Unbiased discovery of cancer pathways and therapeutics using Pathway Ensemble Tool and Benchmark

Supplementary Figures

S1-S7

Figure S1. Benchmark for evaluating and optimizing existing pathway analysis methods. a Cumulative distribution function (CDF) plots comparing indicated methods on shRNA-seq genesets. Median rank for correct pathways, precision@10(P@10), and average precision@(10/AP@10) are shown. Target genesets were defined as the top 200 genes by indicated method (e.g. edgeR, DESeq2) in K562 cells. **b** Histograms showing the overlap of genes (out of 200) between matched IGSs and TGSs from shRNA-seq for each indicated method or random control genesets. The median of overlap sizes is shown in parentheses. **c** Same as **a**; input genesets were extracted as differently expressed genes based on indicated method in HepG2 cells as pre-ranked list or top 200 DEGs when the size is needed e.g. for Enrichr. **d** Effect of geneset size on performance of indicated methods. TGS and IGS were defined according to the best performing settings (**c**-bottom right panel); sizes for both IGS and TGS were set to the same values indicated. For GSEA only the TGS size is needed and the IGS was provided as a pre-ranked list. **e** CDF plots comparing indicated methods on eCLIP-seq (top) or ChIP-seq (bottom) genesets. Target genesets were the top 200 genes by signal p-value, signal value, or q-value in one cell type. Input genesets were differently bound genes by the same metrics in another cell type, pre-ranked or as the top 200 for tools like Enrichr. **f** Same as **b** except IGSs and TGSs are from eCLIP-seq (top panel) and ChIP-seq (bottom panel). **g** False discovery rate (FDR) of indicated methods as a function of the number of biological replicates in each experiment across indicated normal and tumor samples. The expected FDR of 0.05 is highlighted by dashed lines. Error bars denote median ± 95% confidence interval from n=100 randomizations. **h** FDR of indicated methods across prostate cancer samples (n=10 samples per group and 100 randomization). **i** Heatmap showing pathways with FDR<0.05 from the indicated methods and tissues. Example pathways are highlighted. The heatmap color is proportional to the median FDR values from 100 randomizations. ora! is implementation from egsea. *p<0.05; **p < 0.01; **** p< 0.0001 by two-sided Wilcoxon rank sum test. Source data are provided as a Source Data file.

FigureS1

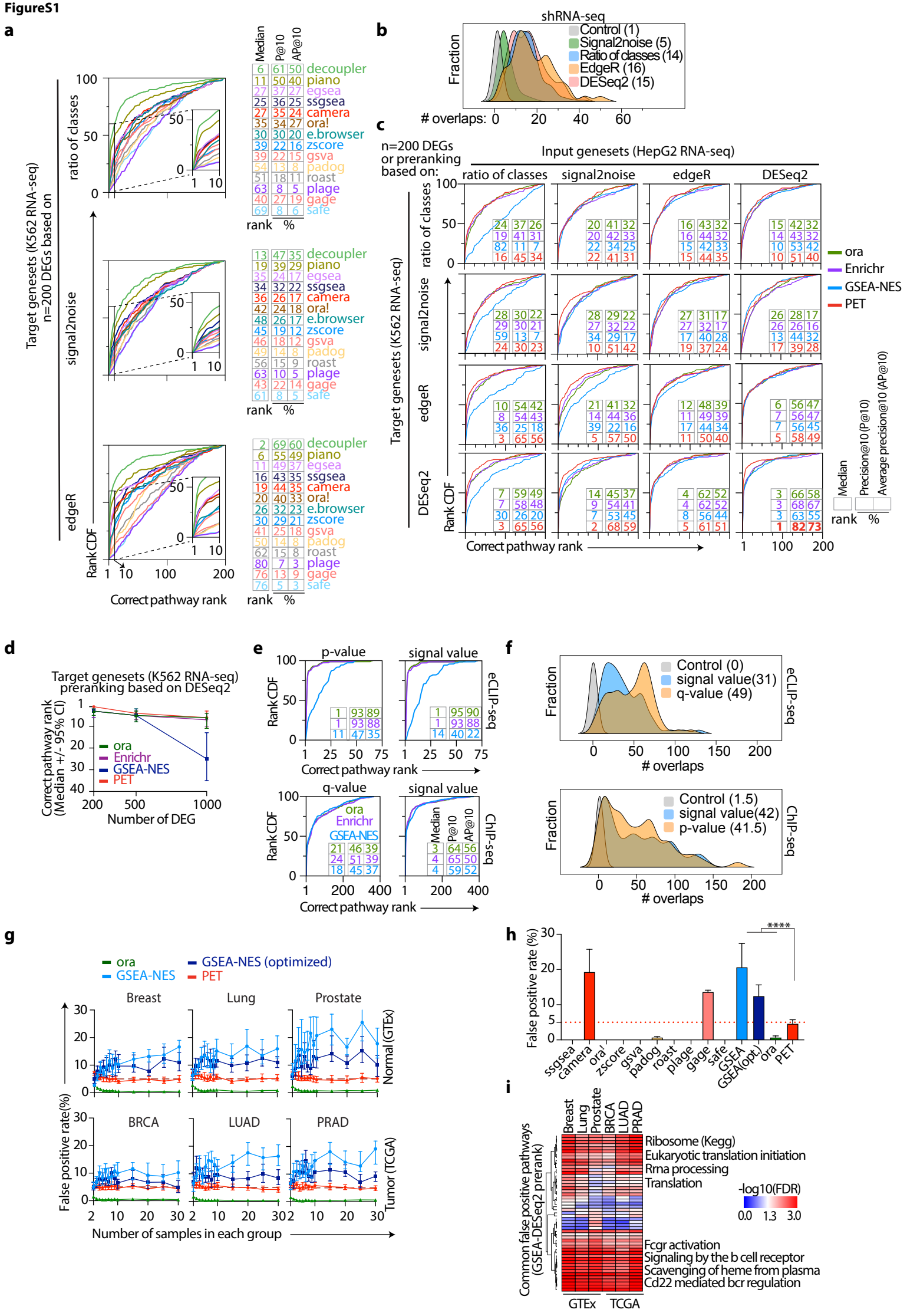

a
Target genesets (K562 RNA-seq)
n=200 DEGs based on
ratio of classes
signal2noise
edgeR
RankCDF
Correct pathway rank
Median P@10 AP@10
decoupler
piano
egsea
ssgsea
camera
ora!
e.browser
zscore
gsva
padog
roast
plage
gage
safe
rank %
b
shRNA-seq
Fraction
Control (1)
Signal2noise (5)
Ratio of classes (14)
EdgeR (16)
DESeq2 (15)
overlaps:
c
n=200 DEGs
or preranking
based on:
Input genesets (HepG2 RNA-seq)
ratio of classes
signal2noise
edgeR
DESeq2
Target genesets (K562 RNA-seq)
RankCDF
Correct pathway rank
ora
Enrichr
GSEA-NES
PET
Median
Precision@10 (P@10)
Average precision@10 (AP@10)
rank %
d
Target genesets (K562 RNA-seq)
preranking based on DESeq2
Correct pathway rank
(Median +/- 95% CI)
Number of DEG
ora
Enrichr
GSEA-NES
PET
e
p-value
signal value
q-value
RankCDF
Correct pathway rank
eCLIP-seq
ChIP-seq
ora
Enrichr
GSEA-NES
Median
P@10
AP@10
f
Fraction
overlaps
Control (0)
signal value(31)
q-value (49)
Control (1.5)
signal value(42)
p-value (41.5)
eCLIP-seq
ChIP-seq
g
ora
GSEA-NES
GSEA-NES (optimized)
PET
Breast
Lung
Prostate
BRCA
LUAD
PRAD
Normal (GTEx)
Tumor (TCGA)
False positive rate(%)
Number of samples in each group
h
False positive rate (%)
****
ssgsea
camera
ora!
zscore
gsva
padog
roast
plage
gage
safe
GSEA
GSEA(opt.)
ora
PET
i
Breast
Lung
Prostate
BRCA
LUAD
PRAD
Common false positive pathways
(GSEA-DESeq2 prerank)
Ribosome (Kegg)
Eukaryotic translation initiation
Rrna processing
Translation
Fcgr activation
Signaling by the b cell receptor
Scavenging of heme from plasma
Cd22 mediated bcr regulation
-log10(FDR)
GTEx
TCGA

Figure S2. PET identifies relevant signaling pathways distinguishing genetic subtypes of cancers. Bar plots show the top 10 pathways reported by PET and their corresponding FDR values when comparing the indicated samples. Bold font highlights the pathways among the top three reported that are supported by the literature. Gene expression data were sourced from CCLE. Source data are provided as a Source Data file.

FigureS2

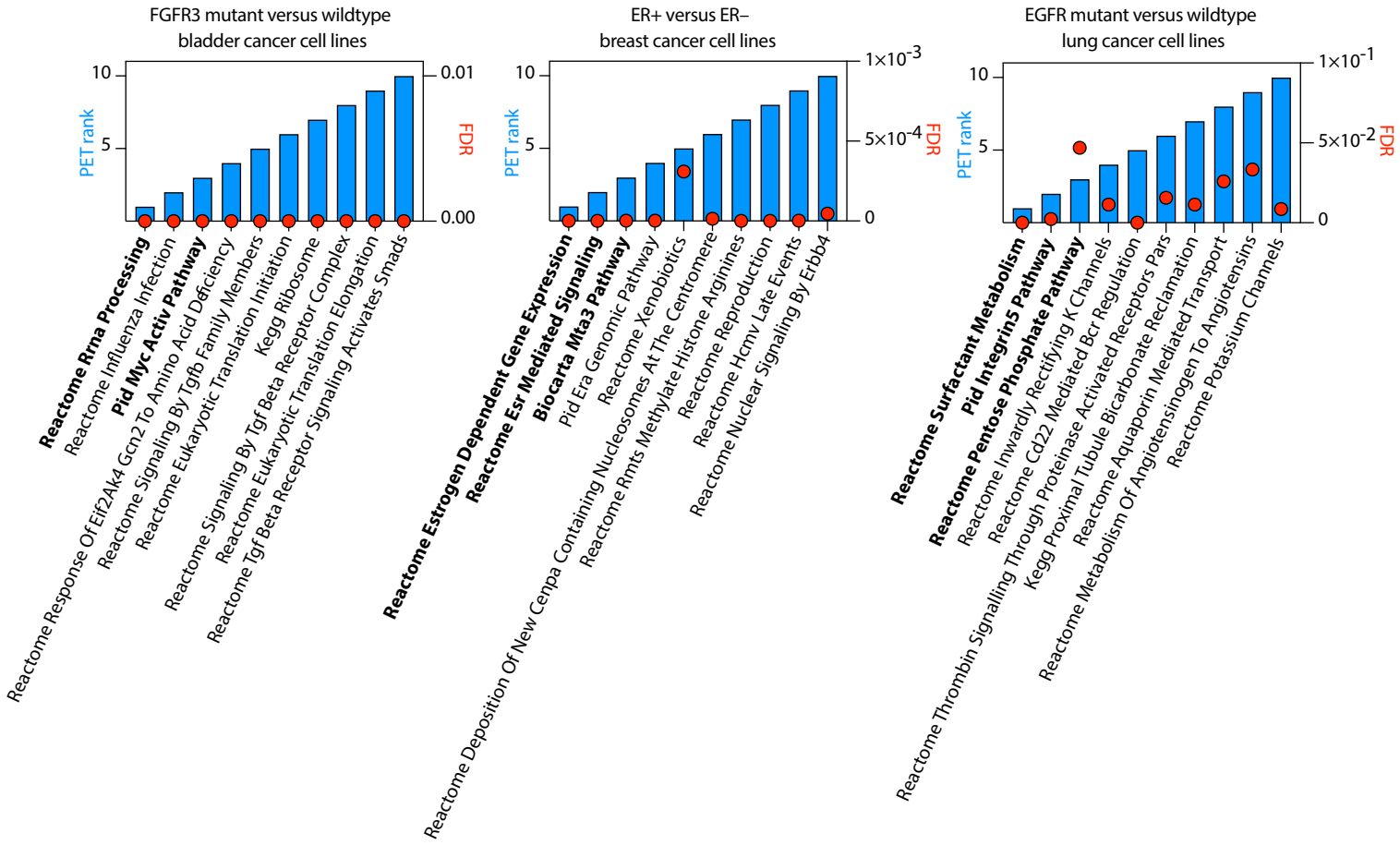

FGFR3 mutant versus wildtype
bladder cancer cell lines
PET rank
FDR
10
5
0.01
0.00
Reactome Rrna Processing
Reactome Influenza Infection
Pid Myc Activ Pathway
Reactome Response Of Eif2Ak4 Gcn2 To Amino Acid Deficiency
Reactome Signaling By Tgfb Family Members
Reactome Eukaryotic Translation Initiation
Kegg Ribosome
Reactome Signaling By Tgf Beta Receptor Complex
Reactome Eukaryotic Translation Elongation
Reactome Tgf Beta Receptor Signaling Activates Smads
ER+ versus ER–
breast cancer cell lines
PET rank
FDR
10
5
1×10-3
5×10-4
0
Reactome Estrogen Dependent Gene Expression
Reactome Esr Mediated Signaling
Biocarta Mta3 Pathway
Pid Era Genomic Pathway
Reactome Xenobiotics
Reactome Deposition Of New Cenpa Containing Nucleosomes At The Centromere
Reactome Rmts Methylate Histone Arginines
Reactome Reproduction
Reactome Hcmv Late Events
Reactome Nuclear Signaling By Erbb4
EGFR mutant versus wildtype
lung cancer cell lines
PET rank
FDR
10
5
1×10-1
5×10-2
0
Reactome Surfactant Metabolism
Pid Integrin5 Pathway
Reactome Pentose Phosphate Pathway
Reactome Inwardly Rectifying K Channels
Reactome Cd22 Mediated Bcr Regulation
Reactome Thrombin Signalling Through Proteinase Activated Receptors Pars
Kegg Proximal Tubule Bicarbonate Reclamation
Reactome Aquaporin Mediated Transport
Reactome Metabolism Of Angiotensinogen To Angiotensins
Reactome Potassium Channels

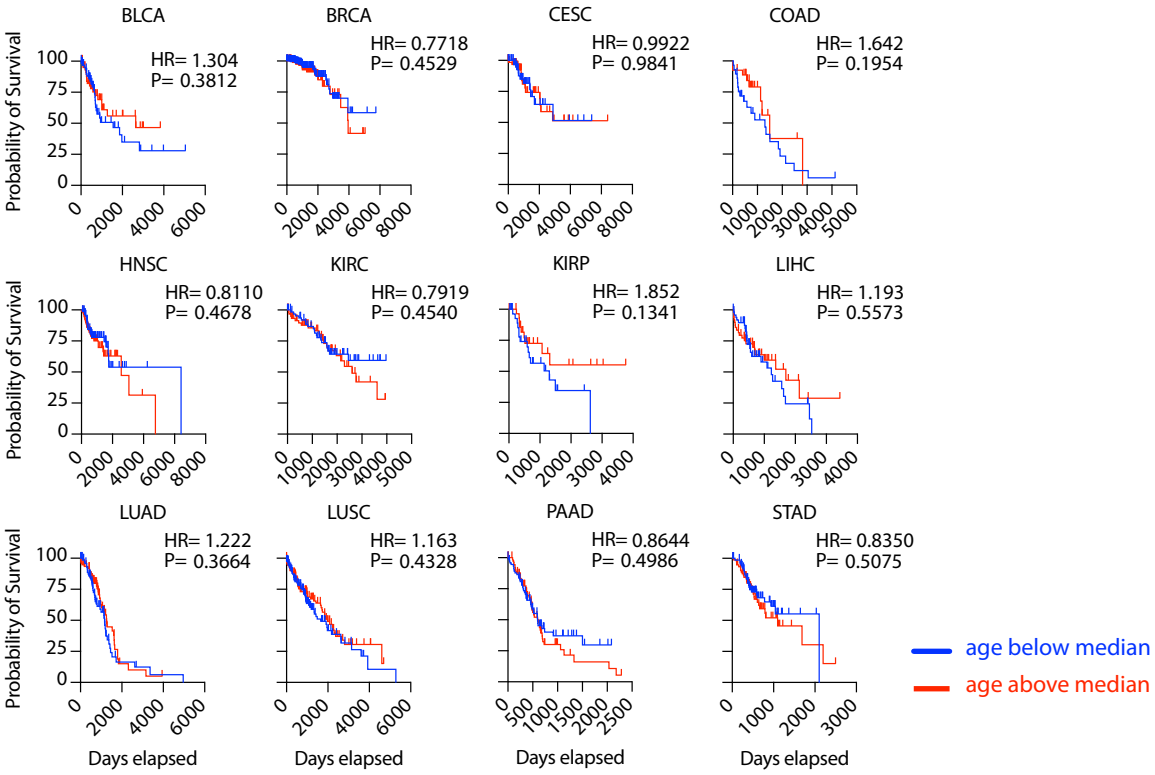
Figure S3. Age and sex are not determinants of cancer outcomes in the selected samples. a Patient demographics. p-values of Fisher exact tests for female:male sex ratios and Wilcoxon rank sum test for age are reported. **b-c** Kaplan-Meier (KM) overall survival curves comparing samples below versus above median age **(b)** and female versus male **(c)** in the indicated cancer types. Hazard ratio (HR) and logrank test p-values are reported. Samples in breast invasive carcinoma and cervical squamous cell carcinoma are biased towards female donors due to the nature of these cancer types. BLCA: bladder carcinoma; BRCA: Breast invasive carcinoma; CESC: Cervical squamous cell carcinoma; COAD: Colon adenocarcinoma; HNSC: Head and Neck squamous cell carcinoma; KIRC: Kidney renal clear cell carcinoma; KIRP: Kidney renal papillary cell carcinoma; LIHC: Liver hepatocellular carcinoma; LAUD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; PAAD: Pancreatic adenocarcinoma; STAD: Stomach adenocarcinoma. Source data are provided as a Source Data file.

FigureS3

a

	Female:Male			Age years(mean ± s.d.)		
	alive	deceased	p-val	alive	deceased	p-val
BLCA	19:65	11:34	0.83	74.4±7.2	70.7±10.2	0.1
BRCA	462:8	33:0	1.00	66.1±8.9	63.9±14.6	0.4
CESC	110:0	26:0	1.00	45.8±12.0	44.8±15.1	0.8
COAD	15:14	13:20	0.44	80.7±3.6	76.3±11.0	0.1
HNSC	23:98	12:35	0.40	65.5±6.6	64.9±14.0	0.9
KIRC	29:60	17:26	0.44	66.0±7.1	67.2±12.2	0.6
KIRP	10:29	9:15	0.40	62.2±13.5	59.1±13.7	0.3
LIHC	19:39	21:23	0.15	66.1±5.8	64.9±11.1	0.6
LAUD	36:40	51:28	0.04	70.0±5.7	67.9±10.4	0.2
LUSC	47:102	24:83	0.12	68.4±8.1	69.2±8.4	0.5
PAAD	26:40	40:45	0.41	64.2±10.4	66.4±10.8	0.2
STAD	42:63	18:35	0.49	66.4±10.6	68.4±9.8	0.3

b



c

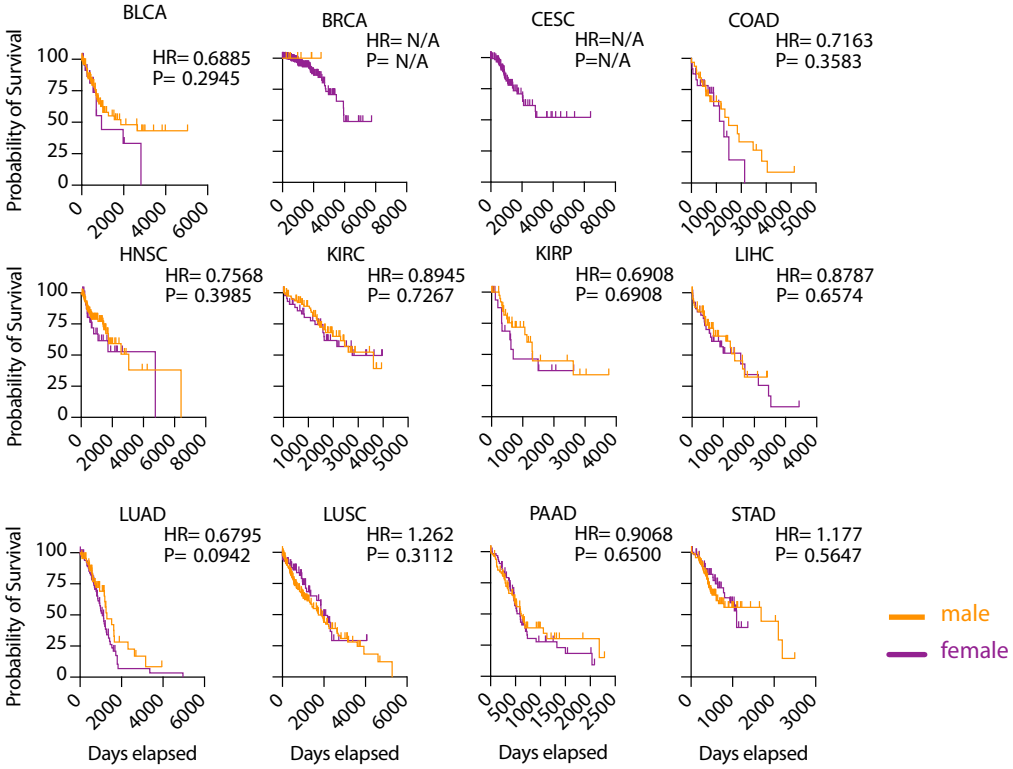
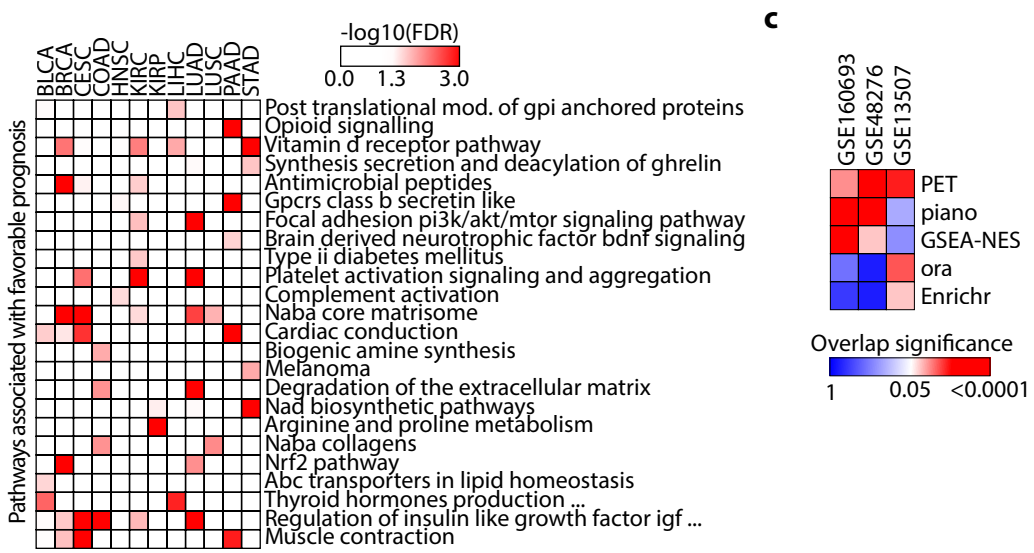


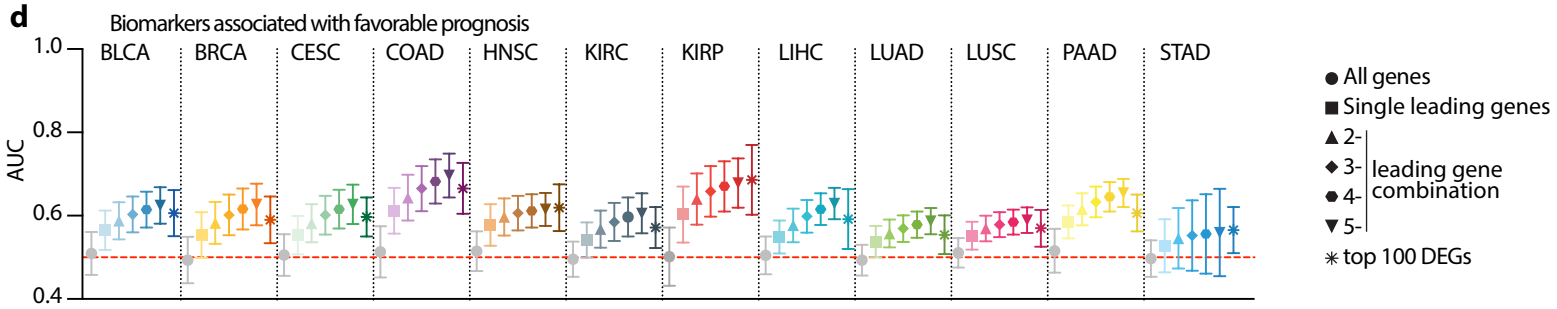
Figure S4. PET identifies pathways and gene combinations associated with favorable prognosis. a *CD8A* mRNA expression and the p-value of Wilcoxon rank sum test comparing alive and deceased samples are provided. **b** Top pathways from each cancer that were associated with favorable survival. FDR values are from PET. **c** Overlap between the top reported pathways associated with unfavorable outcomes in TCGA-BLCA versus three independent datasets using indicated methods. Fisher exact test p-values are displayed. **d** Area under the curve (AUC) distribution of indicated sets of genes or their combination to predict alive status in each cancer type. Leading genes were selected from top 20 significant pathways associated with favorable prognosis. Error bars show mean ± s.d. **e** Elbow plots showing significant AUCs (low to high) from all 1-5 leading gene combinations. Statistics were performed according to FDR-adjusted logrank test of Kaplan–Meier (KM) survival curves. CESC and LUSC did not yield any significant favorable predictors. **f** KM plots showing the power of the top combination biomarker (darker lines) and the top DEG (lighter lines) of favorable survival to stratify overall survival in each cancer type. Samples were evenly split into two groups based on the z-score values of the indicated biomarker. Hazard ratio (HR) and logrank test p-values are reported. * $p<0.01$; ** $p<0.001$; *** $p< 0.0001$. **g** Consistency between significant prognostic biomarkers in TCGA and independent cohorts (see **Methods**). The prognostic biomarkers derived from PET pathways or differentially expressed genes (DEGs) are separated. Independent cohorts are listed in **Supplementary Data2g**. Error bars show mean ± s.d. **$p<0.01$ derived from two-tailed paired t-test. **h** KM plots showing the power of the indicated gene mutation to stratify overall survival in each cancer type. Samples were split into two groups, with one group having the indicated gene mutation and the other group with no indicated mutation. HR and logrank test p-values are reported. * $p<0.01$; ns, not statistically significant. Source data are provided as a Source Data file.

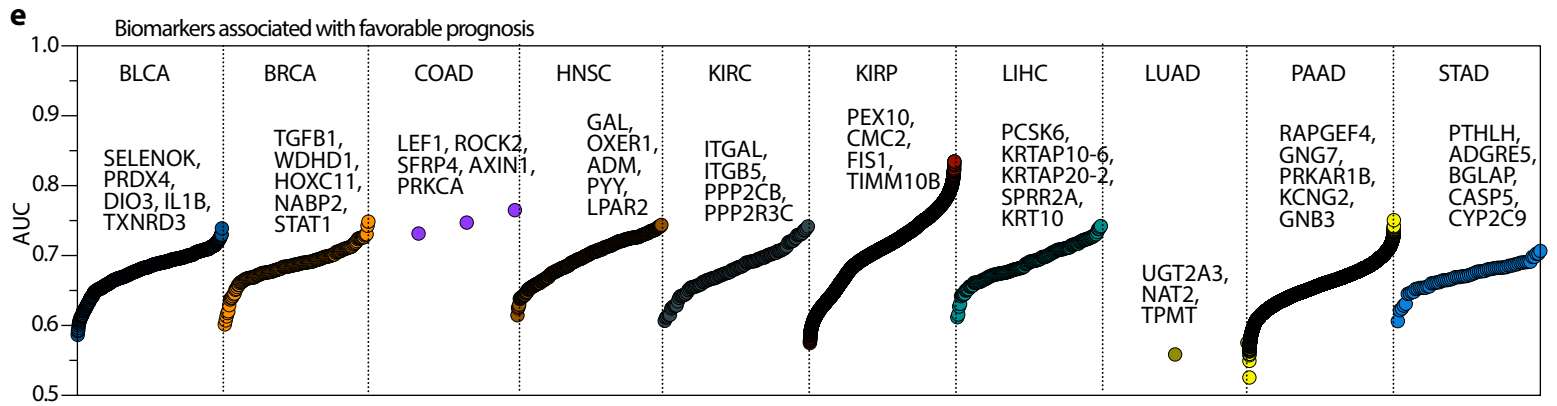
FigureS4

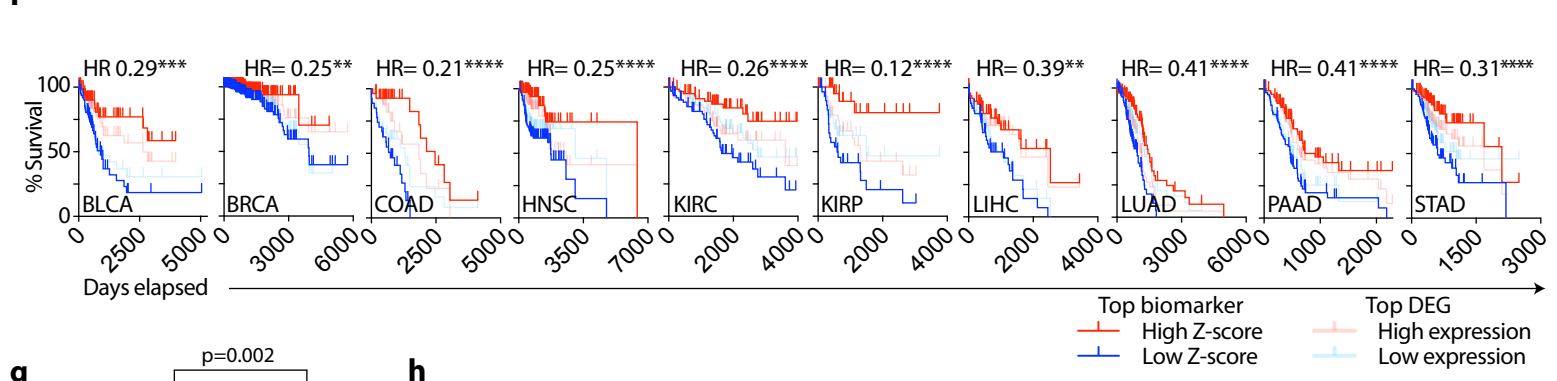
a

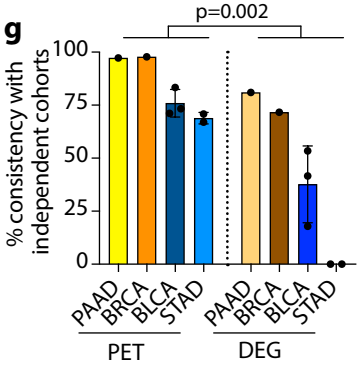
	CD8A FPKM (mean ± s.d.)		
	alive	deceased	p-val
BLCA	4.2±7.0	3.4±5.2	0.55
BRCA	4.1± 5.7	2.9±2.5	0.51
CESC	8.5±14.6	4.6±3.7	0.75
COAD	2.6±3.1	2.5±2.3	0.85
HNSC	4.1±5.5	3.2±3.6	0.40
KIRC	13.0±16.1	10.2±15.1	0.23
KIRP	2.3±3.3	4.1±5.5	0.36
LIHC	1.6±2.6	2.1±4.0	0.30
LAUD	5.0±4.5	5.0±4.9	0.85
LUSC	4.6±4.7	4.3±3.8	0.91
PAAD	3.8±3.9	3.1±3.2	0.39
STAD	4.7±5.5	5.1±5.1	0.27











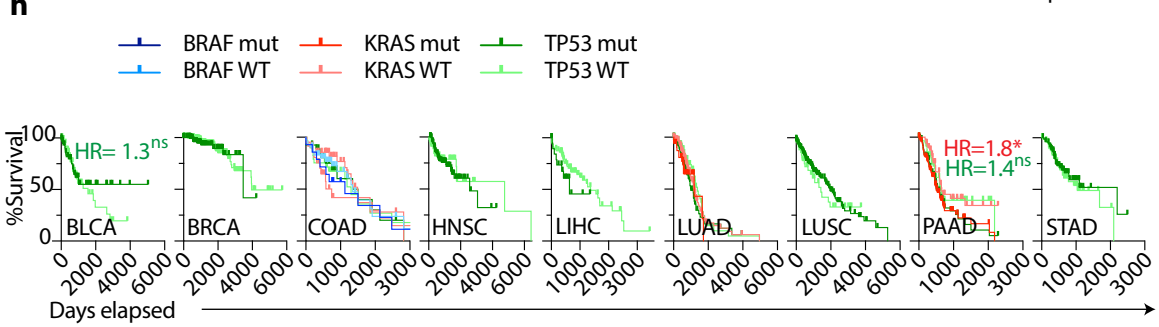


Figure S5. Combinations of favorable and unfavorable prognostic biomarkers offer more precise overall survival stratification. **a** Venn diagrams showing the distribution of patients in indicated cancer type based on expression of the top favorable (labeled in **Fig. S4e**) and the top unfavorable (labeled in **Fig. 2e**) prognostic markers. Kaplan–Meier (KM) overall survival curves for 4 subgroups of patients in each cancer type whose tumors express the indicated combination of prognostic biomarkers. The four subgroups are highlighted with bold font in Venn diagrams. Hazard ratio (HR) and logrank test p-values denote comparisons with the lines in red. ns, not statistically significant; * $p<0.05$; ** $p<0.01$; *** $p<0.001$; **** $p< 0.0001$. Source data are provided as a Source Data file.

FigureS5

a

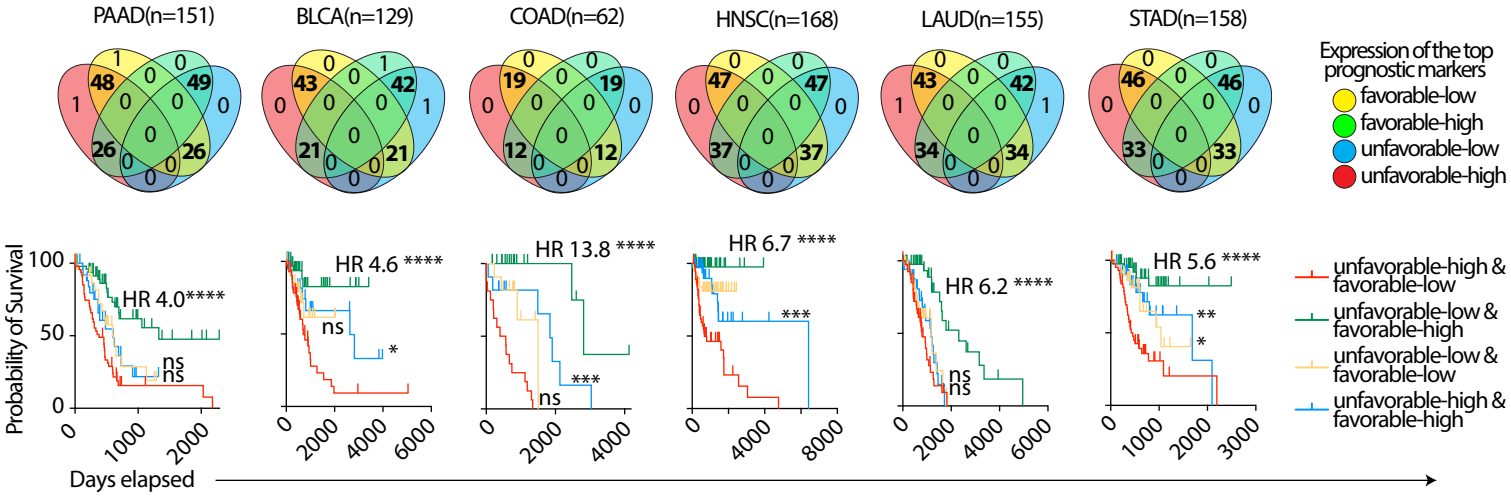
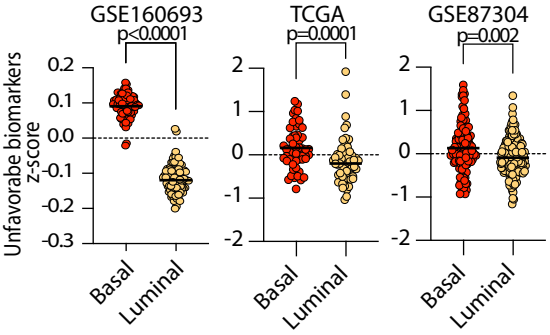

PAAD(n=151)
BLCA(n=129)
COAD(n=62)
HNSC(n=168)
LAUD(n=155)
STAD(n=158)
Expression of the top prognostic markers
favorable-low
favorable-high
unfavorable-low
unfavorable-high
Probability of Survival
HR 4.0****
HR 4.6 ****
HR 13.8 ****
HR 6.7 ****
HR 6.2 ****
HR 5.6 ****
Days elapsed
unfavorable-high & favorable-low
unfavorable-low & favorable-high
unfavorable-low & favorable-low
unfavorable-high & favorable-high

Figure S6. Prognostic biomarkers distinguish existing molecular subtypes of bladder cancer. a Plots showing the expression (indicated by z-score) of top 100 significant unfavorable prognostic biomarkers in RNA-seq of luminal versus basal sub-types of bladder cancer in three indicated cohorts. Each dot represents the mean z-score of an unfavorable prognostic biomarker across all luminal or basal samples. p-values are calculated by two-sided Wilcoxon rank sum test. **b** Heatmap showing expression of leading genes from PET-derived prognostic pathways from of human bladder cancers across n=56 canine tumors (see **Methods**). Canine tumors and annotations have been sourced from GSE110661. Source data are provided as a Source Data file.

FigureS6

a


GSE160693
p<0.0001
TCGA
p=0.0001
GSE87304
p=0.002
Unfavorabe biomarkers
z-score
0.2
0.1
0.0
-0.1
-0.2
-0.3
2
1
0
-1
-2
Basal
Luminal

b

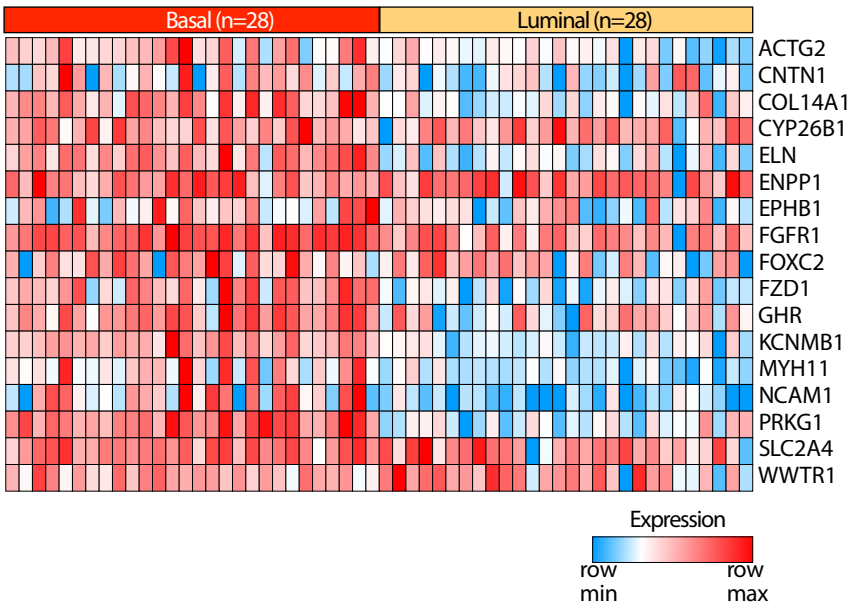
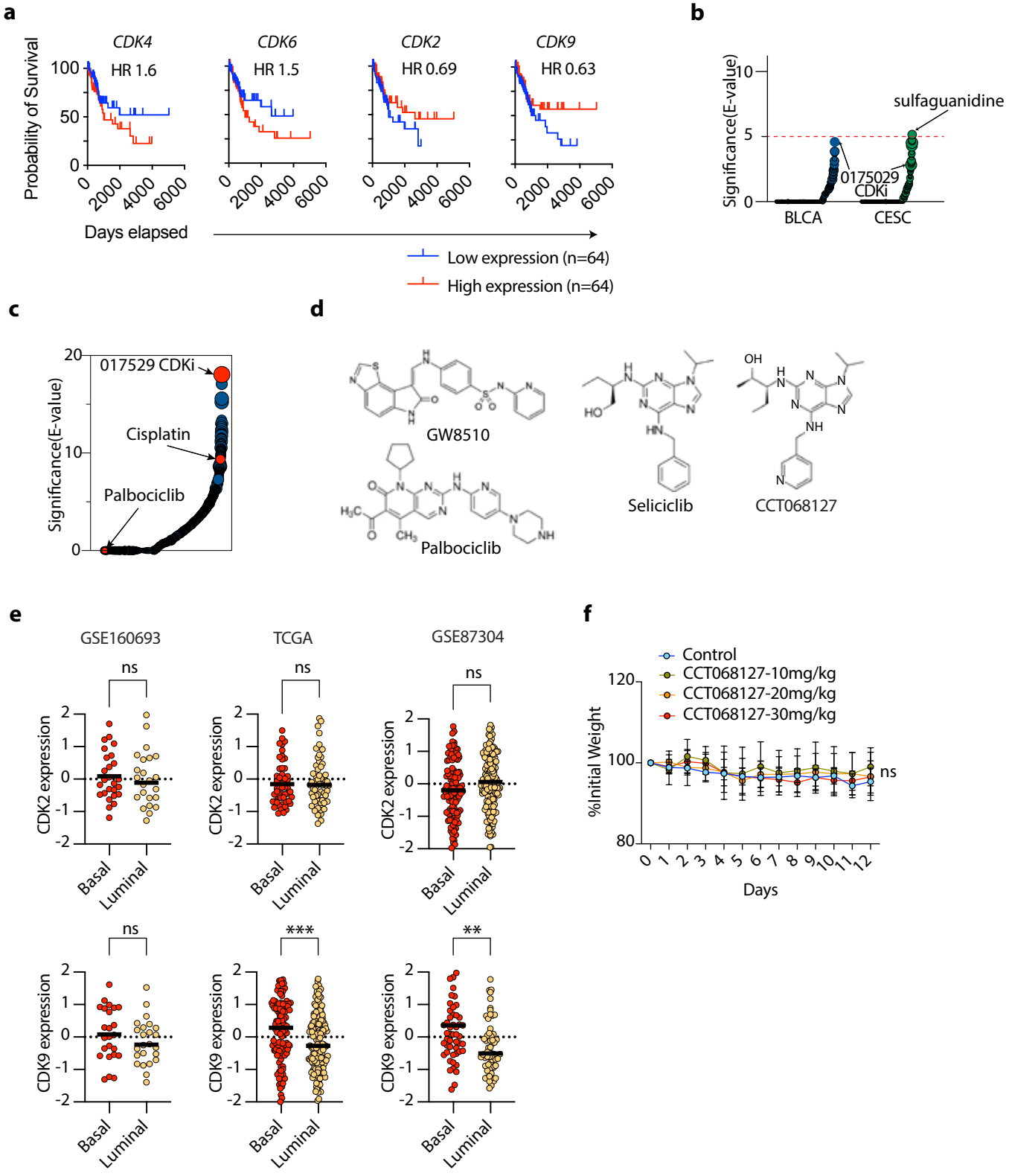

Basal (n=28)
Luminal (n=28)
ACTG2
CNTN1
COL14A1
CYP26B1
ELN
ENPP1
EPHB1
FGFR1
FOXC2
FZD1
GHR
KCNMB1
MYH11
NCAM1
PRKG1
SLC2A4
WWTR1
Expression
row min
row max

Figure S7. *In vitro* and in vivo efficacy of predicted CDK2/9 inhibition in restricting bladder cancers. a Kaplan–Meier (KM) survival curve showing the power of indicated gene expression in stratifying overall survival in BLCA. Samples were evenly split into two groups, one group with samples having the top 50% and the other group having the bottom 50% of expression values. Hazard ratios (HRs) are reported. **b** Elbow plots showing the significance of overlap between drug up (or down) regulated targets with the 200 most differentially expressed genes (DEGs) in bladder and cervical cancers. The top predicted drug for each cancer type is labeled. red dashed line denotes Bonferroni adjusted E-values <0.05 (see Methods). **c** Elbow plots showing the significance of overlap between drug up (or down) regulated genes with leading genes from PET derived pathways with favorable (or unfavorable) prognostic potential in bladder cancer. Included are all drugs from LINC1000[101] and DSigDB[68], sorted based on the significance. Selected drugs are highlighted. **d** Chemical structures of CDK inhibitors used in this study. **e** Plots showing the expression of CDK2 and CDK9 in RNA-seq of luminal versus basal sub-types of bladder cancer in three indicated independent cohorts. Each dot represents a sample in the cohort. y-axes denote z-scores of expression values. **f** Plots showing changes in weight of mice over time during treatment with DMSO control or indicated dose of CCT068127. The DMSO dose is equivalent to 30mg/kg concentration of CCT068127 treatment. p-values in **e** are by two-sided Wilcoxon rank sum test and in **f** are by two-way ANOVA with multiple comparisons adjustment. ns non-significant; ** $p<0.01$; *** $p< 0.001$. Error bars show mean ± s.d. Source data are provided as a Source Data file.

FigureS7


a
CDK4
HR 1.6
CDK6
HR 1.5
CDK2
HR 0.69
CDK9
HR 0.63
Probability of Survival
Days elapsed
Low expression (n=64)
High expression (n=64)
b
Significance(E-value)
sulfaguanidine
0175029 CDKi
BLCA
CESC
c
017529 CDKi
Cisplatin
Palbociclib
Significance(E-value)
d
GW8510
Palbociclib
Seliciclib
CCT068127
e
GSE160693
TCGA
GSE87304
CDK2 expression
CDK9 expression
ns
***
**
Basal
Luminal
f
Control
CCT068127-10mg/kg
CCT068127-20mg/kg
CCT068127-30mg/kg
%Initial Weight
Days
ns