

<https://doi.org/10.1038/s43856-026-01507-8>

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

Check for updates

Annelien Morlion^{1,2}, Philippe Decruyenaere^{1,2,3}, Kathleen Schoofs^{1,2,4}, Jasper Anckaert^{1,2}, Nickolas Johns Ramirez^{1,2}, Justine Nuytens^{1,2}, Eveline Vanden Eynde^{1,2}, Kimberly Verniers^{1,2}, Celine Everaert⁴, Guy Brusselle⁵, Steven Callens⁶, Filomeen Haerynck^{7,8}, Dimitri Hemelsoet⁹, Eric Hoste¹⁰, Jo Lambert¹¹, Nicolaas Lumen¹², Fritz Offner³, Koen Paemeleire⁹, Vanessa Smith^{13,14}, Lies Van den Eynde¹², Jo Van Dorpe¹⁵, Amber Vanhaecke¹⁵, Hans Van Vlierberghe^{7,16,17}, An Mariman^{18,19}, Olivier Thas^{20,21,22}, Jo Vandesompele^{1,2,23} & Pieter Mestdagh^{1,2,23} ✉

Abstract

Background Circulating nucleic acids in blood plasma form an attractive, minimally invasive resource to study human health and disease. In this study, we aimed to identify cell-free RNA alterations that can distinguish cancer patients from cancer-free individuals.

Methods We first performed mRNA capture sequencing on 266 blood plasma samples from cancer patients and controls, including a discovery set of 208 donors across 25 cancer types and a replication set of 58 donors across three cancer types. We first conducted group-level comparisons and then compared individual patient profiles to a reference control population in a one-versus-many approach. This approach was further evaluated in independent cohorts: a prostate cancer plasma cohort ($n = 180$), a non-malignant disease plasma cohort ($n = 125$), a lymphoma plasma cohort ($n = 65$), and a bladder cancer urine cohort ($n = 24$), each including both patients and controls.

Results Here we show that cancer patients exhibit both cancer type-specific and general cell-free RNA alterations. However, differentially abundant RNAs vary widely among patients and across cohorts, hampering robust biomarker identification. By comparing individual patient profiles to control populations, we identify so-called biomarker tail genes, which strongly deviate from controls. The number of these genes per sample distinguishes cancer patients from control samples. Independent cohorts also confirm the potential of this approach.

Conclusions Our findings demonstrate substantial heterogeneity in cell-free RNA alterations among cancer patients and propose that patient-specific changes can be exploited for classification.

Plain language summary

RNAs that circulate in biofluids can provide information about a person's health. In this study, we examined such extracellular RNAs in blood plasma from people with and without cancer to see if specific patterns could help identify cancer. We analyzed over 600 samples from multiple cancer types and compared each patient's RNA patterns to those of healthy individuals. We found that each patient shows unique changes, and that the number of strongly altered RNAs can distinguish cancer patients from healthy people. These findings suggest that studying individual RNA patterns could improve cancer detection and support personalized care.

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 improving patient well-being and outcomes. 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 to 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.

A full list of affiliations appears at the end of the paper. ✉ e-mail: pieter.mestdagh@ugent.be

To date, most studies on circulating nucleic acids in cancer patients have focused on circulating tumor DNA (ctDNA). The amount of tumor DNA in circulation is highly dependent on the tumor type and burden¹, suggesting that ctDNA can be less informative for certain cancer types or early-stage cancers. Cell-free RNA (cfRNA) may complement ctDNA by reflecting dynamic changes in gene expression during cancer development and progression, or upon treatment, in both healthy and diseased cells. A recent study showed that different cell types contribute RNA to the healthy cell-free plasma transcriptome, and that cfRNA allows monitoring of cell-type-specific changes in diseases, such as non-alcoholic fatty liver and Alzheimer's disease². The potential of cfRNA as a prognostic biomarker has also been demonstrated for pre-eclampsia and organ damage in pregnancy³. Some studies have also shown potential of using cfRNA biomarkers in certain cancer types^{4–6}. Larson et al.⁴ specifically looked at genes that are “dark” (no or few detected transcripts) in plasma from healthy controls but abundant in plasma from lung and/or breast cancer patients and found indications of tissue-of-origin signal. Chen et al.⁵ reported the potential of combining human and microbe-derived plasma cfRNA for distinguishing healthy donor and cancer plasma samples (colorectum, stomach, liver, lung, and esophageal cancer) based on profiles of RNAs showing group-level differences. Roskams-Hieter et al.⁶ showed plasma cfRNA classification potential for liver cancer and multiple myeloma patient samples compared to non-cancer controls, with a higher accuracy using learning vector quantization feature selection compared to group-level differential abundance. The authors, however, indicated that liver and bone marrow are major contributors to the plasma transcriptome, which may result in more cancer signals.

Here, we further examine the impact of different cancer types on cfRNA profiles in different cohorts. We first quantify cell-free mRNAs in different cohorts of blood plasma from cancer patients with locally advanced or metastatic tumors and cancer-free donors to uncover tumor or tissue-of-origin mRNA signatures with diagnostic potential. Results reveal substantial heterogeneity in cell-free mRNA abundance among cohorts, questioning the biomarker potential of cell-free mRNA signatures. However, patient-specific alterations, reflected by the number of cell-free mRNAs with differential abundance in a patient sample compared to a cancer-free control population, enable to accurately distinguish cancer patients from healthy individuals.

Methods

Sample cohorts

The study was performed in accordance with the Declaration of Helsinki. Demographic and clinical information can be found in Supplementary Data 1.

Pan-cancer plasma and three-cancer cohort. Blood plasma samples of 266 human donors were obtained from Proteogenex (CA, USA) with informed consent from all donors. The use of these samples was approved by the ethics committee of Ghent University Hospital, Belgium (no. B670201734362). The pan-cancer discovery cohort covers 25 different cancer types and a cancer-free control group. For each type, plasma was collected from 8 different donors. The three-cancer cohort consisted of prostate cancer (PRAD), ovarian cancer (OV), and endometrial cancer (UCEC) plasma samples, with 12 donors per type. This cohort also included plasma samples from the 8 control donors of the pan-cancer cohort, as well as from 22 additional control donors. The control donors in both cohorts were volunteers without (diagnosed) malignant disease. For independently checking the reproducibility of pan-cancer results, only the control samples from the additional control donors were used. For tail gene discovery, all control samples in the three-cancer cohort were used.

All cancer patients in both cohorts had locally advanced (stage 3) to metastatic cancer (stage 4), except for 6 seminoma patients (stage 1). Blood was drawn before treatment with 6 exceptions: 4 glioblastoma patients (grade 4–5) already received chemo and radiation therapy, one prostate

cancer patient with metastasis received hormone therapy, and one melanoma patient with metastasis received immunotherapy.

Prostate cancer and non-malignant cohort. Samples were collected from 48 control donors without diagnosed disease, 132 patients diagnosed with prostate cancer, and 88 patients diagnosed with a non-malignant disease. Plasma samples were collected and used following approval by the ethics committee of Ghent University Hospital, Belgium (no. B670201733120 and no. B6702024000092), and written informed consent was obtained from all patients.

Lymphoma cohort. The lymphoma cohort consisted of plasma samples from 65 human donors: 30 DLBCL patients, 13 PMBCL patients, and 22 control donors. The control donors were volunteers without (diagnosed) malignant disease and without indications of abnormalities in routine blood analyses in the clinical lab. Plasma samples were collected following approval by the ethics committee of Ghent University Hospital, Belgium (no. B670201733701), and written informed consent was obtained from all patients.

Bladder cancer cohort. The bladder cancer cohort is described in more detail elsewhere⁷. This cohort consists of urine samples from 12 muscle-invasive bladder cancer patients and 12 control donors. Samples were collected following informed consent and approval by the Masaryk Memorial Cancer Institute, Brno, Czech Republic (no. 14-08-27-01 and no. MOU190814).

Blood collection and plasma preparation

Blood from the pan-cancer and three-cancer cohort donors was collected in EDTA vacutainer tubes (Becton Dickinson, 367525). Blood was stored at 4 °C until plasma preparation, which started by gently inverting the EDTA tube 10 times and centrifugating for 10 min at 1500 × g (without brake at 4 °C). Supernatants were then transferred to 15 mL centrifuge tubes and centrifuged again for 10 min at 1500 × g (without brake at 4 °C). The resulting platelet-depleted plasma was transferred to 2 mL cryovials, frozen and stored at –80 °C within 4 h of blood collection. All plasma preparations were done in the same lab. Plasma samples were then shipped on dry ice and only thawed on ice immediately before RNA isolation.

For the prostate cancer cohort, 3 times 10 mL of blood was collected from each donor in EDTA vacutainer tubes (Becton Dickinson, 367525). Blood was centrifuged for 10 min at 815 × g (without brake at 4 °C). Supernatants were then transferred to Eppendorf tubes and centrifuged again for 10 min at 2500 × g (without brake at 4 °C). The resulting platelet-depleted plasma was transferred to 2 mL cryovials, snap frozen in liquid nitrogen and stored at –80 °C within 4 h of blood collection.

For the non-malignant cohort, 2 times 9 mL of blood was collected from each donor in EDTA vacutainer tubes (Becton Dickinson, 367525). Blood was processed to platelet-depleted plasma in a three-spin protocol (without brake at room temperature). First, blood was centrifuged for 20 min at 400 × g. The supernatants were subsequently transferred to 15 mL centrifuge tubes and centrifuged again for 15 min at 2500 × g. The resulting supernatants were centrifuged again for another 15 min at 2500 × g. The final platelet-depleted plasma was snap frozen in liquid nitrogen and stored at –80 °C within 4 h of blood collection.

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

RNA isolation and gDNA removal

RNA isolation and subsequent library preparation were done by one lab technician per cohort. For the prostate cancer and non-malignant cohort, RNA was isolated from 1000 µL and 500 µL plasma, respectively, with the iCatcher cfDNA/cfRNA isolation 1000 kit (LabHelp Robotics, AC30100-

36). A total of 12.5 and 6.25 μL of Sequin spike-in controls⁸ were added to the sample lysates, respectively (1/1300,000 of stock solution mix A, Garvan Institute of Medical Research; also available via Sequin.bio, Australia), and 2 μL of External RNA Control Consortium (ERCC) spike-in controls (Thermo Fisher Scientific, 4456740) were added to 12 μL RNA eluate.

For all other cohorts, RNA was isolated from 200 μL of plasma with the miRNeasy Serum/Plasma kit (Qiagen, 217184), according to the manufacturer's instructions. For the pan-cancer and three-cancer cohort, 2 μL Sequin spike-in controls (1/1,300,000 of stock solutions mix A, Garvan Institute of Medical Research) were added to the lysate during RNA isolation as previously described⁹. For the lymphoma cohort, 2 μL of Sequin spike-in controls were added to the sample lysates (1/5000 of stock solution mix A, Garvan Institute of Medical Research), and 2 μL of External RNA Control Consortium (ERCC) spike-in controls (Thermo Fisher Scientific, 4456740) were added to 12 μL RNA eluate.

Genomic DNA was removed by adding 1 μL HL-dsDNase (ArcticZymes, 70800-202) and 1.6 μL reaction buffer (ArcticZymes, 66001) to 12 μL RNA eluate, 10 min incubation at 37 °C, followed by 5 min incubation at 55 °C. RNA was stored at -80 °C and thawed on ice immediately before library preparation.

Messenger RNA capture library preparation, sequencing, and quantification for the pan-cancer and three-cancer cohort

MRNA capture library preparation started from 8.5 μL DNase-treated RNA eluate. cDNA synthesis was performed using TruSeq RNA Library Prep for Enrichment (Illumina, 20020189) as previously described⁹. Briefly, RNA was fragmented, and first-strand cDNA was generated using random priming. RNA templates were subsequently removed and replaced by a newly synthesized second strand of cDNA. AMPure XP beads (Beckman Coulter Life Sciences, A63881) were used for purifying the blunt-ended double-stranded cDNA. A total of 30 μL cDNA of each sample was then used as input for Illumina DNA Prep with Enrichment (previously Nextera Flex for Enrichment; Illumina, 20025524). Tagmentation and amplification steps were done according to the manufacturer's instructions. Quality of resulting pre-enriched libraries was assessed using a high-sensitivity Small DNA Fragment Analysis Kit (Agilent Technologies, DNF-477-0500). The libraries were randomly pooled per 8 or 6 samples, for the pan-cancer and three-cancer cohort, respectively, based on their relative concentrations determined using qPCR for reference genes YWHAZ, ACTB, B2M, and UBC. Each multiplex pool was concentrated to a volume of 15 μL with AMPure XP beads (Beckman Coulter Life Sciences, A63881). Finally, enrichment was performed using probes from the Illumina Exome Panel (Illumina, 20020183), probes complementary to the spike-in controls, and blocking probes against globin (anti-CEX) as previously described⁹. At the end of the enrichment workflow, equimolar library pools were prepared based on qPCR quantification with the KAPA Library Quantification Kit (Roche Diagnostics, KK4854).

Paired-end sequencing of the pan-cancer and three-cancer cohort was performed (2×100 nucleotides) on a NovaSeq 6000. For the pan-cancer cohort, a NovaSeq S2 kit (Illumina, 20028315) was used with standard workflow loading of 1.55 nM (1% PhiX). For the three-cancer cohort, a NovaSeq S1 kit (Illumina, 20028318) was used with Xp workflow loading of 1.25 nM (2% PhiX).

Adapter trimming and removal of reads shorter than 20 nucleotides was done with cutadapt (v1.18). Read quality was assessed with FastQC (v0.11.9), and low-quality reads were filtered: only reads where at least 80% of bases in both mates have a quality score ≥ 20 (99% accuracy) were kept (between 87 and 99% of reads/sample). Samples with very few reads (< 2 M) were removed from all analyses. More specifically, 4 samples were removed in the pan-cancer plasma cohort (1 lymphoma, 1 bladder, 1 head and neck, and 1 lung cancer sample) and 4 samples in the three-cancer cohort (3 control samples, of which 1 repeated from the first cohort, and 1 ovarian cancer) as indicated in Supplementary Data 1. Quality filtered reads were then mapped with STAR¹⁰ (v2.6.0) using the default parameters (except for `--twopassMode Basic`, `--outFilterMatchNmin 20` and `--outSAMprimaryFlag AllBestScore`). The reference files for all analyses were based on genome build GRCh38, and

transcriptome build Ensembl v91, complemented with spike annotations. Finally, gene counts were determined with HTSeq¹¹ (v0.11.0) in non-stranded mode, only considering uniquely mapping reads.

Messenger RNA capture library preparation, sequencing, and quantification for the prostate cancer and non-malignant cohort

MRNA capture library preparation started from 4.25 μL DNase-treated RNA eluate and was done with Illumina RNA Prep with Enrichment (Illumina, 20040537), following the manufacturer's instructions. Briefly, RNA was fragmented, and first-strand cDNA was generated using random priming. RNA templates were subsequently removed and replaced by a newly synthesized second strand of cDNA. AMPure XP beads (Beckman Coulter Life Sciences, A63881) were used for purifying the blunt-ended double-stranded cDNA. Tagmentation and amplification steps were done according to the manufacturer's instructions. Quality of resulting pre-enriched libraries was assessed using a high-sensitivity Small DNA Fragment Analysis Kit (Agilent Technologies, DNF-477-0500). The libraries were randomly pooled per 3 samples, based on their relative concentrations determined using qPCR for reference genes ACTB and B2M. Each multiplex pool was concentrated to a volume of 30 μL with AMPure XP beads (Beckman Coulter Life Sciences, A63881). Finally, enrichment was performed using probes from the Illumina Exome Panel (Illumina, 20020183), probes complementary to the spike-in controls, and blocking probes against globin (anti-CEX) as previously described⁹. At the end of the enrichment workflow, equimolar library pools were prepared based on qPCR quantification with KAPA Library Quantification Kit (Roche Diagnostics, KK4854).

Paired-end sequencing of both cohorts was performed (2×50 nucleotides) on a NovaSeq X using a NovaSeq X 10B flow cell (Illumina, 20085596). For the prostate cancer cohort, a loading concentration of 150 pM with 2% PhiX was used. For the non-malignant cohort, a loading concentration of 172.5 pM with 2% PhiX was used.

Adapter trimming and removal of reads shorter than 20 nucleotides were done with cutadapt (v3.5). Read quality was assessed with FastQC (v0.11.9), and low-quality reads were filtered: only reads where at least 80% of bases in both mates have a quality score ≥ 20 (99% accuracy) were kept (between 87 and 99% of reads/sample). Duplicate removal was done with clumpify dedupe (BBMap v38.98). Filtered reads were then mapped with STAR¹⁰ (v2.7.9) using the default parameters (except for `--twopassMode Basic`, `--outFilterMatchNmin 20` and `--outSAMprimaryFlag AllBestScore`). The reference files for all analyses were based on genome build GRCh38, and transcriptome build Ensembl v109, complemented with spike annotations. Finally, gene counts were determined with HTSeq¹¹ (v0.11.3) in non-stranded mode, only considering uniquely mapping reads.

Total RNA library preparation, sequencing, and quantification

Total RNA sequencing libraries of the lymphoma cohort were prepared starting from 8 μL of RNA eluate using the SMARTer Stranded Total RNA-Seq Kit v3 – Pico Input Mammalian (Takara, 634487) according to the manufacturer's protocol. Equimolar library pools were prepared based on qPCR quantification with the KAPA Library Quantification Kit (Roche Diagnostics, KK4854). The libraries were paired-end sequenced (2×100 nucleotides) on a NovaSeq 6000 instrument using a NovaSeq S2 kit (Illumina, 20028315) with standard workflow loading of 0.65 nM (2% PhiX).

After adapter trimming with cutadapt (v1.18) and quality control with FASTQC (v0.11.9), reads were mapped with STAR¹⁰ (v2.7.3) using default options and the GRCh38 reference described above. Resulting BAM files were deduplicated using UMI-tools (v1.0.0) based on the unique molecular identifier (UMI) sequences in the Pico v3 SMART UMI adapters. Gene counts were determined with HTSeq¹¹ (v0.11.0) in reverse-stranded mode, only considering uniquely mapping reads.

Data analyses

Most data processing and visualization were done in R (v4.2.1) using tidyverse (v1.3.2) and pheatmap (v1.0.12). Other tools and R packages used

for specific analyses are mentioned below. Color schemes are based on Paul Tol's technical note¹².

DNA contamination check. To assess possible DNA contamination, we looked at splicing and exonic coverage. For each cohort, the number of reads going to spliced junctions was determined using samtools (v1.16.1) and regional coverage was determined using a combination of samtools (v1.16.1) and bedtools (v2.30.0) with BED files containing intronic, intergenic and exonic regions. The summary of splicing and regional coverage results can be found in Supplementary Table 2.

Detection threshold. To remove noisy datapoints, we applied a threshold of at least 10 counts. This threshold was based on the median threshold that removes at least 95% of single positive genes¹³—genes that have counts in one replicate but not in the other—between control samples.

Spike counts and mRNA concentration. Spike counts and Spearman correlations are shown in Supplementary Analysis 4 (Supplementary Figs. 12–17). mRNA concentration was determined based on spikes as previously described⁹. Briefly, the mass of a specific Sequin spike-in control was calculated based on the input concentration and volume of spike-in mix added to the sample. The corresponding mRNA concentration was then estimated by multiplying the spike mass by the ratio of reads mapped to the human genome versus the number of reads mapped to the specific spike, and finally dividing the obtained mass by the plasma volume of the sample. In this study, all Sequin spikes above the detection threshold were considered and the geometric mean per sample was taken as the sample's mRNA concentration.

Fusion gene detection. FusionCatcher¹⁴ (v1.30) was used with default settings for fusion transcript identification. Stringent filtering was applied to exclude potential false positives: transcripts with a FusionCatcher label indicative of false positives, transcripts with reads simultaneously mapping to both fusion partners, and transcripts with fusion partners less than 100 kbp apart were filtered out. Only exon-exon fusions were included and detected fusions in plasma were compared to fusions in the Fusion Gene annotation DataBase (FusionGDB)¹⁵, focusing on fusion genes identified in tissue samples of The Cancer Genome Atlas¹⁶.

Differential abundance and principal component analysis. DESeq2 (v1.36.0) was used for normalization and differential abundance analysis. MRNAs were pre-filtered by requiring gene counts equal to or greater than the detection threshold in at least half of the tumor or half of the control samples. In the DESeq2 result table, genes with a Benjamini–Hochberg corrected p value (q) below 0.05 and an absolute \log_2 fold change above 1 were considered differentially abundant in cancer versus control (higher abundant: $q < 0.05$ and \log_2 fold change > 1 ; lower abundant: $q < 0.05$ and \log_2 fold change < -1). Principal component analysis was performed with DESeq2 plotPCA on the 500 most variable genes after applying variance stabilizing transformation. Volcano plots were visualized using Enhanced-Volcano (v1.24.0) and area-proportional Euler diagrams were made using eulerr (v6.1.1). Differential abundance analysis results are freely available at Zenodo (doi:10.5281/zenodo.7953707).

Gene set enrichment and overrepresentation. Fgsea¹⁷ (v1.22.0) was used for preranked gene set enrichment analysis based on \log_2 -transformed fold changes between cancer and controls obtained from DESeq2 differential abundance analysis. Significant enrichment was defined by a false discovery rate ≤ 0.05 . Hallmark¹⁸ and Canonical Pathways gene sets derived from Kyoto Encyclopedia of Genes and Genomes (KEGG)¹⁹, <http://www.pathway.jp>, were obtained via the Molecular Signatures Database, MSigDB (v2022.1)²⁰. Besides these MSigDB gene sets, custom tissue and cell-type-specific gene sets were made using inhouse and external references as described below. Gene set enrichment analysis

results and custom gene sets are available at Zenodo (doi:10.5281/zenodo.7953707). Functional analysis of mRNA sets was done using the functional annotation tool of DAVID (v2021)^{21,22} and—for custom gene sets—the for function of fgsea. For each functional analysis, a background was provided, which included all mRNAs satisfying the count threshold in at least one sample.

RNA Atlas cell type gene sets. These cell-type gene sets were defined based on total RNA sequencing data of cells and cell lines in the RNA Atlas²³. Protein coding gene inclusion criteria were an expression of at least 16 counts after variance stabilizing transformation and an expression fold change of at least 10 compared to the cell type with the second highest expression. Thirty-two different cell type sets were defined, of which 15 had more than 5 genes, yet the largest gene set only included 10 genes.

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

GTEX tissue gene sets. These tissue gene sets were defined using SPECS specificity scores²⁴ of GTEx v7²⁵ tissue samples, summarized in <https://specs.cmgg.be>. Criteria for protein-coding gene inclusion in a certain tissue gene set were a SPECS score ≥ 0.98 and a median TPM ≥ 5 . Large gene sets were reduced to the top 200 genes based on the highest SPECS score. Twenty-three out of 27 defined tissues gene sets included more than 5 genes and the median gene set size was 50.

HPA tissue gene sets. These tissue gene sets were defined based on expression data from the Human Protein Atlas²⁶ (HPA, v23, <https://v23.proteinatlas.org/>). Human Protein Atlas protein coding genes needed to have a TPM value of at least 5 and expression that is at least 10 times higher in the tissue with the highest expression compared to the tissue with the second highest expression. Finally, large gene sets were reduced to the top 200 genes based on the highest fold changes. Gene sets for 30 tissues were defined, with 21 gene sets including more than 5 genes and a median gene set size of 8.

TCGA cancer tissue gene sets. These gene sets were defined similarly to the GTEx gene sets but using expression in The Cancer Genome Atlas (TCGA) cancer tissue samples, generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. Protein-coding genes with a SPECS score of at least 0.98 in a certain tissue were included in the gene set with a limit of 200 genes, based on the highest SPECS score. 23 cancer tissue gene sets were obtained with a median gene set size of 21 genes. 19 out of 23 gene sets included more than 5 genes.

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

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

Tail genes were then defined as genes with at least 40 counts after DESeq2 normalization and deviating more than 3 standard deviations from the mean of the control reference ($|z| > 3$).

For each gene in a sample of interest:

$$z_{gene} = \frac{\log_2(\text{normalized counts}_{gene} + 1) - \overline{ref}_{gene}}{SD_{ref_{gene}}}$$

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

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

Binary logistic regression, R glm function with binomial family, was used to classify a sample as cancer or control based on the number of biomarker tail genes. 10×5-fold cross-validation was used: each time, 80% of the data (stratified partitioning) was used for BTG identification and model training, and the remaining 20% was used to test the classification performance with that particular BTG set. Note that the reference used for z -score calculation never included controls from the test set. For sex-specific cancer types, controls of both sexes were used in the control reference for z -score calculation, but only samples from control donors with the same sex were used to assess performance. Based on these predictions, the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) were calculated and the mean was visualized using pROC (v1.18.0)²⁷.

Cross-validation resulted in different BTG sets. Consensus BTG were defined as BTG identified in at least half of the cross-validation iterations (≥ 25 out of 50). The number of tail genes per sample that belong to this consensus set was visualized, where the z -scores were again based on the reference with all controls from the same cohort, except the sample itself.

For validation in the prostate cancer cohort, z -scores for cancer and control samples in the validation set (30%) were calculated using the control samples from the train-test set (70%) as reference. Then, consensus BTG and the classification threshold previously identified in the train-test set were applied to the validation set to assess classification performance. Similarly, the control samples from the train-test set were used as a reference for the non-malignant cohort samples after joint normalization. The prostate cancer consensus BTG and classification threshold were then applied to the non-malignant sample cohort to assess cancer specificity.

For internal verification, all sample labels were randomly reshuffled using the sample function in R without replacement. Next, tail gene identification and evaluation were done based on the new class labels (using the new control samples as reference). This process was repeated 20 times. For 10 iterations of label swapping, we also did biomarker tail gene identification and evaluation with 10 × 5-fold cross-validation.

Statistics and reproducibility

Kruskal–Wallis tests were used to compare multiple groups, and Wilcoxon rank-sum tests were used to compare two groups. Corresponding statistical significance, effect sizes and confidence intervals were calculated with rstatix (v0.7.2). For individual testing, p values smaller than 0.05 were considered significant. In case of multiple testing, the Benjamini–Hochberg procedure was used to calculate false discovery rate adjusted p values (q values) and significance was defined as q smaller than 0.05.

Significance of overlap between gene sets was determined by Fisher's exact test in GeneOverlap v1.32.0. The universe consisted of all genes considered for the corresponding differential abundance or tail gene analyses (see filtering above). A Jaccard index was calculated to assess similarity between the gene sets, with 0 indicating no similarity and 1 indicating identical sets. The Jonckheere–Terpstra test was used to assess whether there was a significant trend between the cancer stage and the number of biomarker tail genes in cancer samples.

The optimal classification threshold for the number of biomarker tail genes in the train-test set of the prostate cancer cohort was determined based on Youden's Index with pROC v.1.18.0.

Sample sizes were optimized to include a variety of cancers with biological replicates, while taking into account sample availability and costs. Given the small sample size per type in the pan-cancer cohort ($n = 208$, 25 cancer types and a cancer-free control group with each time plasma samples from 8 donors), an independent (three-cancer) cohort was used to test reproducibility ($n = 66$, 3 cancer types with plasma samples from 12 donors per type and a cancer-free control group with plasma samples from 8 donors also included in the pan-cancer cohort and 22 independent donors. Note that the 8 samples from control donors also included in the pan-cancer cohort were excluded for assessing independent reproducibility. The prostate cancer cohort ($n = 180$, plasma samples from 132 prostate cancer patients and 48 cancer-free controls) was used for identifying and validating prostate cancer biomarker tail genes using a 70–30 split. The non-malignant cohort ($n = 125$, plasma samples from 88 patients with non-malignant disease conditions and 37 healthy controls) was used to test the cancer-specificity of the prostate cancer biomarker tail genes. Finally, the lymphoma cohort ($n = 65$, plasma samples from 30 diffuse large B-cell lymphoma patients, 13 primary mediastinal B-cell lymphoma patients, and 22 cancer-free control donors) and the bladder cancer cohort ($n = 24$, urine samples from 12 bladder cancer patients and 12 cancer-free controls) were used for testing the applicability of the tail gene concept for another processing method or another biofluid type, respectively. An overview of sample sizes per cohort can also be found in Table 1.

Results

A diverse repertoire of cell-free mRNAs in plasma

To characterize the blood plasma cell-free mRNA transcriptome, we examined liquid biopsy cfRNA profiles in different cohorts (Table 1). We first started from a pan-cancer blood plasma discovery cohort consisting of 208 individuals covering 25 different cancer types and age-matched healthy controls (Fig. 1a, and Supplementary Data 1). mRNA capture sequencing of the pan-cancer plasma cohort resulted in 3 to 56 M paired-end reads per sample (median 16 M) after quality filtering with no significant differences among groups (Kruskal–Wallis test $p > 0.05$) (Supplementary Fig. 1a). On average 84% (95% CI [82.59, 84.66]) of uniquely mapped reads were assigned to protein coding genes and most other reads originated from mitochondrially encoded ribosomal RNAs (Supplementary Fig. 1b). Per sample, between 7439 and 13,039 unique mRNAs (median 10,932) were detected with at least 10 counts—we use the term mRNA to refer to all mRNA reads mapping to the same gene (Fig. 1b). All plasma samples had 5263 mRNAs in common, and 9908 mRNAs were found in at least 75% of the samples. Of note, we detected more mRNAs in plasma from acute myeloid leukemia (AML) patients compared to plasma from controls and from patients with solid cancers (Kruskal–Wallis $p = 6.6E-4$, effect size = 0.163, post hoc two-sided Wilcoxon rank-sum $q < 0.05$ for AML vs. controls and 16 out of 24 solid cancer types). Although some AML samples were sequenced deeper, sequencing depth alone did not explain the higher number of detected mRNAs (Supplementary Fig. 1c). We next determined the mRNA concentration using synthetic RNA spikes that were added to each lysed plasma sample (Fig. 1c). mRNA concentrations ranged from 0.003 to 8.302 ng/ml with an average concentration of 0.486 ng/ml. Despite the higher number of detected mRNAs, the mRNA concentration in AML plasma was not significantly higher than controls or other cancer types. Moreover, though the largest median mRNA concentration (READ, 0.648 ng/ml) was more than 10 times higher than the smallest median (ANA, 0.063 ng/ml), no significant concentration differences were found among the groups (Kruskal–Wallis test $p = 0.0275$, effect size = 0.087, with all post hoc two-sided Wilcoxon rank-sum $q > 0.05$). Besides the small sample sizes, this may be related to the considerable cfRNA concentration variability between patients, which is also reported at cfDNA level²⁸.

We then searched for evidence of tumor-derived cfRNA by investigating known cancer fusion transcripts in individual plasma samples. In

Table 1 | Cohort overview

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 ^a	Plasma	mRNA capture (Illumina DNA Prep with Enrichment)	Replication pan-cancer analyses & discovery tail gene concept
Prostate cancer cohort	132 (1)	48	Plasma	mRNA capture (Illumina RNA Prep with Enrichment)	Independent validation tail gene concept
Non-malignant cohort	0 (0)	37 healthy, 88 non-malignant	Plasma	mRNA capture (Illumina RNA Prep with Enrichment)	Specificity of tail genes
Lymphoma cohort	43 (2)	22	Plasma	total RNA (SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian)	Independent demonstration tail gene concept
Bladder cancer cohort	12 (1)	12	Urine	mRNA capture (TruSeq RNA Exome)	Independent demonstration tail gene concept

Name, number of donors (one sample per donor), biofluid type, library preparation method and application of the cohort in this study.

^aEight control donors overlap with the pan-cancer cohort; data from these eight donors were only used for discovery of the tail gene concept, not for replication of pan-cancer analyses (see "Methods").

every cancer plasma sample, we only considered fusion transcripts that had been documented in tumor tissue of that same cancer type. Such fusion transcripts were only identified in plasma samples of AML patients. The *PML::RARA* fusion transcript²⁹ was found in two AML plasma samples with several reads supporting the fusion (7 to 19 spanning pairs and 6 to 9 unique spanning reads). The reciprocal *RARA::PML* fusion was found in one of both AML plasma samples (19 pairs and 17 unique reads). In plasma samples of solid cancer types, we only detected fusion transcripts that had not been reported in the tumor tissue of these cancer types (Supplementary Table 1). Of note, these fusion transcripts had fewer supporting reads (1 to 2 pairs and 2 to 5 unique reads) than the *PML::RARA/RARA::PML* fusions in AML and may thus represent false positives. Moreover, other fusions with few supporting reads were also identified in healthy control plasma samples. While the two AML samples with the *PML::RARA* fusion transcripts were sequenced deeper than other samples, the fusions were still detected when downsampling to the median number of reads of the pan-cancer cohort. These analyses confirm the presence of tumor cell-derived cfRNA in plasma samples from AML patients.

Differential mRNA abundance in cancer patients' plasma compared to controls

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

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

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

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

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 hallmark¹⁸, Kyoto Encyclopedia of Genes and Genomes (KEGG)¹⁹, and custom tissue- and cell type gene sets (Fig. 3a and Supplementary Fig. 4a, b). We observed a negative enrichment of various immune-related signals (interleukin signaling, interferon gamma and inflammatory response, and neutrophils) across all cancer types except UCEC, with $q < 0.05$ in at least 21 out of 25 cancer types. We also found a negative enrichment of the KEGG ribosome gene set in all solid cancer types ($q < 0.05$ in 19 out of 24 types) (Supplementary Fig. 4b). These results are in line with a

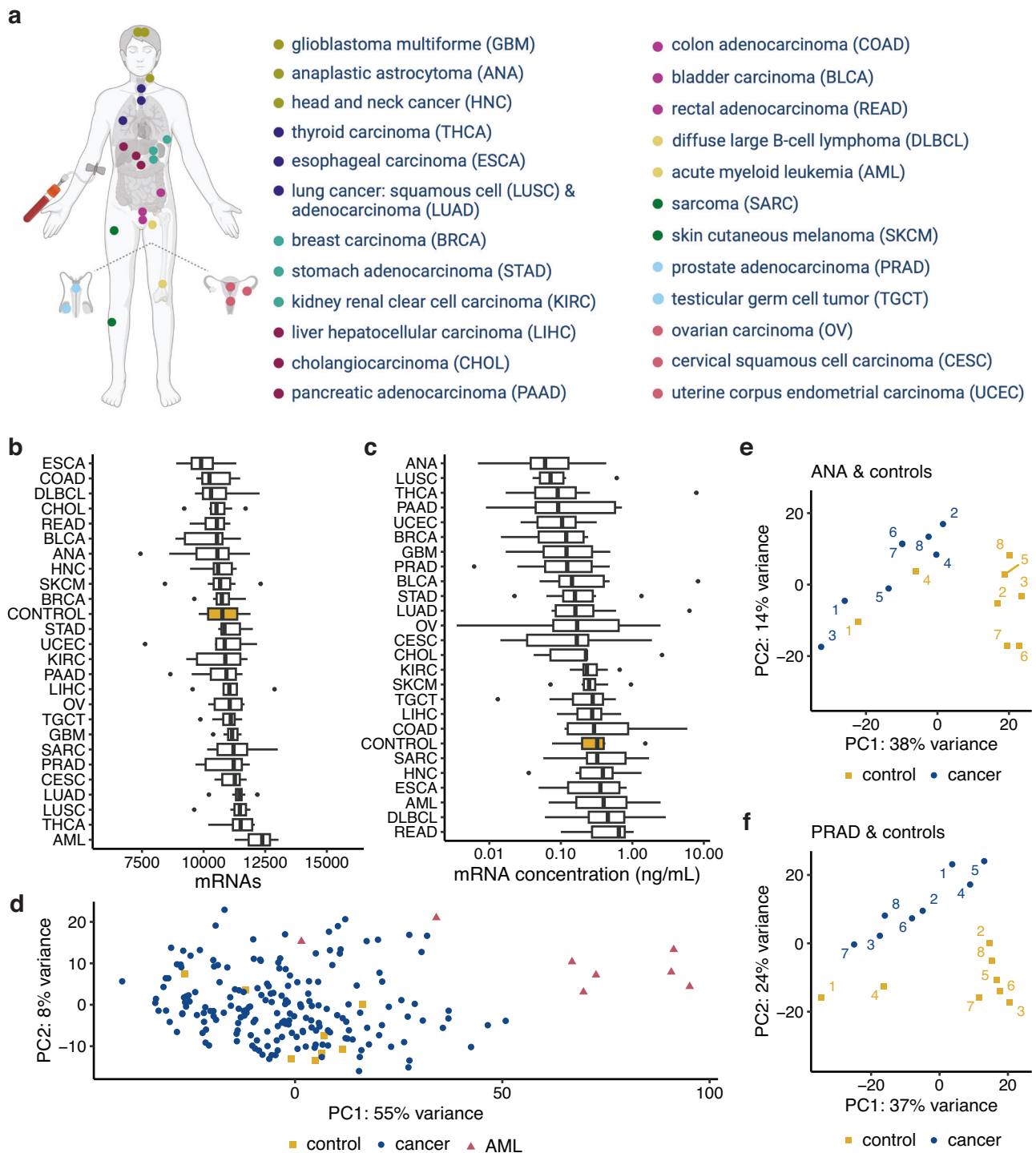


Fig. 1 | Pan-cancer plasma cohort overview, mRNA content and sample clustering. **a** Primary sites of the 25 cancer types in the pan-cancer cohort. Primary sites are indicated with colored dots, and corresponding cancer types are listed (top to bottom) with the (sub)type abbreviations between brackets. Details can be found in Supplementary Data 1. Figure created with BioRender.com. **b** Number of plasma mRNAs, detected with at least 10 counts, across cancer types and healthy controls. Groups ordered according to median. 204 samples in total included (4 of the original 208 samples—biological replicates for each of 25 cancer types and the control group—were excluded due to low read counts; see Methods and Supplementary Data 1). **c** Variability in plasma mRNA concentration across cancer types and healthy controls. Groups ordered according to median. 201 samples in total were included (in addition to the 4 samples excluded for low read counts, 3 samples were further

excluded because their Sequin spike-in counts were <10). **b, 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, 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.

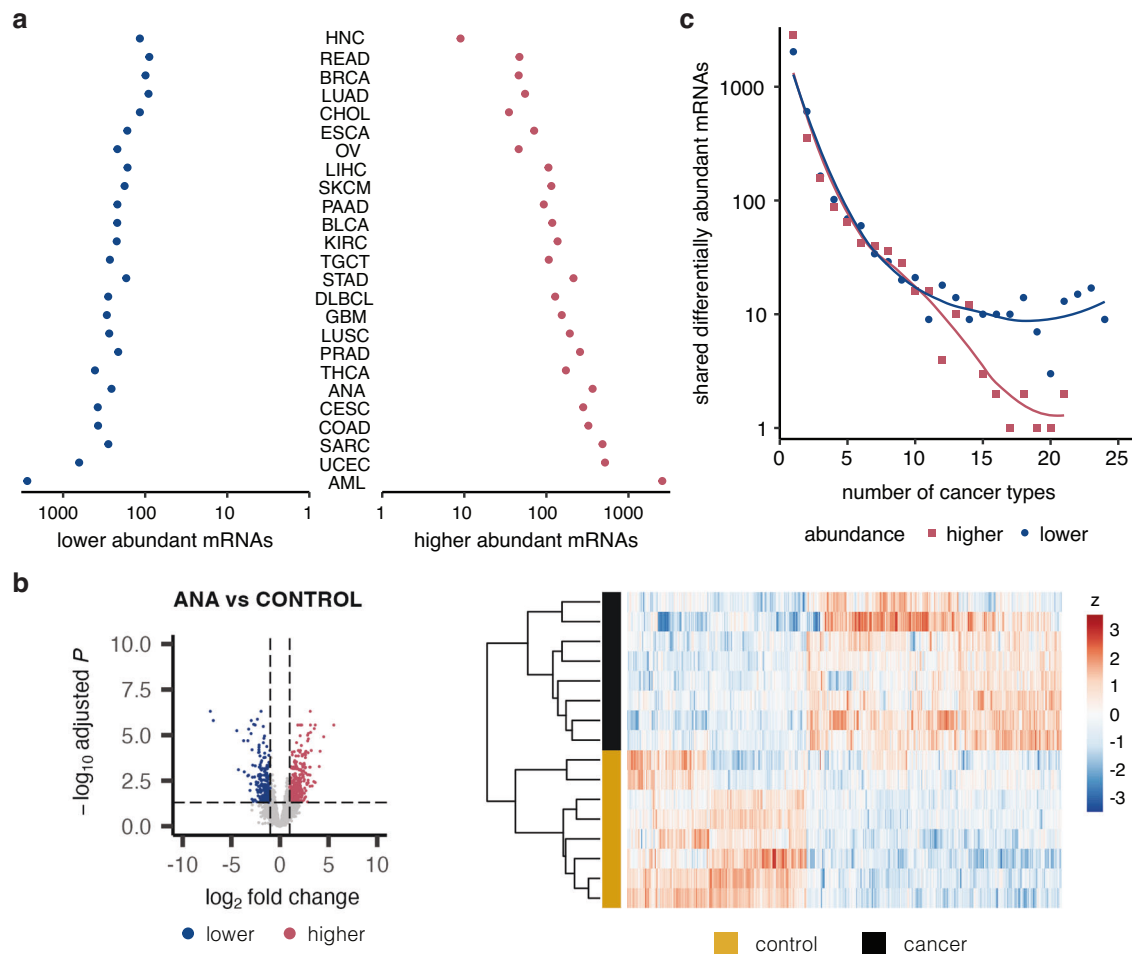


Fig. 2 | Differentially abundant mRNAs in cancer versus control. **a** Differentially abundant mRNAs in plasma from cancer patients versus controls. Cancer types ordered based on total number of differentially abundant mRNAs ($q < 0.05$ and $|\log_2$ fold change > 1). **b** Volcano plot and heatmap of differentially abundant mRNAs in anaplastic astrocytoma vs. controls (other cancer-control comparisons in Supplementary Fig. 3). Differential abundance was assessed using DESeq2 (two-sided Wald test with Benjamini–Hochberg correction for multiple testing). 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 the relative abundance of differentially abundant mRNAs (columns) in individual plasma samples (rows). Heatmap colors represent $\log_2(\text{normalized counts} + 1)$ 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.

previous study by Chen et al.⁵ on plasma cfRNA profiles of five early-stage solid cancer types. Other negatively enriched gene sets with $q < 0.05$ in at least half of the cancer types in our cohort were blood-related (Fig. 3a and Supplementary Fig. 4a, b). Conversely, ‘extracellular matrix receptor’ and ‘focal adhesion’ gene sets were significantly positively enriched in at least half of the cancer types (Supplementary Fig. 4b), confirming results by Chen et al. Other positively enriched gene sets included those related to KRAS signaling, epithelial-mesenchymal transition, myogenesis, adipose tissue, brain function, and hematopoietic stem cells (Fig. 3a and Supplementary Fig. 4a).

Cancer type-specific enrichment of cfRNA reflecting the primary tumor site was only observed for liver cancer patients. ‘Liver tissue’, ‘liver cancer tissue’, and ‘hepatocyte’ gene sets were all positively enriched in liver cancer patients’ plasma ($q < 0.05$, Fig. 3b) and not in plasma from any other cancer type. A ‘bronchial epithelial cells’ gene set was significantly enriched in lung adenocarcinoma and lung squamous cell carcinoma (LUAD $q = 0.027$; LUSC $q = 0.026$), but this enrichment was also found in four other cancer types (ANA, CESC, CHOL, and UCEC).

We then applied computational deconvolution to enumerate the cell type fractions contributing to plasma cfRNA profiles (Supplementary Data 3). Platelets, erythrocytes/erythrocyte progenitors, monocytes, macrophages, and B-cells formed the top-5 of most abundant cell types

that contributed to plasma cfRNA of both controls and patients with solid tumors (Fig. 3c). Except for macrophages, these cell types were also reported as the most abundant cell types in a previous study on healthy donor plasma². Plasma from AML patients showed a very different composition. Several cell type contributions—including hematopoietic stem cells, myeloid progenitor cells, and monocytes—were increased or even exclusively present in AML plasma samples, in line with the blood cell counts typically observed in these patients³⁰. Myeloid progenitor cells, normally not present in peripheral blood, were an abundant contributor in the AML plasma samples (mean 10.7%, or 8.5% when downsampling AML reads to the cohort median), with almost no signal in other plasma samples (mean 0.1%) (Supplementary Fig. 4c). Moreover, the platelet fraction in AML plasma was significantly lower compared to other groups (two-sided Wilcoxon rank sum $p = 3.8E-5$, effect size = 0.289) with on average 18.3% in AML compared to 62.3% in other samples, reflecting the reduced thrombocyte counts commonly found in leukemia patients³⁰.

In plasma from patients with solid tumors, the platelet fraction was significantly higher compared to controls (Supplementary Fig. 4c, two-sided Wilcoxon rank-sum $p = 2.3E-2$, effect size = 0.163), which is in line with the prognostic value of platelet counts and the frequent occurrence of thrombocytosis in solid tumor patients^{31,32}. For neutrophils, erythrocyte/

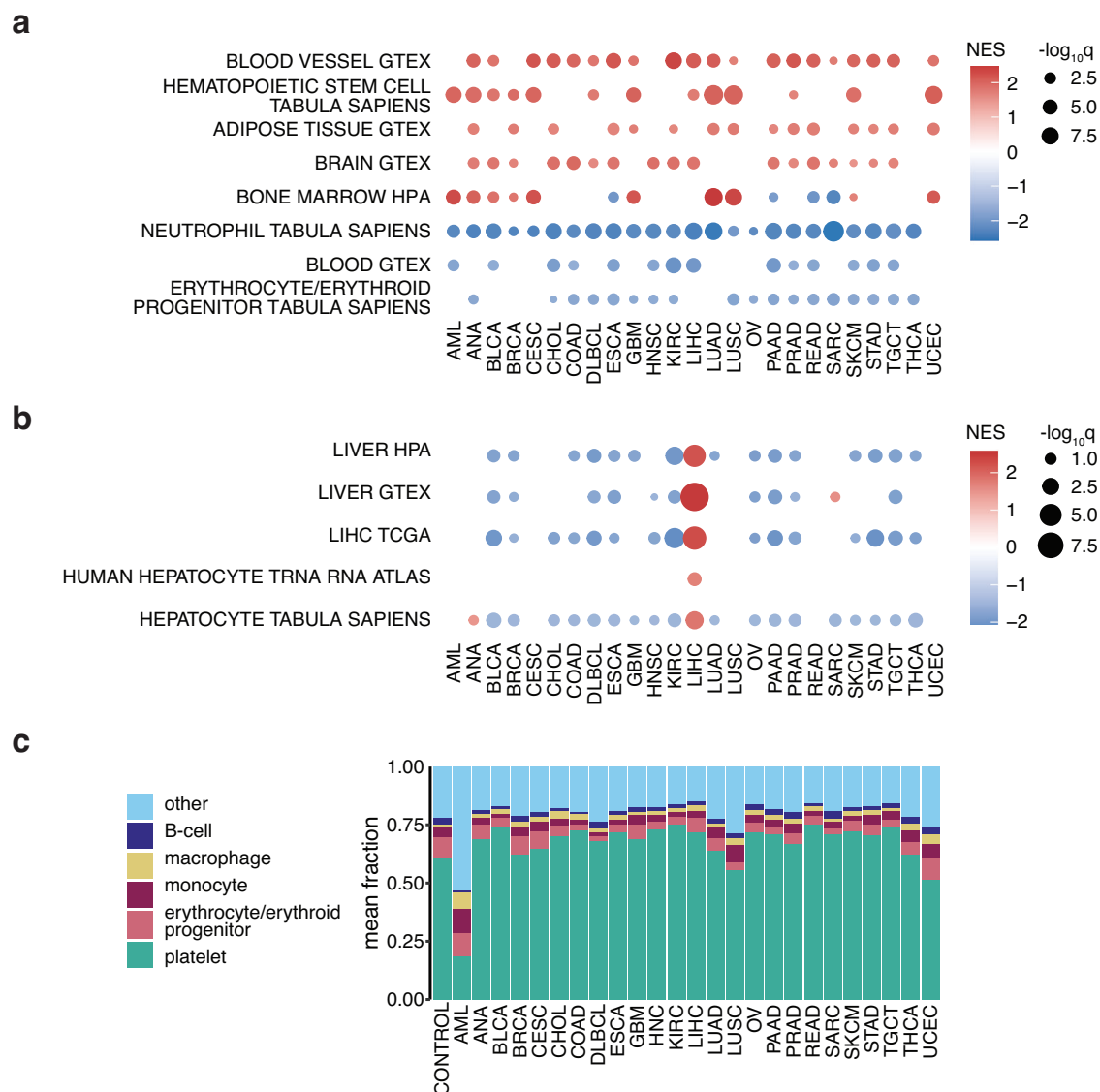


Fig. 3 | Gene set enrichment and cell type deconvolution reveal pan-cancer and cancer-type-specific signals. **a** Gene set enrichment analysis results for custom tissue and cell type sets. Only sets that are significantly enriched in at least half of the cancer types are shown. Gene sets with $q < 0.05$ colored according to normalized enrichment scores (NES). **b** Liver-related gene set enrichment in plasma from liver

cancer patients, gene sets with $q < 0.25$ colored according to NES. Detailed information about gene sets can be found in “Methods”. **c** Average fraction of the five most abundant cell types based on deconvolution. Other: cumulative sum of other cell types. The five most abundant cell types were selected based on mean ranking in all groups.

erythroid progenitor cells, thymocytes, and T-cells, the fraction was significantly lower in cancer patients’ plasma compared to controls (two-sided Wilcoxon rank-sum $p = 1.8E-2$ (effect size = 0.169), $p = 1.4E-2$ (effect size = 0.176), $p = 1.8E-2$ (effect size = 0.164), $p = 4.3E-4$ (effect size = 0.252), respectively).

Independent cohort confirms systemic signal and cfRNA heterogeneity

To verify our findings, we processed an independent (three-cancer) cohort of plasma samples, which, after removing samples with low read depth, included samples from 20 cancer-free control donors and 35 cancer patients with one of three different solid cancer types: prostate (PRAD, $n = 12$), ovarian (OV, $n = 11$), and uterine (UCEC, $n = 12$) cancer. Seven to 61 M paired-end reads were obtained per sample (median 28 M) with no significant difference in sequencing depth among groups (Kruskal–Wallis $p = 0.7858$) (Supplementary Fig. 5a). In line with the pan-cancer cohort, we detected between 8642 and 11,944 mRNAs per sample (Supplementary Fig. 5b), with 6194 mRNAs shared across all plasma samples of this three-

cancer cohort. mRNA concentrations ranged from 0.007 to 0.414 ng/ml (Supplementary Fig. 5c).

Most cancer samples could be separated from controls in the first principal component, especially for prostate and ovarian cancer (Fig. 4a). Based on differential abundance analyses between cancer and control samples, we found the highest number of differentially abundant mRNAs in OV (995 higher abundant, 1644 lower abundant), followed by PRAD (848 higher abundant, 1043 lower abundant) (Fig. 4b). The lowest number of differentially abundant mRNAs was found for UCEC (243 higher abundant, 349 lower abundant) even though this cancer type had the most differentially abundant mRNAs in the pan-cancer cohort. Hierarchical clustering based on differential mRNA abundance in this three-cancer cohort mostly separated cancer and control samples but again revealed variability in mRNA abundance within groups (Supplementary Fig. 5d).

A total of 189 and 299 mRNAs were significantly higher and lower abundant, respectively, in plasma from all three cancer types compared to controls (Fig. 4b). The lower abundant mRNAs included all 26 mRNAs that were identified as lower abundant in most cancers in the pan-cancer cohort.

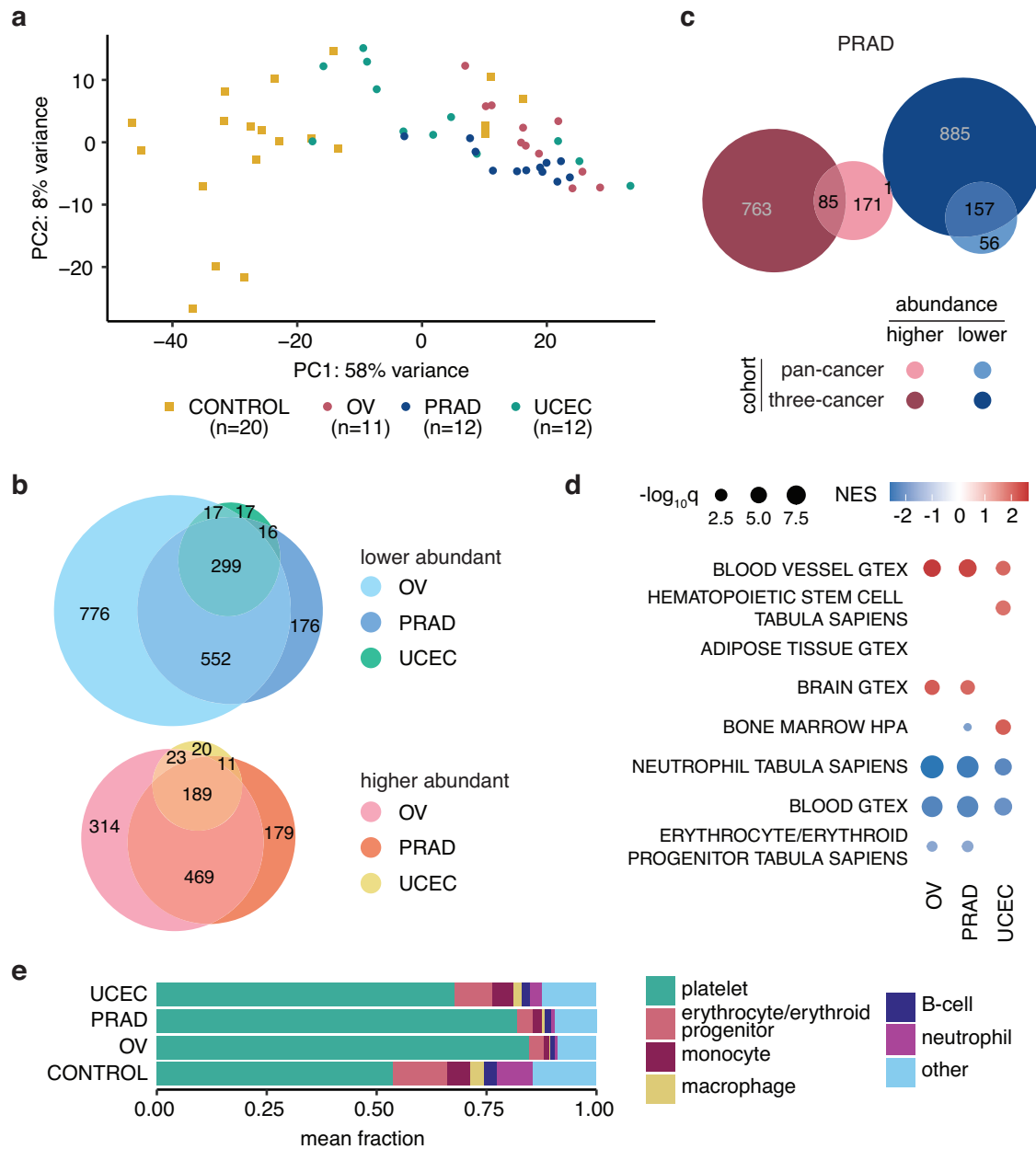


Fig. 4 | Three-cancer cohort confirms findings from pan-cancer cohort and reveals significant, yet limited, overlap in differentially abundant mRNAs. **a** Principal component analysis on top 500 most variable mRNAs shows distinct clustering of most cancer and control samples (variance stabilizing transformation 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. Two-sided 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 abundant cell types and macrophages based on cell type deconvolution. Other: cumulative sum of other cell types. Differential abundance: $q < 0.05$ and $|\log_2 \text{fold change}| > 1$. OV ovarian cancer, PRAD prostate cancer, UCEC uterine cancer.

There was a significant, although limited, overlap in differentially abundant mRNAs between the three-cancer and the pan-cancer cohort for each of the three cancer types (Fisher’s exact test on lower abundant mRNAs: OV $p = 8.4E-118$ (odds ratio = 30.0), PRAD $p = 1.1E-155$ (odds ratio = 59.3), UCEC $p = 7.6E-13$ (odds ratio = 4.2); on higher abundant mRNAs: OV $p = 4.0E-4$ (odds ratio = 4.6), PRAD $p = 2.9E-52$ (odds ratio = 12.2), UCEC $p = 2.2E-5$ (odds ratio = 3.3); Jaccard index [0.009–0.143]) (Fig. 4c and Supplementary Fig. 6a). Despite the uniform collection and processing of samples from patients with the same cancer type, the low Jaccard indices indicate high heterogeneity between the two cohorts, as illustrated with

probabilistic modeling for varying cohort sizes (Supplementary Analysis 1, Supplementary Fig. 11). Respectively 38%, 49%, and 95% of the differentially abundant mRNAs in OV, PRAD, and UCEC in the pan-cancer cohort could not be confirmed in the three-cancer cohort. Moreover, classifiers built on differentially abundant genes did not show robust performance across cohorts (Supplementary Analysis 2, Supplementary Tables 3 and 4). Notably, mRNAs with lower abundance showed a higher overlap between both cohorts than mRNAs with higher abundance.

In line with the pan-cancer cohort, there was no enrichment of cancer type-specific gene sets reflecting tumor tissue of origin in the OV, PRAD and

UCEC samples in the three-cancer cohort. However, enriched gene sets shared among at least half of the cancer types in the pan-cancer cohort were confirmed in the three-cancer cohort, except for adipose tissue (Fig. 4d and Supplementary Fig. 6b, c).

Cell type deconvolution confirmed platelets, erythrocyte/erythroid progenitor cells, monocytes and B-cells among the top-5 cell types contributing cRNA to plasma (Fig. 4e and Supplementary Data 3), with neutrophils being ranked as the fourth most abundant cell type. The higher platelet fraction in cancer patient plasma was even more pronounced in the three-cancer cohort (two-sided Wilcoxon rank-sum cancer versus control $p = 5.8E-9$, effect size = 0.679), especially for ovarian and prostate cancer (Supplementary Fig. 6d). Neutrophils, erythrocyte/erythroid progenitor cells, thymocytes and T-cells again had a lower plasma fraction in cancer patients compared to controls (two-sided Wilcoxon rank-sum $p = 4.4E-6$ (effect size = 0.556), $p = 3.2E-7$ (effect size = 0.610), $p = 6.4E-5$ (effect size = 0.419), $p = 4.1E-2$ (effect size = 0.261), respectively).

Taken together, the limited overlap of differentially abundant mRNAs between the pan-cancer and three-cancer cohort highlights plasma transcriptome variability, yet enrichment and cell type contribution analyses confirm the systemic cRNA changes in plasma from cancer patients.

cRNA tail genes distinguish cancer samples from controls

Apart from liver cancer and AML, our analyses did not reveal evidence for a tumor or tissue of origin signal among mRNAs that were more abundant in cancer patient versus control plasma. Moreover, higher abundant mRNAs were less recurrent across cancer types and showed lower reproducibility between cohorts compared to mRNAs with lower abundance in cancer versus control plasma. Our analyses suggest that the latter represent a systemic immune response rather than a tumor-derived signal. Based on these observations, we hypothesized that (tumor derived) cRNA signals may be heterogeneous among patients, and that classic differential mRNA and pathway abundance analysis between groups may not reveal such non-uniform signals. We therefore decided to investigate cRNA profiles in individual cancer patients using samples from the three-cancer cohort.

To capture aberrations in the cRNA profile of an individual cancer patient, we compared the mRNA abundance of a cancer plasma sample to the mRNA abundance distribution in the entire control group, which served as a reference. More specifically, a z -score was calculated for each mRNA based on the mean and standard deviation of the \log_2 -normalized count distribution in the control group. Abundant mRNAs with an absolute z -score of at least 3 were defined as “tail genes” for that patient, referring to their position in one of the tails of the control z -score distribution (Fig. 5a).

We searched for tail genes in each cancer patient from the three-cancer cohort (where we could take advantage of a large control group) and in each control sample, always using all other control samples as reference (Supplementary Fig. 7a, b). We identified a total of 3311 unique tail genes in this cohort and observed a significantly higher number of tail genes in plasma samples from cancer patients compared to controls (Kruskal–Wallis $p = 0.0012$, effect size = 0.222, Fig. 5b). A total of 74% of the tail genes were not identified as differentially abundant in cancer samples versus controls (Fig. 5c). Moreover, for 40.0% of cancer sample tail genes that were differentially abundant between cancer and controls, the direction (i.e. positive or negative z -score) in at least one cancer sample was opposite to the direction reported by the differential abundance analysis, further highlighting the heterogeneity among patients. Notably, the recurrence of tail genes was much higher in cancer samples than in controls; 98% of the control tail genes were identified in only one control sample, while 47% of cancer tail genes were identified in more than one cancer sample (Fig. 5d).

To determine the significance of the observed results, we randomly swapped sample labels prior to tail gene identification and repeated this procedure 20 times. Doing so, we did not observe a significant difference in number of tail genes between the newly labeled cancer samples and controls in any of the repeats (all Kruskal–Wallis $p \geq 0.05$), confirming that an increased number of tail genes is specifically associated with cancer samples.

Tail genes detected in one cancer type showed significant overlap with tail genes detected in another cancer type (Fisher’s exact tests $p \leq 5.4E-168$, odds ratio: [11.1–15.2]), yet limited similarity (Jaccard index: [0.197–0.348]) (Fig. 5e). The 1182 unique control tail genes also showed significant overlap with any of the cancer tail gene subsets (pairwise Fisher exact tests p : [5.9E–15 to 8.5E–32], odds ratio: [2.1–2.5]) with less similarity (Jaccard index: [0.074–0.119]). A total of 63% of control tail genes were not identified as tail genes in any of the cancer samples.

Seven Cancer Gene Census oncogenes were identified as tail genes ($z > 3$) in at least one sample, which is significantly more than expected by chance (Fisher’s exact $p = 0.019$, odds ratio = 3.13). While most oncogenes were only identified as tail gene in a single sample, *KAT6A* was found in seven cancer samples (4 OV, 2 PRAD, and 1 UCEC) but also in one control sample. Overrepresentation analyses (Supplementary Fig. 8 and Supplementary Data 4) demonstrated that tail genes that are higher abundant in at least one cancer sample ($z > 3$) were involved in different processes, including platelet activation and signaling, actin binding and dynamics, and extracellular matrix. Tail genes that are lower abundant in at least one cancer sample ($z < -3$) were mainly related to mRNA splicing and translation. Significantly enriched processes in the cancer tail gene sets were not enriched in the control tail gene set, and processes enriched in the lower-abundant (cancer) tail genes showed higher recurrence across the three cancer types compared to those enriched in the higher-abundant tail genes.

Given the higher number of tail genes in cancer versus controls, we further assessed the binary (cancer/control) classification potential of the number of tail genes per sample. We used a 10×5 -fold cross-validation in which we first selected a subset of tail genes, further coined biomarker tail genes (BTG), that were significantly associated to either the cancer or control state using Fisher’s exact tests on the training set (see “Methods”). This resulted in sets of 43 to 549 BTG. None of these BTG were recurrent tail genes in the control samples, and for BTG identified in at least half of the cross-validation folds, 43% were not identified as differentially abundant in any cancer type compared to controls (Supplementary Fig. 8c). Applying the same biomarker tail gene selection strategy to 10 iterations of randomly swapping sample labels resulted in considerably fewer or no biomarker tail genes at all: between 0 and 6 BTG (median: 0) instead of 43 to 549.

When selecting biomarker tail genes in a specific cancer group versus controls, we obtained 17 to 271 biomarker tail genes for OV, 6 to 82 for PRAD, and 2 to 16 for UCEC. We then built a classifier for discriminating cancer and control samples based on the number of tail genes belonging to the BTG set in a particular sample. Using 10×5 -fold cross-validation, we obtained a high binary classification for ovarian or prostate cancer samples and controls (mean AUC = 0.901 and 0.964, resp.), yet a lower classification performance for uterine cancer samples and controls (mean AUC = 0.586) (Fig. 5f). Classification performance was still high when combining cancer samples for BTG identification (mean AUC = 0.879), with a better performance for male donors compared to female donors (Supplementary Data 5).

The consensus set of BTG present in at least half of the 10×5 -fold cross-validations, consisted of 38 BTG for OV, 18 for PRAD, 3 for UCEC, and 63 for all three cancer types combined. OV and UCEC consensus BTG sets did not overlap. The overlap for OV and PRAD, as well as for PRAD and UCEC, was significant, but the similarity was limited (Fisher’s exact $p \leq 1.6E-2$; odds ratio: 18.5 and 94.7; Jaccard index: 0.06 and 0.05) (Fig. 5h).

The difference in number of consensus biomarker tail genes between cancer and control samples per biomarker tail gene subset was more pronounced than based on all tail genes (two-sided Wilcoxon rank-sum: for OV $p = 7.3E-8$ and effect size = 0.876; for PRAD $p = 1.0E-8$ and effect size = 0.920; for UCEC $p = 9.4E-5$ and effect size = 0.629; for all cancers combined $p = 1.8E-10$ and effect size = 0.810) (Fig. 5b, g).

Together, these analyses suggest that individual cancer samples can be distinguished from controls based on the number of biomarker tail genes in their plasma cRNA profiles.

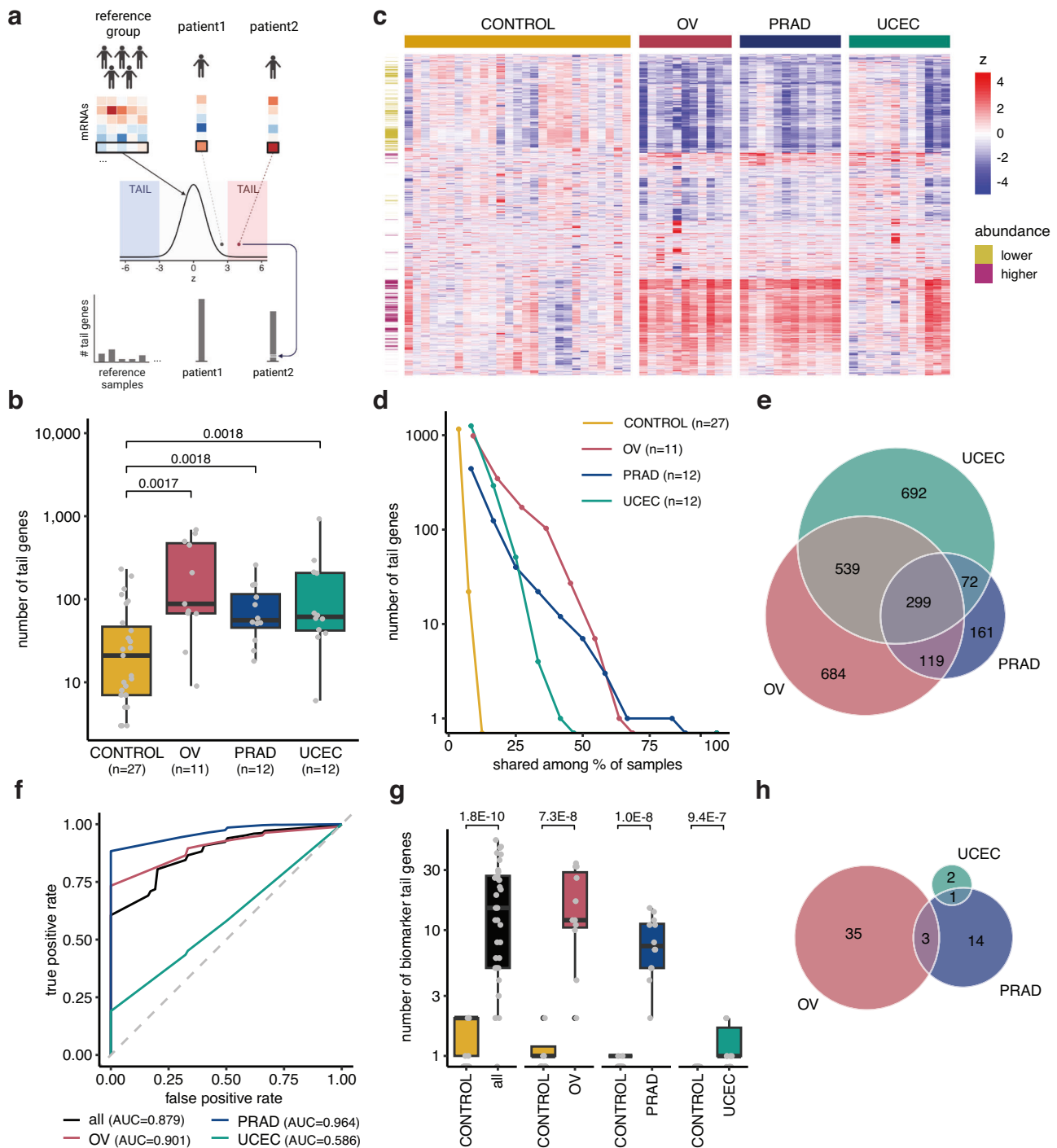


Fig. 5 | Higher number of (biomarker) tail genes in cancer patient plasma samples compared to controls. **a** Tail gene identification workflow. A gene in a particular sample for which the abundance deviates more than 3 standard deviations from the reference group abundance distribution is a tail gene for that sample. The number of tail genes is used as a metric to determine cancer/control status. Figure created with BioRender.com. **b** Boxplot of the number of tail genes per group, with gray 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 are indicated in the plot. **c** Z-score distribution of tail genes in the different samples, all 3311 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 genes detected in the different cancer types. **f** mean ROC curves of 10×5 -fold cross-validation for binary cancer/control classification in test sets based on the number of biomarker tail genes

for control and ovarian cancer samples (OV), prostate cancer samples (PRAD), uterine cancer samples (UCEC), or all cancer samples (all) using the respective biomarker tail gene sets in the training sets. **g** Boxplot of the number of 50% consensus biomarker tail genes per subset in control (CONTROL) versus the respective cancer samples. Total number of biomarker tail genes considered for each comparison: 63 for all cancer ($n = 35$) versus control, 38 for OV versus control, 18 for PRAD versus control, 3 for UCEC versus control. Gray dots represent individual sample counts. Two-sided Wilcoxon rank-sum p values are indicated in plot. **h** Overlap of cancer type-specific biomarker tail genes. 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 $1.5 \times$ inter-quartile range (Q3–Q1) from that quartile. More extreme points are plotted as individual black dots. 62 plasma samples were considered in total from 27 cancer-free controls (CONTROL), 11 ovarian cancer (OV), 12 prostate cancer (PRAD), and 12 uterine cancer patients (UCEC).

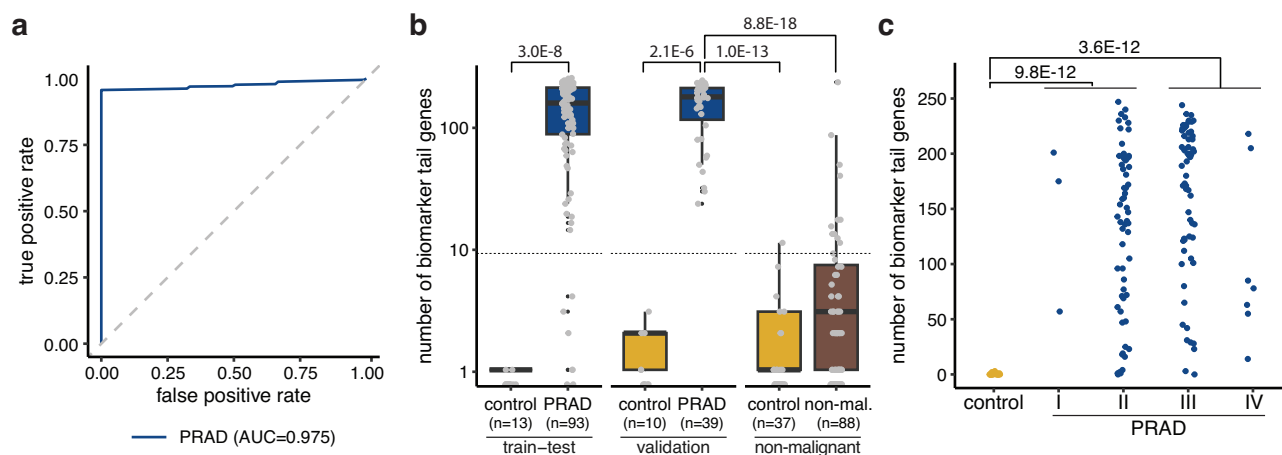


Fig. 6 | Biomarker tail genes separate prostate cancer from healthy controls and non-malignant controls. **a** Mean ROC curves of 10×5 -fold cross-validation for binary cancer/control classification in test sets based on the number of biomarker tail genes for control and prostate cancer samples (PRAD) using the respective biomarker tail gene sets in the training sets. **b** Boxplot of the number of consensus biomarker tail genes ($n = 260$), with gray dots representing individual sample counts. The prostate cancer cohort samples are split in the train-test cohort (used for consensus biomarker tail gene identification) and the independent validation cohort. Only male controls are shown for the prostate cancer cohorts (13 in train-test, 10 in validation); female controls (21 in train-test, 4 in validation) were part of the mixed-

sex reference for z-score calculations. Horizontal dashed line: classification threshold; PRAD: plasma from prostate cancer patients; non-mal.: plasma samples from patients with a non-malignant disease. **c** number of biomarker tail genes of train-test and validation samples visualized per cancer stage. Pairwise comparisons in **b** and **c** were performed using two-sided Wilcoxon rank-sum tests, with p values adjusted for multiple testing using the Benjamini–Hochberg method. 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 $1.5 \times$ interquartile range (Q3–Q1) from that quartile. More extreme points are plotted as individual black dots.

High sensitivity and specificity of prostate cancer biomarker tail genes

Given the limited number of samples in the initial analyses, we further assessed the potential for prostate cancer detection in a larger, independent cohort comprising 132 prostate cancer patient plasma samples and 48 healthy control plasma samples (23 male donors). mRNA capture sequencing of the prostate cancer plasma cohort resulted in 35 to 51 M paired-end reads per sample (median 43 M) after quality filtering (Supplementary Fig. 9a–c). Due to differences in isolation and library preparation protocols compared to the original three-cancer cohort, we first identified prostate cancer BTG in this cohort and established a fixed classification threshold using 70% of the samples (test-train set). The 10×5 -fold cross-validation, using only male donor samples, resulted in 240 to 335 BTG and a mean AUC of 0.975 over all test sets (Fig. 6a and Supplementary Data 5). The consensus set consisted of 260 BTG and the classification threshold was set at 9 BTG per sample (92.5% sensitivity and 100% specificity). In the validation set (remaining 30% of the cohort), we used the controls from the test-train set as a reference for z-score calculation in every sample and used the same consensus BTG and threshold. This resulted in 100% sensitivity and 100% specificity for prostate cancer detection in male donor samples (Fig. 6b).

Note that the number of BTG detected in cancer patient samples was not associated with cancer stage (Jonckheere–Terpstra $p = 0.1021$, Fig. 6c). Although the cancer-free controls were generally younger than the prostate cancer patients in the cohort, classification performance remained high when using an age-matched subset of male control samples as the validation group (97.1% sensitivity, 100% specificity, Supplementary Fig. 9d, e), suggesting that age did not influence classification outcomes. The prostate cancer BTG also resulted in a significant difference between the prostate cancer and control group in the three-cancer cohort (Wilcoxon $p = 0.0009$; Supplementary Fig. 9f). However, the established threshold could not be directly applied to this cohort due to differences in pre-analytical variables (plasma preparation, RNA extraction input volume, and library preparation method), highlighting the known need for workflow standardization when comparing samples.

To analyze cancer specificity of the prostate cancer BTG classification algorithm, we analyzed a similarly processed cohort of plasma samples from

healthy individuals ($n = 37$) and patients with non-malignant conditions ($n = 88$), including inflammatory and autoimmune conditions, respiratory and liver diseases, and monogenic or neurological disorders (Fig. 6b). The number of prostate cancer BTGs in the newly collected control cohort was similar to that in the control cohorts of the prostate cancer train-test and validation cohort (Wilcoxon $q = 0.35$, Fig. 6b). Samples from patients with non-malignant conditions showed more prostate cancer BTGs, but there was still a significant difference in number of BTG between the non-malignant and the prostate cancer group (Wilcoxon $q = 8.8E-18$). Applying the previously established prostate cancer threshold yielded a specificity of 89.6% (1/37 control and 12/88 non-malignant patient samples misclassified). False positives included samples from patients with systemic sclerosis (2/14), cirrhosis (3/10), COPD (2/8), asthma (1/10), vasculitis (1/1), dermatitis (1/1), pneumonia (1/1), and sarcoidosis (1/1). Importantly, none of the five samples from patients diagnosed with benign prostatic hyperplasia (BPH) were misclassified as prostate cancer.

Together, these results further underscore the potential of cancer classification based on biomarker tail genes in plasma.

Classification potential in other cancer types and biofluids

To independently assess the classification potential of the number of biomarker tail genes in other cancer types, we generated additional plasma cfRNA profiles (lymphoma cohort), including plasma samples from 22 cancer-free control donors and 43 lymphoma patients, more specifically 30 diffuse large B-cell lymphoma (DLBCL) and 13 primary mediastinal large B-cell lymphoma (PMBCL) patients. We identified 5374 unique tail genes, again with more tail genes in cancer samples compared to controls (Kruskal–Wallis $p = 2.9E-5$, effect size = 0.305, Supplementary Fig. 10a). All control tail genes were identified in only one control sample while 45.9% of DLBCL and 44.2% of PMBCL tail genes were identified in more than one DLBCL and PMBCL sample, respectively (Supplementary Fig. 10b). When swapping sample labels (20 repeats) prior to tail gene identification, no significant differences in the number of tail genes between cancer and control samples were observed in any of the repeats (Kruskal–Wallis $p \geq 0.05$).

Among the tail genes, 64 to 326 lymphoma BTG, 21 to 220 DLBCL BTG, and 87 to 677 PMBCL BTG were identified using Fisher’s exact tests in

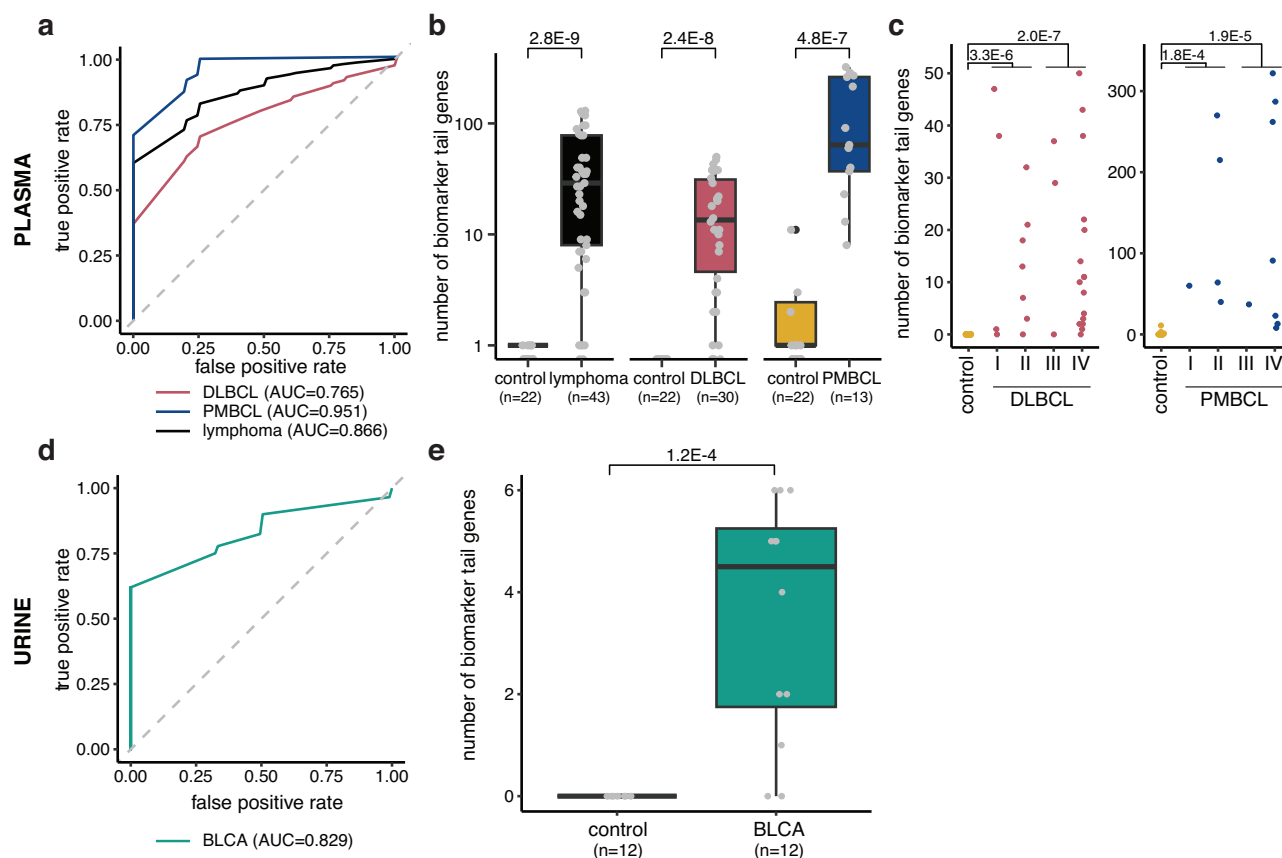


Fig. 7 | Number of biomarker tail genes is significantly higher in cancer versus control samples in an independent blood plasma and urine cohort.

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 10 × 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 the 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. Gray dots represent individual sample counts. **c** Number of biomarker tail genes of **b** visualized per cancer stage. Pairwise comparisons in **b** and **c** were performed using two-sided Wilcoxon rank-sum tests, with *p* values adjusted for multiple testing using the Benjamini–Hochberg method. **d** Mean ROC curve of 10 × 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 are considered in total. Comparison between control and cancer samples was performed using a two-sided Wilcoxon rank-sum test. Boxplots: lower quartile, median, upper quartile, whiskers of 1.5* interquartile range; more extreme points are indicated by black dots.

10 × 5-fold cross-validation. The binary classifiers (cancer/control) based on biomarker tail genes in the training sets resulted in a mean AUC of 0.765 for DLBCL, 0.951 for PMBCL, and 0.866 for all lymphoma in the test sets (Fig. 7a). Again, we observed a better performance for samples from male compared to female donors (Supplementary Data 5). When looking at biomarker tail genes that were identified in at least half of the iterations, we found 60 consensus BTG for DLBCL, 371 for PMBCL, and 151 for lymphoma. The number of tail genes belonging to these sets was again significantly higher in cancer compared to control samples (Fig. 7b). 59% of these consensus BTG were not differentially abundant in any lymphoma subtype versus controls. Notably, early-stage lymphoma patients also displayed an increased number of biomarker tail genes (Fig. 7c) and the number of detected BTG per sample was not significantly linked to cancer staging (Jonckheere–Terpstra *p* = 0.9374 for DLBCL and *p* = 0.8949 for PMBCL).

To test if our cancer-control classification strategy based on biomarker tail genes is also applicable to other biofluids, we applied it to a cohort of urine cfRNA profiles (bladder cancer cohort)⁷. This cohort included urine samples of 12 bladder cancer patients (BLCA) and 12 cancer-free control donors. We identified 1152 tail genes in total and again only observed recurrence in cancer samples (Supplementary Fig. 10c, d). A total of 10 × 5-fold cross-validation yielded 2 to 93 biomarker tail genes and the binary

classifiers based on these biomarker tail gene sets resulted in a mean AUC of 0.829 in the test sets (Fig. 7d). Six biomarker tail genes were present in at least half of the iterations and the number of tail genes belonging to this 50% consensus set was significantly higher in bladder cancer versus control samples (Fig. 7e). In line with the other datasets, sample label swapping prior to tail gene identification did not result in significant differences in the number of tail genes between the newly labeled cancer samples and controls.

Taken together, our results confirm that the number of biomarker tail genes in blood plasma cfRNA enables accurate classification of cancer patients from controls and suggest that the concept may also be applicable to cfRNA profiles from other biofluids.

Discussion

This study demonstrated that cancer induces measurable alterations in blood plasma cfRNA profiles, although the detectability and specificity of tumor-derived signals vary substantially across cancer types. Results from our pan-cancer discovery cohort revealed that AML was the only cancer type for which there is direct evidence for tumor-derived cfRNA in plasma in the form of *PML::RARA* fusion transcripts. Also, deconvolution of AML plasma cfRNA revealed a high fraction of myeloid progenitor cells, providing additional—albeit indirect—evidence for tumor-derived cfRNA. For solid tumor types, we identified differentially abundant mRNAs in plasma

for every cancer versus control comparison, but a cancer type-specific signal was only observed in plasma from liver cancer patients through the enrichment of liver-specific gene sets. Liver is a major contributor to the plasma transcriptome⁴ and its high vascularization and blood flow may facilitate mRNA release in blood upon liver cancer progression. Increased cfDNA contributions from the liver have also been observed in liver cancer patients³³. Note that, despite the enrichment of liver-related mRNAs, the total cfRNA concentration was not increased in plasma from liver cancer patients, indicating that this signal was still strongly diluted.

Considering the organ contribution and dilution argument, other types of liquid biopsies may be better suited to monitor tumor-derived cfRNA for other cancer types. For instance, we could not detect enrichment of prostate tissue-related nor prostate cancer-specific transcripts in blood plasma from prostate cancer patients, in contrast to earlier findings in seminal plasma samples³⁴.

We did observe several general trends in patients' blood plasma transcriptomes in the pan-cancer cohort, which were confirmed in an independent cohort. The lower abundance of immune-related mRNAs may be linked to the suppressed immune system status in cancer patients. It was recently shown that local tumors can have a systemic impact on the immune system, and that different tumor sites of origin resulted in distinct relative abundances and activity profiles of immune cell types³⁵. This may be linked to the observed variations in immune-related mRNA abundance, both between and within cancer types. The (immune) cell-related mRNA contributions agreed with reported cell abundance trends found in blood from several cancer patients³⁶, with the exception of neutrophil-related mRNAs that are lower abundant in plasma. However, neutrophils can have a dual role in cancer³⁷ and reduced neutrophil cell-free mRNA abundance may result from neutrophils specifically recruited to tumor sites. Finally, the negative enrichment of blood/erythrocyte gene sets may be linked to anemia, a common condition in cancer patients. Among the processes that we found positively enriched in cancer patient plasma, some are linked to hallmarks for cancer progression and metastasis, including epithelial-to-mesenchymal transition and increased blood vessel signal that stem from leaky vasculature and angiogenesis³⁸.

A limitation of the pan-cancer cohort is that control samples were not always perfectly age and sex-matched. The anaplastic astrocytoma, cervical cancer, and testicular germ cell tumor patients were younger, as these types predominantly appear at a younger age. A dedicated cohort with matched control samples may be better for revealing cancer-related changes in these cancer types. Similarly, while male and female donors were balanced in the control group, this was not always the case for some cancer types due to sex-specificity (e.g., prostate, testicular, breast, cervical, and ovarian cancer) or sample availability. Nevertheless, at most 4 out of 8 donors were not sex-matched in cancer vs. control group comparisons in the pan-cancer cohort, and our main conclusions were not impacted when only using controls of the same sex in the three-cancer cohort. Another limitation is the relatively small sample size per cancer type, which might contribute to false positive results or insufficient power to detect true differences.

The variable mRNA abundance in plasma, combined with the limited overlap of differentially abundant mRNAs across cohorts of the same cancer type, prompted an alternative approach that can effectively deal with patient heterogeneity. The tail gene concept satisfies this requirement by identifying mRNAs whose abundance in individual samples significantly deviates from the abundance distribution in a reference group. While the dark channel biomarker gene concept⁴ is in part based on a more patient-centric approach, it also requires differential abundance at the group level to be considered as potential biomarkers. Moreover, the dark channel concept looks explicitly at genes that are absent (dark) in control samples and therefore cannot capture changes in transcript abundance of genes that are also present in healthy plasma transcriptomes. Of note, 19 dark channel biomarker genes were identified for lung and breast cancer by Larson et al.⁴, but five of those genes could be readily detected in control samples from our cohorts, indicating that the "darkness" (i.e., absence of gene transcripts in control samples) should be re-evaluated in different cohorts.

Tail gene counts are a simple metric for classifying a biofluid sample as originating from a cancer patient or control. Considering that it is not based on gene signatures with weighted contributions of individual features, we propose it is more robust and compatible with observed patient heterogeneity. Nevertheless, it requires a large and representative group of cancer-free control samples. A control group only needs to be established once, if the entire workflow—from sample collection to RNA-sequencing—is standardized across cancer and control samples. This approach should minimize the introduction of batch effects that may skew cfRNA distributions between a pre-established cohort of cancer-free controls and newly profiled cancer or control samples. In this regard, our lab has demonstrated that preanalytical variables (blood collection tube, time between blood collection and plasma preparation, RNA extraction kit) can indeed dramatically impact cfRNA profiles³⁹.

Deviations in the number of tail genes are not necessarily restricted to cancer patients, nor to blood plasma. Indeed, cfRNA from urine provides a good source of biomarker tail genes to identify bladder cancer patients. Some of the identified biomarker tail genes have previously been linked to cancer, with examples provided in Supplementary Analysis 3. Notably, other diseases may also be characterized by an increase in tail genes, though the specific (biomarker) tail genes may differ.

Given the limited sample sizes for some cancer types, we likely did not capture all variability in the current tail gene sets, nor did we identify all possible tail genes. Follow-up studies with larger cohorts are recommended to establish robust gene sets. For prostate cancer, we identified 260 biomarker tail genes that perfectly separated plasma samples from prostate cancer patients and healthy controls in the validation cohort. Evaluating the performance of prostate cancer BTG in a first cohort of non-malignant diseases resulted in a specificity of 89.6%. Note that the classification threshold could be raised to further improve specificity in this cohort without affecting performance in the validation cohort. Moreover, the non-malignant cohort included various conditions that are often relatively rare in the target population for prostate cancer testing. We therefore hypothesize that specificity in broader, population-based screening settings may be higher. However, the applicability and overlap of tail genes in non-malignant disease cohorts should still be examined further. Similarly, while our results show that biomarker tail genes are also significantly more numerous in low-stage cancer patients (Figs. 6c and 7c), additional follow-up studies comprising different stages across various cancer types could reveal whether the concept is suited for early-stage cancer screening or monitoring.

This study served to evaluate the feasibility of a cfRNA-based biomarker strategy for cancer detection, and in particular the tail gene concept, rather than to establish a clinically validated assay. Within this framework, the prostate cancer cohort represents an informative test case. Collected over four years by multiple study nurses and across several operating rooms, this cohort incorporates meaningful real-life collection variability while maintaining harmonized pre-analytical protocols. In this technically consistent setting, internal independent validation confirmed that the tail gene signal is analytically robust.

The study design aligns with current methodological recommendations for biomarker prediction models and omics-based classifiers, where internal validation through train-test splits or cross-validation is regarded as an essential component of model development. External validation on fully independent cohorts remains the gold standard for establishing generalizability, but was beyond the scope of the present work due to data constraints. However, an exploratory transfer test shows that the underlying biological pattern (an increased number of tail genes in prostate cancer samples) is preserved across independent cohorts, yet absolute classifier thresholds do not transfer directly owing to substantial pre-analytical differences among these cohorts. Also, almost all control samples (36/37) in the cross-cohort evaluation were correctly classified when analytical conditions were comparable, but performance decreased under divergent pre-analytical protocols. These findings confirm that the BTG signal is preserved under comparable analytical conditions, but that broader protocol

differences shift the overall cfrRNA distribution, necessitating harmonization of pre-analytical variables for reliable cross-site application.

Conclusion

This study demonstrates the potential of the tail gene concept as a cfrRNA-based strategy for cancer detection. By quantifying sample-specific deviations, this approach appears compatible with patient heterogeneity while retaining analytical robustness under harmonized conditions.

Future studies using larger, prospectively collected, and pre-analytically standardized multi-center cohorts will be essential to externally validate these findings, refine clinically applicable thresholds, and determine the utility of tail gene metrics for early detection and disease monitoring applications.

Data availability

Raw RNA-sequencing data are available under controlled access in the European Genome-phenome Archive (EGA) due to the presence of potentially identifiable genetic information and in accordance with participant informed consent and the European General Data Protection Regulation (GDPR). The pan-cancer and three-cancer cohort FASTQs can be found under study EGAS00001006755 (dataset EGAD00001009713). The prostate cancer cohort FASTQs can be found under study EGAS50000001265 (datasets EGAD50000001805). The non-malignant cohort FASTQs can be found under study and EGAS50000001264 (dataset EGAD50000001806). The lymphoma cohort FASTQs can be found under study EGAS00001007127 (dataset EGAD00001010259). The bladder cancer cohort FASTQs can be found under study EGAS00001003917 (dataset EGAD00001005439). Access to these controlled datasets is granted by the relevant Data Access Committee (DAC) following submission and approval of a data access request through the EGA portal. Researchers must provide a research proposal and agree to the terms of a Data Access Agreement restricting use to approved research purposes and prohibiting re-identification attempts. Requests are typically reviewed within 4–8 weeks. Differential abundance and gene set enrichment analysis results are available (open access) in a Zenodo repository (<https://doi.org/10.5281/zenodo.7953707>)⁴⁰. An overview of tail genes per type and 50% consensus biomarker tail genes can be found in Supplementary Data 6. Source data for Figs. 1–7 can be found in Supplementary Data 7.

Code availability

Software versions used for data analysis were: bedtools v2.30.0; cutadapt v1.18; FASTQC v0.11.9; FusionCatcher v1.30; HTSeq v0.11.0; samtools v1.16.1; STAR v2.7.3; UMI-tools v1.0.0; and R v4.2.1. The following R packages were used for downstream analyses: biomaRt v2.62.0; caret v6.0.94; data.table v1.16.4; DESeq2 v1.36.0; EnhancedVolcano v1.24.0; eulerr v6.1.1; fgsea v1.22.0; GeneOverlap v1.32.0; ggpubr v0.6.0; gridExtra v2.3; here v1.0.1; pheatmap v1.0.12; pROC v1.18.0; rstatix v0.7.2; tidyverse v1.3.2. Custom R scripts used for differential abundance analysis, tail gene identification, and downstream statistical analyses are available at <https://github.com/OncoRNALab/tailgenes> and have been archived on Zenodo (v1.0.0) with <https://doi.org/10.5281/zenodo.18669199>⁴¹. The archived version corresponds to the code used for this study.

Received: 29 January 2026; Accepted: 24 February 2026;

Published online: 07 March 2026

References

- Pessoa, L. S., Heringer, M. & Ferrer, V. P. ctDNA as a cancer biomarker: a broad overview. *Crit. Rev. Oncol. Hematol.* **155**, 103109 (2020).
- Vorperian, S. K., Moufarrej, M. N. & Quake, S. R. Cell types of origin of the cell-free transcriptome. *Nat. Biotechnol.* **40**, 855–861 (2022).
- Moufarrej, M. N. et al. Early prediction of preeclampsia in pregnancy with cell-free RNA. *Nature* **602**, 689–694 (2022).
- Larson, M. H. et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).
- Chen, S. et al. Cancer type classification using plasma cell-free RNAs derived from human and microbes. *eLife* **11**, e75181 (2022).
- Roskams-Hieter, B. et al. Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. *npj Precis. Oncol.* **6**, 1–11 (2022).
- Hulstaert, E. et al. Charting extracellular transcriptomes in the human biofluid RNA atlas. *Cell Rep.* **33**, 108552 (2020).
- Hardwick, S. A. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).
- Hulstaert, E. et al. Messenger RNA capture sequencing of extracellular RNA from human biofluids using a comprehensive set of spike-in controls. *STAR Protoc.* **2**, 100475 (2021).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
- Tol, P. Colour schemes, Technical Note SRON/EPS/TN/09-002. (2012).
- Mestdagh, P. et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods* **11**, 809–815 (2014).
- Nicorici, D. et al. FusionCatcher—a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Preprint at <https://doi.org/10.1101/011650> (2014).
- Kim, P. & Zhou, X. FusionGDB: fusion gene annotation DataBase. *Nucleic Acids Res.* **47**, D994–D1004 (2019).
- Gao, Q. et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238.e3 (2018).
- Korotkevich, G. et al. Fast gene set enrichment analysis. Preprint at <https://doi.org/10.1101/060012> (2021).
- Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
- Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
- Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
- Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).
- Lorenzi, L. et al. The RNA Atlas expands the catalog of human non-coding RNAs. *Nat. Biotechnol.* **39**, 1453–1465 (2021).
- Everaert, C., Volders, P.-J., Morlion, A., Thas, O. & Mestdagh, P. SPECS: a non-parametric method to identify tissue-specific molecular features for unbalanced sample groups. *BMC Bioinforma.* **21**, 58 (2020).
- Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
- Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
- Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).
- Bryzgunova, O. E., Konoshenko, M. Y. & Laktionov, P. P. Concentration of cell-free DNA in different tumor types. *Expert Rev. Mol. Diagn.* **21**, 63–75 (2021).
- Liquori, A. et al. Acute promyelocytic leukemia: a constellation of molecular events around a single PML-RARA fusion gene. *Cancers* **12**, 624 (2020).

30. Khwaja, A. et al. Acute myeloid leukaemia. *Nat. Rev. Dis. Prim.* **2**, 1–22 (2016).
31. Giannakeas, V. et al. Analysis of platelet count and new cancer diagnosis over a 10-year period. *JAMA Netw. Open* **5**, e2141633 (2022).
32. Haemmerle, M., Stone, R. L., Menter, D. G., Afshar-Kharghan, V. & Sood, A. K. The platelet lifeline to cancer: challenges and opportunities. *Cancer Cell* **33**, 965–983 (2018).
33. Sun, K. et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. USA* **112**, E5503–E5512 (2015).
34. Hulstaert, E. et al. Exploring the extracellular transcriptome in seminal plasma for non-invasive prostate cancer diagnosis. Preprint at <https://doi.org/10.1101/2021.05.11.21256306> (2021).
35. Allen, B. M. et al. Systemic dysfunction and plasticity of the immune macroenvironment in cancer models. *Nat. Med.* **26**, 1125–1134 (2020).
36. Hiam-Galvez, K. J., Allen, B. M. & Spitzer, M. H. Systemic immunity in cancer. *Nat. Rev. Cancer* **21**, 345–359 (2021).
37. Xiong, S., Dong, L. & Cheng, L. Neutrophils in cancer carcinogenesis and metastasis. *J. Hematol. Oncol.* **14**, 173 (2021).
38. Lugano, R., Ramachandran, M. & Dimberg, A. Tumor angiogenesis: causes, consequences, challenges and opportunities. *Cell. Mol. Life Sci.* **77**, 1745–1770 (2020).
39. The exRNAQC Consortium et al. Blood collection tube and RNA purification method recommendations for extracellular RNA transcriptome profiling. *Nat. Commun.* **16**, 4513 (2025).
40. Morlion, A. Differential abundance and gene set enrichment in plasma of cancer patients versus controls. *Zenodo* <https://doi.org/10.5281/zenodo.12721611> (2024).
41. Morlion, A. OncoRNALab/tailgenes: code for ‘Patient-specific alterations in blood plasma cfRNA profiles enable accurate classification of cancer patients and controls’. *Zenodo* <https://doi.org/10.5281/zenodo.18669199> (2026).

Acknowledgements

This work was supported by Ghent University (BOF; BOF19/DOC/228), Ghent University Industrial Research Fund (IOF; F2023/IOF-ConceptTT/068), Ghent University Hospital, Fund for Scientific Research Flanders (FWO; 11C1623N predoctoral fellowship to A.M., 11H7523N predoctoral fellowship to P.D., G0B2820N), the Belgian Foundation Against Cancer (STK; 2025-085 postdoctoral fellowship to A.M.), and Kom Op Tegen Kanker (Stand up to Cancer, the Flemish cancer society). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government—department EWI. We would like to thank Tim Mercer for providing the Sequin spikes, Eva Hulstaert and Francisco Avila Cobos for their support during the first phase of the project, Anneleen Decock for her feedback on the initial draft, and Scott Ailliet, Sofie Roelandt, and Deyna Keppens for their assistance with blood collection. We would also like to thank the Biostatistics Unit of the Faculty of Medicine and Health Sciences (Ghent University) for their feedback on the tail gene performance assessment.

Author contributions

Conceptualization: A.Mo., C.E., J.V., and P.M.; Data curation: A.mo.; Formal analysis: A.Mo. and O.T.; Funding acquisition: A.Mo., J.V., J.V.D., and P.M.; Investigation: E.V., J.N., K.V., and N.J.R.; Methodology: A.Mo., J.V., and P.M.; Project administration: A.Mo., J.V., and P.M.; Resources: A.Ma., A.V., D.H., E.H., F.H., F.O., G.B., H.V., J.L., J.V., J.V.D., K.P., K.S., L.V., N.L., P.D., P.M., S.C., and V.S.; Software: A.Mo. and J.A.; Supervision: J.V. and P.M.; Validation: A.Mo. and O.T.; Visualization: A.mo.; Writing—original draft: A.Mo., J.V., and P.M.; Writing—review & editing: all authors.

Competing interests

A.M., J.V., and P.M. are inventors on the international PCT application “Method to screen, diagnose or monitor treatment using increased numbers of deviating analytes” (WO2024200811; published 03/10/2024) covering the tail gene concept. They are also inventors on the European priority application “Blood plasma RNA profiles enable accurate diagnosis of prostate cancer” (EP26151329.5; filed 12/01/2026) covering specific prostate cancer biomarker tail genes from the prostate cancer cohort in this manuscript. Both applications are assigned to Ghent University.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43856-026-01507-8>.

Correspondence and requests for materials should be addressed to Pieter Mestdagh.

Peer review information *Communications Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹Department of Biomolecular Medicine, Ghent University, Ghent, Belgium. ²OncoRNALab, Cancer Research Institute Ghent (CRIG), Ghent, Belgium. ³Department of Hematology, Ghent University Hospital, Ghent, Belgium. ⁴TOBI Lab, Center for Medical Biotechnology, VIB-UGent, Zwijnaarde, Belgium. ⁵Department of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium. ⁶Department of General Internal Medicine, Ghent University Hospital, Ghent, Belgium. ⁷Department of Internal Medicine and Pediatrics, Ghent University, Ghent, Belgium. ⁸Department of Pediatric Respiratory and Infectious Medicine, Ghent University Hospital, Ghent, Belgium. ⁹Department of Neurology, Ghent University Hospital, Ghent, Belgium. ¹⁰Department of Intensive Care Medicine, Ghent University Hospital, Ghent, Belgium. ¹¹Department of Dermatology, Ghent University Hospital, Ghent, Belgium. ¹²Department of Urology, Ghent University Hospital, Ghent, Belgium. ¹³Department of

Rheumatology, Ghent University Hospital, Ghent, Belgium. ¹⁴Unit for Molecular Immunology and Inflammation, Inflammation Research Center (IRC), VIB-UGent, Zwijnaarde, Belgium. ¹⁵Department of Pathology, Ghent University Hospital, Ghent, Belgium. ¹⁶Hepatology Research Unit, Liver Research Center, Ghent, Belgium. ¹⁷Department of Gastroenterology and Hepatology, Ghent University Hospital, Ghent, Belgium. ¹⁸Department of Physical Medicine and Rehabilitation, Center for Integrative Medicine, Ghent University Hospital, Ghent, Belgium. ¹⁹Department of Head and Skin, Ghent University, Ghent, Belgium. ²⁰Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. ²¹Data Science Institute, I-BioStat, Hasselt University, Hasselt, Belgium. ²²National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, NSW, Australia. ²³These authors jointly supervised this work: Jo Vandesompele, Pieter Mestdagh. ✉e-mail: pieter.mestdagh@ugent.be