[60] 意大利比萨 Pisana per La Scienza 基金会 AIRC 启动小组癌症药理学实验室

61荷兰马斯特里赫特马斯特里赫特大学生理学[系]

62荷兰格罗宁根大学格罗宁根大学医学中心病理学[系]

63 [阿姆斯特丹] UMC 所在地：荷兰阿姆斯特丹 Boelelaan 1117 号阿姆斯特丹自由大学临床遗传学系

64 [阿姆斯特丹]生殖与发育研究所，荷兰阿姆斯特丹

65Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Clinical Chemistry, Boelelaan 1117, Amsterdam, Netherlands

[66Department] of Pathology, Princess Ma'xima Center for Pediatric Oncology and University Medical Center Utrecht, Utrecht, Netherlands [67资深]作者

68[这些]作者做出了同等贡献

69 [牵头]联络人

*通信：m.g.best@amsterdamumc.nl (M.G.B.), t.wurdinger@amsterdamumc.nl (T.W.) https://doi.org/10.1016/j.ccell.2022.08.006

在治疗效果更佳的早期阶段（Cho 等人，2014 年）。

血小板被认为是检测癌症的另一种生物资源。它们在癌症中的作用早在一个多世纪前就已确立（Sabrkhany 等人，2019 年；In 't Veld 和 Wur- dinger，2019 年）。除了凝血功能外，血小板参与炎症、癌症进展和转移的情况也得到了广泛研究（Haemmerle 等人，2018 年；Jiang 等人，2017 年；McAllister 和 Weinberg，2014 年）。血小板大量存在于血液中，很容易分离。它们没有细胞核，但含有巨核细胞衍生的前 mRNA 转录本，这些转录本在受到刺激后可剪接成成熟的 mRNA（Denis 等人，2005 年），并转录成数千种不同的蛋白质（Nassa 等人，2018 年）。此外，血小板还能封存（突变的）肿瘤衍生 RNA（Nilsson 等人，2011 年）。尽管靶向前 mRNA 剪接的确切机制及其驱动线索在很大程度上仍不为人所知，但它为血小板提供了大量潜在的剪接-RNA 生物标记物和替代 RNA 图谱，可用于检测肿瘤。研究表明，肿瘤诱导血小板（TEP）衍生的RNA图谱确实可用于区分几种（个别）肿瘤类型的早期和晚期癌症患者与健康对照（Best等人，2015年，2019年；Hein- huis等人，2020年；Pastuszak等人，2021年；Sabrkhany等人，2017年；Shen等人，2021年；In 't Veld和Wurdinger，2019年；Vernooij等人，2009年；Xing等人，2019年）。在这里，我们展示了 TEP 衍生的 RNA 图谱在检测多达 18 种不同癌症类型方面的潜力。

结果

用于泛癌症检测的血小板采集

由于循环血小板具有容纳和剪接约 5,500 种不同 RNA 的独特能力（Bray 等人，2013 年；Nassa 等人，2018 年；Rowley 等人，2011 年），它们拥有一组宝贵的高度多序列生物标记物，其中最相关和最具鉴别性的剪接 RNA 水平可通过智能选择来选出

软件（Best 等人，2017 年，2019 年）。生物标记物面板选择过程中使用的样本越多，面板就越简洁和精确，所需的计算能力和时间也就越多。因此，我们通过迭代建模确定了用于泛癌算法训练（训练系列）的 20 个最佳样本，以及用于算法优化（评估系列；图 1）的另外 20 个样本。因此，我们从欧洲和北美人群中收集了 2400 多份不同年龄（18-92 岁）和性别的血小板样本，这些样本代表了 18 种不同的肿瘤类型、无症状对照（AC）或有症状对照（SC）（表 1、S1 和 S2）。在血小板 RNA 测序后进行了严格的质量控制，共纳入 2351 份样本进行分析（约 3% 的遗漏率；图 S1A-S1D）。癌症系列样本（n = 1,628）包括最常见的肿瘤类型（表 1、S1 和 S2）。血液样本是在诊断时或治疗期间采集的。部分样本的肿瘤分期不明（n.a.；n = 124）或不详（n.i.；如胶质瘤和多发性骨髓瘤；n = 132；总计 n = 256；占所有癌症的 16%）。无症状对照组包括普通人群中各年龄段的男性和女性，他们均表示没有癌症或其他严重疾病的病史或体征（n = 390）。症状对照组被诊断患有特定症状疾病，包括心血管疾病、良性肿块或炎症，但没有癌症诊断（n = 333）。血小板样本是在抽血后 48 小时内使用标准化的差速离心方案分离出来的，核细胞污染少，血小板活化程度低（Best 等人，2017 年，2019 年）。我们注意到，不同样本上报机构的无症状对照组血小板RNA组成略有不同（图S1E），这可能是由于样本处理方式不同，残留血细胞和/或血浆cfDNA污染所致（Chebbo等人，2022年）。为了尽量减少这种影响，我们加入了数据校正步骤（Best等人，2017年）（图S1F），并遵循逐步标准化的方案（Best等人，2019年）。

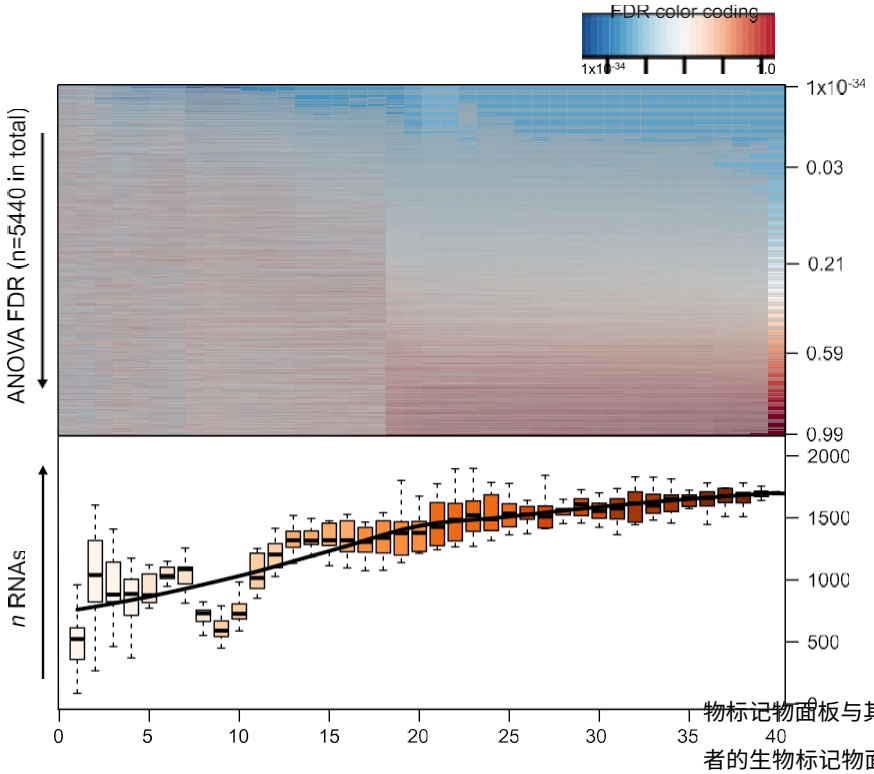


图 1.利用来自 244 例无症状对照和 532 例 癌症患者的 TEP RNA 数据迭代确定生物标记物面板饱 和度 。每次迭代（x 轴）都会从每种肿瘤类型中添加一个新样本，再加上类似数量的无症状对照组样本进行方差分析比较。顶部显示的是方差分析错误发现率（FDR）值，高值（朝向 1，红色）和低值（朝向 0，蓝色）用颜色编码。底部表示方框图中 10 次重复实验中生物标志物面板的大小，通过使用 loess 回归（黑线）对平均面板大小进行求和。方框图报告了 25%（下边 框 ）、50%（中位数）和 75%（上边框）的量化值。下侧线表示大于或等于下侧线 = 1.5 3 四分位数间距的 最小观测值；上侧线表示小于或等于上侧线 = 1.5 3 四分位数间距的 最大观测值。
1.5 3 四分位数间距。另见图 S1。

物标记物面板与其他血小板 RNA 研究中发现的各种类型癌症患者的生物标记物面板之间存在不同程度的重叠（图 S2B-S2D）。泛癌症血栓质谱算法随后被用于

泛癌症检测试验的开发与验证

完整数据集（n = 2,351 个样本）被分成年龄匹配的训练系列（n = 391 个）和评估系列（n = 385 个），以迭代训练泛癌症 thromboSeq 算法（见 STAR 方法）。这些序列包括 18 种肿瘤类型中的 16 种（训练序列 n = 270；评估序列 n = 262）和无症状对照（训练序列 n = 121；评估序列 n = 123）。根据迭代建模得出的结论（图 1），训练系列和评估系列合计每种肿瘤类型约有 40 个样本，具体数量视可用性而定。训练和评估系列中不包括有症状的对照组，因为与无症状对照组相比，无症状对照组的 发病率更高。在训练过程中，算法以特异性达到 99% 为目标，以减少人群筛查所需的假阳性检测结果，并在粒子群优化（PSO）的指导下提高检测灵敏度（图 S1G-S1H）。剩余的 1,575 份样本被分配到验证系列（n = 1,096 名癌症患者、n = 146 名无症状对照组和 n = 333 名有症状对照组）。这一训练过程产生了 493 个泛癌症-血小板 RNA 生物标记物面板（训练系列：曲线下面积 (AUC) 0.91，95% 置信区间 (CI) 0.88-0.94，n = 391）、虚线；评估系列：AUC 0.87，95% CI 0.84-0.91，n = 385，灰线；图 2A 和 S2A）。我们观察到 493 个泛癌症血小板 RNA 生

在验证系列中进行了验证，结果显示无症状对照组的特异性为 99%（n = 146；95% CI 95%-100%），总体灵敏度为 64%（n = 1,096；95% CI 61%-66%）和 46%-72%。

n iterations

四个肿瘤分期的检测准确率（I期为46% [n = 65; 95% CI 34%-59%]，II期为47% [n = 112; 95% CI 38%-57%]，III期为54% [n = 175; 95% CI 46%-61%], IV期为72% [n = 617; 95% CI
68%-75%]，61%为未知阶段[n.a./n.i.；n = 127；95% CI 52%-70%]，验证系列：AUC 0.91; 95% CI 0.89-0.92; n = 1,242; 红线；图 2A-2C 和 S2E-S2F）。值得注意的是，在检测患有各种非癌症疾病（如心血管疾病、良性肿块或炎症）的个体（即无症状对照组）的血小板 RNA 图谱时，Pan.RNA 的 AUC 为 0.91；95 CI 为 0.89-0.92；n = 1,242；红线；图 2A-2C 和 S2E-S2F、图 2D 和 S2G-S2H），这表明泛癌症血栓串联质谱算法可能会导致潜在疾病患者的假阳性检测结果增加，或者在训练过程中加入无症状对照组后检测准确性降低。我们不能排除血栓串联质谱检测出癌症阳性的两个无症状对照组可能患有临床未检测出的癌症。

对验证系列中的癌症样本和对照样本进行的样本供应机构亚组分析表明，该算法可以在一个主要参与训练过程的机构（"13号机构"）和一个只有两个样本参与训练过程的机构（"3号机构"；图3）采集的样本中得到准确验证，这表明泛癌症检验具有普适性。值得注意的是，在锁定生物标记物面板和验证系列的同时，从同一数据集中随机选择样本并使用 1000 个独特的训练和评估系列组成进行算法训练，显示出相似的分类强度（验证系列的 AUC 中位数为 0.87；（四分位数之间））：0.87; (四分位距 (IQR) 0.01), as

表 1.纳入的肿瘤类型、患者特征和泛癌症 thromboSeq 检测结果概览

					预测率				
组别 (n)	性别（女/男/女未知）	中位数年龄（IQR）	验证 AUC (95%) CI; n)	预测率 (95% CI; n)	第二阶段 (95% CI; n)	(95% CI; n)	第三阶段 (95% CI; n)	第四阶段 (95% CI; n)	未知阶段 (95% CI; n)
BRCA (93)	100%, 0%, 0%	58 (15.5)	0.81 (0.74-0.88; 53)	40% (0.26-0.54; 53)	0% (0. (0.02-0.45; 6)	00-17% 0.48; 12)	50% (0.01-0.99; 2)	52% (0.33-0.70; 31)	100% (0.15-1.00; 2)
CHOL (85)	55%, 45%, 0%	68 (14.5)	0.91 (0.86-0.97; 46)	59% (0.43-0.73; 46)	50% (0.01-0.99; 2)	50% (0.21-0.79; 12)	67% (0.22-0.96; 6)	58% (0.36-0.78; 24)	100% (0.16-1.00; 2)
儿童权利委员会 (85)	41%, 59%, 0%	67 (15.5)	0.88 (0.83-0.94; 46)	50% (0.35-0.65; 46)	0% (0. 0.97; 1)	00-50% (0.1-0.99; 2)	33% (0.09-0.99; 3)	50% (0.32-0.68; 32)	62% (0.24-0.91; 8)
ENDO (39)	100%, 0%, 0%	64 (14)	0.78 (0.63-0.93; 12)	42% (0.15-0.72; 12)	57% (0.18-0.90; 7)	0% (0.00-0.97; 1)	25% (0.006-0.80; 4)	不适用	不适用
欧空局 (15)	20%, 80%, 0%	68 (13)	0.80 (0.68-0.92; 15)	40% (0.16-0.67; 15)	不 (0.00-)	适用0% 0.97; 1)	40% (0.12-0.73; 10)	0% (0.00-0.97; 1)	67% (0.09-0.99; 3)
GLIO (132)	32%, 68%, 0%	53 (23)	0.87 (0.82-0.93; 73)	51% (0.38-0.62; 73)	不适用	不适用	不适用	不适用	51% (0.38-0.62; 73)
HCC (23)	26%, 74%, 0%	63 (12)	0.96 (0.89-1.00; 8)	87% (0.47-0.99; 8)	不 (0.02-)	适用100% 1.00; 1)	100% (0.15-1.00; 2)	100% (0.29-1.00; 3)	50% (0.01-0.99; 2)
HNSSC (101)	28%, 72%, 0%	63 (13)	0.92 (0.88-0.96; 61)	57% (0.44-0.70; 61)	50% (0. (0.29-0.99; 2)	1-100% 1.00; 3)	35% (0.15-0.59; 20)	67% (0.49-0.81; 36)	不适用
LYM (20)	45%, 55%, 0%	43 (32)	0.92 (0.83-1.00; 20)	70% (0.45-0.88; 20)	50% (0. 0.99; 2)	1-80% (0.28-0.99; 5)	80% (0.28-0.99; 5)	67% (0.22-0.95; 6)	50% (0.01-0.99; 2)
梅拉 (68)	35%, 65%, 0%	62 (23)	0.90 (0.83-0.96; 28)	57% (0.37-0.75; 28)	不适用	不适用	0% (0.00-0.97; 1)	61% (0.40-0.80; 26)	0% (0.00-0.97; 1)
MM (31)	48%, 52%, 0%	59 (13)	0.99 (0.97-1.00; 11)	91% (0.58-0.99; 11)	不适用	不适用	不适用	不适用	91% (0.58-0.99; 11)
NSCLC (522)	45%, 54.2%, 0.8%	64 (13)	0.94 (0.92-0.95; 482)	74% (0.70-0.78; 482)	50% (0.24-0.75; 16)	70% (0.44-0.89; 17)	63% (0.49-0.74; 62)	77% (0.73-0.81; 372)	73% (0.45-0.92; 15)
OVCAR (144)	100%, 0%, 0%	62 (15)	0.89 (0.84-0.93; 104)	59% (0.48-0.68; 104)	48% (0.28-0.68; 25)	50% (0.18-0.81; 10)	58% (0.40-0.74; 36)	69% (0.50-0.83; 32)	100% (0.02-1.00; 1)
PDAC (126)	40.5%, 59.5%, 0%	68 (14)	0.81 (0.76-0.87; 86)	42% (0.31-0.52; 86)	0% (0. (0.26-0.97; 1)	00-40% 0.55; 47)	30% (0.11-0.54; 20)	61% (0.36-0.82; 18)	不适用
PRCA (35)	0%, 100%, 0%	70 (7)	0.98 (0.93-1.00; 12)	92% (0.61-0.99; 12)	不适用	不适用	不适用	100% (0.48-1.00; 5)	86% (0.42-0.99; 7)
RCC (28)	43%, 57%, 0%	62.5 (16)	0.87 (0.74-1.00; 9)	66% (0.30-0.99; 9)	不适用	不适用	不适用	67% (0.30-0.99; 9)	不适用
SARC (53)	49%, 51%, 0%	60 (17.5)	0.96 (0.91-1.00; 21)	76% (0.53-0.92; 21)	100% (0.29-1.00; 3)	0% (1.00-0.97; 1)	100% (0.39-1.00; 4)	69% (0.38-0.91; 13)	不适用
URO (28)	32%, 68%, 0%	65 (16.5)	0.99 (0.97-1.00; 9)	89% (0.52-0.99; 9)	不适用	不适用	不适用	89% (0.52-0.99; 9)	不适用
交流电	55.6%,	52 (26)	不适用（不适用；	99% (0.95-	不适用	不适用	不适用	不适用	不适用

(390)	40.6%, 3.8%		146)	0.99; 146)					
SC	55.5%,	53 (24)	不适用（不适用；	78% (0.73-	不适用	不适用	不适用	不适用	不适用
(333)	44.2%, 0.3%		333)	0.82; 333)					
癌症	50.2%,	63 (15)	0.91 (0.89-	63% (0.61-	46% (0.34-	47% (0.38-	54% (0.46-	72% (0.68-	61% (0.52-
(1,628)	49.5%, 0.3%		0.92; 1,096)	0.66; 1,096)	0.59; 65)	0.57; 112)	0.61; 175)	0.75; 617)	0.70; 127)

BRCA，乳腺癌；CHOL，胆管癌；CRC，结直肠癌；ENDO，子宫内膜癌；ESO，食管癌；GLIO，胶质瘤；HCC，肝细胞癌；HNSSC，头颈部鳞状细胞癌；LYM，淋巴瘤；MELA，黑色素瘤；MM，多发性骨髓瘤；NSCLC，非小细胞肺癌；OVCAR，卵巢癌；PDAC，胰腺导管腺癌；PRCA，前列腺癌；RCC，肾细胞癌；SARC，肉瘤；URO，尿路上皮癌；AC，无症状对照样本；SC，有症状对照样本。

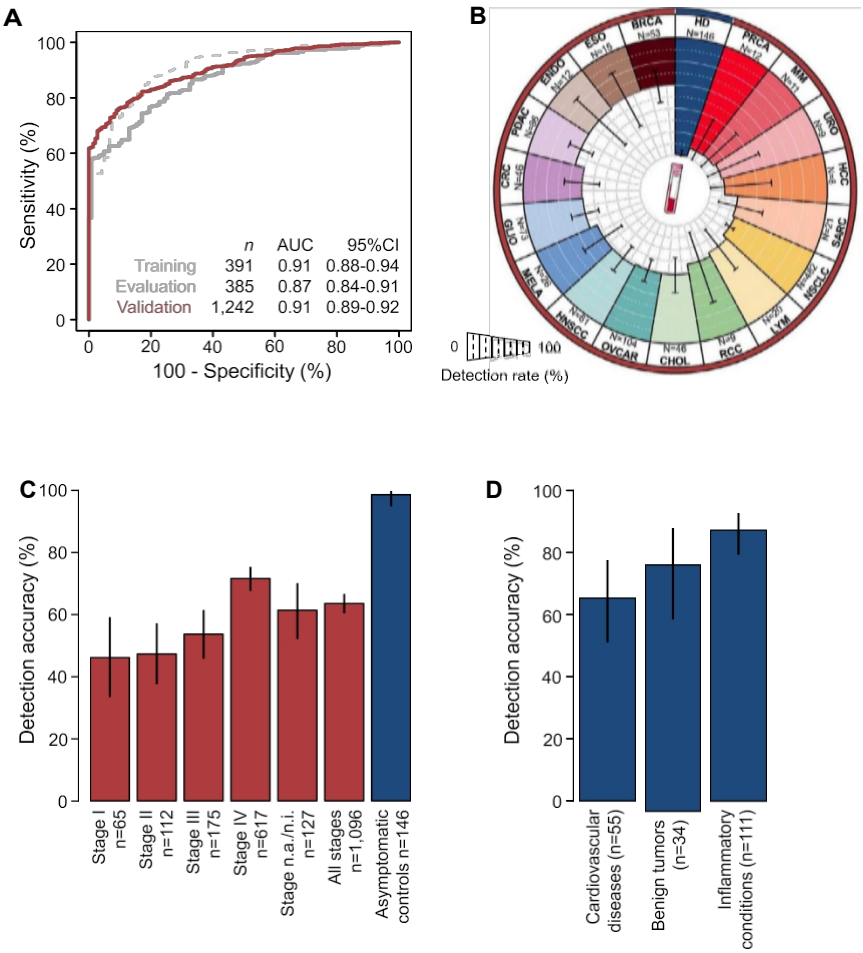


图 2.早期和晚期癌症的泛癌症检测使用 TEP RNA 检测晚期肿瘤

(A) 泛癌症血栓质粒算法的训练（灰虚线）、评估（灰线）和验证（红线）系列的接收者操作特征曲线。图中显示的是样本数、AUC 值和 95% 置信区间。

(B) 在特异性为 99% 的无症状对照组中，验证系列中每种肿瘤类型的检测准确率的 Coxcomb 图。每组的 95% CI 均有标注。AC，无症状对照组；PRCA，前列腺癌；MM，多发性骨髓瘤；URO，尿路上皮癌；HCC，肝细胞癌；SARC，肉瘤；NSCLC，非小细胞肺癌；LYM，淋巴瘤；RCC、肾细胞癌；CHOL，胆管癌；OVCAR，卵巢癌；HNSCC，头颈部癌。鳞状细胞癌；MELA，黑色素瘤；GLIO，胶质瘤；CRC，结直肠癌；PDAC，胰腺导管腺癌；ENDO，内膜癌；ESO，食管癌；BRCA，乳腺癌。

(C) 在特异性为 99% 的情况下，I、II、III 和 IV 期以及 n.a. 和所有阶段的泛癌症血栓质粒算法结果柱状图。图中标出了检测准确率以及 95% 的置信区间。

(D) 无症状对照样本的检测准确率（泛癌算法，特异性 99%）分为心血管疾病 65%（95% CI 51%-78%）；良性肿瘤 79%（95% CI 62%-78%）；恶性肿瘤 65%（95% CI 51%-78%）。

91%）；炎症，87%（95% CI 80%-93%）。另见图 S2。

与随机分类相比（AUC 验证系列中位数：0.50；IQR 0.07；$p < 0.001$）。

此外，为了估计生物标记物面板的鲁棒性，在保持 PSO 选择的参数不变的情况下随机选择训练和评估序列，并对算法进行再训练，结果是生物标记物面板的重叠率约为 40%-50%，而从全部血小板 RNA 库中随机选择 493 条 RNA 后，生物标记物面板的重叠率仅为约 10%。这凸显了 PSO 参数选择对生物标记物面板组合的附加价值。前列腺癌是检出率最高的肿瘤类型，12 例中有 11 例（92%）的特异性达到 99%，而乳腺癌的检出率约为 40%，这表明并非所有肿瘤类型的检出率都相同（图 2B 和 S2E，表 1）。对验证序列的*事后*统计建模显示，RNA 测序文库的大小与算法的输出结果之间没有相关性。不过，我们观察到临床变量 "年龄 "和 "性别 "对算法输出的贡献，前者是癌症患者和无症状患者的泛癌症评分平均增加，后者在乳腺癌患者中最为明显，而可能的技术分析前变量 "样本提供机构 "可能会增强算法输出（图 S3A-S3C）。由此看来，乳腺癌的检测本质上更为复杂，这一点在其他液体活检生物技术中也有所体现。

来源，包括 cfDNA（Cohen 等人，2018 年；Klein 等人，2021 年）。尽管如此，将这些因素迭代添加到以癌症存在作为输出指标的线性生成模型中，结果表明这些因素都没有改变该算法癌症评分的强大预测能力。不过，不能排除这些潜在的混杂变量对算法的影响，需要在后续研究中进行全面评估。值得注意的是，由于样本数量较少，训练过程中未包括淋巴瘤（n = 20）或食管癌（n = 15）患者的样本，但这些样本在验证系列中的分类结果良好（图 2B），这表明已识别出一般的血小板 RNA 泛癌症特征，并能检测出不限于训练过程中包括的癌症类型。同样明显的是，晚期癌症的检出率高于早期癌症（图 2C）。最后，将全血样本储存不同时间（少于 3 到超过 48 小时）以及在 24 或 48 小时内通过邮寄转移全血管，都不会对泛癌测试分类的血小板 RNA 图谱测量结果造成显著干扰（与分离 <3 小时相比，所有比较结果的 p 均大于 0.05，但分离 <8 小时除外，$p < 0.05$，分类得分较低[即癌症信号较少]，学生 t 检验，图 S3D）。这些结果表明，全血样本可在样本加工前运送。总之，我们开发了一种基于血小板 RNA 的泛癌症检测试剂盒。

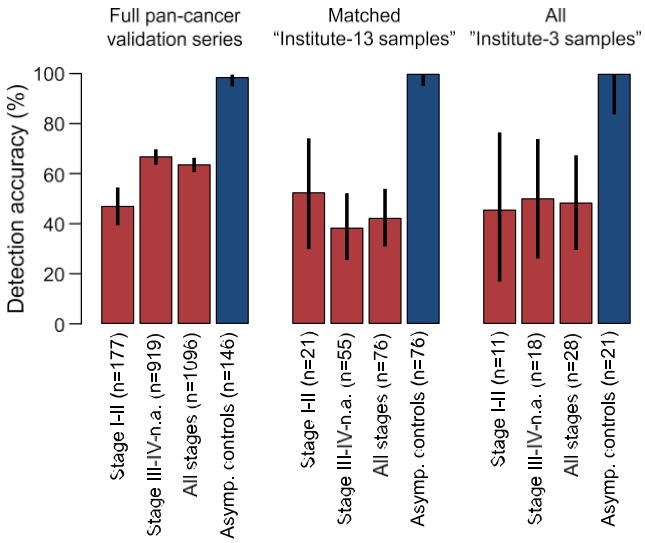


图 3.样本供应机构分组分析

在 99% 的特异性条件下，泛癌症 ThromboSeq 算法的结果条形图：I 期和 II 期合并、III 期、IV 期和不可用/无信息（n.a./n.i.）期合并，以及所有泛癌症验证系列的所有阶段合并（左图）、第 13 研究所（包含在训练过程中，中图）采集的年龄和性别匹配样本，以及第 3 研究所（大部分不包含在训练过程中，右图）采集的所有样本。另见图 S3。

开发肿瘤原发地分类器

接下来，我们试图在 TEP RNA 图谱中识别肿瘤类型特异性图谱，从而确定肿瘤的起源部位。为了简化高度复杂的机器学习任务，构建一种与二元泛癌测试相比对多个分类组（即与二元泛癌测试相比，为了简化高度复杂的机器学习任务，构建对多个分类组（即需要相互区分的组）具有区分性的算法，并保持每组样本的合理数量（最少 100 个样本），我们纳入了非小细胞肺癌（n = 522）、卵巢癌（n = 144）、胶质瘤（n = 132）、胰腺癌（n = 126）和头颈癌（n = 101；共 1,025 个癌症样本）等肿瘤类型。此外，为了提高算法的训练能力，我们采用了 5 倍交叉验证方法，将 80% 的样本分配到训练和评估系列，剩余 20% 的样本用于验证。通过报告第一和第二算法的预测结果，肿瘤原发地血栓素序列算法的交叉验证结果最为理想，总体准确率达到 85%（图 4A-4C，n = 208 个验证序列，95% CI 79%-89%；5 倍交叉验证的中位数为 85% [最小-最大值为 84%-86%]; 仅第一种预测的总体准确率为 68%，95% CI 为 61%-75%；随机分类 [n = 1,000] 验证序列的中位数准确率为 65%，IQR 为 3%，P < 0.001；随机样本选择[n = 1,000] 中位准确率验证系列，82%，IQR 3%）。值得注意的是，肿瘤原发地血栓素基因算法包括了数据集中样本数量较少的肿瘤类型，要求将一些解剖学上密切相关的肿瘤归为一类，结果第一次和第二次预测的分类结果相似（n = 323 个验证序列，95% CI 67%-77%；5 倍交叉验证的中位数）。

70% [最小-最大值 67%-72%]; 随机分类 [n = 1,000] 准确率中位数验证序列，47%，IQR 1%, p < 0.001; 随机样本选择 [n = 1,000] 准确率中位数验证序列，66%，IQR 4%; 图 S4A-S4B）。在此，我们不能排除由于分类模型中样本数量偏斜而造成部分分类混淆的可能性。五组肿瘤原发地血栓素算法提高了转移性肿瘤的分类准确率（I-III 期 [n = 67]，75%，95% CI 63%-84%；IV 期 [n = 67]，75%，95% CI 63%-84%；IV 期 [n = 67]，75%，95% CI 63%-84%）。

[n = 109]，89%，95% CI 82%-94%；n.a.阶段 [n = 32]，91%、95% CI 75%-98%；图 4B）。肿瘤原发地测试的 93 条 RNA 生物标记物与泛癌测试的 493 条 RNA 生物标记物之间只有极少的重叠（10 条 RNA）（表 S3-S4），这是可以预料到的，因为蜂群智能的特性是针对不同的目的优化生物标记物。在肿瘤原发地算法训练过程中加入与器官相关的症状性疾病会使癌症样本的分类准确率略有提高，这表明症状性疾病也会导致血小板 RNA 图谱的特定表型（图 S4C）。我们的结论是，血小板 RNA 可用于识别原发性肿瘤的原发部位。

脑转移瘤患者的血小板可能同时受到原发肿瘤和转移肿瘤部位的教育

由于血小板教育的全身性，转移性癌症患者的 TEP RNA 图谱可能受到原发肿瘤和转移肿瘤部位的教育。因此，我们研究了转移到脑部的原发肿瘤的分类得分是否与原发脑肿瘤（即胶质瘤）的分类得分相关（图 5A）。我们观察到，与没有脑转移的患者相比，有脑转移的患者指向胶质瘤的分类得分平均更高（n = 57；0.15 对 0.08，p = 0.04，学生 t 检验，图 5B）。接下来，我们对132名胶质瘤患者、93名脑转移患者和299名转移肿瘤患者进行了三组方差分析。最后一组转移性肿瘤患者包括原发器官肿瘤患者，这些原发器官肿瘤在 93 例脑转移患者中也有体现（即非小细胞肺癌[NSCLC]、黑色素瘤、乳腺癌、结直肠癌、食管癌、胰腺癌和肾细胞癌）。结果，共有 1,322 条 RNA（误发现率 [FDR] < 0.05）显示出每种情况下 RNA 水平的逐渐增高或降低（图 5C）。随后通过蜂群优化（Best 等人，2019 年）对该 RNA 小组进行分层聚类，结果显示癌症主要起源于脑部的样本与起源于颅外的样本之间存在区别，而脑部转移的样本则分散聚类在两者之间（p < 0.0001；费雪精确检验，图 5D）。这表明，至少对于脑转移瘤而言，血小板 RNA 图谱可能同时受到原发肿瘤和转移瘤的影响。

本研究中描述的测试结果与其他已发表的液体活检测试结果一致（表 S5）。包括数千人在内的大型研究

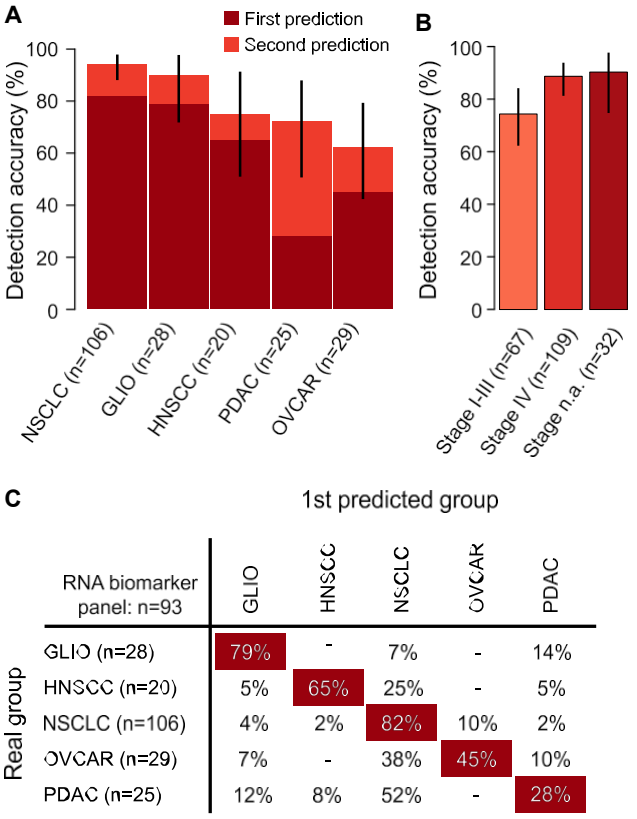


图 4 ThromboSeq 肿瘤原发地算法性能

(A) 五个肿瘤部位的检测准确率。图中显示的是算法的第一次（深红色）和第二次（浅红色）分类以及 95% 的置 信 区 间 。

(B) 每个肿瘤分期（I-III、IV 和 n.a.）五个肿瘤部位的检测准确率。检测准确率和 95% 置信区间已标出。

(C) thromboSeq 肿瘤原发位点算法对第一种算法预测结果的混淆矩阵。真实组（行）和预测组（列）均已标明。百分比表示正确分类样本的百 分 比 。 另见图 S4 及表 S3 和 S4。

Chen等人，2020年；Cohen等人，2018年；Gao等人，2021年；Klein等人，2021年；Lennon等人，2020年；Liu等人，2020年；Stackpole等人，2021年）。大多数研究都是前瞻性的，但没有一项研究深入调查了非癌症疾病的潜在假阳性检测结果（表 S5）。已发表的研究中包含的肿瘤类型数量也不尽相同。与本研究中观察到的情况类似，某些肿瘤类型似乎天生就难以检测，如乳腺癌，这可能与器官的生理学或癌症内在机制有关。血小板 RNA 分析是否能与其他生物大分子互补或互换，还有待研究。

讨论

炎症条件会对泛癌症血栓质谱检测的特异性产生负面影响。进一步深入分析

因此，应在模拟癌症筛查人群的无症状人群中，如在癌症筛查人群中，进行应用泛癌症血栓栓塞序列检测的前瞻性验证研究。因此，应在模拟癌症筛查人群的无症状人群（如 50 岁以上人群）中开展应用泛癌症血栓素检测的前瞻性验证研究，以排除谱系偏倚（Young 等，2018 年）。不过，考虑到人群的癌症发病率，首次验证研究可侧重于因癌症易感综合征（如李-弗劳米尼综合征或 BRCA1/2 基因突变携带者）而接受监测的个体。

在机器学习技术的推动下，泛癌症测试的不断优化将有可能使随着血小板 RNA 样本数量的增加，可以更准确地识别癌症患者。这一理论依据是

此外，还需要进一步研究 thromboSeq 是否能正确诊断在算法开发阶段未包括在内的肿瘤类型以及组织学和分子癌症亚型。此外，还需要进行更多的研究，以调查 thromboSeq 是否能正确诊断算法开发阶段未包括的肿瘤类型以及组织学和分子学癌症亚型。由于非癌症患者的检测结果假阳性率较高，目前的泛癌症检测实际上只适用于无症状的人。一旦将更多非癌症患者的参考数据点添加到持续的泛癌检测系统中，这一问题就有可能得到解决。

血小板 RNA 可用于基于血液的癌症检测和肿瘤原发部位鉴定。非癌症疾病，包括

优化泛癌症测试。此类优化过程可采用相同的 PSO 增强型机器学习算法。不过，由于其耗费时间和计算资源的特性，也可以考虑采用其他方法，例如随机森林和线性回归算法。

由于未经处理的血液可储存 48 小时，因此血液样本可在分离程序之前运送。这样就可以在当地和中央机构进行血液处理。需要注意的是，分离过程必须小心谨慎，以减少红细胞、白细胞和残留血浆（包括细胞外囊泡和 cfDNA）可能造成的污染，并防止血小板在分离、储存和运输过 程 中可能被活化。另外，血栓质谱也可以完全自动化，包括使用专用湿实验室试剂、软件和硬件进行血小板分离过程，最大限度地减少分析前变量对血小板 RNA 图谱的影响，同时排除 "已知诊断偏倚"。最后，应开展后续研究，进一步破译血小板 RNA 图谱的来源和教育，包括巨核细胞衍生 RNA、血小板亚群、替代剪接程序、剪接线索和 RNA 结合蛋白模式的相对贡献。

虽然我们的目标是排除不同医院在样本采集过程中不同的样本处理方式可能造成的偏差，但也不能排除癌症患者、无症状对照组和无症状对照组之间的其他系统性差异对血小板 RNA 原文件造成的额外影响。这些差异包括，例如，使用特定的

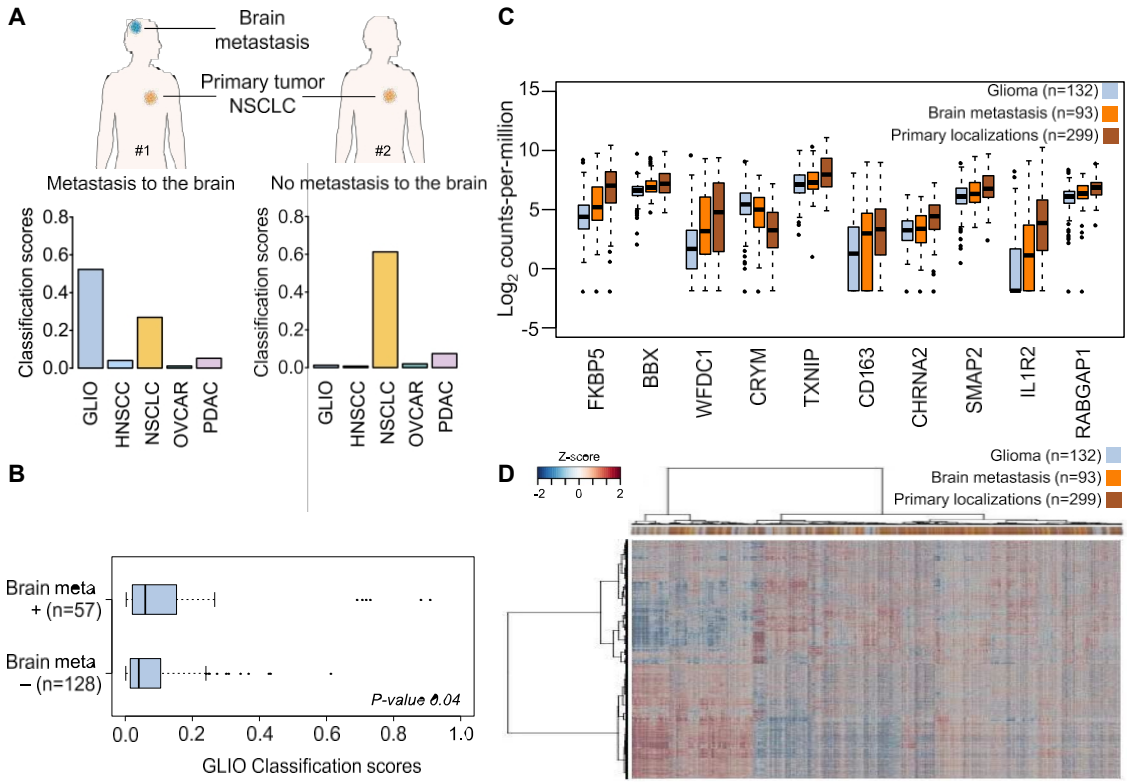


图 5.脑转移瘤定位可能会教育 TEPs

(A) 转移分析示意图。样本来自两名 IV 期转移 NSCLC 患者。1 号患者有脑转移，而 2 号患者没有脑转移。肿瘤原发地算法得出的分类（概率）分数显示，1 号患者的 NSCLC 和胶质瘤均为高值，而 2 号患者的分类分数中没有脑源性信号。

(B) 向大脑转移（+；n = 57）或未向大脑转移（-；n = 128）的样本的胶质母细胞瘤（GLIO）分类得分方框图。

(C) 胶质瘤患者（n = 132；蓝色）或脑转移患者（n = 93；橙色）中最显着富集或减少的 RNA 的表达值（在 x 轴上表示）的单个方框图，以及确有脑转移的肿瘤的原发部位（n = 299；棕色）。图 5B 和图 5C 中的方框显示了 25%（下边框）、50%（中位数）和 75%（上边框）的量化值。下侧线表示大于或等于下侧线=1.5 3四分位数间距的最小观测值；上侧线表示小于或等于上侧线=1.5 3四分位数间距的最大观测值。

(D) 胶质瘤患者（n = 132；蓝色）或脑转移患者（n = 93；橙色）中具有不同剪接RNA水平的RNA的热图和无监督分层聚类分析，以及确有脑转移的肿瘤的原发部位（n = 299；棕色）。列表示样本，行表示 RNA，颜色强度表示经 Z 值转换的 RNA 表达值。样本聚类显示出非随机分区（ P < 0.00001，费雪精确检验）。另见图 S5。

药物、体育锻炼、饮食和精神状态，包括最近诊断出患有癌症。这些潜在的干扰因素应在前瞻性临床试验中进一步研究，并在采血过程中加以规范。

总之，需要对泛癌症血栓质谱检测进行大规模的外部验证，这种验证需要进行专门的、有充分证据支持的、盲法的、以人群为目标的前瞻性临床试验。此类试验还应调查血液检测对临床结果参数（如诊断时的肿瘤分期和/或存活率）的额外益处，同时考虑到准备时间的偏差。血小板 RNA 可作为其他液体生物检查生物资源和生物分子的补充，用于早期癌症检测。

星星+方法

本文的在线版本提供了详细的方法，包括以下内容：

d 量化和统计分析

B 方差分析迭代建模

B 对潜在混杂变量进行事后统计建模

补充资料

补充信息可在线查阅：https://doi.org/10.1016/j. ccell.2022.08.006。

致谢

我们感谢所有献血者愿意参与这项研究。我们感谢 Krzysztof Pastuszak（波兰格但斯克理工大学）在数据处理方面提供的反馈意见，感谢 NKI-AVL 分子病理学和生物库核心设施（CFMPB）的合作者和团队提供的 NKI-AVL 生物库材料和实验室支持，感谢 Amster- dam UMC 液体活检中心核心设施、阿姆斯特丹癌症中心基金会和 Sebastiaan van de Sand（SIT B.V.）提供的计算资源。欧洲研究理事会 713727 和 336540 (T.W.)、荷兰科学研究组织 91711366 (T.W.)、 Stichting STOPHersentumoren.nl (M.G. Best, N. Sol, T.W.)、荷兰癌症协会 (H.M.H.., H.M.V., T.W.) 提供了资金支持、H.M.V., T.W., E.G., G.K., T.W.), the Bennink Foundation 2002262 (L.L.M., T.Y.S.L.L, E.G., G.K.), the European Union's Ho- rizon 2020 research and innovation program under the Marie Sk1odowska-Cu-
国家科学中心（National Science Centre）、美国国家科学院（National Science Centre）、美国国家科学院（National Science Centre）、美国国家科学院（National Science Centre）、美国国家科学院（National Science Centre
2018/02/X/NZ5/01408（A.S.）和格但斯克医科大学法定基金 ST-23，02-0023/07（J.J.）。此外，这项工作还得到了荷兰心血管研究计划（Netherlands Cardiovascular Research Initiative）的支持，该计划得到了荷兰心脏基金会（Dutch Heart Foundation）的支持（CVON2017-4 DOLPHIN-GENESIS, CVON2012-08 PHAEDRA; H-J.B.）。此外，本研究还利用了荷兰癌症协会/Alpe d'Huzes（资助金 VU-2013-5930）资助的荷兰头颈癌生活质量和生物医学队列研究（NET-QUBIC）项目的研究基础设施。资助机构未参与本研究的设计或数据的收集、分析或交互，也未参与手稿的撰写。

作者贡献

构思，S.G.J.G.I.t.V.、N. Sol、M.G. Best、T.W.；资金获取，T.W.、M.G. Best、N.S.、H.M.H.、H.M.V.、E.G.、G.K.、L.L.M.、T.Y.S.L.L.、E.G.、G.K.、D.K.L.、R.R.、H-J.B.；资源、D.C.L.V.、M.M.、A-L.N.M.、J.T.、L.L.M.、T.Y.S.L.L.、G.M.、N.E . W . 、 K.M.H.、S.v.W.、A.J.S.、E . E . E . D.、E.R.、C.E.L、L-A.T.K.F., I.B., A.J.d.L., E.F.S., M.M.v.d.H., K.J.H., M . J.E.K., M . G.A.o.E.. 、A.W.G., R . R., T.J.N.H., E.L.L., K.B.L., P.C.d.W.H., M.K., J.C.R., W.P.J.L.、A.H., I.M.V-d.L., C.R.L., R.J.B.d.J., C.H.J.T., R.P.T., J.A.L., S.C.d.J., A.O.K.. 、G.P.、M.S.、J .A.S.、S.q-S.、A .q.、A.J.Z_ 、H.L.、N.W.C.J.v.d.D.、I.N.、H . J.P、J.M.Z.、S.I.、J.C.B.、C.E.T.、J.K.、M.G. Besselink、L.B.、T.B-H.、F.M.、J.T.M.P、M.H.、Q.d.M.、T.L.、D.M.P.、H-J.B.、J.J.、A.S.、N.M.、W.G.、C.D.d.K.、C.A.R.L、J.M.J.P.、N. Steeghs、W.J.v.H.、R.H.B.、G.S.S.、H.M.V.、E.G.、G.K.、S . S、E.S.、E.A.S.、R.W.、H.M-H.、J.D.、C.O.、B.Y.、B.A.W.、D.v.d.B.、D.K.L.、P.W、R.J.A.N.、W.P.V.、D.P.N.、B.A.T.、N. Sol、M.G. Best；正式分析（湿实验室），E.P.、M.A.F.、S.D.A.、A.V.、M.J.K.、L.E.V.、R.H.、A.G.、M.W.S.、L.E.W.、J.R、K.Z., H.V., N. Sol, M.G. Best; formal analysis (dry lab), software, and data cu- ration, S.G.J.G.I.t.V., E.P.. 、 N. Sol M.G. Best ； 方法和可视化，S.G.J.G.I.t.V.、M.A.、E.P.、M.A.F.、S.D.A.、L.V.、E.J.v.d.L.、G.D.、N. Sol、M.G. Best。Best, T.W.; writing - original draft, S.G.J.G.I.t.V, M.G. Best, T.W.; writing - re- view & editing, all authors.

利益申报

M.G. Best、R.J.A.N. 和 T.W. 是相关专利申请（PCT/NL2011/050518 和 PCT/NL2018/050110）的发明人。R.J.A.N. 和 T.W. 是的股东。M.H. 是 Nurish.Me, Inc. 和 Camelina Sun LLC 的首席配方官，并持有这些公司的股份（其业务活动与本研究无关）。D.M.P. 和 D.K.L. 是 ExBiome BV 公司的股 东 。

收到：2021 年 12 月 6 日

修订：2022 年 5 月 6 日

接受：2022 年 8 月 8 日

出版日期2022 年 9 月 1 日

参考文献

Anders, S., Pyl, P.T., and Huber, W. (2015)。HTSeq-A Python framework to work with high-throughput sequencing data.Bioinformatics *31*, 166-169.

Andy Bunn, M.K. (2017)。用于统计 计算的语言和环境。R Found.Stat.Comput.*10*, 11-18.

Best, M.G., In 't Veld, S.G.J.G., Sol, N., and Wurdinger, T. (2019)。使用拼接血小板 RNA 进行基于血液的疾病诊断的 RNA 测序和蜂群智能增强分类算法开发。Nat.Protoc.*14*, 1206-1234.

Best, M.G., Sol, N., In 't Veld, S.G.J.G., Vancura, A., Muller, M., Niemeijer、A.L.N., Fejes, A.V., Tjon Kon Fat, L.A., Huis In 't Veld, A.E., Leurs, C., et al. (2017).使用肿瘤教育血小板的群集智能增强非小细胞肺癌检测 。Cancer Cell *32*, 238-252.

Best, M.G., Sol, N., Kooi, I., Tannous, J., Westerman, B.A., Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Koster, J., et al. (2015).基于血液的泛癌症、多类和 分子通路癌症诊断。Cancer Cell *28*, 666-676.

Bolger, A.M., Lohse, M., and Usadel, B. (2014).Trimmomatic：用于 Illumina 序列数据的灵活的 修剪器。Bioinformatics *30*, 2114-2120.

Bray, P.F., McKenzie, S.E., Edelstein, L.C., Nagalla, S., Delgrosso, K., Ertel, A., Kupper, J., Jing, Y., Londin, E., Loher, P., et al. (2013)。无核人血小板的复杂转录- tional landscape of the anucleate human platelet.BMC Genom.*14*, 1.

Chebbo, M., Assou, S., Pantesco, V., Duez, C., Alessi, M.C., Chanez, P., and Gras, D. (2022)。血小板纯化是转录组分析的关键步骤- ysis.Int.J. Mol.*23*, 3100.

Chen, X., Gole, J., Gore, A., He, Q., Lu, M., Min, J., Yuan, Z., Yang, X., Jiang, Y., Zhang, T., et al. (2020)。利用血液测试在传统诊断前四年对癌症进行无创早期检测 。Nat.Nat.*11*, 3475.

Cho，H.、Mariotto，A.B.、Schwartz，L.M.、Luo，J. 和 Woloshin，S. (2014)。癌症生存率的变化何时意味着进步？ 发病率和死亡率的启示。J. Natl. Cancer Inst. Monogr. *2014*, 187-197.

Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A., et al. (2018)。利用多分析物血液检验检测和定位 可手术切除的癌症。Science *359*, 926-930.

Denis, M.M., Tolley, N.D., Bunting, M., Schwertz, H., Jiang, H., Lindemann, S., Yost, C.C., Rubner, F.J., Albertine, K.H., Swoboda, K.J., et al. (2005).Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate plate- lets.Cell *122*, 379-391.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013)。STAR: ultrafast universal RNA-seq aligner.Bioinformatics *29*, 15-21.

Gao, Q., Li, B., Cai, S., Xu, J., Wang, C., Su, J., Fang, S., Qiu, F., Wen, X.、Zhang, Y., et al. (2021)。基于血液的甲基化分析（ELSA-seq）对多种癌症的早期检测和定位。J. Clin.Oncol.*39*, 459.

Haemmerle, M., Stone, R.L., Menter, D.G., Afshar-Kharghan, V., and Sood、A.K. (2018)。癌症的血小板生命线：挑战与机遇。Cancer Cell *33*, 965-983.

Heinhuis, K.M., In 't Veld, S.G.J.G., Dwarshuis, G., van den Broek, D., Sol, N., Best, M.G., Coevorden, F.V., Haas, R.L., Beijnen, J.H., van Houdt, W.J., et al. (2020)。肿瘤教育血小板的 RNA 序列，基于血液的肉瘤诊断的新型生物标志物。Cancers *12*, 1372.

Heitzer, E., Haque, I.S., Roberts, C.E.S., and Speicher, M.R. (2019)。基因组学驱动的肿瘤学中液体活检的当前 和未来展望。Nat.Rev. Genet.*20*, 71-88.

In 't Veld, S.G.J.G. and Wurdinger, T. (2019)。肿瘤教育血小板。血液 *133*, 2359-2364.

Jiang, X., Wong, K.H.K., Khankhel, A.H., Zeinali, M., Reategui, E., Phillips, M . J., Luo, X., Aceto, N., Fachin, F., Hoang, A.N., et al. (2017).血小板覆盖循环肿瘤细胞的微流控分离。Lab Chip *17*, 3498-3503.

Klein, E.A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., Chung, G., Clement, J., Gao, J., Hunkapiller, N., et al. (2021)。基于靶向甲基化的多癌症早期检测试验的临床验证（使用内依赖验证集）。Ann.Oncol.*32*, 1167-1177.

Lennon, A.M., Buchanan, A.H., Kinde, I., Warren, A., Honushefsky, A., Cohain, A .T., Ledbetter, D.H., Sanfilippo, F., Sheridan, K., Rosica, D., et al. (2020)。血液检测结合 PET-CT 筛查癌症并指导干预的可行性。Science *369*, eabb9601.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009)。序列对齐/映射格式和 SAMtools。Bioinformatics *25*, 2078-2079.

Liu, M.C., CCGA Consortium, Oxnard, G.R., Klein, E.A., Swanton, C., Seiden, M . V., Liu, M.C., Oxnard, G.R., Klein, E.A., Smith, D., Richards, D., et al. (2020).利用无细胞 DNA 中的甲基化特征进行灵敏、特异的多癌检测和定位。Ann.Oncol.*31*, 745-759.

McAllister, S.S. and Weinberg, R.A. (2014)。肿瘤诱导的全身环境是癌症进展和转移的关键调节因素。Nat.*16*, 717-727.

Nassa, G., Giurato, G., Cimmino, G., Rizzo, F., Ravo, M., Salvati, A., Nyman, T.A., Zhu, Y., Vesterlund, M., Lehtio¨, J., et al. (2018).受生理刺激激活后的血小板复合物前 mRNA 的剪接会导致功能相关的蛋白质组修饰。Sci. Rep. *8*, 498.

Nilsson, R.J.A., Balaj, L., Hulleman, E., Van Rijn, S., Pegtel, D.M., Walraven, M., Widmark, A., Gerritsen, W.R., Verheul, H.M., Vandertop, W.P., et al. (2011)。血小板含有肿瘤衍生的 RNA 生物标记物。血液 *118*，3680-3683。

Pastuszak, K., Supernat, A., Best, M.G., Stokowy, T., In 't Veld, S.G.J.G., qa-pin´ska-Szumczyk, S., qqjkowska, A., Ro´ z_ an´ski, R., Z_ aczek, A.J., Jassem, J., and Wu€rdinger, T. (2021).
生物标志物图谱使基于血液的癌症诊断成为可能。Mol.Oncol.*15*, 2688-2701.

Quinlan, A.R., and Hall, I.M. (2010).BEDTools: a flexible suite of utilities for comparing genomic features.Bioinformatics *26*, 841-842.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014)。利用对照基因或样本的因子分析对 RNA seq 数据进行归一化。Nat.Biotechnol.*32*, 896-902.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011)。RNA-seq 数据的 GC-content 归一化。BMC Bioinf.*12*, 480.

Robinson, M.D. 和 Oshlack, A. (2010)。用于 RNA-seq 数据差异表达分析的比例归一化方法。Genome Biol.

Rowley, J.W., Oler, A.J., Tolley, N.D., Hunter, B.N., Low, E.N., Nix, D.A., Yost, C.C., Zimmerman, G.A., and Weyrich, A.S. (2011)。人类和小鼠血小板转录组的全基因组 RNA-seq 分析。血液 *118*，e101-e111。

RStudio（2015）。RStudio。

Sabrkhany, S., Kuijpers, M.J.E., Griffioen, A.W., and oude Egbrink, M.G.A. (2019)。血小板：癌症血液生物标志物研究的圣杯？Angiogenesis *22*, 1-2.

Sabrkhany, S., Kuijpers, M.J.E., van Kuijk, S.M.J., Sanders, L., Pineda, S., Olde Damink, S.W.M., Dingemans, A.M.C., Griffioen, A.W., and oude Egbrink, M.G.A. (2017)。结合血小板特征可检测早期癌症 。Eur. J. Cancer 80, 5-13.J. Cancer *80*, 5-13.

Shen, Y., Lai, Y., Xu, D., Xu, L., Song, L., Zhou, J., Song, C., and Wang, J. (2021)。基于血小板RNA-seq的 支持向量机算法诊断甲状腺肿瘤。内分泌 *72*，758-783。

Smits, A.J., Arkani, M., In 't Veld, S.G.J.G., Huis In 't Veld, A.E., Sol, N., Groeneveldt, J.A., Botros, L., Braams, N.J., Jansen, S.M., Ramaker, J., et al. (2022)。肺动脉亢 进 患者的血小板 RNA 特征。Ann.Am.Am. Thorac.Soc. https://doi.org/10.1513/AnnalsATS.202201- 085OC。

Sol, N., In 't Veld, S.G.J.G., Vancura, A., Tjerkstra, M., Leurs, C., Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Zwaan, K., et al. (2020a)。用于胶质母细胞瘤检测和（假）进展监测的肿瘤教育血小板 RNA 。Cell Rep. Med.*1*, 100101.

Sol, N., Leurs, C.E., Veld, S.G.I. 't, Strijbis, E.M., Vancura, A., Schweiger, M.W., Teunissen, C.E., Mateen, F.J., Tannous, B.A., Best, M.G., et al. (2020b)。血小板 RNA 能够检测多发性硬化症。Mult.Scler.J. Exp. Transl.Clin.*6*. 2055217320946784.

Stackpole, M., Zeng, W., Li, S., Liu, C.-C., Zhou, Y., He, S., Yeh, A., Wang, Z., Sun, F., Li, Q., et al. (2021).摘要 24：在 无细胞 DNA 上进行多特征集合学习以准确检测和定位癌症。Cancer Res. *81*, 24.

Tolson, B.A., and Shoemaker, C.A. (2007).用于高效计算流域模型校准的动态维度搜索算法。Water Resour.Res. *43*.

Vernooij, F., Heintz, A.P.M., Coebergh, J.W., Massuger, L.F.A.G., Witteveen, P.O., and van der Graaf, Y. (2009)。荷兰卵巢癌治疗的 专业化和高容量护理带来更好的疗效。Gynecol.Oncol.*112*, 455-461.

Xing，S.、Zeng，T.、Xue，N.、He，Y.、Lai，Y.Z.、Li，H.L.、Huang，Q.、Chen，S.L.和Liu，W.L.（2019）。通过RNA-seq开发和验证肿瘤教育血板让整合素α2b（ITGA2B）RNA用于非小 细胞肺癌的诊断和预后。Int.J. Biol.*15*, 1977-1992.

Young, R.P., Christmas, T., and Hopkins, R.J. (2018)。常见癌症的多分析测定和早期检测。J. Thorac.Dis.*10*，S2165-S2167。

星星+方法

关键资源表

试剂或资源	来源	标识符
2 351 份血小板样本	本研究	见表 S2
生物样本		
化学品、肽和重组蛋白质		
RNAlater 稳定溶液	Ambioncat.AM7020	
酸性苯酚：氯仿	安必昂	
RNAseZapSigma-Aldrichcat.R2020		
Agencourt AMPure XP PCR 纯化系统	Beckman Coultercat.A63880	
无核酸酶 H_2	Thermo Fisher	catno.AM9937
关键商业检测		
mirVana miRNA 分离试剂盒	Ambioncat.AM1560	
用于 Illumina 测序 v3 的 SMARTer 超低 RNA 试剂盒	Clontech Laboratoriescat.	
TruSeq Nano DNA 文库制备试剂盒	Illuminacat.FC-121-4001	
安捷伦 RNA 6000 Pico 套件和试剂，2100 生物分析仪	Agilent Technologiescat.5067-1513	
安捷伦高灵敏度 DNA 套件和试剂，2100 生物分析仪	Agilent Technologiescat.5067-4626	
安捷伦 DNA 7500 套件和试剂、2100 生物分析仪	Agilent Technologiescat.5067-1506	
存入数据		
原始和处理过的 RNA-seq 数据	本研究	GEO：GSE183635
软件和算法		
Trimmomatic (version 0.22)	(Bolger et al., 2014)http://www.usadellab.org/cms/? page=trimmomatic	
STAR（2.3.0 版）	（Dobin 等人，2013 年）	https://github.com/alexdobin/STAR
HTSeq (version 0.6.1)	(Anders et al., 2015)http://www-huber.embl.de/HTSeq/doc/ overview.html	
Picardtools（1.115 版）美国	布罗德研究所	https://broadinstitute.github.io/picard/
Samtools（1.115 版）	（Li 等人，2009 年）	http://samtools.sourceforge.net
Bedtools （2.17.0 版）	（Quinlan 和 Hall，2010 年）	http://bedtools.readthedocs.io/en/latest/
MATLAB (version R2015b)	The MathWorks Inc., USA	https://nl.mathworks.com/products/matlab.html
R（3.3.0 版）	（Andy Bunn，2017 年）	https://www.r-project.org
R-studio （版本 0.99.902）	(RStudio, 2015）	https://www.rstudio.com

Bioconductor 软件包 edgeR（3.12.1 版）	(罗宾逊和奥什拉克，2010 年）。	https://bioconductor.org/packages/release/bioc/html/edgeR.html
生物诱导包 EDASeq（2.4.1 版）	(Risso et al., 2011)http://bioconductor.org/packages/release/ bioc/html/EDASeq.html	
Bioconductor 软件包 PPSO（版本 0.9-9991）	(Tolson 和 Shoemaker，2007 年）。	https://www.rforge.net/ppso/
Bioconductor 软件包 RUVSeq（1.4.0 版）	(Risso et al., 2014)http://bioconductor.org/packages/release/ bioc/html/RUVSeq.html	
R 包 e1071（版本 1.6-7）	CRAN	https://cran.r-project.org/web/packages/e1071/index.html
R 包 Caret（版本 6.0-71）	CRAN	https://cran.r-project.org/web/packages/caret/index.html

(接下页)

续

试剂或资源	来源	标识符
R 软件包 pROC（1.8 版）	在CRAN中	https://cran.r-project.org/web/packages/pROC/index.html
R 软件包 ROCR（版本 1.0-7）	在CRAN中	https://cran.r-project.org/web/packages/ROCR/index.html

资源可用性

主要联系人

如需了解更多信息以及索取资源和试剂，请联系主要联系人 Thomas Wurdinger (t.wurdinger@amsterdamumc.nl) 并由其负责处理。

材料供应

这项研究没有产生新的独特材料。

数据和代码的可用性

- 原始测序数据 FASTQ 文件已存入 NCBI GEO 数据库，登录号为 GEO：GSE183635，自发表之日起可公开获取。在该数据库中，作为分析输入的计数表以 "TEP_Count_Matrix.RData "的形式提供。
- 用于生成 thromboSeq 算法的代码，包括 thromboSeq 干实验室管道和复制主要手稿数字的代码，可通过 GitHub (https://github.com/MyronBest/thromboSeq_source_code_v1.5 和 https://github.com/MyronBest/InTVeld_Pancancer_TSOO)获取，自发布之日起可用，仅供研究之 用 。
- 如需重新分析本文所报告的数据，可向主要联系人索取所需的任何其他信息。

实验模型和受试者详情

临床样本采集

通过静脉穿刺抽取了欧洲和美国多家医疗机构的癌症患者、炎症和其他非癌症患者以及无症状者的外周全血。采集的全血分别装在 4 毫升、6 毫升或 10 毫升含抗凝剂 EDTA 的紫盖 BD 真空采血器中。癌症患者通过临床、放射和病理检查确诊，并在采血时确认有可检测到的肿瘤负荷。训练、评估和独立验证系列的样本都是同时采集和处理的。年龄匹配以回顾性方式进行，通过排除和纳入癌症患者和无症状对照组来反复匹配样本，目的是使各组之间的中位年龄和年龄范围相似。表 S1 提供了所纳入样本的详细概述、人口统计学特征、原籍医院以及样本用于哪个系列（即训练、评估或验证）的概述。无症状和有症状的对照组在采血时或之前未被诊断出患有癌症，但也未接受额外的癌症检验。本研究按照《赫尔辛基宣言》的原则进行。本研究获得了各参与医院的机构审查委员会和伦理委员会的批准。由于根据医院的伦理规定对样本进行了匿名处理，因此无法提供无症状对照组的临床随访情况。纳入的样本中有一部分是之前发表的研究的一部分（Best 等人，2015、2017、2019；Heinhuis 等人，2020；Smits 等人，2022；Sol 等人，2020a、2020b）。

方法细节

全血处理

如前所述（Best 等人，2019 年），全血样本在采血后 48 小时内使用标准化血栓素序列（thromboSeq）方案进行处理。为了分离血小板，先用 20 分钟 1203g 离心步骤将富血小板血浆（PRP）与有核血细胞分离，然后再用 20 分钟 3603g 离心步骤将血小板凝集。去除 9/10th 富血小板血浆时要小心谨慎，以减少血小板部分被有核细胞污染的风险。离心在室温下进行。血小板颗粒在 RNAlater（Life Technologies 公司）中仔细重悬，在 4° C 孵育过夜后冷冻在 -80 C。°

血小板 RNA 分离、扩增和标记血栓质谱

测序样本的制备是分批进行的，每批样本包括多种临床情况。所有样本均采用相同的标准化血栓测序方案，包括 SMARTer cDNA 扩增。在分离血小板 RNA 时，将冷冻血小板在冰上解冻，使用 mirVana miRNA 分离试剂盒（Ambion，Thermo Scientific，cat nr. AM1560）分离总 RNA。血小板 RNA 在 30 ㎖ 洗脱缓冲液中洗脱。我们使用 RNA 6000 Picochip（Bioanalyzer 2100，安捷伦）对血小板 RNA 质量进行了评估，并将 RIN 值大于 7 和/或 rRNA 曲线明显的血小板 RNA 样品作为后续实验的质量标准。所有 Bioanalyzer 2100 的质量和数量测量值都是在对参考梯形图（数量、外观和斜率）进行严格评估后，使用默认设置从自动生成的 Bioanalyzer 结果报告中收集的。用于 Illumina 测序的 Truseq cDNA 标记协议要求
~1 mg 输入 cDNA。为了获得足够的血小板 cDNA 以进行稳健的 RNA-seq 文库制备，使用用于 Illumina 测序 v3 的 SMARTer 超低 RNA 试剂盒（Clontech，cat.nr. 634853）对样本进行 cDNA 合成和扩增。扩增前，所有样本均稀释至 ~500 ㎎/㎖ 总 RNA，并再次使用 Bioanalyzer Picochip 对质量进行测定和定量。对于总 RNA 含量低于 400 ㎎/㎖ 的样品，则使用两微升或更多微升的总 RNA（最多约 500 pg 总 RNA）作为 SMARTer 扩增的输入。扩增 cDNA 的质量控制使用带有 DNA 高灵敏度芯片的 Bioanalyzer 2100（安捷伦）进行测量。所有 SMARTer cDNA 合成和扩增均与阴性对照一起进行，阴性对照必须是 Bioanalyzer 分析所要求的阴性。在 300-7500 碱基对 (bp) 区域可检测到片段的样本将被选作进一步处理。为了标记用于测序的血小板 cDNA，所有扩增的血小板 cDNA 首先要经过超声波核酸剪切（Covaris Inc），然后使用 Truseq Nano DNA 样品制备试剂盒（Illumina，cat nr. FC-121-4001）标记用于 Illumina 测序的单个索引条形码。由于血小板 cDNA 输入浓度较低，所有珠子清理步骤均采用 15 分钟的珠子-DNA 结合步骤和 10 个循环的富集 PCR。所有其他步骤均按照生产商的规程进行。使用 DNA 7500 芯片或 DNA 高灵敏度芯片（安捷伦）测量标记血小板 DNA 文库的质量和数量。将产物大小在 300-500 bp 之间的高质量样本以等摩尔浓度汇集（每个汇集 12-19 个样本），用于浅层血栓测序，并在 Illumina Hiseq 2500 或 4000 平台上使用第 4 版测序再制剂进行 100 bp 单读测序。需要对条形码样本库进行精确定量，并仔细进行等摩尔混合，以获得所有样本相同的总测序读数。

通过运输和培养血管评估分析前变量

为了评估几种储存条件的影响，我们设计了一个实验，在该实验中，血液样本与从无症状对照组和 IV 期非小细胞肺癌患者处收集的全血一起，在 EDTA 涂层试管中进行多种环境和运动。血液按照常规抽血程序采集。之后，血液在工作台上的保存时间分别为<3小时（21人）、<8小时（14人）、<12小时（10人）、<24小时（44人）和<48小时（38人），或通过邮寄转移一晚（24小时，5人）或在周末（48小时，7人）。在后一种情况下，样本还受到不规则移动的影响，模拟样本从外周血抽取地点转移到中央处理实验室的过程。全血采用与上文和前文（Best 等人，2019 年）所述相同的血小板分离和 RNA 测序方案。样本按照泛癌症血栓素算法进行分类，分类得分见方框图。

处理原始 RNA 序列数据

以FASTQ文件编码的血小板原始RNA测序数据经过了标准化的RNA测序比对管道，如前所述（Best等人，2015，2017，2019；Heinhuis等人，2020；Sol等人，2020a，2020b）。 总之，RNA测序读数经过了Trimmomatic（0.22版）（Bolger等人，2014）的修剪和序列适配器剪切，使用STAR（2.3.0版）（Dobin等人，2013）映射到人类参考基因组（hg19），并使用HTSeq（0.3.0版）进行汇总、2014），使用 STAR（2.3.0 版）（Dobin 等人，2013）映射到人类参考基因组（hg19），并在 Ensembl 基因注释 75 版（Anders 等人，2015）的指导下使用 HTSeq（0.6.1 版）进行汇总。所有后续统计和分析均在 R（3.3.0 版）和 R-studio（0.99.902 版）中进行。样本筛选是通过评估文库复杂性来进行的，文库复杂性与内含子跨度读数文库大小有部分关联。首先，我们剔除了泛癌症训练和评估系列样本中90%以上的数据集中内含子跨度读数小于30的基因。这一筛选步骤随后应用于验证系列。为确保某一分类组（如特定肿瘤类型）中唯一存在的 RNA 不会在这一筛选步骤后被排除，这一筛选规则被分别应用于每一组（即肿瘤类型）。接

下来，我们对每个样本中至少有一个跨内含子读数被映射的基因数量进行量化，并排除了唯一检测到的高置信度基因小于 1500 个的样本。为了剔除样本间相关性低的血小板样本，我们进行了一次样本剔除交叉相关性分析。数据归一化后，对于数据集训练和评估系列中的每个样本，除 "测试样本 "外的所有样本都被用来计算每个基因的每百万表达计数中位数（参考图谱）。然后，通过皮尔逊相关性确定测试样本与参考集的可比性。相关性小于 0.5 的样本被排除在外。使用 R 软件包（Best 等人，2019 年）中的默认 "perform.RUVg.correction "算法对数据进行校正，使用 "lib.size "作为 "评估变量"，阈值为 0.8。

泛癌症和肿瘤原发地分类器的开发

血栓质谱分类软件

泛癌症血栓质谱算法采用了之前介绍过的方法（Best 等人，2019 年）。简而言之，该算法利用训练和评估系列进行基因面板选择和算法开发，其中特定的选择参数通过 PSO 进行优化。训练序列中的样本作为迭代校正模块的参考样本，该模块旨在通过 RUV 归一化减少混杂因素对数据集的影响（Risso 等人，2014 年）。接下来，通过似然比方差分析测试，利用该训练序列进行基因面板选择。随后，筛选出初步生物标记物面板中高度相关的 RNA。接下来，采用递归特征消除算法训练初步 SVM 分类算法，并识别和过滤对该算法贡献最大的 RNA。随后，在建立最终的 SVM 算法之前，通过网格搜索对 SVM 算法中的*成本*和*伽马*参数进行优化。利用 PSO，我们对通用分类算法的四个步骤进行了优化，即：(i) 用于在文库规模中选择确定为稳定基因的迭代校正模块阈值；(ii) 应用于似然方差分析检验结果的差异剪接过滤器中的 FDR 阈值；(iii) 排除似然方差分析检验后选择的高度相关基因；(iv) 通过递归特征消除算法的基因数量。本研究提出的每项分类任务都向 PSO 算法提交了预定义范围。在 PSO 的每次迭代中，前一次迭代的输出都被用来优化输入变量，模拟黄昏时分鸟群飞入空中的情景。分配给验证系列的样本对血小板 RNA 过滤和质量控制步骤以及算法开发过程没有任何影响。

研究分类软件的优化步骤

在本研究中，为了实现高度特异性泛癌症算法的这一特定目的，实施了几个优化步骤；1）为了过滤低丰度 RNA，现在对训练和评估序列中的每个分类组分别进行评估，主要是为了在多分类肿瘤原发地算法中确保不会错误地过滤掉在某一组中富集的 RNA，2）在质量控制步骤（thromboSeqQC- 功能）中仅使用训练和评估序列作为参考组，从而确保验证序列完全独立于分析和算法训练、4）在支持向量机（SVM）训练过程中引入类权重，以校正不平衡的组规模，这对泛癌分类器尤其有利，因为在训练序列中癌症样本的数量几乎是无症状对照组的两倍。该算法既能处理二元比较（如无症状对照组与癌症），也能处理多类比较（如肿瘤-原发地）。在后一过程中，采用的是一对一方差分析比较。

分类组设置和算法设置

训练、评估和验证系列的样本分配是根据每种肿瘤类型的可用样本总数分层随机进行的，目的是使样本的年龄、性别、肿瘤类型和肿瘤分期特征分布均匀。每个肿瘤类型最好至少有 40 个样本同时用于训练和评估系列，但如果这会导致验证系列中没有该特定肿瘤类型的样本，则前一个系列中的样本数量会减少。淋巴瘤和食管癌这两种肿瘤类型未被纳入训练和评估系列，以便评估算法在未纳入训练过程的肿瘤类型上的性能，同时也因为样本太少而无法将其纳入所有三个系列。因此，该算法对 18 种肿瘤类型中的 16 种进行了训练。表 S2 分别列出了训练、评估和验证系列中的样本 ID。无症状对照组的数量在三个系列中各占一半。我们的目标是尽可能获得年龄匹配的系列样本，但需要注意的是，由于某些癌症类型的固有特性，与其他肿瘤类型和对照组相比，某些肿瘤类型的患者平均年龄较小。泛癌症算法的蜂群变量为lib.size"、"fdr"、"correlatedTranscripts "和 "rankedTranscripts"。采用的边界分别为-0.1-1.0、50 - FDR<0.005、0.5-1.0 和 50 - FDR<0.005。对规则分类器进行了训练，优化了训练过程，使灵敏度达到最高，特异性达到 99%。使用 R 软件包 ggplot2（3.3.5 版）的极坐标系创建了 coxcombplot（图 2B）。

在五组肿瘤原发地算法中，纳入了至少有 100 个样本的肿瘤类型，即非小细胞肺癌、胶质瘤、卵巢癌、头颈部肿瘤和胰腺癌。

对于 11 组肿瘤-原发部位算法，我们决定将解剖位置接近的肿瘤和血液恶性肿瘤归为一组，从而形成更大的分类组，用于算法训练和验证。我们将以下肿瘤部位分组，即多发性骨髓瘤加淋巴瘤、前列腺癌加肾癌加尿路上皮细胞癌、肝细胞癌加胆管癌加胰腺导管腺癌、子宫内膜癌加卵巢癌。食管癌患者因样本数量较少（15 人）而未包括在内，男性未确诊为乳腺癌、子宫内膜癌或卵巢癌。

对于肿瘤原发地算法，除了 FDR 值降至 1 3 10^{-10} 之外，采用了与泛癌症算法相同的蜂群变量。对于泛癌症算法和肿瘤原发地算法，都采用了 60 个蜂群粒子，泛癌症算法迭代 8 次，肿瘤原发地算法迭代 6 次。所有其他设置均沿用先前发布的默认设置（Best 等人，2019 年）。

分类器的输出以灵敏度、特异性、接收器操作曲线（ROC 曲线）下的面积和精确度-召回曲线等指标进行总结，所有指标均使用 R 软件包 ROCR (v.1.0-7)。

算法控制实验

为了支持所开发算法的可解释性，我们进行了多次对照实验。首先，作为对内部可重复性的控制，我们随机抽取了训练和评估序列，同时保留了验证序列和原始分类器的蜂群引导基因面板，并执行了 1000 次训练和分类程序。这样做的理想结果应该是分类准确率相近，从而强调生物标志物面板的正确性。其次，作为对随机分类的控制，在保留原始分类器的群引导基因列表的同时，对 SVM 算法用于训练支持向量的样本的类标签进行随机排列。这一过程进行了 1000 次，理想情况下，分类准确率应该会降低，这表明所包含样本的真实标签具有附加值。第三，作为对 493 个 RNA 生物标记物面板稳健性的控制，使用相同的组大小组合并从相同的训练和评估样本池中选择新的训练和评估系列，然后根据泛癌症算法中设置的 PSO 参数进行基因面板选择和算法训练。随后，将得出的生物标志物面板与 493 个 RNA 生物标志物面板以及从全部血小板 RNA 再样本（n = 5440 个 RNA）中随机选择的 1000 个面板进行叠加。理想情况下，这应显示真正的生物标记物面板与迭代开发的新生物标记物面板之间的重叠，而与从所有血小板 RNA 中随机选择的生物标记物面板之间的重叠则很少。据此计算 P 值。

脑转移分析

为了确定脑转移瘤患者与胶质瘤患者的血小板RNA图谱的相似性，我们选择了所有患有转移性疾病且已知有脑转移瘤的患者，并将他们从五组肿瘤原发地分析中得出的GLIO分类得分与抽血时已知无脑转移瘤的患者进行了比较。两组之间的比较采用学生 t 检验。通过方差分析统计，确定了所有胶质瘤患者、已知有脑转移的患者（包括全部数据集）和肿瘤类型与 "脑转移 "组中的肿瘤类型相似但没有已知脑转移的 IV 期患者的血小板 RNA 特征差异。通过沃德聚类和皮尔逊距离对热图的行和列树枝图进行无损分层聚类。使用费雪精确检验（Fisher.test-function in R）确定非随机分区和无监督分层聚类的相应 p 值，其中 RNA 面板选择的最佳阈值由 PSO 优化。

量化和统计分析

所有统计分析均在 R (v. 3.3.0) 或 MATLAB (v. R2015b) 中进行。连续数据的比较采用学生 t 检验。95% 置信区间使用二项式统计计算。基因面板使用方差分析统计计算。分类器的输出结果用灵敏度、特异性、接收者操作特征曲线（ROC 曲线）的曲线下面积和精确度-召回曲线等指标进行总结，所有指标都使用 R 软件包 ROCR。根据 Delonge 的方法，使用 R 软件包 pROC 计算 ROC 曲线的 95% 置信区间。p 值和 FDR 值小于 0.05 即为具有统计学意义。

分析的统计细节可参见具体实验的结果部分和图例。

方差分析迭代建模

为了通过增加每个条件的样本数量来研究生物标记物面板的稳定性，我们进行了迭代分析。为此，纳入了分配给训练系列（n = 391）和评估系列（n = 385）的所有样本。最初，一组癌症样本包含每种肿瘤类型的一个样本（其中至少有 20 个样本），对这些癌症样本和无症状对照样本进行方差分析比较。随后，在每次迭代过程中，在初始数据集中加入每个肿瘤类型的一个样本和类似数量的无症状对照样本。这样总共进行了 40 次迭代。由于本研究的总样本量较小，因此在这些系列中样本量少于 20 个的肿瘤类型将根据其在数据集中的流行程度纳入其中。每增加一个癌症样本，就增加相同数量的无症状对照组，直到迭代 20 达到这些系列中最多 244 个无症状对照组为止。方差分析比较使用我们软件包（Best 等人，2019 年）中的默认 thromboSeq.ANOVA 函数进行，"lib.size "为 "varia- ble.to.assess"（阈值：0.8）。每次方差分析都会存储 FDR 输出。这一过程重复了 10 次，每次重复都会对每种肿瘤类型的样本进行重新排序。图中显示的是具有代表性的热图，行中包括所有 5440 条检测到的 RNA，在最后一次迭代中根据方差分析 FDR 的递减进行排序。方框图汇总了重复 10 次过程中 FDR<0.05 的 RNA 数量，并用 R 中的 loess 函数拟合了每次迭代的中位值趋势线。

潜在混杂变量的事后统计建模

为了评估 RNA 测序文库的大小、年龄和性别是否可能成为泛癌症血栓质谱算法输出中的混杂变量，我们采用了一个线性模型。在这些分析中，选择了验证系列中的所有癌症样本和无症状对照。排除了患者年龄和/或性别状态未知的样本（14 例），最后得出 1107 例癌症样本和 120 例无症状对照。建立的线性模型以泛癌症 ThromboSeq 算法得分作为结果。预判据包括潜在混杂因素的固定项、组别（癌症或对照）的固定项以及组别与组别之间的交互作用。

混杂因素和组别。其次，我们使用 R 中的 emmeans 软件包（emmeans_1.6.2-1）估算并可视化了组间比较和相关交互作用的边际均值。为了估计潜在的混杂因素（即年龄、性别、样本提供机构和 RNA 测序文库大小）是否对算法的总体预测价值有影响，我们拟合了一个广义线性模型（GLM），将这些因素和算法得分都列为预测因素，并将是否存在癌症列为结果（R-base 软件包"stats"；R 版本 3.6.1）。为此，只纳入了来自同时分离出无症状对照组和癌症样本的研究所的样本，以便进行研究所校正，限制仅来自一个研究所的样本类型造成的过度拟合。分析包括 521 份癌症样本和 100 份无症状对照样本，分别从五个不同机构（即 VUMC、AMC、RAD、VIENNA 和 UMEA）分离出来。根据这一选择，共包括 15 种不同类型的肿瘤（即乳腺癌、胆管癌、结直肠癌、食管癌、头颈部鳞状细胞癌、淋巴瘤、黑色素瘤、多发性骨髓瘤、非小细胞肺癌、卵巢癌、胰腺导管腺癌、前列腺癌、肉瘤、尿道癌和神经胶质瘤）。