Title

Patient-specific alterations in blood plasma cfRNA profiles enable accurate classification of cancer patients and controls

Authors

Annelien Morlion (1,2), Philippe Decruyenaere (1,2,3), Kathleen Schoofs (1,2,4), Jasper Anckaert (1,2), Justine Nuytens (1,2), Eveline Vanden Eynde (1,2), Kimberly Verniers (1,2), Celine Everaert (4), Fritz Offner (3), Jo Van Dorpe (5), Olivier Thas (6,7,8), Jo Vandesompele° (1,2), Pieter Mestdagh°* (1,2)

°Both authors contributed equally to this work

*Corresponding author

(1) Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

(2) OncoRNALab, Cancer Research Institute Ghent (CRIG), Ghent, Belgium

(3) Department of Hematology, Ghent University Hospital, Ghent, Belgium

(4) TOBI Lab, Center for Medical Biotechnology, VIB-UGent, Zwijnaarde, Belgium

(5) Department of Pathology, Ghent University Hospital, Ghent, Belgium

(6) Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

(7) Data Science Institute, I-BioStat, Hasselt University, Hasselt, Belgium

(8) National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia

Author contributions

Conceptualization: AM, CE, JV, PM

Data Curation: AM

Formal analysis: AM, OT

Funding acquisition: AM, JV, JVD, PM

Investigation: EV, JN, KV

Methodology: AM, JV, PM

Project administration: AM, JV, PM

Resources: FO, JV, JVD, KS, PD, PM

Software: AM, JA

Supervision: JV, PM

Validation: AM, OT

Visualization: AM

Writing – original draft: AM, JV, PM

Writing – review & editing: all authors

Abstract

Circulating nucleic acids in blood plasma form an attractive resource to study human health and disease. Here, we applied mRNA capture sequencing of blood plasma cell-free RNA from 266 cancer patients and cancer-free controls (discovery n = 208, 25 cancer types; replication n = 58, 3 types). We observed cancer type-specific as well as pan-cancer alterations in cell-free transcriptomes compared to controls. Differentially abundant RNAs were heterogenous among patients and among cohorts, hampering identification of robust cancer biomarkers. Therefore, we developed a novel method that compares each individual cancer patient to a reference control population to identify so-called biomarker tail genes. The number of biomarker tail genes in a sample enables distinguishing an individual cancer patient from controls. The potential of this novel approach was confirmed in additional cohorts of 65 plasma donors (2 lymphoma types) and 24 urine donors (bladder cancer). Together, our findings demonstrate heterogeneity in cell-free RNA alterations among cancer patients and propose that case-specific alterations can be exploited for classification purposes.

Introduction

Cancer is a complex disease with considerable heterogeneity, both between as well as within cancer types. Finding robust and easily accessible biomarkers is crucial to improve patient well-being and outcome. Such biomarkers can either assist in early detection of cancer, enabling treatment to start at an earlier stage, or they can support diagnosis, treatment selection or response monitoring. As an alternative for biomarkers that require invasive tissue biopsies, liquid biopsy-based biomarkers such as circulating nucleic acids have emerged as a promising tool for cancer diagnosis and monitoring.

60 To date, most studies on circulating nucleic acids in cancer patients have focused on
61 circulating tumor DNA (ctDNA). The amount of tumor DNA in circulation is highly dependent
62 on the tumor type and burden[1], suggesting that ctDNA can be less informative for certain
63 cancer types or early-stage cancers. Cell-free RNA (cfRNA) may complement ctDNA by
64 reflecting dynamic changes in gene expression during cancer development and progression,
65 or upon treatment, in both healthy and diseased cells. A recent study showed that different
66 cell types contribute RNA to the healthy cell-free plasma transcriptome, and that cfRNA
67 allows monitoring of cell-type-specific changes in diseases, such as non-alcoholic fatty liver
68 and Alzheimer's disease[2]. The potential of cfRNA as prognostic biomarker has also been
69 demonstrated for pre-eclampsia and organ damage in pregnancy[3]. Some studies have also
70 shown potential of using cfRNA biomarkers in certain cancer types[4–6]. Larson et al.[4]
71 specifically looked at genes that are "dark" (no or few detected transcripts) in plasma from
72 healthy controls but abundant in plasma from lung and/or breast cancer patients and found
73 indications of tissue-of-origin signal. Chen et al.[5] reported potential of combining human and
74 microbe-derived plasma cfRNA for distinguishing healthy donor and cancer plasma samples
75 (colorectum, stomach, liver, lung, and esophageal cancer) based on profiles of RNAs showing
76 group-level differences. Roskams-Hieter et al.[6] showed plasma cfRNA classification potential
77 for liver cancer and multiple myeloma patient samples compared to non-cancer controls, with
78 a higher accuracy using learning vector quantization feature selection compared to group-
79 level differential abundance. The authors, however, indicated that liver and bone marrow are
80 major contributors to the plasma transcriptome which may result in more cancer signal.
81 Here, we further examined the impact of different cancer types on plasma cfRNA profiles in
82 different cohorts (Table 1). We first quantified cell-free mRNAs in different cohorts of blood
83 plasma from cancer patients with locally advanced or metastatic tumors and cancer-free
84 donors to uncover tumor or tissue-of-origin mRNA signatures with diagnostic potential.
85 Results revealed substantial heterogeneity in cell-free mRNA abundance among cohorts,
86 questioning the biomarker potential of cell-free mRNA signatures. However, patient-specific
87 alterations, reflected by the number of cell-free mRNAs with differential abundance in a
88 patient sample compared to a cancer-free control population, enabled to accurately
89 distinguish cancer patients from healthy individuals.

90 ***Table 1:*** *Cohort overview. Name, number of donors (one sample per donor), biofluid type, library preparation*
91 *method and application of the cohort in this study. (*) eight control donors overlap with pan-cancer cohort, data*
92 *from these eight donors was only used for discovery of the tail gene concept, not for replication of pan-cancer*
93 *analyses (see Methods).*

cohort name	cancer patients (cancer types)	controls	biofluid	library preparation	purpose
pan-cancer cohort	200 (25)	8	plasma	mRNA capture (Illumina DNA Prep with Enrichment)	discovery
three-cancer cohort	36 (3)	30*	plasma	mRNA capture (Illumina DNA Prep with Enrichment)	replication pan-cancer analyses & discovery tail gene concept
lymphoma cohort	43 (2)	22	plasma	total RNA (SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian)	tail gene concept in independent plasma cohort
bladder cancer cohort	12 (1)	12	urine	mRNA capture (TruSeq RNA Exome)	tail gene concept in independent urine cohort

94

95 # Results

96 ## A diverse repertoire of cell-free mRNAs in plasma

97 To characterize the blood plasma cell-free mRNA transcriptome, we collected blood plasma
98 from a pan-cancer discovery cohort consisting of 208 individuals covering 25 different cancer
99 types and aged-matched healthy controls (Fig.1a, Supplementary Table 1). MRNA capture
100 sequencing of the pan-cancer plasma cohort resulted in 3 to 56M paired-end reads per
101 sample (median 16M) after quality filtering with no significant differences among groups
102 (Kruskal-Wallis test $p > 0.05$) (Extended Data Fig.1a). On average 84% (95% CI [82.59, 84.66])
103 of uniquely mapped reads were assigned to protein coding genes and most other reads
104 originated from mitochondrially encoded ribosomal RNAs (Extended Data Fig.1b). Per sample,
105 between 7439 and 13,039 unique mRNAs (median 10,932) were detected with at least 10
106 counts – we use the term mRNA to refer to all mRNA reads mapping to the same gene (Fig.1b).
107 All plasma samples had 5263 mRNAs in common and 9908 mRNAs were found in at least 75%
108 of the samples. Of note, we detected more mRNAs in plasma from acute myeloid leukemia
109 (AML) patients compared to plasma from controls and from patients with solid cancers
110 (Kruskal-Wallis $p = 6.6E-4$, effect size = 0.163, *post hoc* two-sided Wilcoxon rank-sum $q < 0.05$
111 for AML vs controls and 16 out of 24 solid cancer types). Although some AML samples were
112 sequenced deeper, sequencing depth alone did not explain the higher number of detected
113 mRNAs (Extended Data Fig.1c). We next determined the mRNA concentration using synthetic

114 RNA spikes that were added to each lysed plasma sample (Fig.1c). MRNA concentrations
115 ranged from 0.003 to 8.302 ng/ml with an average concentration of 0.486 ng/ml. Despite the
116 higher number of detected mRNAs, the mRNA-concentration in AML plasma was not
117 significantly higher than controls or other cancer types. Moreover, though the largest median
118 mRNA concentration (READ, 0.648 ng/ml) was more than 10 times higher than the smallest
119 median (ANA, 0.063 ng/ml), no significant concentration differences were found among the
120 groups (Kruskal-Wallis test $p = 0.0275$, effect size = 0.087, with all *post hoc* two-sided
121 Wilcoxon rank-sum $q > 0.05$). Besides the small sample sizes, this may be related to the
122 considerable cfRNA concentration variability between patients, which is also reported at
123 cfDNA level[7].
124 We then searched for evidence of tumor-derived cfRNA by investigating known cancer fusion
125 transcripts in individual plasma samples. In every cancer plasma sample, we only considered
126 fusion transcripts that had been documented in tumor tissue of that same cancer type. Such
127 fusion transcripts were only identified in plasma samples of AML patients. The *PML::RARA*
128 fusion transcript[8] was found in two AML plasma samples with several reads supporting the
129 fusion (7 to 19 spanning pairs and 6 to 9 unique spanning reads). The reciprocal *RARA::PML*
130 fusion was found in one of both AML plasma samples (19 pairs and 17 unique reads). In
131 plasma samples of solid cancer types, we only detected fusion transcripts that had not been
132 reported in the tumor tissue of these cancer types (Supplementary Table 2). Of note, these
133 fusion transcripts had fewer supporting reads (1 to 2 pairs and 2 to 5 unique reads) than the
134 *PML::RARA*/*RARA::PML* fusions in AML and may thus represent false positives. Moreover,
135 other fusions with few supporting reads were also identified in healthy control plasma
136 samples. While the two AML samples with the *PML::RARA* fusion transcripts were sequenced
137 deeper than other samples, the fusions were still detected when downsampling to the
138 median number of reads of the pan-cancer cohort. These analyses confirm the presence of
139 tumor cell-derived cfRNA in plasma samples from AML patients.

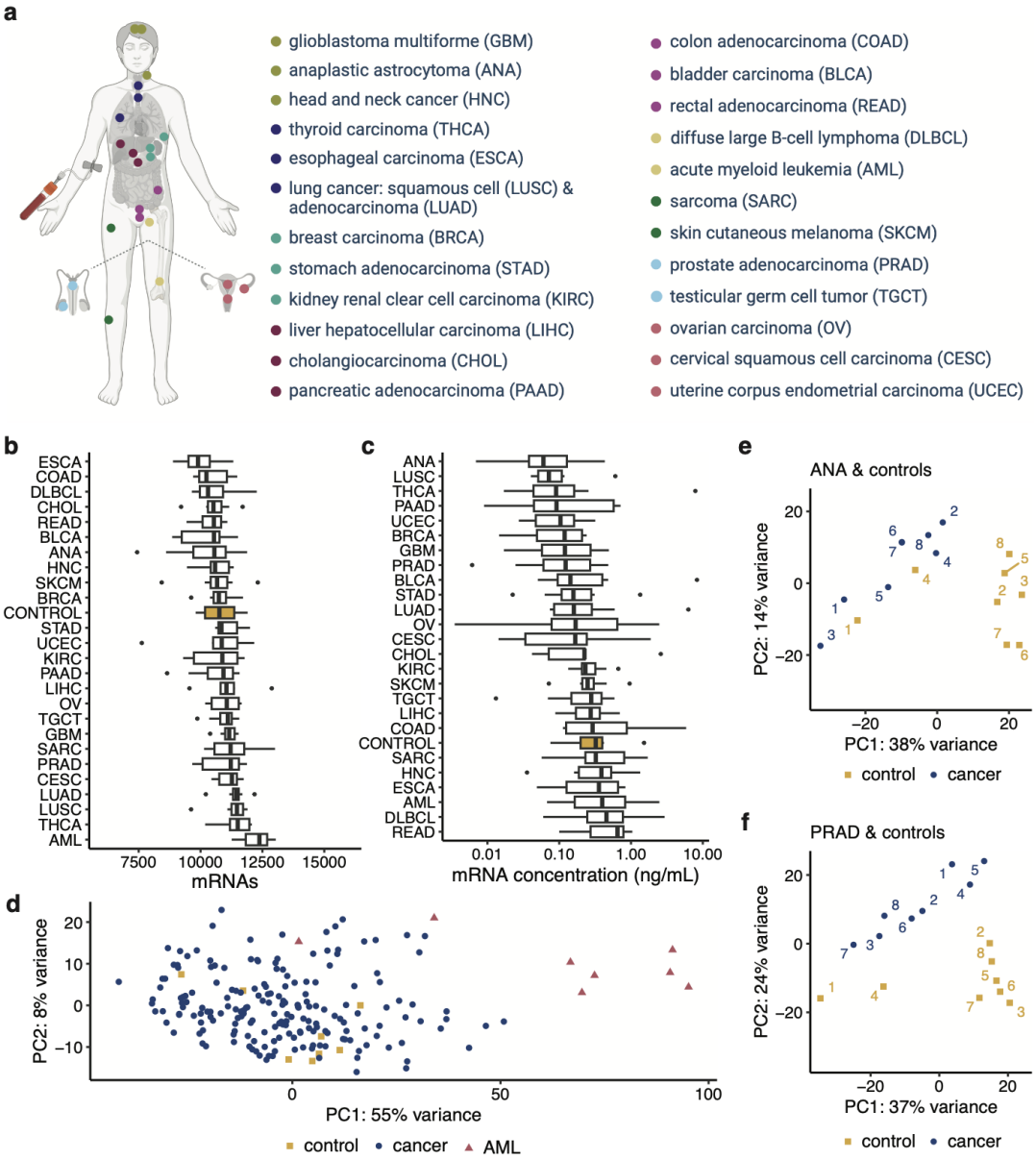


Fig.1: Pan-cancer plasma cohort overview, mRNA content and sample clustering. ***a***, *Primary sites of the 25 cancer types in the pan-cancer cohort. Primary sites are indicated with colored dots and corresponding cancer types are listed (top to bottom) with the (sub)type abbreviations between brackets. Details can be found in Supplementary Table 1.* ***b***, *Number of plasma mRNAs, detected with at least 10 counts, across cancer types and healthy controls. Groups ordered according to median.* ***c***, *Variability in plasma mRNA concentration across cancer types and healthy controls. Groups ordered according to median.* ***b*** *and* ***c***, *Boxplots show lower quartile (Q1), median, and upper quartile (Q3). Whiskers extend from the lower and upper quartile to the smallest and largest value, respectively, within at most 1.5*interquartile range (Q3-Q1) from that quartile. More extreme points are plotted as individual dots.* ***d***, *Principal component analysis using the 500 most variable mRNAs in the pan-cancer cohort (with variance stabilizing transformation DESeq2). Yellow squares: control samples; red triangles: acute myeloid leukemia samples; blue dots: non-leukemia cancer samples.* ***e*** *and* ***f***, *Principal component analysis using the 500 most variable mRNAs in controls and anaplastic astrocytoma (ANA) or prostate cancer (PRAD) patients, respectively. Replicate numbers are indicated in the plot. PC: principal component.*

154 **Differential mRNA abundance in cancer patients' plasma compared to controls**

155 Principal component analysis revealed a clear separation between AML and other plasma
156 samples for six out of eight AML patients, with and without downsampling AML sample reads
157 to the median of the cohort (Fig.1d & Extended Data Fig.1d). Control samples mostly clustered
158 in between samples of solid tumors. Analysis of individual solid tumor types revealed that
159 most patient samples could be separated from controls in the first two principal components
160 (Fig.1e and f, Extended Data Fig.2), suggesting that plasma cfRNA profiles of cancer patients
161 differ from that of controls.

162 Differentially abundant mRNAs were identified in every cancer-control comparison, with
163 numbers varying from 10 to 2583 (median 128) for higher abundant mRNAs and from 89 to
164 2727 (median 221) for lower abundant mRNAs in cancer compared to control (Fig.2a). Most
165 differentially abundant mRNAs were found in AML (5310 mRNAs, or 5347 mRNAs when
166 downsampling reads to the median reads in cohort). Uterine cancer (UCEC) and sarcoma
167 (SARC) completed the top 3 with 1154 and 763 mRNAs, respectively. Head and neck (HNC),
168 rectal (READ) and breast cancer (BRCA) had the lowest number of differentially abundant
169 mRNAs (127, 136, and 145 respectively).

170 In total, 6758 unique mRNAs were differentially abundant in one or more cancer-control
171 comparisons (b) and the median absolute fold change was 3.1. One quarter of differentially
172 abundant mRNAs had an absolute fold change of at least 4.6. Hierarchical clustering based on
173 differential mRNA abundance mostly separated cancer and control samples but also revealed
174 variability in gene abundance within groups (Fig.2b, Extended Data Fig.3).

175 4442 mRNAs (66%) were differentially abundant in one cancer type only (Fig.2c), and most
176 (3680 mRNAs) were identified in AML. While many mRNAs were differentially abundant in
177 plasma of only one or a few cancer types, 26 differentially abundant mRNAs were recurrent
178 in 23 out of 25 cancer-control comparisons (Fig.2c). These mRNAs were all lower abundant in
179 cancer compared to controls and showed enrichment for inflammatory and immune response
180 pathways (Supplementary Table 3).

181

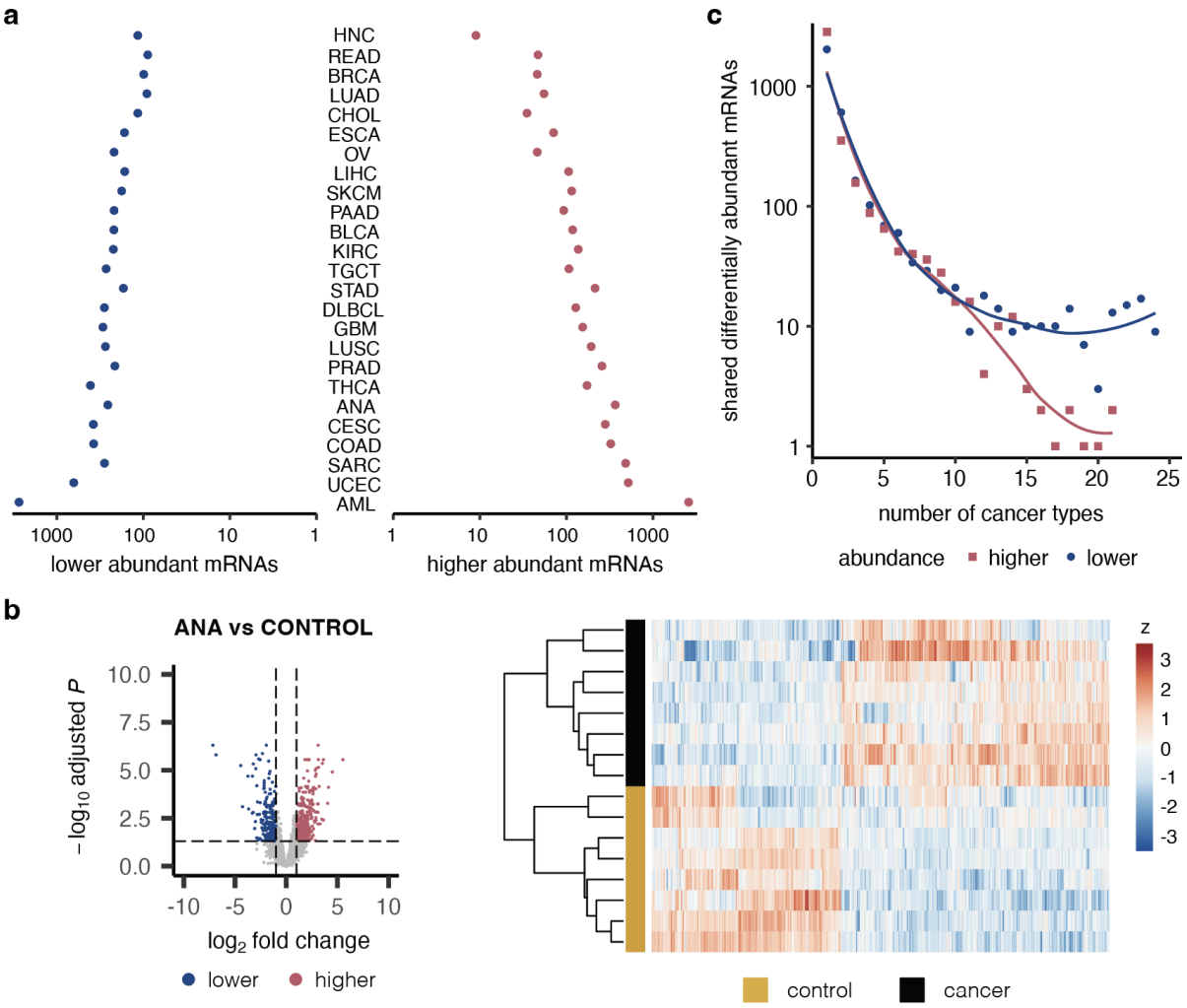


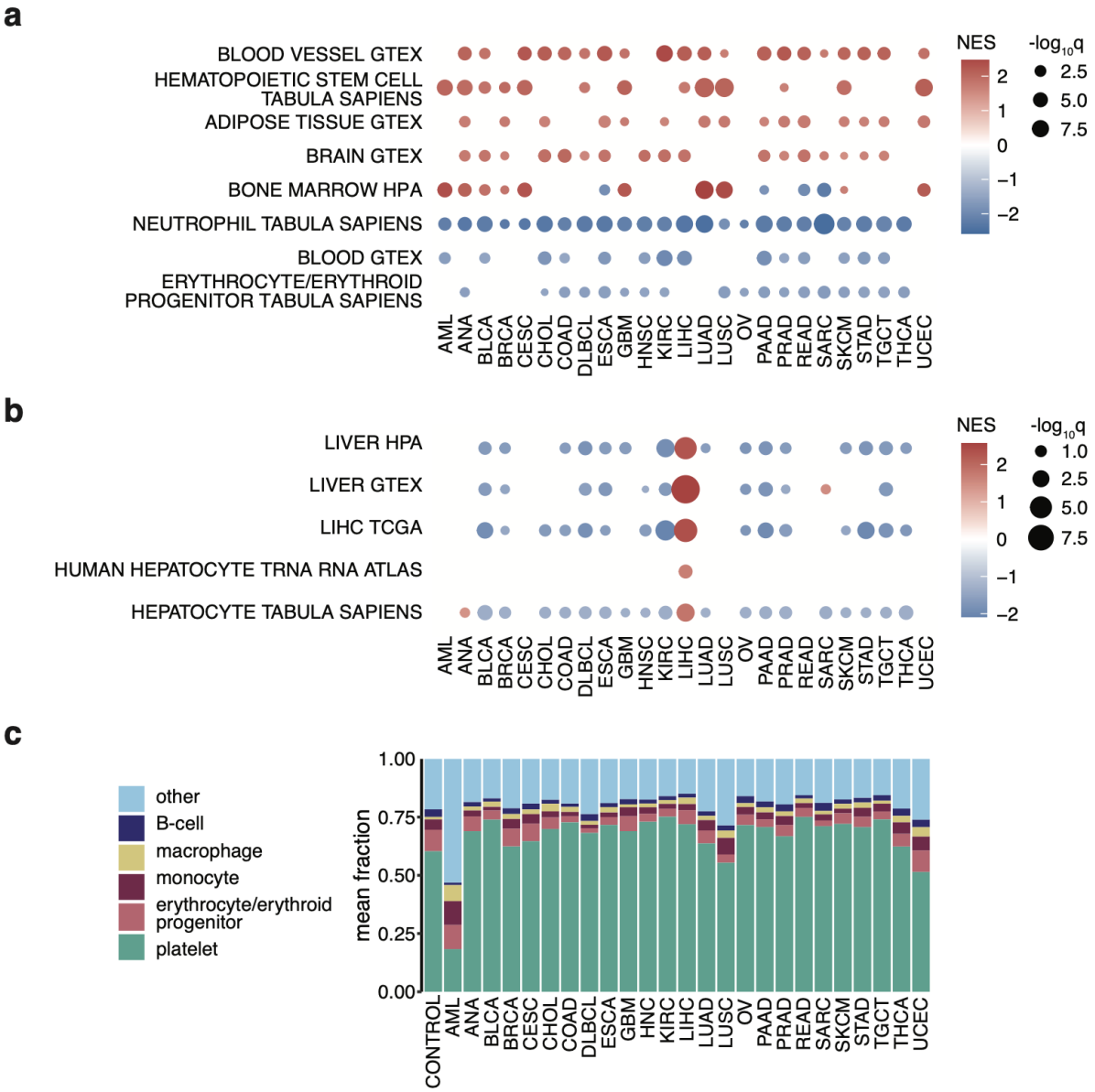
***Fig.2: Differentially abundant mRNAs in cancer versus control**. **a,** Differentially abundant mRNAs in plasma from cancer patients versus controls. Cancer types ordered based on total number of differentially abundant mRNAs ($q < 0.05$ and $|\log_2 \text{fold change}| > 1$). **b,** Volcano plot and heatmap of differentially abundant mRNAs in anaplastic astrocytoma vs controls (other cancer-control comparisons in Extended Data Fig.3). Volcano plots have a horizontal line at $q = 0.05$ and vertical lines at $\log_2$ fold change of -1 and 1, resp. Higher (red): higher abundant in cancer compared to control samples ($q < 0.05$ and $\log_2$ fold change > 1); lower (blue): lower abundant in cancer compared to control samples ($q < 0.05$ and $\log_2$ fold change < -1). Heatmaps show relative abundance of differentially abundant mRNAs (columns) in individual plasma samples (rows). Heatmap colors represent log2(normalized counts+1) followed by z-score transformation per gene (z). Clustering of genes and samples based on Pearson correlation. **c,** Recurrence of differentially abundant mRNAs (shown in a) across cancer types.*

Enrichment of pathways and cell types in cancer patients' plasma

To explore the pathways, cell types and tissues underlying the differentially abundant mRNAs in cancer plasma, we first applied gene set enrichment analyses using hallmark[9], Kyoto Encyclopedia of Genes and Genomes (KEGG)[10], and custom tissue- and cell type gene sets (Fig.3a, Extended Data Fig.4a & b). We observed a negative enrichment of various immune-

198 related signals (interleukin signaling, interferon gamma and inflammatory response, and
199 neutrophils) across all cancer types except UCEC, with q < 0.05 in at least 21 out of 25 cancer
200 types. We also found a negative enrichment of the KEGG ribosome gene set in all solid cancer
201 types (q < 0.05 in 19 out of 24 types) (Extended Data Fig.4b). These results are in line with a
202 previous study by Chen et al.[5] on plasma cfRNA profiles of five early-stage solid cancer types.
203 Other negatively enriched gene sets with q < 0.05 in at least half of the cancer types in our
204 cohort were blood related (Fig.3a, Extended Data Fig.4a & b). Conversely, 'extracellular matrix
205 receptor' and 'focal adhesion' gene sets were significantly positively enriched in at least half
206 of the cancer types (Extended Data Fig.4b), confirming results by Chen et al. Other positively
207 enriched gene sets included those related to KRAS signaling, epithelial-mesenchymal
208 transition, myogenesis, adipose tissue, brain function, and hematopoietic stem cells. (Fig.3a,
209 Extended Data Fig.4a).
210 Cancer type-specific enrichment of cfRNA reflecting the primary tumor site was only observed
211 for liver cancer patients. 'Liver tissue', 'liver cancer tissue', and 'hepatocyte' gene sets were
212 all positively enriched in liver cancer patients' plasma (q < 0.05, Fig.3b) and not in plasma from
213 any other cancer type. A 'bronchial epithelial cells' gene set was significantly enriched in lung
214 adenocarcinoma and lung squamous cell carcinoma (LUAD q = 0.027; LUSC q = 0.026), but this
215 enrichment was also found in four other cancer types (ANA, CESC, CHOL, and UCEC).
216 We then applied computational deconvolution to enumerate the cell type fractions
217 contributing to plasma cfRNA profiles (Supplementary Table 4). Platelets,
218 erythrocytes/erythrocyte progenitors, monocytes, macrophages, and B-cells formed the top-
219 5 of most abundant cell types that contributed to plasma cfRNA of both controls and patients
220 with solid tumors (Fig.3c). Except for macrophages, these cell types were also reported as the
221 most abundant cell types in a previous study on healthy donor plasma[2]. Plasma from AML
222 patients showed a very different composition. Several cell type contributions - including
223 hematopoietic stem cells, myeloid progenitor cells, and monocytes - were increased or even
224 exclusively present in AML plasma samples, in line with the blood cell counts typically
225 observed in these patients[11]. Myeloid progenitor cells, normally not present in peripheral
226 blood, were an abundant contributor in the AML plasma samples (mean 10.7%, or 8.5% when
227 downsampling AML reads to the cohort median), with almost no signal in other plasma
228 samples (mean 0.1%) (Extended Data Fig.4c). Moreover, the platelet fraction in AML plasma
229 was significantly lower compared to other groups (two-sided Wilcoxon rank sum p = 3.8E-5,

230 effect size = 0.289) with on average 18.3% in AML compared to 62.3% in other samples,
231 reflecting the reduced thrombocyte counts commonly found in leukemia patients[11].
232 In plasma from patients with solid tumors, the platelet fraction was significantly higher
233 compared to controls (Extended Data Fig.4c, two-sided Wilcoxon rank-sum p = 2.3E-2, effect
234 size = 0.163), which is in line with the prognostic value of platelet counts and the frequent
235 occurrence of thrombocytosis in solid tumor patients[12,13]. For neutrophils,
236 erythrocyte/erythroid progenitor cells, thymocytes, and T-cells, the fraction was significantly
237 lower in cancer patients' plasma compared to controls (two-sided Wilcoxon rank-sum p =
238 1.8E-2 (effect size = 0.169), p = 1.4E-2 (effect size=0.176), p = 1.8E-2 (effect size=0.164), p =
239 4.3E-4 (effect size = 0.252), respectively).



240

241 ***Fig.3: Gene set enrichment and cell type deconvolution reveal pan-cancer and cancer type-specific signals**. **a**,*
242 *Gene set enrichment analysis results for custom tissue and cell type sets. Only sets that are significantly enriched*
243 *in at least half of the cancer types are shown. Gene sets with q < 0.05 colored according to normalized enrichment*
244 *scores (NES).* ***b****, Liver-related gene set enrichment in plasma from liver cancer patients, gene sets with q < 0.25*
245 *colored according to NES. Detailed information about gene sets can be found in Methods.* ***c****, Average fraction of*
246 *the five most abundant cell types based on deconvolution. Other: cumulative sum of other cell types. The five*
247 *most abundant cell types were selected based on mean ranking in all groups.*

248 **Independent cohort confirms systemic signal and cfRNA heterogeneity**

249 To verify our findings, we processed an independent (three-cancer) cohort of plasma samples
250 which, after removing samples with low read depth, included samples from 20 cancer-free
251 control donors and 35 cancer patients with one of three different solid cancer types: prostate
252 (PRAD, n = 12), ovarian (OV, n = 11), and uterine (UCEC, n = 12) cancer. 7 to 61M paired-end
253 reads were obtained per sample (median 28M) with no significant difference in sequencing
254 depth among groups (Kruskal-Wallis p = 0.7858) (Extended Data Fig.5a). In line with the pan-
255 cancer cohort, we detected between 8642 and 11,944 mRNAs per sample (Extended Data
256 Fig.5b), with 6194 mRNAs shared across all plasma samples of this three-cancer cohort. MRNA
257 concentrations ranged from 0.007 to 0.414 ng/ml (Extended Data Fig.5c).
258 Most cancer samples could be separated from controls in the first principal component,
259 especially for prostate and ovarian cancer (Fig.4a). Based on differential abundance analyses
260 between cancer and control samples, we found the highest number of differentially abundant
261 mRNAs in OV (995 higher abundant, 1644 lower abundant), followed by PRAD (848 higher
262 abundant, 1043 lower abundant) (Fig.4b). The lowest number of differentially abundant
263 mRNAs was found for UCEC (243 higher abundant, 349 lower abundant) even though this
264 cancer type had the most differentially abundant mRNAs in the pan-cancer cohort.
265 Hierarchical clustering based on differential mRNA abundance in this three-cancer cohort
266 mostly separated cancer and control samples but again revealed variability in mRNA
267 abundance within groups (Extended Data Fig.5d).
268 189 and 299 mRNAs were significantly higher and lower abundant, respectively, in plasma
269 from all three cancer types compared to controls (Fig.4b). The lower abundant mRNAs
270 included all 26 mRNAs that were identified as lower abundant in most cancers in the pan-
271 cancer cohort. There was a significant, although limited, overlap in differentially abundant
272 mRNAs between the three-cancer and the pan-cancer cohort for each of the three cancer

273 types (Fisher's exact test on lower abundant mRNAs: OV p = 8.4E-118 (odds ratio = 30.0),
274 PRAD p = 1.1E-155 (odds ratio = 59.3), UCEC p = 7.6E-13 (odds ratio = 4.2); on higher abundant
275 mRNAs: OV p = 4.0E-4 (odds ratio = 4.6), PRAD p = 2.9E-52 (odds ratio = 12.2), UCEC p = 2.2E-
276 5 (odds ratio = 3.3); Jaccard index [0.009-0.143]) (Fig.4c and Extended Data Fig.6a). Despite
277 the uniform collection and processing of samples from patients with the same cancer type,
278 the low Jaccard indices indicate high heterogeneity between the two cohorts, as illustrated
279 with probabilistic modelling for varying cohort sizes (Supplementary Analysis 1). Respectively
280 38%, 49%, and 95% of the differentially abundant mRNAs in OV, PRAD, and UCEC in the pan-
281 cancer cohort could not be confirmed in the three-cancer cohort. Notably, mRNAs with lower
282 abundance showed a higher overlap between both cohorts than mRNAs with higher
283 abundance.
284 In line with the pan-cancer cohort, there was no enrichment of cancer type-specific gene sets
285 reflecting tumor tissue of origin in the OV, PRAD and UCEC samples in the three-cancer
286 cohort. However, enriched gene sets shared among at least half of the cancer types in the
287 pan-cancer cohort were confirmed in the three-cancer cohort, except for adipose tissue
288 (Fig.4d, Extended Data Fig.6b and c).
289 Cell type deconvolution confirmed platelets, erythrocyte/erythroid progenitor cells,
290 monocytes and B-cells among the top-5 cell types contributing cfRNA to plasma (Fig.4e,
291 Supplementary Table 4), with neutrophils being ranked as the fourth most abundant cell type.
292 The higher platelet fraction in cancer patient plasma was even more pronounced in the three-
293 cancer cohort (two-sided Wilcoxon rank-sum cancer versus control p = 5.8E-9, effect size =
294 0.679), especially for ovarian and prostate cancer (Extended Data Fig.6d). Neutrophils,
295 erythrocyte/erythroid progenitor cells, thymocytes and T-cells again had a lower plasma
296 fraction in cancer patients compared to controls (two-sided Wilcoxon rank-sum p=4.4E-6
297 (effect size = 0.556), p = 3.2E-7 (effect size = 0.610), p = 6.4E-5 (effect size = 0.419), p = 4.1E-
298 2 (effect size = 0.261), respectively).
299 Taken together, the limited overlap of differentially abundant mRNAs between the pan-
300 cancer and three-cancer cohort highlights plasma transcriptome variability, yet enrichment
301 and cell type contribution analyses confirm the systemic cfRNA changes in plasma from
302 cancer patients.

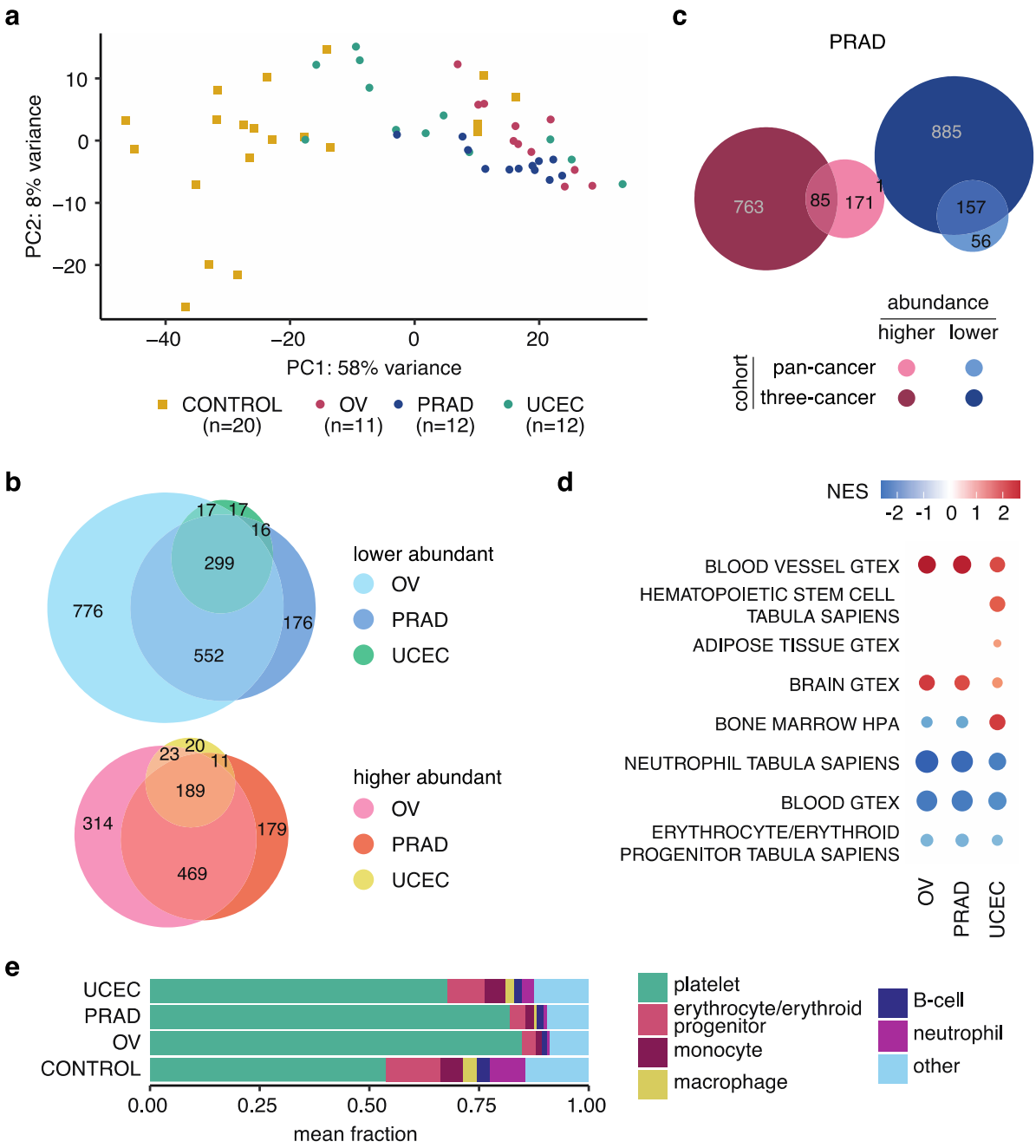


Fig.4: Three-cancer cohort confirms findings from pan-cancer cohort and reveals significant, yet limited, overlap in differentially abundant mRNAs. a*, Principal component analysis on top 500 most variable mRNAs shows distinct clustering of most cancer and control samples (variance stabilizing transformation DESeq2).* ***b****, Overlap between differentially abundant mRNAs detected in individual cancer-control comparisons. Circles are proportional to the number of genes.* ***c****, Overlap between differentially abundant genes in prostate cancer versus control for the pan-cancer and three-cancer cohort. Fisher's exact test: p = 2.9E-52, odds ratio = 12.2, Jaccard index = 0.083 for higher abundant mRNAs in cancer versus control (higher); p = 1.1E-155, odds ratio = 59.3, Jaccard index = 0.143 for lower abundant mRNAs (lower). Circles are proportional to the number of genes.* ***d****, Gene set enrichment analysis results for custom tissue and cell type sets. Only pathways of Fig.3a shown. Gene sets with q < 0.05 colored according to normalized enrichment scores (NES).* ***e****, Average fraction of the five most*

314 *abundant cell types and macrophages based on cell type deconvolution. Other: cumulative sum of other cell*
315 *types. Differential abundance: $q < 0.05$ & $|\log_2 \text{fold change}| > 1$. OV: ovarian cancer; PRAD: prostate cancer;*
316 *UCEC: uterine cancer.*

317 **cfRNA tail genes distinguish cancer samples from controls**

318 Apart from liver cancer and AML, our analyses did not reveal evidence for a tumor or tissue
319 of origin signal among mRNAs that were more abundant in cancer patient versus control
320 plasma. Moreover, higher abundant mRNAs were less recurrent across cancer types and
321 showed lower reproducibility between cohorts compared to mRNAs with lower abundance
322 in cancer versus control plasma. Our analyses suggest that the latter represent a systemic
323 immune response rather than a tumor derived signal. Based on these observations, we
324 hypothesized that (tumor derived) cfRNA signals may be heterogeneous among patients, and
325 that classic differential mRNA and pathway abundance analysis between groups may not
326 reveal such non-uniform signals. We therefore decided to investigate cfRNA profiles in
327 individual cancer patients using samples from the three-cancer cohort.
328 To capture aberrations in the cfRNA profile of an individual cancer patient, we compared the
329 mRNA abundance of a cancer plasma sample to the mRNA abundance distribution in the
330 entire control group, that served as a reference. More specifically, a z-score was calculated
331 for each mRNA based on the mean and standard deviation of the $\log_2$ normalized count
332 distribution in the control group. Abundant mRNAs with an absolute z-score of at least 3 were
333 defined as "tail genes" for that patient, referring to their position in one of the tails of the
334 control z-score distribution (Fig.5a).
335 We searched for tail genes in each cancer patient from the three-cancer cohort (where we
336 could take advantage of a large control group) and in each control sample, always using all
337 other control samples as reference (Extended Data Fig.7a & b). We identified a total of 3311
338 unique tail genes in this cohort and observed a significantly higher number of tail genes in
339 plasma samples from cancer patients compared to controls (Kruskal-Wallis $p = 0.0012$, effect
340 size = 0.222, Fig.5b). 74% of the tail genes were not identified as differentially abundant in
341 cancer samples versus controls (Fig.5c). Moreover, for 40.0% of cancer sample tail genes that
342 were differentially abundant between cancer and controls, the direction (i.e. positive or
343 negative z-score) in at least one cancer sample was opposite to the direction reported by the
344 differential abundance analysis, further highlighting the heterogeneity among patients.
345 Notably, the recurrence of tail genes was much higher in cancer samples than controls; 98%

346 of the control tail genes were identified in only one control sample while 47% of cancer tail
347 genes were identified in more than one cancer sample (Fig.5d).
348 To determine the significance of the observed results, we randomly swapped sample labels
349 prior to tail gene identification and repeated this procedure 20 times. Doing so, we did not
350 observe a significant difference in number of tail genes between the newly labeled cancer
351 samples and controls in any of the repeats (all Kruskal-Wallis $p \geq 0.05$), confirming that an
352 increased number of tail genes is specifically associated with cancer samples.
353 Tail genes detected in one cancer type showed significant overlap with tail genes detected in
354 another cancer type (Fisher's exact tests $p \leq 5.4E-168$, odds ratio: [11.1-15.2]), yet limited
355 similarity (Jaccard index: [0.197-0.348]) (Fig.5e). The 1182 unique control tail genes also
356 showed significant overlap with any of the cancer tail gene subsets (pairwise Fisher exact tests
357 p: [5.9E-15 - 8.5E-32], odds ratio: [2.1-2.5]) with less similarity (Jaccard index: [0.074-0.119]).
358 63% of control tail genes were not identified as tail gene in any of the cancer samples.
359 Seven Cancer Gene Census oncogenes were identified as tail genes ($z > 3$) in at least one
360 sample, which is significantly more than expected by chance (Fisher's exact $p = 0.019$, odds
361 ratio = 3.13). While most oncogenes were only identified as tail gene in a single sample, *KAT6A*
362 was found in seven cancer samples (4 OV, 2 PRAD, and 1 UCEC) but also in one control sample.
363 Overrepresentation analyses (Extended Data Fig.8, Supplementary Table 5) demonstrated
364 that tail genes that are higher abundant in at least one cancer sample ($z > 3$) were involved
365 in different processes including platelet activation and signaling, actin binding and dynamics,
366 and extracellular matrix. Tail genes that are lower abundant in at least one cancer sample (z
367 < -3) were mainly related to mRNA splicing and translation. Significantly enriched processes
368 in the cancer tail gene sets were not enriched in the control tail gene set, and processes
369 enriched in the lower abundant (cancer) tail genes showed higher recurrence across the three
370 cancer types compared to those enriched in the higher abundant tail genes.
371 Given the higher number of tail genes in cancer versus controls, we further assessed the
372 binary (cancer/control) classification potential of the number of tail genes per sample. We
373 used a 10x 5-fold cross-validation in which we first selected a subset of tail genes, further
374 coined biomarker tail genes (BTG), that were significantly associated to either the cancer or
375 control state using Fisher's exact tests on the training set (see Methods). This resulted in sets
376 of 43 to 549 BTG. None of these BTG were recurrent tail genes in the control samples, and for
377 BTG identified in at least half of the cross-validation folds, 43% were not identified as

378 differentially abundant in any cancer type compared to controls (Extended Data Fig.9).
379 Applying the same biomarker tail gene selection strategy to 10 iterations of randomly
380 swapping sample labels resulted in considerably fewer or no biomarker tail genes at all:
381 between 0 and 6 BTG (median: 0) instead of 43 to 549.
382 When selecting biomarker tail genes in a specific cancer group versus controls, we obtained
383 17 to 271 biomarker tail genes for OV, 6 to 82 for PRAD, and 2 to 16 for UCEC. We then built
384 a classifier for discriminating cancer and control samples based on the number of tail genes
385 belonging to the BTG set in a particular sample. Using 10x 5-fold cross-validation, we obtained
386 a high binary classification for ovarian or prostate cancer samples and controls (mean AUC =
387 0.901 and 0.964, resp.), yet a lower classification performance for uterine cancer samples and
388 controls (mean AUC = 0.586) (Fig.5f). Classification performance was still high when
389 combining cancer samples for BTG identification (mean AUC = 0.879), with a better
390 performance for male donors compared to female donors (Supplementary Table 6).
391 The consensus set of BTG present in at least half of the 10x 5-fold cross-validations, consisted
392 of 38 BTG for OV, 18 for PRAD, 3 for UCEC, and 63 for all three cancer types combined. OV
393 and UCEC consensus BTG sets do not overlap. The overlap for OV and PRAD as well as for
394 PRAD and UCEC is significant but the similarity is limited (Fisher's exact $p \leq 1.6E-2$; odds ratio:
395 18.5 and 94.7; Jaccard index: 0.06 and 0.05) (Fig.5h).
396 The difference in number of consensus biomarker tail genes between cancer and control
397 samples per biomarker tail gene subset was more pronounced than based on all tail genes
398 (two-sided Wilcoxon rank-sum: for OV $p = 7.3E-8$ and effect size = 0.876; for PRAD $p = 1.0E-8$
399 and effect size = 0.920; for UCEC $p = 9.4E-5$ and effect size = 0.629; for all cancers combined
400 $p = 1.8E-10$ and effect size = 0.810) (Fig.5b and g).
401 Together, these analyses suggest that individual cancer samples can be distinguished from
402 controls based on the number of biomarker tail genes in their plasma cfRNA profiles.

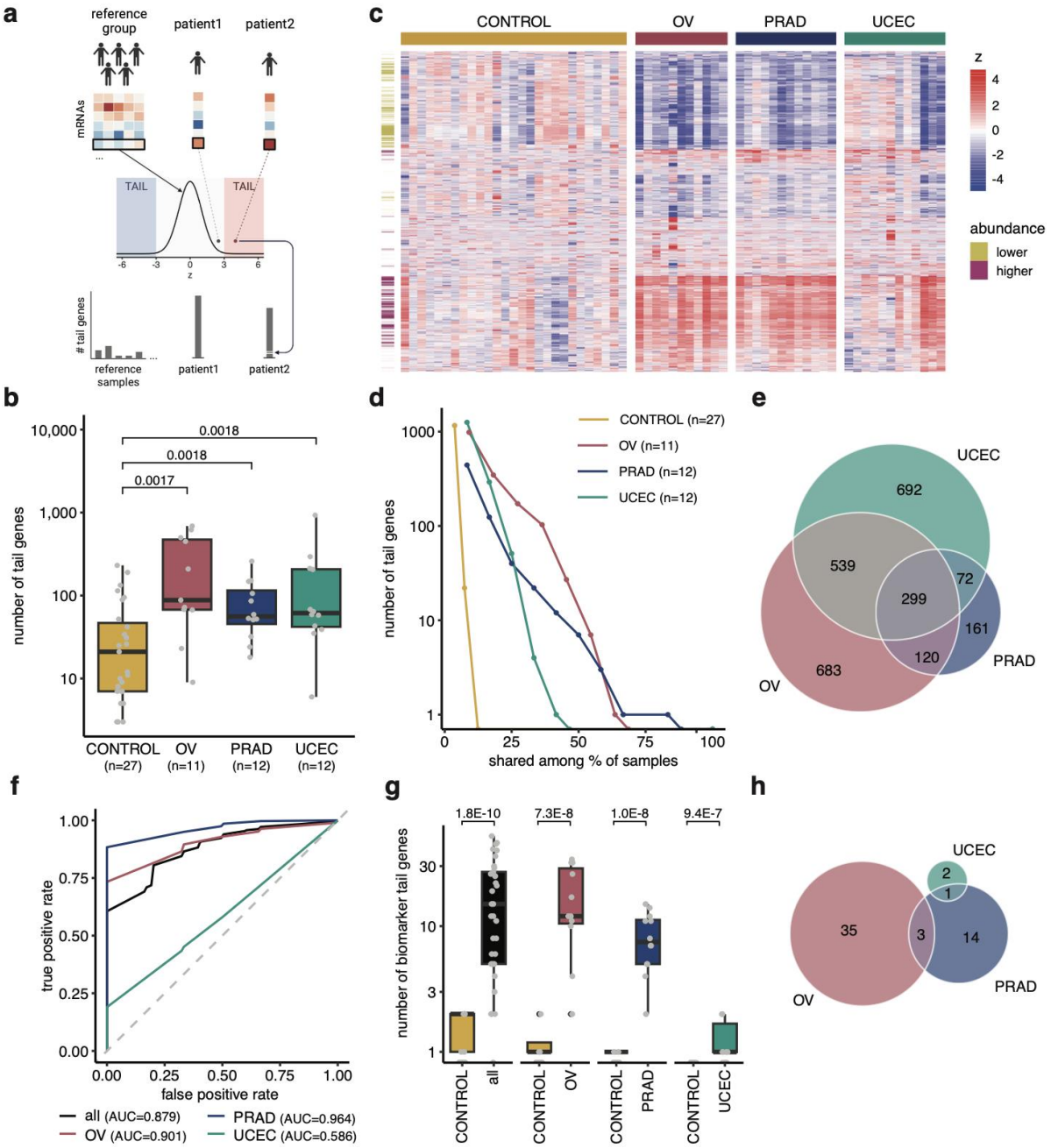


***Fig.5: Higher number of (biomarker) tail genes in cancer patient plasma samples compared to controls**. **a**, Tail gene identification workflow. A gene in a particular sample for which the abundance deviates more than 3 standard deviations from the reference group abundance distribution is a tail gene for that sample. The number of tail genes is used as a metric to determine cancer/control status. **b**, Boxplot of the number of tail genes per group, with grey dots representing individual sample counts. Kruskal-Wallis p = 0.001, effect size = 0.222. Two-sided Wilcoxon rank-sum q-values for specific cancer versus control comparisons indicated in the plot. **c**, Z-score distribution of tail genes in the different samples, all 3312 genes that fulfilled tail gene criteria in at least one of the samples are shown. For visualization purposes, z-scores above 5 or below -5 were reduced to 5 and -5, respectively. Genes (rows) clustered based on Euclidean distance. **d**, Number of tail genes shared by a certain fraction of samples of the same type. Number of samples indicated between brackets. **e**, Overlap between tail*

414 *genes detected in the different cancer types.* ***f****, mean ROC curves of 10x 5-fold cross-validation for binary*
415 *cancer/control classification in test sets based on the number of biomarker tail genes for control and ovarian*
416 *cancer samples (OV), prostate cancer samples (PRAD), uterine cancer samples (UCEC), or all cancer samples (all)*
417 *using the respective biomarker tail gene sets in the training sets.* ***g****, Boxplot of number of 50% consensus*
418 *biomarker tail genes per subset in control (CONTROL) versus the respective cancer samples. Total number of*
419 *biomarker tail genes considered for each comparison: 63 for all cancer (n = 35) versus control, 38 for OV versus*
420 *control, 18 for PRAD versus control, 3 for UCEC versus control. Grey dots represent individual sample counts.*
421 *Two-sided Wilcoxon rank-sum p-values indicated in plot.* ***h****, Overlap of cancer type-specific biomarker tail genes.*
422 *Boxplots show lower quartile (Q1), median, and upper quartile (Q3). Whiskers extend from the lower and upper*
423 *quartile to the smallest and largest value, respectively, within at most 1.5*interquartile range (Q3-Q1) from that*
424 *quartile. More extreme points are plotted as individual black dots. 62 plasma samples considered in total from*
425 *27 cancer-free controls (CONTROL), 11 ovarian cancer (OV), 12 prostate cancer (PRAD), and 12 uterine cancer*
426 *patients (UCEC)).*

427 To independently assess the classification potential of the number of biomarker tail genes,
428 we generated additional plasma cfRNA profiles (lymphoma cohort), including plasma samples
429 from 22 cancer-free control donors and 43 lymphoma patients, more specifically 30 diffuse
430 large B-cell lymphoma (DLBCL) and 13 primary mediastinal large B-cell lymphoma (PMBCL)
431 patients. We identified 5374 unique tail genes, again with more tail genes in cancer samples
432 compared to controls (Kruskal-Wallis p = 2.9E-5, effect size = 0.305, Extended Data Fig.10a).
433 All control tail genes were identified in only one control sample while 45.9% of DLBCL and
434 44.2% of PMBCL tail genes were identified in more than one DLBCL and PMBCL sample,
435 respectively (Extended Data Fig.10b). When swapping sample labels (20 repeats) prior to tail
436 gene identification, no significant differences in the number of tail genes between cancer and
437 control samples were observed in any of the repeats (Kruskal-Wallis p≥0.05).
438 Among the tail genes, 64 to 326 lymphoma BTG, 21 to 220 DLBCL BTG, and 87 to 677 PMBCL
439 BTG were identified using Fisher's exact tests in 10x 5-fold cross-validation. The binary
440 classifiers (cancer/control) based on biomarker tail genes in the training sets resulted in a
441 mean AUC of 0.765 for DLBCL, 0.951 for PMBCL, and 0.866 for all lymphoma in the test sets
442 (Fig.6a). Again, we observed a better performance for samples from male compared to female
443 donors (Supplementary Table 6). When looking at biomarker tail genes that were identified
444 in at least half of the iterations, we found 60 consensus BTG for DLBCL, 371 for PMBCL, and
445 151 for lymphoma. The number of tail genes belonging to these sets was again significantly
446 higher in cancer compared to control samples (Fig.6b). 59% of these consensus BTG were not

447 differentially abundant in any lymphoma subtype versus controls. Notably, early-stage
448 lymphoma patients also displayed an increased number of biomarker tail genes (Fig.6c) and
449 the number of detected BTG per sample was not significantly linked to cancer staging
450 (Jonckheere-Terpstra p = 0.9374 for DLBCL and p = 0.8949 for PMBCL).
451 To test if our cancer-control classification strategy based on biomarker tail genes is also
452 applicable to other biofluids, we applied it to a cohort of urine cfRNA profiles (bladder cancer
453 cohort)[14]. This cohort included urine samples of 12 bladder cancer patients (BLCA) and 12
454 cancer-free control donors. We identified 1152 tail genes in total and again only observed
455 recurrence in cancer samples (Extended Data Fig.10d). 10x 5-fold cross-validation yielded 2
456 to 93 biomarker tail genes and the binary classifiers based on these biomarker tail gene sets
457 resulted in a mean AUC of 0.829 in the test sets (Fig.6d). 6 biomarker tail genes were present
458 in at least half of the iterations and the number of tail genes belonging to this 50% consensus
459 set was significantly higher in bladder cancer versus control samples (Fig.6e). In line with the
460 other datasets, sample label swapping prior to tail gene identification did not result in
461 significant differences in the number of tail genes between the newly labeled cancer samples
462 and controls.
463 Taken together, our results confirm that the number of biomarker tail genes in blood plasma
464 cfRNA enables accurate classification of cancer patients from controls and suggest that the
465 concept may also be applicable to cfRNA profiles from other biofluids.

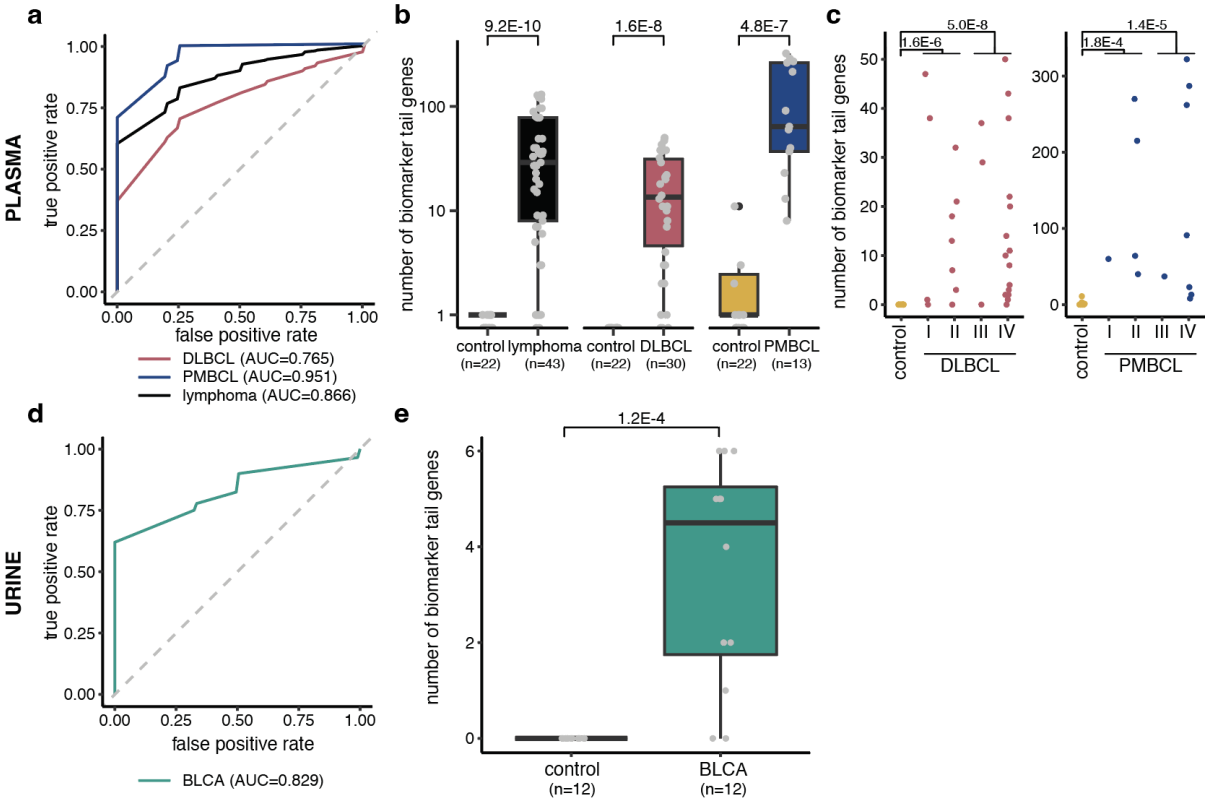


Fig.6: Number of biomarker tail genes is significantly higher in cancer versus control samples in an independent blood plasma and urine cohort. ***a-c***, *lymphoma cohort: plasma samples from 22 control, 30 diffuse large B-cell lymphoma (DLBCL), and 13 primary mediastinal large B-cell lymphoma (PMBCL) patients.* ***d-e***, *bladder cancer cohort: urine samples from 12 control (CONTROL) and 12 bladder cancer (BLCA) patients.* ***a***, *mean ROC curve of 10x 5-fold cross-validation for binary cancer/control classification based on the number of biomarker tail genes for control and DLBCL, or control and PMBCL using the respective biomarker tail gene sets.* ***b***, *boxplot of number of 50% consensus biomarker tail genes in control versus the respective cancer samples. Total number of biomarker tail genes considered for each comparison: 60 for DLBCL versus control, 371 for PMBCL versus control, 151 for both lymphoma types versus control. Grey dots represent individual sample counts.* ***c***, *number of biomarker tail genes of panel b visualized per cancer stage****d***, *mean ROC curve of 10x 5-fold cross-validation for binary cancer/control classification based on the number of biomarker tail genes.* ***e***, *boxplot of number of 50% consensus biomarker tail genes in control versus cancer samples. 6 biomarker tail genes considered in total. Wilcoxon rank-sum p-values indicated in the plots. Boxplots: lower quartile, median, upper quartile, whiskers of 1.5*interquartile range; more extreme points are indicated by black dots.*

481 ## Discussion

482 In this study, we explored cancer-induced changes in blood plasma cfRNA profiles and
483 investigated if these changes could be exploited for classification purposes. Results from our
484 pan-cancer discovery cohort revealed that AML was the only cancer type for which there is
485 direct evidence for tumor-derived cfRNA in plasma in the form of *PML::RARA* fusion
486 transcripts. Also, deconvolution of AML plasma cfRNA revealed a high fraction of myeloid
487 progenitor cells, providing additional – albeit indirect – evidence for tumor-derived cfRNA.
488 For solid tumor types, we identified differentially abundant mRNAs in plasma for every cancer
489 versus control comparison, but a cancer type-specific signal was only observed in plasma from
490 liver cancer patients through the enrichment of liver-specific gene sets. Liver is a major
491 contributor to the plasma transcriptome[4] and its high vascularization and blood flow may
492 facilitate mRNA release in blood upon liver cancer progression. Increased cfDNA contributions
493 from the liver have also been observed in liver cancer patients[15]. Note that, despite the
494 enrichment of liver-related mRNAs, the total cfRNA concentration was not increased in
495 plasma from liver cancer patients, indicating that this signal was still strongly diluted.
496 Considering the organ contribution and dilution argument, other types of liquid biopsies may
497 be better suited to monitor tumor-derived cfRNA for other cancer types. For instance, we
498 could not detect enrichment of prostate tissue-related nor prostate cancer-specific
499 transcripts in blood plasma from prostate cancer patients, in contrast to earlier findings in
500 seminal plasma samples[16].
501 We did observe several general trends in patients' blood plasma transcriptomes in the pan-
502 cancer cohort, which were confirmed in an independent cohort. The lower abundance of
503 immune-related mRNAs may be linked to the suppressed immune system status in cancer
504 patients. It was recently shown that local tumors can have a systemic impact on the immune
505 system, and that different tumor sites of origin resulted in distinct relative abundances and
506 activity profiles of immune cell types[17]. This may be linked to the observed variations in
507 immune-related mRNA abundance, both between and within cancer types. The (immune)
508 cell-related mRNA contributions agreed with reported cell abundance trends found in blood
509 from several cancer patients[18], with the exception of neutrophil-related mRNAs that are
510 lower abundant in plasma. However, neutrophils can have a dual role in cancer[19] and reduced
511 neutrophil cell-free mRNA abundance may result from neutrophils specifically recruited to

512 tumor sites. Finally, the negative enrichment of blood/erythrocyte gene sets may be linked to
513 anemia, a common condition in cancer patients. Among the processes that we found
514 positively enriched in cancer patient plasma, some are linked to hallmarks for cancer
515 progression and metastasis, including epithelial-to-mesenchymal transition and increased
516 blood vessel signal that stem from leaky vasculature and angiogenesis[20].
517 A limitation of the pan-cancer cohort is that control samples were not always perfectly age
518 and sex-matched. The anaplastic astrocytoma, cervical cancer, and testicular germ cell tumor
519 patients were younger as these types predominantly appear at younger age. A dedicated
520 cohort with matched control samples may be better for revealing cancer-related changes in
521 these cancer types. Similarly, while male and female donors were balanced in the control
522 group, this was not always the case for some cancer types due to sex-specificity (e.g. prostate,
523 testicular, breast, cervical, and ovarian cancer) or sample availability. Nevertheless, at most 4
524 out of 8 donors were not sex-matched in cancer vs control group comparisons in the pan-
525 cancer cohort, and our main conclusions were not impacted when only using controls of the
526 same sex in the three-cancer cohort. Another limitation is the relatively small sample size per
527 cancer type, which might contribute to false positive results or insufficient power to detect
528 true differences.
529 The variable mRNA abundance in plasma, combined with the limited overlap of differentially
530 abundant mRNAs across cohorts of the same cancer type prompted an alternative approach
531 that can effectively deal with patient heterogeneity. The tail gene concept satisfies this
532 requirement by identifying mRNAs whose abundance in individual samples significantly
533 deviates from the abundance distribution in a reference group. While the dark channel
534 biomarker gene concept[4] is in part based on a more patient-centric approach, it also requires
535 differential abundance at group level to be considered as potential biomarkers. Moreover,
536 the dark channel concept looks explicitly at genes that are absent (dark) in control samples
537 and therefore cannot capture changes in transcript abundance of genes that are also present
538 in healthy plasma transcriptomes. Of note, 19 dark channel biomarker genes were identified
539 for lung and breast cancer by Larson et al.[4] but five of those genes could be readily detected
540 in control samples from our cohorts, indicating that the “darkness” (i.e. absence of gene
541 transcripts in control samples) should be re-evaluated in different cohorts.
542 Tail gene counts are a simple metric for classifying a biofluid sample as originating from a
543 cancer patient or control. Considering that it is not based on gene signatures with weighted

544 contributions of individual features, we propose it is more robust and compatible with
545 observed patient heterogeneity. Nevertheless, it requires a large and representative group of
546 cancer-free control samples. A control group only needs to be established once, if the entire
547 workflow – from sample collection to RNA-sequencing – is standardized across cancer and
548 control samples. This approach should minimize the introduction of batch effects that may
549 skew cfRNA distributions between a pre-established cohort of cancer-free controls and newly
550 profiled cancer or control samples. In this regard, our lab has demonstrated that preanalytical
551 variables (blood collection tube, time between blood collection and plasma preparation, RNA
552 extraction kit) can indeed dramatically impact cfRNA profiles[21].
553 Deviations in the number of tail genes are not necessarily restricted to cancer patients, nor
554 to blood plasma. Indeed, cfRNA from urine provides a good source of biomarker tail genes to
555 identify bladder cancer patients. Some of the identified biomarker tail genes have previously
556 been linked to cancer, with examples provided in Supplementary Analysis 3. Notably, other
557 diseases may also be characterized by an increase in tail genes, though the specific
558 (biomarker) tail genes may differ.
559 Given the limited sample sizes, we likely did not capture all variability in the current tail gene
560 sets, nor did we identify all possible tail genes. Follow-up studies with larger cohorts are
561 recommended to establish robust gene sets. Also, the applicability and overlap of tail genes
562 in non-malignant disease cohorts needs to be examined further. Similarly, while we have
563 preliminary evidence that biomarker tail genes are also significantly more numerous in low
564 stage cancer patients, additional follow-up studies comprising different stages across various
565 cancer types could reveal whether the concept is suited for early-stage cancer screening or
566 monitoring.

567 ## Methods

568 **Sample cohorts**

569 The study was approved by the ethics committee of Ghent University Hospital, Belgium (no.
570 B670201734362) and performed in accordance with the Declaration of Helsinki.

571 <u>Pan-cancer plasma and three-cancer cohort</u>

572 Blood plasma samples of 266 human donors were obtained from Proteogenex (CA, USA) with
573 informed consent from all donors. The pan-cancer discovery cohort covers 25 different cancer
574 types and a cancer-free control group. For each type, plasma was collected from 8 different

575 donors. The three-cancer cohort consisted of prostate cancer (PRAD), ovarian cancer (OV),
576 and endometrial cancer (UCEC) plasma samples, with 12 donors per type. This cohort also
577 included plasma samples from the 8 control donors of the pan-cancer cohort as well as from
578 22 additional control donors. The control donors in both cohorts where volunteers without
579 (diagnosed) malignant disease. For independently checking reproducibility of pan-cancer
580 results, only the control samples from the additional control donors were used. For tail gene
581 discovery, all control samples in the three-cancer cohort were used.
582 All cancer patients in both cohorts had locally advanced (stage 3) to metastatic cancer (stage
583 4), except for 6 seminoma patients (stage 1). Blood was drawn before treatment with 6
584 exceptions: 4 glioblastoma patients (grade 4-5) already received chemo and radiation
585 therapy, one prostate cancer patient with metastasis received hormone therapy, and one
586 melanoma patient with metastasis received immunotherapy. Demographic and clinical
587 information can be found in Supplementary Table 1.
588 Lymphoma cohort
589 The lymphoma cohort consisted of plasma samples from 65 human donors: 30 DLBCL
590 patients, 13 PMBCL patients, and 22 control donors. The control donors were volunteers
591 without (diagnosed) malignant disease and without indications of abnormalities in routine
592 blood analyses in the clinical lab. Plasma samples were collected following approval by the
593 ethics committee of Ghent University Hospital, Belgium (B670201733701), and written,
594 informed consent was obtained from all patients.
595 Bladder cancer cohort
596 The bladder cancer cohort is described in more detail elsewhere[14]. This cohort consists of
597 urine samples from 12 muscle-invasive bladder cancer patients and 12 control donors.
598 **Blood collection and plasma preparation**
599 Blood from the pan-cancer and three-cancer cohort donors was collected in EDTA vacutainer
600 tubes (Becton Dickinson, 367525). Blood was stored at 4 °C until plasma preparation which
601 started by gently inverting the EDTA tube 10 times and centrifugating for 10 min at 1500 g
602 (without brake at 4 °C). Supernatans was then transferred to 15 mL centrifuge tubes and
603 centrifuged again for 10 min at 1500 g (without brake at 4 °C). The resulting platelet depleted
604 plasma was transferred to 2 mL cryovials, frozen and stored at -80 °C within 4 hours of blood
605 collection. All plasma preparations were done in the same lab. Plasma samples were then
606 shipped on dry ice and only thawed on ice immediately before RNA isolation.

607 For the lymphoma cohort, 2.5 mL of blood was collected from each donor in PAXgene Blood
608 DNA Tubes (BD Biosciences, 761165). Plasma was prepared in one lab by a one-step 15 min
609 centrifugation at 1900 g without brake (room temperature) and stored at -80 °C within 4
610 hours of blood collection.

611 **RNA isolation and gDNA removal**

612 RNA isolation and subsequent library preparation was done by one lab technician per cohort.
613 RNA was isolated from 200 µL of plasma with the miRNeasy Serum/Plasma kit (Qiagen,
614 217184), according to the manufacturer's instructions. For the pan-cancer and three-cancer
615 cohort, 2 µL Sequin spike-in controls (1/1,300,000 of stock solutions mix A, Garvan Institute
616 of Medical Research, Australia) was added to the lysate during RNA isolation as previously
617 described[22]. For the lymphoma cohort, 2 µL of Sequin spike-in controls were added to the
618 sample lysates (1/5000 of stock solution mix A (Garvan Institute of Medical Research) and 2
619 µL of External RNA Control Consortium (ERCC) spike-in controls (ThermoFisher Scientific,
620 4456740) were added to 12 µl RNA eluate.
621 Genomic DNA was removed by adding 1 µL HL-dsDNase (ArcticZymes, 70800-202) and 1.6 µL
622 reaction buffer (ArcticZymes, 66001) to 12 µL RNA eluate, 10 min incubation at 37 °C, followed
623 by 5 min incubation at 55 °C. RNA was stored at -80 °C and thawed on ice immediately before
624 library preparation.

625 **Messenger RNA capture library preparation, sequencing, and quantification**

626 MRNA capture library preparation in pan-cancer and three-cancer cohorts started from 8.5
627 µL DNase treated RNA eluate. cDNA synthesis was performed using TruSeq RNA Library Prep
628 for Enrichment (Illumina, 20020189) as previously described[22]. Briefly, RNA was fragmented,
629 and first strand cDNA was generated using random priming. RNA templates were
630 subsequently removed and replaced by a newly synthesized second strand of cDNA. AMPure
631 XP beads (Beckman Coulter Life Sciences, A63881) were used for purifying the blunt-ended
632 double stranded cDNA. 30 µL cDNA of each sample was then used as input for Illumina DNA
633 Prep with Enrichment (previously Nextera Flex for Enrichment; Illumina, 20025524).
634 Tagmentation and amplification steps were done according to manufacturer's instructions.
635 Quality of resulting pre-enriched libraries was assessed using a high sensitivity Small DNA
636 Fragment Analysis Kit (Agilent Technologies, DNF-477-0500). The libraries were randomly
637 pooled per 8 or 6 samples, for the pan-cancer and three-cancer cohort, respectively, based

638 on their relative concentrations determined using qPCR for reference genes YWHAZ, ACTB,
639 B2M, and UBC. Each multiplex pool was concentrated to a volume of 15 µL with AMPure XP
640 beads (Beckman Coulter Life Sciences, A63881). Finally, enrichment was performed using
641 probes from the Illumina Exome Panel (Illumina, 20020183), probes complementary to the
642 spike-in controls, and blocking probes against globin (anti-CEX) as previously described[22]. At
643 the end of the enrichment workflow, equimolar library pools were prepared based on qPCR
644 quantification with KAPA Library Quantification Kit (Roche Diagnostics, KK4854).
645 Paired-end sequencing of both cohorts was performed (2x100 nucleotides) on a NovaSeq
646 6000. For the pan-cancer cohort, a NovaSeq S2 kit (Illumina, 20028315) was used with
647 standard workflow loading of 1.55 nM (1% PhiX). For the three-cancer cohort, a NovaSeq S1
648 kit (Illumina, 20028318) was used with Xp workflow loading of 1.25 nM (2% PhiX).
649 Adapter trimming and removal of reads shorter than 20 nucleotides was done with cutadapt
650 (v1.18). Read quality was assessed with FastQC (v0.11.9) and low-quality reads were filtered:
651 only reads where at least 80% of bases in both mates have a quality score ≥20 (99% accuracy)
652 were kept (between 87 and 99% of reads/sample). Samples with very few reads (< 2M) were
653 removed from all analyses. More specifically, 4 samples were removed in the pan-cancer
654 plasma cohort (1 lymphoma, 1 bladder, 1 head and neck, and 1 lung cancer sample) and 4
655 samples in the three-cancer cohort (3 control samples, of which 1 repeated from first cohort,
656 and 1 ovarian cancer) as indicated in Supplementary Table 1. Quality filtered reads were then
657 mapped with STAR[23] (v2.6.0) using the default parameters (except for --twopassMode Basic,
658 --outFilterMatchNmin 20 and --outSAMprimaryFlag AllBestScore). The reference files for all
659 analyses were based on genome build GRCh38 and transcriptome build Ensembl v91,
660 complemented with spike annotations. Finally, gene counts were determined with HTSeq[24]
661 (v0.11.0) in non-stranded mode, only considering uniquely mapping reads.

662 **Total RNA library preparation, sequencing, and quantification**

663 Total RNA sequencing libraries of the lymphoma cohort were prepared starting from 8 µL of
664 RNA eluate using the SMARTer Stranded Total RNA-Seq Kit v3 – Pico Input Mammalian
665 (Takara, 634487) according to the manufacturer's protocol. Equimolar library pools were
666 prepared based on qPCR quantification with KAPA Library Quantification Kit (Roche
667 Diagnostics, KK4854). The libraries were paired-end sequenced (2x100 nucleotides) on a
668 NovaSeq 6000 instrument using a NovaSeq S2 kit (Illumina, 20028315) with standard
669 workflow loading of 0.65 nM (2% PhiX).

670 After adapter trimming with cutadapt (v1.18) and quality control with FASTQC (v0.11.9),
671 reads were mapped with STAR[23] (v2.7.3) using default options and the GRCh38 reference
672 described above. Resulting BAM files were deduplicated using UMI-tools (v1.0.0) based on
673 the unique molecular identifier (UMI) sequences in the Pico v3 SMART UMI adapters. Gene
674 counts were determined with HTSeq[24] (v0.11.0) in reverse stranded mode, only considering
675 uniquely mapping reads.

676 **DNA contamination check**

677 To assess possible DNA contamination, we looked at splicing and exonic coverage. For each
678 cohort, the number of reads going to spliced junctions was determined using samtools
679 (v1.16.1) and regional coverage was determined using a combination of samtools (v1.16.1)
680 and bedtools (v2.30.0) with BED files containing intronic, intergenic and exonic regions. The
681 summary of splicing and regional coverage results can be found in Supplementary Table 8.

682 **Data analyses**

683 Most data processing and visualization was done in R (v4.2.1) using tidyverse (v1.3.2) and
684 pheatmap (v1.0.12). Other tools and R packages used for specific analyses are mentioned
685 below. Color schemes are based on Paul Tol's technical note[25].

686 <u>Detection threshold</u>

687 To remove noisy datapoints, we applied a threshold of at least 10 counts. This threshold was
688 based on the median threshold that removes at least 95% of single positive genes[26] - genes
689 that have counts in one replicate but not in the other - between control samples.

690 <u>Spike counts and mRNA concentration</u>

691 Spike counts and Spearman correlations are shown in Supplementary Analysis 3. MRNA
692 concentration was determined based on spikes as previously described[22]. Briefly, the mass of
693 a specific Sequin spike-in control was calculated based on the input concentration and volume
694 of spike-in mix added to the sample. The corresponding mRNA concentration was then
695 estimated by multiplying the spike mass by the ratio of reads mapped to the human genome
696 versus the number of reads mapped to the specific spike, and finally dividing the obtained
697 mass by the plasma volume of the sample. In this study, all Sequin spikes above detection
698 threshold were considered and the geometric mean per sample was taken as the sample's
699 mRNA concentration.

700 <u>Fusion gene detection</u>

701 FusionCatcher[27] (v1.30) was used with default settings for fusion transcript identification.
702 Stringent filtering was applied to exclude potential false positives: transcripts with a
703 FusionCatcher label indicative for false positives, transcripts with reads simultaneously
704 mapping to both fusion partners, and transcripts with fusion partners less than 100 kbp apart
705 were filtered out. Only exon-exon fusions were included and detected fusions in plasma were
706 compared to fusions in the Fusion Gene annotation DataBase (FusionGDB)[28], focusing on
707 fusion genes identified in tissue samples of The Cancer Genome Atlas[29].

708 Differential abundance and principal component analysis

709 DESeq2 (v1.36.0) was used for normalization and differential abundance analysis. MRNAs
710 were pre-filtered by requiring gene counts equal to or greater than the detection threshold
711 in at least half of the tumor or half of the control samples. In the DESeq2 result table, genes
712 with a Benjamini-Hochberg corrected p-value (q) below 0.05 and an absolute $\log_2$ fold change
713 above 1 were considered differentially abundant in cancer versus control (higher abundant:
714 $q < 0.05$ and log2 fold change > 1; lower abundant: $q < 0.05$ and log2 fold change < -1).
715 Principal component analysis was performed with DESeq2 plotPCA on the 500 most variable
716 genes after applying variance stabilizing transformation. Volcano plots were visualized using
717 EnhancedVolcano (v1.14.0) and area-proportional Euler diagrams were made using eulerr
718 (v6.1.1). Differential abundance analysis results are freely available at Zenodo
719 (doi:10.5281/zenodo.7953707).

720 Gene set enrichment and overrepresentation

721 Fgsea[30] (v1.22.0) was used for preranked gene set enrichment analysis based on $\log_2$
722 transformed fold changes between cancer and controls obtained from DESeq2 differential
723 abundance analysis. Significant enrichment was defined by a false discovery rate ≤ 0.05.
724 Hallmark[9] and Canonical Pathways gene sets derived from Kyoto Encyclopedia of Genes and
725 Genomes (KEGG)[10], http://www.pathway.jp, were obtained via the Molecular Signatures
726 Database, MSigDB (v2022.1)[31]. Besides these MSigDB gene sets, custom tissue and cell type
727 specific gene sets were made using inhouse and external references as described below. Gene
728 set enrichment analysis results and custom gene sets are available at Zenodo
729 (doi:10.5281/zenodo.7953707). Functional analysis of mRNA sets was done using the
730 functional annotation tool of DAVID (v2021)[32,33] and - for custom gene sets - the fora function
731 of fgsea. For each functional analysis, a background was provided which included all mRNAs
732 satisfying the count threshold in at least one sample.

733 *RNA Atlas cell type gene sets*

734 These cell type gene sets were defined based on total RNA sequencing data of cells and cell
735 lines in the RNA Atlas[34]. Protein coding gene inclusion criteria were an expression of at least
736 16 counts after variance stabilizing transformation and an expression fold change of at least
737 10 compared to the cell type with second highest expression. 32 different cell type sets were
738 defined, of which 15 had more than 5 genes, yet the largest gene set only included 10 genes.

739 *Tabula Sapiens cell type gene sets*

740 These cell type gene sets were based on the Tabula Sapiens v1 basis matrix defined by
741 Vorperian et al[2]. Protein coding genes with at least 5 CPM and an expression fold change of
742 at least 10 compared to the cell type with the second highest expression were retained. In
743 total, 39 gene sets were obtained, of which 13 sets had more than 5 genes. The largest gene
744 set contained 26 genes.

745 *GTEX tissue gene sets*

746 These tissue gene sets were defined using SPECS specificity scores[35] of GTEx v7[36] tissue
747 samples, summarized in https://specs.cmgg.be. Criteria for protein coding gene inclusion in a
748 certain tissue gene set were a SPECS score ≥ 0.98 and a median TPM ≥ 5. Large gene sets were
749 reduced to the top 200 genes based on the highest SPECS score. 23 out of 27 defined tissues
750 gene sets included more than 5 genes and the median gene set size was 50.

751 *HPA tissue gene sets*

752 These tissue gene sets were defined based on expression data from the Human Protein Atlas[37]
753 (HPA, v23, https://v23.proteinatlas.org/). Human Protein Atlas protein coding genes needed
754 to have a TPM value of at least 5 and expression that is at least 10 times higher in the tissue
755 with highest expression compared to the tissue with second highest expression. Finally, large
756 gene sets were reduced to the top 200 genes based on the highest fold changes. Gene sets
757 for 30 tissues were defined, with 21 gene sets including more than 5 genes and a median gene
758 set size of 8.

759 *TCGA cancer tissue gene sets*

760 These gene sets were defined similarly to the GTEX gene sets but using expression in The
761 Cancer Genome Atlas (TCGA) cancer tissue samples, generated by the TCGA Research
762 Network: https://www.cancer.gov/tcga. Protein coding genes with a SPECS score of at least
763 0.98 in a certain tissue were included in the gene set with a limit of 200 genes, based on

764 highest SPECS score. 23 cancer tissue gene sets were obtained with a median gene set size of
765 21 genes. 19 out of 23 gene sets included more than 5 genes.

766 Cell type deconvolution

767 NuSVR deconvolution using the Tabula Sapiens v1 basis matrix was performed as described
768 by Vorperian et al.[2] and the accompanying GitHub repository,
769 https://github.com/sevahn/deconvolution. As recommended by the authors, data was CPM
770 normalized and no log transformation was performed. Note that the erythrocyte fraction is
771 underestimated due to the absence of hemoglobin probes in mRNA capture sequencing.

772 Tail genes

773 To identify tail genes, transcript counts of every sample were first converted to a customized
774 z-score based on the mean and standard deviation in the reference group, including all control
775 samples except the sample of interest. More specifically, a pseudocount was added to all
776 normalized transcript counts followed by $\log_2$ transformation. Then, for each gene, the mean
777 and standard deviation of these transformed counts in the reference group was used to scale
778 the transformed counts in the sample of interest.

779 Tail genes were then defined as genes with at least 40 counts after DESeq2 normalization and
780 deviating more than 3 standard deviations from the mean of the control reference (|z| > 3).
781 For each gene in a sample of interest:

782 $$z_{gene} = \frac{log_2\left(normalized\ counts_{gene} + 1\right) - \overline{ref}_{gene}}{SD_{ref\ gene}}$$

783 where $\overline{ref}_{gene}$ and $SD_{ref\ gene}$ are respectively the mean and standard deviation of $\log_2$
784 transformed counts for this gene over all (other) control samples.

785 Fisher's exact tests (based on tail gene characteristics (|z| > 3 and normalized counts ≥ 40)
786 and donor cancer/control status) were used to select a subset of tail genes, named biomarker
787 tail genes (BTG), that are specifically associated to the disease state.

788 Binary logistic regression, R glm function with binomial family, was used to classify a sample
789 as cancer or control based on the number of biomarker tail genes. 10x 5-fold cross-validation
790 was used: each time 80% of data (stratified partitioning) was used for BTG identification and
791 model training, the remaining 20% was used to test the classification performance with that
792 particular BTG set. Note that the reference used for z-score calculation never included
793 controls from the test set. For sex-specific cancer types, only samples from control donors

794 with the same sex were used for training and testing. Based on these predictions, the Receiver
795 Operating Characteristic (ROC) curve and Area Under Curve (AUC) were calculated and the
796 mean was visualized using pROC (v1.18.0)[38].
797 Cross-validation resulted in different BTG sets. Consensus BTG were defined as BTG identified
798 in at least half of the cross-validation iterations (≥25 out of 50). The number of tail genes per
799 sample that belong to this consensus set was visualized, where the z-scores were again based
800 on the reference with all controls from the same cohort, except the sample itself.
801 For internal verification, all sample labels were randomly reshuffled using the sample function
802 in R without replacement. Next, tail gene identification and evaluation were done based on
803 the new class labels (using the new control samples as reference). This process was repeated
804 20 times. For 10 iterations of label swapping, we also did biomarker tail gene identification
805 and evaluation with 10x 5-fold cross-validation.

806 **Statistics**

807 Kruskal-Wallis tests were used to compare multiple groups, and Wilcoxon rank-sum tests
808 were used to compare two groups. Corresponding statistical significance, effect sizes and
809 confidence intervals were calculated with rstatix (v0.7.2). For individual testing, p-values
810 smaller than 0.05 were considered significant. In case of multiple testing, the Benjamini-
811 Hochberg procedure was used to calculate false discovery rate adjusted p-values (q-values)
812 and significance was defined as q smaller than 0.05.
813 Significance of overlap between gene sets was determined by Fisher's exact test in
814 GeneOverlap v1.32.0. The universe consisted of all genes considered for the corresponding
815 differential abundance or tail gene analyses (see filtering above). A Jaccard index was
816 calculated to assess similarity between the gene sets, with 0 indicating no similarity and 1
817 indicating identical sets. Jonckheere-Terpstra test was used to assess whether there was a
818 significant trend between the cancer stage and the number of biomarker tail genes in cancer
819 samples.

820 **Data availability**

821 Raw RNA-sequencing data is available under restricted access in the European Genome-
822 phenome Archive (EGA) to preserve individuals' privacy and comply with patient consent for
823 data sharing under the European General Data Protection Regulation. The pan-cancer and
824 three-cancer cohort FASTQs can be found under study EGAS00001006755 (dataset

825 EGAD00001009713). The lymphoma cohort FASTQs can be found under study
826 EGAS00001007127 (dataset EGAD00001010259). The bladder cancer cohort FASTQs can be
827 found under study EGAS00001003917 (dataset EGAD00001005439). Differential abundance
828 and gene set enrichment analysis results are available (open access) in a Zenodo repository
829 (doi:10.5281/zenodo.7953707). An overview of tail genes per type and 50% consensus
830 biomarker tail genes can be found in Supplementary Table 7.

831 **Code availability**

832 Scripts used for differential abundance and tail gene analyses are available on GitHub
833 (https://github.com/OncoRNALab/tailgenes).

834 Acknowledgements

835 Figures 1a and 5b were created with BioRender.com. We would like to thank Tim Mercer for
836 providing the Sequin spikes, Eva Hulstaert and Francisco Avila Cobos for their support during
837 the first phase of the project, and Anneleen Decock for her feedback on the initial draft. We
838 would also like to thank the Biostatistics Unit of the Faculty of Medicine and Health Sciences
839 (Ghent University) for their feedback on the tail gene performance assessment.

840 Funding

841 This work was supported by Ghent University (BOF; BOF19/DOC/228), Ghent University
842 Hospital, Fund for Scientific Research Flanders (FWO; 11C1623N to A.M., 11H7523N to P.D.,
843 G0B2820N), and Kom Op Tegen Kanker (Stand up to Cancer, the Flemish cancer society).

844 Conflict of interest

845 A.M., J.V. and P.M. are inventors on a patent application regarding the tail gene concept
846 (WO2024200811A1).

847 References

848 1. Pessoa, L. S., Heringer, M. & Ferrer, V. P. ctDNA as a cancer biomarker: A broad overview.
849 *Critical Reviews in Oncology/Hematology* **155**, 103109 (2020).
850 2. Vorperian, S. K., Moufarrej, M. N. & Quake, S. R. Cell types of origin of the cell-free
851 transcriptome. *Nat Biotechnol* **40**, 855–861 (2022).

852 3. Moufarrej, M. N. *et al.* Early prediction of preeclampsia in pregnancy with cell-free RNA.
853 *Nature* **602**, 689–694 (2022).
854 4. Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome
855 reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat Commun* **12**,
856 2357 (2021).
857 5. Chen, S. *et al.* Cancer type classification using plasma cell-free RNAs derived from human
858 and microbes. *eLife* **11**, e75181 (2022).
859 6. Roskams-Hieter, B. *et al.* Plasma cell-free RNA profiling distinguishes cancers from pre-
860 malignant conditions in solid and hematologic malignancies. *npj Precis. Onc.* **6**, 1–11
861 (2022).
862 7. Bryzgunova, O. E., Konoshenko, M. Y. & Laktionov, P. P. Concentration of cell-free DNA in
863 different tumor types. *Expert Review of Molecular Diagnostics* **21**, 63–75 (2021).
864 8. Liquori, A. *et al.* Acute Promyelocytic Leukemia: A Constellation of Molecular Events
865 around a Single PML-RARA Fusion Gene. *Cancers (Basel)* **12**, 624 (2020).
866 9. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
867 collection. *Cell Syst* **1**, 417–425 (2015).
868 10. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*
869 *Research* **28**, 27–30 (2000).
870 11. Khwaja, A. *et al.* Acute myeloid leukaemia. *Nat Rev Dis Primers* **2**, 1–22 (2016).
871 12. Giannakeas, V. *et al.* Analysis of Platelet Count and New Cancer Diagnosis Over a 10-Year
872 Period. *JAMA Network Open* **5**, e2141633 (2022).
873 13. Haemmerle, M., Stone, R. L., Menter, D. G., Afshar-Kharghan, V. & Sood, A. K. The Platelet
874 Lifeline to Cancer: Challenges and Opportunities. *Cancer Cell* **33**, 965–983 (2018).
875 14. Hulstaert, E. *et al.* Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas.
876 *Cell Reports* **33**, 108552 (2020).
877 15. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for
878 noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the*
879 *National Academy of Sciences* **112**, E5503–E5512 (2015).
880 16. Hulstaert, E. *et al.* Exploring the extracellular transcriptome in seminal plasma for non-
881 invasive prostate cancer diagnosis. 2021.05.11.21256306 Preprint at
882 https://doi.org/10.1101/2021.05.11.21256306 (2021).

883 17. Allen, B. M. *et al.* Systemic dysfunction and plasticity of the immune macroenvironment
884 in cancer models. *Nat Med* **26**, 1125–1134 (2020).
885 18. Hiam-Galvez, K. J., Allen, B. M. & Spitzer, M. H. Systemic immunity in cancer. *Nat Rev*
886 *Cancer* **21**, 345–359 (2021).
887 19. Xiong, S., Dong, L. & Cheng, L. Neutrophils in cancer carcinogenesis and metastasis.
888 *Journal of Hematology & Oncology* **14**, 173 (2021).
889 20. Lugano, R., Ramachandran, M. & Dimberg, A. Tumor angiogenesis: causes, consequences,
890 challenges and opportunities. *Cell. Mol. Life Sci.* **77**, 1745–1770 (2020).
891 21. exRNAQC Consortium *et al.* Blood collection tube and RNA purification method
892 recommendations for extracellular RNA transcriptome profiling: A large-scale pre-
893 analytical variable performance evaluation. Preprint at https://doi.org/10.21203/rs.3.rs-
894 2409038/v2 (2024).
895 22. Hulstaert, E. *et al.* Messenger RNA capture sequencing of extracellular RNA from human
896 biofluids using a comprehensive set of spike-in controls. *STAR Protocols* **2**, 100475 (2021).
897 23. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
898 24. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-
899 throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
900 25. Tol, P. Colour Schemes, Technical Note SRON/EPS/TN/09-002. (2012).
901 26. Mestdagh, P. *et al.* Evaluation of quantitative miRNA expression platforms in the
902 microRNA quality control (miRQC) study. *Nature Methods* **11**, 809–815 (2014).
903 27. Nicorici, D. *et al.* FusionCatcher – a tool for finding somatic fusion genes in paired-end
904 RNA-sequencing data. 011650 Preprint at https://doi.org/10.1101/011650 (2014).
905 28. Kim, P. & Zhou, X. FusionGDB: fusion gene annotation DataBase. *Nucleic Acids Research*
906 **47**, D994–D1004 (2019).
907 29. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of
908 Human Cancers. *Cell Reports* **23**, 227-238.e3 (2018).
909 30. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at
910 https://doi.org/10.1101/060012 (2021).
911 31. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
912 interpreting genome-wide expression profiles. *Proceedings of the National Academy of*
913 *Sciences* **102**, 15545–15550 (2005).

914 32. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large
915 gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).
916 33. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and
917 functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216–W221
918 (2022).
919 34. Lorenzi, L. *et al.* The RNA Atlas expands the catalog of human non-coding RNAs. *Nat*
920 *Biotechnol* **39**, 1453–1465 (2021).
921 35. Everaert, C., Volders, P.-J., Morlion, A., Thas, O. & Mestdagh, P. SPECS: a non-parametric
922 method to identify tissue-specific molecular features for unbalanced sample groups. *BMC*
923 *Bioinformatics* **21**, 58 (2020).
924 36. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585
925 (2013).
926 37. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
927 38. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC
928 curves. *BMC Bioinformatics* **12**, 77 (2011).
929