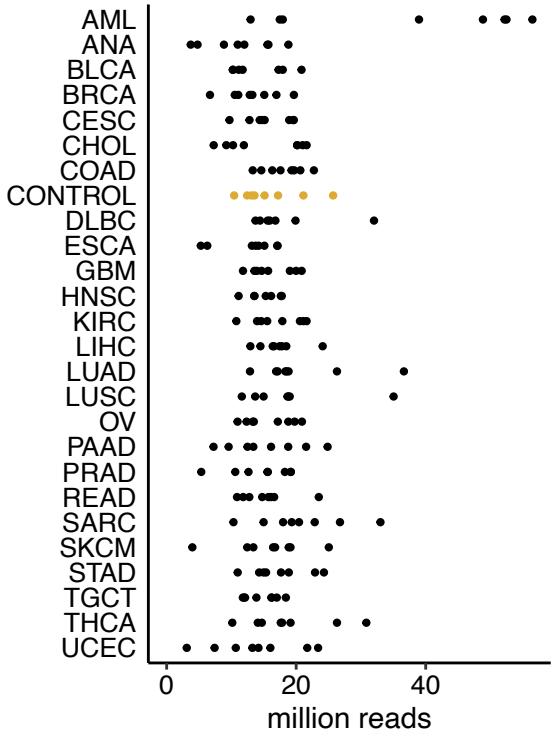
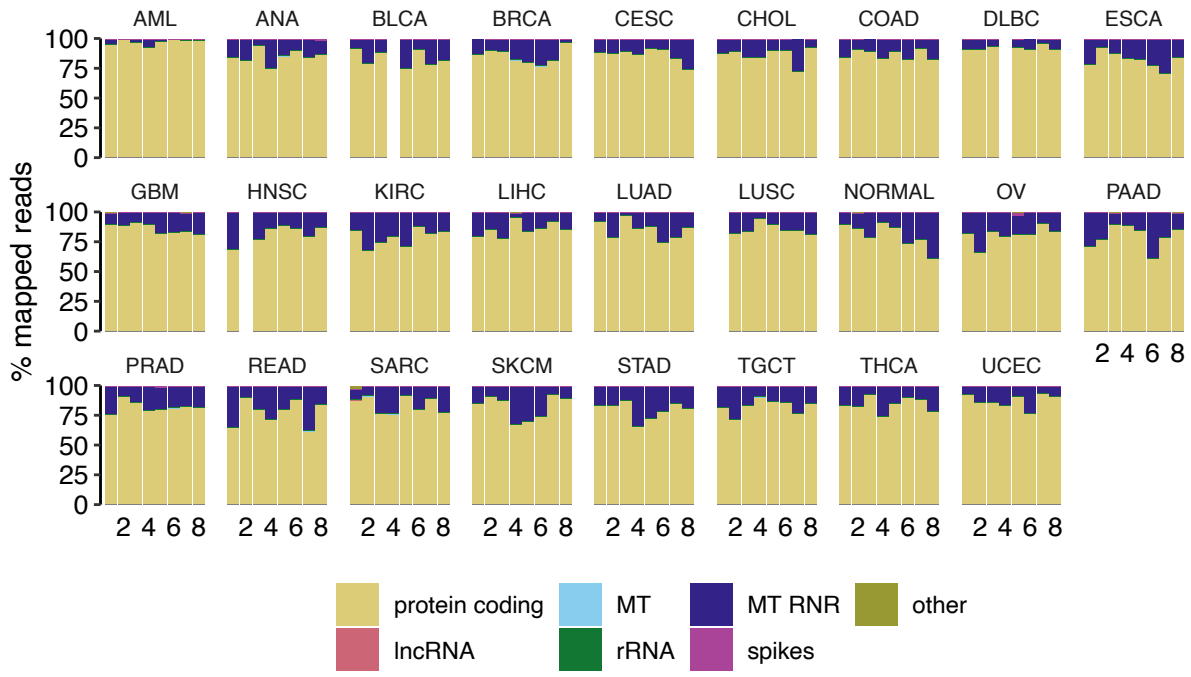
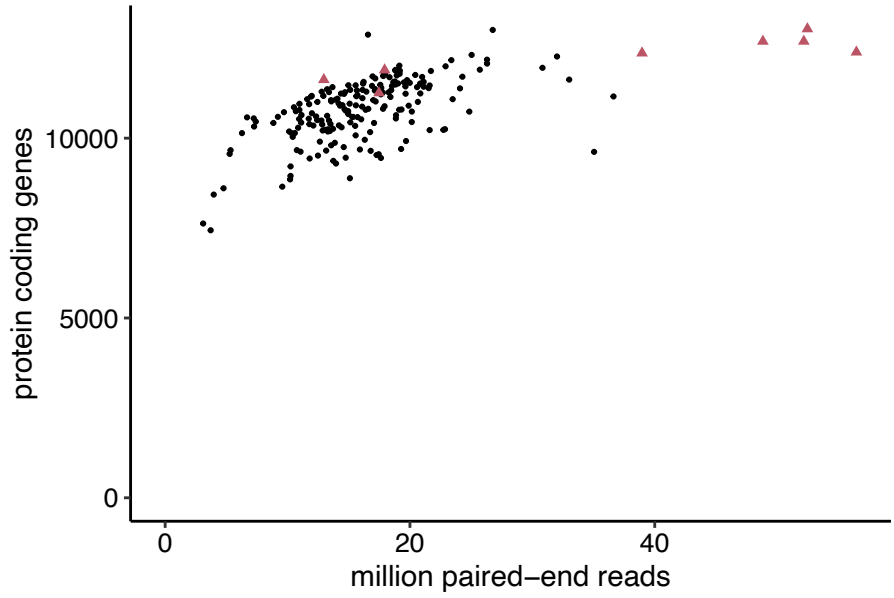
Supplemental figures



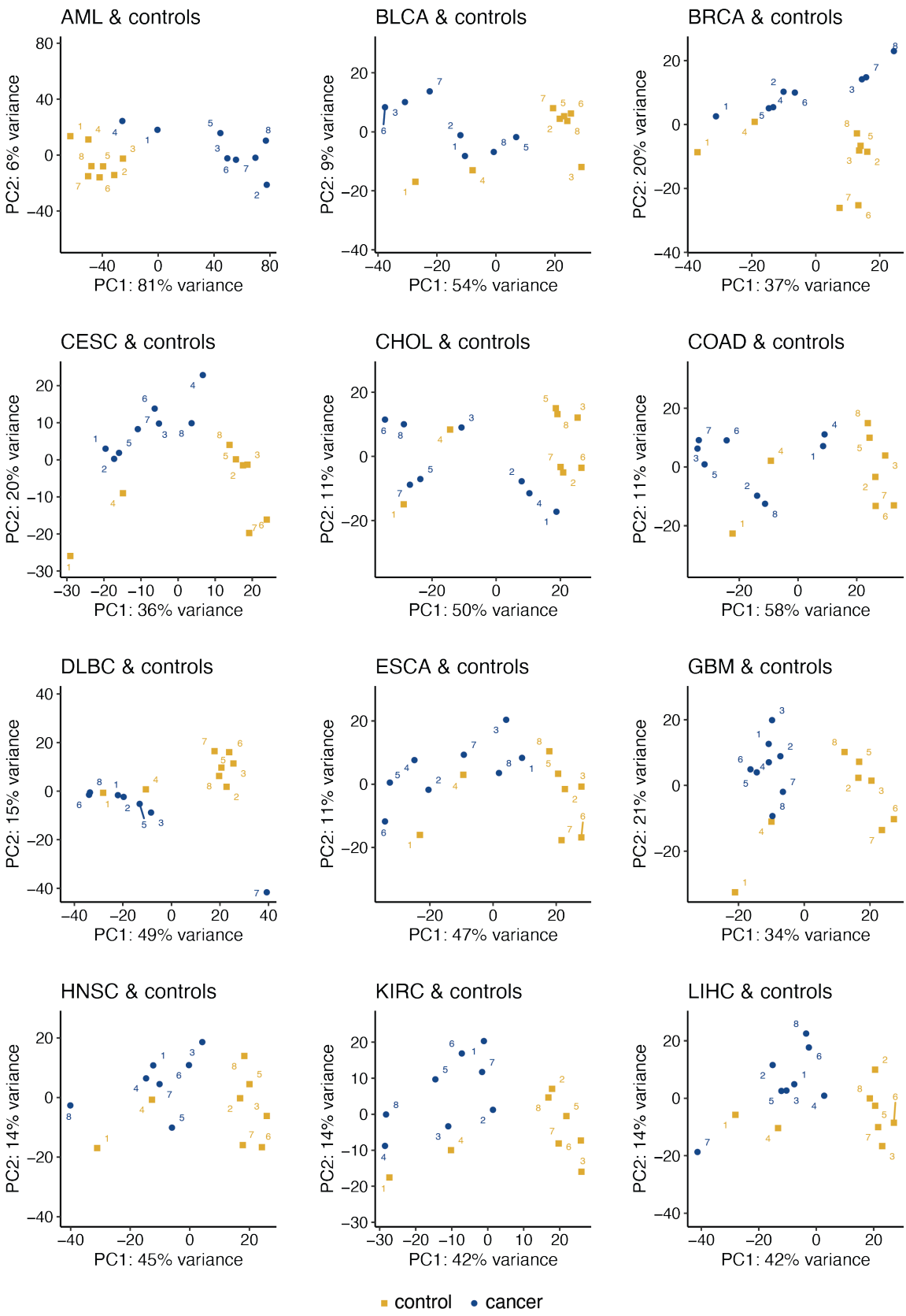
Supplemental Figure 1: High-quality paired end reads per sample. *Plot shows million paired end reads per sample after quality filtering. Groups ordered alphabetically. No significant difference between groups: Kruskal-Wallis test p-value = 0.0503, with moderate effect size (0.071), smallest post-hoc two-sided Wilcoxon rank-sum test q-value = 0.354.*

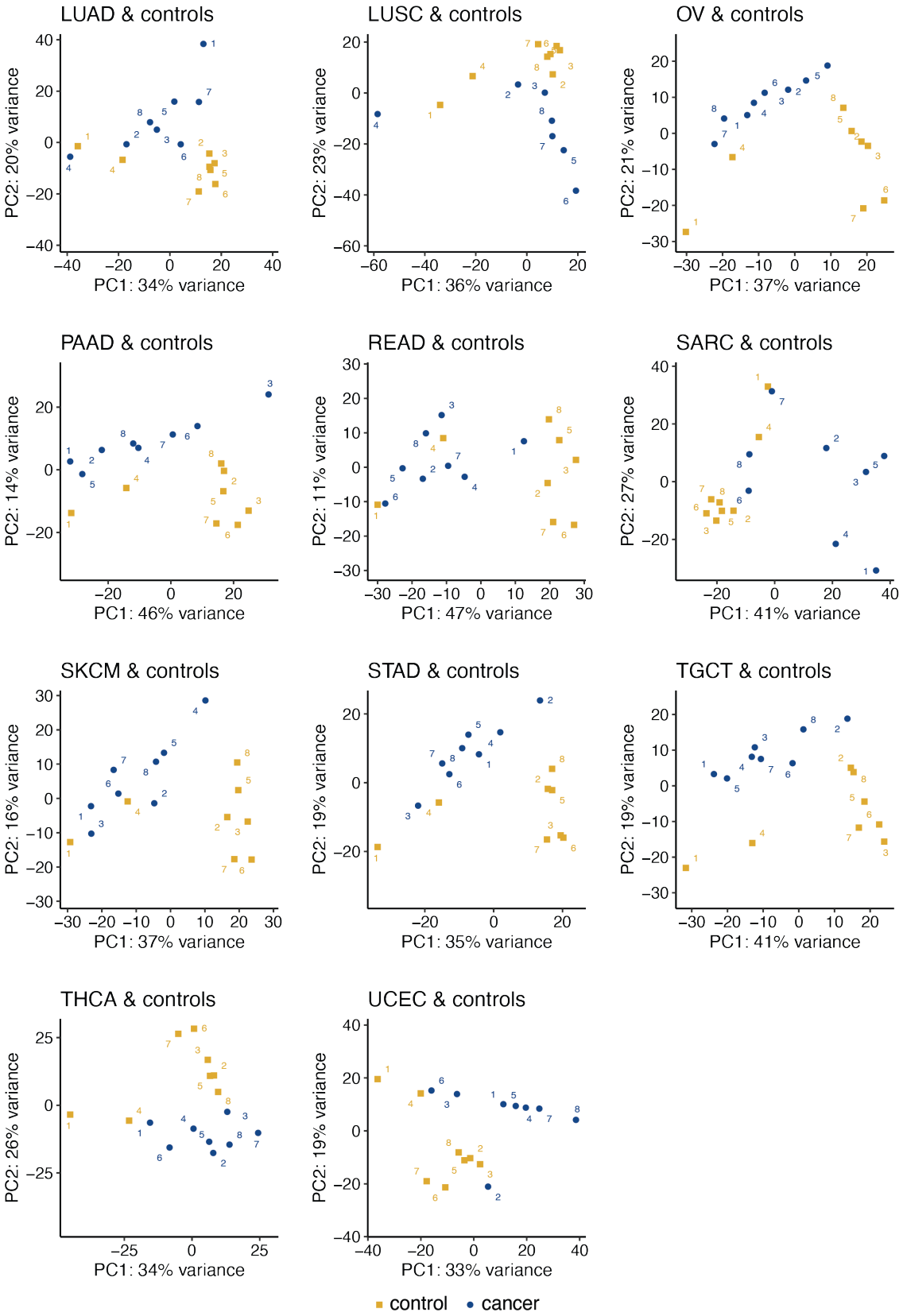


Supplemental Figure 2: Capture sequencing resulted in mRNA-enriched plasma cfRNA profiles. *RNA biotype distribution of mapped reads in pan-cancer cohort. Each bar represents one sample. Samples with less than 2M reads were excluded. Protein coding: protein coding gene transcripts (mRNA); lncRNA: long non-coding RNA; MT: mitochondrial RNA; MT RNR: mitochondrially encoded ribosomal RNA; rRNA: ribosomal RNA; spikes: Sequin spike-in RNA; other: other RNA transcripts.*

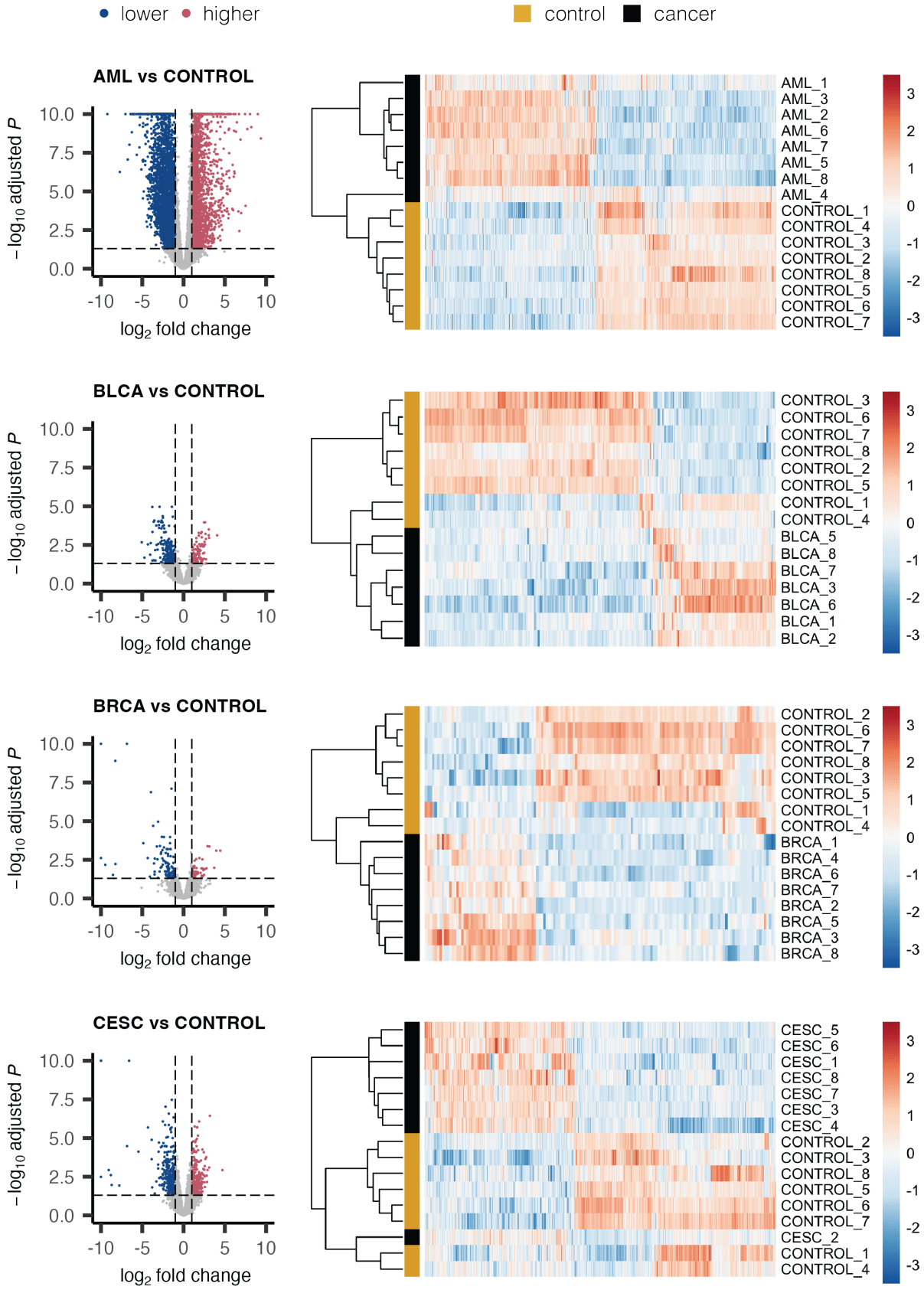


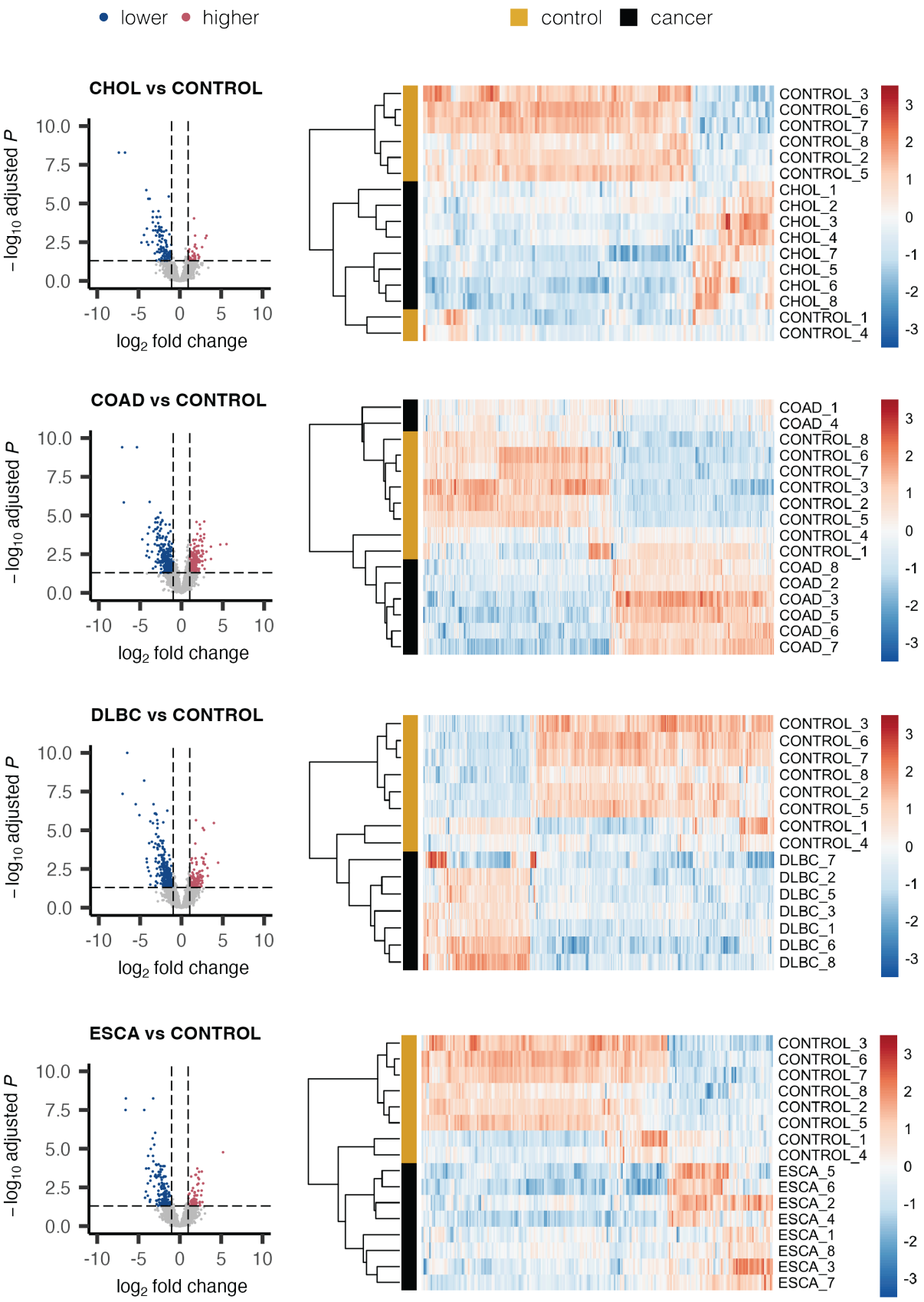
Supplemental Figure 3: Protein coding gene diversity in plasma is not only linked to sequencing depth*. Sequencing depth and number of protein coding genes are moderately correlated (spearman correlation r = 0.63, p < 2.2E-16). Red triangles: acute myeloid leukemia samples; black dots: non-leukemia cancer and control samples.*

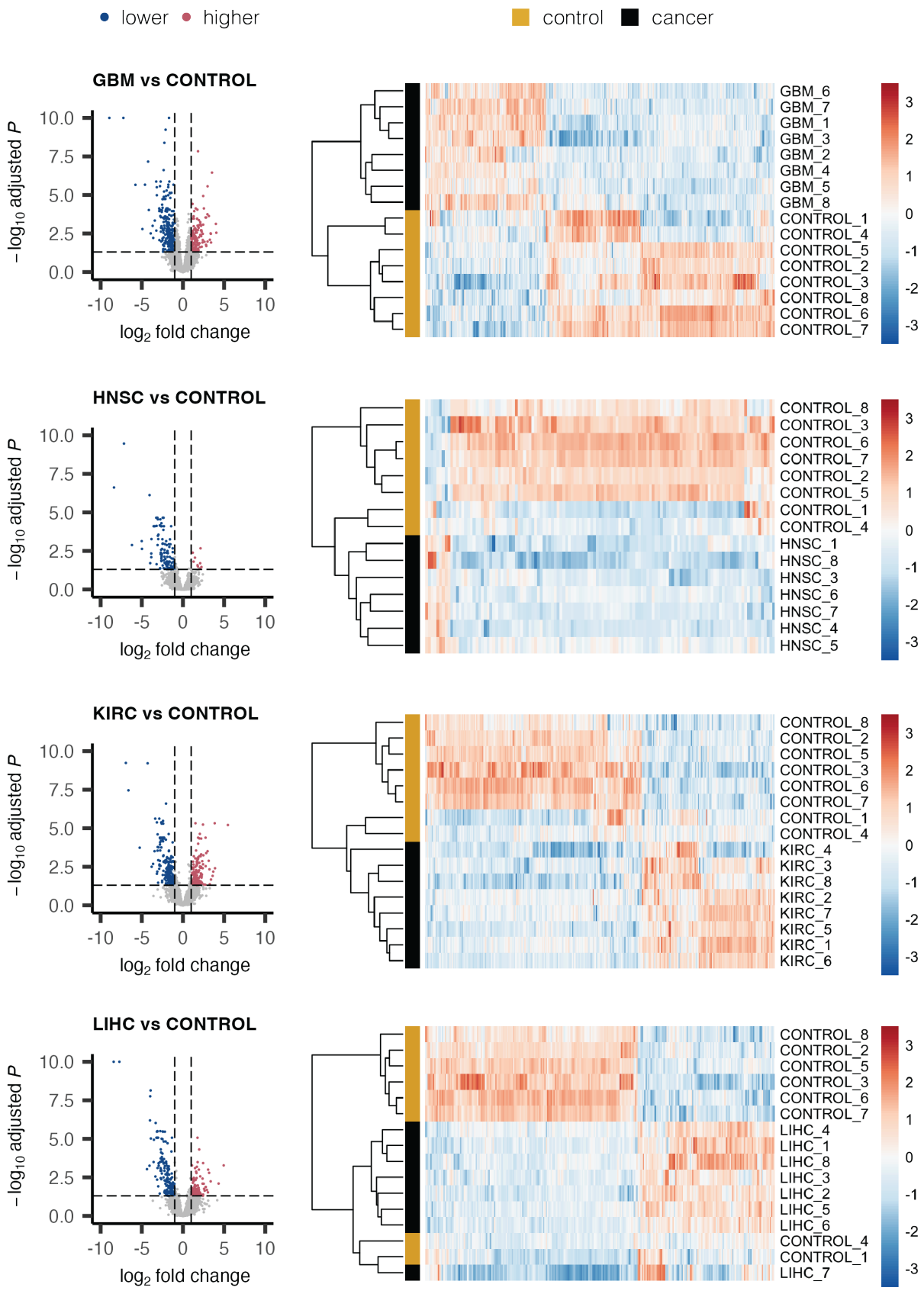

AML & controls
PC2: 6% variance
PC1: 81% variance
BLCA & controls
PC2: 9% variance
PC1: 54% variance
BRCA & controls
PC2: 20% variance
PC1: 37% variance
CESC & controls
PC2: 20% variance
PC1: 36% variance
CHOL & controls
PC2: 11% variance
PC1: 50% variance
COAD & controls
PC2: 11% variance
PC1: 58% variance
DLBC & controls
PC2: 15% variance
PC1: 49% variance
ESCA & controls
PC2: 11% variance
PC1: 47% variance
GBM & controls
PC2: 21% variance
PC1: 34% variance
HNSC & controls
PC2: 14% variance
PC1: 45% variance
KIRC & controls
PC2: 14% variance
PC1: 42% variance
LIHC & controls
PC2: 14% variance
PC1: 42% variance
control
cancer

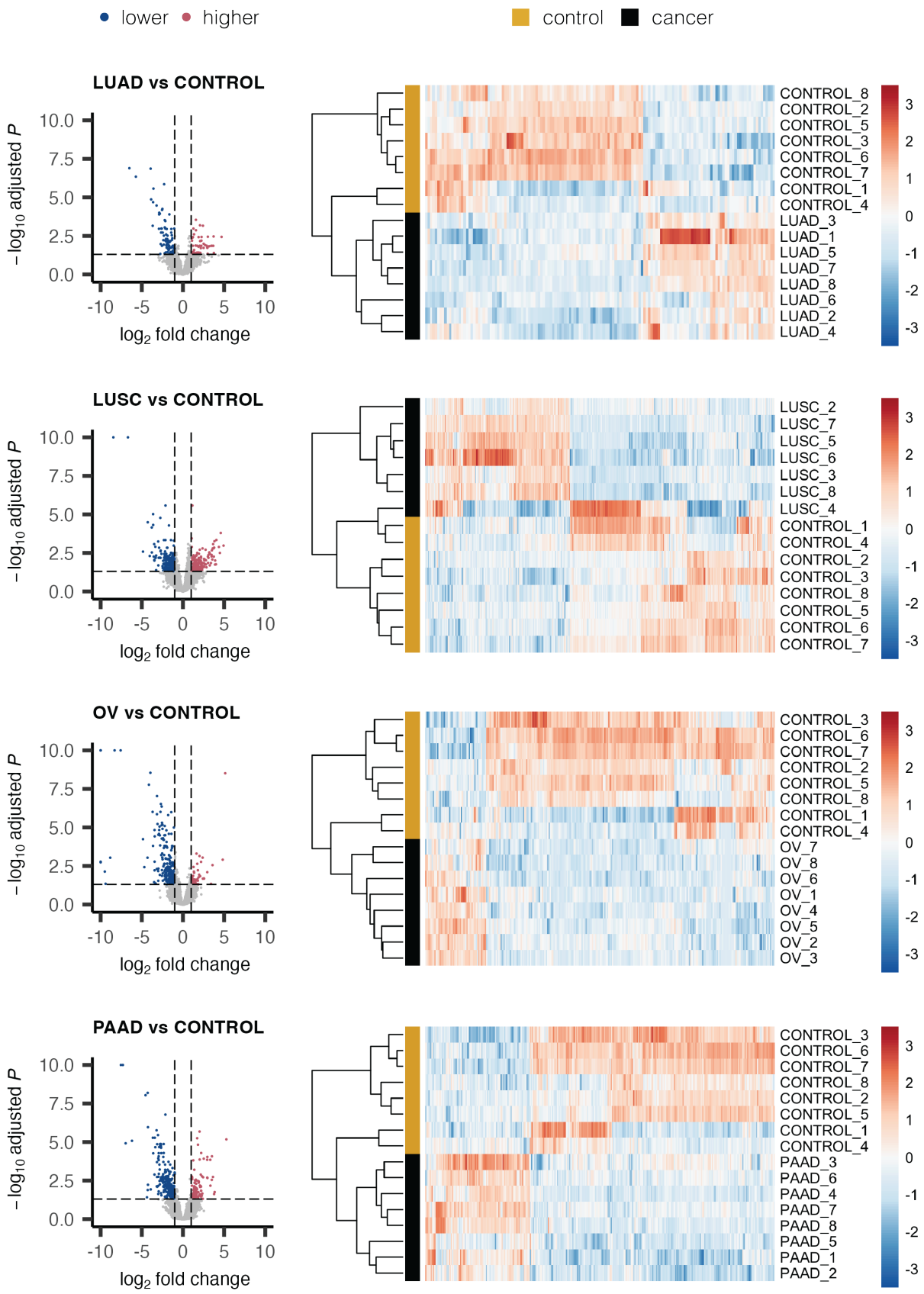


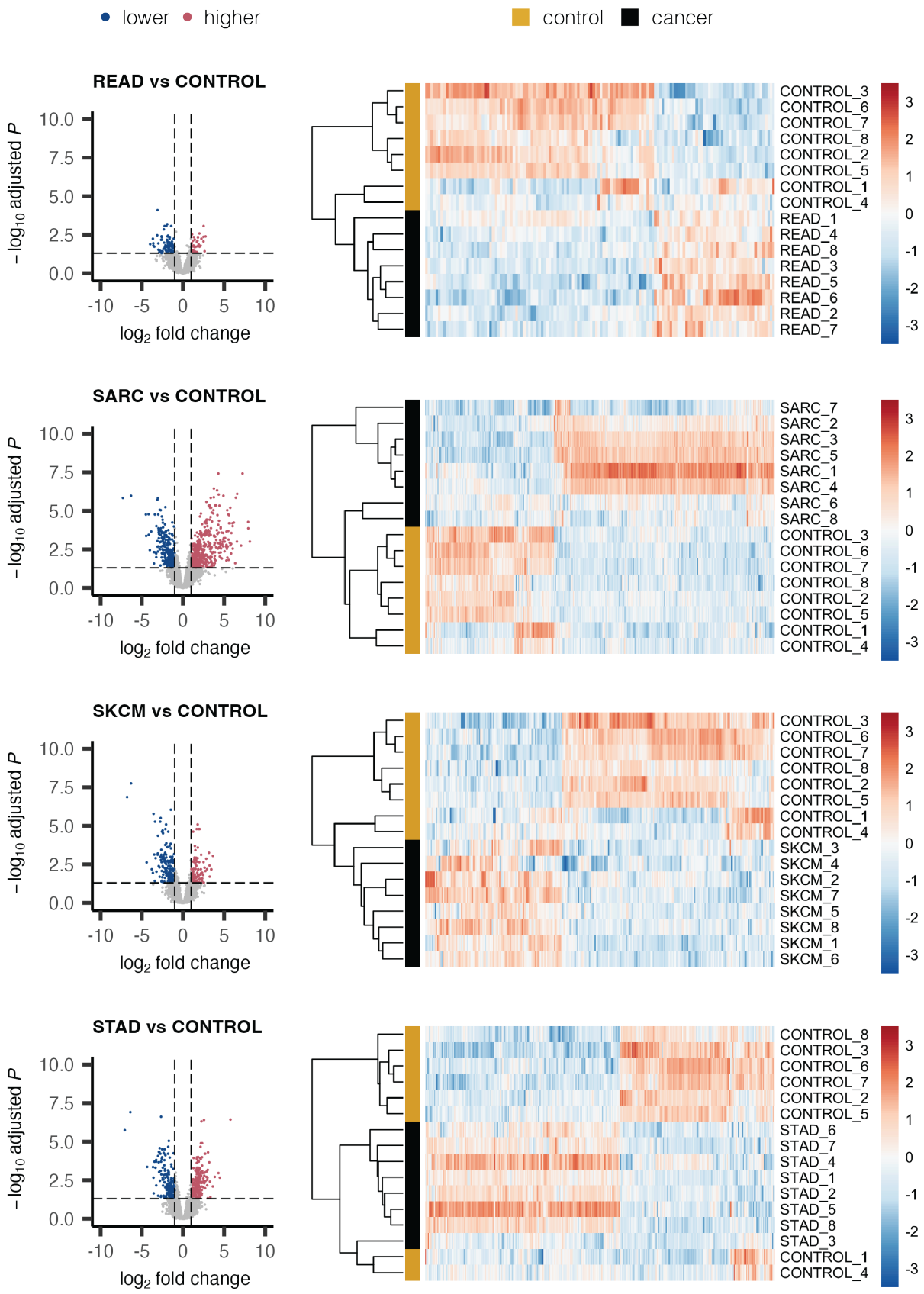
Supplemental Figure 4: Most cancer and control samples are separated in the first two principal components when considering individual types. *Principal component analysis using the 500 most variable mRNAs based on samples of control group and one cancer type (with variance stabilizing transformation DESeq2). Replicate numbers are indicated in the plot. Yellow squares: control samples; blue dots: cancer samples.*

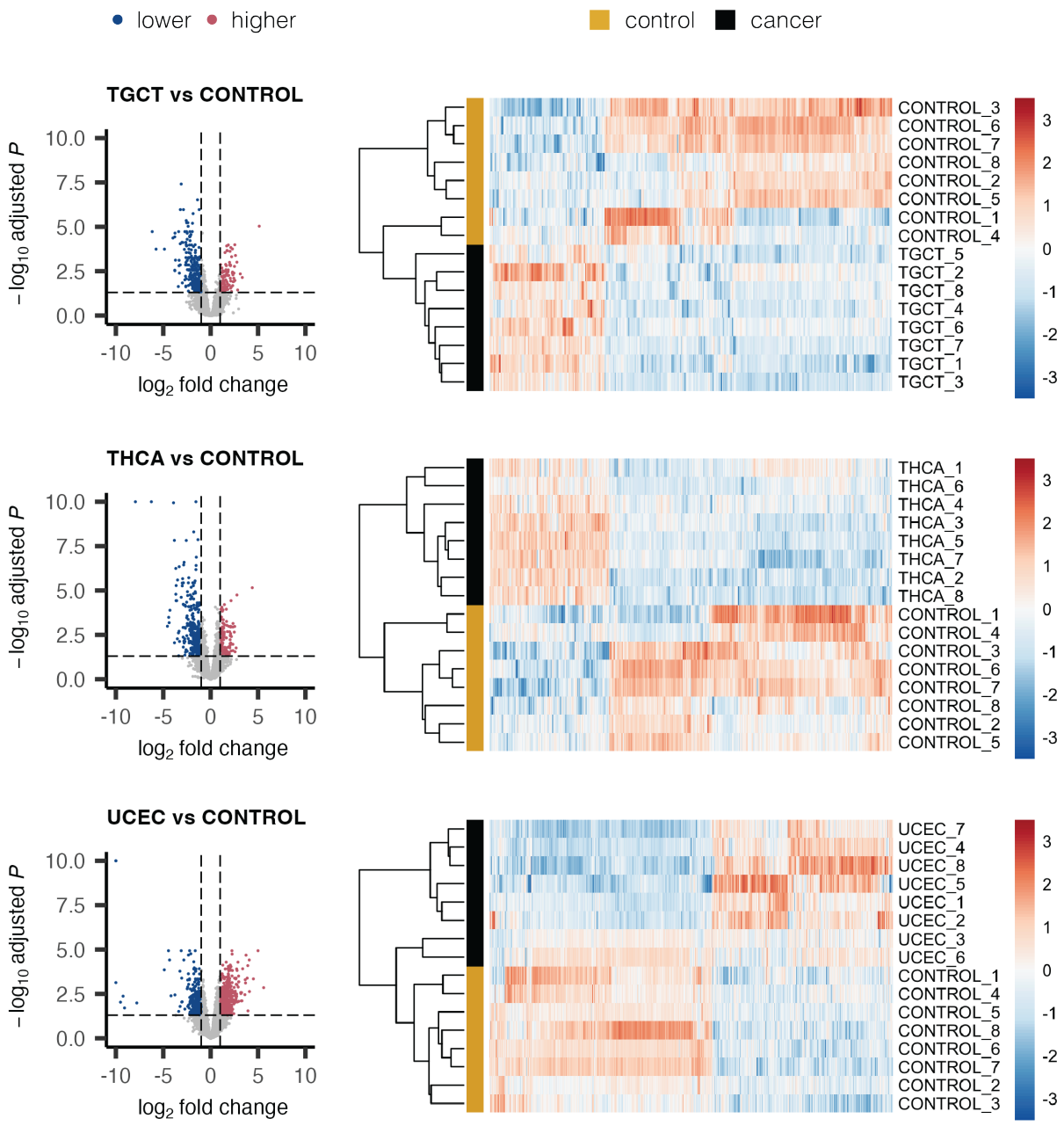

lower
higher
control
cancer
AML vs CONTROL
BLCA vs CONTROL
BRCA vs CONTROL
CESC vs CONTROL
−log10 adjusted P
log2 fold change
10.0
7.5
5.0
2.5
0.0
-10
-5
0
5
10
AML_1
AML_3
AML_2
AML_6
AML_7
AML_5
AML_8
AML_4
CONTROL_1
CONTROL_4
CONTROL_3
CONTROL_2
CONTROL_8
CONTROL_5
CONTROL_6
CONTROL_7
CONTROL_3
CONTROL_6
CONTROL_7
CONTROL_8
CONTROL_2
CONTROL_5
CONTROL_1
CONTROL_4
BLCA_5
BLCA_8
BLCA_7
BLCA_3
BLCA_6
BLCA_1
BLCA_2
CONTROL_2
CONTROL_6
CONTROL_7
CONTROL_8
CONTROL_3
CONTROL_5
CONTROL_1
CONTROL_4
BRCA_1
BRCA_4
BRCA_6
BRCA_7
BRCA_2
BRCA_5
BRCA_3
BRCA_8
CESC_5
CESC_6
CESC_1
CESC_8
CESC_7
CESC_3
CESC_4
CONTROL_2
CONTROL_3
CONTROL_8
CONTROL_5
CONTROL_6
CONTROL_7
CESC_2
CONTROL_1
CONTROL_4
3
2
1
0
-1
-2
-3


lower
higher
control
cancer
CHOL vs CONTROL
COAD vs CONTROL
DLBC vs CONTROL
ESCA vs CONTROL
−log10 adjusted P
log2 fold change
10.0
7.5
5.0
2.5
0.0
-10
-5
0
5
10
CONTROL_3
CONTROL_6
CONTROL_7
CONTROL_8
CONTROL_2
CONTROL_5
CONTROL_1
CONTROL_4
CHOL_1
CHOL_2
CHOL_3
CHOL_4
CHOL_7
CHOL_5
CHOL_6
CHOL_8
COAD_1
COAD_4
COAD_8
COAD_2
COAD_3
COAD_5
COAD_6
COAD_7
DLBC_7
DLBC_2
DLBC_5
DLBC_3
DLBC_1
DLBC_6
DLBC_8
ESCA_5
ESCA_6
ESCA_2
ESCA_4
ESCA_1
ESCA_8
ESCA_3
ESCA_7
3
2
1
0
-1
-2
-3

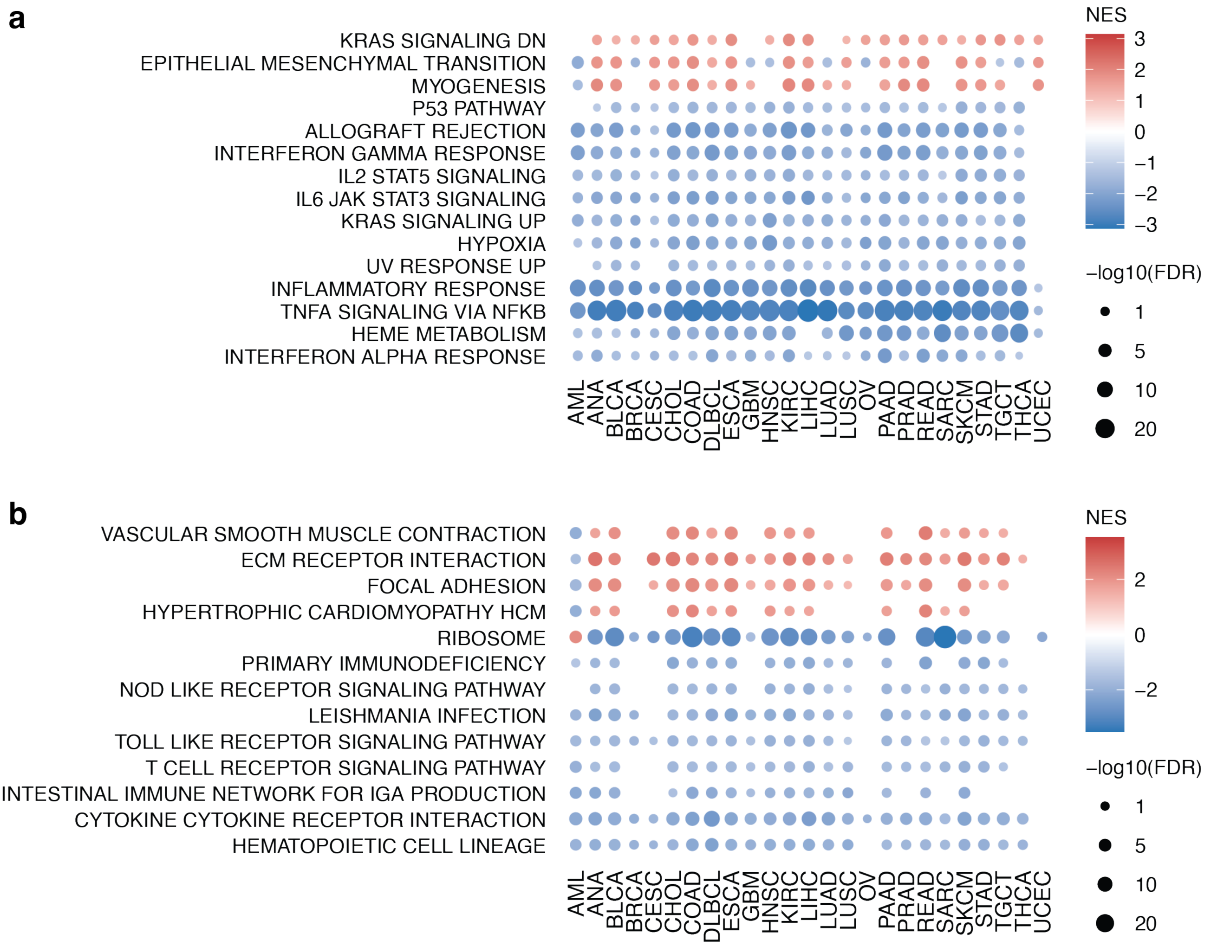

lower
higher
control
cancer
GBM vs CONTROL
HNSC vs CONTROL
KIRC vs CONTROL
LIHC vs CONTROL
−log10 adjusted P
log2 fold change
10.0
7.5
5.0
2.5
0.0
-10
-5
0
5
10
GBM_6
GBM_7
GBM_1
GBM_3
GBM_2
GBM_4
GBM_5
GBM_8
CONTROL_1
CONTROL_4
CONTROL_5
CONTROL_2
CONTROL_3
CONTROL_8
CONTROL_6
CONTROL_7
CONTROL_8
CONTROL_3
CONTROL_6
CONTROL_7
CONTROL_2
CONTROL_5
CONTROL_1
CONTROL_4
HNSC_1
HNSC_8
HNSC_3
HNSC_6
HNSC_7
HNSC_4
HNSC_5
CONTROL_8
CONTROL_2
CONTROL_5
CONTROL_3
CONTROL_6
CONTROL_7
CONTROL_1
CONTROL_4
KIRC_4
KIRC_3
KIRC_8
KIRC_2
KIRC_7
KIRC_5
KIRC_1
KIRC_6
CONTROL_8
CONTROL_2
CONTROL_5
CONTROL_3
CONTROL_6
CONTROL_7
LIHC_4
LIHC_1
LIHC_8
LIHC_3
LIHC_2
LIHC_5
LIHC_6
CONTROL_4
CONTROL_1
LIHC_7
3
2
1
0
-1
-2
-3


lower
higher
control
cancer
LUAD vs CONTROL
LUSC vs CONTROL
OV vs CONTROL
PAAD vs CONTROL
−log10 adjusted P
log2 fold change
10.0
7.5
5.0
2.5
0.0
-10
-5
0
5
10
CONTROL_8
CONTROL_2
CONTROL_5
CONTROL_3
CONTROL_6
CONTROL_7
CONTROL_1
CONTROL_4
LUAD_3
LUAD_1
LUAD_5
LUAD_7
LUAD_8
LUAD_6
LUAD_2
LUAD_4
LUSC_2
LUSC_7
LUSC_5
LUSC_6
LUSC_3
LUSC_8
LUSC_4
OV_7
OV_8
OV_6
OV_1
OV_4
OV_5
OV_2
OV_3
PAAD_3
PAAD_6
PAAD_4
PAAD_7
PAAD_8
PAAD_5
PAAD_1
PAAD_2
3
2
1
0
-1
-2
-3

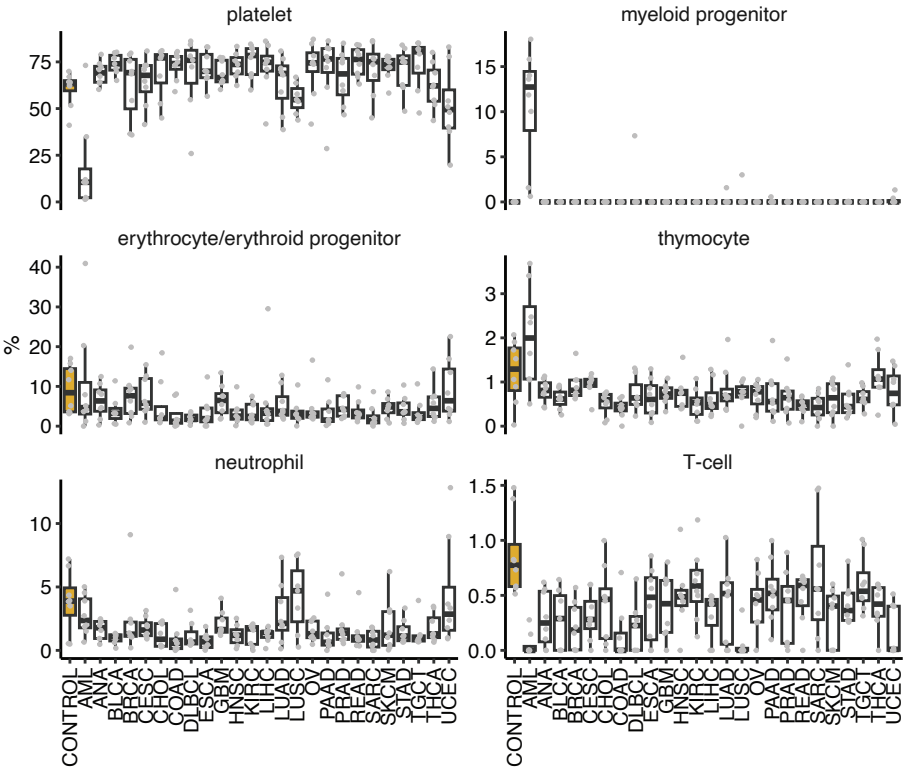

lower
higher
control
cancer
READ vs CONTROL
SARC vs CONTROL
SKCM vs CONTROL
STAD vs CONTROL
−log10 adjusted P
log2 fold change
10.0
7.5
5.0
2.5
0.0
-10
-5
0
5
10
CONTROL_3
CONTROL_6
CONTROL_7
CONTROL_8
CONTROL_2
CONTROL_5
CONTROL_1
CONTROL_4
READ_1
READ_4
READ_8
READ_3
READ_5
READ_6
READ_2
READ_7
SARC_7
SARC_2
SARC_3
SARC_5
SARC_1
SARC_4
SARC_6
SARC_8
SKCM_3
SKCM_4
SKCM_2
SKCM_7
SKCM_5
SKCM_8
SKCM_1
SKCM_6
STAD_6
STAD_7
STAD_4
STAD_1
STAD_2
STAD_5
STAD_8
STAD_3
3
2
1
0
-1
-2
-3



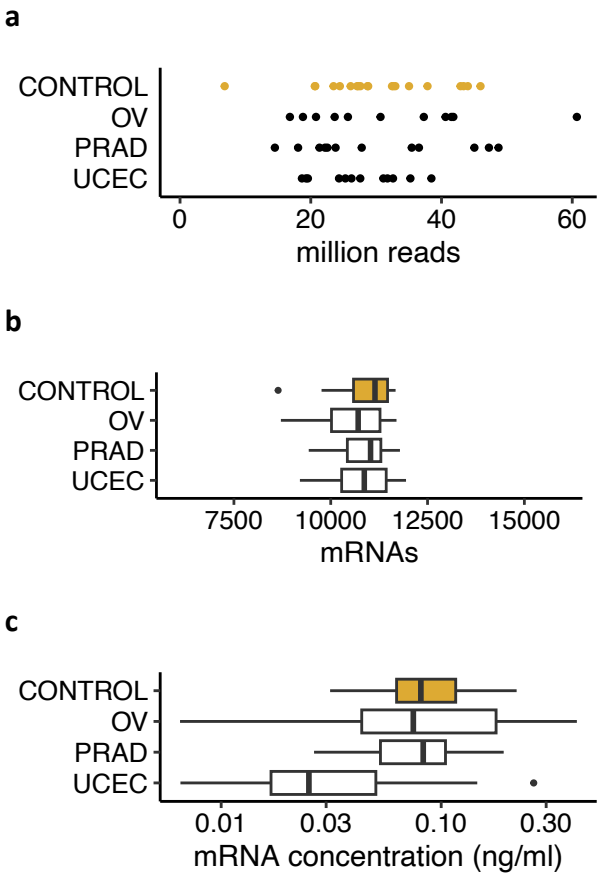
Supplemental Figure 5: Differentially abundant mRNAs in plasma of cancer patients compared to controls. *Volcano plots show fold change and multiple testing corrected p-values of mRNAs in cancer versus control. Horizontal line at adjusted p-value of 0.05 and vertical lines at log2 fold change of -1 and 1, resp. Higher (red): higher abundant in cancer compared to control samples ($q < 0.05$ and $\log_2$ fold change > 1); lower (blue): lower abundant in cancer compared to control samples ($q < 0.05$ and $\log_2$ fold change < -1). Heatmaps show relative abundance of differentially abundant genes (columns) in individual plasma samples (rows) based on log2(normalized counts +1) followed by z-score transformation per gene. Clustering of genes and samples based on Pearson correlation.*



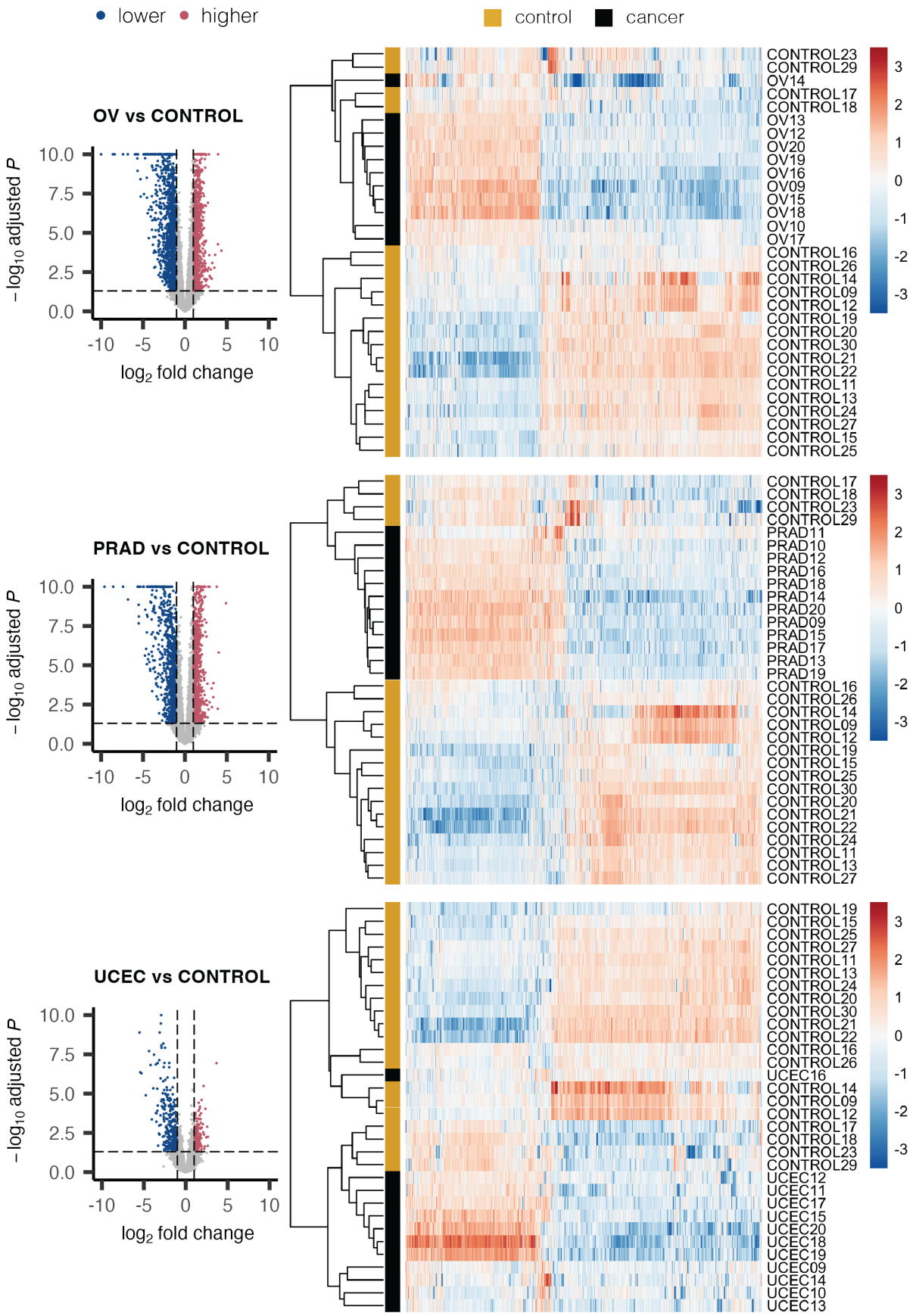
Supplemental Figure 6: Common patterns in gene set enrichment analysis for hallmark and KEGG gene sets. *Only pathways that are significantly enriched in at least half of the cancer types are shown. Colored according to normalized enrichment scores (NES) for gene sets with q<0.05. Gene sets obtained from Human Molecular Signatures Database: hallmark gene sets (**a**) and KEGG subcollection of canonical pathway gene (**b**).*



Supplemental Figure 7: Cell type contributions to plasma cfRNA profiles show distinct leukemic signals and lower immune cell fractions in cancer samples*. Boxplots of specific cell fractions based on deconvolution. 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. Yellow: control samples; grey dots: individual sample fractions. Cell type fractions obtained by nuSVR deconvolution using Tabula Sapiens v1.0 as basis matrix.*

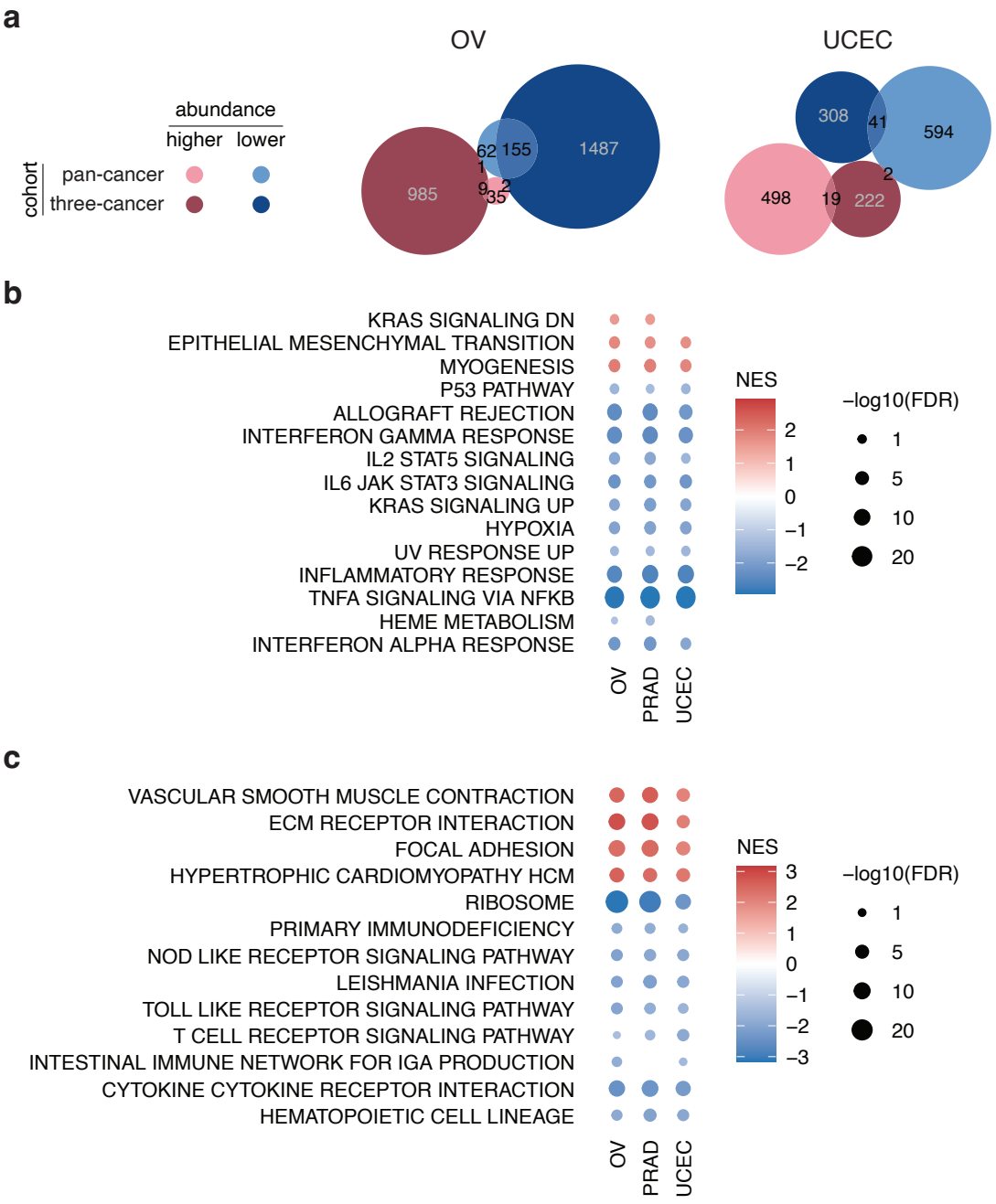


Supplemental Figure 8: Sequencing depth and number of unique mRNAs similar between types. a*, Plot shows million paired end reads after quality filtering with no significant difference between groups (Kruskal-Wallis p-value = 0.7858).* ***b****, Number of messenger RNAs (mRNAs) with at least 10 counts, no significant difference between groups (Kruskal-Wallis p = 0.8116).* ***c****, mRNA concentration in plasma based on Sequin spikes with no significant difference between groups (Kruskal-Wallis p = 0.05902).* ***b*** *and* ***c****, Boxplots show lower quartile (Q1), median, and upper quartile (Q3). Whiskers extend from the lower and upper quartile to the smallest and largest value, respectively, within at most 1.5*interquartile range (Q3-Q1) from that quartile. More extreme points are plotted as individual dots. Groups ordered alphabetically and controls indicated in yellow.*

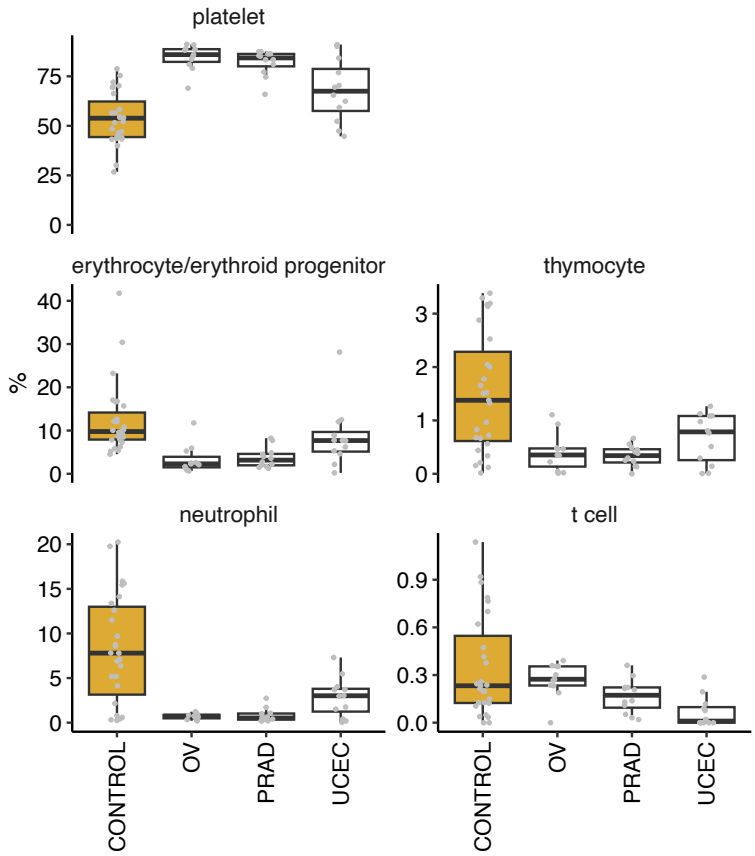


Supplemental Figure 9: Differentially abundant mRNAs in plasma of cancer patients compared to controls of the validation cohort. *Volcano plots show fold change and multiple testing corrected p-values of mRNAs in*

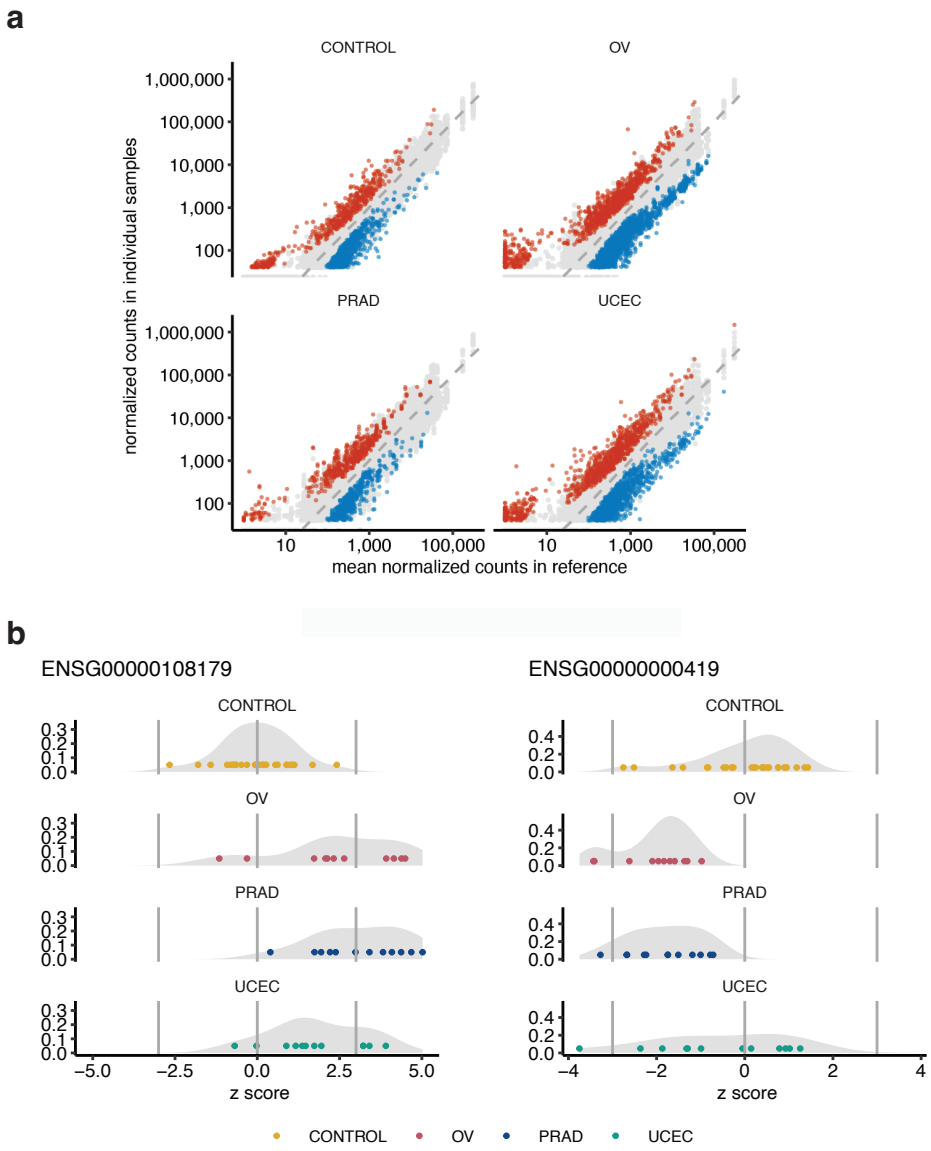
cancer versus control. Horizontal line at adjusted p-value of 0.05 and vertical lines at fold change 0.5 and 2. More (red): more abundant in cancer compared to control samples (adjusted p < 0.05 and fold change > 2); less (blue): less abundant in cancer compared to control samples (adjusted p < 0.05 and fold change < 0.5). Heatmaps show relative abundance of differentially abundant genes (columns) in individual plasma samples (rows) based on log2(normalized counts +1) followed by z-score transformation per gene. Clustering of genes and samples based on Pearson correlation.



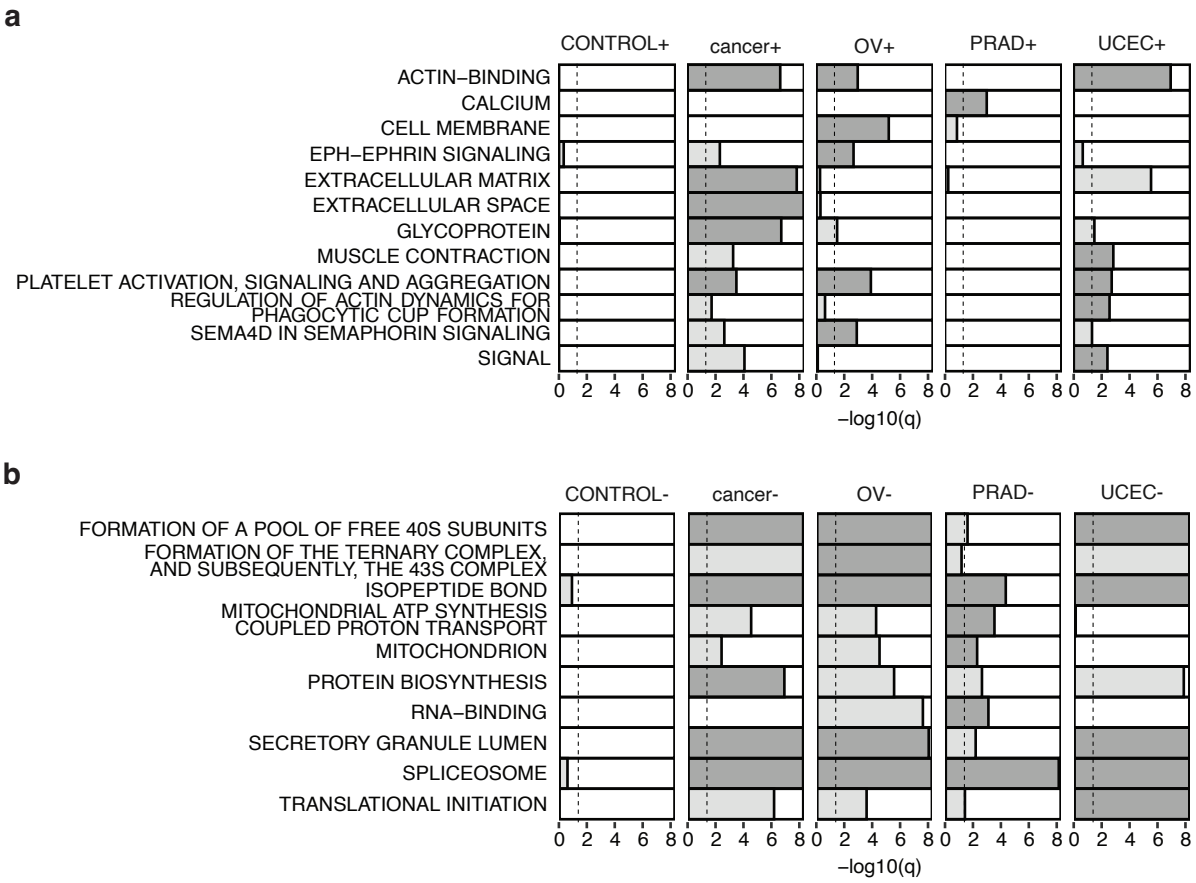
Supplemental Figure 10: Significant, but small, overlap in differentially abundant genes between cohorts and recurrence of enrichment patterns. a, *Overlap between differentially abundant genes in cancer versus control between pan-cancer and validation cohort of ovarian (OV) and uterine cancer (UCEC), respectively. Fisher's exact test OV: p=4.0E-4, odds ratio=4.6, Jaccard index=0.9% for higher abundant genes (higher); p=8.4E-118, odds ratio=30.0, Jaccard index=9.1% for lower abundant genes (lower). Fisher's exact test UCEC: p=2.2E-5, odds ratio=3.3, Jaccard index=2.6% for higher abundant genes (higher); p=7.6E-13, odds ratio=4.2, Jaccard index=4.3% for lower abundant genes (lower). Circles are proportional to the number of genes within one cancer vs control comparison. Differential abundance: q<0.05 & $|\log_2$ fold change$|$>1.* ***b,*** *Gene set enrichment in validation cohort based on hallmark and KEGG gene sets, respectively, obtained from Human Molecular Signatures Database. Only gene sets significantly enriched in at least half of the cancer types of the pan-cancer cohort (cf. Supplemental Figure 6) are shown. Coloring according to normalized enrichment scores (NES) for gene sets with q<0.05.*



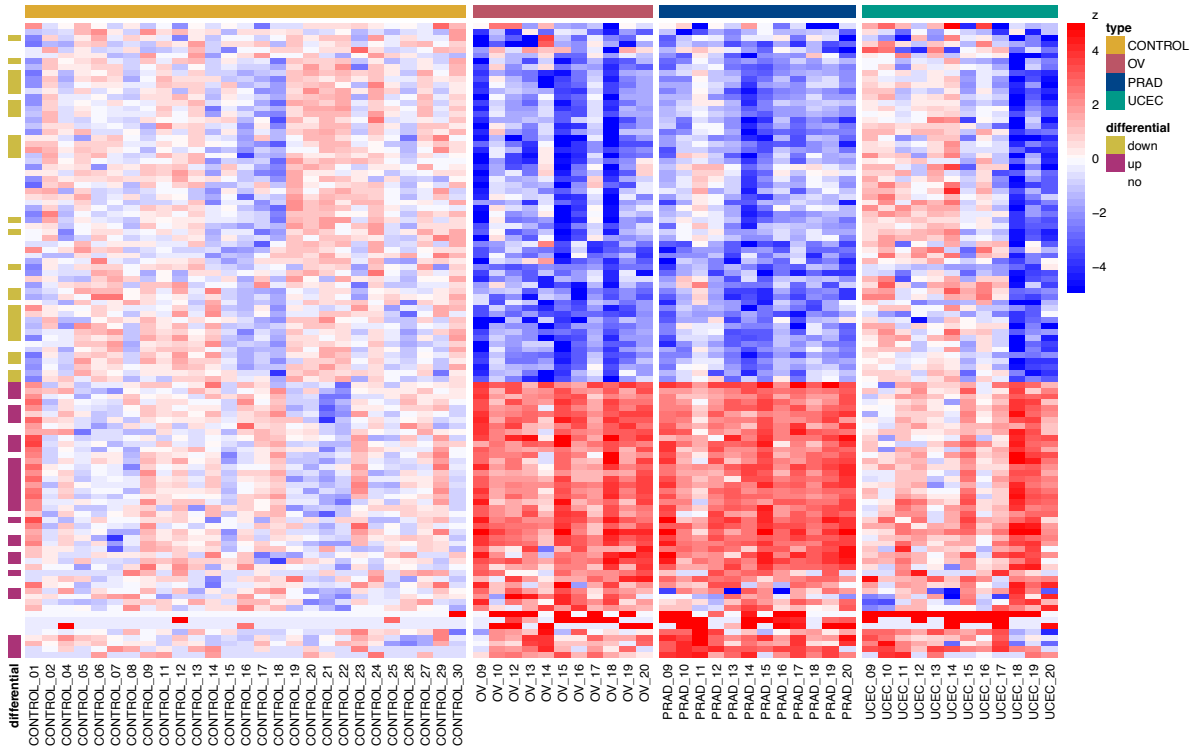
Supplemental Figure 11: Cell type contributions to plasma cfRNA show increased platelet and lower immune cell fractions in cancer samples compared to controls. *Boxplots of specific cell fractions obtained from nuSVR deconvolution using Tabula Sapiens v1.0 as basis matrix. Boxplots show lower quartile (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. Yellow: control samples; grey dots: individual sample fractions.*



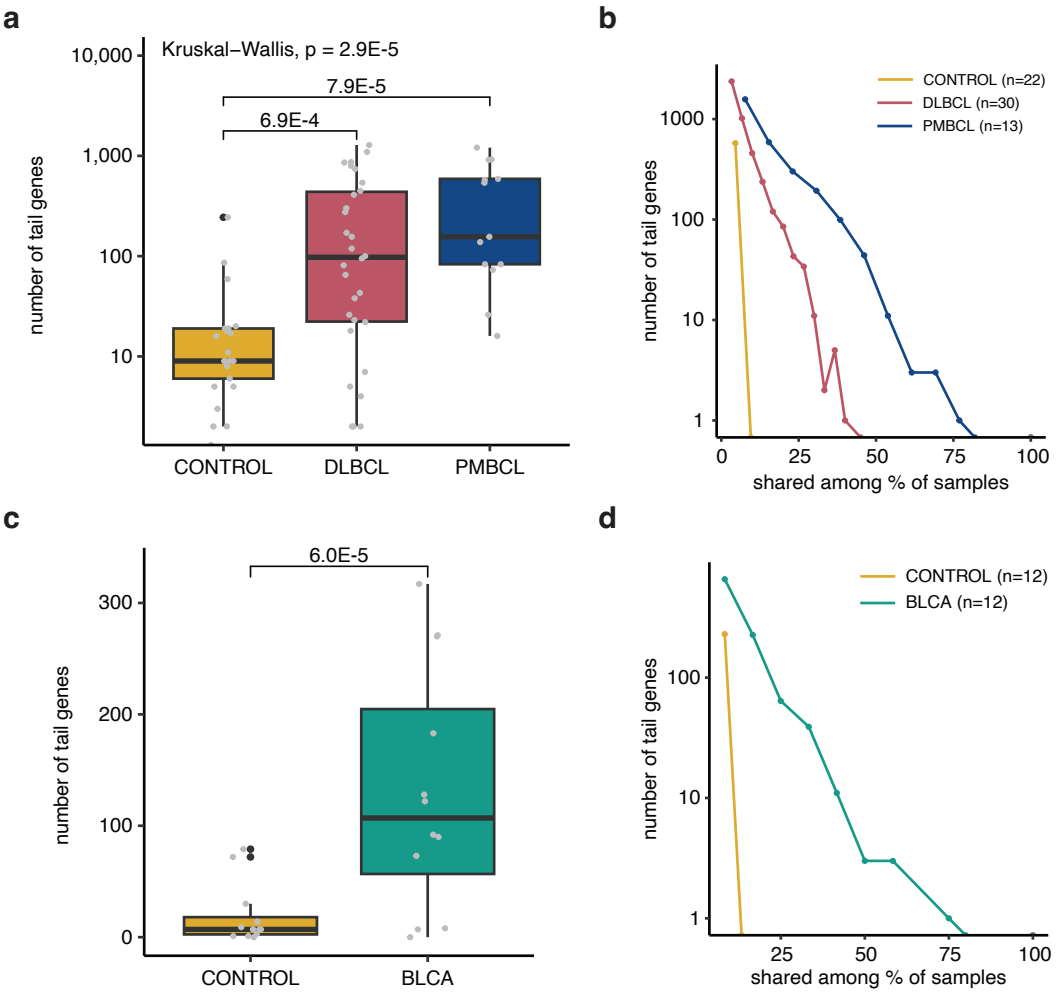
Supplemental Figure 12: Tail gene counts and z score distribution per type. a, *normalized tail gene counts in individual samples of a certain type compared to the mean of the control reference. Red (colored dots above diagonal): z>3, blue (colored dots below diagonal): z<-3.* ***b,*** *z scores per type for two unique tail genes where individual dots represent the exact z scores for that gene in a particular sample (compared to control reference) and a grey density plot. The genes are identified as tail gene in samples where the respective z score is smaller than -3 (left vertical line) or above 3 (right vertical line).*



Supplemental Figure 13: Recurrence of top overrepresented patterns in cancer sample tail genes and absence in control sample tail genes. a *and* ***b,*** *top-5 most enriched clusters (q<0.05) for positive (z>3) and negative (z<-3) tail genes, respectively. Dark grey: -log10(q-value) of the most significant annotation (Benjamini correction) for each of the 5 most enriched clusters per group of tail genes (CONTROL, OV, PRAD, UCEC, or 'cancer', i.e. all three cancer types combined) based on DAVID functional clustering analysis. Light grey: -log10(q-value) of top-5 annotations of other groups that are not in top-5 of the OV, PRAD, UCEC, all cancer, or CONTROL group. -log10(q) topped of at 8. Dashed vertical lines at q=0.05.*



Supplemental Figure 14: Z score distribution of selected tail genes in the three-cancer validation cohort. *Heatmap shows z scores of the 108 consensus tail genes, genes significant in all leave-one-sample-out Fisher's exact tests. Genes (rows) clustered based on Euclidean distance.*



Supplemental Figure 15: Tail genes in the lymphoma plasma and bladder cancer urine validation cohort. ***a*** *and* ***b****, lymphoma plasma cohort: plasma samples from 22 control (CONTROL), 30 diffuse large B-cell lymphoma (DLBCL), and 13 primary mediastinal large B-cell lymphoma (PMBCL) patients.* ***a****, Boxplot shows number of tail genes (n=5374) per group (lower quartile, median, upper quartile, whiskers of 1.5*interquartile range, more extreme points indicated by black dots), grey dots represent individual sample counts. Wilcoxon rank-sum q-values for specific cancer versus control comparisons indicated in the plot.* ***b****, Number of tail genes shared by a certain fraction of samples of the lymphoma or control group. Number of samples indicated between brackets.* ***c*** *and* ***d****, urine cohort: urine samples from 12 control (CONTROL) and 12 bladder cancer (BLCA) patients.* ***c****, Boxplot shows number of Tail genes (n=1152) per group (lower quartile, median, upper quartile, whiskers of 1.5*interquartile range, more extreme points indicated by black dots), grey dots represent individual sample counts.* ***d****, Number of tail genes shared by a certain fraction of samples of the bladder cancer or control group. Number of samples indicated between brackets.*

Supplemental tables

Supplemental Table 1: Overview of donor, disease, and sample information*. Cohort: sample cohort (Table1) to which a certain plasma sample belongs; DonorID: donor code; RNAID: sample code; Cancer: cancer type; Abbreviation: cancer type abbreviation; ReplicateNr: replicate number; CollectionTube: blood collection tube; Biofluid: biofluid type of sample; Age, Sex, Ethnicity: age, sex (F for female, M for male), and ethnicity of the sample donor; TumorLocation; HistologicalDiagnosis; Grade; TNM; Stage; Treatment: ongoing treatment; Excluded: indicates samples excluded based on insufficient sequencing reads (<2M), see Methods.*

Supplemental Table 2: Fusion genes of TCGA found in individual plasma samples. *Fusion: fusion gene; TCGA: cancer type(s) for which fusion gene was detected in TCGA; PlasmaID: plasma sample id (cancer type followed by replicate number) for which the fusion transcript is detected by FusionCatcher; SpanningPairs: Count of pairs of reads supporting the fusion; SpanningUniqueReads: Count of unique reads mapping on the fusion junction.*

Fusion	TCGA	PlasmaID	SpanningPairs	SpanningUniqueReads
PML::RARA	AML	AML_5	19	6
RARA::PML	AML	AML_5	19	17
PML::RARA	AML	AML_8	7	9
SOS1::MAP4K3	TGCT, THCA	CESC_2	1	3
SOS1::MAP4K3	TGCT, THCA	CESC_6	2	5
PPA2::TBCK	LUAD	STAD_2	1	4
SPAG9::MBTD1	OV, STAD	LIHC_2	1	2
LTBP1::BIRC6	BRCA, AML, LUSC, PRAD	LIHC_7	1	2
TAB3::DMD	SARC	KIRC_1	1	3
MITF::FOXP1	SKCM	KIRC_4	1	2
TNRC6A::PRKCB	GBM	LUAD_7	1	3
CLEC16A::TXNDC11	GBM	PAAD_2	1	3
LTBP1::BIRC6	BRCA, AML, LUSC, PRAD	READ_8	1	3
ASAP2::MBOAT2	LUAD	SARC_4	2	3
SNX6::BAZ1A	STAD	SARC_6	1	2
DENND3::PTK2	GBM	TGCT_7	1	2
CTBP2::MGMT	LGG	PRAD_18	1	3

Supplemental Table 3: Overrepresentation of Reactome, KEGG and Gene Ontology gene sets in recurrent lower abundant mRNAs. *DAVID functional annotation chart for the (26) mRNAs with q < 0.05 and log2 fold change < -1 in cancer versus control for at least 23 out of 25 cancer types of the pan-cancer cohort. Category: original database/resource where the term orients, Term: enriched term associated with input gene list, Count: number of input genes involved in the term, Percentage: involved genes divided by total input genes, P-value: EASE Score – modified Fisher Exact p-value, Genes: involved gene names, List total: number of input genes that are annotated to at least one term in the category. Pop hits: number of background genes annotated to the specific term. Pop total: number of background genes annotated to at least one term in the category, Fold Enrichment, Bonferroni, Benjamini, FDR (false discovery rate).*

Supplemental Table 4: Cell type deconvolution results. *Fractions obtained by nuSVR deconvolution using Tabula Sapiens v1.0 as basis matrix. Pan-cancer: deconvolution results for the plasma samples of the pan-cancer cohort. Three-cancer: deconvolution results for the independent plasma samples of the validation cohort.*

Supplemental Table 5: DAVID functional annotation clustering based on tail gene lists. *Anycancer+, OV+, PRAD+, UCEC+, CONTROL+: based on tail genes with z>3 in any cancer, OV, PRAD, UCEC, CONTROL sample of the three-cancer cohort, respectively. Anycancer-, OV-, PRAD-, UCEC-, CONTROL-: based on tail genes with z<-3 in any cancer, OV, PRAD, UCEC, CONTROL sample of the three-cancer cohort, respectively. Category: original database/resource where the term orients, Term: enriched term associated with input gene list, Count: number of input genes involved in the term, Percentage: involved genes divided by total input genes, P-value: EASE Score – modified Fisher Exact p-value, Genes: involved gene names, List total: number of input genes that are annotated to at least one term in the category. Pop hits: number of background genes annotated to the specific term. Pop total: number of background genes annotated to at least one term in the category, Fold Enrichment, Bonferroni, Benjamini, FDR (false discovery rate).*