

Cell Genomics, Volume 3

Supplemental information

**Systematic assessment of prognostic
molecular features across cancers**

Balaji Santhanam, Panos Oikonomou, and Saeed Tavazoie

Supplemental figures and legends

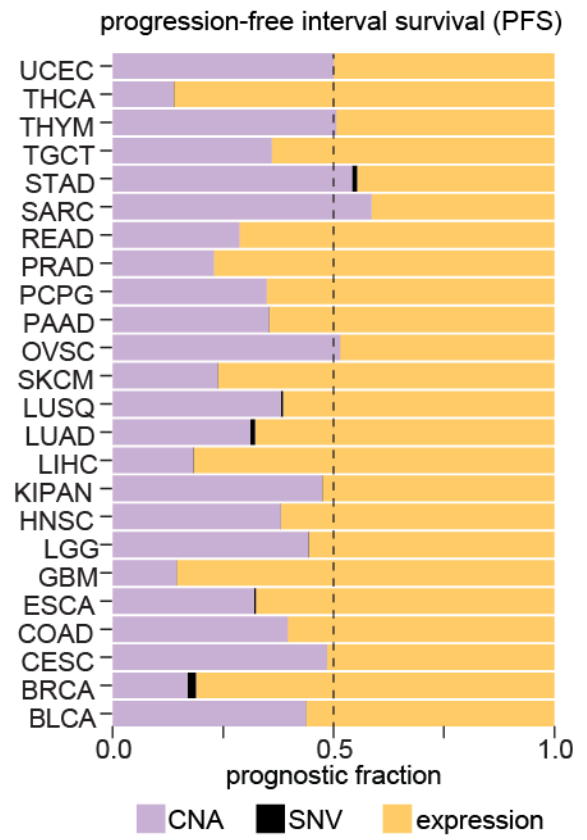
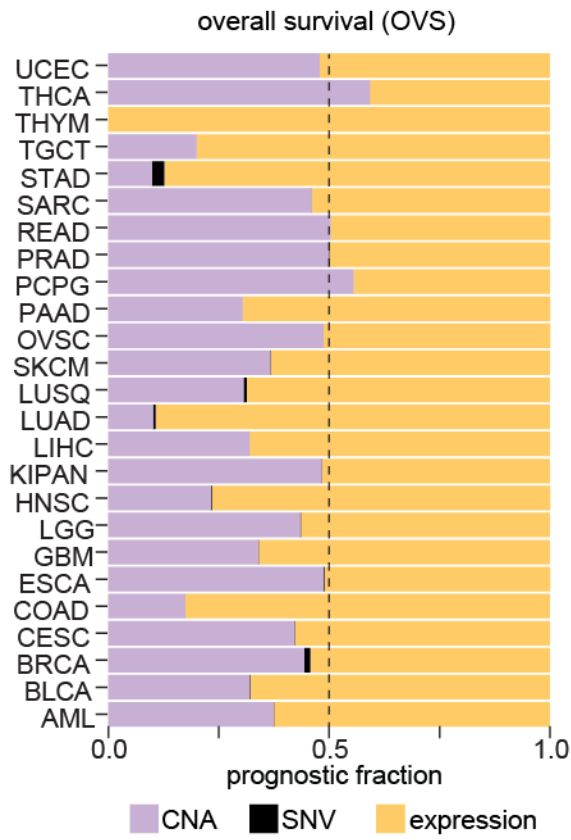


Figure S1. Prognosis specified by genomic and transcriptomic observations on individual genes, Related to Figure 1.

(a) Proportion of genes prognostic based on their copy-number aberrations (purple), mutation statuses (black) and expression changes (yellow) in each cohort (Y-axis) for overall survival (left panel) and progression-free interval survival (right panel).

overall survival

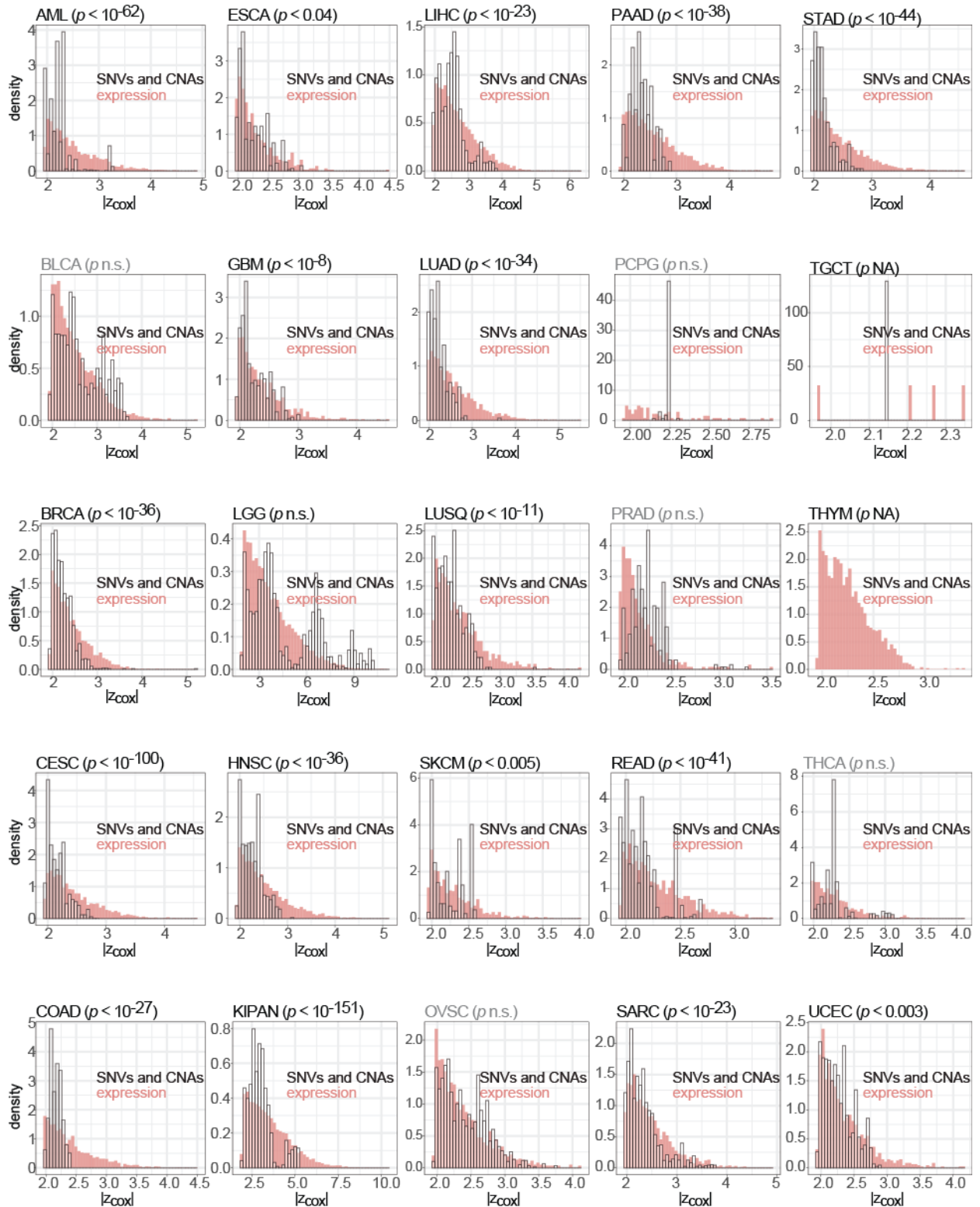


Figure S2. Gene expression changes are more strongly prognostic for overall survival than either copy-number aberrations or mutations in individual genes, Related to Figure 1.

Distributions of absolute standardized significance scores (X-axis; Wald statistic) in patient stratifications based on gene expression changes or copy-number aberrations and mutation statuses are shown for each cohort in TCGA (color key indicated). For each gene's expression and copy-number, patients were stratified into groups with relatively high and low expression (positive and negative z-scored gene expression) and, high and low copy-number (positive and negative germline-subtracted segmented somatic copy-number) respectively, to quantify differences in overall survival. For mutations, patients were grouped into those with observed mutations and those with no observed mutations to quantify differences in overall survival. One-sided Mann-Whitney test p -values are indicated for each cohort.

progression-free interval survival

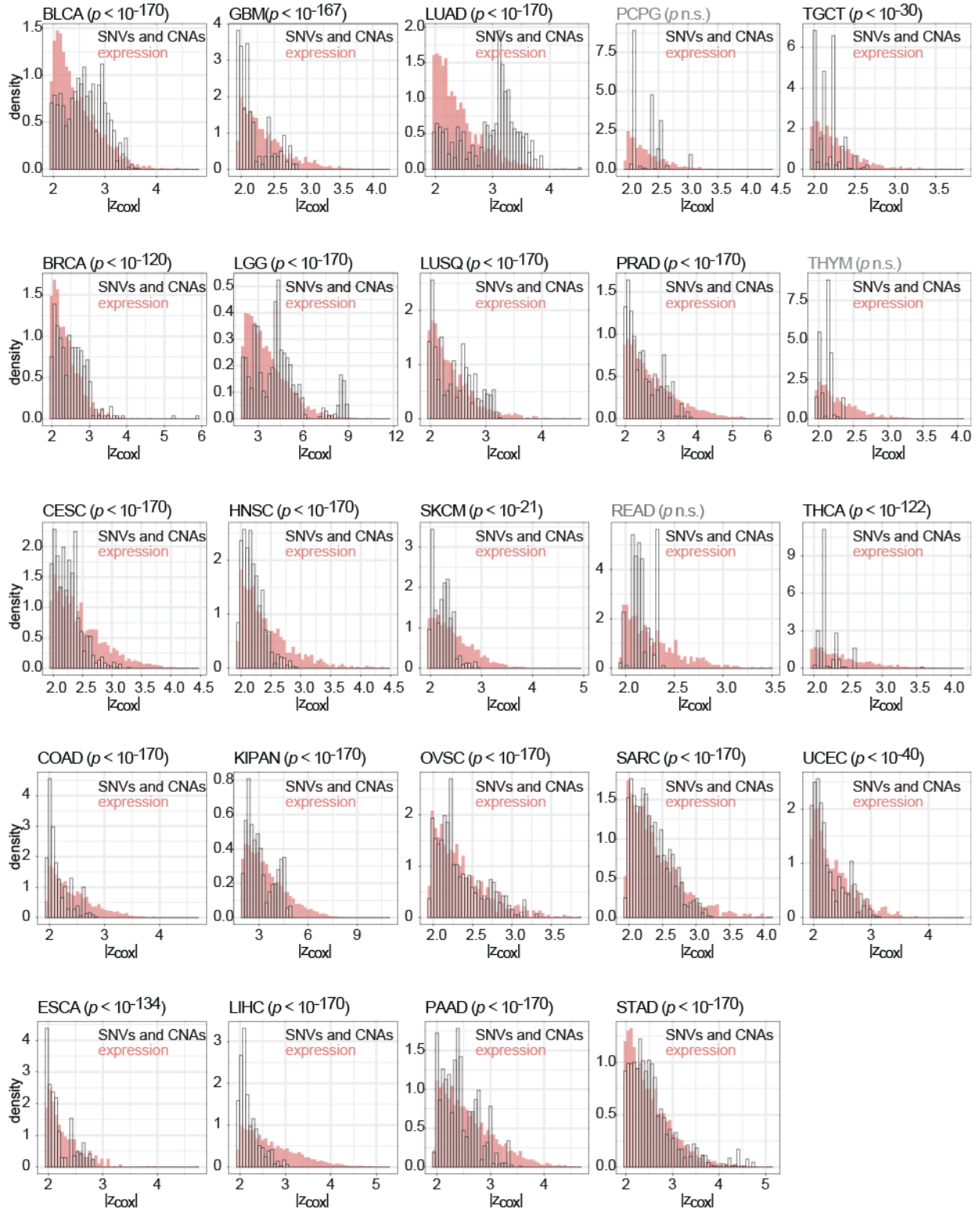


Figure S3. Gene expression changes are more strongly prognostic for progression-free interval survival than either copy-number aberrations or mutations in individual genes, Related to Figure 1.

Distributions of absolute standardized significance scores (X-axis; Wald statistic) in patient stratifications based on gene expression changes or copy-number aberrations and mutation statuses are shown for each cohort in TCGA (color key indicated). For each gene's expression and copy-number, patients were stratified into groups with relatively high and low expression (positive and negative z-scored gene expression) and, high and low copy-number (positive and negative germline-subtracted segmented somatic copy-number) respectively, to quantify differences in progression-free interval survival. For mutations, patients were grouped into those with observed mutations and those with no observed mutations to quantify differences in progression-free interval survival. One-sided Mann-Whitney test p -values are indicated for each cohort.

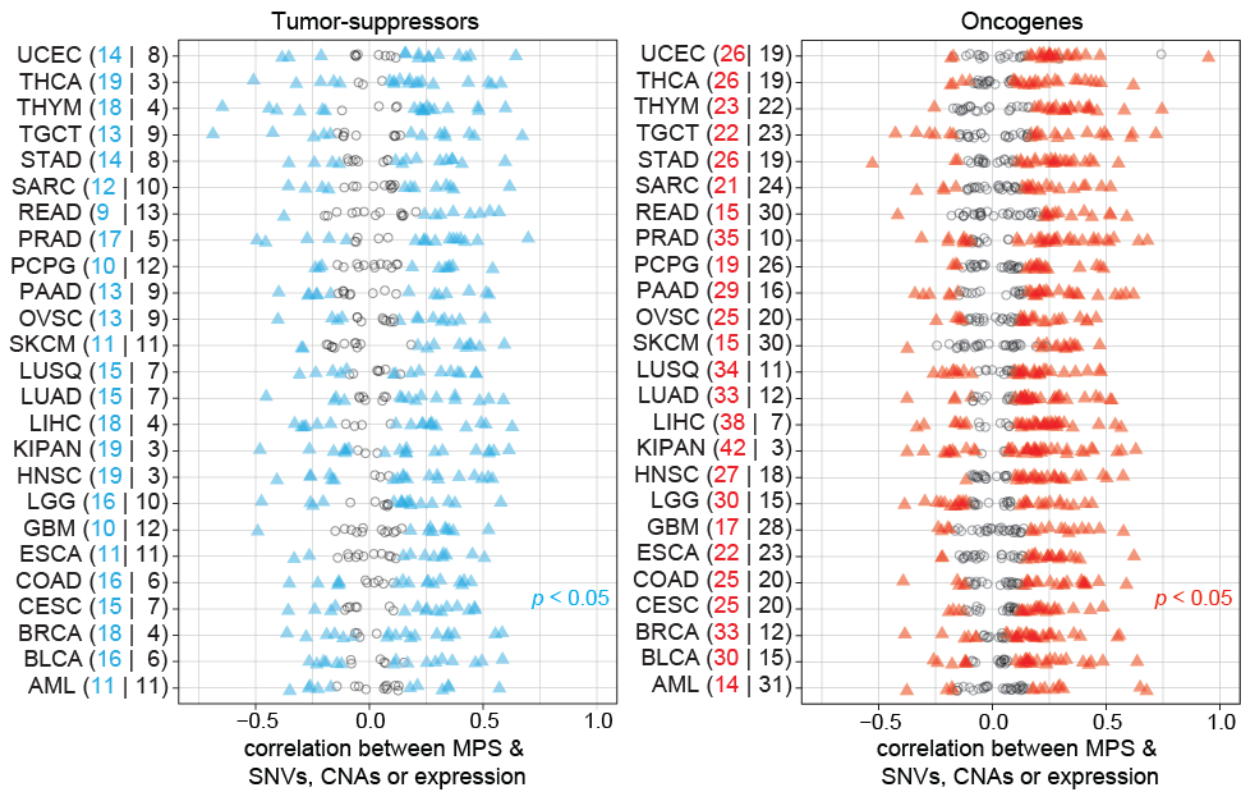


Figure S4. Molecular measurements on genes encoding tumor-suppressors and oncogenes are largely correlated with perturbations in their associated modules, Related to Figure 1.

Distributions of correlations coefficients between module perturbations of 67 cancer-drivers¹ with mutations, copy-number aberrations or expression changes in genes encoding them across 25 cancers from TCGA (Y-axis). For every cohort, there are 22 points corresponding to tumor-suppressors (left panel) and 45 points corresponding to oncogenes¹. Each dot represents the strongest correlation between module perturbations and each of the three single-gene measurements. Spearman's correlation coefficients were computed between copy-number aberrations and expression changes with module perturbations whereas Pearson's correlation coefficients were used for mutation statuses. Numbers of significant (p -value < 0.05 ; blue or red triangles) and not significant (p -value > 0.05 ; gray dots) correlation coefficients are indicated for each cohort.

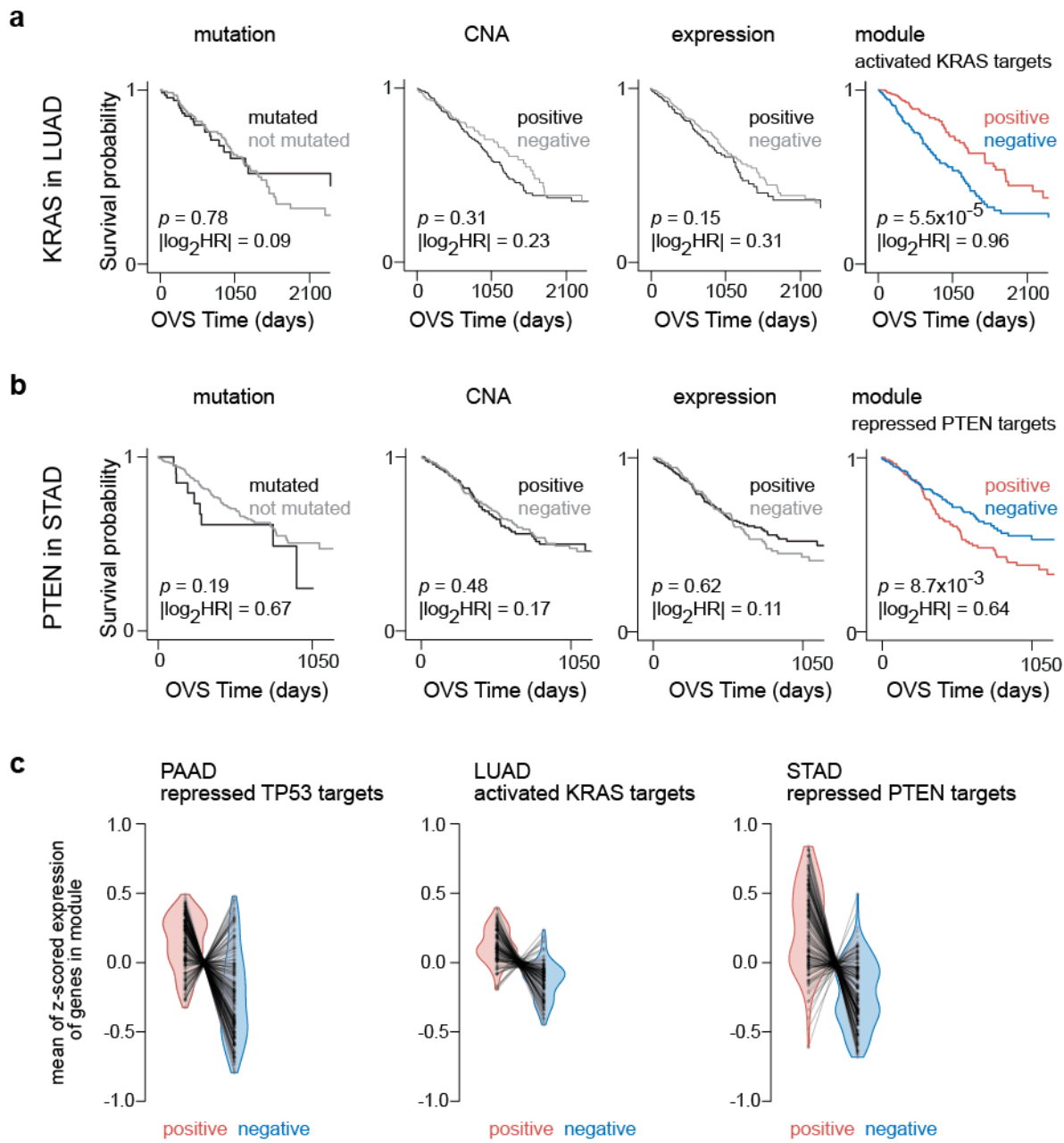


Figure S5. Perturbations in modules associated with KRAS and PTEN provide better patient survival stratification than measurements on individual genes, Related to Figure 1.

(a-b) Patients stratified based on mutation status, copy-number aberrations and expression changes in (a) KRAS in lung adenocarcinoma (LUAD) and, (b) PTEN in stomach cancer (STAD). Also shown are patient stratifications based on perturbations in module comprising (a) genes up-regulated in lung cancer cells over-expression KRAS (MSigDB² M12860) and, (b) genes up-regulated upon knockdown of PTEN (MSigDB² M2787). (c) Violin plots show distributions of the mean expression of genes that make up the modules corresponding to repressed targets of TP53 (MSigDB M2698; 198 genes) in pancreatic cancer (left panel), activated targets of KRAS (MSigDB M12860; 139 genes) in lung adenocarcinoma (middle panel) and repressed targets of PTEN (MSigDB M2787; 186 genes) in stomach cancer cohort (right panel). Patients have been stratified into groups with positive (red) and negative (blue) perturbation scores for these modules. Genes that constitute the modules are shown (black dots) and their mean expression values in the MPS positive and negative groups are linked (black lines). Slopes of the lines indicate whether genes are up- (negative slope) or down-regulated (positive slope) in the MPS positive compared to the MPS negative patient groups. Each module shown comprises of genes with both high and low average expression, while its perturbation scores is statistically significant and it significantly discriminates patients based on survival. Stratification of patients based on the perturbation scores of the three modules shown in this sub-panel are plotted in Fig 1d, Fig S5a (right panel) and Fig S5b (right panel), respectively.

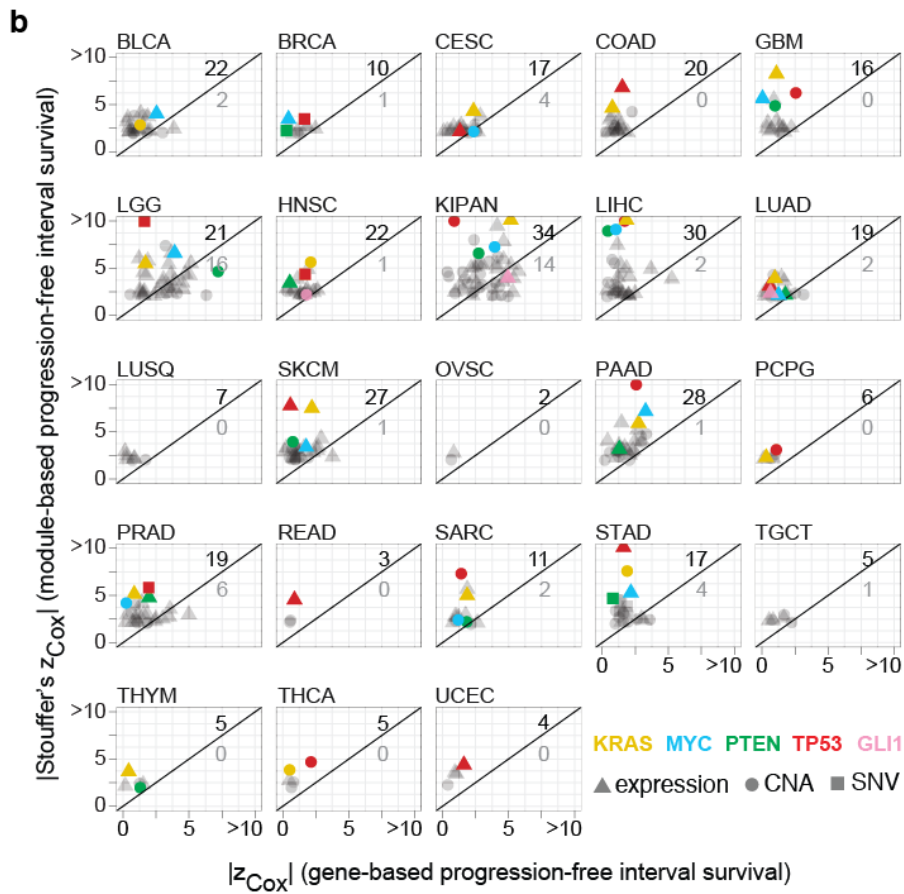
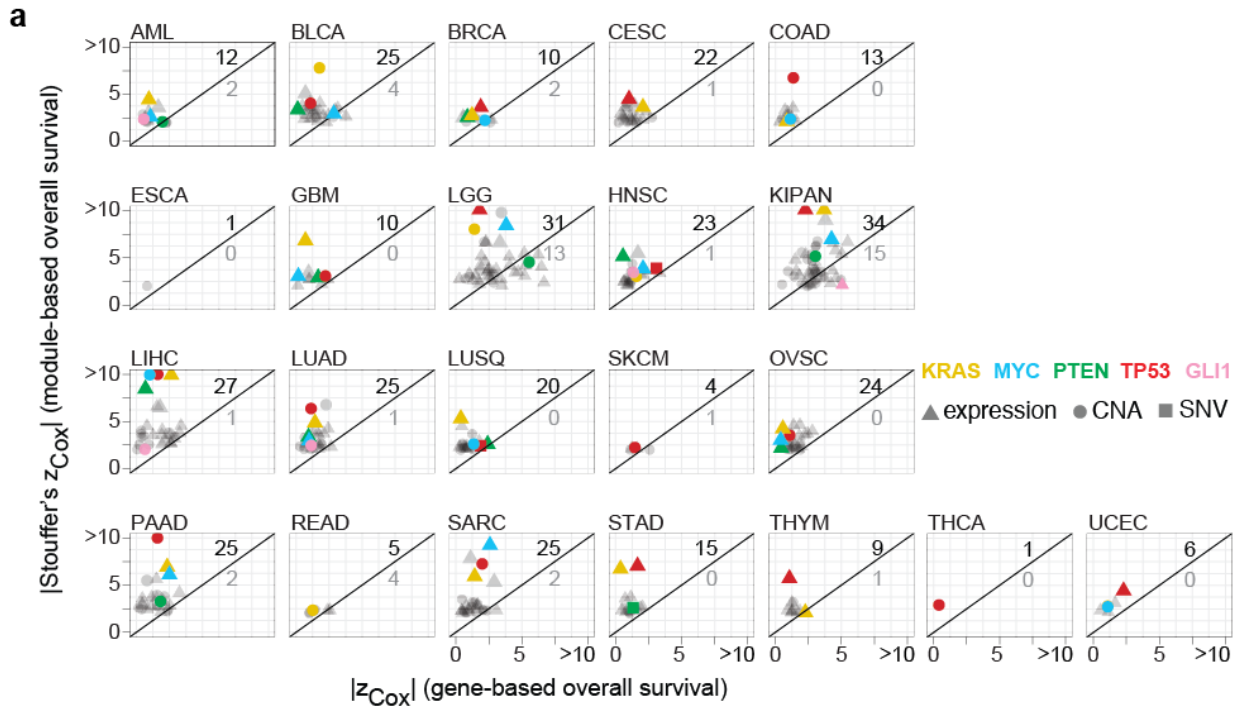


Figure S6. Perturbation scores of modules associated with oncogenes and tumor-suppressors provide better patient survival stratification than measurements on individual genes, Related to Figure 1.

(a-b) Scatterplots show comparisons between absolute standardized significance scores (Wald statistic) individual genes and their corresponding modules (Wald statistic summarized by Stouffer's method) in individual cohorts (labels indicated on top). For individual gene measurements of each oncogene or tumor-suppressor, we report the absolute values of the standardized significance scores for best gene-based stratification (mutation, copy-number or expression; shown in legend). The shape of each dot indicates the nature of measurements used. The modules whose prognostic value is better than the individual gene measurements are above the reference line $y=x$. The number of oncogenes and tumor-suppressors whose modules have better (worse) prognostic value than individual gene measurements are shown above (below) the reference line. A handful of prominent oncogenes and tumor-suppressors are highlighted (color legend shown). Comparisons are shown for (a) overall survival and, (b) progression-free interval survival.

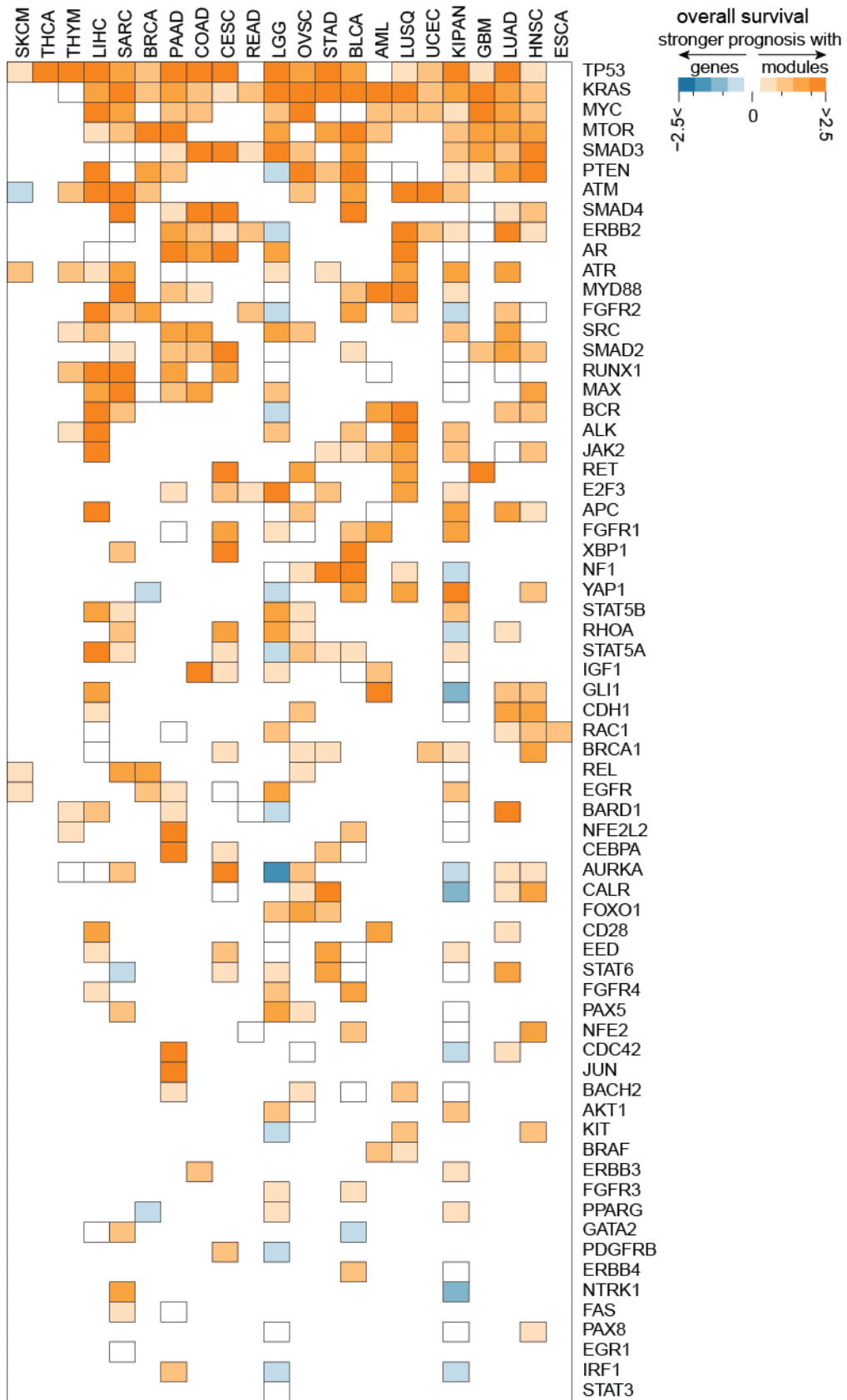


Figure S7. Perturbation scores of modules associated with oncogenes and tumor-suppressors convey stronger overall survival prognosis than measurements on individual genes, Related to Figure 1.

The \log_2 -ratio of the absolute standardized significance of modules associated with cancer-drivers¹ and measurements on these individual genes (in rows) are visualized across cancers from TCGA (columns). Standardized significance (Wald statistic) for individual genes were chosen to be the maximum from expression, copy-number and mutation based patient stratifications in each cohort. For the corresponding modules, standardized scores were summarized using Stouffer's method (Methods). Rows and columns are clustered using hierarchical clustering using Euclidean distance with optimal ordering of leaves.

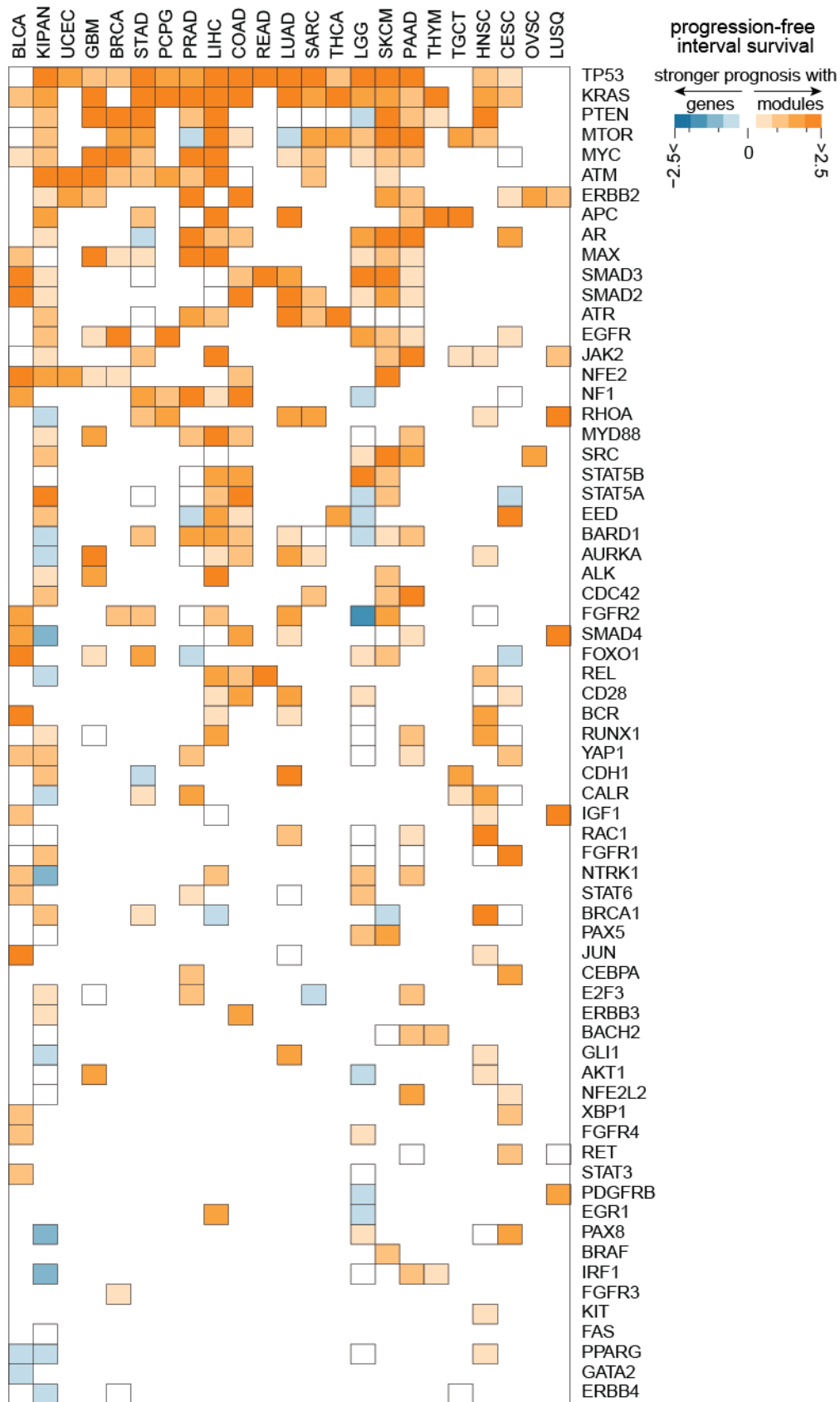


Figure S8. Perturbation scores of modules associated with oncogenes and tumor-suppressors convey stronger progression-free interval survival prognosis than measurements on individual genes, Related to Figure 1.

The \log_2 -ratio of the absolute standardized significance of modules associated with cancer-drivers¹ and measurements on these individual genes (in rows) are visualized across cancers from TCGA (columns). Standardized significance (Wald statistic) for individual genes were chosen to be the maximum from expression, copy-number and mutation based patient stratifications in each cohort. For the corresponding modules, standardized scores were summarized using Stouffer's method (Methods). Rows and columns are clustered using hierarchical clustering using Euclidean distance with optimal ordering of leaves.

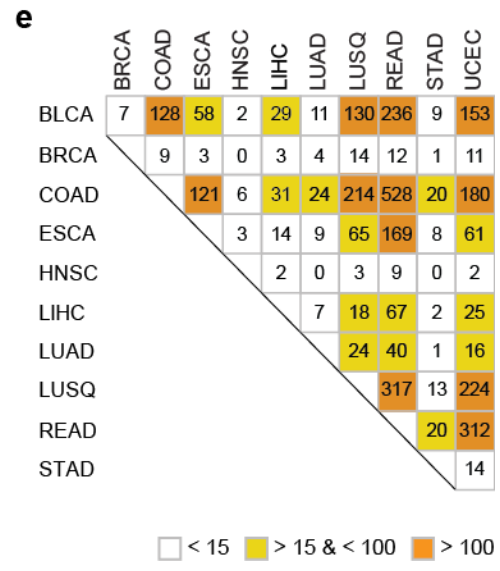
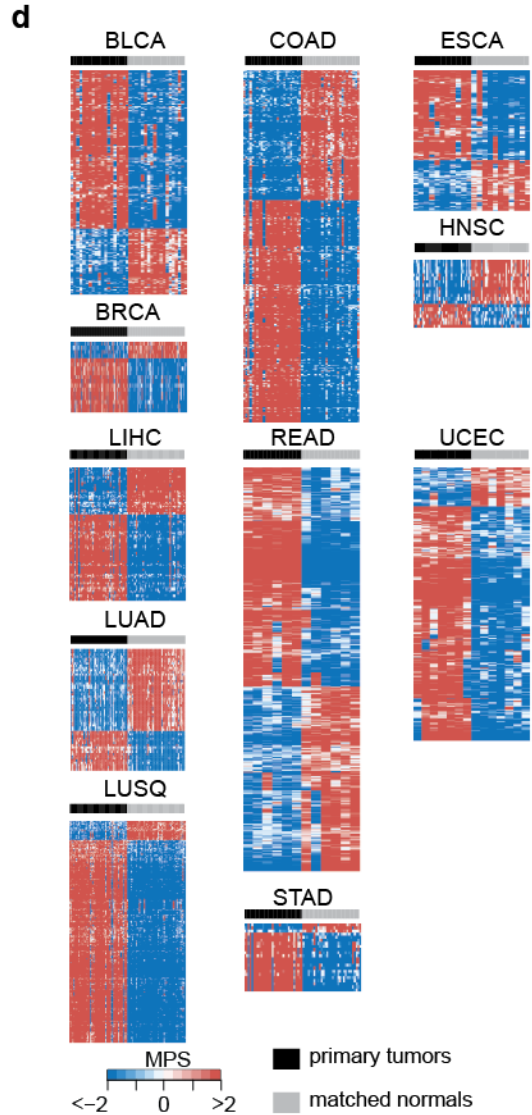
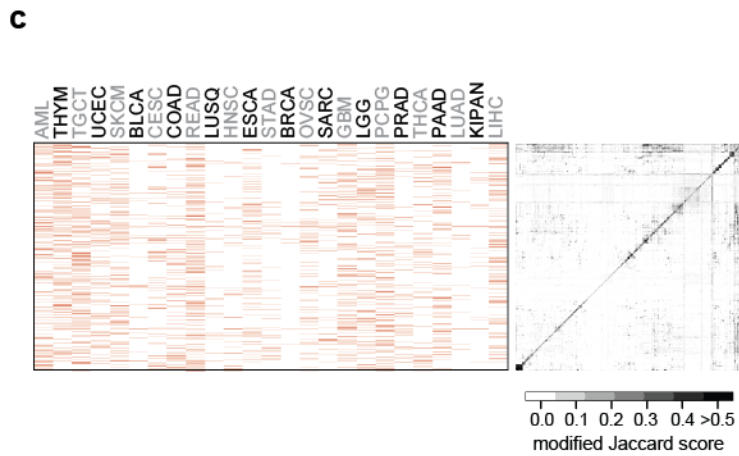
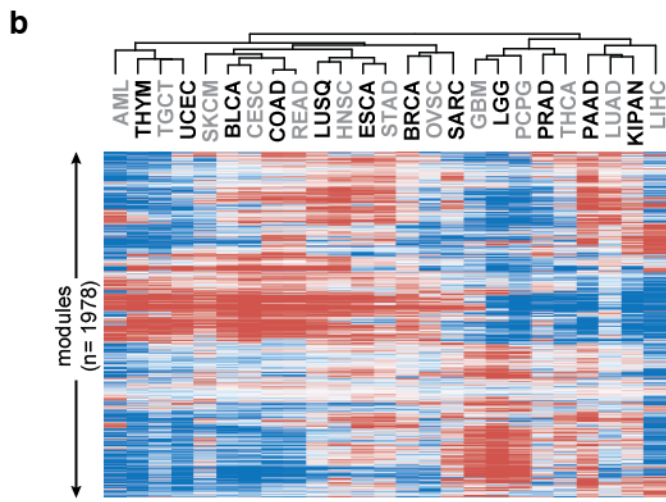
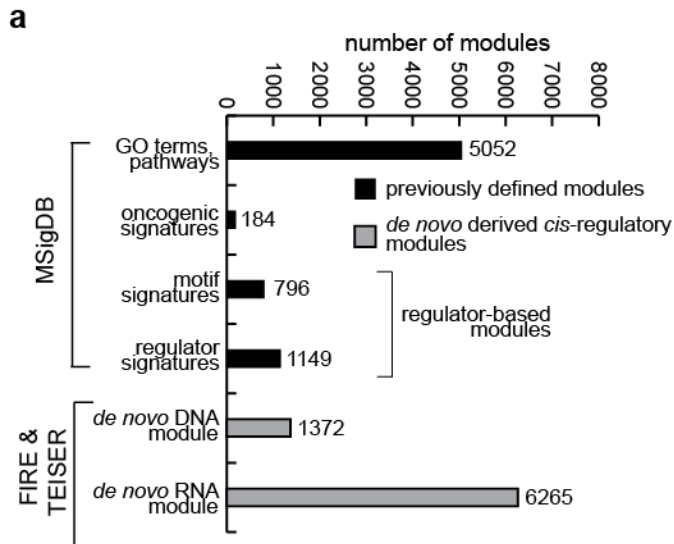


Figure S9. Module perturbation scores are reflective of cohort-specific states and tumor-normal differences, Related to STAR Methods.

(a) Number of all modules considered in this study per category. Previously defined gene modules include well-defined pathways, ontology attributes (from KEGG, PID, REACTOME, Gene Ontology collections), oncogenic signatures (MSigDB) and DNA/RNA sequence motifs characterized to be bound by known regulatory factors²⁻⁵. Modules associated with *de novo* discovered DNA, linear and structural RNA sequence motifs were discovered using transcriptome analyses of patient biopsies by applying the linear motif discovery algorithm FIRE⁶ and the structural motif discovery algorithm TEISER⁷. (b) MPS patterns of modules dysregulated in cohort-specific ways (see Methods). Rows (modules) are clustered based on patterns of MPS (scale indicated) in these samples. The median module perturbation scores of cohort-specific modules across samples within each cohort are visualized (scale indicated). (c) Heatmaps show the membership of cohort-specific modules in cancers (left panel; Methods section) and the overlap between their constituent genes (right panel). The binary matrix (left panel) indicates modules identified in a cancer in red. The extent of overlap is quantified using a modified Jaccard score (scale indicated; Methods). In both heatmaps, rows (modules) are ordered based on the overlap matrix and columns (cancers) in the binary matrix (left panel) have the same order as in panel (b). (d) Tumor-specific MPS patterns of modules dysregulated in primary tumors and matched normal samples in 11 different cohorts (labels shown). Rows (modules) are clustered based on patterns of MPS (scale indicated) in these samples. (e) Matrix shows number of shared modules dysregulated in primary tumors compared to matched normal samples across 11 different cancer cohorts. The colors indicate the number of shared modules between cancers (color key indicated).

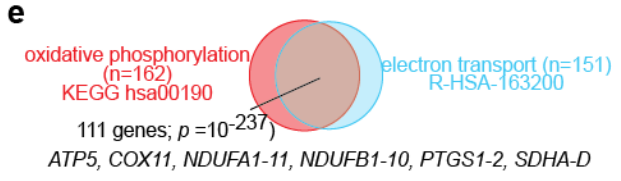
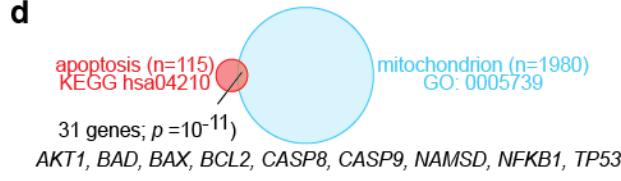
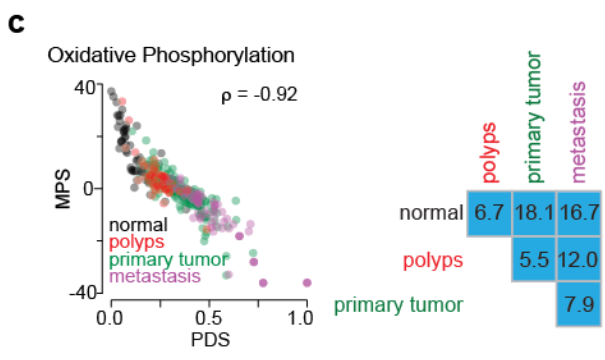
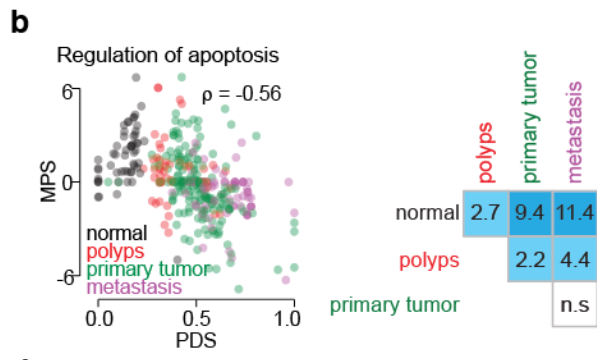


Figure S10. Module perturbation scores are correlated with cancer stage, Related to STAR Methods, Data Sheet 3.

(a) Number of modules that are significantly correlated with cancer stage in each module category (Y-axis; module categories indicated) across different TCGA cohorts (labels on top).

(b-c) Scatterplots (left panel) show comparisons between Pathway Deregulation Scores (PDS; X-axis) and Module Perturbation Scores (MPS; Y-axis) for 342 samples obtained from patients with colonic neoplasms (GEO accession GSE41258)^{8,9} for modules associated with (b) regulation of apoptosis (KEGG database; MSigDB M8492¹⁰) and (c) oxidative phosphorylation (KEGG database; MSigDB M19540¹⁰). Each dot corresponds to a patient and colors indicate tissue-types (key indicated) and the correlation coefficient between PDS and MPS of these two modules are shown. Significant correlation coefficients between the two measures suggest that that MPS of these modules also capture disease progression just like their PDS, as reported originally by Drier and colleagues^{8,11}. We compared the distributions of MPS between the biopsied tissue-types on the independent cohort and the statistics of these comparisons are shown as a heatmap (right panel; $-\log_{10}$ of two-sided Mann-Whitney test p -values shown).

(d) Overlap between apoptosis pathway and module with highest similarity to it (based on hypergeometric test) is shown as a Venn diagram. Statistics of the overlap (hypergeometric test) and some of the prominent overlapping genes are indicated. The module associated with GO term mitochondrion (GO:0005739; 1980 genes) was one of the progression-associated modules discovered in TCGA's colon cancer cohort (panel (a)).

(e) Overlap between oxidative phosphorylation pathway and module with highest similarity to it (based on hypergeometric test) is shown as a Venn diagram. Statistics of the overlap (hypergeometric test) and some of the prominent overlapping genes are indicated. The module associated with electron transport respiratory chain (Reactome R-HSA-163200; 151 genes) was one of the progression-associated modules discovered to be associated with colon cancer stage in TCGA (panel (a)).

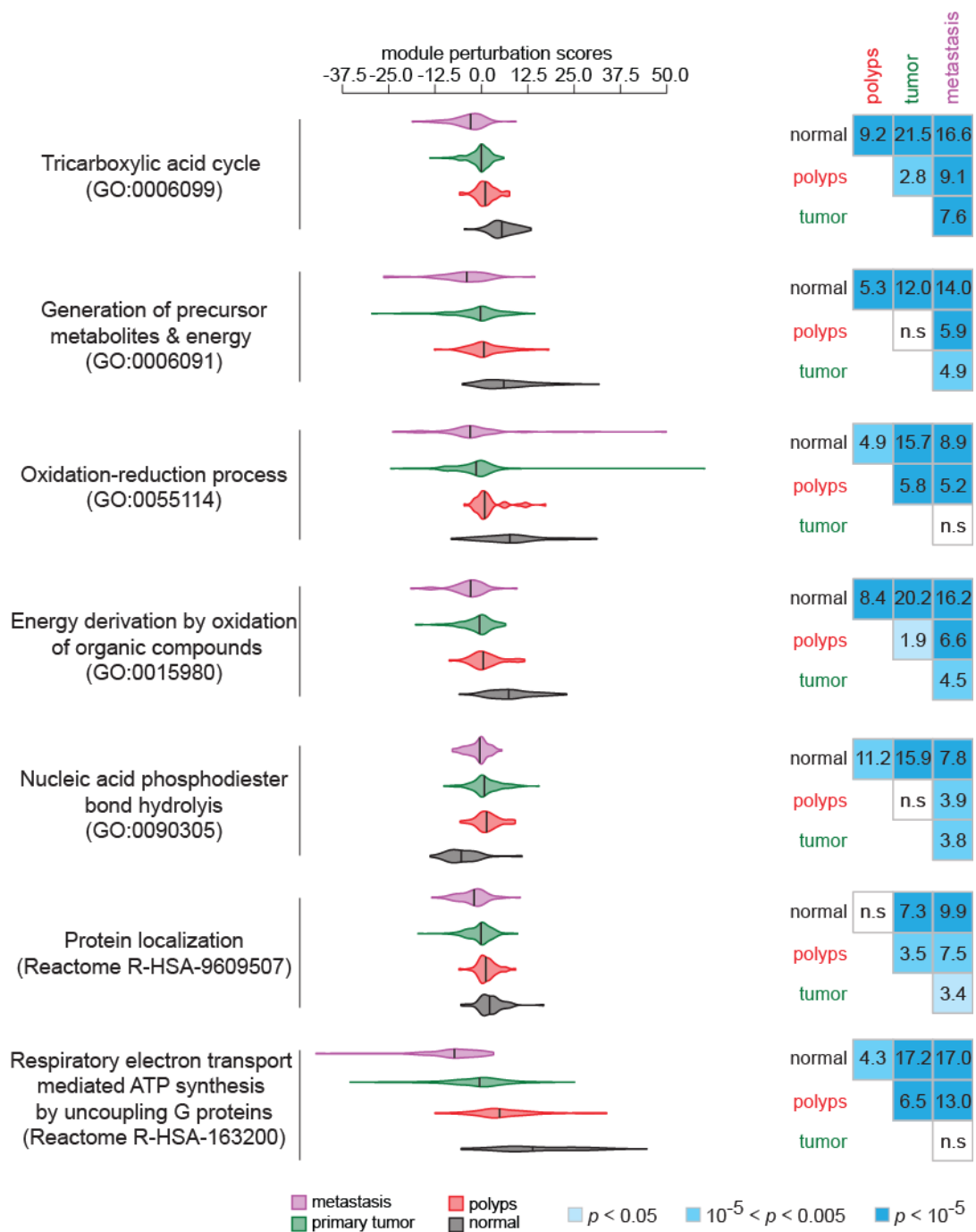


Figure S11. Modules discovered on TCGA's colon cancer cohort are correlated with disease progression on an independent colon cancer cohort, Related to STAR Methods, Data Sheet 3.

For progression-associated modules discovered on TCGA's colon cancer cohort, the distributions of perturbation scores (modules indicated) in metastatic, primary tumor, polyps and normal tissues biopsied from patients with colonic neoplasms (GEO accession GSE41258)^{8,9} are shown as violin plots (left panel). For each module, heatmap (right panel) shows statistics of the comparisons between MPS distributions in the biopsied tissue-types (right panel; two-sided Mann-Whitney test p -values shown).

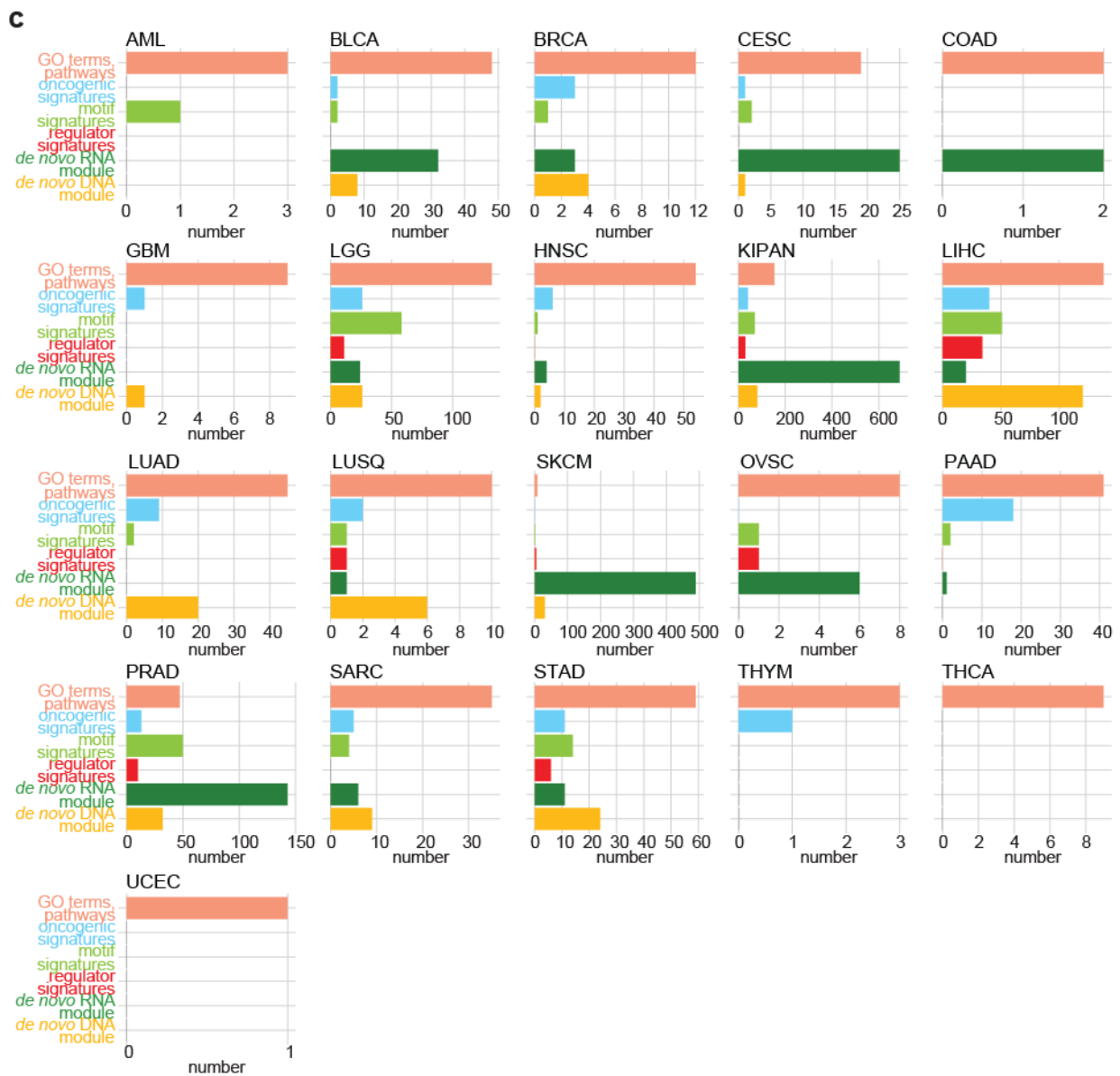
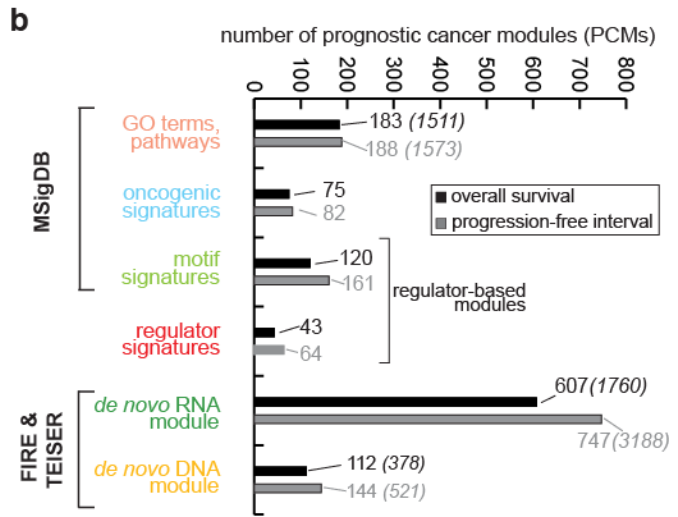
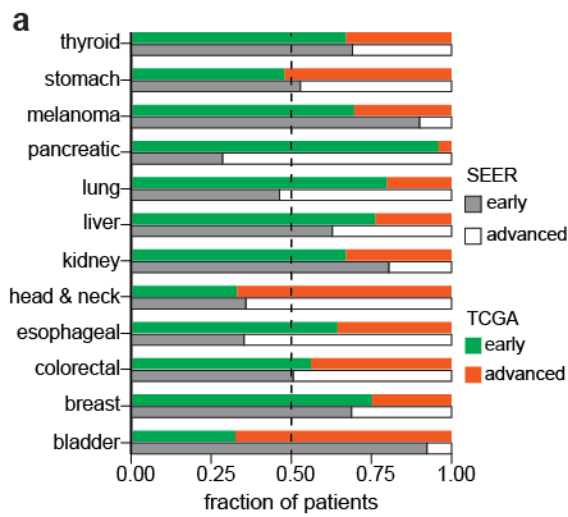


Figure S12. Numbers of prognostic cancer modules across different module categories for different cancers, Related to STAR Methods, Data Sheet 4.

(a) Fraction of patients with early or advanced stage cancer at time of diagnosis from the SEER database and in tissue-matched TCGA cohorts (Methods). Cancers are shown in rows. **(b)** Number of modules in each module category that specify significant prognostic value for overall survival (black bars) and progression-free interval survival (gray bars). For GO terms and pathways, *de novo* discovered DNA and RNA modules, numbers refer to non-redundant modules (as defined through Affinity Propagation¹²) and the numbers in parentheses, the full set of redundant modules (see Methods). **(c)** Number of prognostic cancer modules (both overall survival and progression-free interval survival) in each module category (Y-axis; module categories indicated) across different TCGA cohorts (labels on top).

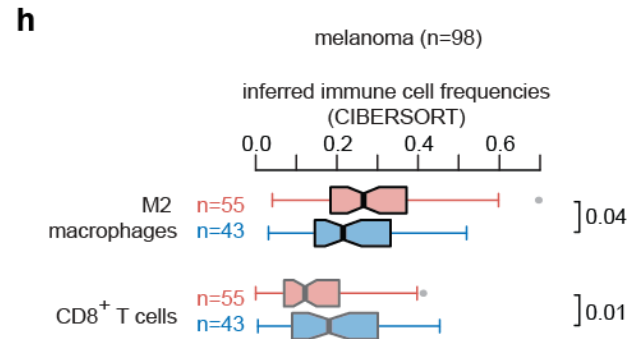
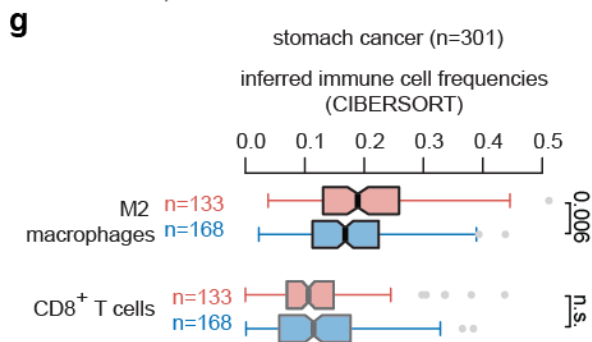
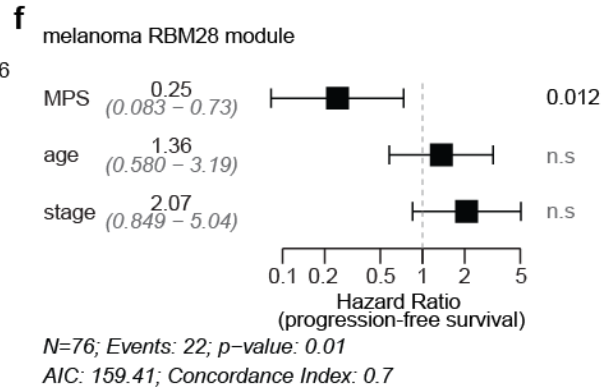
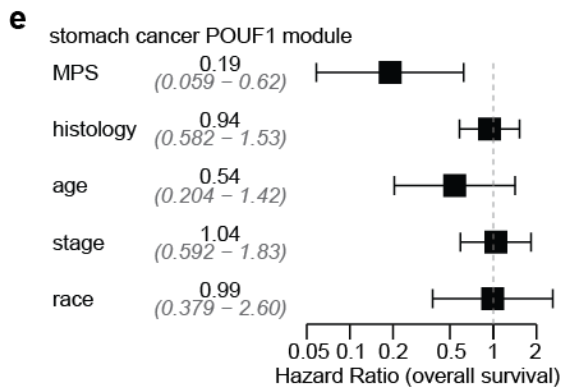
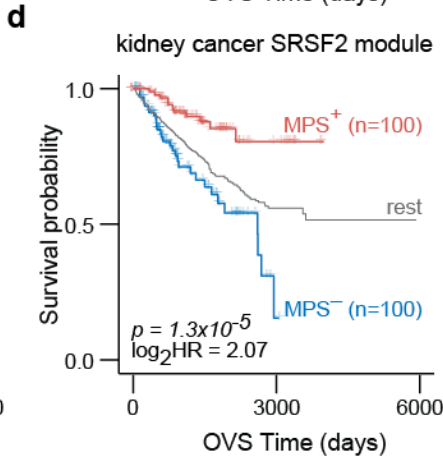
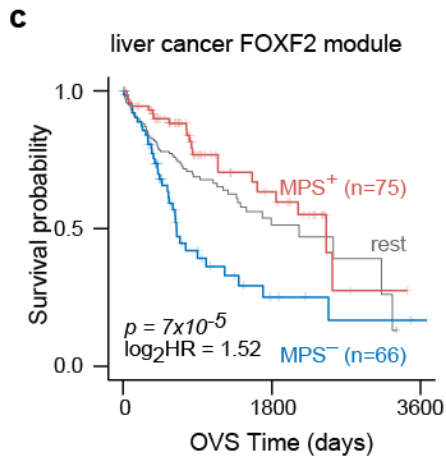
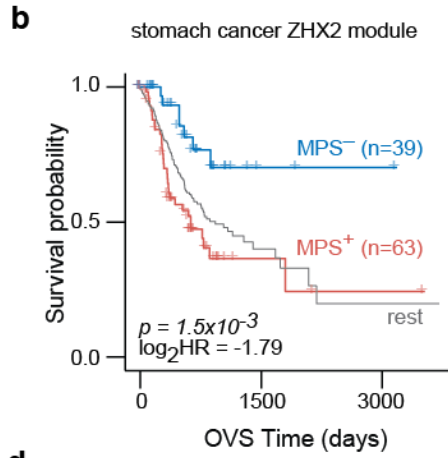
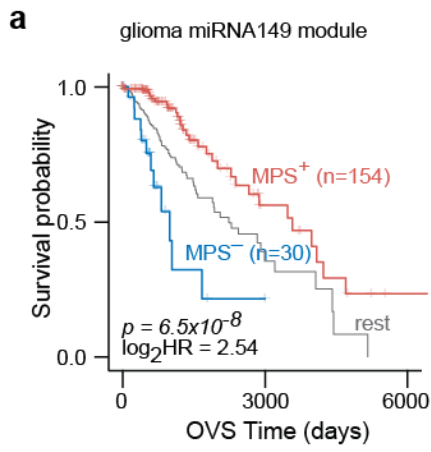
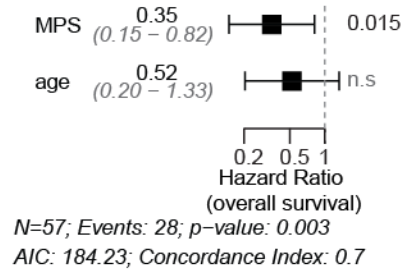
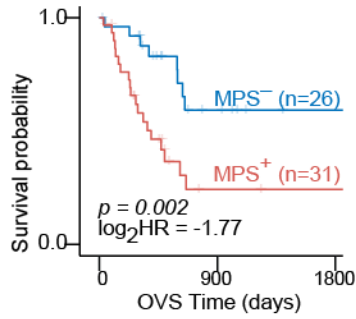


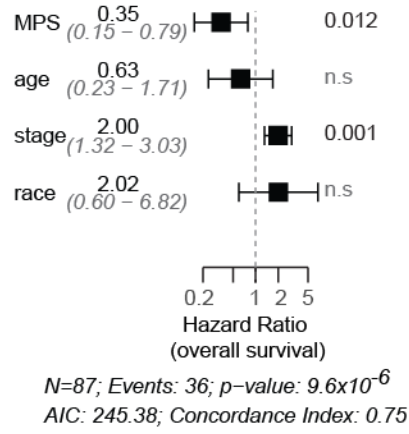
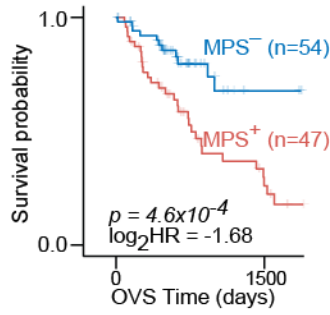
Figure S13. Patient survival stratification using previously defined modules associated with transcription factors, miRNAs and RNA binding proteins, Related to Figure 2, Data Sheet 4.

(a) Glioma patients with significant module activation (MPS⁺; red) for transcripts carrying at least one instance of the motif GAGCCAG in their 3'UTRs (miRNA-149 module; MSigDB² M2014; 149 genes) showed better overall survival (OVS) prognosis compared to samples with significant module repression (MPS⁻; blue). (b) Stomach cancer patients with significant module activation (MPS⁺; red) for genes carrying at least one instance of ZHX2 binding sites in their promoters (MSigDB² M14351; 271 genes) showed worse overall survival (OVS) prognosis compared to samples with significant module repression (MPS⁻; blue). (c) Liver cancer patients with significant module activation (MPS⁺; red) for genes carrying at least one instance of FOXF2 binding sites in their promoters (MSigDB² M3746; 939 genes) showed better overall survival (OVS) prognosis compared to samples with significant module repression (MPS⁻; blue). (d) Kidney cancer patients with significant module activation (MPS⁺; red) for genes carrying at least one instance of SRSF2 binding sites (CISBP-RNA⁴; 671 genes) showed better overall survival (OVS) prognosis compared to samples with significant module repression (MPS⁻; blue). (e) Forest plot for stomach cancer patients with a module corresponding to genes carrying at least one instance of the binding site for POU1F1. Shown are results from multivariate Cox analysis incorporating module perturbation scores (MPS), histological types, age, stage and race. For overall survival prognosis, hazard ratios (horizontal axis) with 95% confidence intervals and *p*-values are shown for each variable. (f) Forest plot for melanoma patients with a module corresponding to genes carrying at least one instance of RBM28 binding sites in their 3'UTRs. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), age and stage. For progression-free interval survival prognosis, hazard ratios (horizontal axis) with 95% confidence intervals and *p*-values are shown for each variable. (g) Boxplots show the distributions of inferred frequencies of M2 Macrophages and CD8⁺ T cells (CIBERSORT¹³) in stomach cancer patients with significant activation (red) or repression (blue) of the module associated with POU1F1. One-sided Mann-Whitney test *p*-values are indicated. (h) Boxplots show the distributions of inferred frequencies of M2 Macrophages and CD8⁺ T cells (CIBERSORT¹³) in melanoma patients with significant activation (red) or repression (blue) of the module associated with RBM28. One-sided Mann-Whitney test *p*-values are indicated. For all KM plot comparisons, statistics (median survival times, log₂ hazard ratio and *p*-value) are indicated and survival of the rest of the samples is shown in gray.

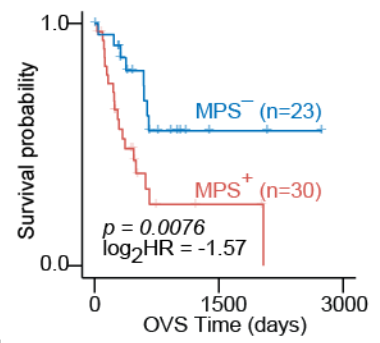
a pancreatic cancer RUNX2 module



b lung adenocarcinoma RUNX2 module



c pancreatic cancer RUNX2 module (only stage I/IA/IB, IIA/IIB)



d lung adenocarcinoma RUNX2 module (only stage I/IA/IB)

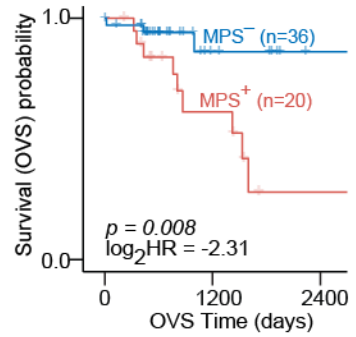
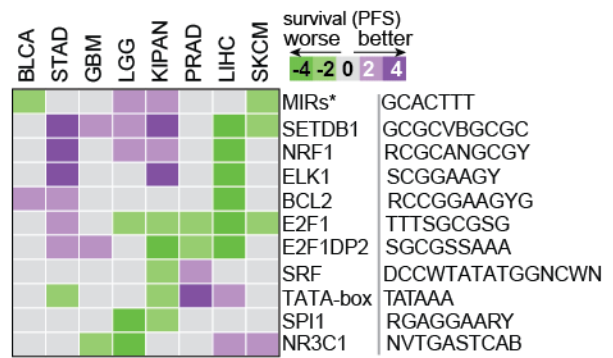
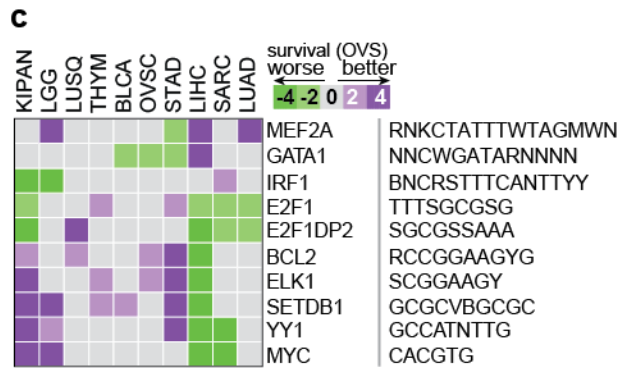
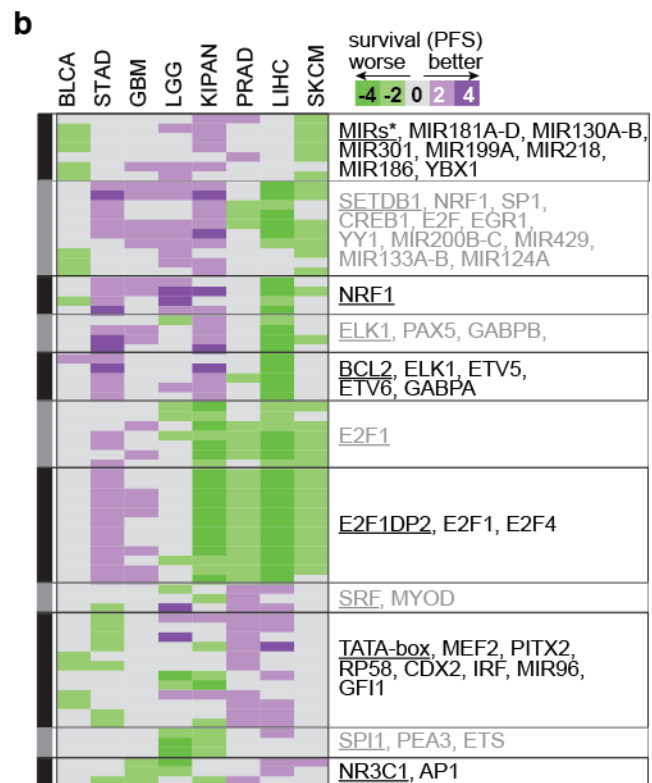


Figure S14. RUNX2 associated module conveys clinical significance in lung and pancreatic cancers, Related to Figure 2.

(a) Patients with significant module activation (MPS⁺; red) for genes that are transcriptional targets of RUNX2 (MSigDB² M27785; 125 genes) have worse overall survival (OVS) than patients with significant module repression (MPS⁻; blue) in pancreatic cancer patients (top panel). Forest plot shows results from multivariate cox analysis incorporating module perturbation scores (MPS) and age (bottom panel). (b) Patients with significant module activation (MPS⁺; red) for genes that are transcriptional targets of RUNX2 (MSigDB² M27785; 125 genes) have worse overall survival (OVS) than patients with significant module repression (MPS⁻; blue) in lung adenocarcinoma patients (top panel). Forest plot shows results from multivariate cox analysis incorporating module perturbation scores (MPS) and age (bottom panel). For visualization, time axes of KM curves are trimmed when the percentage of samples in MPS⁺ or MPS⁻ groups fall below 5%. In forest plots, hazard ratios (horizontal axis) with 95% confidence intervals and *p*-values are shown for each variable. (c-d) For the module associated with RUNX2, MPS-based patient stratification in specified sub-cohorts to only include early-stage tumors in (c) pancreatic cancer and, (d) lung adenocarcinoma patients. Module activation (MPS⁺; red) was associated with worse overall survival outcomes than module repression (MPS⁻; blue). Statistics of the comparisons are indicated. For all KM plot comparisons, statistics (median survival times, log₂ hazard ratio and *p*-value) are indicated.



* MIR175P, MIR20A, MIR106A, MIR106B, MIR20B, MIR519D

Figure S15. Identifying regulator-based modules conveying clinical significance in multiple cancers, Related to Figure 2.

(a-b) Standardized significance of MPS based patient survival (Wald statistic) of recurrently prognostic regulator-based modules (rows) shown across cancer cohorts (columns) for (a) overall survival (OVS) and (b) progression-free interval survival (PFS). Positive (or negative) values indicate worse (or better) survival of patients with significant module activation compared to repression. For each module in every cohort, samples were stratified into MPS⁺ and MPS⁻ groups to quantify survival differences between patients in the two groups. Stratifications that result in significant survival differences (KM test p -value < 0.05) are retained. Recurrently prognostic regulator-based PCMs are clustered using Affinity Propagation¹² to identify groups of modules that share substantial overlap between them. Similarities between modules were quantified using a modified Jaccard score (Methods). Rows are grouped together based on the clusters identified (indicated) and regulators associated with the exemplar module within each cluster are indicated. (c) Standardized significance of MPS based patient survival (Wald statistic) of exemplar regulator-associated modules (rows; motif IUPAC symbols indicated) specifying significant overall survival (OVS; top panel) and progression-free interval survival (PFS; bottom panel) across TCGA cohorts (columns).

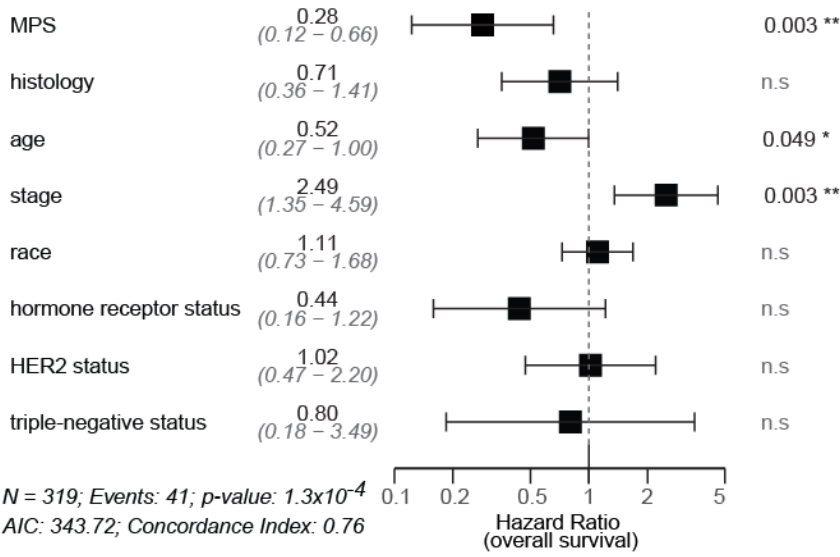
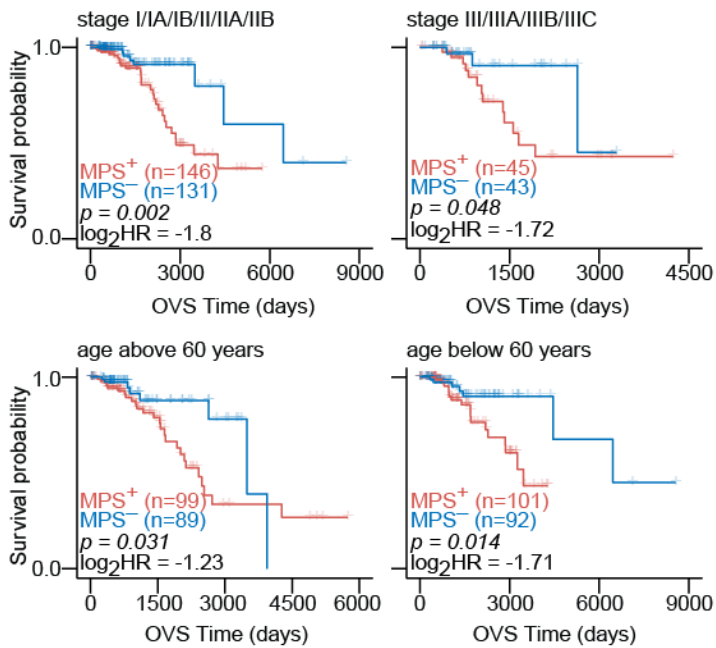
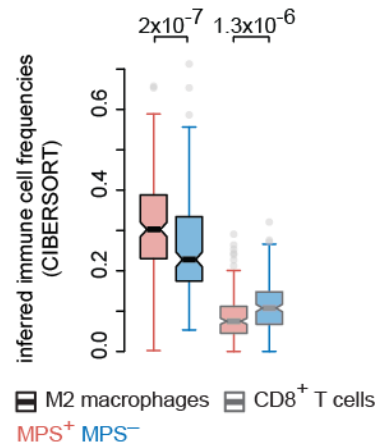

abreast cancer
(HWRTNACGH DNA) **b****c**

Figure S16. A *de novo* discovered DNA-based *cis*-regulatory module specifies significant prognosis in breast cancer even after controlling for confounds, Related to Figure 3.

(a) Forest plot for breast cancer patients with a *de novo* discovered, *cis*-regulatory prognostic cancer module corresponding to transcripts carrying at least one instance of the DNA motif HWRTNACGH (logo shown) within the first 1KB of their promoters. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), histological type, age, stage, race, hormone receptor status, HER2 status and triple-negative status. (b) MPS-based patient stratification for this PCM (previous panel) in specified sub-cohorts of the breast cancer patients including stage I and II patients (top; left panel), stage III patients (top; right panel), patients whose age is above 60 years at time of diagnosis (bottom; left panel) and patients whose age is less than 60 years at time of diagnosis (bottom; right panel). Module activation (MPS⁺; red) was associated with worse overall survival outcomes than module repression (MPS⁻; blue). (c) Boxplots show the inferred frequencies of M2 Macrophages and CD8⁺ T cells (CIBERSORT¹³) in breast cancer patients with significant activation (MPS⁺; red) or repression (MPS⁻; blue) of the module associated with the DNA motif HWRTNACGH (module in (a-b)). One-sided Mann-Whitney test *p*-values are indicated.

Figure S17. A *de novo* discovered DNA-based *cis*-regulatory module specifies significant prognosis in prostate cancer even after controlling for confounds, Related to Figure 3.

(a) Forest plot for prostate cancer patients with a module corresponding to transcripts carrying at least one instance of the DNA motif DTTTMCAM (logo shown) within the first 1KB of their promoters. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), age, race, Gleason score and prostate specific antigen (PSA) levels. (b) MPS-based patient stratification for this PCM (previous panel) in specified sub-cohorts of the prostate cancer patients including patients whose tumors have Gleason score less than 8 (left panel) and Gleason score that is 8 and above (right panel). Module activation (MPS⁺; red) was associated with better overall survival outcomes than module repression (MPS⁻; blue). (c) Boxplots show the inferred frequencies of M2 Macrophages and CD8⁺ T cells (CIBERSORT¹³) in prostate cancer patients with significant activation (MPS⁺; red) or repression (MPS⁻; blue) of the module associated with the DNA motif DTTTMCAM (module in (a-b)). One-sided Mann-Whitney test *p*-values are indicated.

a stomach cancer
(WSUUMCAMP RNA) 

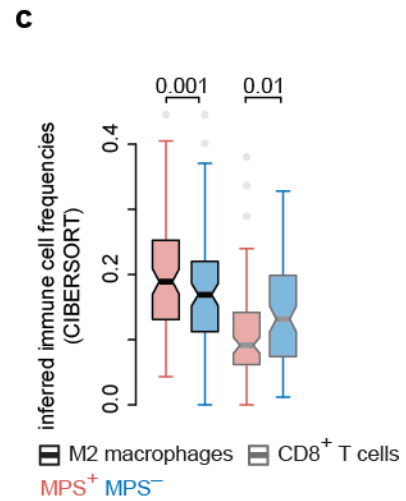
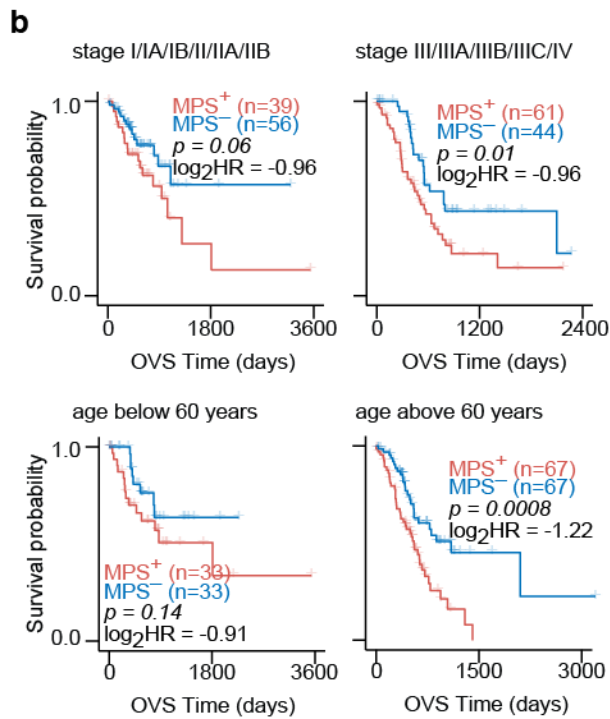
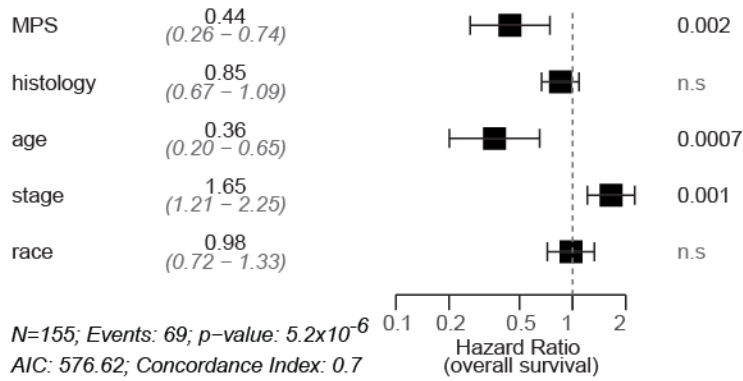


Figure S18. A *de novo* discovered linear RNA-based *cis*-regulatory module specifies significant prognosis in stomach cancer even after controlling for confounds, Related to Figure 3.

(a) Forest plot for stomach cancer patients with a module corresponding to transcripts carrying at least one instance of the linear RNA motif WSUUCAMR (logo shown) within the first 1KB of their 3'UTRs. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), histological types, age, stage and race. (b) MPS-based patient stratification for this PCM (previous panel) in specified sub-cohorts of the stomach cancer patients including patients whose tumors are stage I or stage II (top; left panel), stage III or stage IV (top; right panel), patients whose age at diagnosis is below 60 years (bottom; left panel) or above 60 years (bottom; right panel). Module activation (MPS⁺; red) was associated with worse overall survival outcomes than module repression (MPS⁻; blue). (c) Boxplots show the inferred frequencies of M2 Macrophages and CD8⁺ T cells (CIBERSORT¹³) in stomach cancer patients with significant activation (MPS⁺; red) or repression (MPS⁻; blue) of the module associated with the RNA motif WSUUCAMR (module in (a-b)). One-sided Mann-Whitney test *p*-values are indicated.

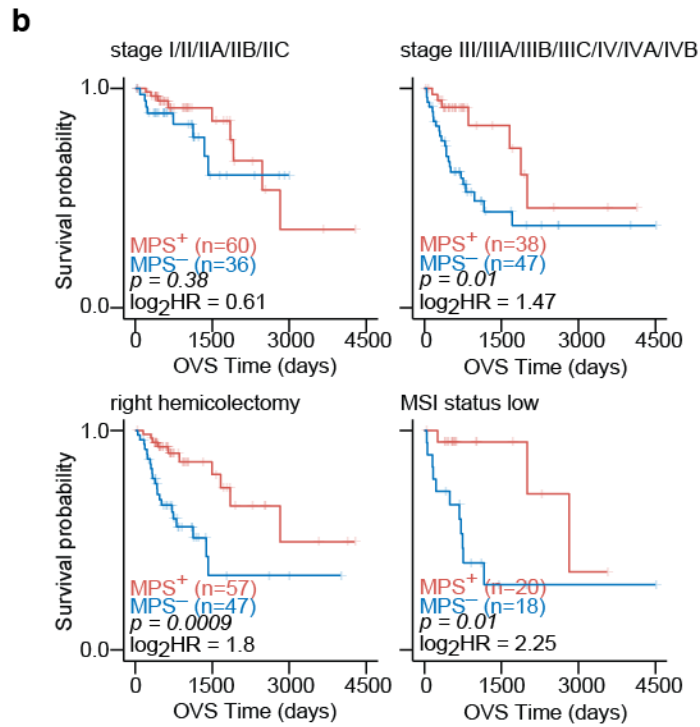
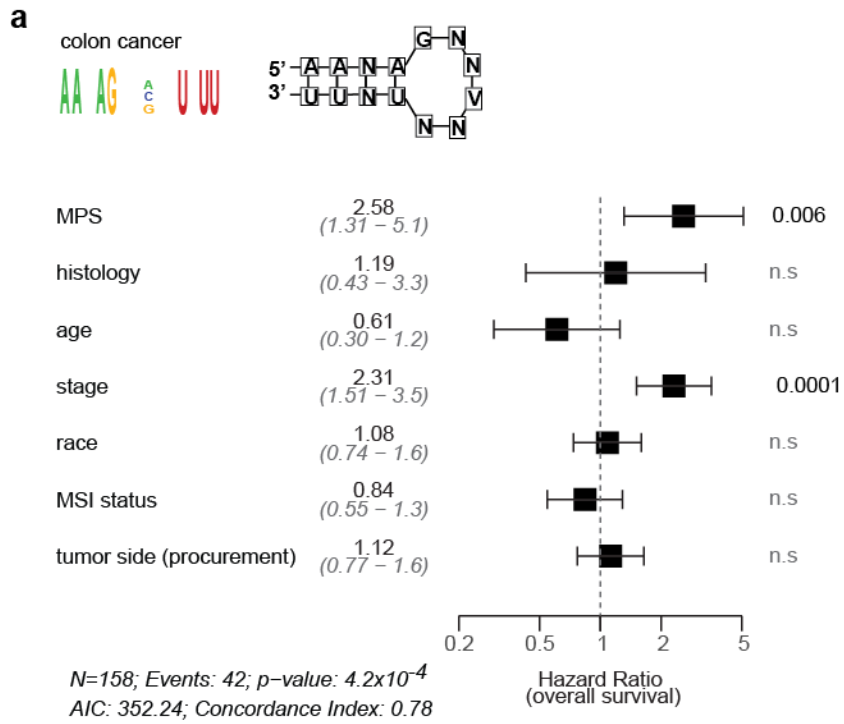


Figure S19. A *de novo* discovered structural RNA-based *cis*-regulatory module specifies significant prognosis in colon cancer even after controlling for confounds, Related to Figure 3.

(a) Forest plot for colon cancer patients with a module corresponding to transcripts carrying at least one instance of the structural RNA motif (logo and putative structure shown) within the first 1KB of their 3'UTRs. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), histological types, age, stage, race, micro-satellite instability (MSI) status and side of colectomy. (b) MPS-based patient stratification for this PCM (previous panel) in specified sub-cohorts of the stomach cancer patients including stage I or stage II tumors (top; left panel), stage III or stage IV tumors (top; right panel), patients who had a right hemicolectomy (bottom; left panel) and tumors characterized to be MSI low (bottom; right panel). Module activation (MPS⁺; red) was associated with better overall survival outcomes than module repression (MPS⁻; blue). In forest plots, hazard ratios (horizontal axis) with 95% confidence intervals and *p*-values are shown for each variable and in KM plots, statistics of survival comparisons are indicated.

De novo discovered cis-regulatory gene modules DNA, linear RNA, structural RNA

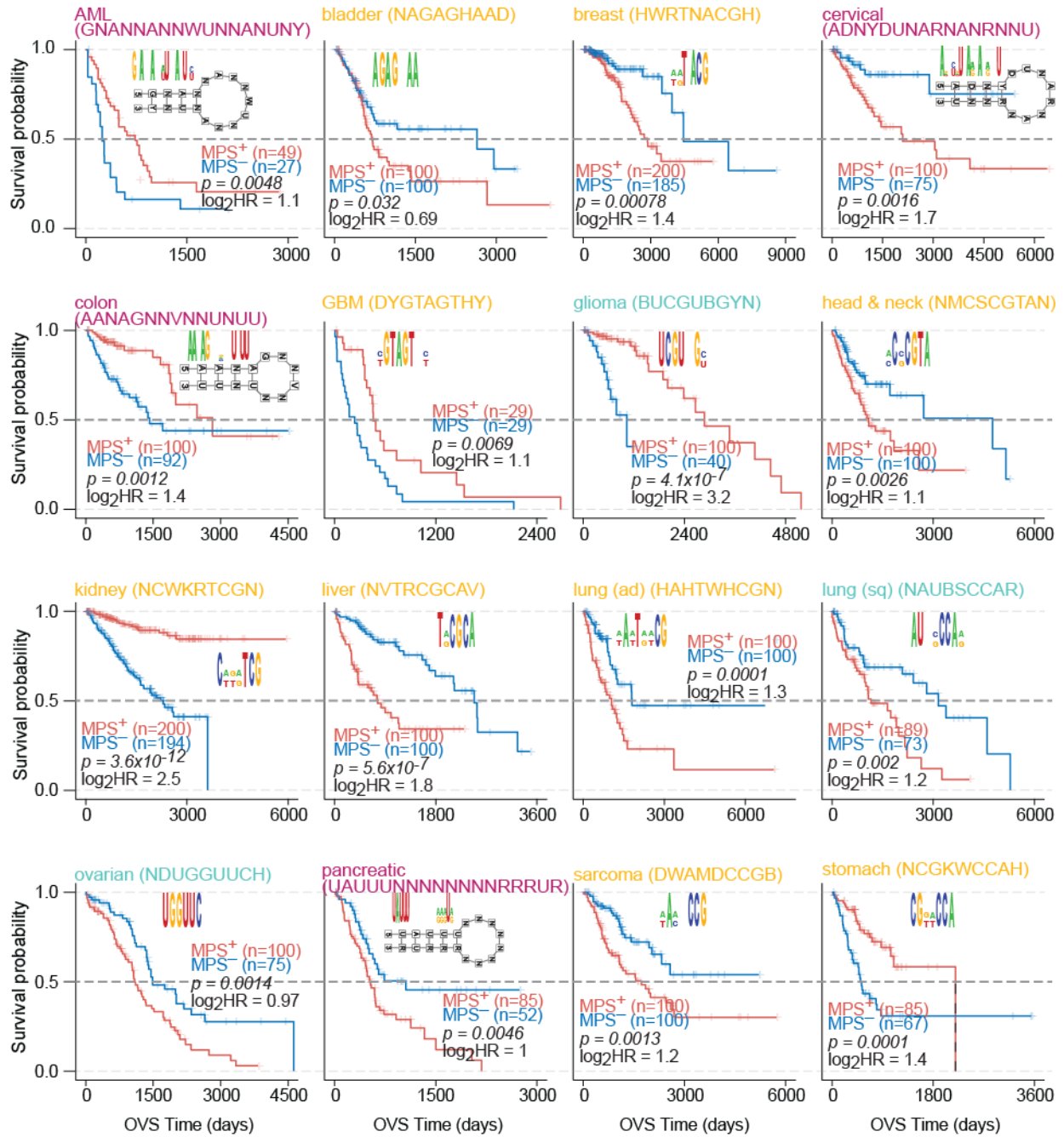


Figure S20. Subset of highly significant *de novo* discovered PCMs for overall survival, Related to STAR Methods, Data Sheet 4.

Kaplan-Meier (KM) plots showcase highly significant *de novo* discovered *cis*-regulatory PCMs for overall survival (OVS) in individual cohorts (labels shown on top). The colors of labels correspond to the categories of the modules- DNA (yellow), linear RNA (green) and structural RNA (purple). KM plots show fraction of surviving patients (Y-axis) and time in months (X-axis) as well as statistics of the comparisons between patients in whom the module is significantly activated (MPS⁺) and repressed (MPS⁻). The modules were selected such that the associated motifs have no similarities (analyses using TOMOTOM¹⁴) to previously known transcription factor or RNA binding protein binding sequences or micro-RNA seeds (Methods).

De novo discovered *cis*-regulatory gene modules DNA, linear RNA, structural RNA

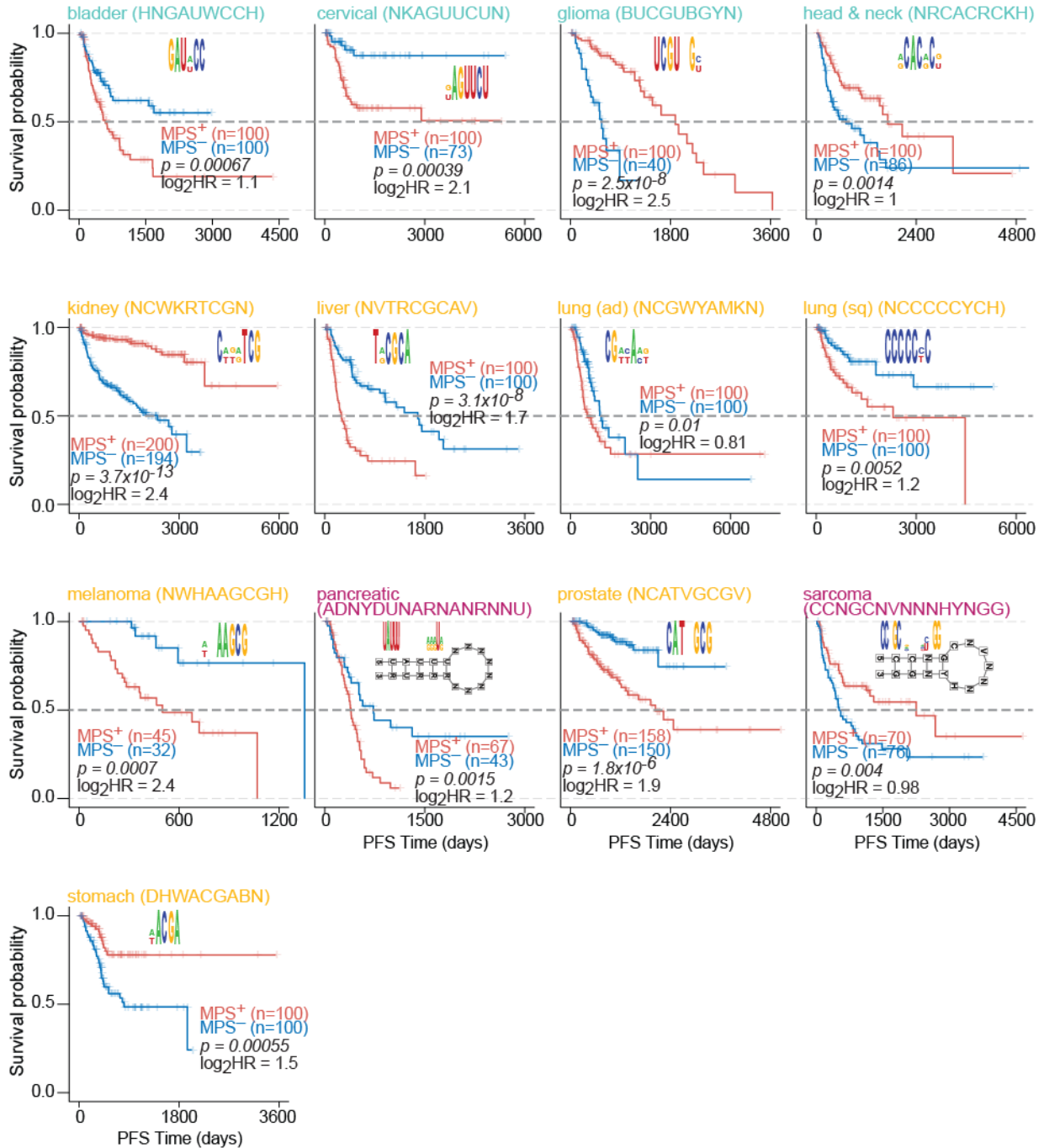
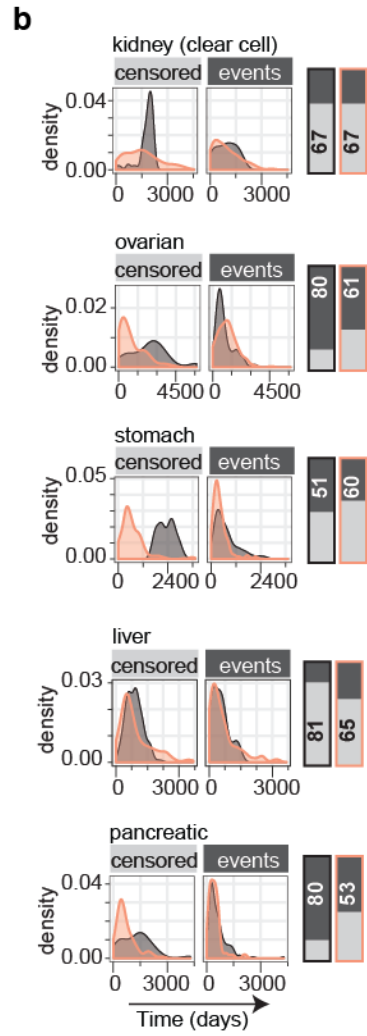
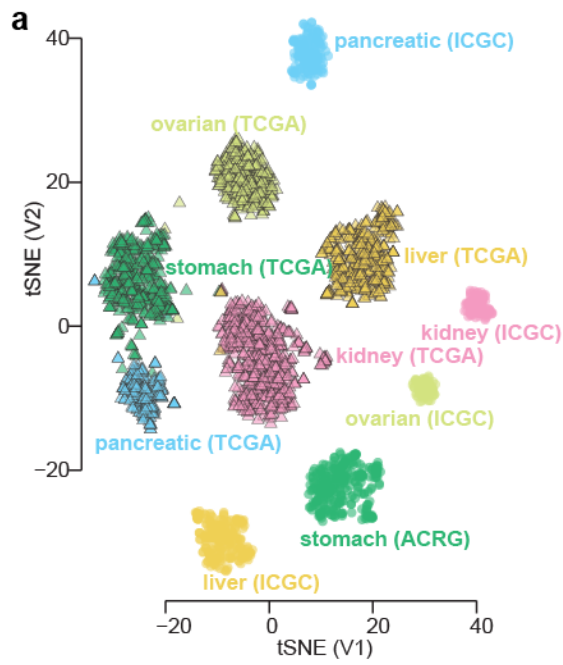


Figure S21. Subset of highly significant *de novo* discovered PCMs for progression-free interval survival, Related to STAR Methods, Data Sheet 4.

Kaplan-Meier (KM) plots showcase highly significant *de novo* discovered *cis*-regulatory PCMs for progression-free interval survival (PFS) in individual cohorts (labels shown on top). The colors of labels correspond to the categories of the modules- DNA (yellow), linear RNA (green) and structural RNA (purple). KM plots show fraction of surviving patients (Y-axis) and time in months (X-axis) as well as statistics of the comparisons between patients in whom the module is significantly activated (MPS⁺) and repressed (MPS⁻). The modules were selected such that the associated motifs have no similarities (analyses using TOMOTOM¹⁴) to previously known transcription factor or RNA binding protein binding sequences or micro-RNA seeds (Methods).



■ external ■ TCGA

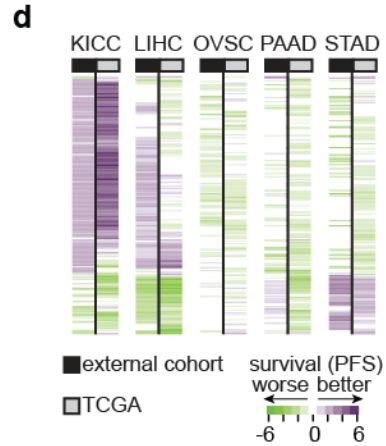
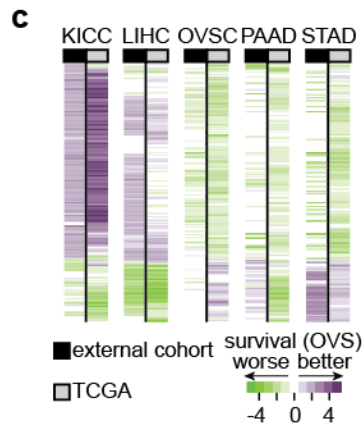


Figure S22. Conserved prognostic potential of module perturbations in external cohorts, Related to STAR Methods.

(a) Visualization of high-dimensional data using tSNE applied to transcriptome data (z-scored) of primary tumor samples from clear cell kidney cancer (TCGA; n=528 and ICGC-Europe; n=91), liver cancer (TCGA; n=360 and ICGC-Riken; n=223), ovarian cancer (TCGA; n=283 and ICGC-Australia; n=76), pancreatic cancer (TCGA; n=177 and ICGC-Canada; n=163) and stomach cancer (TCGA; n=383 and ACRG; n=300) cohorts. Each dot corresponds to a primary tumor sample from TCGA (triangles) or the external cohort (dots) from tissue-matched cohorts (color, labels shown). (b) Clinical characteristics of tissue-matched primary tumors analyzed through TCGA or independent external cohorts. Density plots show the distribution of follow-up times for patients that are censored (light gray; label on top) or event times for the remainder of the cohort (dark gray; label on top). Stacked barplots show percent of events (dark gray) and patients that are censored (light gray) in the external cohort (black) or in TCGA (orange) for liver, ovarian, stomach, clear cell kidney and pancreatic cancer cohorts. (c-d) Standardized significance of MPS based patient survival in external cohorts (from ICGC and ACRG) and TCGA cohorts (labels indicated on top) of clear cell kidney cancer (KICC), liver cancer (LIHC), ovarian cancer (OVSC), pancreatic cancer (PAAD) and stomach cancer (STAD) samples. Modules that specify recurrent prognostic value in the external cohorts are shown and the values (color key indicated) correspond to the Wald statistic computed using the Cox proportional-hazards model for (c) overall survival and (d) progression-free interval survival.

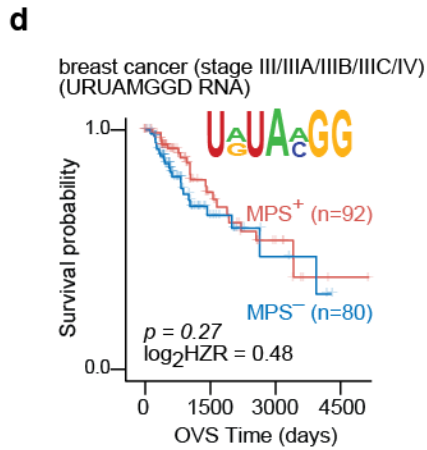
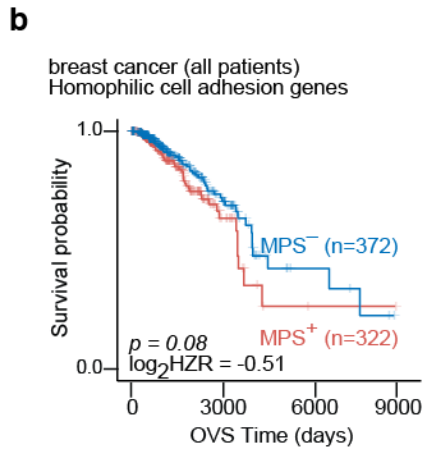
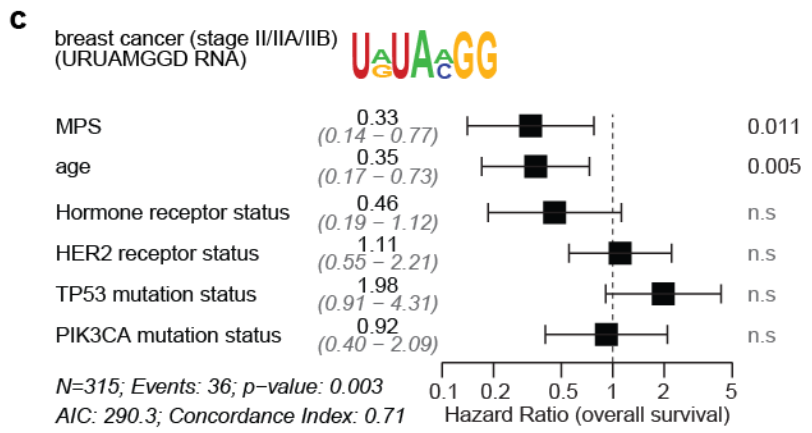
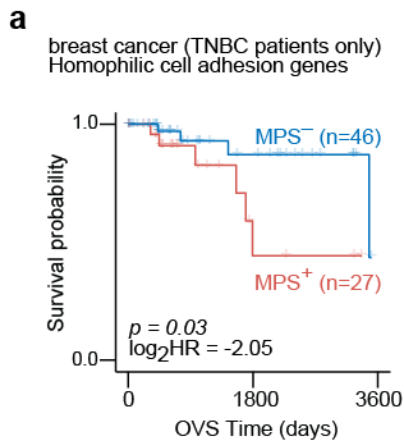


Figure S23. *A priori* incorporation of histological types and tumor stage identifies significant prognostic modules within the specified sub-cohorts, Related to Figure 4, Data Sheet 6.

(a) In triple negative breast cancer patients, significant activation of module (MPS⁺; red) corresponding to homophilic cell adhesion specified worse prognosis than patients with significant module repression (MPS⁻; blue). (b) On the full breast cancer cohort, the module corresponding to homophilic cell adhesion (MSigDB² Gene Ontology; 173 genes) did not stratify patients based on overall survival. (c) Forest plot for stage II/IIA/IIB breast cancer patients with a *de novo* discovered *cis*-regulatory gene module corresponding to transcripts carrying at least one instance of the linear RNA motif URUAMGGD (logo shown; 1082 genes) within the first 1KB of their 3'UTRs. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), patient age, tumor stage, PIK3CA and TP53 mutation status, hormone receptor status and HER2 status. For overall survival prognosis, hazard ratios (horizontal axis) with 95% confidence intervals and *p*-values are shown for each predictor. (d) For this module (previous panel), stratification of breast cancer patients with stage III/IIIA/IIIB/IIIC/IV cancers into MPS⁺ and MPS⁻ groups did not provide significantly different survival trajectories.

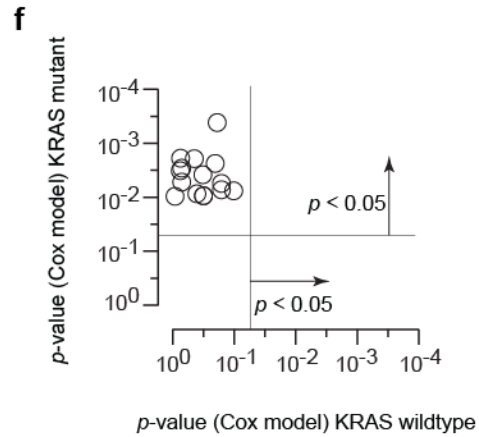
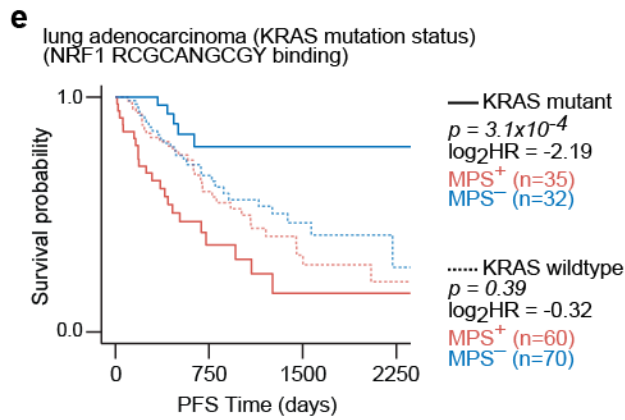
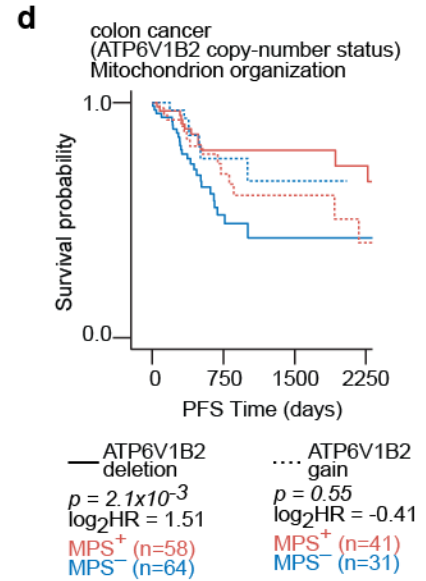
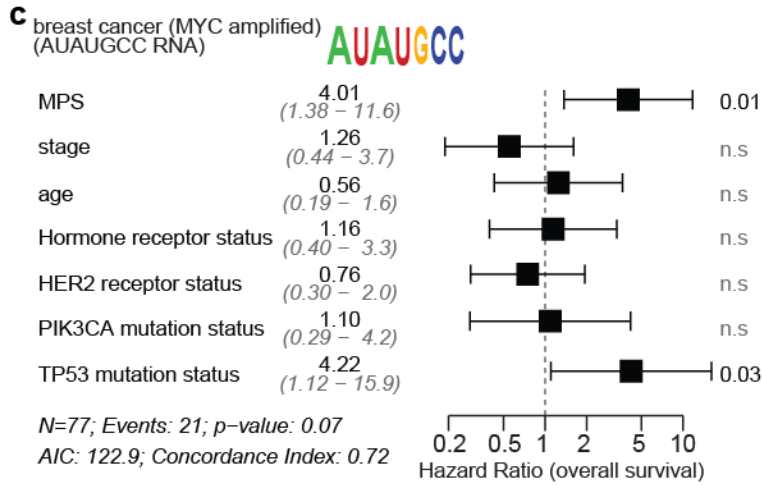
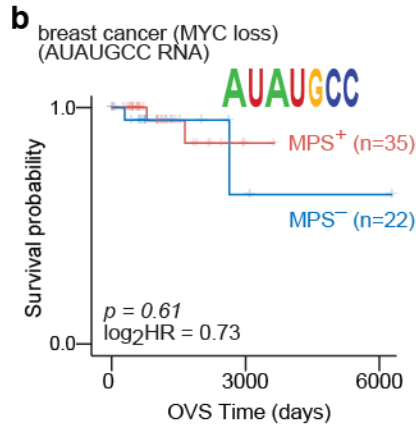
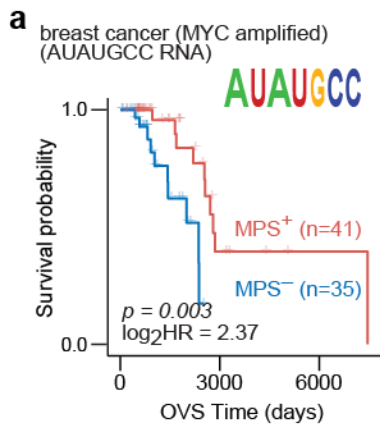


Figure S24. A priori incorporation of prominent genomic aberrations identifies significant prognostic modules within the specified sub-cohorts, Related to Figure 4, Data Sheet 6.

(a) Breast cancer patients with deep amplification of MYC (Methods) and with significant module activation (MPS⁺; red) for transcripts harboring at least one instance of the RNA motif AUAUGCC (logo shown; 529 genes) within the first 1KB of their 3'UTRs, showed better overall survival (OVS) prognosis than samples with significant module-repression (MPS⁻; blue). (b) For this module (previous panel), stratification of breast cancer patient tumors that harbor deletions or deep deletions at the MYC locus (Methods) into MPS⁺ and MPS⁻ groups did not provide significantly different survival trajectories. (c) Forest plot for breast cancer patients harboring deep amplification of the locus spanning MYC (Methods) with a *de novo* discovered *cis*-regulatory gene module corresponding to transcripts carrying at least one instance of the linear RNA motif AUAUGCC (logo shown) within the first 1KB of their 3'UTRs. Shown are results from multivariate cox analysis incorporating module perturbation scores (MPS), patient age, tumor stage, PIK3CA and TP53 mutation status, hormone receptor status and HER2 status. For overall survival prognosis, hazard ratios (horizontal axis) with 95% confidence intervals and *p*-values are shown for each predictor. (d) Colon cancer tumor samples harboring deletions at the ATP6V1B2 locus with significant activation of the module corresponding to genes annotated to be involved in mitochondrion organization (MSigDB²; 509 genes) showed better progression-free interval survival (PFS) than samples with significant repression. For this module, stratification of colon cancer patients harboring amplification or deep amplification of the ATP6V1B2 locus into MPS⁺ and MPS⁻ groups did not provide significantly different survival trajectories (dashed lines). (e) Lung adenocarcinoma patients with mutated KRAS and with significant module activation (MPS⁺; red) for genes harboring at least one instance of binding sites for NRF1 transcription factor (MSigDB² M9394; 958 genes) showed worse prognosis than patients with significant module repression (MPS⁻; blue). For this module, stratification of lung adenocarcinoma patients that are wildtype for KRAS into MPS⁺ and MPS⁻ groups did not provide significantly different survival trajectories (dashed lines). (f) Survival stratifications in lung adenocarcinoma patients harboring mutated or wildtype KRAS based on perturbation scores of modules identified to be significantly prognostic in patients whose tumors harbor mutated KRAS. Scatterplot shows comparisons between *p*-values (Cox proportional hazards model) in survival stratification (MPS⁺ vs. MPS⁻) of lung adenocarcinoma patients harboring wildtype (X-axis) and mutated (Y-axis) KRAS.

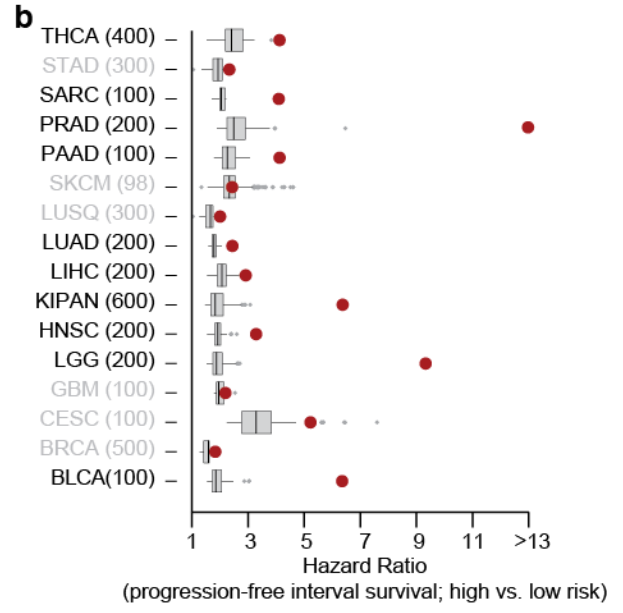
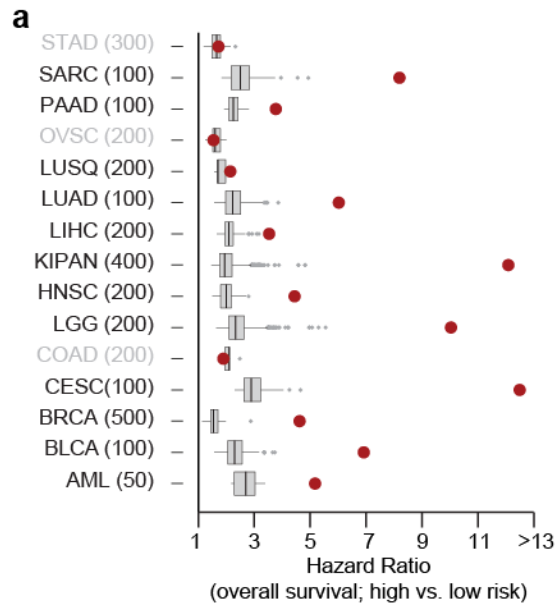


Figure S25. Predictions from ensemble learning models that utilize combinations of PCMs provide stronger patient stratification than individual PCMs, Related to Figure 5, STAR Methods.

(a-b) Hazard ratio distributions obtained by stratifying patients using the perturbation scores of prognostic cancer modules in individual cohorts. Red dots correspond to patient stratifications derived from random survival forest models¹⁵ trained on PCMs for **(a)** overall survival (OVS) or **(b)** progression-free interval survival (PFS) in individual cohorts (see Methods). Cohorts in which the random survival forest models provided stronger stratification compared to every individual PCM are in bold font and number of samples in each survival comparison is indicated.

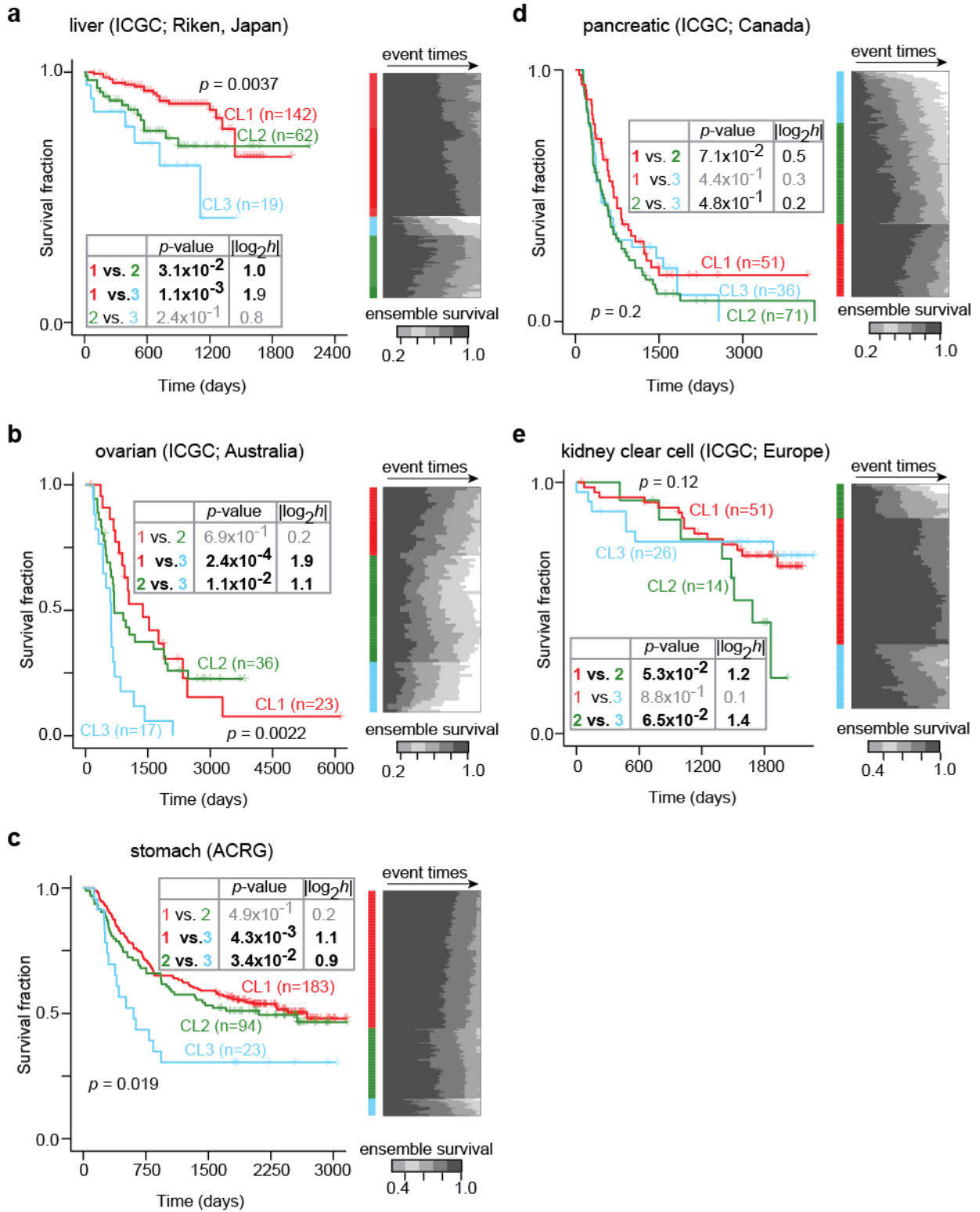


Figure S26. Predictions from ensemble learning models that utilize combinations of PCMs are conserved in tissue-matched external cohorts, Related to STAR Methods.

(a-e) Random survival forest models trained on TCGA data are used to make survival predictions on the full set of patients in tissue-matched external cohorts. The predicted ensemble survival probabilities of individual patients at event times (unique death times) are shown (heatmap; right panel). Rows (patients) are clustered using k-means clustering with Euclidean distance and the optimal cluster size (3; NbClust¹⁶) is chosen based on Krzanowski and Lai¹⁷, Calinski and Harbasz¹⁸, silhouette¹⁹ and gap statistic²⁰ metrics. Kaplan-Meier plots (left panel) show the survival trajectories of patients in these clusters (colors indicated). Across the cohorts, there was clear concordance between predicted survival probabilities and survival trajectories of patients in the cohorts. Statistics for survival comparison as well as pairwise survival comparisons are listed for **(a)** liver, **(b)** ovarian, **(c)** stomach, **(d)** pancreatic and **(e)** clear cell kidney cancer cohorts.

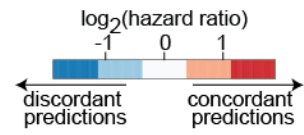
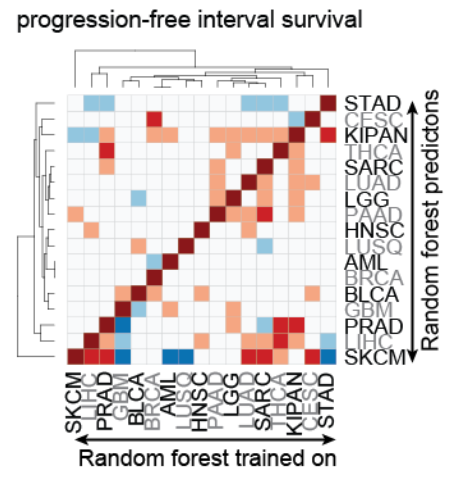
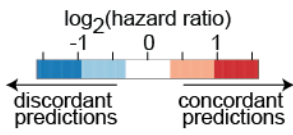
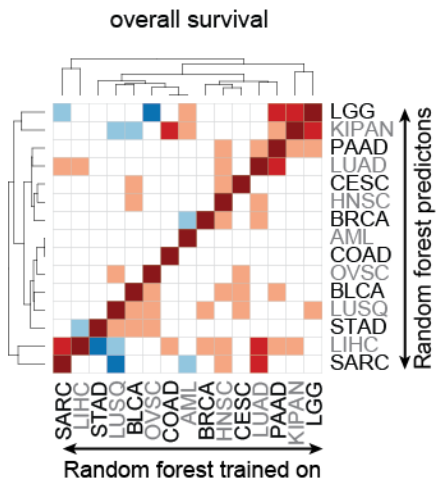
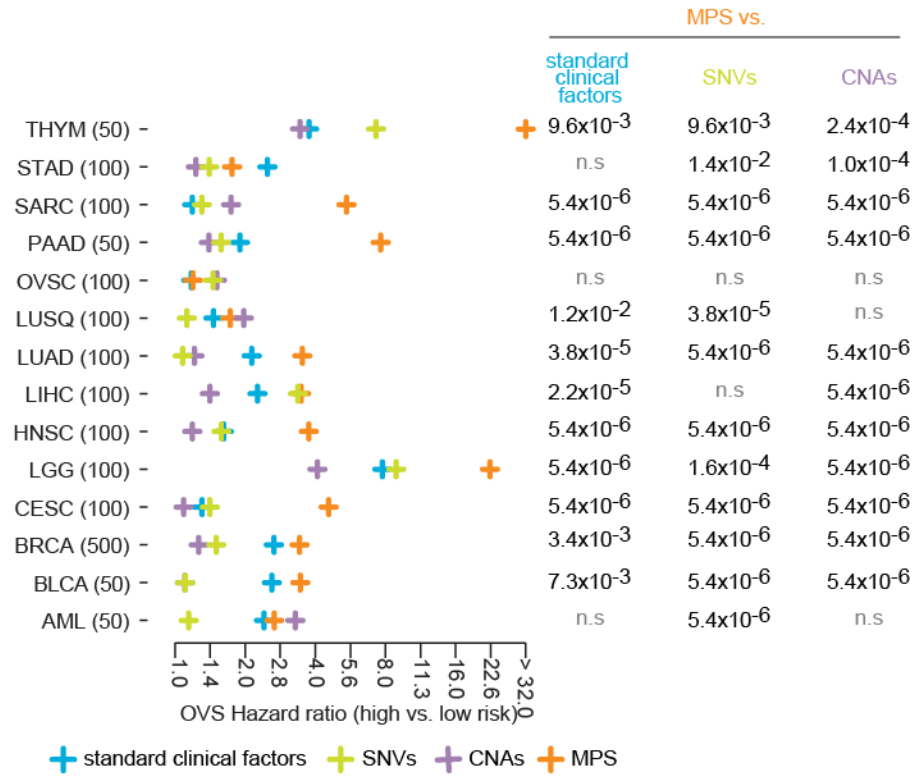


Figure S27. Predictions from ensemble learning models that utilize combinations of PCMs are conserved in some cancers despite diverse tissues-of-origin, Related to STAR Methods.

For all possible pairs of cohorts, the prognostic predictive value of random survival forest models trained on PCMs in one cohort (columns) was tested on every other cohort (rows). Survival comparisons in the test cohort were made based on predicted risk (top and bottom 40%). Colors correspond to hazard ratios and indicate if predictions in the test cohorts exhibit concordance (red) or discordance (blue) with training cohort. Diagonal entries of the heatmap (same training and test cohort and thus, concordant predictions) are in dark red.

a



b

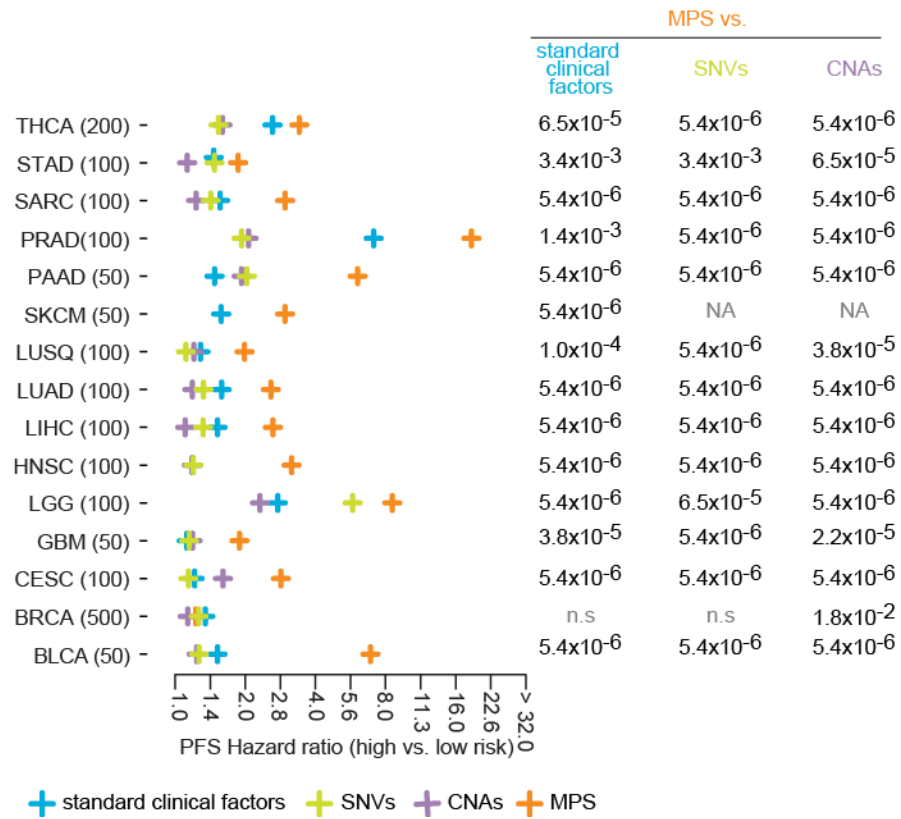


Figure S28. Random survival forest models trained on module perturbations provide more prognostic value than models trained on conventional clinical factors, Related to Figure 5, STAR Methods.

Performance of random survival forest models trained on standard clinical factors (blue), module perturbation scores (orange), recurrent single-nucleotide variants (SNV; green) and prominent copy-number aberrations (CNA; purple). For each cohort (rows), random forest models were trained in 10-fold cross-validation using prognostic cancer modules (PCMs) in each cohort for (a) overall survival (OVS) and (b) progression-free interval survival (PFS). Hazard ratios were obtained using Cox proportional hazards model on patients stratified into high and low risk groups based on predictions from the random survival forest models¹⁵. Number of patients used in each comparison is specified. Crosses correspond to the median hazard ratio calculated from 10 repetitions of the random survival forest predictions with randomized initializations. One-sided Mann-Whitney test *p*-values for comparisons of hazard ratios from models trained on MPS and each of standard clinical factors, SNVs and CNAs are also indicated for each cohort.

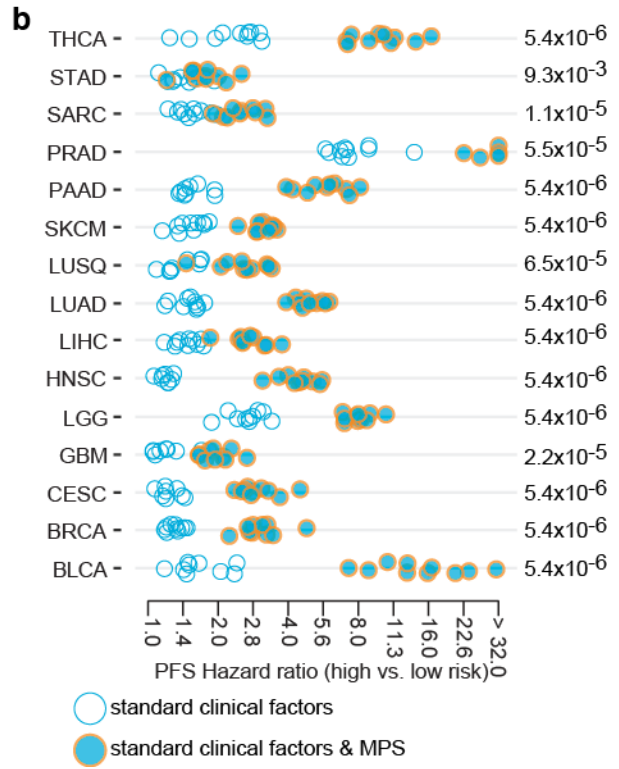
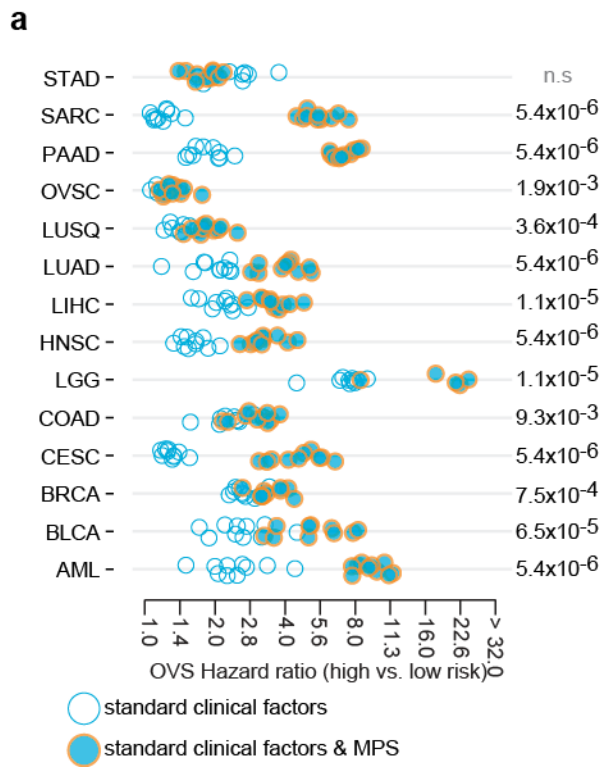


Figure S29. Random survival forest models trained on module perturbations provide additional prognostic value to models trained on conventional clinical factors, Related to STAR Methods.

(a-b) Performance of random survival forest models trained on only standard clinical factors (open blue circles) or standard clinical factors and perturbation scores of PCMs (blue dots with orange outline) for (a) overall survival (OVS) and, (b) progression-free interval survival (PFS). Each dot corresponds to the hazard ratio of a 10-fold cross-validated random survival forest model. For every cohort, distributions of hazard ratios obtained from 10 randomly initialized instances of the random survival forests are shown. Comparisons between hazard ratios obtained from the random survival forest models are shown (p -values from one-sided Mann-Whitney test).

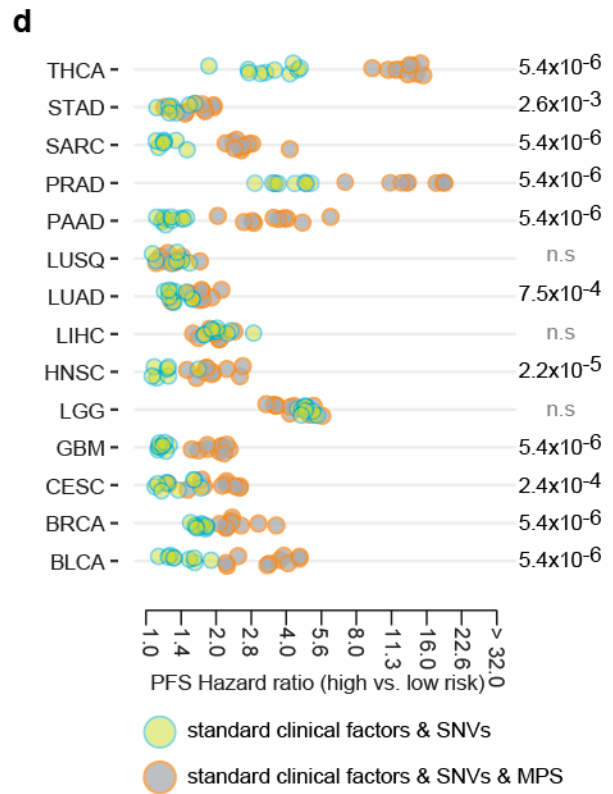
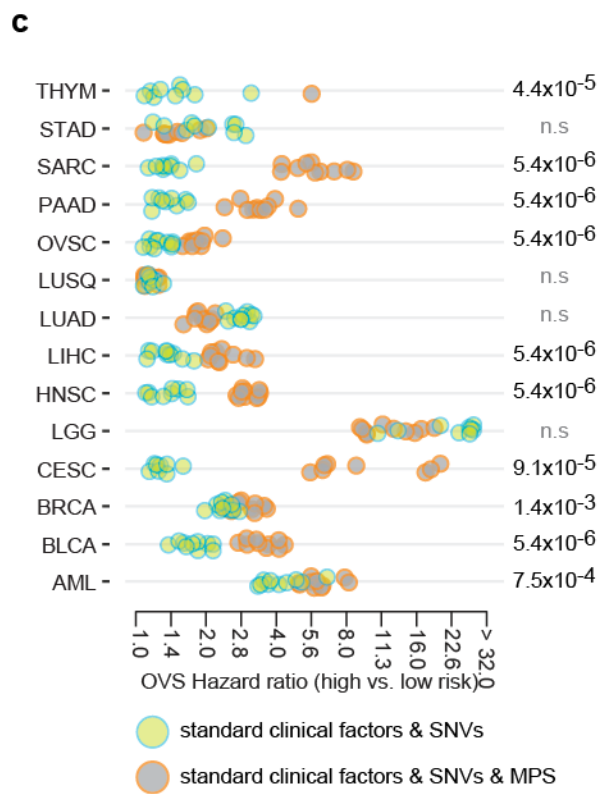
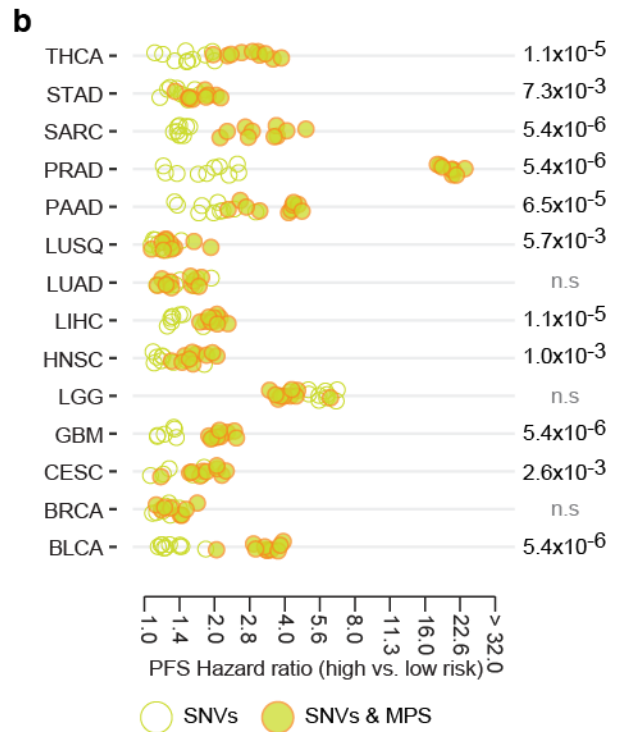
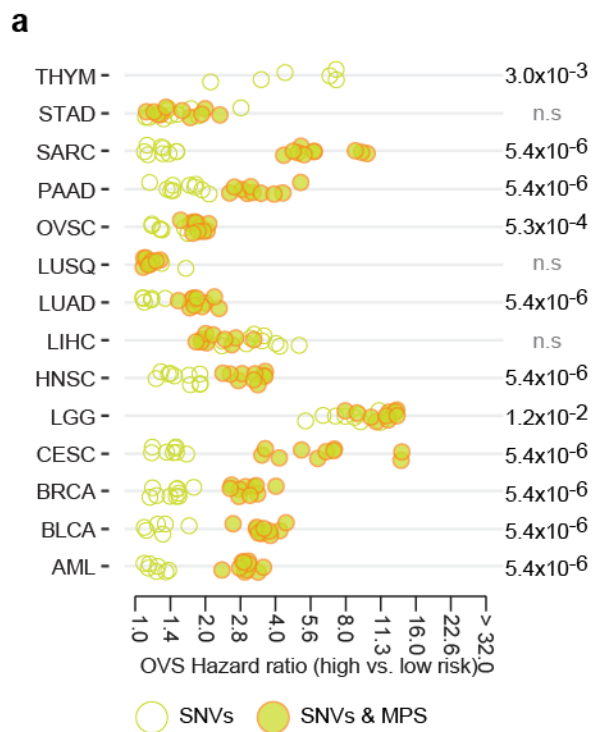


Figure S30. Random survival forests trained on prognostic modules provide significantly improved patient stratification compared to prominent SNVs and standard clinical factors, Related to Figure 5, STAR Methods.

(a-b) Performance of random survival forest models trained on prominent single-nucleotide variants (SNVs) (open green circles) or SNVs and module perturbation scores of PCMs in each cohort (green dots with orange outline) for (a) overall survival (OVS) and, (b) progression-free interval survival (PFS). (c-d) Performance of random survival forest models are shown for models trained on prominent SNVs and standard clinical factors with (gray dots with orange outline) or without (green dots with blue outline) module perturbation scores for (c) overall survival (OVS) and, (d) progression-free interval survival (PFS). Hazard ratios were obtained using Cox proportional hazards model on patients stratified into high and low risk groups based on predictions from the random survival forest models¹⁵. Each dot corresponds to the predicted hazard ratio of a 10-fold cross-validated random forest model. For every cohort, distributions of hazard ratios obtained from 10 randomly initialized instances are shown. Comparisons between hazard ratios obtained from the two sets of random survival forest models are shown (p -values from one-sided Mann-Whitney test).

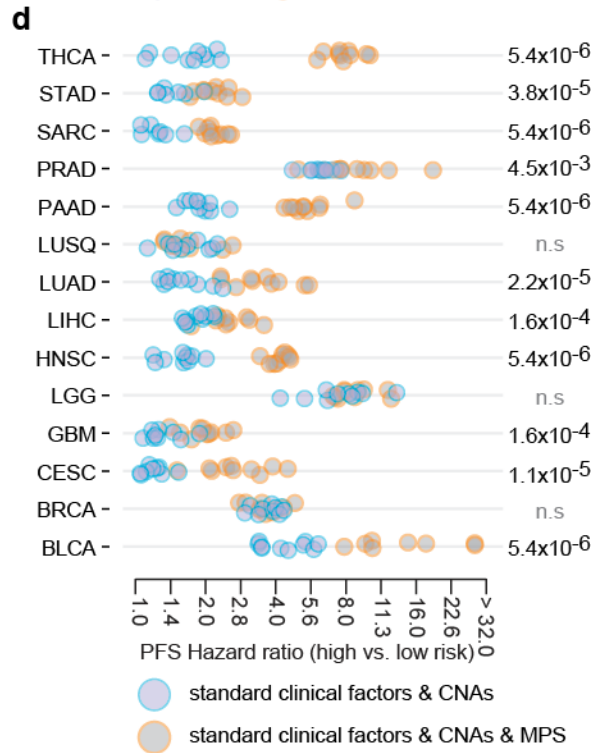
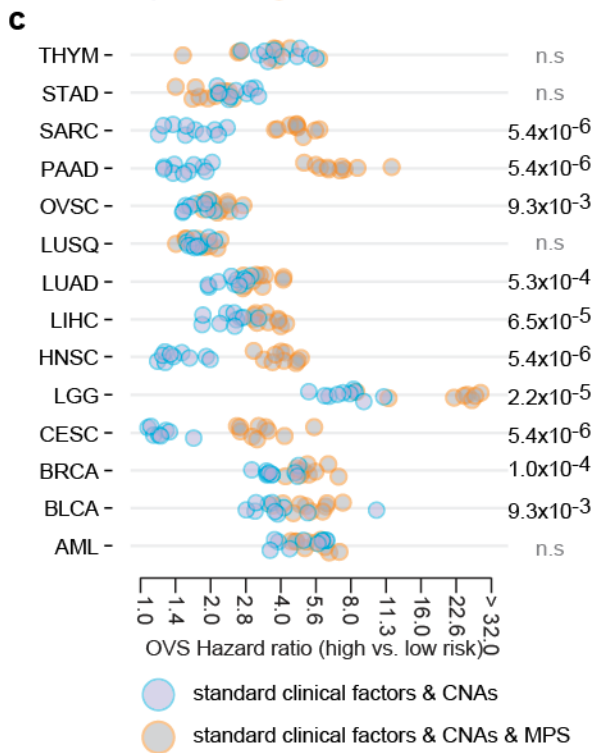


Figure S31. Random survival forests trained on prognostic modules provide significantly improved patient stratification compared to prominent CNAs and standard clinical factors, Related to STAR Methods.

(a-b) Performance of random survival forest models trained on prominent copy-number aberrations (CNAs) (open purple circles) or CNAs and module perturbation scores of PCMs in each cohort (purple dots with orange outline) for (a) overall survival (OVS) and, (b) progression-free interval survival (PFS). (c-d) Performance of random survival forest models are shown for models trained on prominent CNAs and standard clinical factors) with (gray dots with orange outline) or without (purple dots with blue outline) module perturbation scores for (c) overall survival (OVS) and, (d) progression-free interval survival (PFS). Hazard ratios were obtained using Cox proportional hazards model on patients stratified into high and low risk groups based on predictions from the random survival forest models¹⁵. Each dot corresponds to the predicted hazard ratio of a 10-fold cross-validated random forest model. For every cohort, distributions of hazard ratios obtained from 10 randomly initialized instances are shown. Comparisons between hazard ratios obtained from the two sets of random survival forest models are shown (p -values from one-sided Mann-Whitney test).

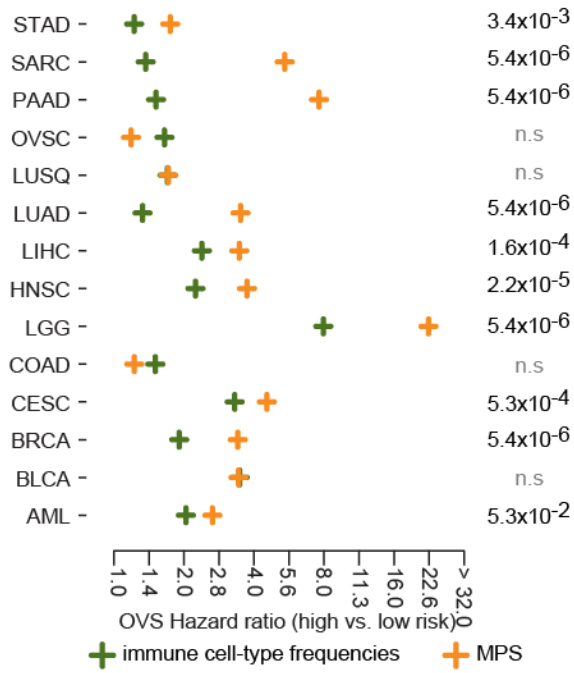
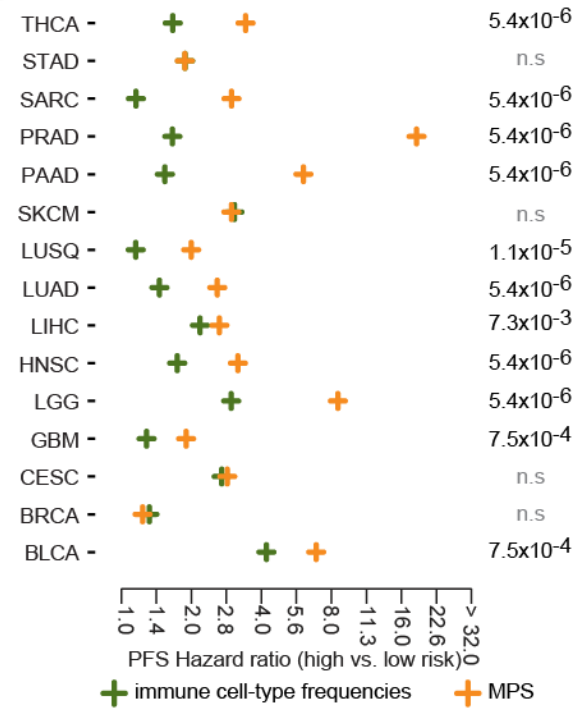
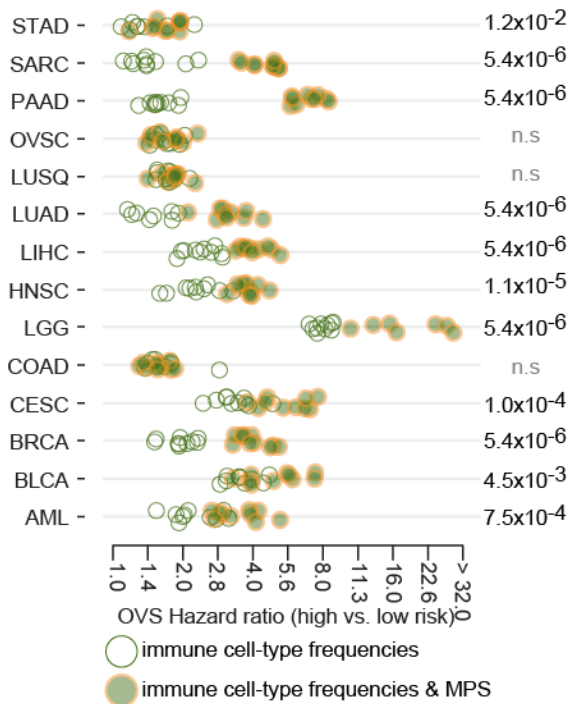
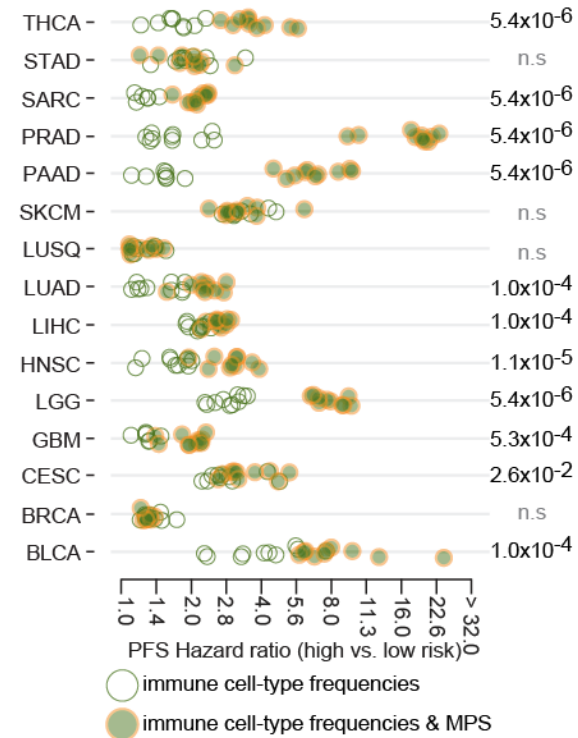
a**b****c****d**

Figure S32. Random survival forests trained on prognostic modules provide significantly improved patient stratification compared to inferred immune cell-type frequencies, Related to STAR Methods.

(a-b) Performance of random survival forest models trained on module perturbation scores of PCMs (orange) and inferred immune cell-type frequencies (dark green). For each cohort (rows), random survival forest models were trained in a 10-fold cross-validated setup for (a) overall survival (OVS) and (b) progression-free interval survival (PFS). Crosses correspond to the median hazard ratios calculated from 10 repetitions of the random forest predictions with randomized initializations. (c-d) Performance of random survival forest models are shown for random survival forest models trained on prominent inferred immune cell-type frequencies (open dark green circles) or inferred immune cell-type frequencies and PCMs (dark green dots with orange outlines) for (c) overall survival (OVS) and, (d) progression-free interval survival (PFS). Hazard ratios were obtained using Cox proportional hazards model on patients stratified into high and low risk groups¹⁵. Each dot corresponds to the hazard ratio of a 10-fold cross-validated random forest model. For every cohort, distributions of hazard ratios obtained from 10 randomly initialized instances are shown. Comparisons between hazard ratios obtained from the two sets of random survival forest models are shown (*p*-values from one-sided Mann-Whitney test).

References

1. Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* 1, 1–16. 10.1200/PO.17.00011.
2. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 1, 417–425. 10.1016/j.cels.2015.12.004.
3. Bryne, J.C., Valen, E., Tang, M.-H.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 36, D102-106. 10.1093/nar/gkm955.
4. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–177. 10.1038/nature12311.
5. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. 10.1016/j.cell.2014.08.009.
6. Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* 28, 337–350. 10.1016/j.molcel.2007.09.027.
7. Goodarzi, H., Najafabadi, H.S., Oikonomou, P., Greco, T.M., Fish, L., Salavati, R., Cristea, I.M., and Tavazoie, S. (2012). Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* 485, 264–268. 10.1038/nature11013.
8. Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* 110, 6388–6393. 10.1073/pnas.1219651110.
9. Martin, M.L., Zeng, Z., Adileh, M., Jacobo, A., Li, C., Vakiani, E., Hua, G., Zhang, L., Haimovitz-Friedman, A., Fuks, Z., et al. (2018). Logarithmic expansion of LGR5+ cells in human colorectal cancer. *Cell. Signal.* 42, 97–105. 10.1016/j.cellsig.2017.09.018.
10. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 1, 417–425. 10.1016/j.cels.2015.12.004.
11. Drier, Y. (2017). pathifier. 10.18129/B9.BIOC.PATHIFIER.
12. Frey, B.J., and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science* 315, 972–976. 10.1126/science.1136800.
13. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. 10.1038/nmeth.3337.

14. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8, R24. 10.1186/gb-2007-8-2-r24.
15. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S. (2008). Random survival forests. *Ann. Appl. Stat.* 2, 841–860. 10.1214/08-AOAS169.
16. Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). **NbClust** : An *R* Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* 61. 10.18637/jss.v061.i06.
17. Krzanowski, W.J., and Lai, Y.T. (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics* 44, 23. 10.2307/2531893.
18. Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods* 3, 1–27. 10.1080/03610927408827101.
19. Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. 10.1016/0377-0427(87)90125-7.
20. Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 63, 411–423. 10.1111/1467-9868.00293.