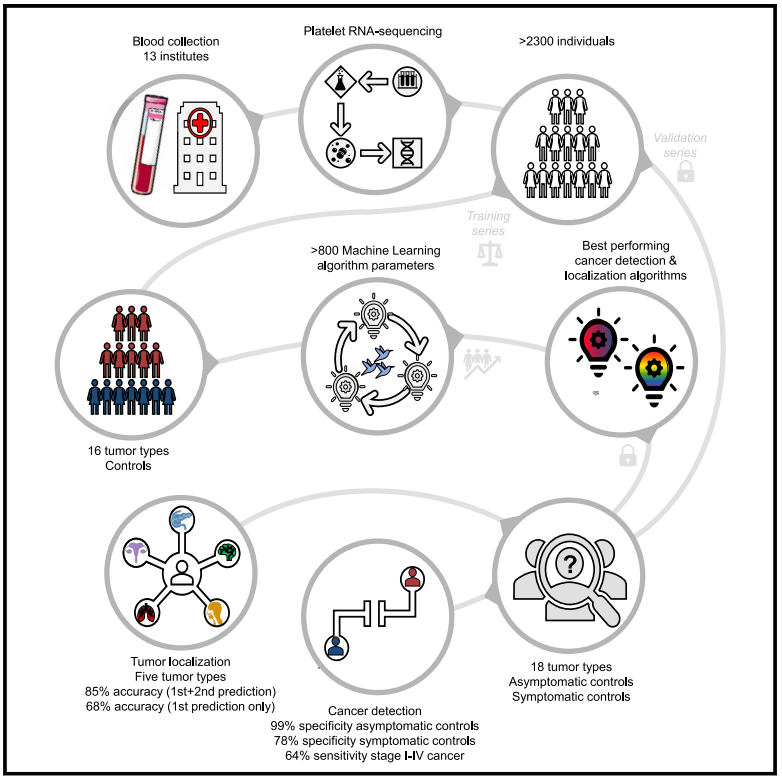
Cancer Cell

Article

Detection and localization of early- and late-stage cancers using platelet RNA

Graphical abstract



Highlights

- Eighteen tumor types are identified by blood platelet RNA analysis with high specificity
- Tumor-type-associated platelet RNA profiles allow for tumor-site-of-origin analysis
- Platelets may be educated by multiple locations of tumor activity
- Platelet RNAs may complement the field of liquid biopsies

Authors

Sjors G.J.G. In 't Veld, Mohammad Arkani, Edward Post, ..., Nik Sol, Myron G. Best, Thomas Wurdinger

Correspondence

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

In brief

In 't Veld et al. employ blood platelet RNA profiles to develop a highly specific pan-cancer blood test covering 18 different tumor types and enabling localization of the primary tumor. This study highlights the value of platelets for early cancer detection and can serve as a complementary biosource for "liquid biopsies."

In 't Veld et al., 2022, Cancer Cell *40*, 999–1009
September 12, 2022
https://doi.org/10.1016/j.ccell.2022.08.006

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Article

Detection and localization of early- and late-stage cancers using platelet RNA

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]

(Author list continued on next page)

[1]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Neurosurgery, Boelelaan 1117, Amsterdam, the Netherlands
[2]Cancer Center Amsterdam and Liquid Biopsy Center, Amsterdam, the Netherlands
[3]Brain Tumor Center Amsterdam, Amsterdam, the Netherlands
[4]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Clinical Chemistry, Neurochemistry Lab, Boelelaan 1117, Amsterdam, the Netherlands
[5]Neuroscience Campus Amsterdam, Amsterdam, the Netherlands
[6]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Pulmonary Medicine, Boelelaan 1117, Amsterdam, the Netherlands
[7]Department of Laboratory Medicine, the Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands
[8]Department of Thoracic Oncology, the Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands
[9]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[10]Neuroscience Program, Harvard Medical School, Boston, MA, USA
[11]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Surgery, Boelelaan 1117, Amsterdam, the Netherlands
[12]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Medical Oncology, Boelelaan 1117, Amsterdam, the Netherlands
[13]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Otolaryngology and Head and Neck Surgery, Boelelaan 1117, Amsterdam, the Netherlands

(Affiliations continued on next page)

SUMMARY

Cancer patients benefit from early tumor detection since treatment outcomes are more favorable for less advanced cancers. Platelets are involved in cancer progression and are considered a promising biosource for cancer detection, as they alter their RNA content upon local and systemic cues. We show that tumor-educated platelet (TEP) RNA-based blood tests enable the detection of 18 cancer types. With 99% specificity in asymptomatic controls, thromboSeq correctly detected the presence of cancer in two-thirds of 1,096 blood samples from stage I–IV cancer patients and in half of 352 stage I–III tumors. Symptomatic controls, including inflammatory and cardiovascular diseases, and benign tumors had increased false-positive test results with an average specificity of 78%. Moreover, thromboSeq determined the tumor site of origin in five different tumor types correctly in over 80% of the cancer patients. These results highlight the potential properties of TEP-derived RNA panels to supplement current approaches for blood-based cancer screening.

INTRODUCTION

Several sequencing technologies enable in-depth analysis of protein and nucleic acids circulating in blood, including plasma-derived cell-free (cf) DNA and RNA molecules that are also used for minimally invasive cancer detection. However, in patients with early-stage cancer, the level of plasma-derived mutant cfDNA is relatively low, depending on the cancer type, and its detection is complicated by the natural presence of non-cancerous cfDNA variants attributed to aging-related processes (Heitzer et al., 2019). Consequently, complementary liquid biosources are desired to enable detection of cancer in

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,[2,13,35] C. René 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 Łapińska-Szumczyk,[44] Anna Łojkowska,[44] Anna J. Żaczek,[45] Henk Lokhorst,[2,46] Niels W.C.J. van de Donk,[2,46] Inger Nijhof,[2,46] Henk-Jan Prins,[2,46] Josée 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]

(Author list continued on next page)

[14]Department of Medical Oncology, the Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands
[15]Department of Clinical Pharmacology, the Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands
[16]Department of Medical Oncology, Radboud University Medical Center, Nijmegen, the Netherlands
[17]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Pathology, Boelelaan 1117, Amsterdam, the Netherlands
[18]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Neurology, Boelelaan 1117, Amsterdam, the Netherlands
[19]MS Center Amsterdam, Amsterdam, the Netherlands
[20]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden
[21]Department of Respiratory Diseases, Radboud University Medical Center, Nijmegen, the Netherlands
[22]Department of Thoracic Surgery, the Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands
[23]Department of Biochemistry, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, the Netherlands
[24]Thrombosis Expertise Centre, Heart and Vascular Centre, Maastricht University Medical Centre, Maastricht, the Netherlands
[25]Department of Physiology, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, the Netherlands
[26]Translational Research Unit, Dr. Rosell Oncology Institute, Quirón Dexeus University Hospital, Barcelona, Spain
[27]Pangaea Biotech SL, Barcelona, Spain
[28]Catalan Institute of Oncology, Hospital Germans Trias i Pujol, Barcelona, Spain
[29]Molecular Oncology Research (MORe) Foundation, Barcelona, Spain
[30]University of Groningen, Department of Pulmonary Diseases, University Medical Center Groningen, Groningen, the Netherlands
[31]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[32]Department of Neurology, Stichting Epilepsie Instellingen Nederland (SEIN), Heemstede, the Netherlands
[33]Department of Biochemistry, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands
[34]Department of Medical Oncology, School for Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, the Netherlands
[35]Department of Clinical, Neuro- and Developmental Psychology, Faculty of Behavioral and Movement Sciences & Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[36]Department of Otolaryngology and Head and Neck Surgery, Erasmus MC Cancer Institute, Rotterdam, the Netherlands
[37]Department of Radiotherapy, University Medical Center Utrecht, Utrecht, the Netherlands
[38]Department of Otorhinolaryngology and Head and Neck Surgery, Radboud University Medical Center, Nijmegen, the Netherlands
[39]Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[40]Department of Experimental Cardiology, University Medical Center Utrecht, Utrecht, the Netherlands
[41]Department of Cardiology, Division of Heart and Lungs, Utrecht University Medical Center, Utrecht, the Netherlands
[42]Urological Research Laboratory, Radboud University Medical Center, Nijmegen, the Netherlands
[43]Department of Urology, Radboud University Medical Center, Nijmegen, the Netherlands
[44]Department of Gynaecology, Gynaecological Oncology and Gynaecological Endocrinology, Medical University of Gdańsk, Gdańsk, Poland
[45]Laboratory of Translational Oncology, Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Gdańsk, Poland
[46]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Hematology, Boelelaan 1117, Amsterdam, the Netherlands
[47]Department of Surgery, Division of General Surgery, Medical University of Vienna, Vienna, Austria
[48]Clinical Institute of Laboratory Medicine, Medical University of Vienna, Vienna, Austria
[49]Department of Surgery, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
[50]Department of Pharmaceutics, Jiaxing Key Laboratory for Photonanomedicine and Experimental Therapeutics, College of Medicine, Jiaxing University, Jiaxing, Zhejiang, PR China
[51]Department of Pathology, Laboratory Experimental Oncology, Erasmus MC, Rotterdam, the Netherlands
[52]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, the Netherlands
[53]Surgical Research Laboratory, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
[54]Department of Oncology and Radiotherapy, Medical University of Gdańsk, Gdańsk, Poland
[55]Department of Obstetrics and Gynaecology, Leiden University Medical Center, Leiden, the Netherlands
[56]Department of Gynaecological Oncology, the Netherlands Cancer Institute – Antoni van Leeuwenhoek, Amsterdam, the Netherlands
[57]Center of Gynaecologic Oncology Amsterdam, the Netherlands Cancer Institute – Antoni van Leeuwenhoek, Amsterdam, the Netherlands

(Affiliations continued on next page)

CellPress
OPEN ACCESS

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,[2,12,60] 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]Department of Obstetrics and Gynaecology and Catharina Cancer Institute, Catharina Hospital, Eindhoven, the Netherlands
[59]Department of Surgical Oncology, the Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands
[60]Cancer Pharmacology Lab, AIRC Start-Up Unit, Fondazione Pisana per La Scienza, Pisa, Italy
[61]Department of Physiology, Maastricht University, Maastricht, the Netherlands
[62]Department of Pathology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
[63]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Clinical Genetics, Boelelaan 1117, Amsterdam, the Netherlands
[64]Amsterdam Reproduction & Development Research Institute, Amsterdam, the Netherlands
[65]Amsterdam UMC Location Vrije Universiteit Amsterdam, Department of Clinical Chemistry, Boelelaan 1117, Amsterdam, the Netherlands
[66]Department of Pathology, Princess Máxima Center for Pediatric Oncology and University Medical Center Utrecht, Utrecht, the Netherlands
[67]Senior authors
[68]These authors contributed equally
[69]Lead contact
*Correspondence: m.g.best@amsterdamumc.nl (M.G.B.), t.wurdinger@amsterdamumc.nl (T.W.)
https://doi.org/10.1016/j.ccell.2022.08.006

an early stage, when treatment outcomes are more favorable (Cho et al., 2014).

Platelets are considered as an alternative biosource for the detection of cancer. Their role in cancer was established more than a century ago (Sabrkhany et al., 2019; In 't Veld and Wurdinger, 2019). Apart from their function in blood clotting, the involvement of platelets in inflammation, cancer progression, and metastasis has been extensively studied (Haemmerle et al., 2018; Jiang et al., 2017; McAllister and Weinberg, 2014). Platelets are present in the bloodstream in large numbers and can be easily isolated. They lack a nucleus but do contain megakaryocyte-derived pre-mRNA transcripts that, upon stimulation, can be spliced into mature mRNA (Denis et al., 2005) and translated into thousands of different proteins (Nassa et al., 2018). In addition, platelets can sequester (mutant) tumor-derived RNAs (Nilsson et al., 2011). Although the exact mechanisms of targeted pre-mRNA splicing and its driving cues remain largely unknown, it provides platelets with an abundance of potential spliced-RNA biomarkers and surrogate RNA profiles for the detection of cancer. It was demonstrated that indeed tumor-educated platelet (TEP)-derived RNA profiles can be employed to differentiate early- and late-stage cancer patients from healthy controls for several (individual) tumor types (Best et al., 2015, 2019; Heinhuis et al., 2020; Pastuszak et al., 2021; Sabrkhany et al., 2017; Shen et al., 2021; In 't Veld and Wurdinger, 2019; Vernooij et al., 2009; Xing et al., 2019). Here, we show the potential of TEP-derived RNA profiles for the detection of up to 18 different cancer types.

RESULTS

Platelet collection for pan-cancer detection

Due to the unique capability of circulating platelets to harbor and splice ~5,500 different RNAs (Bray et al., 2013; Nassa et al., 2018; Rowley et al., 2011), they have a valuable set of highly multiplexed biomarkers, of which the most relevant and discriminating spliced RNA levels can be selected by intelligent selection software (Best et al., 2017, 2019). The more samples that are employed in the biomarker panel selection process, the more concise and precise the panel will be and the more computational power and time are required. Therefore, we determined by iterative modeling an optimum of 20 samples for pan-cancer algorithm training (training series) and another 20 samples for algorithm optimization (evaluation series; Figure 1). Therefore, we collected platelet samples from over 2,400 individuals from all ages (range 18–92) and both sexes from European and North American populations representing 18 different tumor types, asymptomatic controls (ACs), or symptomatic controls (SCs) (Tables 1, S1, and S2). Following stringent quality controls after platelet RNA sequencing, 2,351 samples were included for analysis (~3% dropout rate; Figures S1A–S1D). The cancer series (n = 1,628) included the most prevalent tumor types (Tables 1, S1, and S2). The blood samples were collected at the moment of diagnosis or during treatment. For a subset of samples, tumor stages were unknown (n.a.; n = 124) or not informative (n.i.; e.g., gliomas and multiple myeloma; n = 132; in total n = 256; 16% of all cancers). The asymptomatic controls included male and female individuals from all ages from the general population who reported having no history or signs of cancer or other severe diseases (n = 390). The symptomatic controls were diagnosed with specific symptomatic diseases, including a cardiovascular disease, a benign mass, or an inflammatory condition, but did not have a diagnosis of cancer (n = 333 in total). The platelet samples were isolated using a standardized differential centrifugation protocol within 48 h after blood draw, with low nucleated-cell contamination and low platelet activation (Best et al., 2017, 2019). We noticed slightly different platelet RNA compositions from asymptomatic controls between the different sample-supplying institutions (Figure S1E), potentially attributable to different sample-handling manners, with residual blood cell and/or plasma cfDNA contamination (Chebbo et al., 2022). To minimize this effect, we included data correction steps (Best et al., 2017) (Figure S1F) and followed a step-by-step standardized protocol (Best et al., 2019).

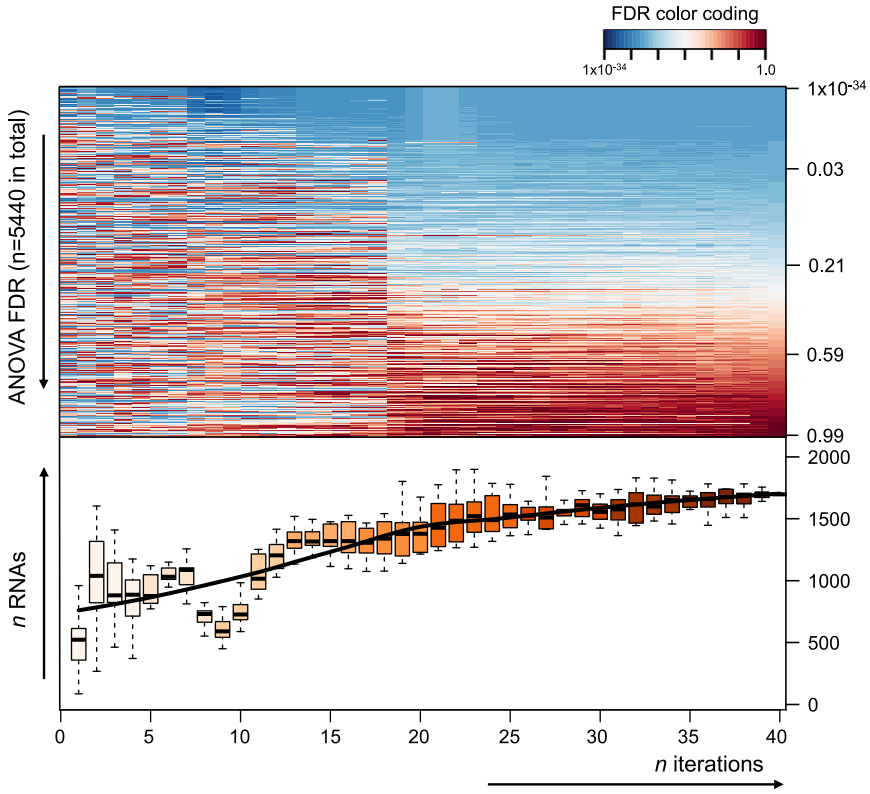


Figure 1. Iterative modeling to estimate the number of tumor samples required for thromboSeq pan-cancer algorithm training

Iterative determination of biomarker panel saturation using TEP RNA data from 244 asymptomatic controls and 532 cancer patients. Per iteration (x axis), a new sample from each tumor type was added, plus a similar number of asymptomatic controls for ANOVA comparison. The top indicates the ANOVA false discovery rate (FDR) values color coded for high values (toward 1, red) and low values (toward 0, blue). The bottom indicates the size of the biomarker panels in the boxplots among 10 repetitions of this iterative experiment, summarized by an average panel size using loess regression (black line). The boxplots report the 25% (lower hinge), 50% (median), and 75% quantiles (upper hinge). The lower whiskers indicate the smallest observation greater than or equal to the lower hinge = 1.5 × interquartile range; the upper whiskers indicate the largest observation less than or equal to the upper hinge = 1.5 × interquartile range. See also Figure S1.

Development and validation of a pan-cancer detection test

The full dataset (n = 2,351 samples) was split into age-matched training (n = 391) and evaluation (n = 385) series to iteratively train the pan-cancer thromboSeq algorithm (see STAR Methods). These series included 16 of 18 tumor types (n = 270 for training series; n = 262 for evaluation series) and asymptomatic controls (n = 121 for training series; n = 123 for evaluation series). The training and evaluation series together had, depending on availability, approximately 40 samples per tumor type included, as concluded from iterative modeling (Figure 1). Symptomatic controls were not included in the training and evaluation series because of their overrepresented prevalence as opposed to the asymptomatic controls, compared with a real-world setting. During the training process, the algorithm was geared toward 99% specificity, to reduce false-positive test results as required for population-based screening, with particle-swarm optimization (PSO)-guided enhancement of the detection sensitivity (Figures S1G–S1H). The remaining 1,575 samples were assigned to the validation series (n = 1,096 cancer patients, n = 146 asymptomatic controls, and n = 333 symptomatic controls). This training process resulted in a 493-pan-cancer-platelet RNA biomarker panel (training series: area under the curve (AUC) 0.91, 95% confidence interval (CI) 0.88–0.94, n = 391, dashed line; evaluation series: AUC 0.87, 95% CI 0.84–0.91, n = 385, gray line; Figures 2A and S2A). We observed variable overlay between this 493-pan-cancer-platelet RNA biomarker panel and those from patients with various types of cancer previously identified in other platelet RNA studies (Figures S2B–S2D). The pan-cancer thromboSeq algorithm was subsequently validated in the validation series, resulting in a specificity of 99% in asymptomatic controls (n = 146; 95% CI 95%–100%), overall sensitivity of 64% (n = 1,096; 95% CI 61%–66%), and 46%–72% detection accuracy in the four tumor stages (46% for stage I [n = 65; 95% CI 34%–59%], 47% for stage II [n = 112; 95% CI 38%–57%], 54% for stage III [n = 175; 95% CI 46%–61%], 72% for stage IV [n = 617; 95% CI 68%–75%], 61% for unknown stage [n.a./n.i.; n = 127; 95% CI 52%–70%], validation series: AUC 0.91; 95% CI 0.89–0.92; n = 1,242; red line; Figures 2A–2C and S2E–S2F). Of interest, when testing the platelet RNA profiles of individuals with various non-cancerous diseases, e.g., cardiovascular disease, a benign mass, or an inflammatory condition (i.e., symptomatic controls), the pan-cancer thromboSeq algorithm performance showed a decreased specificity of 78% (n = 333; 95% CI 73%–82%; Figures 2D and S2G–S2H), indicating that the pan-cancer thromboSeq algorithm may result in increased false-positive test results in patients with an underlying disease or reduced detection accuracy once symptomatic controls are included in the training process. We cannot rule out that the two asymptomatic controls who tested positive for cancer in thromboSeq may have clinically undetected cancer.

Sample-supplying-institute subgroup analysis of both cancer and control samples in the validation series indicated that the algorithm can be accurately validated in samples collected in an institute that has primarily contributed to the training process (“institute 13”), as well as an institute from which only two samples contributed to the training process (“institute 3”; Figure 3), indicating the generalizability of the pan-cancer test. Of note, random sample selection and algorithm training employing 1,000 unique compositions of the training and evaluation series from the same dataset, while locking the biomarker panel and validation series, showed similar classification strength (median AUC validation series: 0.87; (interquartile range (IQR) 0.01), as

Table 1. Overview of included tumor types, patient characteristics, and performance in pan-cancer thromboSeq test

					Prediction rate				
Group (n)	Sex (F/M/ unknown)	Median age (IQR)	Validation AUC (95% CI; n)	Prediction rate (95% CI; n)	Stage I (95% CI; n)	Stage II (95% CI; n)	Stage III (95% CI; n)	Stage IV (95% CI; n)	Unknown stages (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% (0.00–0.45; 6)	17% (0.02–0.48; 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)
CRC (85)	41%, 59%, 0%	67 (15.5)	0.88 (0.83–0.94; 46)	50% (0.35–0.65; 46)	0% (0.00–0.97; 1)	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)	N/A	N/A
ESO (15)	20%, 80%, 0%	68 (13)	0.80 (0.68–0.92; 15)	40% (0.16–0.67; 15)	N/A	0% (0.00–0.97; 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)	N/A	N/A	N/A	N/A	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)	N/A	100% (0.02–1.00; 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% (0.1–0.99; 2)	100% (0.29–1.00; 3)	35% (0.15–0.59; 20)	67% (0.49–0.81; 36)	N/A
LYM (20)	45%, 55%, 0%	43 (32)	0.92 (0.83–1.00; 20)	70% (0.45–0.88; 20)	50% (0.1–0.99; 2)	80% (0.28–0.99; 5)	80% (0.28–0.99; 5)	67% (0.22–0.95; 6)	50% (0.01–0.99; 2)
MELA (68)	35%, 65%, 0%	62 (23)	0.90 (0.83–0.96; 28)	57% (0.37–0.75; 28)	N/A	N/A	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)	N/A	N/A	N/A	N/A	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% (0.00–0.97; 1)	40% (0.26–0.55; 47)	30% (0.11–0.54; 20)	61% (0.36–0.82; 18)	N/A
PRCA (35)	0%, 100%, 0%	70 (7)	0.98 (0.93–1.00; 12)	92% (0.61–0.99; 12)	N/A	N/A	N/A	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)	N/A	N/A	N/A	67% (0.30–0.99; 9)	N/A
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)	N/A
URO (28)	32%, 68%, 0%	65 (16.5)	0.99 (0.97–1.00; 9)	89% (0.52–0.99; 9)	N/A	N/A	N/A	89% (0.52–0.99; 9)	N/A
AC (390)	55.6%, 40.6%, 3.8%	52 (26)	N/A (N/A; 146)	99% (0.95–0.99; 146)	N/A	N/A	N/A	N/A	N/A
SC (333)	55.5%, 44.2%, 0.3%	53 (24)	N/A (N/A; 333)	78% (0.73–0.82; 333)	N/A	N/A	N/A	N/A	N/A
Cancer (1,628)	50.2%, 49.5%, 0.3%	63 (15)	0.91 (0.89–0.92; 1,096)	63% (0.61–0.66; 1,096)	46% (0.34–0.59; 65)	47% (0.38–0.57; 112)	54% (0.46–0.61; 175)	72% (0.68–0.75; 617)	61% (0.52–0.70; 127)

BRCA, breast cancer; CHOL, cholangiocarcinoma; CRC, colorectal cancer; ENDO, endometrial cancer; ESO, esophageal cancer; GLIO, glioma; HCC, hepatocellular carcinoma; HNSSC, head and neck squamous cell carcinoma; LYM, lymphoma; MELA, melanoma; MM, multiple myeloma; NSCLC, non-small cell lung cancer; OVCAR, ovarian cancer; PDAC, pancreatic ductal adenocarcinomas; PRCA, prostate cancer; RCC, renal cell carcinoma; SARC, sarcoma; URO, urothelial carcinoma; AC, asymptomatic control samples; SC, symptomatic control samples.

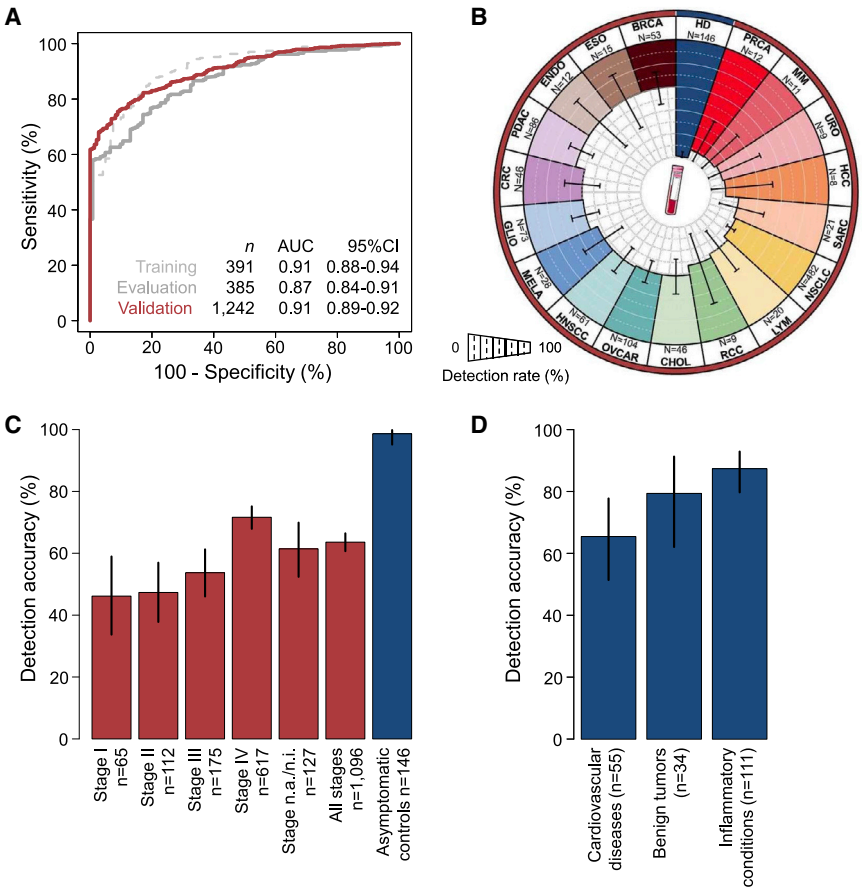


Figure 2. Pan-cancer detection of early- and late-stage tumors using TEP RNA

(A) Receiver operating characteristics curve of the pan-cancer thromboSeq algorithm of the training (dashed gray line), evaluation (gray line), and validation (red line) series. Indicated are sample numbers, AUC values, and the 95% confidence intervals.

(B) Coxcomb plot of the detection accuracies per tumor type included in the validation series, at 99% specificity in the asymptomatic controls. For each group, the 95% CI is indicated. AC, asymptomatic controls; PRCA, prostate cancer; MM, multiple myeloma; URO, urothelial carcinoma; HCC, hepatocellular carcinoma; SARC, sarcoma; NSCLC, non-small cell lung cancer; LYM, lymphoma; RCC, renal cell carcinoma; CHOL, cholangiocarcinoma; OVCAR, ovarian cancer; HNSCC, head and neck squamous cell carcinoma; MELA, melanoma; GLIO, glioma; CRC, colorectal cancer; PDAC, pancreatic ductal adenocarcinoma; ENDO, endometrial cancer; ESO, esophageal cancer; BRCA, breast cancer.

(C) Bar plot of the pan-cancer thromboSeq algorithm results shown at 99% specificity for stages I, II, III, and IV and n.a. and all stages included. The detection accuracy with 95% confidence intervals is indicated.

(D) Detection accuracy (pan-cancer algorithm at 99% specificity) of symptomatic control samples grouped into cardiovascular disease, 65% (95% CI 51%–78%); benign tumors, 79% (95% CI 62%–91%); and inflammatory conditions, 87% (95% CI 80%–93%). See also Figure S2.

opposed to random classification (median AUC validation series: 0.50; IQR 0.07; $p < 0.001$).

Also, in order to estimate the robustness of the biomarker panel, random selection of training and evaluation series while maintaining the same PSO-selected parameters, and retraining of the algorithm, resulted in ~40%–50% overlay of the biomarker panels, compared with only ~10% overlay with the biomarker panels once 493 random RNAs were selected from the full platelet RNA repertoire. This highlights the added value of parameter selection by PSO for biomarker panel composition. Prostate cancer was the most abundantly detected tumor type, with 11 of 12 cases (92%) detected at 99% specificity, whereas breast cancer was detected in approximately 40% of the cases, indicating that not all tumor types were detected at the same rate (Figures 2B and S2E, Table 1). *Post hoc* statistical modeling of the validation series showed there is no correlation between the RNA-sequencing library size and the algorithm's output. However, we observed a contribution of the clinical variables "age," with older individuals having on average increased pan-cancer scores for cancer patients and asymptomatic controls separately, and "sex," which is most pronounced in patients with breast cancer, potentially enhanced by a possible technical pre-analytical variable, "sample-supplying institution," to the algorithm's output (Figures S3A–S3C). Here, it seems that breast cancer is intrinsically more complicated to detect, as had also been noted for other liquid biopsy biosources, including cfDNA (Cohen et al., 2018; Klein et al., 2021). Despite this, iterative addition of these factors to a generalized linear model with the cancer presence as output measure showed that none of these factors changed the strong predictive power of the algorithm's cancer score. However, a contribution of such potential confounding variables to the algorithm cannot be ruled out and requires thorough evaluation in follow-up studies. Of note, whereas samples from patients with lymphoma (n = 20) or esophageal cancer (n = 15) were not included in the training process due to the low number of samples available, they classified well in the validation series (Figure 2B), thus indicating that a general platelet RNA pan-cancer profile was identified and enabled the detection of cancer types not limited to those included in the training process. It was also evident that late-stage cancers exhibited higher detection rates compared with early-stage cancers (Figure 2C). Finally, storage of whole blood samples for different times (less than 3 to over 48 h) and transfer of whole blood tubes via mail within 24 or 48 h did not result in significant disturbance of the measured platelet RNA profiles as classified by the pan-cancer test (all comparisons $p > 0.05$ compared with isolation <3 h, except for isolation <8 h, $p < 0.05$ with lower classification scores [i.e., less cancer signal], Student's t test, Figure S3D). These results indicate that whole-blood samples may be shipped before sample processing. In all, we developed a platelet RNA-based test for pan-cancer detection.

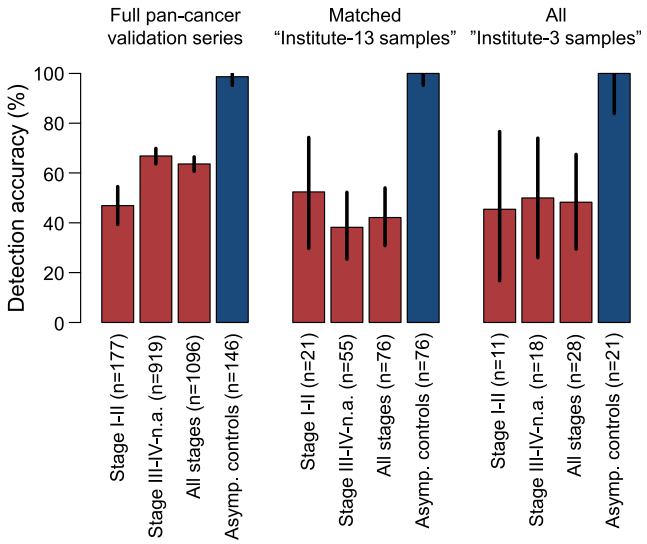


Figure 3. Sample-supplying-institute subgroup analysis

Bar plots of the pan-cancer thromboSeq algorithm results shown at 99% specificity for stages I and II combined, stages III, IV, and not available/not informative (n.a./n.i.) combined, and all stages combined for the full pan-cancer validations series (left), for an age- and sex-matched selection of samples collected at institute 13 (included in the training process, middle), and all samples collected at institute 3 (largely excluded from the training process, right). See also Figure S3.

Development of a tumor-site-of-origin classifier

Next, we sought to identify tumor-type-specific profiles in the TEP RNA profiles so as to identify the tumor site of origin. To simplify the highly complex machine learning task to construct an algorithm discriminative for multiple classification groups (i.e., groups that need to be discriminated from one another) compared with the binary pan-cancer test and to maintain a reasonable number of samples per group (minimum of 100 samples), we included the tumor types non-small cell lung cancer (n = 522), ovarian cancer (n = 144), glioma (n = 132), pancreatic cancer (n = 126), and head-and-neck cancer (n = 101; n = 1,025 cancer samples in total). In addition, to improve the algorithm's training power, we adopted a 5-fold cross-validation approach in which 80% of the samples were assigned to a training and evaluation series and the remaining 20% of the samples were used for validation purposes. The tumor-site-of-origin thromboSeq algorithm resulted in the most optimal cross-validation with an overall accuracy of 85% by reporting the first and second algorithm predictions (Figures 4A–4C, n = 208 validation series, 95% CI 79%–89%; median of 5-fold cross-validation 85% [min-max 84%–86%]; overall accuracy from first prediction only, 68%, 95% CI 61%–75%; random classification [n = 1,000] median accuracy validation series, 65%, IQR 3%, $p < 0.001$; random sample selection [n = 1,000] median accuracy validation series, 82%, IQR 3%) by employing a biomarker RNA panel of 93 RNAs in total. Of note, a tumor-site-of-origin thromboSeq algorithm including tumor types with a lower number of samples in the dataset, requiring some anatomically closely related tumors to be grouped together, resulted in similar classification performance combining first and second predictions (n = 323 validation series, 95% CI 67%–77%; median of 5-fold cross-validation 70% [min-max 67%–72%]; random classification [n = 1,000] median accuracy validation series, 47%, IQR 1%, $p < 0.001$; random sample selection [n = 1,000] median accuracy validation series, 66%, IQR 4%; Figures S4A–S4B). Here, we cannot exclude that part of the classifications can be confounded due to skewed sample numbers in the classification model. The five-group tumor-site-of-origin thromboSeq algorithm resulted in increased classification accuracies among the metastasized tumors (stages I–III [n = 67], 75%, 95% CI 63%–84%; stage IV [n = 109], 89%, 95% CI 82%–94%; stage n.a. [n = 32], 91%, 95% CI 75%–98%; Figure 4B). There was only minimal overlap (10 RNAs) between the 93-RNA biomarker panel for the tumor-site-of-origin test with the 493-RNA biomarker panel from the pan-cancer test (Tables S3–S4), as can be expected because of the nature of swarm intelligence to optimize the biomarker panel for different purposes. Inclusion of organ-relevant symptomatic diseases in the tumor-site-of-origin algorithm training process resulted in slightly improved classification accuracies for the cancer samples, which suggests that the symptomatic conditions can also result in a specific phenotype in the platelet RNA profiles (Figure S4C). We conclude that platelet RNA may be employed for identifying the primary tumor site of origin.

Platelets from patients with a brain metastasis may be educated by both the primary and the metastatic tumor sites

Due to the systemic nature of platelet education, the TEP RNA profiles in patients with metastasized cancer could be educated by both the primary and the metastatic tumor sites. Therefore, we investigated whether the classification score of a primary tumor with a metastasis to the brain was correlated to the classification score of a primary brain tumor, i.e., glioma (Figure 5A). We observed that the classification score pointing toward glioma was higher on average for patients with a brain metastasis compared with patients without a brain metastasis (n = 57; 0.15 versus 0.08, $p = 0.04$, Student's t test, Figure 5B). Next, we performed a three-group ANOVA differential RNA-level analysis between 132 glioma patients, 93 patients with a metastasis to the brain, and 299 cancer patients with a metastasized tumor. The last group of metastasized tumors included patients with a tumor in a primary organ that is also represented in the 93 patients with a metastasis to the brain (i.e., non-small cell lung cancer [NSCLC], melanoma, breast cancer, colorectal cancer, esophageal cancer, pancreatic cancer, and renal cell carcinoma). This resulted in a total of 1,322 RNAs (false discovery rate [FDR] < 0.05) that showed a gradually increasing or decreasing RNA level per condition (Figure 5C). Subsequent hierarchical clustering of this RNA panel, enhanced by swarm optimization (Best et al., 2019), showed a distinction between samples with cancer originating primarily from the brain and those originating extracranially, and the samples with a brain metastasis diffusely clustered in between ($p < 0.0001$; Fisher's exact test, Figure 5D). Altogether, this indicates that, at least for brain metastases, platelet RNA profiles may be influenced by both the primary tumor and the metastasis.

The performance of the test described in this study shows results in line with other published liquid biopsy tests (Table S5). Large studies including thousands of individuals are available

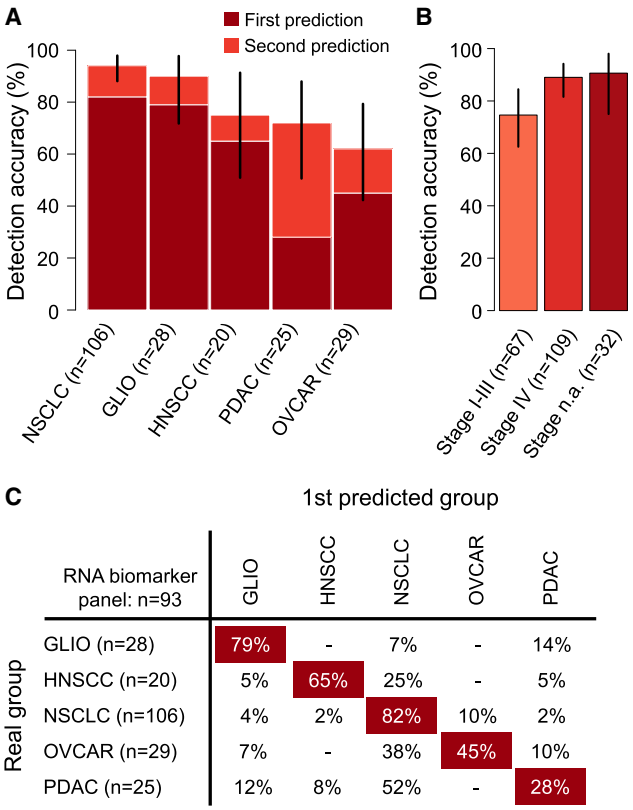


Figure 4. thromboSeq tumor-site-of-origin algorithm performance

(A) Detection accuracy of five tumor sites. Indicated are first (dark red) and second (light red) classifications of the algorithm and 95% confidence intervals.

(B) Detection accuracy of the five tumor sites per tumor stage (I–III, IV, and n.a.). Detection accuracy and 95% confidence interval are indicated.

(C) Confusion matrix of the thromboSeq tumor-site-of-origin algorithm predictions for the first algorithm's prediction. The real groups (rows) and predicted groups (columns) are indicated. Percentages indicate the percentage of correctly classified samples. See also Figure S4 and Tables S3 and S4.

for cfDNA analysis, both mutation analysis and copy number analysis, and methylation patterns, potentially in combination with protein biomarkers (Chen et al., 2020; Cohen et al., 2018; Gao et al., 2021; Klein et al., 2021; Lennon et al., 2020; Liu et al., 2020; Stackpole et al., 2021). Most studies were performed prospectively, although none have investigated in depth the potential false-positive test results from non-cancerous diseases (Table S5). The number of tumor types included also varies among the published studies. Similar to what was observed in this study, certain tumor types seem to be inherently difficult to detect, such as breast cancer, possibly related to the organ's physiology or cancer-intrinsic mechanisms. It remains to be investigated whether platelet RNA analysis is complementary to or interchangeable with the other biomolecules.

DISCUSSION

Platelet RNA enables blood-based cancer detection and tumor-site-of-origin identification. Non-cancerous diseases, including inflammatory conditions, negatively affect the specificity of the pan-cancer thromboSeq test. Further in-depth analysis of various (a)symptomatic populations is warranted, potentially including a broader selection of non-cancerous diseases adjusted to the prevalence in the intended screening population. Therefore, a prospective validation study applying the pan-cancer thromboSeq test should be performed in asymptomatic settings that mimic the cancer screening population, e.g., people older than 50 years of age, to rule out spectrum bias (Young et al., 2018). However, taking into account the population's prevalence of cancer, a first validation study may focus on individuals already being monitored because of a cancer predisposition syndrome, such as individuals with Li-Fraumeni syndrome or BRCA1/2 mutation carriers.

The continuous optimization of the pan-cancer test, facilitated by machine learning technologies, will potentially enable more accurate identification of cancer patients as the repository of platelet RNA samples increases. This rationale is supported by the correct identification of the two tumor types that are not included in the pan-cancer test development. Also, additional research is required to investigate whether thromboSeq may correctly diagnose tumor types and histological and molecular cancer subtypes that are not included during the algorithm development phase. Due to the nature of high false-positive test results in patients with non-cancerous diseases, the current pan-cancer test is practically applicable only to asymptomatic individuals. Potentially, this may be extended once more reference data points from non-cancer-afflicted individuals are added to the continuously optimizing pan-cancer test. Such optimization processes may be guided by the same PSO-enhanced machine learning algorithms. However, due to its time- and computational resource-consuming nature, other approaches can also be considered, for example, random forest and linear regression algorithms.

As unprocessed blood can be stored up to 48 h, the blood samples may be shipped before the isolation procedure. This enables blood processing at both local and central institutes. It should be noted that the isolation has to be performed carefully to reduce potential contamination with erythrocytes and leucocytes and residual plasma, including extracellular vesicles and cfDNA, and to prevent potential activation of platelets during the isolation procedure, storage, and transport. Alternatively, thromboSeq may be fully automatized, including the platelet isolation process, using dedicated wet-lab reagents, software, and hardware, minimizing the influence of pre-analytical variables on the platelet RNA profiles and also to rule out "known diagnosis bias." And last, follow-up research should be performed to further decipher the origin and education of the surrogate platelet RNA profiles, including the relative contributions of megakaryocyte-derived RNAs, blood platelet subpopulations, alternative splicing programs, splicing cues, and RNA-binding protein patterns.

Although we aimed to rule out potential bias introduced by different sample handling during the collection procedure at different hospitals, additional variation in the platelet RNA profiles introduced by other systemic differences between cancer patients, symptomatic, and asymptomatic controls cannot be excluded. These include, for example, the use of specific

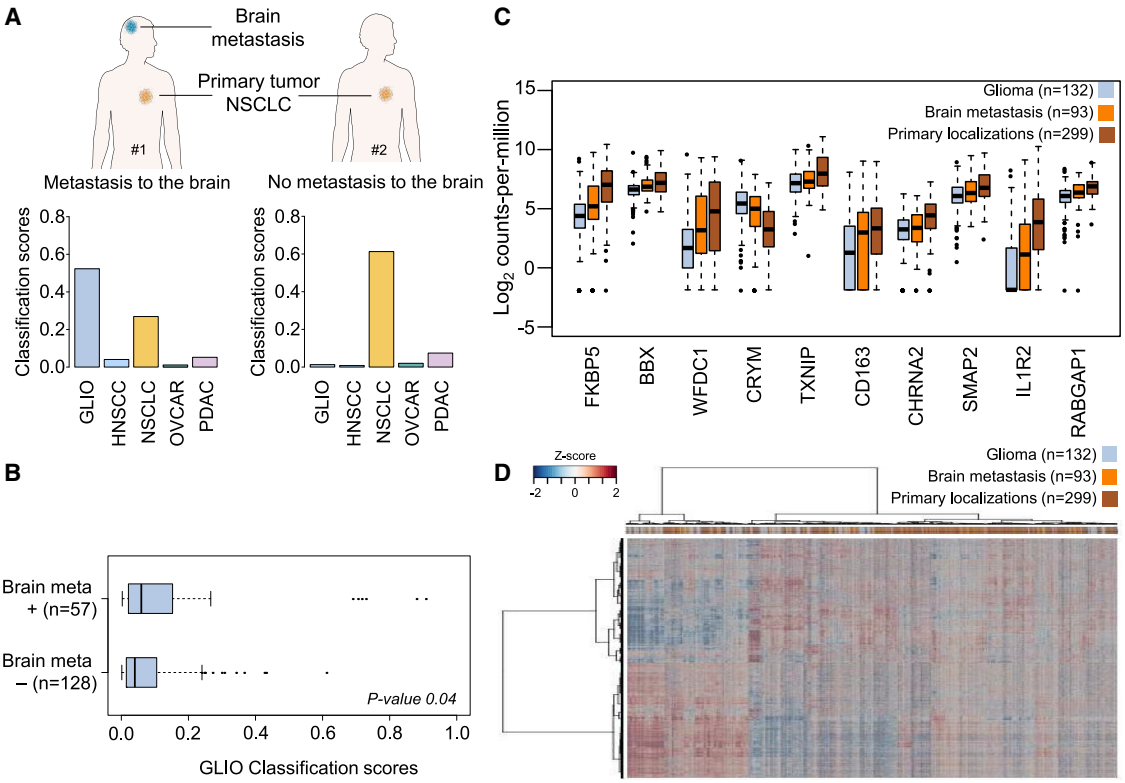


Figure 5. Brain metastatic localization may educate the TEPs

(A) Schematic representation of the metastasis analysis. Samples from two patients with stage IV metastasized NSCLC were used. Patient #1 has a brain metastasis, while patient #2 does not have a brain metastasis. Classification (probability) scores from the tumor-site-of-origin algorithm show both a high value for NSCLC and glioma in patient #1, whereas no brain-derived signal is present in the classification scores for patient #2.

(B) Boxplot of classification scores for glioblastoma (GLIO) of samples metastasizing toward the brain (+; n = 57) or not (–; n = 128).

(C) Individual boxplots of the expression values of most significantly enriched or decreased RNAs (indicated on the x axis) among patients with glioma (n = 132; blue) or brain metastasis (n = 93; orange), and the primary localizations of tumors that did have metastasis to the brain (n = 299; brown). The boxes in Figures 5B and 5C report the 25% (lower hinge), 50% (median), and 75% quantiles (upper hinge). The lower whiskers indicate the smallest observation greater than or equal to the lower hinge = 1.5 × interquartile range; the upper whiskers indicate the largest observation less than or equal to the upper hinge = 1.5 × interquartile range.

(D) Heatmap and unsupervised hierarchical clustering analysis of RNAs with differentially spliced RNA levels among patients with glioma (n = 132; blue) or brain metastasis (n = 93; orange), and the primary localizations of tumors that did have metastasis to the brain (n = 299; brown). Columns indicate samples, rows indicate RNAs, and color intensity represents the *Z*-score transformed RNA expression values. Clustering of samples showed non-random partitioning ($p < 0.00001$, Fisher's exact test). See also Figure S5.

medications, physical exercise, diets, and mental status, including a recent diagnosis of having cancer. Such potential confounding factors should be further addressed in a prospective clinical trial and should be standardized during the blood collection process.

Taken together, large-scale external validation in a dedicated, well-powered, blinded, and population-targeted prospective clinical trial of the pan-cancer thromboSeq test is required. Such trial should also investigate the added benefit of a blood test on clinical outcome parameters such as tumor stage at diagnosis and/or survival taking lead-time bias into account. Platelet RNAs may supplement other liquid biopsy biosources and biomolecules for early cancer detection.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Clinical sample collection
- METHOD DETAILS
  - Whole blood processing
  - Platelet RNA isolation, amplification and labelling for thromboSeq
  - Assessment of pre-analytical variables by transport and incubating blood tubes
  - Processing of raw RNA-sequencing data
  - Pan-cancer and tumor-site-of-origin classifier development
  - Algorithm control experiments
  - Brain metastasis analysis

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ANOVA iterative modeling
  - Post-hoc statistical modeling of potential confounding variables

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2022.08.006.

ACKNOWLEDGMENTS

We thank all the blood donors for their willingness to participate in this study. We are thankful to Krzysztof Pastuszak (Gdańsk University of Technology, Poland) for his feedback regarding data processing, the collaborators and the team of the NKI-AVL Core Facility Molecular Pathology and Biobanking (CFMPB) for supplying NKI-AVL Biobank material and lab support, the Amsterdam UMC Liquid Biopsy Center Core Facility, the Cancer Center Amsterdam Foundation, and Sebastiaan van de Sand (SIT B.V.) for computational resources. Financial support was provided by European Research Council 713727 and 336540 (T.W.), the Dutch Organisation of Scientific Research 91711366 (T.W.), Stichting STOPHersentumoren.nl (M.G. Best, N. Sol, T.W.), the Dutch Cancer Society (H.M.H., 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 Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 765492 (D.K.L., R.R., T.W.), the National Science Centre 2018/02/X/NZ5/01408 (A.S.), and the Medical University of Gdańsk statutory grant ST-23, 02-0023/07 (J.J.). In addition, this work was supported by the Netherlands Cardiovascular Research Initiative, an initiative with the support of the Dutch Heart Foundation (CVON2017-4 DOLPHIN-GENESIS, CVON2012-08 PHAEDRA; H-J.B.). Also, this study was carried out using the research infrastructure within the Netherlands Quality of Life and Biomedical Cohort Study in Head and Neck Cancer (NET-QUBIC) project funded by the Dutch Cancer Society/Alpe d'Huzes (grant VU-2013-5930). The funding bodies had no role in the design of the study or the collection, analysis, or interpretation of data nor in writing the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, S.G.J.G.I.t.V., N. Sol, M.G. Best, T.W.; funding acquisition, 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.; resources, 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.Ł-S., A.Ł., A.J.Ż., 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; formal analysis (wet lab), 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 curation, S.G.J.G.I.t.V., E.P., N. Sol M.G. Best; methodology and visualization, 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, T.W.; writing – original draft, S.G.J.G.I.t.V, M.G. Best, T.W.; writing – review & editing, all authors.

DECLARATION OF INTERESTS

M.G. Best, R.J.A.N., and T.W. are inventors on relevant patent applications (PCT/NL2011/050518 and PCT/NL2018/050110). R.J.A.N. and T.W. are shareholders of Illumina, Inc. M.H. is chief formulation officer at Nurish.Me, Inc., and Camelina Sun LLC and has equity in those companies (whose business activities are unrelated to the present work). D.M.P. and D.K.L. are shareholders of ExBiome BV.

Received: December 6, 2021
Revised: May 6, 2022
Accepted: August 8, 2022
Published: September 1, 2022

REFERENCES

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). A language and environment for statistical computing. R Found. Stat. Comput. *10*, 11–18.

Best, M.G., In 't Veld, S.G.J.G., Sol, N., and Wurdinger, T. (2019). RNA sequencing and swarm intelligence–enhanced classification algorithm development for blood-based disease diagnostics using spliced blood platelet RNA. 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). Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. 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). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell *28*, 666–676.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. 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). The complex transcriptional 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). Platelets purification is a crucial step for transcriptomic analysis. Int. J. Mol. Sci. *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). Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat. Commun. *11*, 3475.

Cho, H., Mariotto, A.B., Schwartz, L.M., Luo, J., and Woloshin, S. (2014). When do changes in cancer survival mean progress? The insight from population incidence and mortality. 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). Detection and localization of surgically resectable cancers with a multi-analyte blood test. 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 platelets. 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). Early detection and localization of multiple cancers using a blood-based methylation assay (ELSA-seq). J. Clin. Oncol. *39*, 459.

Haemmerle, M., Stone, R.L., Menter, D.G., Afshar-Kharghan, V., and Sood, A.K. (2018). The platelet lifeline to cancer: challenges and opportunities. 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). RNA-sequencing of tumor-educated platelets, a novel biomarker for blood-based sarcoma diagnostics. Cancers *12*, 1372.

Heitzer, E., Haque, I.S., Roberts, C.E.S., and Speicher, M.R. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat. Rev. Genet. *20*, 71–88.

In 't Veld, S.G.J.G., and Wurdinger, T. (2019). Tumor-educated platelets. Blood *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). Microfluidic isolation of platelet-covered circulating tumor cells. 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). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. 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). Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. 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). The sequence alignment/map format and 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). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann. Oncol. *31*, 745–759.

McAllister, S.S., and Weinberg, R.A. (2014). The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. Nat. Cell Biol. *16*, 717–727.

Nassa, G., Giurato, G., Cimmino, G., Rizzo, F., Ravo, M., Salvati, A., Nyman, T.A., Zhu, Y., Vesterlund, M., Lehtiö, J., et al. (2018). Splicing of platelet resident pre-mRNAs upon activation by physiological stimuli results in functionally relevant proteome modifications. 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). Blood platelets contain tumor-derived RNA biomarkers. Blood *118*, 3680–3683.

Pastuszak, K., Supernat, A., Best, M.G., Stokowy, T., In 't Veld, S.G.J.G., Łapińska-Szumczyk, S., Łojkowska, A., Różański, R., Żaczek, A.J., Jassem, J., and Würdinger, T. (2021). imPlatelet classifier: image-converted RNA biomarker profiles enable blood-based cancer diagnostics. 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). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-seq data. BMC Bioinf. *12*, 480.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

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). Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. Blood *118*, e101–e111.

RStudio (2015). RStudio.

Sabrkhany, S., Kuijpers, M.J.E., Griffioen, A.W., and oude Egbrink, M.G.A. (2019). Platelets: the holy grail in cancer blood biomarker research? 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). A combination of platelet features allows detection of early-stage cancer. Eur. J. Cancer *80*, 5–13.

Shen, Y., Lai, Y., Xu, D., Xu, L., Song, L., Zhou, J., Song, C., and Wang, J. (2021). Diagnosis of thyroid neoplasm using support vector machine algorithms based on platelet RNA-seq. Endocrine *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). Distinct platelet RNA signatures in patients with pulmonary hypertension. Ann. 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). Tumor-educated platelet RNA for the detection and (Pseudo)progression monitoring of glioblastoma. 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). Blood platelet RNA enables the detection of multiple sclerosis. 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). Abstract 24: multi-feature ensemble learning on cell-free dna for accurately detecting and locating cancer. Cancer Res. *81*, 24.

Tolson, B.A., and Shoemaker, C.A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. 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). Specialized and high-volume care leads to better outcomes of ovarian cancer treatment in The Netherlands. Gynecol. Oncol. *112*, 455–461.

Xing, S., Zeng, T., Xue, N., He, Y., Lai, Y.Z., Li, H.L., Huang, Q., Chen, S.L., and Liu, W.L. (2019). Development and validation of tumor-educated blood platelets integrin Alpha 2b (ITGA2B) RNA for diagnosis and prognosis of non-small-cell lung cancer through RNA-seq. Int. J. Biol. Sci. *15*, 1977–1992.

Young, R.P., Christmas, T., and Hopkins, R.J. (2018). Multi-analyte assays and early detection of common cancers. J. Thorac. Dis. *10*, S2165–S2167.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
2,351 blood platelet samples	This study	See Table S2
Chemicals, peptides, and recombinant proteins		
RNAlater stabilization solution	Ambion	cat. no. AM7020
Acid–phenol:chloroform	Ambion	
RNAseZap	Sigma-Aldrich	cat. no. R2020
Agencourt AMPure XP PCR purification system	Beckman Coulter	cat. no. A63880
Nuclease-free H_2O	Thermo Fisher	cat. no. AM9937
Critical commercial assays		
*mir*Vana miRNA Isolation Kit	Ambion	cat. no. AM1560
SMARTer Ultra Low RNA Kit for Illumina sequencing v3	Clontech Laboratories	cat. nos. 634,848–634853
TruSeq Nano DNA Library Prep kit	Illumina	cat. no. FC-121-4001
Agilent RNA 6000 Pico Kit and reagents, Bioanalyzer 2100	Agilent Technologies	cat. no. 5067–1513
Agilent High Sensitivity DNA kit and reagents, Bioanalyzer 2100	Agilent Technologies	cat. no. 5067–4626
Agilent DNA 7500 kit and reagents, Bioanalyzer 2100	Agilent Technologies	cat. no. 5067–1506
Deposited data		
Raw and processed RNA-seq data	This study	GEO: GSE183635
Software and algorithms		
Trimmomatic (version 0.22)	(Bolger et al., 2014)	http://www.usadellab.org/cms/?page=trimmomatic
STAR (version 2.3.0)	(Dobin et al., 2013)	https://github.com/alexdobin/STAR
HTSeq (version 0.6.1)	(Anders et al., 2015)	http://www-huber.embl.de/HTSeq/doc/overview.html
Picardtools (version 1.115)	Broad Institute, USA	https://broadinstitute.github.io/picard/
Samtools (version 1.115)	(Li et al., 2009)	http://samtools.sourceforge.net
Bedtools (version 2.17.0)	(Quinlan and Hall, 2010)	http://bedtools.readthedocs.io/en/latest/
MATLAB (version R2015b)	The MathWorks Inc., USA	https://nl.mathworks.com/products/matlab.html
R (version 3.3.0)	(Andy Bunn, 2017)	https://www.r-project.org
R-studio (version 0.99.902)	(RStudio, 2015)	https://www.rstudio.com
Bioconductor package edgeR (version 3.12.1)	(Robinson and Oshlack, 2010)	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Bioconductor package EDASeq (version 2.4.1)	(Risso et al., 2011)	http://bioconductor.org/packages/release/bioc/html/EDASeq.html
Bioconductor package PPSO (version 0.9–9991)	(Tolson and Shoemaker, 2007)	https://www.rforge.net/ppso/
Bioconductor package RUVSeq (version 1.4.0)	(Risso et al., 2014)	http://bioconductor.org/packages/release/bioc/html/RUVSeq.html
R-package e1071 (version 1.6–7)	CRAN	https://cran.r-project.org/web/packages/e1071/index.html
R-package Caret (version 6.0–71)	CRAN	https://cran.r-project.org/web/packages/caret/index.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R-package pROC (version 1.8)	CRAN	https://cran.r-project.org/web/packages/pROC/index.html
R-package ROCR (version 1.0–7)	CRAN	https://cran.r-project.org/web/packages/ROCR/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Thomas Wurdinger (t.wurdinger@amsterdamumc.nl).

Materials availability

This study did not generate new unique materials.

Data and code availability

- The raw sequencing data FASTQ-files are deposited in the NCBI GEO database under accession number GEO: GSE183635 and is publicly available as of the date of publication. Within this repository, a count table that served as input for the analyses is available as 'TEP_Count_Matrix.RData'.
- The code used to generate the thromboSeq algorithms including the thromboSeq dry-lab pipeline and a code reproducing the main manuscripts' figures is available via GitHub (https://github.com/MyronBest/thromboSeq_source_code_v1.5 and https://github.com/MyronBest/InTVeld_Pancancer_TSOO), is available as of the date of publication, and is for research purposes only.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Clinical sample collection

Peripheral whole blood was drawn by venipuncture from cancer patients, patients with inflammatory and other non-cancerous conditions, and asymptomatic individuals at various medical institutions in Europe and the USA. Whole blood was collected in 4-, 6-, or 10-mL EDTA-coated purple-capped BD Vacutainers containing the anti-coagulant EDTA. Cancer patients were diagnosed by clinical, radiological and pathological examination, and were confirmed to have at moment of blood collection detectable tumor load. Samples for both training, evaluation, and independent validation series were collected and processed similarly and simultaneously. Age-matching was performed retrospectively, iteratively matching samples by excluding and including patients with cancer and asymptomatic controls, aiming at a similar median age and age-range between groups. A detailed overview of the included samples, demographic characteristics, the hospital of origin, as well as an overview for which series (i.e. training, evaluation or validation) the samples were used is provided in Table S1. Asymptomatic and symptomatic controls were at the moment of blood collection, or previously, not diagnosed with cancer, but were not subjected to additional tests confirming the absence of cancer. This study was conducted in accordance with the principles of the Declaration of Helsinki. Approval for this study was obtained from the institutional review board and the ethics committee at each participating hospital. Clinical follow-up of asymptomatic controls is not available due to anonymization of these samples according to the ethical rules of the hospitals. A subset of the samples included were part of previously published studies (Best et al., 2015, 2017, 2019; Heinhuis et al., 2020; Smits et al., 2022; Sol et al., 2020a, 2020b).

METHOD DETAILS

Whole blood processing

Whole blood samples were processed using the standardized thromboSeq protocol within 48 h after blood collection, as described previously (Best et al., 2019). To isolate platelets, platelet rich plasma (PRP) was separated from nucleated blood cells by a 20-min 120×g centrifugation step, after which the platelets were pelleted by a 20-min 360×g centrifugation step. Removal of 9/10th of the platelet rich plasma was performed carefully to reduce the risk of contamination of the platelet fraction with nucleated cells, pelleted in the buffy coat. Centrifugations were performed at room temperature. Platelet pellets were carefully resuspended in RNAlater (Life Technologies) and after overnight incubation at 4°C frozen at −80°C.

Platelet RNA isolation, amplification and labelling for thromboSeq

Preparation of samples for sequencing was performed in batches, and included per batch a mixture of clinical conditions. All samples have been subjected to the identical standardized thromboSeq protocol, including SMARTer cDNA amplification. For platelet RNA isolation, frozen platelets were thawed on ice and total RNA was isolated using the mirVana miRNA isolation kit (Ambion, Thermo Scientific, cat nr. AM1560). Platelet RNA was eluated in 30 μL elution buffer. We evaluated the platelet RNA quality using the RNA 6000 Picochip (Bioanalyzer 2100, Agilent), and included as a quality standard for subsequent experiments only platelet RNA samples with a RIN-value >7 and/or distinctive rRNA curves. All Bioanalyzer 2100 quality and quantity measures were collected from the automatically generated Bioanalyzer result reports using default settings, and after critical assessment of the reference ladder (quantity, appearance, and slope). The Truseq cDNA labelling protocol for Illumina sequencing requires ~1 μg of input cDNA. To have sufficient platelet cDNA for robust RNA-seq library preparation, the samples were subjected to cDNA synthesis and amplification using the SMARTer Ultra Low RNA Kit for Illumina Sequencing v3 (Clontech, cat. nr. 634853). Prior to amplification, all samples were diluted to ~500 pg/μL total RNA and again the quality was determined and quantified using the Bioanalyzer Picochip. For samples with a stock yield below 400 pg/μL, a volume of two or more microliters of total RNA (up to ~500 pg total RNA) was used as input for the SMARTer amplification. Quality control of amplified cDNA was measured using the Bioanalyzer 2100 with DNA High Sensitivity chip (Agilent). All SMARTer cDNA synthesis and amplifications were performed together with a negative control, which was required to be negative by Bioanalyzer analysis. Samples with detectable fragments in the 300–7500 base pair (bp) region were selected for further processing. For labeling of platelet cDNA for sequencing, all amplified platelet cDNA was first subjected to nucleic acid shearing by sonication (Covaris Inc) and subsequently labeled with single index barcodes for Illumina sequencing using the Truseq Nano DNA Sample Prep Kit (Illumina, cat nr. FC-121-4001). To account for the low platelet cDNA input concentration, all bead clean-up steps were performed using a 15-min bead-cDNA binding step and a 10-cycle enrichment PCR. All other steps were according to manufacturer's protocol. Labeled platelet DNA library quality and quantity was measured using the DNA 7500 chip or DNA High Sensitivity chip (Agilent). High-quality samples with product sizes between 300-500 bp were pooled (12–19 samples per pool) in equimolar concentrations for shallow thromboSeq and submitted for 100 bp Single-Read sequencing on the Illumina Hiseq 2500 or 4000 platform using version 4 sequencing reagents. Precise and accurate quantification of the barcoded sample libraries and careful equimolar pooling is required to obtain equal total sequencing reads counts for all samples.

Assessment of pre-analytical variables by transport and incubating blood tubes

To assess the effect of several storage conditions, we designed an experiment in which blood samples were subjected to multiple environments and movements with whole blood collected in EDTA-coated tubes from asymptomatic controls and patients with stage IV non-small-cell lung cancer. Blood was collected according to the regular blood drawl procedures. Following, blood was maintained on the bench for <3 h (n = 21), <8 h (n = 14), <12 h (n = 10), <24 h (n = 44), and <48 h (n = 38), or transferred via mail for one night (24h, n = 5) or during the weekend (48h, n = 7). In the latter conditions, the samples were also subjected to irregular movement, mimicking the transfer of samples from a peripheral blood drawl location towards a central processing laboratory. Whole blood was subjected to the same platelet isolation and RNA-sequencing protocol as described above and previously (Best et al., 2019). Samples were classified in the pan-cancer thromboSeq algorithm and classification scores are reported in the boxplots.

Processing of raw RNA-sequencing data

Raw RNA-sequencing data of platelets encoded in FASTQ-files were subjected to a standardized RNA-sequencing alignment pipeline, as described previously (Best et al., 2015, 2017, 2019; Heinhuis et al., 2020; Sol et al., 2020a, 2020b) In summary, RNA-sequencing reads were subjected to trimming and clipping of sequence adapters by Trimmomatic (version 0.22) (Bolger et al., 2014), mapped to the human reference genome (hg19) using STAR (version 2.3.0) (Dobin et al., 2013), and summarized using HTSeq (version 0.6.1), which was guided by the Ensembl gene annotation version 75 (Anders et al., 2015). All subsequent statistical and analytical analyses were performed in R (version 3.3.0) and R-studio (version 0.99.902). Sample filtering was performed by assessing the library complexity, which is partially associated with the intron-spanning reads library size. First, we excluded the genes that yielded <30 intron-spanning reads in >90% of the dataset for the samples in the pan-cancer training and evaluation series. This filter step was subsequently applied to the validation series. To ensure that RNAs that are uniquely present in a certain classification group (e.g. a specific tumor type) are not excluded following this filtering step, this filter-rule was applied to each group (i.e. tumor type) separately. Next, for each sample we quantified the number of genes for which at least one intron-spanning read was mapped, and excluded samples with <1500 uniquely detected high confidence genes. To exclude platelet samples that have low intersample correlation, we performed a leave-one-sample-out cross-correlation analysis. Following data normalization, for each sample in the training and evaluation series of the dataset, all samples except the 'test sample' were used to calculate the median counts-per-million expression for each gene (reference profile). Following, the comparability of the test sample to the reference set was determined by Pearson's correlation. Samples with a correlation <0.5 were excluded. Principal component analyses were performed using the prcomp-function in R (stats-package), and the RLE-plots were generated using the plotRLE-function available in the RUVSeq-package in R. Data was corrected using the default 'perform.RUVg.correction'-algorithm from our software package (Best et al., 2019), using 'lib.size' as 'variable to assess', and a threshold of 0.8.

Pan-cancer and tumor-site-of-origin classifier development

thromboSeq classification software

For the pan-cancer thromboSeq algorithm, methods previously described were used (Best et al., 2019). In short, the algorithm employs training and evaluation series for gene panel selection and algorithm development, of which specific selection parameters are optimized by PSO. The samples in the training series served as reference samples for the iterative correction module that aims at reducing the influence of confounding factors on the dataset by RUV-normalization (Risso et al., 2014). Next, this training series is employed for gene panel selection by a likelihood ratio ANOVA test. Following, highly correlated RNAs within the preliminary biomarker panel are filtered. Next, a preliminary SVM-classification algorithm is trained and most contributive RNAs to this algorithm are identified and filtered, employing a recursive feature elimination-algorithm. Following, before building the final SVM algorithm, the *cost* and *gamma* parameters within the SVM-algorithms are optimized by a grid search. Using PSO we optimize four steps of the generic classification algorithm, i.e. (i) the iterative correction module threshold used for selection of genes identified as stable genes among the library size, (ii) the FDR-threshold included in the differential splicing filter applied to the results of the likelihood ANOVA test, (iii) the exclusion of highly correlated genes selected after the likelihood ANOVA test, and (iv) number of genes passing the recursive feature elimination algorithm. Predefined ranges were submitted to the PSO-algorithm for every classification task presented in this study. With each PSO iteration, the output of the previous iteration(s) is employed to optimize the input variables, mimicking the swarm of birds into the air at dusk. The samples assigned to the validation series did not have any influence on the platelet RNA filtering and QC-steps and the algorithm development process.

This studies optimization steps of the classification software

In this study, several optimization steps were implemented for this specific purpose of a highly specific pan-cancer algorithm; 1) for filtering of low-abundant RNAs, now each classification group in the training and evaluation series was assessed separately, in order to primarily in the multiclass tumor-site-of-origin algorithm ensure that RNAs that are enriched in especially one group are erroneously filtered out, 2) use of only the training and evaluation series as a reference group during the quality controls steps (thromboSeqQC-function), thereby ensuring that the validation series is fully independent from the analyses and algorithm training, 3) instead of passing a false discovery ratio (FDR)-threshold to the PSO algorithm, absolute gene counts were provided as determined by ANOVA-analysis, and 4) class weights were introduced in the support vector machine (SVM) training process to correct for imbalanced group sizes, especially beneficial in the pan-cancer classifier with nearly twice the number of cancer samples as compared to asymptomatic controls in the training series. The algorithm is able to process both binary comparisons (e.g. asymptomatic controls versus cancer) and multiclass comparisons (e.g. tumor-site-of-origin). In the latter process, a one-versus-one ANOVA comparison is employed.

Classification group setup and algorithm settings

Samples assignment to the training, evaluation, and validation series was performed in a stratified though random way based on the total number of samples available per tumor type, and aiming for equal distribution of the samples characteristics age, sex, tumor type and tumor stage. Preferably at least 40 samples per tumor type were included for the training and evaluation series together, however if that would result in no samples for that particular tumor type available in the validation series, less samples were assigned to the former series. The tumor types lymphoma and esophageal cancer were left out of the training and evaluation series, in order to evaluate the performance of the algorithm on tumor types not included in the training process and because too little samples were available to include them in all three series. Thereby, the algorithm was trained on 16 out of 18 tumor types. The sample IDs included in the training, evaluation and validation series, respectively, are listed in Table S2. The number of asymptomatic controls were equally separated among the series. It was aimed to obtain as much as possible age-matched series, though it should be noted that due to the inherent nature of some cancer types, some tumor types had on average younger patients as compared to the other tumor types and control groups. The swarm-variables for the pan-cancer algorithm were: 'lib.size', 'fdr', 'correlatedTranscripts', and 'rankedTranscripts. The employed boundaries were −0.1–1.0, 50 – FDR <0.005, 0.5–1.0, and respectively 50 – FDR<0.005, respectively. Training of a rule-in classifier was enabled, optimizing the training process towards highest sensitivity, at 99% specificity in the evaluation series. A coxcombplot (Figure 2B) was created using the polar coordinate system of the R-package ggplot2 (version 3.3.5).

For the five-groups tumor-site-of-origin algorithm, the tumor types with at least 100 samples in total available were included, i.e. non-small-cell lung cancer, glioma, ovarian cancer, head and neck cancer, and pancreatic cancer.

For the eleven groups tumor-site-of-origin algorithm, we decided to group anatomically closely located tumors and hematological malignancies together, resulting in larger classification groups for algorithm training and validation. The following tumor sites were grouped together, i.e. multiple myeloma plus lymphoma, prostate cancer plus renal cell carcinoma plus urothelial cell carcinoma, hepatocellular carcinoma plus cholangiocarcinoma plus pancreatic ductal adenocarcinomas, and endometrial cancer plus ovarian cancer. Patients suffering esophageal carcinomas were not included because of low sample numbers (n = 15), and males were not diagnosed with breast, endometrial or ovarian cancer.

For the tumor-site-of-origin algorithm the same swarm variables as for the pan-cancer algorithm were employed except that the FDR-value was decreased to 1×10^{-10}. For both the pan-cancer algorithm and tumor-site-of-origin algorithm 60 swarm particles were employed, with eight iterations for the pan-cancer algorithm and six iterations for the tumor-site-of-origin algorithm. All other settings were following the default settings as previously published (Best et al., 2019).

Output of the classifiers is summarized in the metrics sensitivity, specificity, area-under-the-curve from receiver-operating-curves (ROC-curves), and precision-recall curves, all with the R-package ROCR (v.1.0–7).

Algorithm control experiments

To support interpretability of the develop algorithms, multiple control experiments were performed. First as a control for internal reproducibility, we randomly sampled training and evaluation series, while maintaining the validation series and the swarm-guided gene panel of the original classifier, and perform 1000 training and classification procedures. This should ideally result into similar classification accuracies, emphasizing the correctness of the biomarker panel. Second, as a control for random classification, class labels of the samples used by the SVM-algorithm for training of the support vectors were randomly permutated, while maintaining the swarm-guided gene list of the original classifier. This process was performed 1000 times and should ideally result into diminished classification accuracies, indicating the added value of the true labels of the included samples. Third, as a control for the robustness of the 493 RNA biomarker panel, selection of new training and evaluation series using the same group-size composition and from the same pool of training and evaluation samples was performed, followed by gene panel selection and algorithm training according to the PSO-parameters set in the pan-cancer algorithm. Subsequently, an overlay was made between the resulting biomarker panel and the 493 RNA biomarker panel, and 1000 randomly selected panels from the full platelet RNA repertoire (n = 5440 RNAs). This should ideally show overlay between the true biomarker panel and the iteratively developed new biomarker panels, and very little overlay with a randomly selected biomarker panel from all platelet RNAs. P-values were calculated accordingly.

Brain metastasis analysis

To identify similarities in the platelet RNA profiles in patients who had a metastasis to the brain versus those with a glioma, all patients with metastasized disease and with a known metastasis to the brain were selected and their GLIO classification score derived from the five-groups tumor-site-of-origin-analysis was compared to those without a known metastasis to the brain at moment of blood drawl. Both groups were compared using a Student's t-test. Differential platelet RNA profiles were identified by ANOVA-statistics, employing all patients with a glioma, those with a known metastasis to the brain including the full dataset, and stage IV patients with a tumor type similar as the tumor types included in the 'brain metastasis'-group, but without a known brain metastasis. Unsupervised hierarchical clustering of heatmap row and column dendrograms was performed by Ward clustering and Pearson distances. Non-random partitioning and the corresponding p-value of unsupervised hierarchical clustering was determined using a Fisher's exact test (fisher.test-function in R), of which the most optimal threshold for RNA panel selection was optimized by PSO.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed in R (v. 3.3.0) or MATLAB (v. R2015b). Continuous data was compared using a Student's t-test. 95%-confidence intervals were calculated using binominal statistics. Gene panels were calculated using ANOVA-statistics. The output of the classifiers was summarized in the metrics sensitivity, specificity, area-under-the-curve from receiver-operating-characteristics-curves (ROC-curves), and precision-recall curves, all with the R-package ROCR. 95-% confidence intervals of the ROC-curves was calculated according to the method of Delonge using the R-package pROC. A p-value and FDR-value <0.05 was considered as statistically significant.

Statistical details of the analyses can be found in the results-section of the specific experiment, as well as the figure legend.

ANOVA iterative modeling

To investigate the stability of a biomarker panel by increasing the number of samples per condition, iterative analyses were performed. For this, all samples assigned to the training (n = 391) and evaluation series (n = 385) were included. Initially, a group of cancer samples containing one sample for each tumor type (of which at least 20 samples are available), an ANOVA comparison between these cancer samples and asymptomatic control samples was performed. Subsequently, during each iteration one sample per tumor type and a similar number of asymptomatic control samples were added to the initial dataset. This results into 40 iterations in total. Tumor types that had less than 20 samples in these series due to smaller total group sizes in this study were included according to their prevalence in the dataset. Per cancer sample added, a similar number of asymptomatic controls was added, till a maximum of 244 asymptomatic controls in these series, which is reached in iteration 20. The ANOVA comparison was performed using the default thromboSeq.ANOVA-function from our software package (Best et al., 2019), with 'lib.size' as 'variable.to.assess' (threshold: 0.8). Per ANOVA the FDR output was stored. This process was repeated 10-times with for each repeat a shuffle of the ranking of samples per tumor type. A representative heatmap is shown, with the rows including all 5440 detected RNAs sorted according to a decreasing ANOVA FDR in the final iteration. The number of RNAs with an FDR<0.05 is summarized for the 10-times repeated process in boxplots, and a trend line of the median values per iteration was fitted by the loess-function in R.

Post-hoc statistical modeling of potential confounding variables

To evaluate whether RNA-sequencing library size, age, and sex may be confounding variables in the pan-cancer thromboSeq algorithm output a linear model was employed. For these analyses, all cancer samples and asymptomatic controls of the validation series were selected. Samples with unknown patient age and/or sex status (n = 14) were excluded, resulting into 1,107 cancer samples and 120 asymptomatic controls. Linear models were created with the pan-cancer thromboSeq algorithm score as the outcome. The predictors included a fixed term for the potential confounder, a fixed term for group (cancer or control) and the interaction between the

confounder and group. Second, we estimated and visualized marginal means for the group comparisons and interactions of interest with the emmeans package (emmeans_1.6.2–1) in R. In order to estimate whether potential confounding factors, i.e. age, sex, sample supplying institution, and RNA-sequencing library size, contributed to the overall predictive value of the algorithm, a generalized linear model (GLM) was fitted that included both these factors and the algorithm score as predictors and the presence of cancer as the outcome (R-base package ''stats''; R version 3.6.1). For this, only samples from the institutes that isolated both asymptomatic controls and cancer samples were included to allow for institute correction, limiting overfitting caused by sample types originating from only one institute. The analysis included 521 cancer samples and 100 asymptomatic controls, isolated in five different institutes (i.e. VUMC, AMC, RAD, VIENNA and UMEA). With this selection, 15 different tumor types were included (i.e. breast cancer, cholangiocarcinoma, colorectal cancer, esophageal cancer, head and neck squamous cell carcinoma, lymphoma, melanoma, multiple myeloma, non-small-cell lung cancer, ovarian cancer, pancreatic ductal adenocarcinomas, prostate cancer, sarcoma, urothelial carcinoma, and glioma).